nature genetics

LETTERS

https://doi.org/10.1038/s41588-020-00747-1

Check for updates

Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome

Malte Christoph Rühlemann[1], Britt Marie Hermes[2,3,4], Corinna Bang[1], Shauni Doms[2,3], Lucas Moitinho-Silva[1,5], Louise Bruun Thingholm[1], Fabian Frost[6], Frauke Degenhardt[1], Michael Wittig[1], Jan Kässens[1], Frank Ulrich Weiss[6], Annette Peters[7,8], Klaus Neuhaus[9], Uwe Völker[10], Henry Völzke[10], Georg Homuth[10], Stefan Weiss[6,10], Harald Grallert[7], Matthias Laudes[11], Wolfgang Lieb[12], Dirk Haller[9,13], Markus M. Lerch[6], John F. Baines[2,3] and Andre Franke[1]✉

The intestinal microbiome is implicated as an important modulating factor in multiple inflammatory[1,2], neurologic[3] and neoplastic diseases[4]. Recent genome-wide association studies yielded inconsistent, underpowered and rarely replicated results such that the role of human host genetics as a contributing factor to microbiome assembly and structure remains uncertain[5–11]. Nevertheless, twin studies clearly suggest host genetics as a driver of microbiome composition[11]. In a genome-wide association analysis of 8,956 German individuals, we identified 38 genetic loci to be associated with single bacteria and overall microbiome composition. Further analyses confirm the identified associations of ABO histo-blood groups and FUT2 secretor status with *Bacteroides* and *Faecalibacterium* spp. Mendelian randomization analysis suggests causative and protective effects of gut microbes, with clade-specific effects on inflammatory bowel disease. This holistic investigative approach of the host, its genetics and its associated microbial communities as a 'metaorganism' broaden our understanding of disease etiology, and emphasize the potential for implementing microbiota in disease treatment and management.

We conducted a large, single-country, genome-wide association study (GWAS) of microbial traits followed by Mendelian randomization (MR) analysis to elucidate the genetic link between humans and their associated microbiota. Our study comprised five independent cohorts from German biobanks located in northern Germany (Kiel, Schleswig-Holstein; PopGen[12], $n=724$; FoCus, $n=957$), north-eastern Germany (Greifswald, Mecklenburg-Western Pomerania; SHIP, $n=2,029$; SHIP-TREND, $n=3,382$)[13,14] and southern Germany (Augsburg, Bavaria; KORA, $n=1,864$;[15,16] see Methods for details).

Baseline comparisons show similarities in anthropometric measures, genomic variation and microbial community compositions between cohorts (Fig. 1, Supplementary Fig. 1 and Supplementary Note). Taxonomic groups and sequence similarity clusters included in the univariate analysis, henceforth called microbial features, covered between 98.4% and 98.7% of the whole community at the phylum level and between 77.8% (PopGen) and 82.6% (SHIP-TREND) at the genus level across cohorts. These data indicate that the cohorts share a common core microbiota (cohort-level summaries of microbial features can be found in Supplementary Table 1).

Univariate microbial features were defined based on taxonomic annotations from phylum to genus level. As taxonomic assignments below the genus level do not perform well[17], finer-scale features were defined by sequence similarity clustering (97% and 99% similarity) and amplicon sequence variants (ASVs) to create a comprehensive dataset (Methods). Host features encoded by genetics can possibly influence presence–absence patterns of microorganisms, and also lead to shifts in the relative abundances of such, so both assumptions were tested in the association analysis. In total, 198 and 233 univariate microbial features were analyzed using logistic and linear regression, respectively (see Methods for details). Host–genetic variation might affect multiple community members, so, in addition to univariate analyses, whole-community multivariate association analysis of genus-level Bray–Curtis dissimilarity and weighted UniFrac distance[18] was performed (Methods, Supplementary Note and Supplementary Fig. 2). Per-cohort results were combined in a meta-analysis framework (Methods). To ensure robustness of results, genome-wide significant results ($P_{Meta}<5\times10^{-8}$) were reported when supported by nominal significance ($P<0.05$) in at least two cohorts. In addition, a study-wide significance threshold was defined as $P_{Meta}<1.866\times10^{-10}$ and heterogeneity measures were calculated (Methods).

Accordingly, we identified a total of 44 genome-wide significant associations with microbial features and community composition involving 38 genomic loci (Table 1 and Fig. 2a), among which 4 associations stemmed from the multivariate analysis, 17 from the univariate abundance analysis and 17 from the presence–absence

[1]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany. [2]Evolutionary Genomics, Max Planck Institute for Evolutionary Biology, Plön, Germany. [3]Institute of Experimental Medicine, Kiel University, Kiel, Germany. [4]Lübeck Institute of Experimental Dermatology, University of Lübeck, Lübeck, Germany. [5]Department of Dermatology, Kiel University, Kiel, Germany. [6]Department of Medicine A, University Medicine Greifswald, Greifswald, Germany. [7]Institute of Epidemiology, Helmholtz Zentrum München, Neuherberg, Germany. [8]German Center for Diabetes Research, Neuherberg, Germany. [9]ZIEL—Institute for Food & Health, Technical University of Munich, Freising, Germany. [10]Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany. [11]Department of Internal Medicine 1, Kiel University, Kiel, Germany. [12]Institute of Epidemiology, Kiel University, Kiel, Germany. [13]Chair of Nutrition and Immunology, Technical University of Munich, Freising, Germany. ✉e-mail: a.franke@mucosa.de

