Supplementary Materials:

Genome-wide host-microbiota association analysis of 1,812 individuals identifies vitamin D receptor genetic variation and other host factors shaping the gut microbiota

Jun Wang[1,2,#], Louise B. Thingholm[3,#], Jurgita Skiecevičienė[3,#], Philipp Rausch[1,2], Martin Kummen[4,5,6], Johannes R. Hov[4,5,6,7], Frauke Degenhardt[3], Femke-Anouska Heinsen[3], Malte C. Rühlemann[3], Silke Szymczak[3,†], Kristian Holm[4,5,6], Tönu Esko[8], Jun Sun[9], Mihaela Pricop-Jeckstadt[10], Samer Al-Dury[11], Pavol Bohov[12], Jörn Bethune[3], Felix Sommer[3], David Ellinghaus[3], Rolf K. Berge[12,13], Matthias Hübenthal[3], Manja Koch[14], Karin Schwarz[15], Gerald Rimbach[15], Patricia Hübbe[15], Wei-Hung Pan[3], Raheleh Sheibani[3], Robert Häsler[3], Philipp Rosenstiel[3], Mauro D'Amato[16,17], Katja Cloppenborg-Schmidt[2], Sven Künzel[1], Matthias Laudes[18], Hanns-Ulrich Marschall[11], Wolfgang Lieb[14], Ute Nöthlings[10], Tom H. Karlsen[4,5,6,7,19,] ‡, John F. Baines[1,2,] ‡, Andre Franke[3,] ‡*

Materials and Methods:

Study subjects and sample collection

Two population-based cohorts from Schleswig-Holstein (Germany) were included in the study. Nine hundred and fourteen individuals from the PopGen- and 1115 individuals from the FoCus (Food Chain Plus) cohort were included. These two study cohorts were recruited independently from each other, and maximum number of individuals available was included to increase statistical power for various analyses. All samples, as well as corresponding phenotypic and dietary behavior were obtained from the PopGen biobank (Schleswig-Holstein, Germany)[20]. Study participants collected fecal samples at home in standard fecal tubes. Samples were shipped immediately at room temperature (RT) or brought to the collection center by the participants. Upon arrival into the study center (within 24 hours), they were stored at -80^{o}C until processing. Studies exploring the impact of storage conditions on sample quality and stability of the microbial communities indicate that storage at RT for up to 24 hour is a recommended procedure for preservation[55,56]. Written, informed consent was obtained from all study participants and all protocols were approved by the institutional ethical review committee in adherence with the Declaration of Helsinki Principles, whereas the sample identities are blinded from investigators. Sequence data of the 16S rRNA gene, genotype, nutritional, and phenotype data used for the herein described study has been made available to other scientists through PopGen's biobank general data transfer agreement. A summary of phenotypes used in this paper is given in **Table S1**.

Genotyping data

Samples of the PopGen and FOCUS cohorts were genotyped on different genotyping arrays. The PopGen samples were typed on the Affymetrix 6.0, Affymetrix Axiom, Illumina 550k, the custom Illumina Immunochip and Illumina Metabochip with sample pre-quality control (QC) sizes ranging from 678 to 1,218 and a variant coverage of 196,524 to 934,968 variants. The FoCus samples were typed on the custom Illumina Immunochip and the Omni Express Exome, with overall 1024 and 1713 pre-QC samples and a variant coverage of 195,732 to 964,193 variants. For each cohort, genotype data of each array were quality controlled separately and then merged and imputed. In total 17,017,474 single nucleotide variants (SNVs) were included for the PopGen cohort and 17,340,550 for the FoCus cohort.

Quality control

First, a sample QC was conducted followed by a SNP QC and identification of population outliers by PCA, as well as a batch PCA to identify outliers within each batch. (I) Individuals with missingness >10 %, excessive heterozygosity (not within interval [mean + 3*sd, mean - 3*sd]) (sd: standard deviation), with a kinship coefficient (identity by descent; IBD) > 0.185 and those failing a gender check were removed from the dataset. The IBD was estimated using the R-package SNPRelate (vs. 0.9.19) and a maximum likelihood (MLE) approach. This has proven to

be especially useful for custom arrays, for which moment estimators in this analysis overestimated relatedness.

The gender check was performed either with PLINK (v1.07)[57] or assessed by plotting the average X- and Y-chromosomal variant intensities. In the latter approach samples were gender classified using k-means clustering with two centers and 10 iterations (R package kmeans). Samples with incorrect genders were thus detected by non-equality of expected and observed gender. SNPs with missingness of > 5% and a deviation from Hardy-Weinberg equilibrium P < 0.00001 were excluded. At this step no MAF-threshold was applied. Population outliers were identified by PCA-based mapping against the HapMap III CEU, CHB, JPT and YRI population and excluded from the dataset. The PCA was performed using flashpca (git version f16ac44-dirty)[58]. Finally, a batch PCA was performed. All samples lying outside the rectangle [median(PC1) - 3*IQR(PC1),median(PC1) + 3*IQR(PC1)] x [median(PC2) - 3*IQR(PC1), median(PC2)+ 3*IQR(PC2)] were excluded. For estimation of relatedness and both the population and batch PCAs, a dataset pruned to remove linkage disequilibrium (LD) was used, leaving only SNPs with $r^2 < 0.2$) and a MAF > 0.05. Additionally, regions of high linkage disequilibrium (chromosome 5: 51.5 Mb – 55 Mb, chromosome 6: 25 Mb – 33.5 Mb, chromosome 8: 8 Mb – 12 Mb and chromosome 11: 45 Mb to 57 Mb) were removed. Next, samples from the different chips were merged, excluding one of the related samples (relatedness: IBD > 0.185; exclusion criteria: genotyping quality) across the different chips. For the PCA, only variants with a MAF > 0.05 were included. After quality control, the PopGen dataset comprised 1198 samples and 1,313,548 autosomal variants with MAF > 5%. Between 2 – 15 % of all samples were excluded during sample and population QC and 1 - 20 % of variants during SNP QC (not including the MAF criterion). The FoCus dataset comprised 1182 samples and 684,690 autosomal variants with MAF > 5 %. Between 8 – 13 % of all samples were excluded during sample and population QC and 3 – 6 % of all variants during SNP QC (not including the MAF criterion). In total, 37 % of all variants in the FoCus and 25 % of all variants in the PopGen cohort had MAF < 5 %.

Phasing and Imputation

Before imputation, all variants with a MAF < 0.05 were excluded. The data were phased using SHAPEIT (v2.r727.linux.x64)[59] with default parameters (input-thr 0.9, states 100, windows 2, effective-size 11418, burn 7, prune 8, main 20) without a reference panel. The coordinates published with the 1000 Genome Phase I variant set of March 2012 were used as the genetic map. The imputation was carried out using IMPUTE2 (vs.3.0_x86_64_static)[60] and the 1000 Genome Phase I variant set of March 2012 in 5 - 10 Mb chunks. Each chunk contained at least 200 variants. Again, default parameters were used (Ne 20000, buffer 250, burnin 10, k 80, k_hap 500). Three genotype probabilities were obtained. Only variants with an IMPUTE2-Info Score of more than 0.3 were further analyzed. After imputation, 17,017,474 with IMPUTE2-Info Score > 0.3 were present for the PopGen cohort and 17,340,550 for the FoCus cohort.

HLA alleles were imputed via SNP2HLA (v1.0)[61] using the T1DGCreference panel. Imputation was based on the pre-imputated SNP sets for each respective cohort. SNPs were mapped to the

chromosome 6 (25-34 Mb) of the snp142 set (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/), ATCG variants were excluded, minus strand variants were transformed to plus strand variants, and rsIDs were derived and used for the imputation of HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, and HLADRB1. Only alleles with a posterior probability of at least 0.7 were considered in the following analyses.

Sequencing and processing of bacterial 16S rRNA sequences

Bacterial genomic DNA was extracted using the QIAamp DNA Stool Mini Kit from QIAGEN on a QIAcube system. For all samples, the V1-V2 region of 16S rRNA gene was sequenced on the MiSeq platform, using the 27F/338R primer pair and dual MID indexing (8-nt each on the forward- and reverse primer) as described by Kozich *et al.*[62]. Sequencing was performed with MiSeq Reagent Kits v2. After sequencing, MiSeq fastq files were derived from base calls for read one and two (R1/R2), as well as both indices (I1/I2), using the Bcl2fastq module in CASAVA 1.8.2 (http://support.illumina.com/sequencing/sequencing_software/casava). Stringent demultiplexing was carried out by allowing no mismatches in either index sequence (instead of the default of one mismatch allowed by the MiSeq). Forward and reverse reads were merged with the FLASH software (v1.2)[63] and quality filtering was subsequently performed with the fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), excluding those sequences with >5% nucleotides of quality score <30. Chimeras in sequences were removed using UCHIME (v6.0)[64]. After randomly selecting 10,000 reads for each sample, taxonomical classification and compositional matrices for each taxonomical level were carried out using RDP classifier[65] with the latest reference database (RDP14), where classifications with low confidence at the genus level (<0.8) were organized in an arbitrary taxon of "unclassified_family". Species-level operational taxonomical units (97% similarity, OTUs) were created using the UPARSE routine[66].

Bile acids and fatty acids measurements on human serum samples

Serum bile acids were analyzed for 551 PopGen samples by HPLC-MSMS as recently described[67]. Five bile acids (cholic acid, CA; chenodeoxycholic acid, CDCA; lithocholic acid, LCA; deoxycholic acid, DCA and ursodeoxycholic acid, UDCA) including their Taurinated (T) and glycinated (G) conjugates were measured.

Polyunsaturated fatty acid composition in plasma was analyzed for the same 551 PopGen samples as above as described previously[68]. The following fatty acids were measured: C18:2n-6 (Linoleic acid), C18:3n-3 (Alpha-linolenic acid), C18:3n-6 (Gamma-linolenic acid), C18:4n-3 (Stearidonic acid), C20:2n-6 (Eicosadienoic acid), C20:3n-6 (Dihomo-gamma-linolenic acid), C20:4n-3 (Eicosatetraenoic acid), C20:4n-6 (Arachidonic acid), C20:5n-3 (Eicosapentaenoic acid), C21:5n-3 (Heneicosapentaenoic acid), C22:2n-6 (Docosadienoic acid), C22:4n-6 (Adrenic acid), C22:5n-3(Docosapentaenoic acid), C22:5n-6 (Docosapentaenoic acid), C22:6n-3 (Docosahexaenoic acid).

Cis- and trans-eQTL analysis on human data

For SNPs identified as associated with beta-diversity and/or single bacterial traits, a *cis*- and *trans*-eQTL analysis was performed using data on 2,360 individuals. The analysis design and recourse are described in detail in previous studies[69,70]. In summary, *cis*-eQTL analysis was performed on SNP-probe pairs for cases where the distance was less than 1 Mb. To consider effects of SNPs in LD with disease-associated SNP (trait-SNP), a conditioned analysis was performed by first adjusting the probe expression level for the effect of the strongest associated local SNPs (eSNP), and then repeating the eQTL analysis. Likewise, the p-value for the local best SNP was calculated conditioned on the trait SNP. In order to control for FDR, sample labels were permuted 100 times to obtain a p-value distribution. Expression probes with a significant association (FDR < 5%; two-way conditional analysis for cis-eQTL analysis) to a trait SNP are given in **Table S6**.

Statistical analysis

Correlation between microbiome and metadata

In both cohorts, beta-diversity measures based on genus-level composition were generated using the "*vegdist*" function (Bray-Curtis and Jaccard dissimilarities). Community ordination was performed using principle coordinates analysis (PCoA) based on the calculated dissimilarities using the "*capscale*" function in "vegan" (v2.3)[71]. The "*envfit*" function in "vegan" was used to correlate either categorical data, for which it performs multi-dimensional ANOVA on the ordination, or continuous variables, for which the function tests linear correlations between a given variable and the coordinates of microbial communities.

