Supplementary Information

Genome-wide associations of human gut microbiome variation and causal inference analyses.

David A Hughes, Rodrigo Bacigalupe, Jun Wang, Malte C Rühlemann, Raul Y. Tito, Gwen Falony, Marie Joossens, Sara Vieira-Silva, Liesbet Henckaerts, Leen Rymenans, Chloë Verspecht, Susan Ring, Andre Franke, Kaitlin H. Wade, Nicholas J. Timpson, Jeroen Raes.

Distributions of 16S integer count data for mGWAS taxa

To evaluate the normality of 16S mGWAS taxa count distribution, we estimated the Shapiro's W-statistic, a measure of normality ranging from 0-1, where one is equal to a perfectly normal distribution, for the raw data distribution as well as rank normal transformed (RNT) and log2 transformations of the data. The rank normal transformation where first zero-truncated when the prevalence of 0 counts was greater than 5% in the population and, for the log transformation, all zero-values where removed. The average W-statistic for the raw data set was 0.488 with a range of 0.027 to 0.99. As seen in Supplementary Figure 1, the rank normal transformations in blue performed as expected in normalizing the distributions. Despite this, there were some instances where the distributions appeared to be bimodal. This is the product of an abundance of zero values below the 5% prevalence threshold we used to push a taxa through a hurdle binary (HB) analysis. In these instances, there were many zero or other common values in the data distribution that carry the same rank upon rank normal transformation. An option worth evaluating in the future is randomly ranking tied values, which would eliminate the bimodality of some distributions. However, it is worth considering the effect of randomly ranking tied values when the number of tied values is as large as those seen in zero-inflated microbiota abundances.

We chose to rank normal transform each continuous phenotype because it performed better at normalizing the distributions than log transformation, as seen in Supplementary Figure 2. In fact, in six instances, the log transformation made the distribution less normal than raw distribution as determined by a comparison of the W-statistics (Supplementary Table 17). Moreover, the decision on how to treat zero values is not straight forward for log transformations, as you might remove all zero values or simply add a constant to all observations. Regardless, we deemed the performance of log transformations on ecological count data, such as microbial 16S data, unsatisfactory for the analysis we aimed to carry out and moved forward with rank normal transformations.

Distribution / model choice

Each of the 139 mGWAS worthy raw abundance traits were tested with the R function fitdist() (package fitdistrplus) to find the best negative binomial, gamma, lognormal, and Poisson distribution parameters. We then identified the best fitting distribution, for each raw abundance distribution with a goodness of fit statistic (function gofstat()). Of the 139 traits,

10 were best explained by a gamma distribution, 67 by a lognormal, and 62 with a negative binomial. While these distributions may model the raw data the "best" it does not mean it fits the data well. At the onset of this study we ran 10 MT GWAS using a negative binomial distribution. Six of those 10 MT are "best" explained by a negative binomial (NB) distribution. The taxa run through a negative binomial GWAS are (1) C_Actibobacteria [genomic inflation or λ =1.709], (2) F_Coriobacteriaceae [λ = 1.679], (3) F_Sutterellaceae [λ = 1.092], (4) F_Veillonellaceae [NB; λ = 1.46], (5) G_Alistipes [NB, λ = 1.583], (6) G_Anaerostipes [NB; λ = 1.494], (7) G_Blautia [λ = 1.485], G_Clostridium_IV [NB; λ = 2.2], (9) G_Dorea [NB; λ = 1.167], and (10) F_Bacteroidaceae [NB; λ = 0.9193]. We quickly noticed that the NB distribution was insufficient in its ability to fit the raw data distributions of the FGFP MT data, specifically the skewness of the distribution tails, as highlighted by observed inflated p-values and genomic inflation values (λ), as reported above. Given the numerous distributions that may be needed to model all of the MT in this study, the insufficiency of the NB distribution to model that data in hand, and the fact that logging the data often made distribution less normal than the raw data, we chose to rank normal transform all MT and run linear models in this study.

Heritability

Estimates of "chip-based" heritability was carried out as described in Methods, using GCTA; however, for the purposes of comparing heritability estimates across studies, we performed estimates for not only our RNT but also log2 and box-cox transformations, as used by other studies like Goodrich *et al*[2]. As seen in Extended Data Fig. 1a and 1b, the choice of data transformation (RNT, Box-Cox, or log) can influence on the estimation of heritability, even within the same data set, making comparison across studies even more complicated. Yet, we compare our FGFP heritability estimates with those published by Goodrich *et al.* (Extended Data Fig. 1c) and Davenport *et al.* (Extended Data Fig. 1d). As Goodrich *et al.* used Box-Cox transformation we compare their estimates to our box-cox transformed data estimates, and as Davenport used quantile-quantile normalizations we chose to compare it to our rank normal transformed data. The data used to generate Extended Data Fig. 1 can be found in Table S3. Between FGFP box-cox transformed data and Goodrich *et al*. (2016) four microbial traits have estimates of heritability whose confidence intervals do not overlap with zero in both datasets (Extended Data Fig. 1c). They are G_Bifidobacterium, G_Roseburia,

G_Faecalibacterium, and G_unclassified_ F_Ruminococcaceae. Between FGFP and Davenport *et al.* (2015) the only microbial trait that appears to have an estimate of heritability whose confidence intervals do not overlap with zero in both datasets is C_Gammaproteobacteria, with an estimate of 0.332 (se = 0.178) in FGFP and an estimate of 0.373 (se = 0.249) in Davenport *et al*. (2015).

FGFP sample description

The current Flemish Gut Flora Project (FGFP) 16S fecal microbiome dataset is composed of 2482 individuals, collected in the Flanders region of Belgium. Of those, 2259 individuals also had genotype data and will be described here. A total of 1347 individuals are genetically predicted to be female, 912 as male, with an overall average age, height, weight, and body mass index (BMI) of 52.3 years, 170.30 cm, 72.97 kilograms, and 25.09 kg/m^2, respectively. Average and 95% confidence intervals (CI) for age (years), height (cm), weight (kilograms), and BMI (kg/m^2) are partitioned sex and presented below in Supplementary Table 6, and visually in Supplementary Figure 3.

FGFP 16S microbiome data description

The V4 region of the 16S rRNA was amplified using the 515F/806R primer pair (GTGYCAGCMGCCGCGGTAA and GGACTACNVGGGTWTCTAAT). The DADA2 pipeline was used to partition rarefaction counts (10,000) into taxonomic units. A total of 499 taxonomic units were identified in the FGFP dataset, one (G_Syntrophococcus) exhibited no variation was removed from further analysis, which was composed of 18 phyla, 35 class, 53 order, 105 family, and 287 genera. The most abundant phyla, on average, were Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria, each making up, on average, greater than 1% of the fecal microbiome. Extended Data Fig. 4a presents the 18 phyla by the rank order of their average abundance. Similarly, the only phyla present in greater than 99% of all sampled individuals was Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria. The rank order of prevalence for all phyla is presented in Extended Data Fig. 4b.

Quality control of microbiome individual-level data was performed by estimating an initial non-metric multiple dimension scaling (MDS) using the isoMDS() function and Bray-Curtis

distance derived from the vegdist() function of the vegan package in R, and rarefaction genera-level data. Two individuals presented extreme microbiome profiles (> 5 standard deviation from MDS1 or MDS2 means; Supplementary Figure 4) and were excluded from all further analysis.