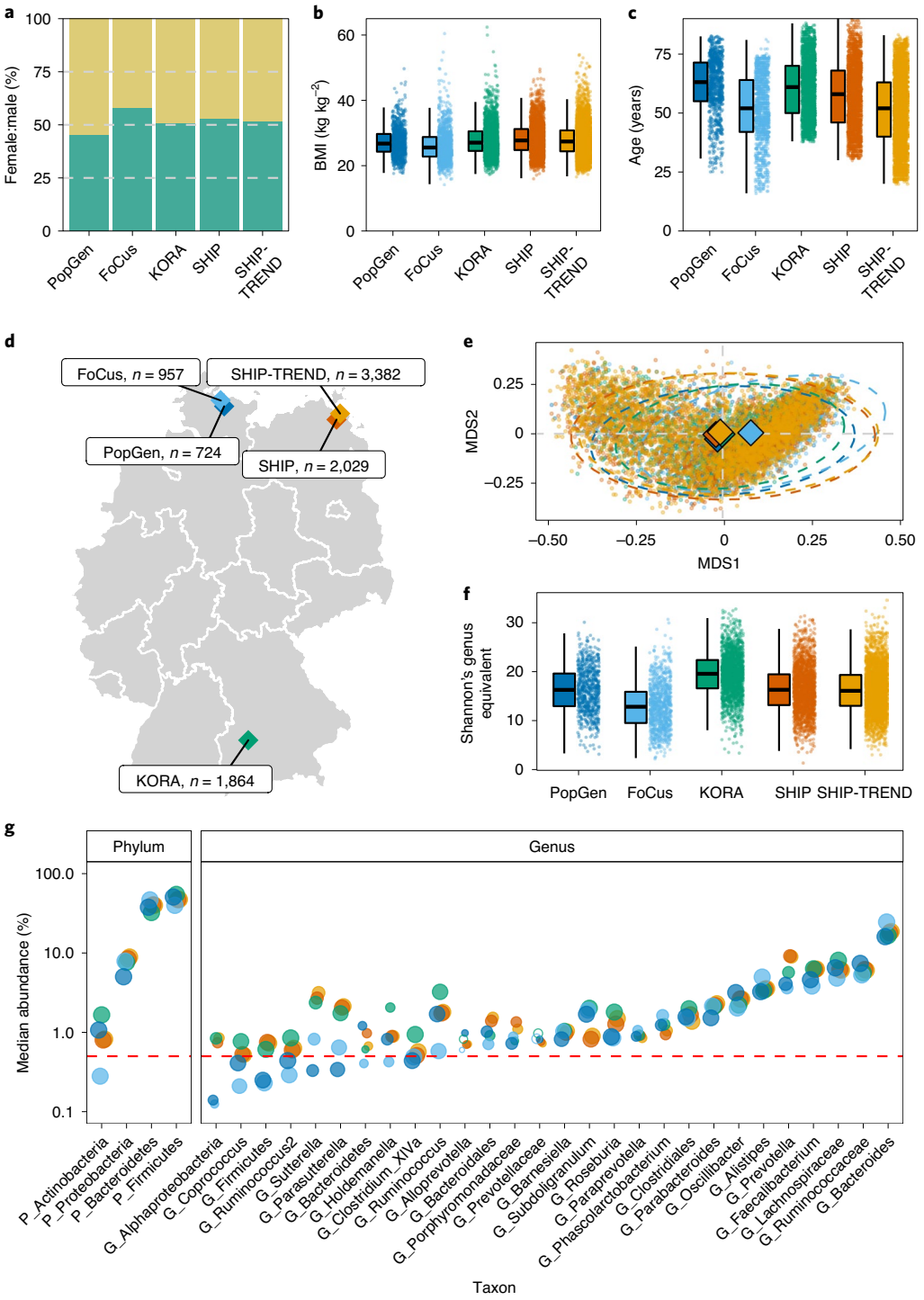

Fig. 1 | Summary of cohort properties. All included cohorts were based in German biobanks: PopGen (Kiel, $n = 724$), FoCus (Kiel, $n = 957$), KORA (Augsburg, $n = 1{,}864$), SHIP (Greifswald, $n = 2{,}029$) and SHIP-TREND (Greifswald, $n = 3{,}382$). **a**, Composition of study participants by sex. **b**,**c**, Distribution of participants' BMI (**b**) and age (**c**). **d**, Map showing biobank/cohort locations in Germany. The map was created using the ggmap package (https://github.com/dkahle/ggmap) and geographic coordinates obtained using the rgdal package (https://CRAN.R-project.org/package=rgdal). **e**, Ordination of all samples based on genus-level Bray–Curtis dissimilarity using principal coordinate analysis. The first and second axes explaining the highest amount of variance, MDS1 and MDS2, are shown. Diamonds represent cohort centroids and dashed ellipses represent the 95% confidence level of multivariate *t*-distributions. **f**, Distribution of α diversities as calculated by Shannon's diversity genus-level equivalent and the number of observed genera. **g**, Comparison of relative abundances of phylum- and genus-level taxonomic groups that met the inclusion criteria for the GWAS in the five analyzed German cohorts. The *y* axis represents the median abundance of the samples with nonzero abundance of the respective taxa; point size is relative to the prevalence of the respective taxon in the cohort. Taxa with cohort prevalence below the inclusion threshold of 20% are displayed as empty circles. The dashed red line represents the abundance threshold of 0.5% for inclusion in the analysis. Taxa are arranged from left to right by the lowest median abundance over all cohorts from high to low. Cohort-level summaries of microbial features can be found in Supplementary Table 1. In **b**, **c** and **f**, the center lines in the box-and-whisker plots represent median values, box limits show the 1st and 3rd quartiles, and whiskers extend from the 1st/3rd quartile to the last datapoint within ±1.5× IQR.

Table 1 | Results summary of the GWAS for β diversity (column 'Analysis': β), logistic regression of presence–absence patterns (LR) and analysis of abundances (NB) ordered, and enumerated by genomic location of the loci

Locus	Analysis	uniqID	rsID	Major	MAF	Features	Chr	Position	Effect allele	*P* value	β	s.e.	*n*	I^2 (lead SNP)	Genes in locus (±100 kb)
1	NB	1:100446046:A:G	rs1571350	A	0.425	G_Alistipes	1	100446046	G	3.945×10^{-8}	−0.1391	0.0253	8538	0.226	*AGL*, ***SLC35A3***, *HIAT*
2	LR	2:71268031:A:T	rs10195730	A	0.244	TestASV_30 (G_Paraprevotella)	2	71268031	T	3.370×10^{-9}	0.4971	0.0841	3487	0	*ATP6V1B1*, *ANKRD53*, *TEX261*, ***OR7E91P***, *NAGK*, *MCEE*, *MPHOSPH10*
3	NB	2:171151691:A:G	rs6751051	G	0.1078	C_Betaproteobacteria, O_Burkholderiales	2	171151691	G	1.597×10^{-8}	−0.1418	0.0252	8381	0	*MYO3B*
4	NB	2:239153765:A:T	rs35093813	T	0.0895	C_Alphaproteobacteria, G_Alphaproteobacteria	2	239153765	T	1.111×10^{-8}	−0.2607	0.0456	2903	0.607	*KLHL30*, *FAM132B*, *ILKAP*, *LOC151174*, *LOC643387*, *HES6*, ***PER2***, *TRAF3IP1*
5	LR	3:60225409:A:C	rs213388	A	0.2411	TestASV_15 (G_Bacteroides)	3	60225409	C	1.492×10^{-8}	−0.2259	0.0399	8930	0	*FHIT*
6	NB	4:40481757:C:T	rs424048	C	0.5460	TestASV_27 (F_Ruminococcaceae)	4	40481757	T	4.588×10^{-8}	−0.2146	0.0393	1345	0.682	*RBM47*
7	NB	4:60918325:A:G	rs11729574	G	0.0717	OTU97_11 (G_Parabacteroides)	4	60918325	G	1.193×10^{-8}	0.2267	0.0398	4832	0.449	-
8	LR	4:82818818:A:G	rs10000747	G	0.0617	OTU97_11 (G_Parabacteroides)	4	82818818	G	8.398×10^{-9}	0.3956	0.0687	7733	0	-
9	β	4:137653726:C:T	rs1585404	C	0.4150	Bray–Curtis	4	137653726	T	4.176×10^{-8}	-	-	8612	-	-
10	LR	5:8438531:C:T	rs340669	C	0.2910	OTU97_80 (G_Ruminococcus), OTU99_92 (G_Ruminococcus)	5	8438531	T	1.710×10^{-8}	−0.2178	0.0400	8889	0	*LOC729506*
11	LR	5:57935865:G:T	rs158966	G	0.1973	OTU97_51 (G_Barnesiella), G_Barnesiella,	5	57935865	T	1.517×10^{-8}	0.2524	0.0446	8914	0	*RAB3C*
12	LR	6:90978161:C:T	rs62408234	C	0.1411	OTU97_27 (G_Bacteroides)	6	90978161	T	4.575×10^{-10}	0.3634	0.0583	5582	0	*BACH2*
13	LR	6:140101119:C:T	rs6900161	T	0.0565	OTU97_23 (G_Faecalibacterium)	6	140101119	T	2.219×10^{-8}	−0.7158	0.1280	7244	0.114	*FILNC1*
14	NB	7:43864699:A:G	rs623108	G	0.3567	OTU99_55 (G_Barnesiella)	7	43864699	G	1.045×10^{-8}	−0.1664	0.0291	2743	0	*COA1*, ***BLVRA***, *MRPS24*, *URGCP*
15	LR	7:85818086:C:T	rs17160658	T	0.1535	TestASV_48 (G_Sutterella)	7	85818086	T	4.438×10^{-9}	0.5136	0.0875	7207	0	-
16	NB	7:117721635:C:T	rs10235327	C	0.4526	OTU99_30 (G_Parasutterella)	7	117721635	T	2.550×10^{-8}	−0.1504	0.0270	2734	0	-
17	NB	8:5719816:A:G	rs12679403	G	0.3337	TestASV_26 (G_Phascolarctobacterium)	8	5719816	G	2.038×10^{-8}	0.3961	0.0706	460	0.545	-
18	LR	8:22906641:A:G	rs67791511	A	0.2907	OTU97_34 (G_Ruminococcus), OTU99_35 (G_Ruminococcus)	8	22906641	G	1.486×10^{-8}	−0.2433	0.0431	5777	0	*RHOBTB2*, ***TNFRSF10B***, *LOC286059*, *LOC254896*, *TNFRSF10C*, *TNFRSF10D*
19	LR	8:112651697:A:G	rs4876786	G	0.1585	G_Sutterella	8	112651697	G	1.829×10^{-8}	0.2495	0.0443	8200	0	-
20	NB	9:7545825:A:G	rs62530076	A	0.1070	OTU97_15 (G_Parasutterella)	9	7545825	G	4.588×10^{-8}	0.1797	0.0329	4825	0	-
21	LR	9:22175188:C:G	rs10965279			TestASV_20 (G_Phascolarctobacterium)	9	22175188	G	4.967×10^{-8}	0.5047	0.0926	8186	0	*CDKN2B-AS1*