We considered a range of reported confounding variables that could shape the human gut microbiome: age, gender, body mass index (BMI), smoking and major nutritional components or food groups derived from diet patterns. Dietary patterns were collected via a validated, self-administered, 112-item food-frequency questionnaire established for German populations[72]. All participants were given the option of completing the questionnaire preferably as a web-based version and, optionally, on paper. Macro- and micronutrient intakes were obtained by using the German Food Code and Nutrient Database (vII.3) and provided by the Department of Epidemiology of the German Institute of Human Nutrition Potsdam-Rehbruecke[73]. Prior to association analysis, all individuals who took antibiotics less than 6 weeks before the stool collection were excluded, in order to remove the possible influences of antibiotic medication. The effect size and significances of the mentioned variables were estimated using "*envfit*", and the variables with significant effects (p<0.05) were further used in the GWAS analysis as co-variates (water, alcohol and all other highly correlated nutritional variables, which were collectively joined under the umbrella 'Total Energy'). The combined effect of host metadata was estimated further using the "*bioenv*" function in the "vegan" package, which calculates the

maximum Pearson correlation of microbial variation (Bray-Curtis dissimilarity) and combined dissimilarity in the selected subset of metadata (denoted by Gower distances). In order to reduce random errors in the lowly abundant taxa, the analysis focused on the "core-measurable-microbiota", which was determined using technical replicates according to Benson *et al.*[38]. Only those taxa with an average >40 reads per sample (thus less error introduce by random processes) were included (**Figure S12**).

Association of individual bacterial traits with human genetic variation

In order to identify human genetic variation associated with the abundance of individual gut bacteria, a statistical test for each combination of SNP and taxon was performed. The abundance of bacteria in the human gut is characterized by increasing number of zeros at lower taxonomic levels, a right screwed distribution often with a long tail and only positive values. Thus, a model assuming a normal distribution of dependent variables could not be fitted to our data. The generalized linear model (GLM) with a negative binomial (negbin) distribution and log link was selected for the statistical analysis as the best fitting model across all bacteria. The hurdle model with a negbin distribution showed an increasingly good fit with increasing number of zeros. The GLM negbin model was therefore selected as a consistent model across all bacteria, while the analysis of species (97% similarity threshold OTUs) was supported with the hurdle model[74] .

Our identified "core measurable microbiota" [38], consists of 64 taxa across five levels (phylum, class, order, family and genus) and 42 species-level OTUs. Taxa with >90% of their counts within the first 5% of range of counts or with >90% of above-zero counts within the first 5% of the above-zero range were excluded, as they performed badly with the selected model(s). Forty OTUs and 58 taxa were used for association study with human SNPs. The analyses were preformed on both cohorts separately (986 samples in FoCus and 826 samples in PopGen). In the analyses, outliers defined as 5*standard deviation (SD) were removed, genetic variants not overlapping between FoCus and PopGen were discarded, while variants with MAF > 0.05 and IMPUTE2 INFO criteria > 0.8 were included. No population stratification is observed between the two cohorts ($\lambda GC=1.00$, **Figure S18**)[75].The covariates BMI, age, gender, genetic principal components 1-3 and nutritional variables alcohol, water and 'total energy' intake were used. The analyses were performed using R Project version 3.2 and the GLM.nb function in the "MASS"[76] package version 7.3 for GLM negbin, and the hurdle function in the package "pscl" (v1.4)[77].

A meta-analysis of GLM negbin hits across the two cohorts was performed using PLINK (v1.9 64-bit)[57], with the command "–meta-analysis +qt", including information on beta coefficients and standard errors. Clumping was performed using PLINK v1.9 with the "–clump" command on SNPs meeting the following filtering criteria: meta-study fixed-effect p-value $< 5 \times 10^{-8}$, single cohort p-value $< 5 \times 10^{-4}$, the same beta-value sign (same direction of association) and AIC (model fit parameter) < 50k. Clumps with at least two SNPs of which at least one SNP was genotyped were selected. For each selected clump, the SNP with the lowest meta-analysis p-value was selected as the tag SNP and for bacteria containing zero counts the hurdle model was

applied as described above. All hits were confirmed to be supported by the count- or zero part of the hurdle model with a p-value < 0.05 in both studies.

Annotation and enrichment

DEPICT[39] was used to annotate and perform tissue and gene set enrichment analyses among the significant associations between human genetic variation and both individual bacteria traits and beta-diversity. DEPICT was used with the following settings a) association_pvalue_cutoff: $1x10^{-5}$, b) nr_repititions: 20, and c) nr_permutations: 500, and all available analysis steps were performed.

For genotype data we used 1000_genomes_project_phase3_CEU/ALL.chr_merged.phase3_shapeit2_mvncall_integrated_v5.2 0130502.genotypes, for the collection file we specified ld0.5_collection_depict_150315.txt.gz and for the reconstituted gene sets file we specified GPL570-GPL96-GPL1261-GPL1355TermGeneZScores-MGI_MF_CC_RT_IW_BP_KEGG_z_z.binary.

Genetic variation correlated with overall community differences

In addition to taxon-oriented association analysis, we also performed analyses aimed at identifying genetic variation that may not necessarily associate with individual bacterial taxa with genome-wide significance, but rather correlate with overall community differences (beta-diversity), by *e.g.* simultaneously influencing numerous taxa. We performed a simulation and treated genotype at each locus as categorical variables (the distribution of each genotype follows Hardy-Weinberg equilibrium), and measured the genotype association using the "*envfit*" function in the "vegan" R package (v2.3)[71]. This approach calculates the community differences associated with different factor levels (in this case, three different genotypes), by comparing the difference in centroids of each group relative to the total variation based on the main axes of the PCoA based on the Bray-Curtis dissimilarity. By shuffling the simulated genotype $>2x10^7$ times, we effectively obtained a large-enough null distribution of effect size. This was performed for six categories of MAF to represent loci with MAF of 5%, 10%, 20%, 30%, 40% and 50% (whereas in case of a real SNP, it is compared to its closest MAF category, **Figure S19**), and if a certain locus displays greater effect sizes than the simulated maximum, they are extremely unlikely to be observed by chance ($p<5\times10^{-8}$) and can be considered genome-wide significant. We have filtered SNPs in a similar fashion as taxa-association mentioned above, that a SNP is considered significant when 1) its effect size surpasses the significance threshold in both cohorts; 2) its effect size surpasses the significance threshold in combined cohort; 3) such SNP is included in a clump with at least two SNPs, and at least one SNP was genotyped.

The additive effect of the significant loci from this analysis was then determined using redundancy analysis based on genus-level composition ("*rda*" in "vegan" package), and the "*ordiR2step*" function in "vegan" package, which optimizes the order of loci in a linear model and sums up the variation of the ordination explained by each additional locus.

HLA analyses were conducted on the respective HLA haplotypes within each locus, coded as carrier or non-carrier of each specific allele. We performed distance based redundancy analysis after correction for host characteristics (see association analysis for factors). These models were then tested using a permutative ANOVA approach (5000 permutations) as implemented in the "vegan" function "*anova.cca*", and the coefficients of determination were extracted via "*RsquareAdj*".

Analysis of gut microbiome data from *Vdr* KO mice

Gut microbiome data from Jin *et al.*[25] was kindly provided by Dr. Jun Sun from the University of Illinois at Chicago, which includes fecal samples from three wild type and five *Vdr* knockout mice for which the V4-V6 region of the 16S rRNA gene was sequenced on the 454 GS-FLX platform. Quality filtering, removing chimeras and classification were performed according to the same procedure as described in the previous section. Statistical tests for the effect of *Vdr* genotype on the microbiome were carried out with "*envfit*" function in "vegan" as described for the analysis with respect to human SNPs (see above). Comparison of specific taxa was carried out by the Wilcoxon test. Results are shown in **Figure S5-6**.

Analysis of association between bile- and fatty acids and the microbiome

In order to identify bacteria associated with the concentration of measured bile acids, including total LCA (the sum of LCA, G.LCA and T.LCA) and total BA (sum of all 15 bile acids), a generalized linear model with an inverse Gaussian distribution and log link was applied, excluding 5*SD outliers of bacteria and bile acids and including the covariates age, gender, BMI, 'total energy' intake, water, alcohol and bile acid batch number. To evaluate significance levels, the Benjamini–Hochberg (BH) correction method was applied for each bile acid analysis (**Table S3**). To identify bacteria associated with omega-3 and -6 fatty acids, a linear regression model was applied with a square root transformation of fatty acids, excluding 5*SD outliers of bacteria and including the covariates age, gender, BMI, total energy intake, water and alcohol. To evaluate significance levels, the BH corrected *p*-values were calculated (**Table S4**). Association with beta-diversity was carried out using "*envfit*" function as described above.

Shotgun metagenomic analysis

HiSeq sequencing

For a subset of 122 individuals, the same DNA extracts used in 16S rRNA gene sequencing were subjected to shotgun metagenomic sequencing. Samples were prepared following the Illumina Nextera DNA Library Preparation Kit and sequenced on the HiSeq Platform as 2×125 bp paired-end reads. Nextera adapter sequences were trimmed using Trimmomatic (v0.32)[78]. Quality control of the sequencing reads was performed with sickle (v1.330, https://github.com/najoshi/sickle) and parameters set to a sliding-window quality threshold of 20 and a minimum length of 60 after quality trimming. DeconSeq[79] was run to identify and remove

human reads from the sequencing file, using the hg19 human genome sequence as reference database. If one of the reads belonging to a read pair was removed at any of the QC steps, the respective paired read was discarded, as well.

HUMAnN2analysis.

The 189 samples that passed QC and for which genetic data is available were analyzed using HUMAnN2 (https://bitbucket.org/biobakery/humann2/) with default settings except '--bt2_ps sensitive' for the analysis of pathway and gene family abundance. Gene families including the term 'bile acid' were selected and four pathways relevant for bile acid metabolism were selected (bile acids degradation, iso-bile acids biosynthesis I + II, bile acid biosynthesis, neutral pathway and glycocholate metabolism (bacteria)). Association with *VDR* genotype (rs7974353) was evaluated using GLM with an inverse Gaussian distribution, the covariates BMI, age, gender, alcohol, water and total energy intake and removal of 5*SD outliers.

Replication in FoCus obesity cohort

SNPs found to be significantly associated with beta-diversity in this study were consequently replicated in an additional "FoCus obesity cohort". The FoCus obesity cohort was recruited from the Obesity Outpatient Centre at the University Hospital in Kiel, which offers both non-surgical and surgical obesity therapies. The pheno- and genotyping profile was similar to the one applied for the FoCus control cohort. The recruitment of the FoCus obesity cohort was approved by the local Ethics Committee (A156/03) and each patient gave their informed consent. To replicate associations of lead SNPs with beta-diversity, the effect size of each SNP was calculated with "*envfit*" and consequent p-values were calculated based on the same empirical null distributions as described above; successful replications are defined as having p-values <0.05/42 (in total 42 SNPs included in the test).

Transcriptome analysis of human colon biopsies

RNAseq was conducted on biopsies obtained from four groups of individuals (59 samples) that were retrieved from the local hospital biobank in Kiel. The individual groups are composed of healthy individuals without significant pathological findings, patients with acute non-inflammatory bowel disease intestinal inflammation (Disease controls), Crohn's disease patients and ulcerative colitis patients. The samples were taken either from the sigmoidal colon or the terminal ileum in two inflammatory stages (inflamed or non-inflamed). Total RNA was extracted using the RNeasy kit (Qiagen, Germany) according to the manufacturer's protocol and sequenced on an Illumina HiSeq2000 using the Illumina total RNA stranded TruSeq protocol. Pre-processed reads were aligned to the hg19/CRCh37 reference genome with TopHat2[80]. Gene expression levels were computed by HTSeq[81] and analyzed with the Bioconductor package DESeq2[82].

To gain insight into the nature of differentially expressed genes in pair-wise comparisons between inflamed versus non-inflamed samples, transcription factor binding sites (TFBS) were obtained using the InnateDB database (www.innatedb.com)[83], integrating predicted transcription factor binding site data from the CisRED database (www.cisred.org) [84]. Genes related to VDR

transcription factor binding sites were subjected to further analysis. Correlations were calculated employing the Spearman rank correlation coefficient while differences between correlations were assessed using the Mann-Whitney U test. Functions associated to *VDR* related transcripts were obtained from Geneontology.org and further summarized into meta-terms. Bacterial abundances from the corresponding samples were generated as previously described[15].