α- and β-diversity metric estimates as well as enterotyping was performed using all rarefaction genera-level data as described in online methods. In Extended Data Fig. 4c, we present the 2-axis inter-individual β-diversity, or specifically the Bray-Curtis non-metric multiple dimension scaling (nMDS) estimates in relationship to the enterotype calls derived from the Dirichlet multinomial mixtures (DMM) method. The two measures are strongly related as illustrated in both the nMDS and the boxplots of each dimension (Extended Data Fig. 4c).

Microbiome genome-wide association studies (mGWAS) taxa, or those taxa we identified for carrying forward into association analysis with genotypes, were identified following two criteria. Firstly, the taxa needed to be present in at least 15% or 339 of the 2259 sampled individuals. Second, the taxa needed to make up a reasonable (5%) proportion of the rarefaction counts in at least one of the sampled individuals. Previously, we had defined genome-wide association study (GWAS) taxa, called core taxa at the time, as those with an average abundance of 40 reads in rarefaction (10,000 reads) data[1] (Extended Data Fig. 5a). We effectively expanded upon this single criterion to include more taxa, specifically those that may be considered to be prevalent (making up at least 5% of the rarefaction data) in at least one individual. The relationship between average abundance in the population, prevalence in the population, and proportional composition within an individual is presented in Extended Data Fig. 5a. It is worth noting that some taxa are highly prevalent in the population but have both a low average abundance and do not make up a substantial proportion of reads in any one individual.

Among all taxon, there are 19 genera, 8 families, 6 orders, 6 classes and 4 phyla that may be considered core microbiome, defined as those with a population-level prevalence of at least 95% (Supplementary Table 15). Among the genera in Supplementary Table 15, only G_Intestinimonas (indicated with "**") was not included in the mGWAS, because it did not make up at least 5% of any one individual's microbiome, consistent with its low mean abundance (Supplementary Table 16).

Those taxa exhibiting high prevalence, but low abundance, and never making up at least 5% of an individual's microbiome (among those in Extended Data Fig. 5a plotted in red) are listed below in Supplementary Table 16. Their high prevalence makes them of interest, but their low abundance makes their accuracy and precision of quantification in rarefaction data low, as described by Wang *et* al.

Among the 139 taxa that met our criteria as mGWAS taxa, there are statistical redundancies in the form of taxon from two or more taxonomic levels that were highly correlated. To illustrate this, we calculated a correlation distance, defined as 1-Spearman's rho among all 139 taxon in a pairwise fashion using the function cor() and flags, method = "sp", and use = "pairwise.complete.obs", and then generated a clustering dendrogram with the function hclust() and flag method = "complete" in R (Extended Data Fig. 5b). We defined statistical redundancy across taxa as those pairs with a Pearson's rho greater than or equal to 0.985. This is illustrated with the red horizontal line in Extended Data Fig. 5b at a value of 0.015.

The level of correlation among the mGWAS taxa is further illustrated in Extended Data Fig. 5c with a correlation plot of Spearman's rho created with the corrplot() function from the corrplot package for R. To aid in the identification of possible clusters, eight cluster groups where drawn in boxes with the corrplot flag "addrect = 8". The clusters minimize the sum of square difference within a cluster and maximize those differences among taxa. Eight clusters are illustrated because there are eight phylum ("P_*") level taxa among the mGWAS taxa.

The taxa retained for the mGWAS were simply identified by working from the genera-level data and going up the taxonomic levels to identify any other taxon that was highly correlated to a lower-level taxon. When the correlation coefficient was greater than or equal to 0.985, we removed the higher taxa from the list of taxa to include in the mGWAS.

Batch variables influence microbiome variation

A total of eight batch related variables were available, to evaluate variation in count data that may have been introduced during laboratory procedures. Those variables included (1) the extraction type, (2) extraction date, (3) who the extraction was performed by, (4) the aliquot date, (5) who performed the aliquot, (6) the PCR plate ID, (7) the position in the plate and (8) the library production data. In a univariate analysis against each of the 92 RNT mGWAS taxa (with zero-truncation when the frequency of zeros in the population exceed 5%) each of these variables was associated with abundance variation at a false discovery rate (FDR; Benjamini-

Hochberg corrected p-value) of 5%. Extraction type influenced 35 taxa, extraction date 9, extraction by 11, aliquot date 42, aliquot by 44, PCR plate 42, position 1, and library date 21 at an FDR of 5% or less. The average variance explained across all 92 abundance traits is presented in Supplementary Figure 5.

Position, and library date had numerous levels containing only one or a few individual samples, and the person performing the extraction was strongly structured by extraction date, leading us to believe that including these as covariates may be over adjusting the data. As such, we chose to include extraction type, extraction date, aliquot date, the person performing the aliquot, and the library PCR plate as covariates in all association analysis with genotypes.

Heritability power estimates

GCTA-GREML (genome wide complex trait analysis – genomic-relatedness-based restricted maximum-likelihood) genetic (co)variation power estimates were calculated using scripts provided on the GCTA website (http://cnsgenomics.com/software/gcta/#GREML powercalculator, Extended Data Fig. 7)**.** When estimating the power of quantitative traits, we iterated over sample sizes of 300 to 2300 at steps of 300, with narrow sense heritability set from 0.01 to 0.7 at steps of 0.01. As such we modified the parameters "n" and "hsq" in the provided R function calcUniQt(). To estimate power for our presence/absence microbial traits we used the provided function calcUniCc(), and modified the parameters "cases", "controls", "hsq", "K", where cases and controls are the number of individuals where the microbe is present and absent, hsq is the heritability estimate and K is the prevalence of the microbial trait in the population. As with the quantitative trait we looped over heritability estimates from 0.01 to 0.7 at steps of 0.01. Prevalence (K) of the microbial trait ranged from 0.15 to 0.5 at steps of 0.05. Cases where then defined as our sample size times prevalence in the population, or 2300 * K, and the sample size of controls was set as 2300 – cases. Alpha was set to 0.05 for all estimates.

Influence of genotype uncertainty on mGWAS parameter estimates

To account for genotype uncertainty in the imputed genotype data, we used the score method, or missing data likelihood score test, as implemented by SNPTEST[3]. This method is computationally fast and is designed to revert to the em or expectation-maximization method

when the score test performs poorly. In checking this, beta estimates derived from the score and em methods were highly correlated, with deviations at the tails (Supplementary Figure 8). We used SNPTEST's score-based method as an initial screening and ranking of variants to carry forward into a targeted meta-analysis in the FoCus and PopGen studies which was then based on the slower but more stable em method. This two-step process (designed to illustrate the feasibility of host/microbiome GWAS) explains the unusual distribution of p-values in the meta-analysis results presented in Figure 2c (as this is a targeted meta-analysis).