Continued

Table 1 | Results summary of the GWAS for β diversity (column 'Analysis': β), logistic regression of presence–absence patterns (LR) and analysis of abundances (NB) ordered, and enumerated by genomic location of the loci (Continued)

Locus	Analysis	uniqID	rsID	Major	MAF	Features	Chr	Position	Effect allele	*P* value	β	s.e.	*n*	I^2 (lead SNP)	Genes in locus (±100 kb)
22	LR	9:136152547:C:T	rs8176632	C	0.1686	OTU97_27 (G_Bacteroides)	9	136152547	T	6.866×10^{-10}	0.3142	0.0509	6100	0	*OBP2B*, ***ABO***, *SURF6*, *MED22*, *RPL7A*, *SURF1*, *SURF2*, *SURF4*, *C9orf96*
23	NB	9:136239399:C:G	rs3758348	G	0.1516	OTU99_16 (G_Faecalibacterium)	9	136239399	G	4.332×10^{-9}	−0.1434	0.0244	6559	0.433	***ABO***, *SURF6*, *MED22*, *RPL7A*, *SURF1*, *SURF2*, ***SURF4***, *C9orf96*, *REXO4*, *ADAMTS13*, *CACFD1*, *SLC2A6* [a]
24	NB	10:112954252:A:G	rs4439447	A	0.3334	C_Clostridia	10	112954252	G	2.861×10^{-8}	0.0878	0.0158	8821	0	–
25	NB	11:119792443:C:T	rs71484153	T	0.2799	TestASV_21 (F_Ruminococcaceae)	11	119792443	T	3.947×10^{-8}	0.1640	0.0299	2776	0	–
26	NB	11:134761316:A:G	rs10894898	A	0.3893	OTU99_4 (G_Alistipes), TestASV_4 (G_Alistipes)	11	134761316	G	1.651×10^{-8}	−0.1147	0.0203	5513	0	–
27	Beta	12:30561406:A:G	rs7954208	A	0.0603	Bray–Curtis	12	30561406	G	4.667×10^{-9}	–	–	8903	–	–
28	NB	12:97768678:C:T	rs12304493	C	0.4571	G_Alloprevotella	12	97768678	T	1.798×10^{-8}	0.1922	0.0341	1738	0.568	*RMST*
29	LR	13:46265207:C:G	rs7981551	G	0.3511	OTU97_56 (F_Ruminococcaceae)	13	46265207	G	3.303×10^{-8}	0.1882	0.0341	8390	0	*FAM194B*, ***SPERT***, *SIAH3*
30	LR	13:107517622:A:G	rs8000335	G	0.09352	OTU97_109 (G_Paraprevotella)	13	107517622	G	4.693×10^{-8}	−0.3318	0.0607	8640	0	–
31	NB	14:38073877:A:T	rs1956429	A	0.3944	F_Rikenellaceae	14	38073877	T	8.316×10^{-9}	−0.0917	0.0159	8464	0	*MIPOL1*, ***FOXA1***, *C14orf25*
32	NB	15:23999122:C:G	rs72703939	G	0.2171	TestASV_37 (F_Ruminococcaceae)	15	23999122	G	3.567×10^{-8}	−0.3657	0.0664	615	0	*NDN*
33	Beta	15:93820994:A:G	rs12592825	A	0.3616	Bray–Curtis	15	93820994	G	4.552×10^{-8}	–	–	7837	–	–
34	NB	16:23372110:G:T	rs185150	G	0.0673	TestASV_16 (G_Bacteroides)	16	23372110	T	2.476×10^{-8}	0.6011	0.1078	669	0.541	***SCNN1B***, *COG7*
35	LR	17:53069650:A:G	rs4794550	A	0.2829	TestASV_16 (G_Bacteroides)	17	53069650	G	1.707×10^{-8}	−0.2978	0.0538	8864	0	*TOM1L1*, *COX11*, ***STXBP4***
36	LR	17:65565305:G:T	rs3760215	G	0.4492	OTU99_94 (G_Bacteroides)	17	65565305	T	2.169×10^{-8}	−0.4044	0.0722	6659	0.5023	*PITPNC1*
37	Beta	19:4855248:C:T	rs11880778	C	0.2103	Bray–Curtis	19	4855248	T	7.974×10^{-9}	–	–	8664	–	*FEM1A*, *TICAM1*, ***PLIN3***, *ARRDC5*, *C19orf31*, *UHRF1*
38	LR	20:34011645:C:T	rs6060406	C	0.0593	OTU97_117 (G_Ruminococcus)	20	34011645	T	2.810×10^{-8}	0.6333	0.1141	7267	0	*UQCC*, ***GDF5***, *GDF5OS*, *CEP260*

[a]In this locus, two signals in weak LD (<0.4) are found, one close to *SURF4*, the other close to *ABO* (Fig. 2e). A single-variant association test was performed for each cohort and each microbial feature, adjusting the respective model for the first ten genetic PCs, age, sex and BMI. All positions are given relative to the human genome assembly hg19 (GRCh37). Results were meta-analyzed, weighted by inverse variance for univariate and sample size in multivariate nonparametric models. For univariate analysis, meta-analysis effect size (β) and s.e. are given with respect to the effect allele, and total sample numbers in the meta-analysis are given in column '*n*'. A genome-wide significance threshold of $P_{Meta}<5\times10^{-8}$ and nominal significance ($P<0.05$) in at least two cohorts was considered to ensure robustness. All tests were performed two sided. I^2 values for the lead SNPs are given as a measure of heterogeneity. Genes of up to 100 kb up- and downstream of the lead SNP are listed; in case multiple genes are found in the locus, the closest gene to the lead SNP is marked in bold.