Supplementary Figures:

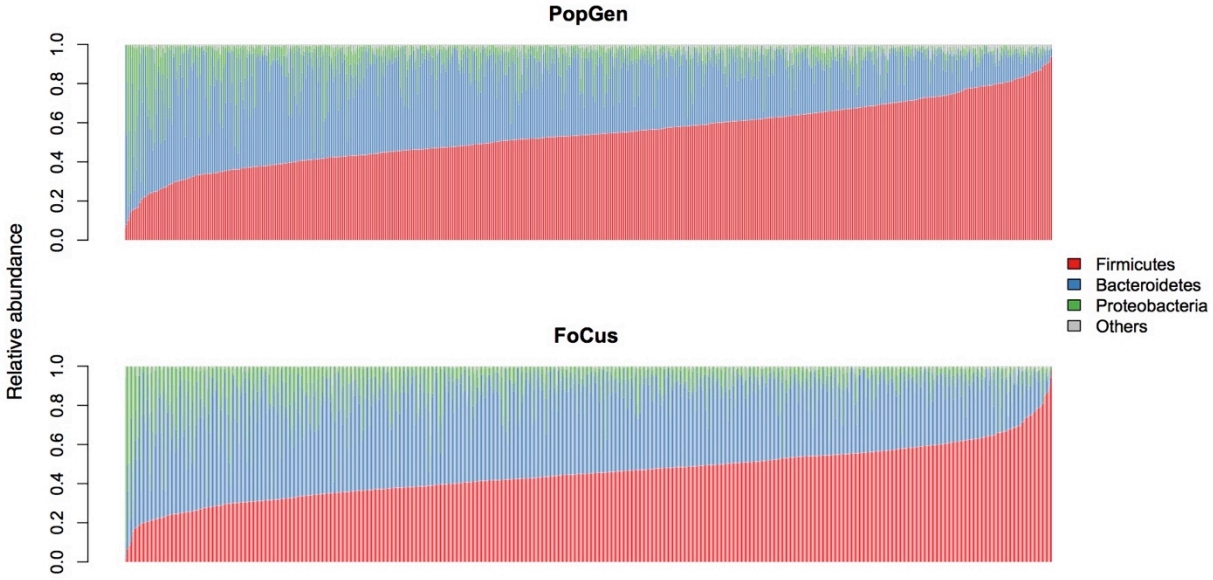
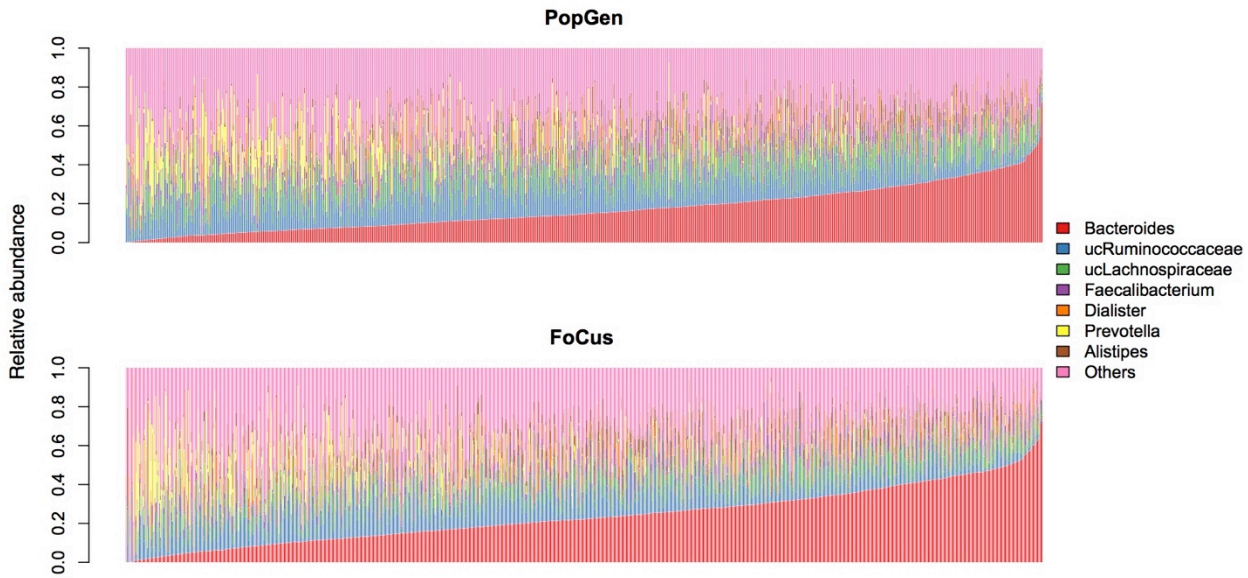


Figure S1: Microbiome composition in both cohorts at the phylum level. The three most abundant phyla are shown, and samples are ordered in ascending relative abundance of the overall most abundant phylum (Firmicutes).



340

Figure S2: Overview of microbiome composition in both cohorts at the genus level. The seven most abundant genera are shown, and samples are ordered in ascending relative abundance
345 of most abundant genus (*Bacteroides*). uc: unclassified.

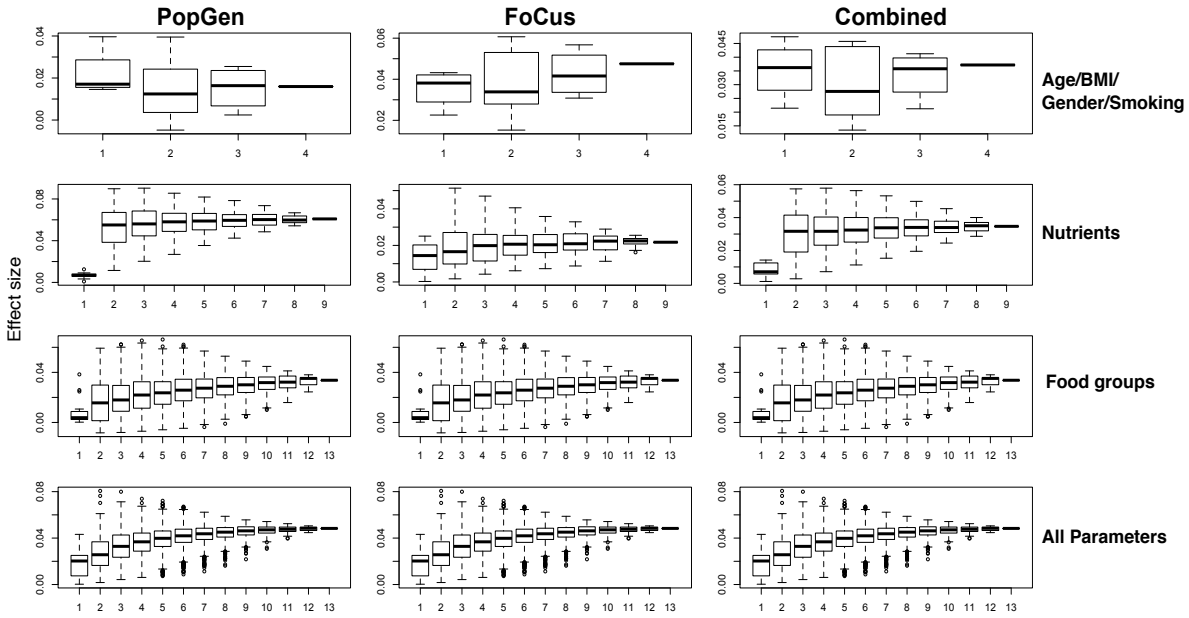
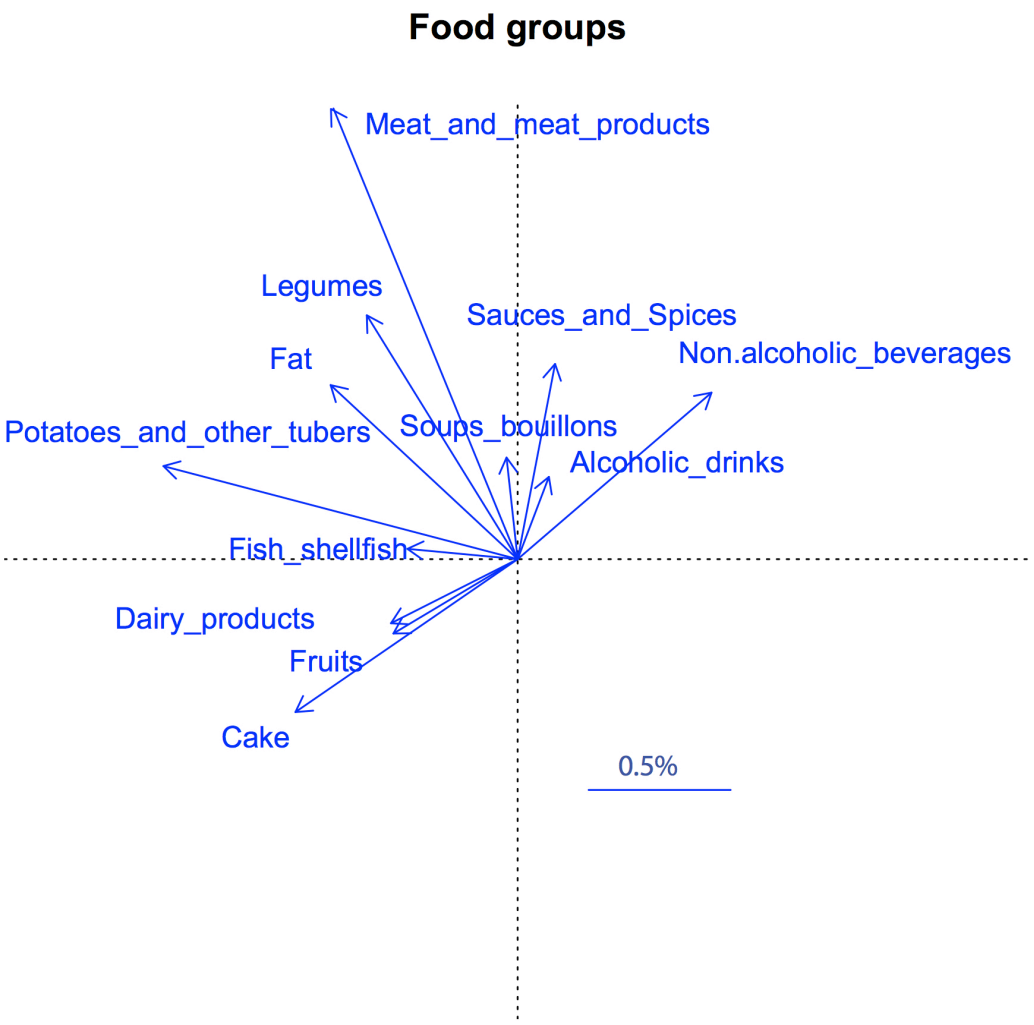


Figure S3: Cumulative effect for host parameters (first row), nutrients (second row), food groups (third row) and all factors combined (bottom row). For each number of incrementing variables included, all possible subsets were used, and the cumulative effects were calculated with "*bioenv*" function, which calculates the association between dissimilarity of the metadata (Gower dissimilarity) and beta-diversity (Bray-Curtis dissimilarity). The largest effect size is usually observed with an optimal combination of several variables instead of all variables. Since nutrients are derived from food groups and both have a similar cumulative effect size, only nutrients were included with anthropometric parameters to calculate the overall cumulative effect.



360

Figure S4: Significant associations between food groups and beta-diversity (PCoA based on Bray-Curtis dissimilarity) in combined cohorts. The ordination of the microbial communities is identical to **Figure 1,** with arrows denoting the effect size (variation explained using "*envfit*" 365 function, see **Methods**).