16S prevalence across cohorts

In this study, we were able to formalize analytical methods across studies but not laboratory protocols. Most importantly, the variable regions targeted between FGFP and the German studies varied. The FGFP study targeted V4 of the 16S rRNA locus, while FoCus and PopGen targeted V1-V2. As a product of this, there is variability in the microbiome variation across cohorts (Supplementary Figure 9). As a product of this, three taxa that were analyzed in FGFP mGWAS were not available for analysis in FoCus or PopGen. They are *Escherichia Shigella*, *Hespellia*, and *Methanobrevibacter.*

Inter-study Catalog

Starting with the supplementary material provided by Rothschild *et al.* (2018), we compiled a table of all of the previously reported mGWAS associated variants and their associated microbial trait (Table S7). We pruned via clumping all variants associated to the same traits but sitting in LD with each other, retaining the association with the smallest p-value. We then add to this data table the FGFP effect estimates (snptest: -method score), for the similar taxonomic trait or when not available one from a higher taxonomic level.

Of the 44 reported variants associated with beta-diversity in previous studies, we had estimates on 35 of them. However, none of these previously reported associations had a p-value less than 0.05 in our MANOVA analysis.

A total of 591 genetic variants have been reported to be associated with microbial abundance in previous studies. We identified, in FGFP, estimates for the same SNP - trait pairs (or higher order taxon) in 545 instances. These 545 SNP – trait pairs involve 492 unique SNPs (rsids). In total 30 SNP – trait pairs exhibit a p-value less than 0.05 in the FGFP (snptest: -

method score) data set. This represents just 5.5% of the tested pairs, a value similar to the randomly expected 5% assuming uniformity in the p-values. However, if we place a Bonferroni correction on the number of tests we are making here (0.05/492 = 1.016 x 10-4) the only associations that pass our Bonferroni p-value threshold are those between *Bifidobacterium* and variants within the block of LD around the *MCM6* (*LCT*) selected locus. This includes the variant (rs6730157) within *RAB3GAP1* (FGFP: beta = 0.145, se = 0.032, p-value = 6.91 x 10^{-6}), the variant with the strongest signal at this locus in Goodrich *et al.* (rs1446585) and sitting within *R3HDM1* (FGFP: beta = 0.120, se = 0.032, p-value = 5.60 x 10^{-5}), and the presumed lactase-persistent variant (rs4988235; FGFP: beta = -0.147, se = 0.033, p-value = 7.86 x 10^{-6}), reported by both Goodrich *et al.* (2015) and Bonder *et al.* (2016) (Supplementary Figure 10). Finally, we note that the associations reported by Blekham *et al.* (2015), using exonic variants within *LCT* and the neighbouring gene *UBXN4* were not replicate in FGFP (Tables S7-S9).

Interestingly, an association reported by Davenport *et al.* (2015) between the chromosome nine variant rs7868228 and family_Succinivibrionaceae in a winter sampling (Davenport: beta = -0.958, se = 0.195, p-value = 3.85 x 10^{-6}) fell just below our Bonferroni p-value threshold in a higher taxonomic level, C_Gammaproteobacteria, in FGFP (beta = -0.142, se = 0.042, p-value = 7.76 x 10^{-4}). However, given that the direction of effect was also consistent we thought it worth noting here.

As performed by Goodrich *et al.* (2016), in their Figure 3b, rather than attempting to compare the exact SNP-trait associated pair, they queried for the strongest association among all tested traits for each previously reported SNP. Effectively asking if previously reported variants associate with any microbial trait available in the new data set. We repeated this procedure here (Supplementary Figure 11) creating a table of 522 unique SNP – top microbiota traits (Tables S8). Using a Bonferroni corrected p-value threshold of 9.578 x 10^{-5} we once again identified Bifidobacterium for the *LCT* locus, an exact SNP-trait match as reported above, but also two other loci. The first, reported by Bonder *et al.* (2016) between genera *Lactococcus* and variant rs2294239, on chromosome 22 within the gene ZNRF3 (beta = -0.186, p-value = 2.92 x 10-6) - is associated with G_unclassified_P_Bacteroidetes_RNT in the FGFP (-method score) data set with estimates: beta = 0.225, se = 0.057, p-value = 7.11 x 10-5. This variant has also been previously reported in a GWAS for interaction between physical activity and adiposity in a multi-ethnic meta-analysis (Graff *et al*., 2017). The second, reported by Goodrich *et al.* (2016) between *Anaerostipes* and the SNP rs10233359 which is

an intergenic SNP on chromosome 7 (beta = -0.211, se = 0.036, p-value = 4.94 x 10-9) near the gene FOXP2 and GPR85. In the FGFP (-method score) data set this SNP is associated with another genera in the Lachnospiraceae family, *Roseburia* (beta = -0.224, se = 0.056, p-value = 5.84 x 10-5).

Supplementary Tables

sex	N	Age	Height	Weight	BMI	W/H	LDL
Female	1347	50.54	165.86	67.41	24.48	0.87	119.04
		(23.21-71.00)	(153.00-180.00)	(49.00-97.73	(18.08-34.68)	(0.70-1.12)	(62.0-193.0)
Male	912	54.89	176.43	81.02	26.03	0.95	120.53
		(26.54-74.55)	(162.00-190.00)	58.00-114.00	(19.37-35.31)	(0.78-1.11)	(65.0-188.35)

Supplementary Table 6: FGFP cohort description. Average and 95% confidence intervals for age, height, weight, body mass index (BMI), waist-hip ratio (W/H) and LDL cholesterol, partitioned by genotype predicted sex in the FGFP cohort.