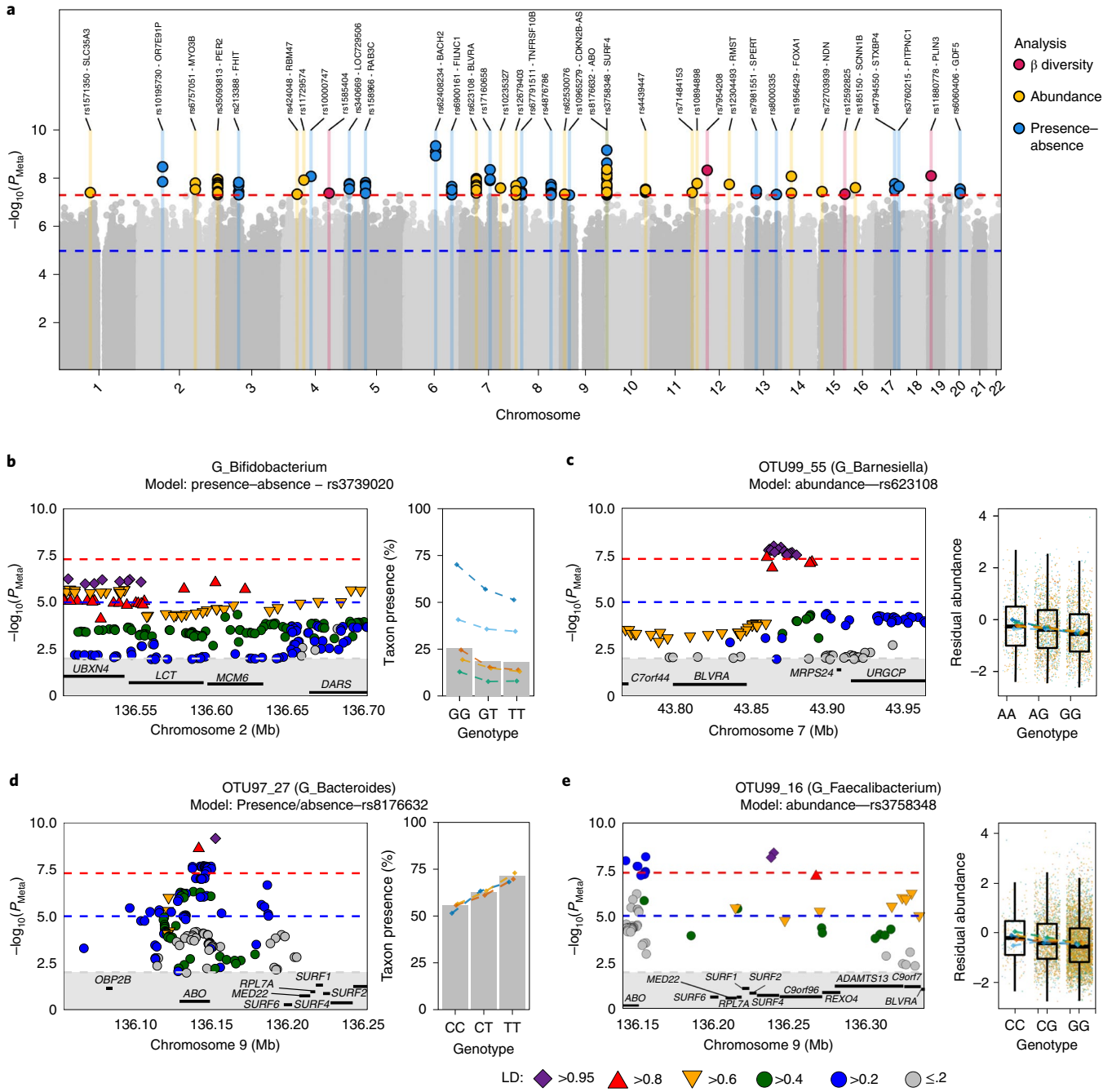

Fig. 2 | Results from the GWAS. a, Manhattan plot of *P* values from the meta-analysis across all tested traits; the lowest *P* value at each position is shown. Color coding is by analysis type: yellow: abundance models; blue: presence–absence models (logistic regression); red: β diversity. **b**, Regional association plot of genus *Bifidobacterium* presence–absence test with variants in the *LCT* gene locus. **c**, OTU97_55 (*Barnesiella* sp.) abundance versus variants at the biliverdin reductase A (*BLVRA*) gene locus (n_{Meta} = 2,743). **d**, OTU99_16 (*Faecalibacterium* sp.) abundance versus variants in the *ABO/SURF4* gene locus. **e**, OTU97_27 (*Bacteroides* sp.) presence–absence versus *ABO* variants (n_{Meta} = 6,100). The per-cohort feature abundance means and presences for each genotype are given by the diamonds in the respective colors. In **b** and **d**, barplots denote overall feature prevalence by genotype across all n = 8,956 samples. In **c** and **e**, all residual abundance for the individual samples is displayed as dots in the respective colors of the cohort. Center lines in box-and-whisker plots represent median values, box limits show 1st and 3rd quartiles, and whiskers extend from the 1st/3rd quartile to the last datapoint within ±1.5× IQR. All *P* values are the results of logistic and linear regression for prevalence- and abundance-based models, respectively. All tests were two sided.

patterns. Most genome associations—including those found in the presence–absence models—showed low heterogeneity (I^2 < 40%), with only six abundance-associated variants showing moderate heterogeneity (I^2 < 60%) and two surpassing this threshold, and thus should be interpreted with caution. The top 10,000 genetic variants for univariate and multivariate analyses are summarized in Supplementary Tables 2–4. All results can be queried via the metabolite (m)GWAS results browser (http://ikmb.shinyapps.io/German_mGWAS_Browser). None of the signals surpassed the conservative threshold of study-wide significance. Univariate signals with overlapping genetic loci in all cases are found from the same taxonomic group at a different taxonomic and/or clustering level.

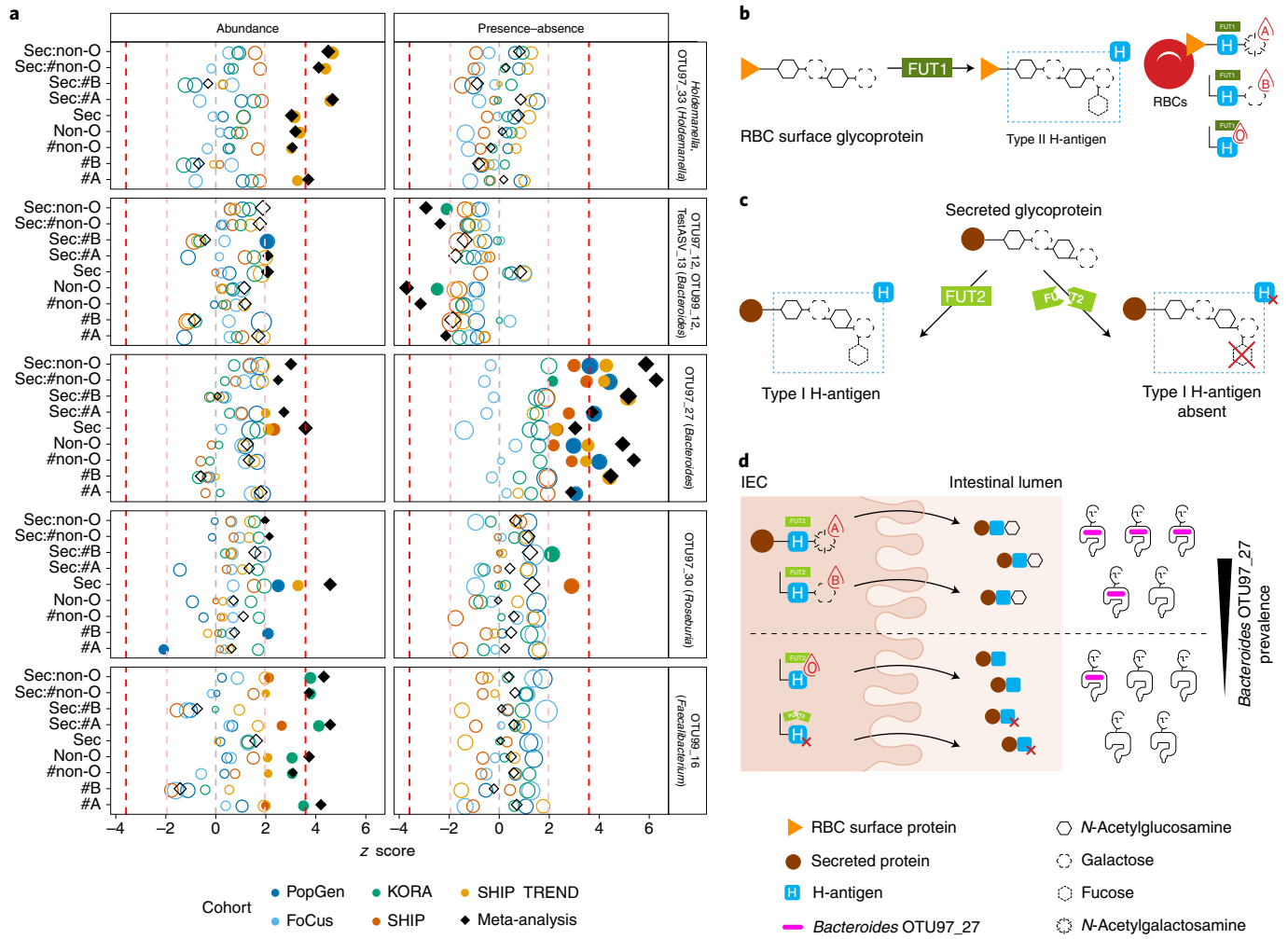

Fig. 3 | ABO histo-blood groups show connection to gut microbial features. a, Results of the Hurdle-model analysis with negative binomial distribution of the nonzero counts of nine models associating feature prevalence and abundance to ABO blood group alleles and FUT2 secretor status (Sec) in 8,956 individuals in five German cohorts. All tests were two sided. All univariate microbial features are shown that have at least one meta-analysis q_{FDR} value <0.05. Taxonomic classifications of OTUs/ASVs are given in brackets. The *y* axis represents the nine models investigated; interactions are indicated by the colon symbol; the *x* axis shows *z*-scores of the respective models. Symbols are colored according to cohort: black diamonds represent the result of the meta-analysis of all five cohorts. Symbol size represents absolute effect size. *P* values <0.05 are displayed as solid shapes. Dashed vertical lines represent *z* values corresponding to nominal significance (light-red line: $P<0.05$; $z=\pm1.96$; two sided) and adjusted significance (dark-red line: $q_{FDR}<0.05$; $z=\pm3.59$; two sided). Complete results can be found in Supplementary Table 5. Non-O: binary trait, blood group O versus non-O blood groups (A, B, AB); #non-O: 'quantitative' trait, number of non-O alleles, i.e., OO = 0, AO = 1, AA = 2. **b**, The type II H-antigen on red blood cells (RBCs) is completed via addition of a fucose sugar by the enzyme FUT1. Subsequently the A- and B-antigens are synthesized by addition of *N*-acetylgalactosamine or galactose, respectively. In individuals with an O histo-blood group, no additional sugars can be added to the H-antigen. **c**, On secreted proteins and mucosal cells, the fucosylated type I H-antigen is synthesized by the enzyme FUT2. In individuals homozygous for the rs601338:G > A missense variant in *FUT2*—also known as nonsecretors—there is no addition of a fucosyl group, resulting in no H-antigen. **d**, Consequently, no additional sugars are added to the precursor of the H-antigen in nonsecretors, irrespective of the individual's histo-blood group genotype at ABO. *Bacteroides* OTU97_27 exhibits higher prevalence in individuals with non-O ABO histo-blood groups and functioning FUT2 compared with individuals with O histo-blood group or FUT2 nonsecretors. IEC, intestinal epithelial cells.