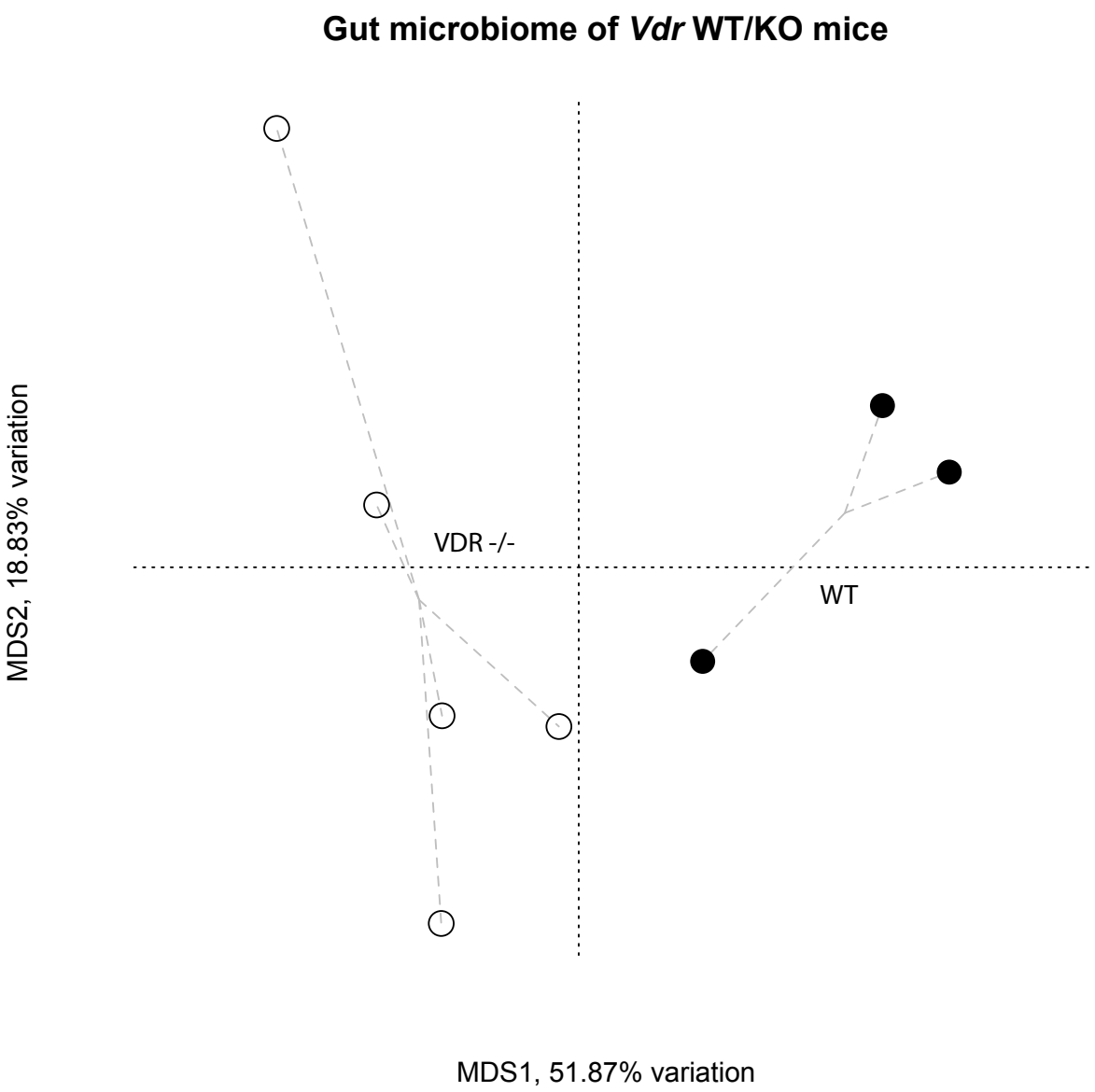
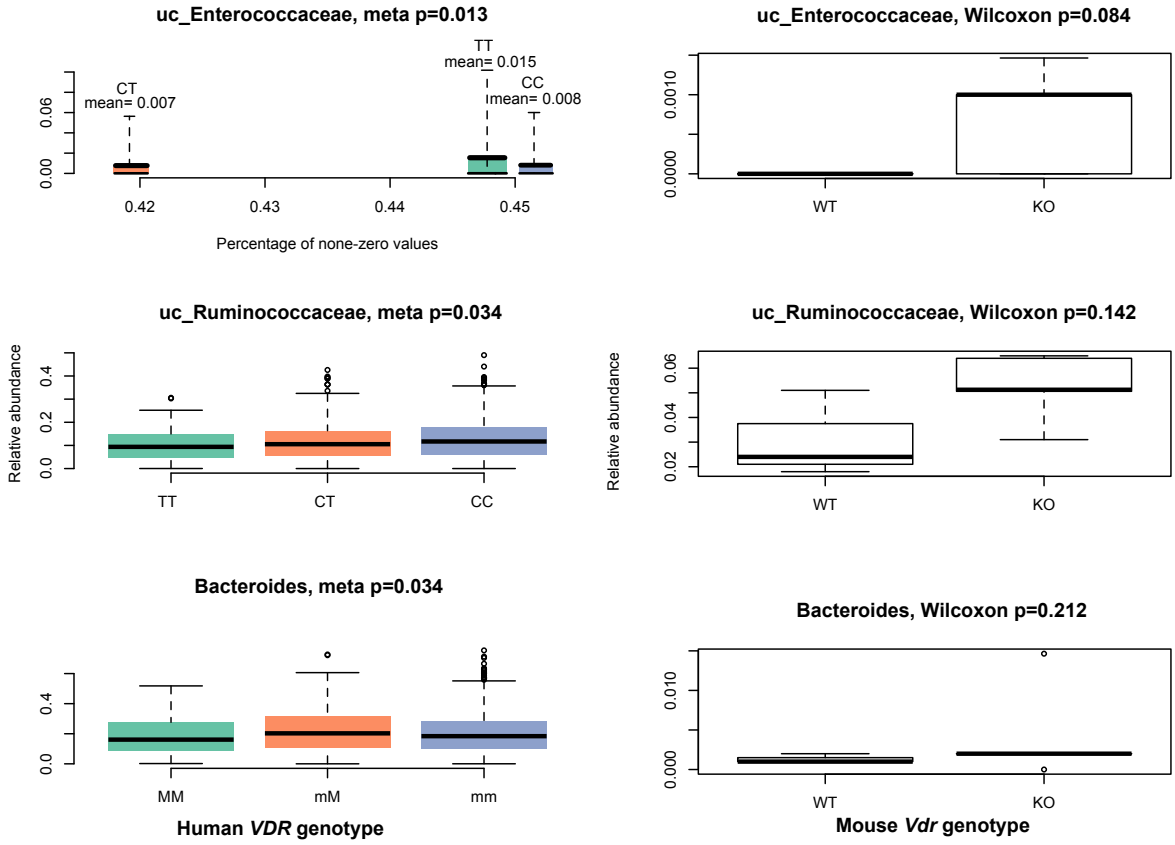


Figure S5: *Vdr* knockout in mice significantly changes the gut microbiome. Ordination shows the fecal microbiome of eight mice based on Bray-Curtis dissimilarity, and lines connect samples of the same genotype ($Vdr^{-/-}$, n= 5, and WT, n=3). Data (V4-V6 region of the 16S rRNA gene) were obtained from Jin *et al.*[25].



375

Figure S6: Replication of *VDR* genotype and specific taxa associations. The left panels show three most significant taxa associations of *VDR* gene after *Parabacteroides* in humans, with *p* values derived from meta-analysis using a GLM (see **Methods**). For uc_Enterococococaceae, both the percentage of non-zero values for each genotype as well as the mean values are shown (see
380 also Figure 3), while the other two taxa are present in nearly all samples, thus boxplots are shown. Right panels display taxa in wild type (WT, n=3) and *Vdr* knockout mice (KO, n=5) for which significant differences exist based on the Wilcoxon test. uc: unclassified.

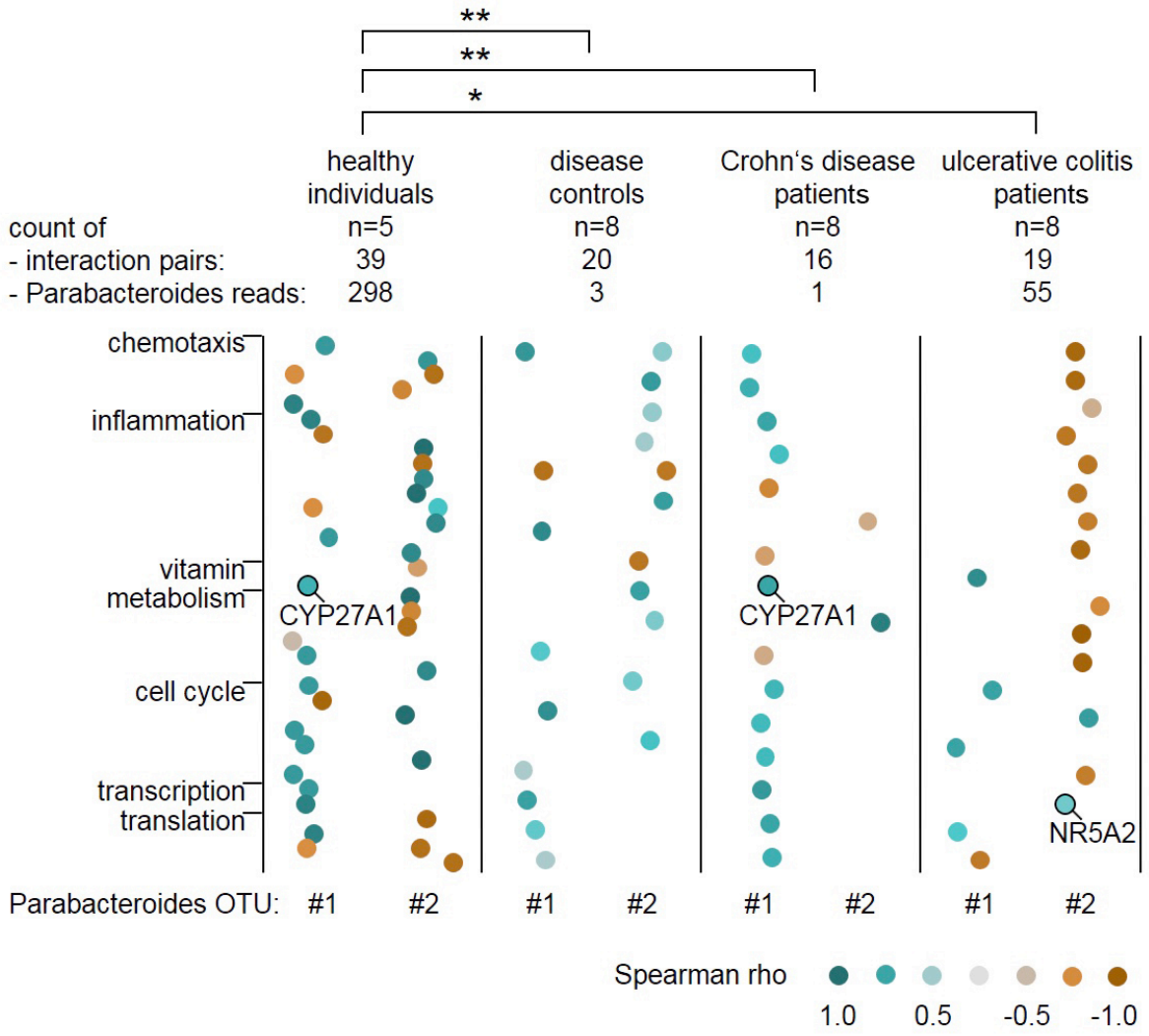


Figure S7: Top 50 interactions between VDR-disease transcripts and *Parabacteroides*. Transcripts from the sigmoidal colon which share a VDR binding motif and are differentially expressed in response to inflammation are shown with their correlation to Parabacteroides: unique correlation pairs with the 50 strongest Spearman rho correlation coefficents are shown. Several transcripts are represented in more than one group. Each dot corresponds to a correlation pair between a transcript and one of the two most abundant Parabacteroides signals (#1 present in 87% of all individuals, #2 present in 67% of all individuals), colour coded by Spearman rank correlation coefficient. The different groups are shown separately: healthy individuals, disease controls, Crohn's disease patients and ulcerative colitis patients. The significance of the differences between the Spearman rank correlation coefficients of the individual groups are indicated by stars (* p<5x10-2; ** p<5x10-4). Two transcripts functionally associated to bile

acid metabolism were highlighted separately(CYP27A1: Cytochrome P450 Family 27 Subfamily A Member 1; NR5A2: Nuclear Receptor Subfamily 5 Group A Member 2). Biological processes indicated to the left refer to: chemotaxis, migration and adhesion (chemotaxis); inflammation, immune process, stress defense, response to bacteria (inflammation); ion transport, vitamin and nutrition process (vitamin); metabolism and energy process (metabolism); proliferation, cell cycle and apoptosis (cell cycle); RNA transcription and splicing (transcription); translation and phosphorylation(translation).