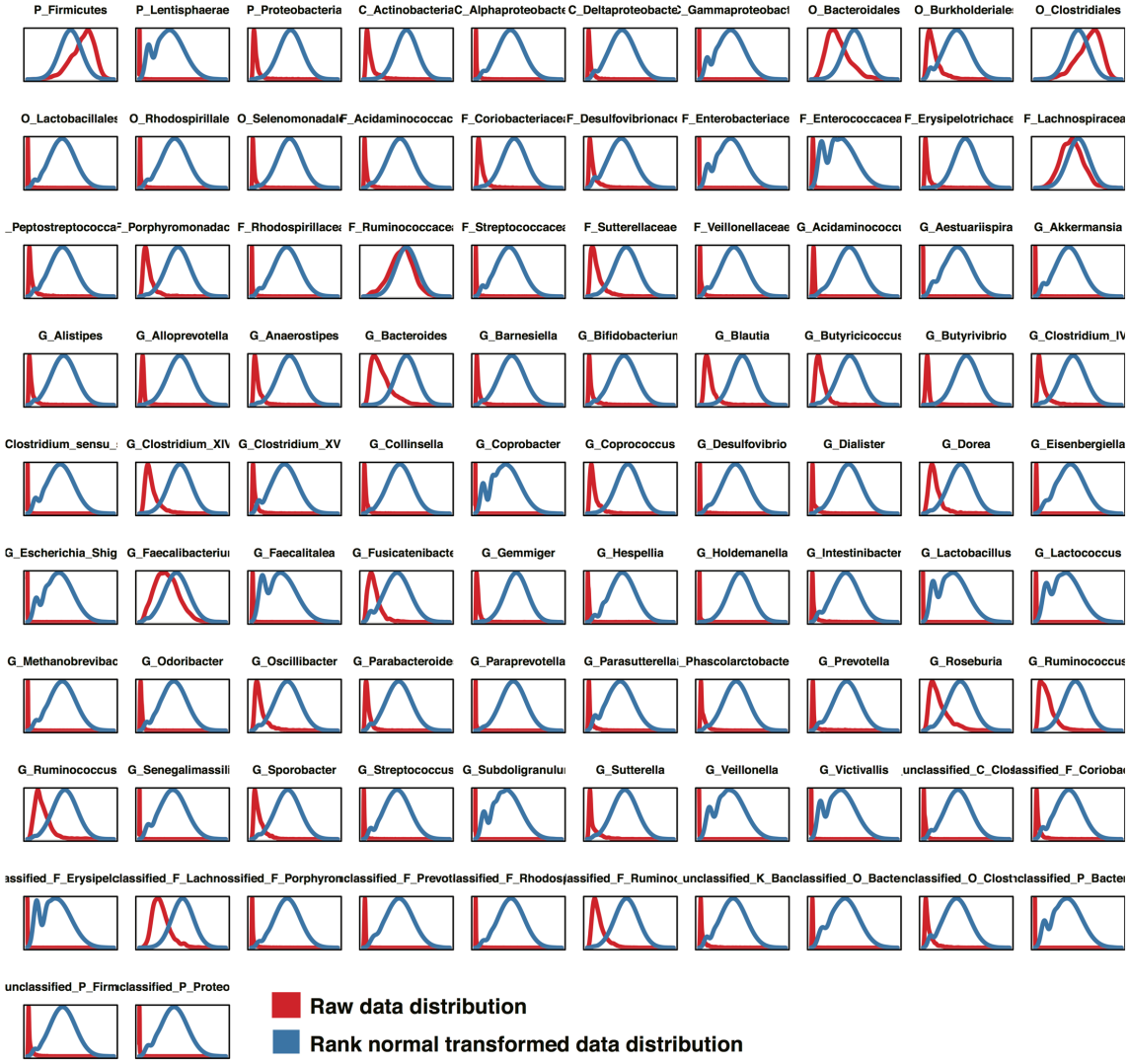
P_Actinobacteria	P_Bacteroidetes	P_Firmicutes	P_Proteobacteria
C_Actinobacteria	C_Bacteroidia	C_Betaproteobacteria	C_Clostridia
C_Erysipelotrichia	C_Negativicutes	O_Bacteroidales	O_Burkholderiales
O_Clostridiales	O_Coriobacteriales	O_Erysipelotrichales	O_Selenomonadales
F_Bacteroidaceae	F_Coriobacteriaceae	F_Erysipelotrichaceae	F_Lachnospiraceae
F_Porphyromonadaceae	F_Rikenellaceae	F_Ruminococcaceae	F_u_O_Clostridiales
G_Alistipes	G_Anaerostipes	G_Bacteroides	G_Blautia
G_Butyricicoccus	G_Clostridium_IV	G_Clostridium_XlVa	G_Clostridium_XVIII
G_Dorea	G_Faecalibacterium	G_Fusicatenibacter	G_Intestinimonas**
G_Oscillibacter	G_Parabacteroides	G_Roseburia	G_Ruminococcus2
G_u_F_Lachnospiraceae	G_u_F_Ruminococcaceae	G_u_O_Clostridiales	

Supplementary Table 15: Core microbial taxa. A table of taxon that may be defined as core taxa, given a prevalence value of 95%, or being present in 95% of all individuals sampled in FGFP. Colors represent taxonomic ranks (genus in blue, family in orange, order in yellow, class in green and phylum in grey).

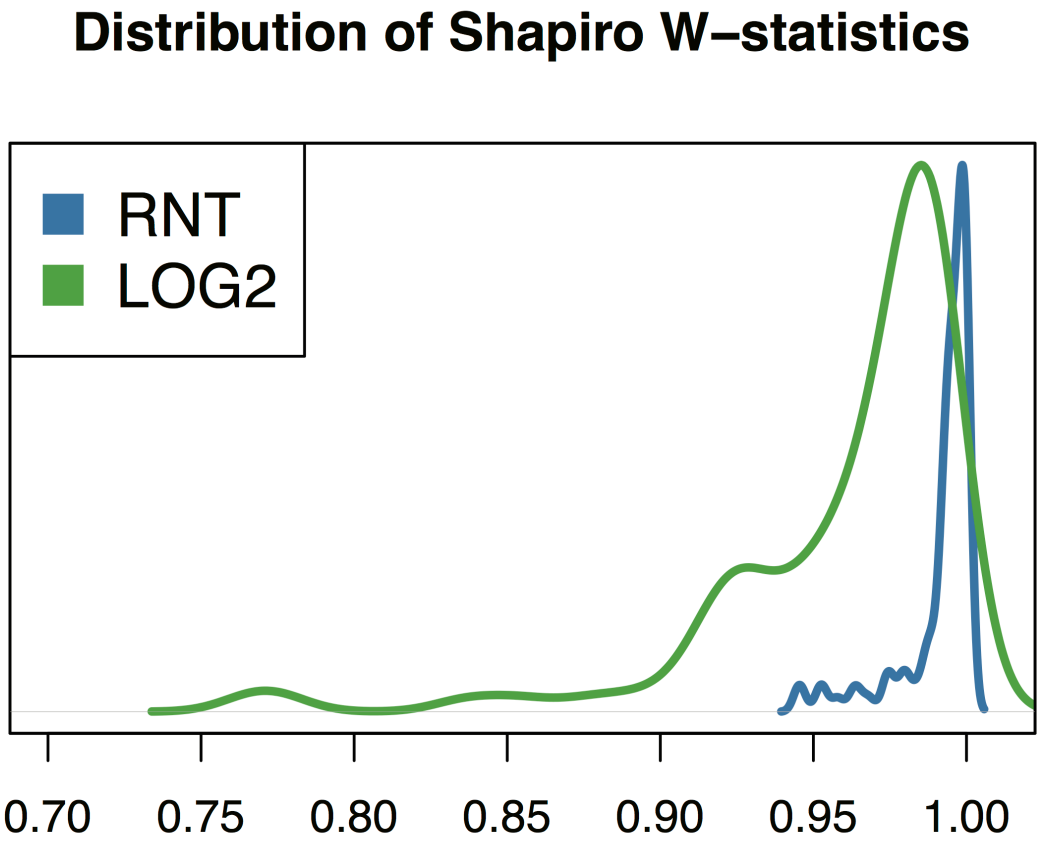
taxon	Avg. Abundance	log10(AvgAb)	Prevalence
G_Intestinimonas	18.599468	1.2695005	0.9525919
G_Clostridium_XlVb	19.887461	1.2985793	0.8825875
F_Clostridiales_Incertae_Sedis_XIII	14.54763	1.1627922	0.8475853
G_Butyricimonas	14.49136	1.1611092	0.8028356
G_Bilophila	13.810811	1.1402192	0.7970758
G_Romboutsia	17.263181	1.2371208	0.7620735
G_Flavonifractor	12.002658	1.0792774	0.7603013
G_u_F_Clostridiales_Incertae_Sedis_XIII	7.396101	0.8690028	0.6951706
G_Adlercreutzia	10.73992	1.0310011	0.6597253
G_Mogibacterium	6.733274	0.8282263	0.5897209

Supplementary Table 16: Prevalent but low abundant taxa. Taxa with prevalence over 50%, but average abundance levels below 40, a value previously shown to indicate poor replicability in rarefaction data.