Although not meeting the initial inclusion criteria (Methods), the genus *Bifidobacterium* was included in the analysis. Its connection with the lactase gene locus (*LCT*) on chromosome 2 is important, because it is the only signal replicating across numerous previous studies[5,9,11]. The meta-analysis shows a clear association peak in the *LCT* locus with 53 variants displaying *P* values lower than the suggested $P<10^{-5}$ threshold, with the lowest for rs3820794 (chr2:136505546; $P_{Meta}=5.62\times10^{-7}$; Fig. 2b). This is supported by nominally significant *P* values in four of the five cohorts, with only the FoCus cohort showing a *P* value above nominal significance ($P_{FoCus}=0.069$), underlining the previously found connection between the *LCT* locus and *Bifidobacterium* spp. and the validity of the herein-used model of choice, although a linkage disequilibrium (LD) structure does not pinpoint the *LCT* gene itself as the location of the primary association. Also, connections to age and consumption of dairy products remain unresolved and need to be investigated through more targeted approaches[11,19].

Our obtained genome-wide association results point to immune-mediated interactions of host and microbiota, for example the association detected for OTU99_55 (*Barnesiella* spp.; OTU: operational taxonomic unit) and variants in the biliverdin reductase A (*BLVRA*; rs623108; $P_{Meta}=1.05\times10^{-8}$; Fig. 2c) locus. Biliverdin reductase A was previously shown to inhibit toll-like receptor 4 (*TLR4*) gene expression[20]. TLR-4 is a pattern recognition receptor

Table 2 | Results from MR analysis

Outcome	Exposure	Analysis	Method	No. of SNPs	*F* (mean)	*F* (min.)	β	s.e.	*P* value	*q* value
Anthropometric										
Extreme BMI || id:85	P_Firmicutes	NB	IVW	2	20.17	18.81	−0.8804	0.2634	8.31×10^{-4}	0.4044
Obesity class 2 || id:91	G_Parabacteroides	NB	IVW	3	18.32	18.09	−0.5683	0.1659	6.14×10^{-4}	0.3350
Autoimmune / inflammatory										
Asthma || id:44	OTU99_84 (G_Prevotella)	LR	WR	1	11.77	11.77	−0.7256	0.2109	5.82×10^{-4}	0.3927
Celiac disease || id:1059	OTU99_62 (F_Ruminococcaceae)	LR	WR	1	21.22	21.22	−0.6131	0.1763	5.07×10^{-4}	0.3764
Crohn's disease || id:10	C_Clostridia	NB	WR	1	19.28	19.28	−1.7060	0.2465	4.46×10^{-12}	4.28×10^{-8}
Crohn's disease || id:10	OTU97_27 (G_Bacteroides)	LR	WR	1	17.45	17.45	−0.5151	0.0867	2.77×10^{-9}	1.33×10^{-5}
Crohn's disease || id:10	TestASV_23 (G_Barnesiella)	LR	WR	1	15.17	15.17	0.2711	0.0599	6.00×10^{-6}	0.0115
Crohn's disease || id:10	TestASV_18 (G_Prevotella)	LR	WR	1	13.42	13.42	−0.2575	0.0579	8.76×10^{-6}	0.0140
Crohn's disease || id:11	F_Porphyromonadaceae	NB	IVW	2	21.01	20.43	3.2134	0.7962	5.44×10^{-5}	0.0745
Crohn's disease || id:11	TestASV_12 (G_Bacteroides)	LR	WR	1	20.44	20.44	−1.0344	0.3160	1.06×10^{-3}	0.5672
IBD || id:293	F_Porphyromonadaceae	NB	IVW	2	21.01	20.43	2.5143	0.5433	3.70×10^{-6}	8.88×10^{-3}
IBD || id:293	TestASV_12 (G_Bacteroides)	LR	WR	1	20.44	20.44	−0.9989	0.2654	1.68×10^{-4}	0.1786
IBD || id:293	OTU99_85 (G_Alistipes)	LR	WR	1	17.18	17.18	0.4096	0.0776	1.29×10^{-7}	4.13×10^{-4}
Cancer										
Ovarian cancer || id:1120	OTU97_27 (G_Bacteroides)	LR	IVW	5	14.54	13.44	−0.1140	0.0341	8.39×10^{-4}	0.4740
Gallbladder cancer || id:1057	OTU97_4 (G_Alistipes)	NB	IVW	4	14.71	13.52	5.8987	1.5071	9.08×10^{-5}	0.1090
Cardiovascular										
Coronary heart disease || id:6	TestASV_23 (G_Barnesiella)	LR	IVW	4	17.78	15.17	0.1497	0.0431	5.10×10^{-4}	0.3764
Psychiatric/neurologic										
Autism || id:802	TestASV_11 (F_Lachnospiraceae)	NB	IVW	7	11.44	10.60	0.4156	0.1151	3.07×10^{-4}	0.2945
Major depressive disorder || id:804	OTU97_51 (G_Barnesiella)	NB	WR	1	16.49	16.49	0.8655	0.2447	4.05×10^{-4}	0.3538
Schizophrenia || id:22	F_Lachnospiraceae	NB	IVW	8	21.45	19.96	0.1687	0.0521	1.20×10^{-3}	0.6069

Only results with $P<1.220\times10^{-3}$ are shown (significance threshold as determined in Methods) and the respective FDR-adjusted *q* values. All SNPs with *F*-statistics >10 and $P<10^{-5}$ in the respective genome-wide association meta-analysis of presence–absence (LR) and abundance (NB) patterns (exposures) were used as instrument variables, and tested for their effects on 41 binary traits (Methods and Supplementary Note). Mean and minimum (min.) *F*-statistics of included instruments are reported. Tests used for MR (Method) were Wald's ratio (WR) in the case of a single instrument variable and IVW analysis in the case of two or more instrument variables (No. of SNPs). All tests were performed two sided. Effect sizes (β) and s.e. of the primary analyses are reported. A complete table of all results from all MR analyses, including results from the sensitive analysis, can be found in Supplementary Table 6.

that initiates an immune response to bacterial lipopolysaccharides present in many Gram-negative bacteria[21]. *Barnesiella* sp., which itself is Gram negative, is negatively associated with lipopolysaccharide-induced interferon-γ production, suggesting a contribution of this commensal to homeostasis by immune or TLR-4 signal modulation.

We identified two independent univariate associations with a locus surrounding the histo-blood group ABO system transferase (*ABO*) gene. One *ABO* gene signal for differential abundance includes OTU99_16 belonging to *Faecalibacterium* sp. (rs3758348; chr9:136155000; $P_{Meta}=6.16\times10^{-9}$; Fig. 2d), which is accompanied by a second signal ~100 kb downstream in the surfeit locus protein 4 (*SURF4*) gene (chr9:136239399; $P_{Meta}=4.33\times10^{-9}$). The second *ABO* association is between rs8176632 allele T and the increased prevalence of a *Bacteroides* OTU (OTU97_27; rs8176632; chr9:136152547; $P_{Meta}=6.87\times10^{-10}$; Fig. 2e). It is interesting that this same *Bacteroides* OTU is also significantly associated with variants at the *BACH2* (BTB domain and CNC homolog 2) gene locus (chr6:90978161; $P_{Meta}=4.58\times10^{-10}$). Moreover, a suggestive association between this *Bacteroides* OTU is present for the *FUT2* (galactoside 2-α-L-fucosyltransferase 2) locus, whereby the strongest signal is from the missense variant rs602662 (chr19:49206985; $P_{Meta}=4.46\times10^{-7}$), which is in strong LD with variant rs601338 ($R^2=0.8898$) encoding the FUT2 secretor phenotype. This variant determines whether the fucosyl precursor for the ABO blood group system is synthesized on mucosal surfaces in the body and secretions. Individuals homozygous for this missense variant do not have the ABO-encoded antigen on mucosal cells, independent

of the ABO allele (that is, display the nonsecretor phenotype; Fig. 3b–d). Variants at *FUT2* and *BACH2*, correlated with *Bacteroides* OTU97_27 in the present study, were previously shown to be associated with inflammatory bowel disease (IBD)[22–25].