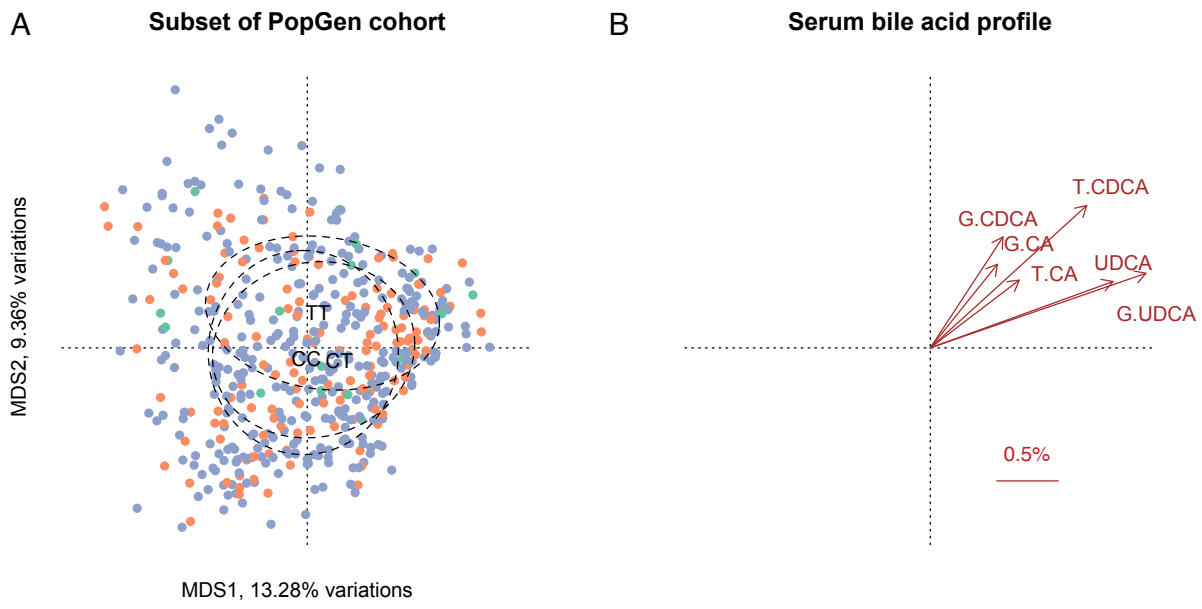
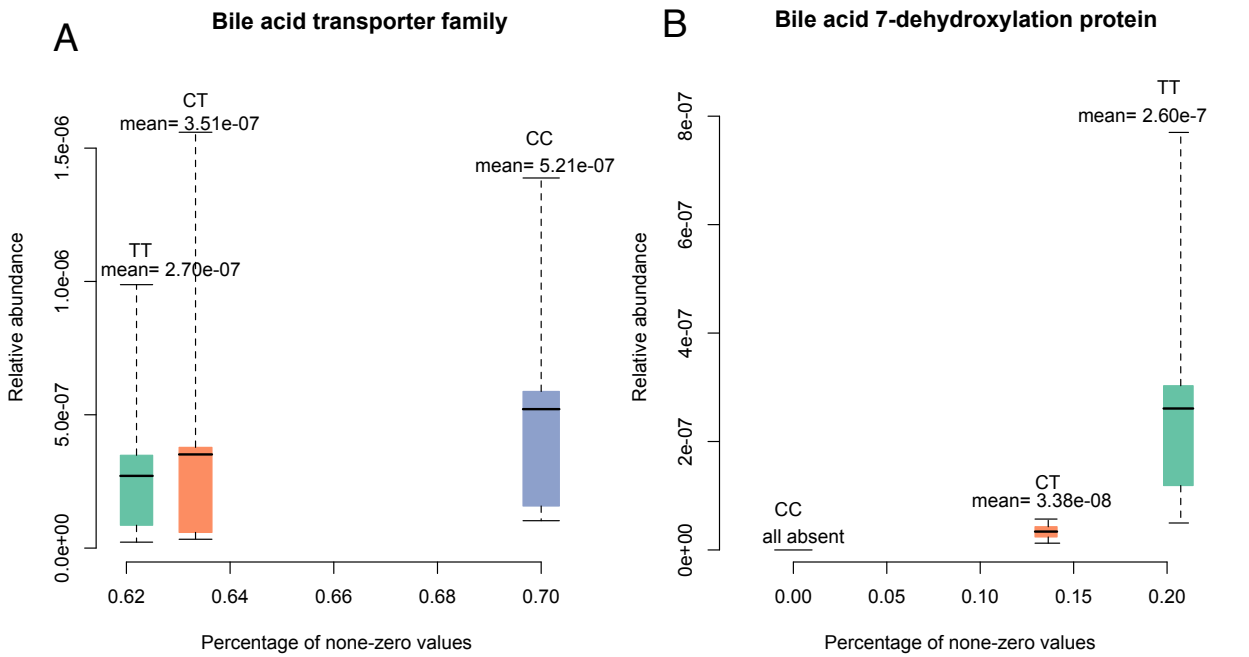


Figure S8: Overview of beta-diversity (Bray-Curtis dissimilarity, A) and significantly associated bile acids (B). Panel **(A)** shows the principle coordinate analysis of a subset of the PopGen cohort (n=551). Samples are shown in different colors according to *VDR* genotype (blue=CC, orange=CT, green=TT), with circles containing 50% of the samples for each group (for visualization). Panel **(B)** displays bile acids with a significant correlation to beta-diversity, with arrows denoting the effect size (variation explained as calculated by "*envfit*" function, see **Methods**).



415

Figure S9: Association between *VDR* genotype (rs7974353) and specific gene-family abundances derived from shotgun metagenomic data. Associations were tested with generalized linear models and two gene families with $p<0.05$ are shown. For each association, both the percentage of non-zero values of each genotype as well as mean values are shown (see also **Figure 3**).

Affinity-based ranking of sequences

Rank	Affinity value	Sequence name	Start Position in the sequence	End Position in the sequence	Sequence length
1	0.160902	rs66589178_G	1	16	16
2	0.00887359	rs66589178_C	1	16	16

Figure S10: In silico prediction of VDR affinity affected by SNP rs66589178. The VDR binding matrix was retrieved from the Homoco database (http://autosome.ru/HOCOMOCO/browseTFs.php, VDR-f1, 8nt), and prediction was carried out using the TRAP algorithm (http://trap.molgen.mpg.de/cgi-bin/trap_pers_receiver.cgi) for different nucleotides at this locus.

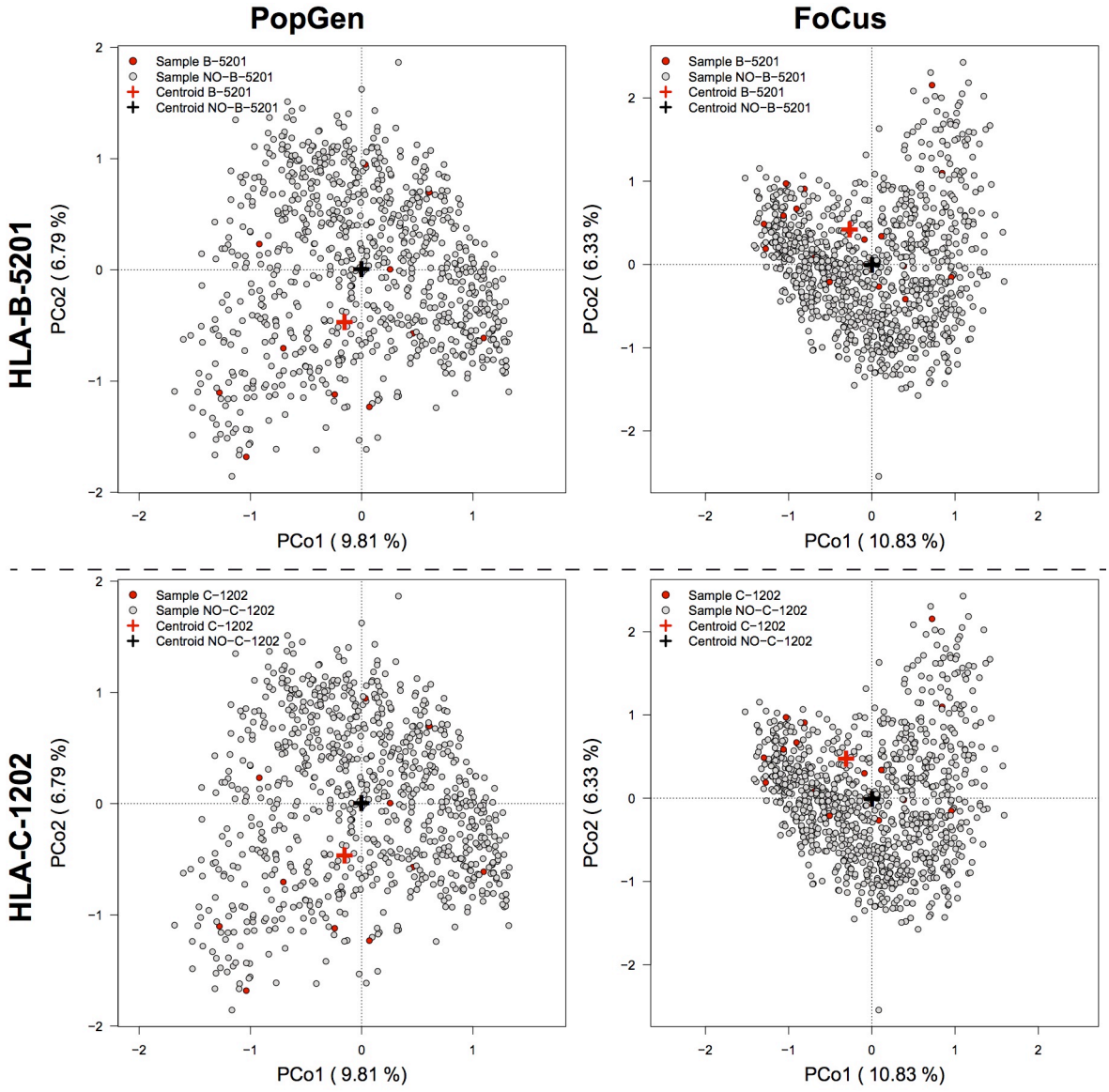


Figure S11: Consistent influences of HLA haplotypes on the beta-diversity (Bray-Curtis dissimilarity) in the two independent cohorts: HLA-B-5201 and –C-1202 alleles display a consistent and significant influence in the PopGen (left panels) and FoCus cohort (right panels). Analyses were carried out to incorporate the influence of anthropometric variables (see methods). Other HLA haplotypes show cohort-specific influences on the microbial community (see **Table S7**).