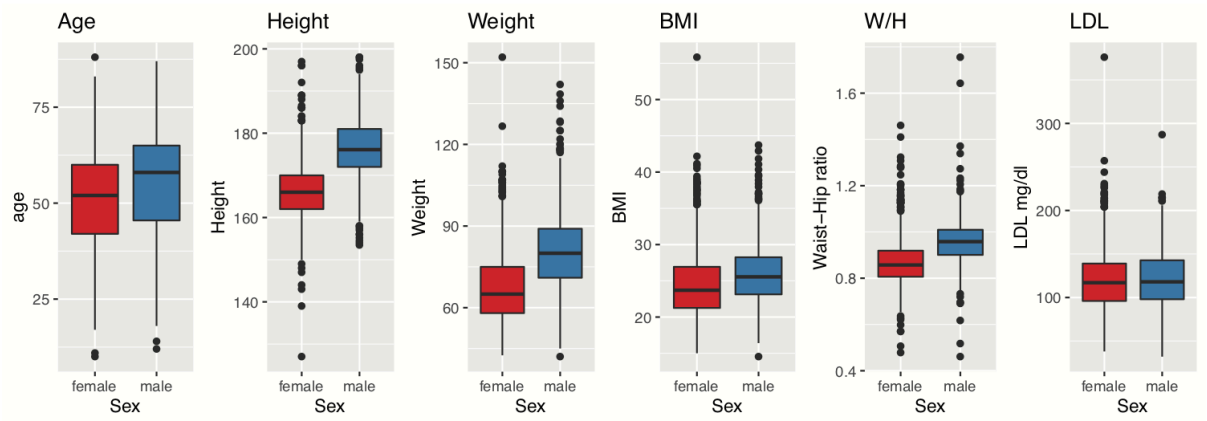
Supplementary Figures



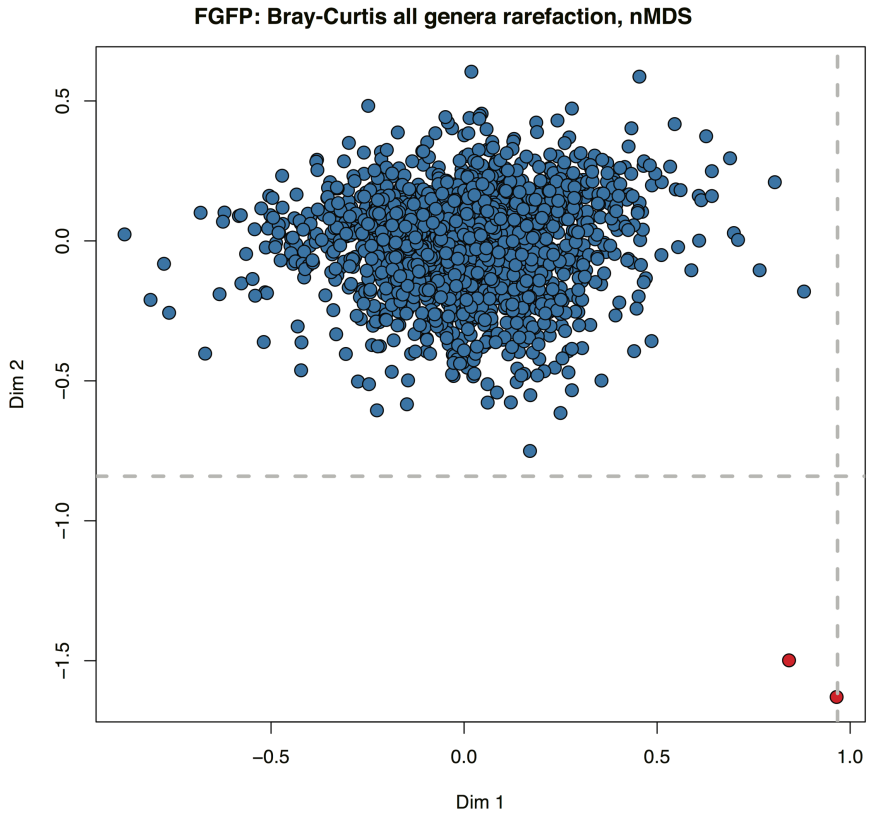
Supplementary Figure 1. Raw and RNT data distributions. Density plots of each taxa used in the mGWAS. Each plot illustrates the raw (red) and rank normal transformed (blue) data distributions. Data used to generate these plots can be found in Table S2.



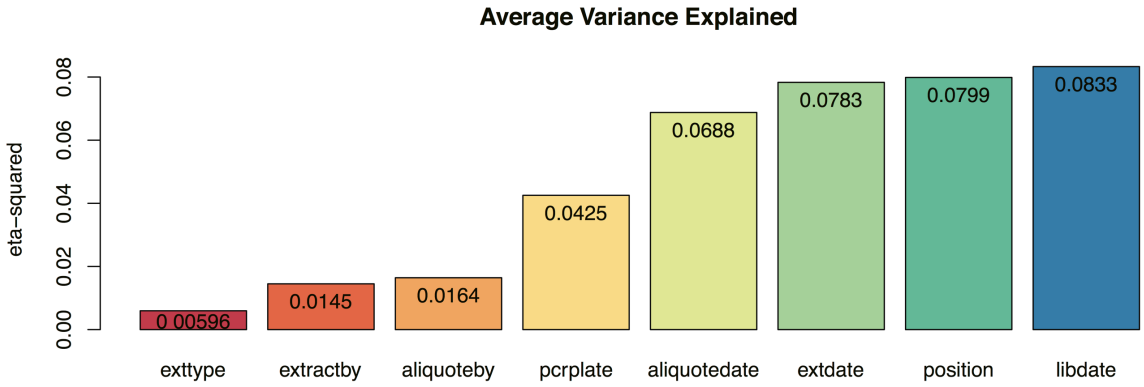
Supplementary Figure 2: RNT and log2 W-statistic estimates. Density plot of the Shapiro W-statistics for rank normal transformed (blue) and log2 transformed (green) abundance data for all taxa used in the mGWAS. The Shapiro W-statistic ranges from 0, uniform, to 1, perfectly normal. The log2 data deviates more strongly from normality. Data used to generate this plot can be found in Table S17.