For a focused evaluation of blood group-dependent associations with microbial features, we investigated ABO histo-blood group and FUT2 secretor status (Methods). The prevalence or abundance of eight taxonomic groups shows at least one false discovery rate (FDR)-corrected significant association ($q < 0.05$) with ABO histo-blood group alleles, secretor status or their interaction (Fig. 3a and Supplementary Table 5). These results demonstrate a positive correlation between non-O blood group and positive secretor status and the prevalence of the aforementioned *Bacteroides* OTU97_27 in four of the five cohorts ($P_{Meta} = 3.65 \times 10^{-10}$). Notably, a different *Bacteroides* branch, represented by OTU97_12, OTU99_12 and TestASV_13, exhibited significant associations with ABO histo-blood group status as well, although, in this case, characterized by an inverse relationship between prevalence and non-O blood group alleles ($P_{Meta} = 2.1 \times 10^{-4}$). Together, these findings suggest histo-blood group-dependent effects on *Bacteroides* subclades.

In addition, the model points to an association between *Faecalibacterium* OTU99_16 and the ABO histo-blood group A allele in interaction with secretor status ($P_{Meta} = 4.7 \times 10^{-6}$). A significant association between *Holdemanella* sp. and ABO is also identified, although the signal is exclusively driven by the SHIP-TREND cohort with only weak support from the remaining cohorts. Furthermore, FUT2 secretor status is associated with differential abundance of *Roseburia* OTU97_30, independent of ABO blood type ($P_{Meta} = 4.79 \times 10^{-6}$). In conclusion, the analyses show a specific impact of the human ABO blood groups and secretor status on members of the intestinal community.

MR has recently become a popular tool to infer causal relationships of complex traits in observational data[26], and recent publications suggest that MR can be used for exploratory inference of causal effects that the microbiome may have on complex host traits[27]. MR analysis was performed for all univariate microbial features as 'exposures' and 41 selected binary traits from the MR-Base database[28] as outcomes (Methods and Supplementary Note). This allows us to assess the potential causal effect of microbial features on disease. A total of 19 comparisons reach the per-trait suggestive threshold of $P < 1.22 \times 10^{-3}$, with five traits falling below the global FDR-correction threshold $q < 0.05$ (Table 2 and Supplementary Table 6). Of 19 suggestive microbial effects on host traits, 9 point to IBD and its subentity Crohn's disease. For example, the presence of the same *Bacteroides* OTU associated with ABO histo-blood group status (OTU97_27) and a *Prevotella* ASV (TestASV_18) appear to significantly protect against Crohn's disease development ($\beta = -0.515$ and $\beta = -0.257$, respectively). Previous work has identified *Bacteroides* and *Prevotella* spp. as the main determinants of gut enterotypes[29–31]. Recent studies applying quantitative microbiome profiling suggested that the protective effects of *Prevotella* sp. dominated communities on Crohn's disease, as well as contradicting connections of IBD to *Bacteroides* subclades, potentially modulated by microbial load[32], supported by additional studies pointing to lower abundances of *Bacteroides* sp. being associated with IBD development[33]. These results once again emphasize the large variability of *Bacteroides* taxa in connection to genetics and disease.

Further results from MR confirm host–microbiome interactions previously described in observational studies. *Parabacteroides* spp. show a protective effect on the 'obesity class 2' trait ($\beta = -0.568$), supporting previous experimental observations of *Parabacteroides* spp. alleviating obesity effects in mice[34]. It is interesting that none of the microbial traits with causal effects reaches genome-wide significance at any locus in the univariate analysis. In addition to MR, replication of previously associated loci and gene-set enrichment and tissue specificity analysis were performed using the FUMA web service[35] (Supplementary Note). The obtained results indicate metabolic interactions between the host and associated microbes, and an enrichment of genes derived from metabolic and inflammatory traits.

Our results highlight the power of combining multiple independent cohorts for genomic association analyses of microbial features, because they allow for robust and replicable results. Although a direct influence of ABO histo-blood group and secretor status on the microbiome is debated[36,37], our results support this interaction, potentially acting as a modulator in diseases for which variants in histo-blood groups and the microbiome were independently reported as risk factors[22,38–40]. The suggested causative role of *Bacteroides* spp. in patients genetically susceptible to IBD development is notable, because multiple independent, and sometimes contrasting, results were previously reported from host–microbe association and MR analyses. The multifaceted role of *Bacteroides* spp. in the human gut microbiome is probably insufficiently captured by 16S ribosomal RNA gene amplicon-based surveys, and may therefore require future in-depth strain-level analysis. Nevertheless, our results suggest an important role of the human ABO histo-blood group antigens as candidates for direct modulation of the human metaorganism in health and disease.

Online content

Any methods, additional references, Nature Research reporting summaries, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00747-1.

Received: 14 January 2020; Accepted: 3 November 2020;
Published online: 18 January 2021

References

1. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
2. Franzosa, E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
3. Cryan, J. F., O'Riordan, K. J., Sandhu, K., Peterson, V. & Dinan, T. G. The gut microbiome in neurological disorders. *Lancet Neurol.* **19**, 179–194 (2019).
4. Wirbel, J. et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
5. Blekhman, R. et al. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).
6. Goodrich, J. K. et al. Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
7. 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).
8. Turpin, W. et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
9. Bonder, M. J. et al. The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
10. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
11. Goodrich, J. K. et al. Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**, 731–743 (2016).
12. Krawczak, M. et al. PopGen: population-based recruitment of patients and controls for the analysis of complex genotype–phenotype relationships. *Community Genet.* **9**, 55–61 (2006).
13. Völzke, H. [Study of Health in Pomerania (SHIP). Concept, design and selected results]. *Bundesgesundheitsblatt—Gesundheitsforschung—Gesundheitsschutz* **55**, 790–794 (2012).
14. Völzke, H. et al. Cohort profile: the study of health in Pomerania. *Int. J. Epidemiol.* **40**, 294–307 (2011).
15. Holle, R., Happich, M., Löwel, H. & Wichmann, H. E., MONICA/KORA Study Group. KORA—a research platform for population based health research. *Gesundheitswesen Bundesverb.* **67**, S19–S25 (2005).
16. Reitmeier, S. et al. Arrhythmic gut microbiome signatures for risk profiling of type-2 diabetes. *Cell Host Microbe* **28**, 258–272.e6 (2020).
17. Johnson, J. S. et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).

18. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
19. Davenport, E. R. et al. Genome-wide association studies of the human gut microbiota. *PLoS ONE* **10**, e0140301 (2015).
20. Wegiel, B. et al. Biliverdin inhibits Toll-like receptor-4 (TLR4) expression through nitric oxide-dependent nuclear translocation of biliverdin reductase. *Proc. Natl Acad. Sci. USA* **108**, 18849–18854 (2011).
21. Schirmer, M. et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167**, 1125–1136.e8 (2016).
22. McGovern, D. P. B. et al. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. *Hum. Mol. Genet.* **19**, 3468–3476 (2010).
23. Jostins, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
24. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
25. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
26. Smith, G. D. & Ebrahim, S. *Mendelian Randomization: Genetic Variants as Instruments for Strengthening Causal Inference in Observational Studies* (National Academies Press, 2008).
27. Wade, K. H. & Hall, L. J. Improving causality in microbiome research: can human genetic epidemiology help? *Wellcome Open Res.* **4**, 199 (2020).
28. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).
29. Arumugam, M. et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
30. Wu, G. D. et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–108 (2011).
31. Costea, P. I. et al. Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3**, 8–16 (2018).
32. Vieira-Silva, S. et al. Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. *Nat. Microbiol.* **4**, 1826–1831 (2019).
33. Zhou, Y. & Zhi, F. Lower level of bacteroides in the gut microbiota is associated with inflammatory bowel disease: a meta-analysis. *BioMed. Res. Int.* **2016**, 5828959 (2016).
34. Wang, K. et al. *Parabacteroides distasonis* alleviates obesity and metabolic dysfunctions via production of succinate and secondary bile acids. *Cell Rep.* **26**, 222–235.e5 (2019).
35. Watanabe, K., Taskesen, E., van. Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
36. Davenport, E. R. et al. ABO antigen and secretor statuses are not associated with gut microbiota composition in 1,500 twins. *BMC Genomics* **17**, 941 (2016).
37. Turpin, W. et al. FUT2 genotype and secretory status are not associated with fecal microbial composition and inferred function in healthy subjects. *Gut Microbes* **9**, 357–368 (2018).
38. Rausch, P. et al. Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (secretor) genotype. *Proc. Natl Acad. Sci. USA* **108**, 19030–19035 (2011).
39. Weiss, F. U. et al. Fucosyltransferase 2 (FUT2) non-secretor status and blood group B are associated with elevated serum lipase activity in asymptomatic subjects, and an increased risk for chronic pancreatitis: a genetic association study. *Gut* **64**, 646–656 (2015).
40. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl. Environ. Microbiol.* **63**, 2802–2813 (1997).

Methods

Cohort description, genotyping and imputation. *PopGen.* The PopGen cohort is a population-based cohort from the area around Kiel, Schleswig-Holstein, Germany[12]. From this cohort, 1,108 individuals were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 covering 906,600 genetic variants. After the initial quality control (QC), which included filtering out variants with a minor allele frequency (MAF) < 1%, per-SNP call rate <95% and deviation from Hardy–Weinberg equilibrium (HWE) with $P < 10^{-5}$, the genotyping data were prepared for imputation following the miQTL cookbook instructions (https://github.com/alexa-kur/miQTL_cookbook#chapter-2-genotype-imputation). Briefly, this Plink-based processing script includes steps to prepare variants to be in consistency with the Haplotype Reference Consortium (HRC) v.1.1 reference panel regarding the order of reference and alternative alleles, variant naming and strand orientation. Finally, all data are converted to VCF files for imputation. Imputation of the autosomal chromosomes was performed using the Michigan Imputation Server with the HRC release v.1.1 from 2016 as reference panel. Eagle v.2.3 was chosen as the phasing algorithm and EUR individuals were selected as the population for QC purposes. The process was started in 'Quality Control & Imputation' mode. After downloading the final data, it was converted to binary plink files and variants with MAF < 1% were removed. Fecal samples were available for 724 of these individuals; they were collected by the participants themselves at their respective homes in standard fecal collection tubes, and mailed to the study center where they were stored at −80 °C until processing. DNA from fecal samples (~200 mg) was extracted using the QIAamp DNA stool mini-kit automated on the QIAcube (QIAGEN).

FoCus. The FoCus cohort was incepted as part of the competence network Food Chain Plus (http://www.focus.uni-kiel.de/component/content/article/88.html). This cohort consists of two parts, one of which is a population-registry-based, cross-sectional cohort including individuals from the area around Kiel, Schleswig-Holstein, Germany. The second part is an outpatient clinic-based cohort including obese individuals (body mass index (BMI) > 30) with and without accompanying disease status. For our study, only the registry-based part of the cohort was included. Cohort participants were genotyped using the Infinium OmniExpressExome array. Data processing, imputation and sampling of fecal material were performed in the same way as in the PopGen cohort. Finally, of 1,583 participants, 957 belonged to the population-based part of the cohort and supplied fecal samples. DNA from fecal samples (~200 mg) was extracted using the QIAamp DNA stool mini-kit automated on the QIAcube.

KORA FF4. KORA (Kooperative Gesundheitsforschung in der Region Augsburg) is a population-based adult cohort study in the Region of Augsburg, southern Germany, which was initiated in 1984 (https://www.helmholtz-muenchen.de/epi/research/cohorts/kora-cohort/objectives/index.html). For the second follow-up study (FF4) of the baseline study S4, 2,279 participants were recruited and the study was conducted in 2013–14, mainly focusing on diabetes, cardiovascular disease, lung disease and links to environmental factors such as the microbiome. Stool-derived DNA samples of 2,136 participants were obtained via the KORA biobank. The DNA had been extracted using a guanidine thiocyanate /*N*-lauroylsarcosine-based buffer[40] and subsequent clean-up with NucleoSpin gDNA Clean-up (Macherey–Nagel) for further analysis. Genotyping was performed using the Affymetrix Axiom array, and initial QC of raw data included MAF filtering <1%, per-SNP call rate <98% and deviation from the HWE with $P < 10^{-4}$. In total, 1,864 samples with genotyping and 16S rRNA gene survey data were included in the association analysis.

SHIP and SHIP-TREND. The Study of Health in Pomerania (SHIP) is a longitudinal population-based cohort study located in the area of West Pomerania (north-east Germany). It consists of the two independent cohorts, SHIP ($n = 4{,}308$; baseline examinations 1997–2001) and SHIP-TREND ($n = 4{,}420$; baseline examinations 2008–12) with regular follow-up examinations every 5 years[13]. Stool samples have been collected since the second follow-up investigation of the SHIP (SHIP-2, 2008–12) and baseline examination of the SHIP-TREND cohort. All fecal samples were collected by the study participants in their home environments, stored in plastic tubes containing stabilizing ethylenediaminetetraacetic acid buffer and shipped to the laboratory where DNA isolation (PSP Spin Stool DNA Kit, Stratec Biomedical AG) was performed as described previously[41]. For a total of 2,029 and 3,382 samples, on 16S rRNA gene survey and genotype on Affymetrix Genome-Wide Human SNP Array 6.0 and Illumina Infinium Global Screening Array, respectively, data were available and included in the association analysis. Initial QC of raw genotyping data included a filter for per-SNP call rate <95% and deviation from the HWE with $P < 10^{-5}$.

All genome positions are given in relation to the hg19/GRCh37 human genome reference. Written, informed consent was obtained from all study participants in all cohorts, and all protocols were approved by the local institutional ethical review committees in adherence with the Declaration of Helsinki principles.

Inference of ABO blood group and secretor status. ABO blood groups were inferred using the phased and imputed genetic data and four variants as proposed by Paré et al.[42], which rs507666, rs687289, rs8176746 and rs8176704 encode for the alleles A1, O, B and A2, respectively. All variants were, depending on the genotyping array used in the respective cohort, either genotyped by the array or showed very high imputation quality scores of between 98.7% and 99.8%. In addition, observed allele frequencies were manually compared with frequencies in public databases to assure the highest-quality blood group assignments. Secretor status was assessed by variant rs601338;G>A on chromosome 19. Individuals homozygous for the A allele were classified as 'nonsecretors'. This variant was genotyped in all cohorts, except for the PopGen cohort. In the present study, the estimated imputation accuracy was 94.6%.