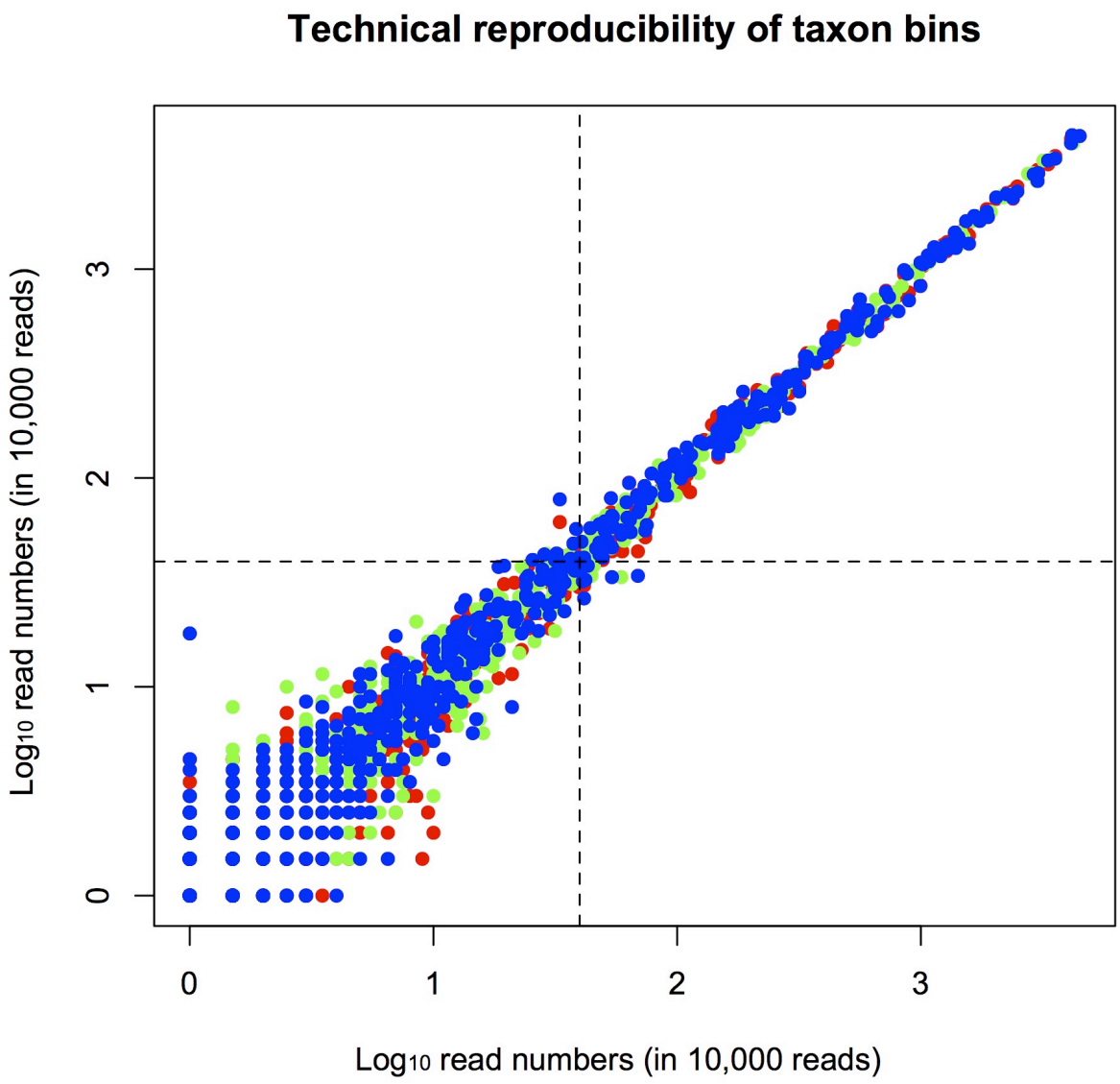
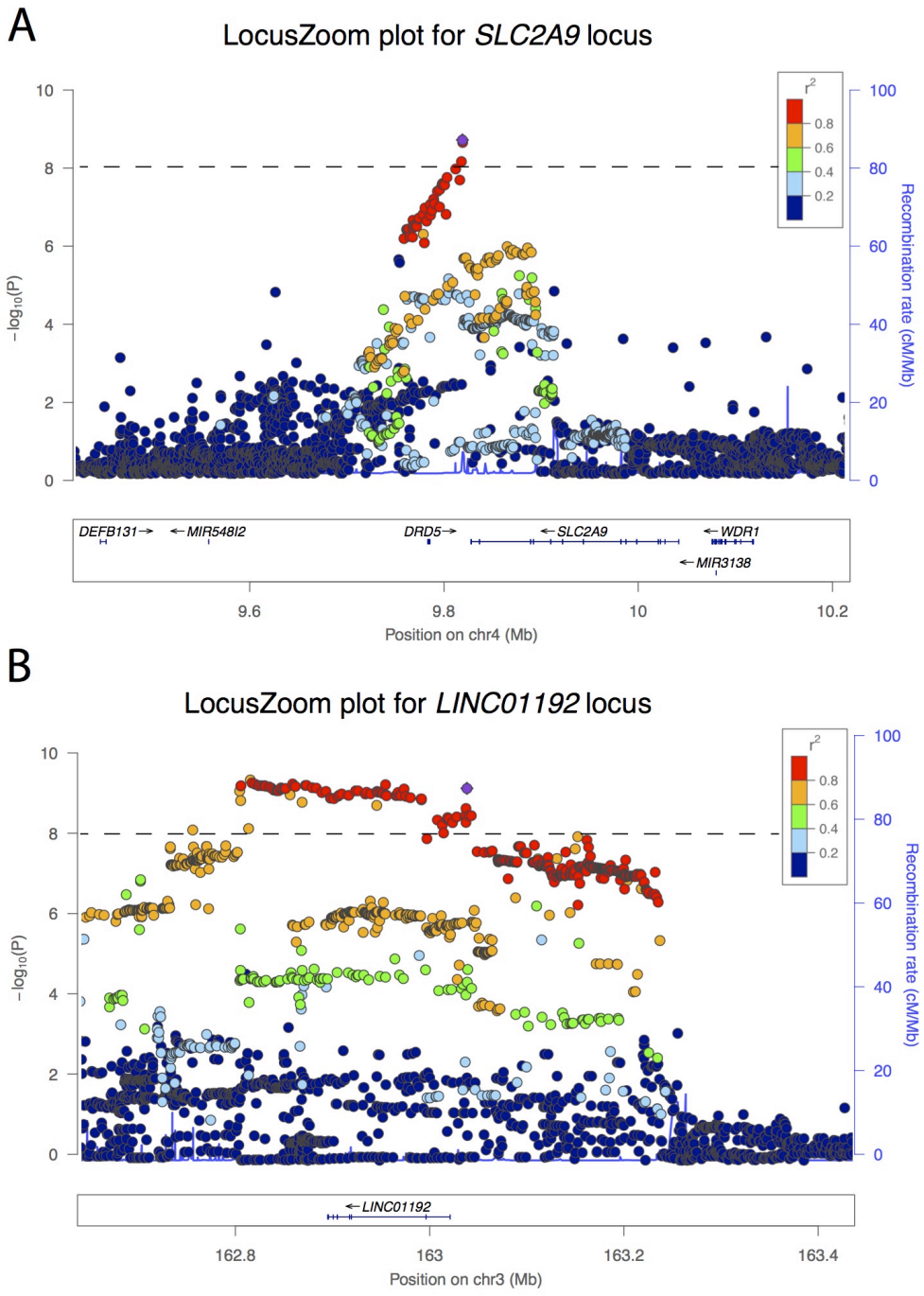


Figure S12: Correlation of taxon bins between technical replicates and determination of the core-measurable-microbiota (CMM). Taxon abundances from three technical replicates generated for the same sample (ten in total) are plotted against each other (log_{10} transformed) according to Benson *et al.*[38]. Highly reproducible taxa (r^2>0.97) can be found for those with higher than 40 reads per replicate (in 10,000 reads, **Table S8**). One color denotes technical replicates from one biological sample.



455

Figure S13: *SLC2A9* and *LINC01192* as examples of genes associated with individual taxa. **(A)** LocusZoom plot showing meta analysis p-values for *SLC2A9* associated with unclassified Porphyromonadaceae (Species), which is a locus containing genes enriched for response to vitamin D (see results). (**B)** LocusZoom plot showing meta analysis p-values of *LINC01192* associated with Lactobacillales (Order). This locus contains genes enriched for response to vitamin A (see **Results**).

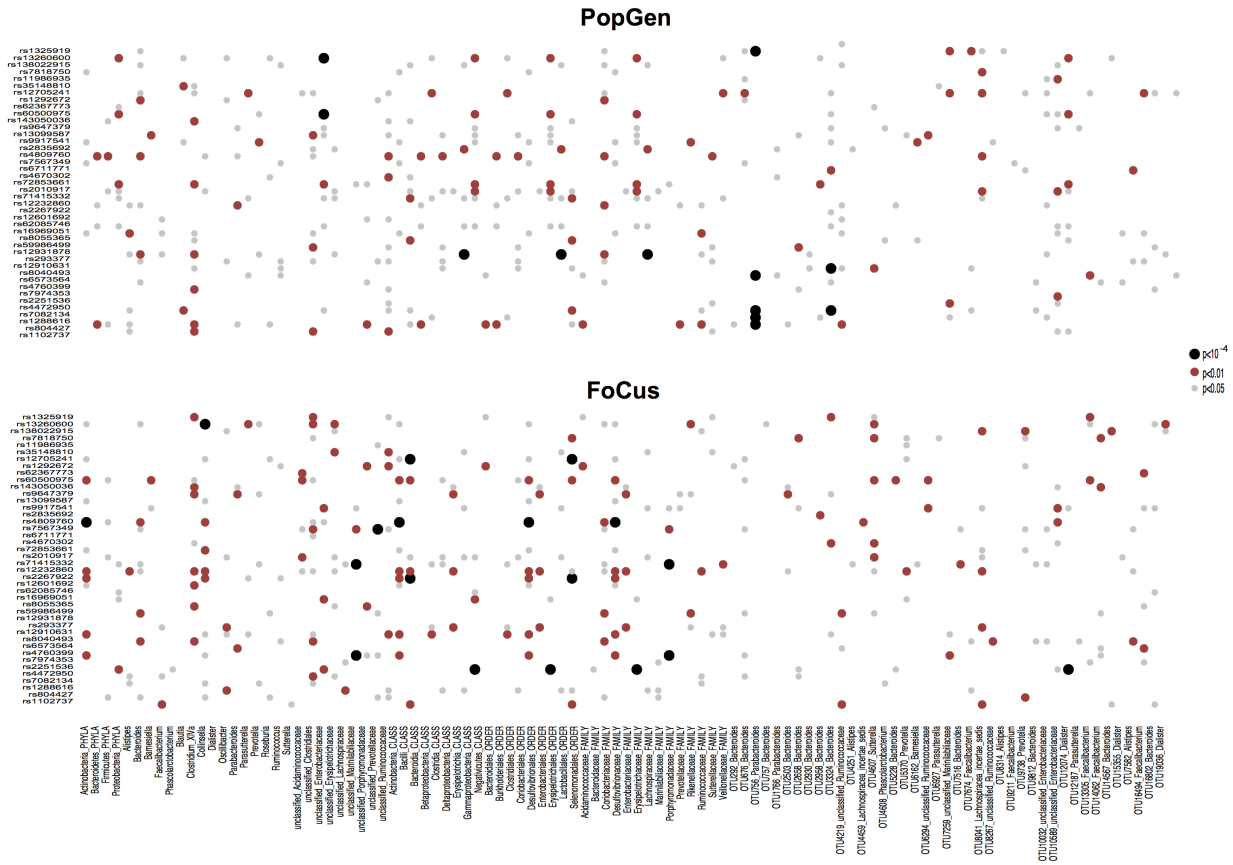


Figure S14: Overview of correlation between measured CMM and 42 loci with significant associations to beta-diversity (Bray-Curtis dissimilarity). GLM results for each cohort are shown separately, and those with significant correlations (p<0.05, grey; p<0.01, dark brown; p<1 × 10^{-4}, black) are shown for each locus represented by the leading SNP.