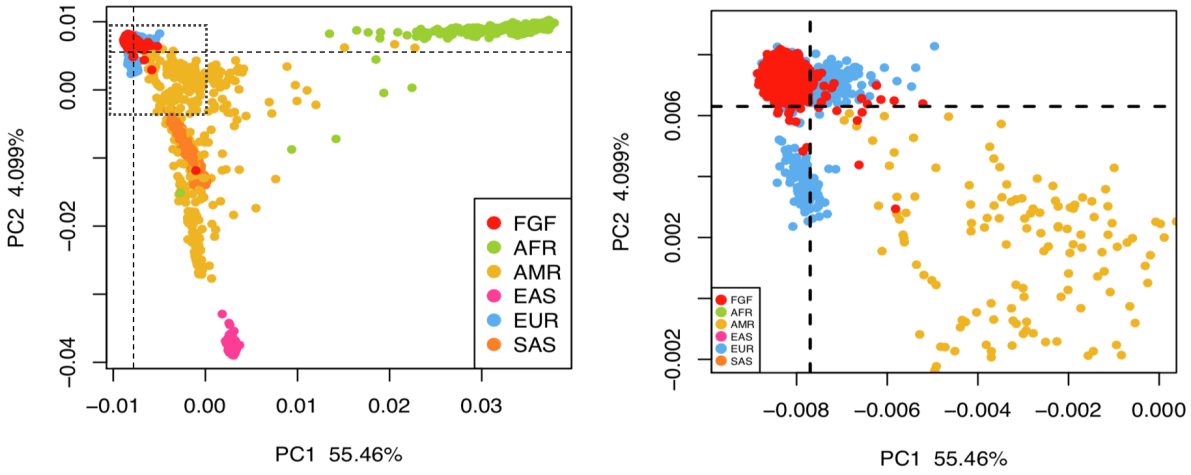
Supplementary Figure 3: Boxplot illustration of FGFP cohort. Boxplots illustrating the distribution of age, height, weight, body mass index, waist-hip ratio and LDL cholesterol in FGFP cohort, partitioned by genotype predicted sex. There are n = 1347 females and n = 912 males in the study cohort. Each box plot presents the mean, first and third quantiles, and 95% confidence intervals of the data distribution. Individual level data used to generate this plot can be accessed upon request from Dr. Jeroen Raes.



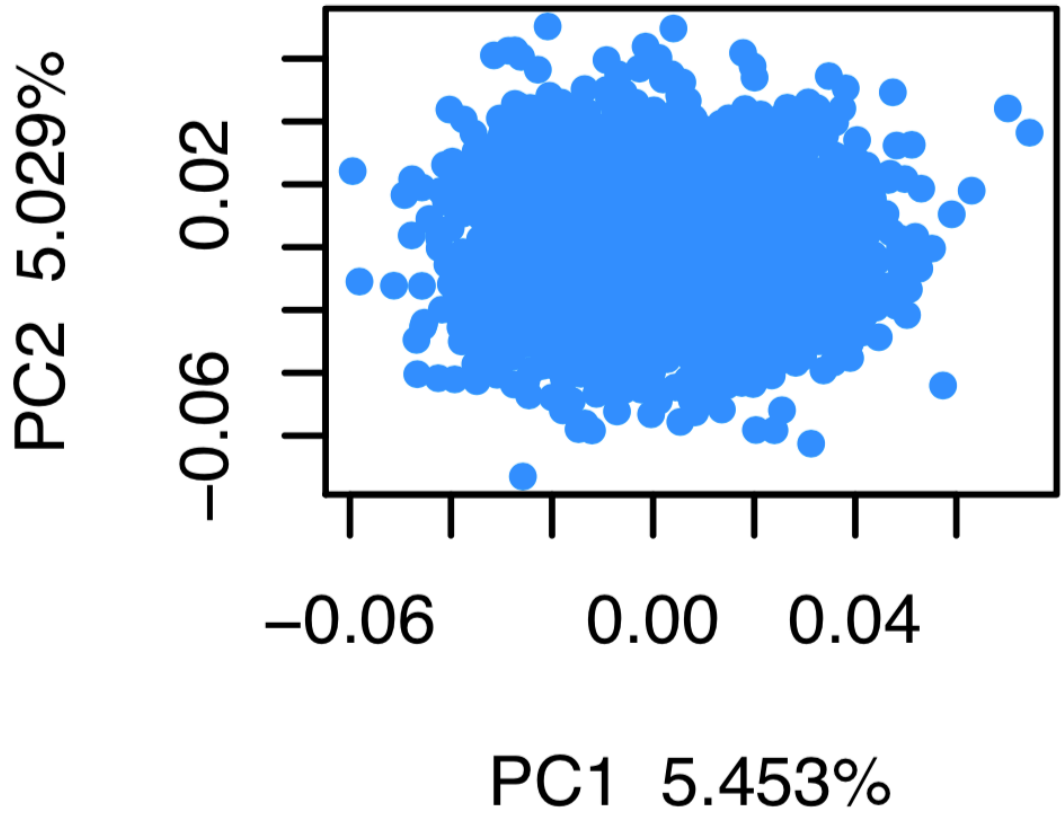
Supplementary Figure 4: nMDS quality control plot identifying outliers. Individual level β-diversity, or non-metric multiple dimension scaling plot of FGFP individuals. The two individuals in red, where identified as outliers and removed from all subsequent analysis. Dashed grey lines indicate five standard deviation from mean estimates on dimension 1 (vertical) and dimension 2 (horizontal). These are the thresholds for identifying spurious outliers. Data used to generate this plot can be found in Table S2.



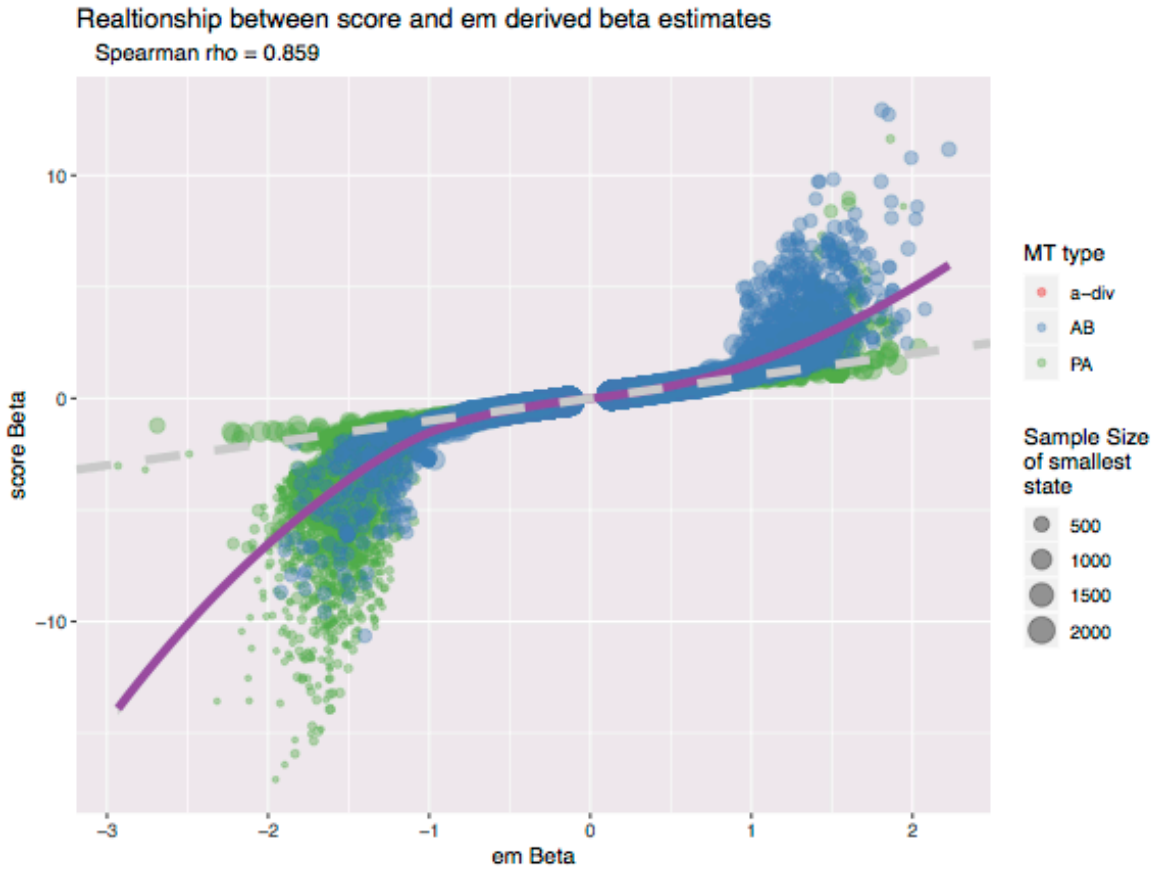
Supplementary Figure 5: Average variation explained by batch variables. Bar-plot for the average variation explained (eta-squared on y-axis), across abundance mGWAS microbial traits, by each batch variable. The variable across the x-axis are extraction type, extraction performed by, aliquot performed by, PCR plate, aliquot date, extraction date, position on PCR plate, and library construction date.



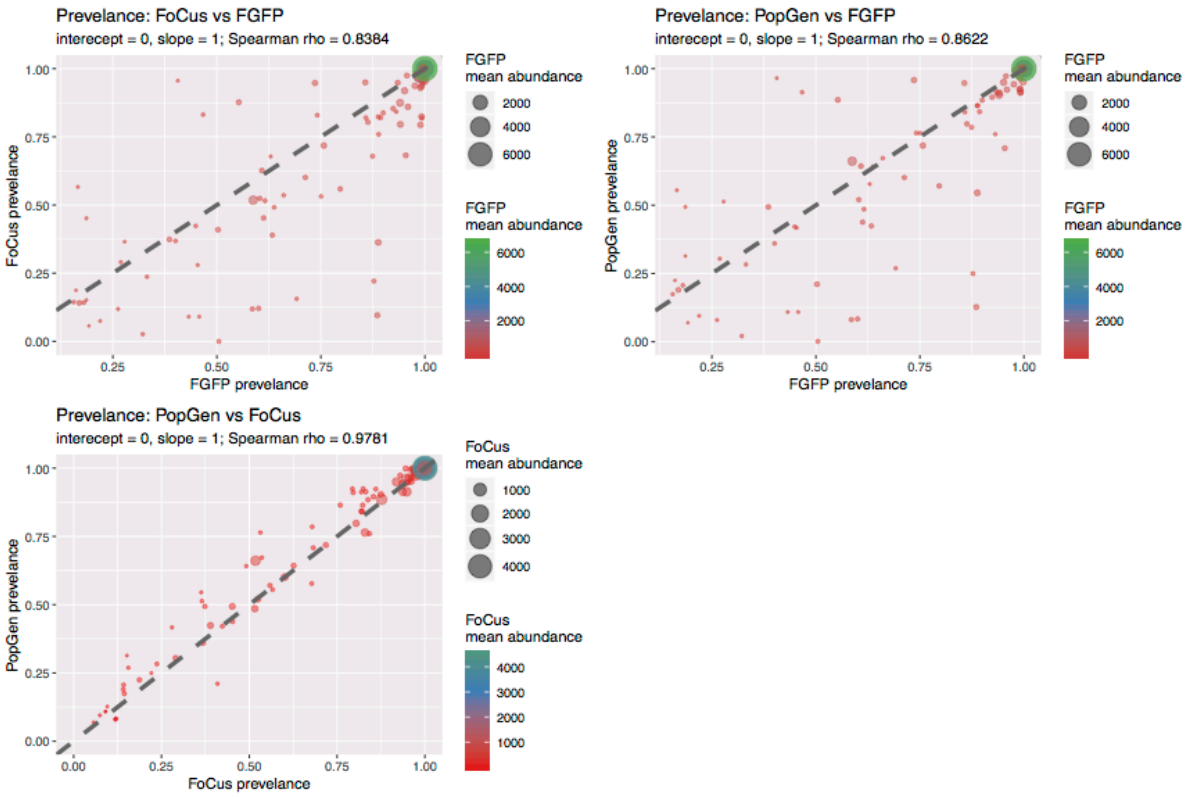
Supplementary Figure 6: FGFP and 1000 Genomes population principal components. Principal component analysis results of array genotype data from (N = 2646) FGFP individuals (red) merged and plotted with the (N = 2504) 1000 genomes data set. The plot on the left illustrates all data points, while the plot on the right is a zoom, of the upper left corner square in the plot on the left, in on the FGFP cohort samples relative to other continental Europe (EUR) samples. Labels are FGFP (FGF), African (AFR), East Asia (EAS), Europe (EUR), and South Asia (SAS). The numerical values on x- and y-axis labels are the proportion of variance explained by that axis. Dashed lines represent PC1 (vertical) and PC2 (horizontal) boundaries for individuals of Western Europe ancestry in the 1000 Genomes data set and where used to identify individuals in the FGFP cohort that were excluded from the GWAS for population structure.



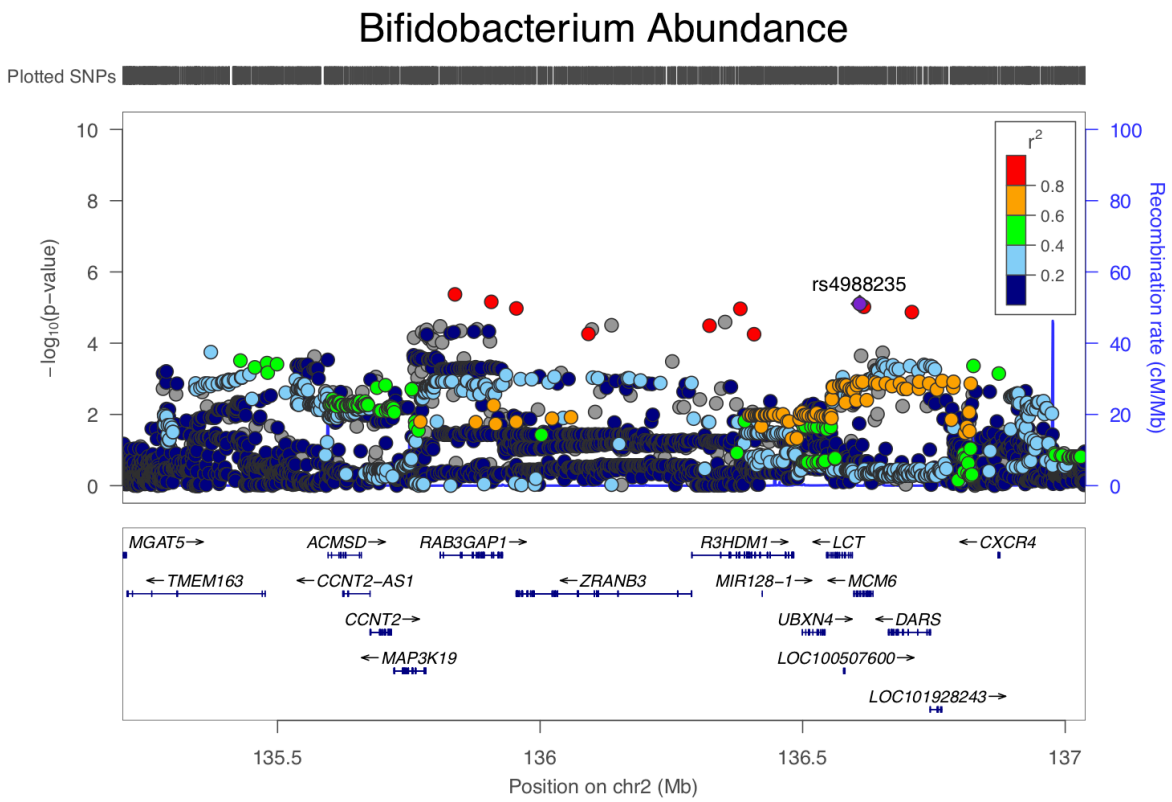
Supplementary Figure 7: FGFP mGWAS principal components. Principal component analysis of the quality controlled (N = 2257) FGFP cohort samples. The numerical values on x- and y-axis labels are the proportion of variance explained by that axis.