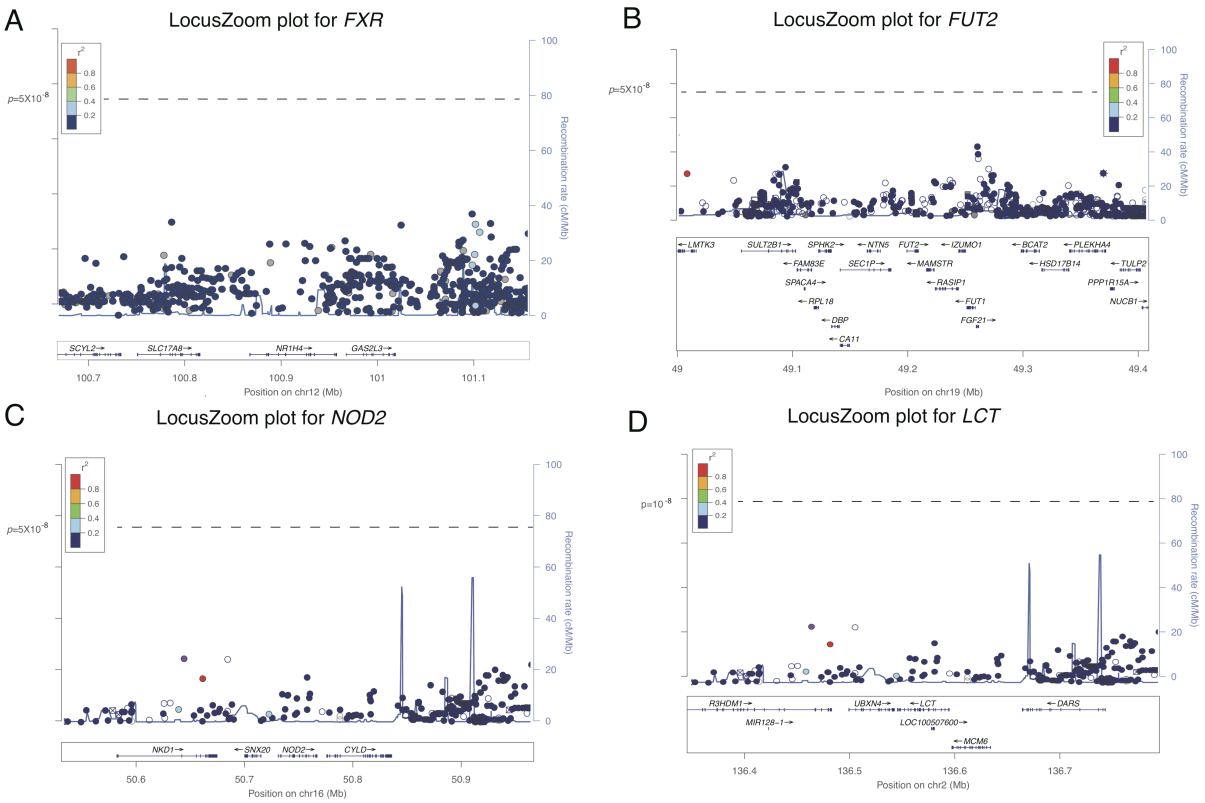


Figure S15: LocusZoom plot showing other regions of interest. Adjusted effect sizes with respect to beta-diversity (actual effect size divided by significance threshold, see **Methods**) are shown for SNPs in each region for four loci: *FXR* (*NR1H4*), *FUT2*[13,86], *NOD2*[18], and *LCT*[19]. Dashed lines denote the genome-wide significance threshold (maximum effect size from null distribution, see methods). No SNP in the region has a large enough effect size to be considered genome-wide significant in any of these four genes.

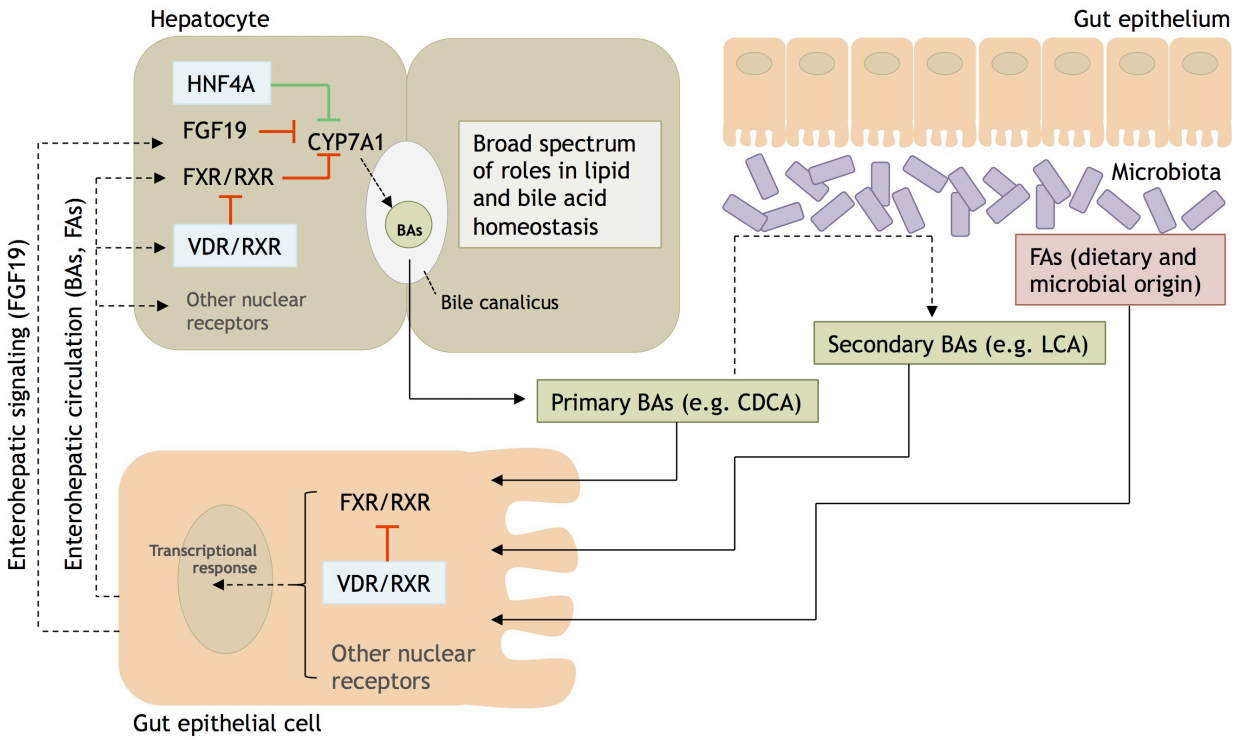


Figure S16: Schematic overview of bile acid and fatty acid signaling within the enterohepatic circulation. Genome-wide significant associations between *VDR* and *HNF4A* (highlighted in blue) polymorphisms and gut microbial community composition are prototype findings of the current article. To contextualize these results, the figure shows key aspects of gut-liver interactions relevant to our findings. The intention is not to claim causality, since by nature all findings of the study design are correlative, but to provide a simplified context for interpretation. In addition to 1α,25-dihydroxyvitamin D3, VDR/RXR ligands include secondary bile acids (lithocholic acids) and essential fatty acids[23]. Primary bile acids are produced in the liver by CYP7A1. They are actively secreted from hepatocytes into the bile canaliculi and are thereafter reabsorbed in the terminal ileum. Through gut bacterial co-metabolism, primary bile acids are converted to secondary bile acids. Primary and secondary bile acids, together with fatty acids (derived both from diet and produced by colonic microbiota from dietary carbohydrates and proteins) are absorbed within ileal enterocytes. Here, they bind and activate the nuclear receptor FXR[86] and other nuclear receptors, the former initiating transcriptional responses leading to: (i) stimulation of FGF19 production and (ii) activation of bile acid and fatty acid recirculation, thus completing the enterohepatic cycle[27]. Activation of nuclear VDR with secondary bile acids or fatty acids, amongst other biological actions, inhibits the activation of FXR and thereby suppresses the signal transduction mediated by FXR. In the liver, FGF19 from the intestine activates signaling pathways that down-regulate bile acid synthesis. Upon reaching the liver, bile acids also activate FXR signaling pathways and other nuclear receptors with broad effects on bile acid and lipid biology. Activation of VDR in the liver causes suppression of FXR signaling amongst other biological effects. HNF4A serves complex biological effects related to

lipid biology, amongst which is involvement in the constitutive expression of CYP7A1. Multiple other findings in the study (see main manuscript) point toward key components of the sterol pathway, substantiating the example made by the VDR findings. Arrows indicate the direction of the processes, green lines indicate stimulatory effects and red lines indicate suppressive effects on target genes. Abbreviations: BA, bile acid; FA, fatty acid; CDCA, chenodeoxycholic acid; LCA, lithocholic acid; FXR, farnesoid X receptor, VDR, vitamin D receptor; RXR, retinoid X receptor; HNF4A, hepatocyte nuclear factor 4, Alpha, FGF19, fibroblast growth factor 19.

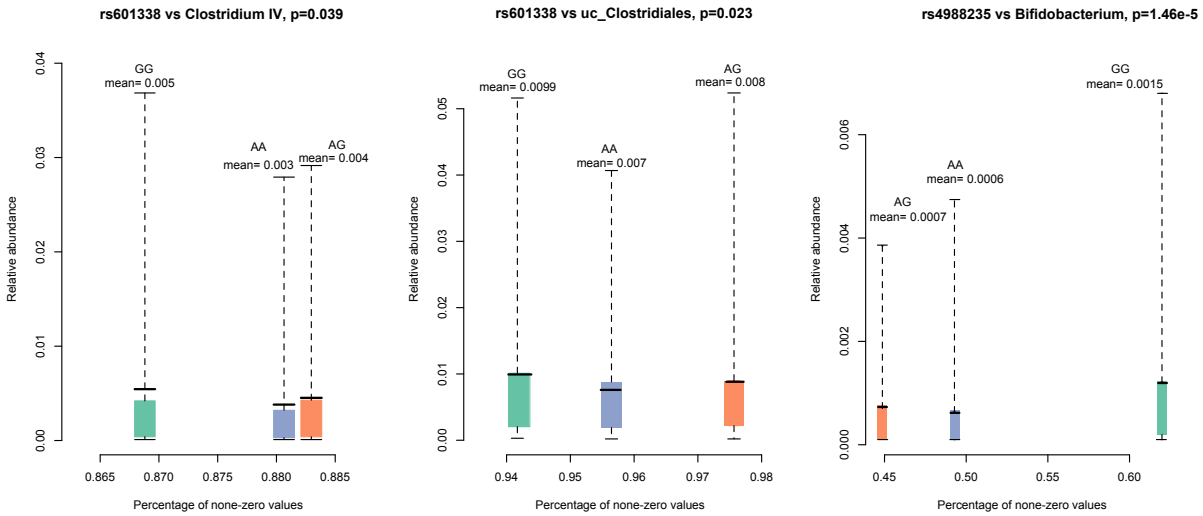


Figure S17: Replication of specific gene-taxon associations for *FUT2* (A, B) and *LCT* (C). Associations were tested with spearman correlations between relative abundances of the taxa and genotype. Weak associations were discovered for *FUT2* with *Clostridium IV* and unclassified Clostridiales, partially agreeing with Wacklin *et al.*[85]. A strong association was found between *LCT* and *Bifidobacterium* (p=1.46 × 10^{-5}), similar to the results reported by Blekhman *et al.*[19] (p=1.16 × 10^{-5}), but still not reaching the genome-wide significance threshold, both with- and without correcting for other confounders. For each association, both the percentage of non-zero values of each genotype as well as mean values are shown (see also **Figure 3**).