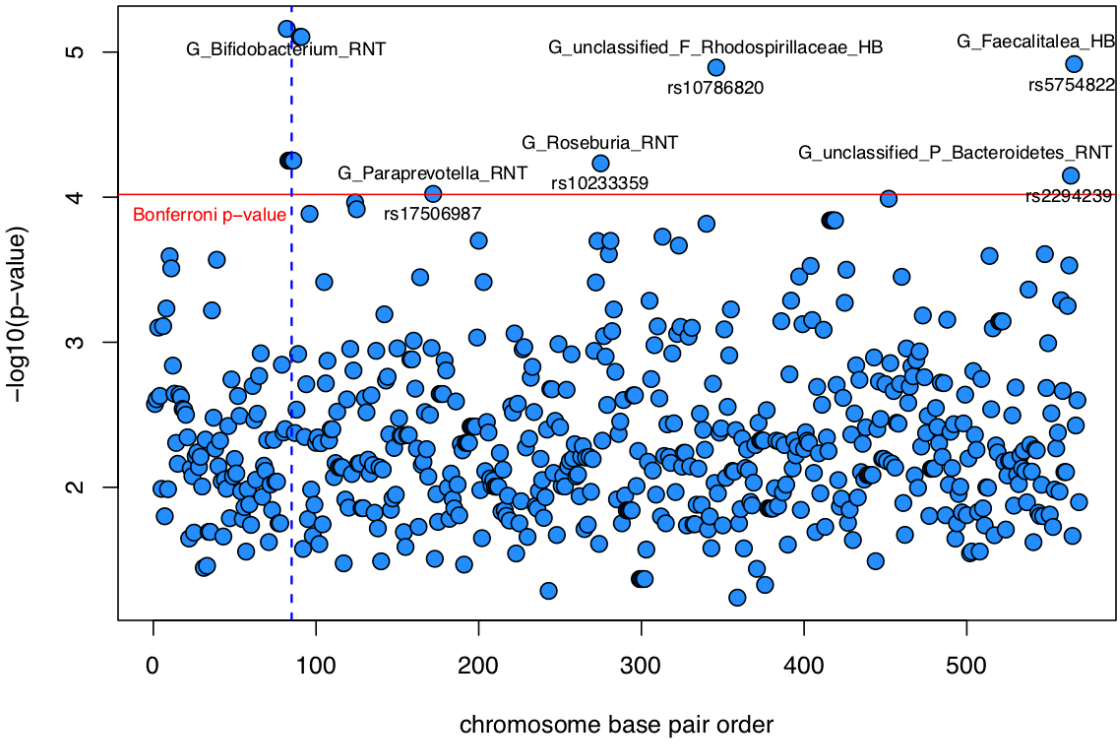
Supplementary Figure 8: Comparison of FGFP beta estimates derived from em and score methods. Scatter plot illustrating the relationship between effect estimates (beta) derived from em (x-axis) and score (y-axis) based methods used to account for genotype uncertainty in imputed data. The size of each point is an indication of (n) the sample size for that SNP-microbial trait analysis, and the color of the dot indicates if it is an abundance (AB), presence/absence (PA) or alpha-diversity microbial trait. All specific microbial trait sample sizes can be found in Table S3. The purple line is a loess curve fit through the data. The grey dashed line has a slope of one and a y-intercept of 0. An estimate of the Spearman's rho correlation coefficient is in the subtitle of the plot.



Supplementary Figure 9. Correlation of prevalence across the three study populations. Scatter plot of taxa prevalence between FGFP (n = 2257) and FoCus (n = 950) (top left), FGFP (n = 2257) and PopGen (n = 717) (top right), and PopGen (n = 717) and FoCus (n = 950) (bottom left). The size and color of each dot is an indication of mean abundance. The grey dashed line is a line with a slope of one and a y-intercept of 0. Data used to generate this plot can be found in Table S18.



Supplementary Figure 10: Locus Zoom plot of the *MCM6*, *LCT* and *RAB3GAP1* region. A locus zoom plot (http://locuszoom.org) of two-sided F-test association p-value with *Bifidobacterium* abundance (n = 1975) on chromosome 2 at the *RAB3GAP1*/*LCT*/*MCM6* locus. Along the x-axis is the -log10 p-values and the y-axis is the base position along chromosome 2. Gene models are below the scatter plot aligning genes to the p-value estimates above. The color coding in the scatter plot indicates the LD (r^2) estimates relative to the tagged variant (rs4988235) in purple. Note however that the variant with the smallest p-value is within the *RAB3GAP1* locus.



Supplementary Figure 11: Previously reported mGWAS SNPs and the most strongly associated MT in FGFP. Scatter plot of negative log10 (two-sided) p-values (F-test for AB traits, chi-squared for P/A traits) for the most associated MT, for each of the previously reported mGWAS associated SNPs. SNPs are ordered by their chromosomal and base position along the x-axis. The smallest p-value observed for each SNP, in the FGFP mGWAS, is indicated by the y-axis The horizontal red line is the Bonferroni corrected p-value threshold given that we extracted estimates for 522 unique SNPs. The vertical blue, dashed line marks the *MCM6*/*LCT* region on chromosome 2. All of the blue dots above the red line and directly adjacent to the blue line are *Bifidobacterium* associations with the *MCM6* region. All other SNP-MT trait associations observed above the Bonferroni threshold are annotated with their FGFP MT and SNP. Sample sizes for each microbial trait and data used to generate this plot can be found in Table S8.

References

1. Wang, J. *et al.* Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* **48,** 1396–1406 (2016).
2. Goodrich, J. K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **19,** 731–743 (2016).
3. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
4. Bonder, M.J., *et al.* The effect of host genetics on the gut microbiome. *Nat. Genet*. **48**, 1407–1412. (2016).
5. Davenport, E.R., *et al.* Genome-wide association studies of the human gut microbiota. *PLoS One* **10**, 1–22. (2015).
6. Rothschild, D., *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215. (2018).
7. Blekhman, R., *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol*. **16**. (2015).
8. Graff M, *et al.* Genome-wide physical activity interactions in adiposity — A meta-analysis of 200,452 adults. *PLoS Genet* **13**(4): e1006528. (2017).