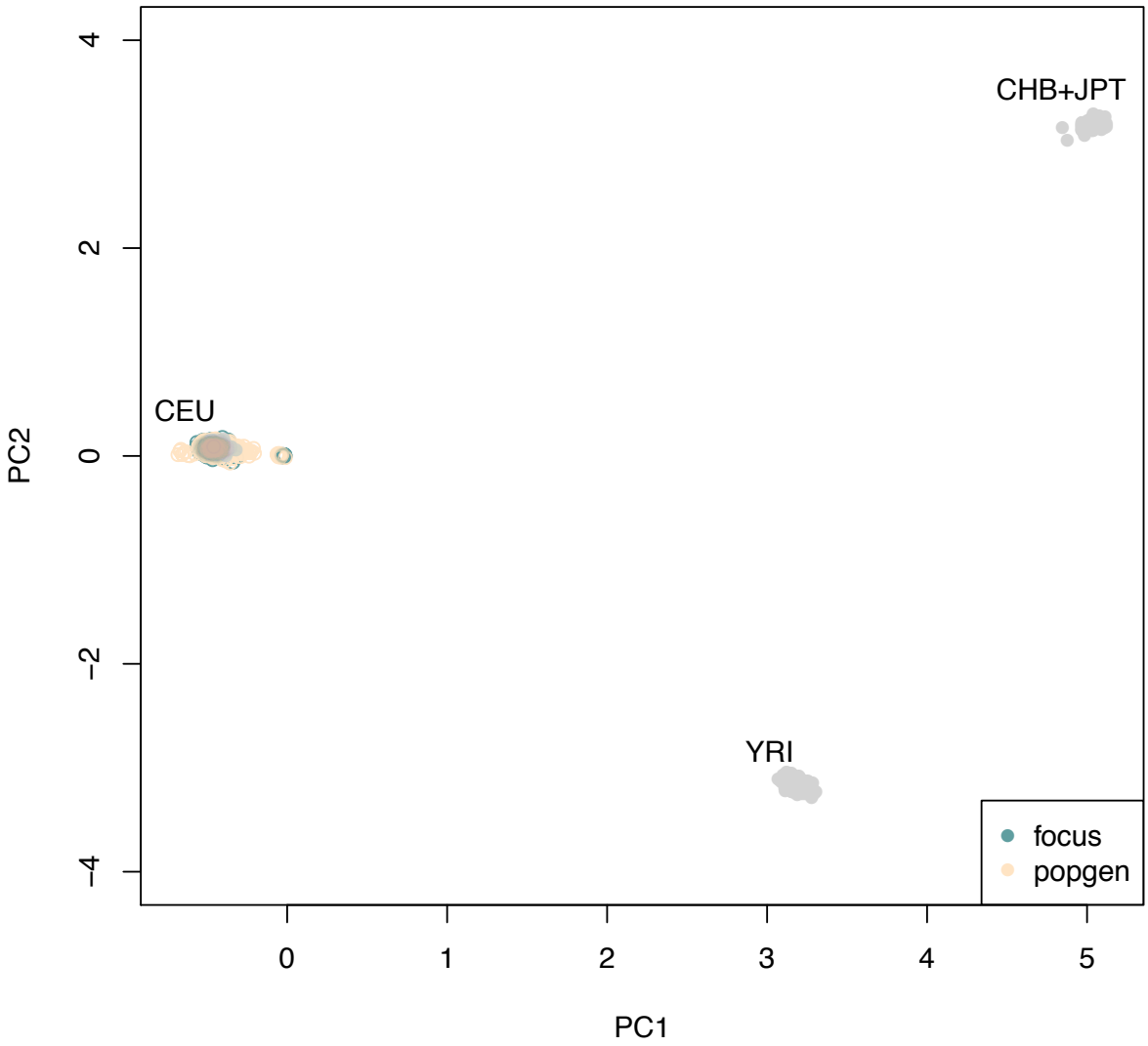


Figure S18: Principle component analysis that shows dense clustering of the PopGen and FoCus cohorts together with the HapMap CEU (Utah residents with Northern and Western European ancestry from the CEPH collection) samples. This suggests a large degree of genetic homogeneity in our study panels.

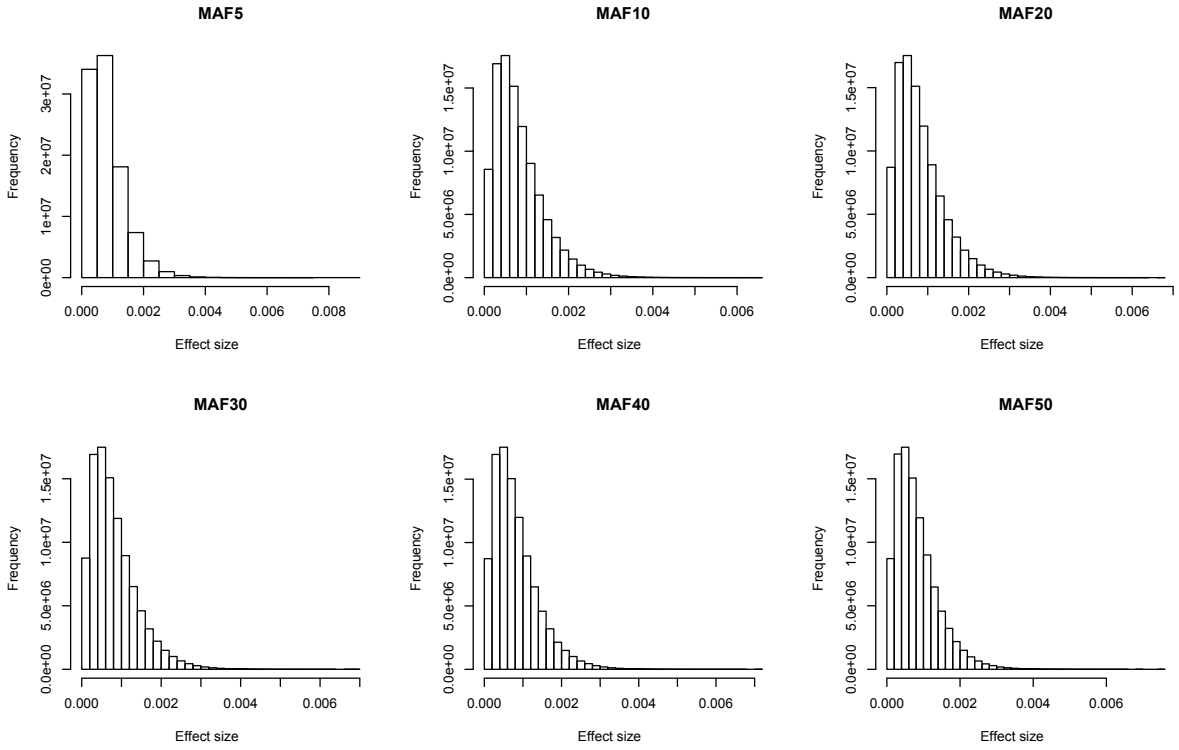


Figure S19: Null distribution of effect sizes for loci with MAF between 5%-50%. For each MAF category, a set of genotypes following Hardy-Weinberg equilibrium were simulated for the whole cohort and then permutated $>2\times10^7$ times to reach a maximum effect size observed by random chance. A real effect size higher than the determined maximum effect size is considered to pass the significance threshold and have $p<5\times10^{-8}$.

Supplementary Tables

Table S6: Significant microbial community differences according to HLA alleles. Results highlighted in bold indicate a significant influence of this allele on microbial communities after correction for anthropogenic confounding factors in distance based redundancy analyses of Bray-Curtis dissimilarity. HLA alleles highlighted with * show significant associations in both cohorts.

	PopGen				FoCus			
HLA-Haplotype	*F*-Value	*P*-Value	r^2	adj. r^2	*F*-Value	*P*-Value	r^2	adj. r^2
A 3001	1.5958	**0.0202**	0.00192	0.00073	0.88704	0.637	0.00094	-0.00012
A 6601	0.93014	0.563	0.00112	-0.00009	1.79968	**9.60×10^{-3}**	0.0019	0.00085
B 1302	1.59457	**0.0170**	0.00192	0.00072	0.64995	0.984	0.00069	-0.00037
B 2705	0.75539	0.903	0.00091	-0.0003	1.46044	**0.0462**	0.00154	0.00049
B 5201 *	1.61313	**0.0174**	0.00194	0.00075	1.48385	**0.0410**	0.00156	0.00051
C 0701	1.73167	**7.20×10^{-3}**	0.00209	0.00089	1.00964	0.404	0.00106	0.00001
C 1202 *	1.61313	**0.0158**	0.00194	0.00075	1.54459	**0.0282**	0.00163	0.00058
C 1402	0.92538	0.574	0.00112	-0.00009	1.47113	**0.0418**	0.00155	0.0005
DPB1 0202	-	-	-	-	1.51547	**0.0378**	0.0016	0.00055
DPB1 1601	1.02477	0.384	0.00124	0.00003	1.43189	**0.0474**	0.00151	0.00046
DQA1 0201	1.46264	**0.0390**	0.00176	0.00056	1.11954	0.249	0.00118	0.00013
DRB1 0701	1.50837	**0.0298**	0.00182	0.00062	1.11954	0.234	0.00118	0.00013
DRB1 1502	1.41422	**0.0202**	0.0017	0.0005	1.0215	0.637	0.00108	0.00002

Additional References:

55 Cardona, S. *et al.* Storage conditions of intestinal microbiota matter in metagenomic analysis. *BMC Microbiol*. 12:158 (2012).

56 Lauber, C.L. Zhou, N. Gordon, J.I., Knight, R. & Fierer, N. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiol. Lett.* 307(1):80-86 (2010).

57 Purcell, S. *et al.* PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81(3): 559–575(2007).

58 Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. *PLoS One*. 9(4):e93766(2014).

59 Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. Nat Methods. 9(2):179-81 (2012).

60 Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3: Genes, Genomics, Genetics.* 1(6): 457-470 (2011).

61 Jia, X. *et al.* Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS One*. 8(6):e64683 (2013).

62 Kozich, J.J. *et al.* Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microb*. 79:5112-5120 (2013).

63 Magoč, T. & Salzberg, S.L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 27:2957–2963 (2011).

64 Edgar, R.C, Haas, B.J., Clemente, J.C., Quince, C., & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27:2194–2200 (2011).

65 Wang Q., Garrity G.M., Tiedje J.M. & Cole J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73(16):5261-5267 (2007).

66 Edgar, R.C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods.* 10:996–998 (2013).

67 Abu-Hayyeh, S. *et al.* Prognostic and mechanistic potential of progesterone sulfates in intrahepatic cholestasis of pregnancy and pruritus gravidarum. *Hepatology.* 63(4):1287-1298 (2016).

68 Bjørndal, B. *et al.* Krill powder increases liver lipid catabolism and reduces glucose mobilization in tumor necrosis factor-alpha transgenic mice fed a high -fat diet. *Metabolism*, 61:1461-72 (2012).

69 Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* advanced online publication (2016).

70 Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* 46:1173-1186 (2014).

71 Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* 14:927-930 (2003).

72 Nöthlings, U. Fitting portion sizes in a self-administered food frequency questionnaire. *J. Nutr.* 137:2781–2786 (2007).

73 Dehne. L.I. The German food code and nutrient data base (BLS II.2). *Eur. J. Epidemiol.* 15:355–359 (1999).

74 Xu, L., Paterson, A.D., Turpin, W. & Xu, W. Assessment and selection of competing models for zero-inflated microbiome data. PLoS One. 10(7):e0129606 (2015).

75 Degenhardt, F. *et al.* Genome-wide association study of serum coenzyme Q10 levels identifies susceptibility loci linked to neuronal diseases. *Hum. Mol. Genet.* doi: 10.1093/hmg/ddw134 (2016).

76 Venables, W.N. & Ripley, B.D. Modern Applied Statistics with S (Fourth edition). *Springer, New York.* ISBN 0-387-95457-0 (2002).

77 Zeileis, A., Kleiber, C. & Jackman, S. Regression Models for Count Data in R. *J. Stat. Soft.* 27(8) (2008).

78 Bolger, A. M., Lohse, M., & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170 (2014).

79 Schmieder, R., & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* 6:e17288. (2011).

80 Trapnell, C. *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7:562–578 (2012).

81 Anders, S., Pyl, P.T., & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinforma. Oxf. Engl.* 31:166–9 (2015).

82 Love, M.I., Huber, W., & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550 (2014).

83 Breuer, K. *et al.* InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic Acids Res.* 41:D1228–1233 (2013).

84 Robertson, G. *et al.* cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res.* 34:D68–73 (2006).

85 Wacklin, P. *et al*. Secretor genotype (FUT2 gene) is strongly associated with the composition of Bifidobacteria in the human intestine. *PLoS One,*.6:e20113 (2011).

86 Evans, R.M. & Mangelsdorf, D.J. Nuclear receptors, RXR, and the Big Bang. *Cell*. 157(1):255-266 (2014).