Microbial data generation and processing. Library preparation and sequencing were performed using a standardized protocol at a single wet lab in Kiel, Germany. DNA amplification by PCR of the bacterial 16S rRNA gene was performed using the 27F/338R primer combination targeting the V1–V2 region of the gene, employing a dual-index strategy to achieve multiplex sequencing of up to 384 samples per sequencing run. After PCR, product DNA was normalized using the SequalPrep Normalization Kit. Sequencing of the libraries was performed on an Illumina MiSeq using v.3 chemistry and generating 2 × 300-bp reads. Demultiplexing was performed allowing no mismatches in the index sequences. Data processing was performed in the R software environment (v.3.5.1)[43], using the DADA2 (v.1.10)[44] workflow for big datasets (https://benjjneb.github.io/dada2/bigdata.html), resulting in abundance tables of ASVs. All sequencing runs underwent QC and error profiling separately. Briefly, forward and reverse reads were trimmed to a length of 230 and 180 bp, respectively, or at the first position with a quality score ⩽5. Low quality read-pairs were discarded when the estimated error in one of the reads exceeded 2 or when ambiguous bases ('N's) were present in the base sequence. Read-pairs that could not be merged due to insufficient overlap or mismatches in their nucleotide sequences were discarded. The complete workflow adjusted for the 16S rDNA V1–V2 amplicon can be found on GitHub: https://github.com/mruehlemann/german_mgwas_code/tree/master/1_preprocess. Finally, all data from the separate sequencing runs were collected in a single abundance table per dataset, followed by chimera filtering. ASVs underwent taxonomic annotation using the Bayesian classifier provided in DADA2 and the Ribosomal Database Project v.6 release[45]. ASV abundance tables and taxonomic annotation were passed on to the phyloseq package[46] for random subsampling to 10,000 sequences per sample (rarefy_even_depth()) and construction of phylum- to genus-level abundance tables (tax_glom()). Samples with fewer than 10,000 clean reads were not included in the analysis. Sequences that were not assignable to genus level were binned into the finest-possible taxonomic classification. As amplicon-based sequencing of the 16S rDNA has clade-dependent taxonomic resolution differences[17], abundance profiles of ASVs and OTUs, based on two widely used similarity cut-offs (97% similarity for a proxy of species level, 99% similarity for strain level), were included in the analysis. This enables an unbiased assessment of genetic effects at a subgenus taxonomic scale. Although similarity cut-offs as a proxy for taxonomic resolution are elements of ongoing discussion[47], clustering still allows the bundling of similar sequences, and by that evolutionarily closely related organisms, into units of probable functional similarity. For this, ASV nucleotide sequences were exported to the FASTA format, including information about their respective abundances as part of the sequence header. The sequences underwent dataset-spanning OTU picking at the 99% and 97% identity levels, using the VSEARCH software[48]. ASVs and OTUs were assigned cross-dataset-consistent IDs for more convenient data handling, 97% and 99% identity-based features being named OTU97 and OTU99 throughout the letter, respectively. ASVs included in the analysis were relabeled to 'TestASV'. OTUs at the 97% identity level were aligned against the SILVA reference alignment (v.132) using the SINA aligner; consistent gaps in the alignment were truncated[49]. The resulting alignment was used to construct a phylogenetic tree using the FastTree (v.2.1.7)[50] software with the flags –nt (input is nucleotide alignment), –gtr (generally time-reversible model) and –gamma (for branch-length rescaling and calculation of gamma20-likelihood).

Statistics for cohort comparisons. Basal phenotypes of age and BMI were compared between cohorts using pairwise Wilcoxon's rank-sum test with the R-base function pairwise.wilcox.test() and the default method 'holm' for correction of *P* values. Within-sample diversity was assessed using the total number of observed genera and Shannon's diversity index calculated at the genus level using the vegan::diversity()[51] function in R. To generate Shannon's genus level equivalents, Shannon's diversity was used as an exponent in the natural exponent function exp(). Differences between cohorts were assessed using pairwise Wilcoxon's rank-sum test implemented in the R-base function pairwise.wilcox.test() and the default method 'holm' for correction of *P* values. Pairwise cohort differences in between-sample diversity (β diversity) were assessed using genus-level Bray–Curtis dissimilarity and a permutational multivariate analysis of variance using distance matrices, as implemented in the vegan::adonis() function. For each comparison, 1,000 permutations were used to assess *P* values.

Statistical framework for GWAS. *Rationale.* The assembly of intestinal microbial communities is a highly complex process, which potentially can be driven by

environmental and lifestyle factors, host genetics[5–10] and disease[1–4]. These biotic and abiotic factors mold niches for specific microorganisms, supplying them with metabolic substrates that can be directly host derived, as with specific glycosylation patterns, or influenced by the host's metabolism, as discussed for the connection between the persistence of lactose hydrolysis and the abundance of bifidobacteria[11]. The univariate statistical frameworks applied in the present study aimed to identify genetic associations with presence–absence and abundance patterns of microbial clades. These associations could be the result of variation in host genes leading to the availability of specific energy sources or metabolic substrates (and the lack thereof, respectively). Such effects would, therefore, facilitate competitive (dis)advantage of the specific bacteria associated with them. Alternatively, an immune response, which is specific to a given microbial feature, could be influenced by genetic variations. In this case, the abundance or the presence of the microorganism in the community would be modulated. In addition, the community as a whole can be influenced by the effect of host genetics, which can act on more than a single clade and also depend on stochastic effects in the initial community assembly[52]. As such, effects would be distributed across multiple features or clades with only small individual effect sizes. Therefore, an association analysis targeting multivariate effects was additionally implemented to identify host genetics-associated shifts at the level of the microbial community.

Feature filtering. All univariate microbial features, defined by either taxonomic annotation or ASV/OTU clustering, independently underwent filtering using the same criteria for inclusion in the association analysis. Within a cohort, a feature had to be present in at least 100 individuals and had to exceed the median abundance of 50 reads, so 0.5%, in the individuals with nonzero counts. For the analysis of differential prevalence, the feature also had to be absent in at least 100 individuals. If these criteria were fulfilled in at least three of the cohorts, the feature was included in the analysis. Summary statistics for all cohorts and microbial features included in the analysis can be found in Supplementary Table 1. Using the described approach, each unique ASV sequence can thus be part of up to eight tested univariate microbial features, five of them defined by the assigned taxonomy (phylum to genus level) and three defined independent of taxonomy and solely by similarity (ASV, 99% OTU and 97% OTU). To account for this, the total number of independent univariate features tested in the association analysis were estimated (Meta-analysis). The filtering resulted in 233 univariate features for the abundance-based analysis, of which 4 were at phylum level, 8 at class level, 6 at order level, 10 at family level, 29 at genus level, and 65, 62 and 49 at 97% OTU, 99% OTU and ASV levels, respectively. For the presence–absence-based analysis, 198 features were included, of which 2 were on class, 1 on order, 2 on family, 17 on genus, and 65, 62 and 49 on 97% OTU, 99% OTU and ASV levels, respectively. In total, 431 univariate microbial features were included in the GWAS.

Prevalence-based analysis. For the analysis of genetic effects on the prevalence of bacterial features, abundance values were recoded into 0 (absence) and 1 (presence). Genetic variants were filtered to an MAF > 5% and coded into numeric features 0 (homozygous for reference allele), 1 (heterozygous) and 2 (homozygous for alternative allele). Taxon prevalence was submitted to a logistic regression employing a generalized linear model with binomial distribution and logit-link-function using the genotype as predictor, including age, sex, BMI and the ten first genetic principal components (PCs) as covariates. All statistical tests were performed two sided.

Abundance-based analysis. For calculating the effects of genetic variants on the zero-truncated abundance of bacterial features, the features were first filtered for extreme outliers, deviating more than 5× the interquartile range (IQR) from the median abundance. Using the glm.nb() function from the MASS package in R, count abundances were fit in a model using previously mentioned covariates age, sex, BMI and the first ten genetic PCs as covariates. Residual variation was extracted using the *residuals*() function and submitted to a linear model estimating the effect of the genetic variants on the residual abundance. Analysis of SNP versus feature abundance, directly using generalized linear models with negative binomial distribution, was tested as well; however, these models' results showed highly inflated λ_{GC} values, and so were discarded for the GWAS. All statistical tests were performed two sided.

Analysis of β diversity. In addition to the single-feature-based analyses, we analyzed the effects of genetic variants on the β diversity. For this, the genus-level abundance tables were used to calculate the pairwise Bray–Curtis dissimilarity between the individual microbial communities. In addition, weighted, normalized UniFrac distance was calculated based on 97% identity OTU abundances using the UniFrac() function in phyloseq. Distance matrices were submitted to a distance-based redundancy analysis using the vegan::capscale() function and the same previously mentioned covariates. The residual variance of the model was extracted using the residuals() function, resulting in a distance matrix adjusted for these possibly confounding factors. This distance matrix was used in a procedure to estimate the effect of genetic variants based on a distance-based *F*-test using moment matching[53]. The calculations were implemented to run on a graphs processing unit for further speed-up, especially in the larger cohorts (see Supplementary Note and Supplementary Fig. 2 for benchmark). As calculations for large cohorts with $n > 1{,}000$ individuals (with tables of size $n \times n$) still could not be finished in a reasonable time, we employed a stepwise calculation of results for the cohorts (estimating, from single central processing unit usage, that processing time of 7×10^6 variants for the SHIP-Trend dataset would take 61 years; even using one graphs processing unit instance, processing would take ~94 d). The stepwise calculation process was as follows: for the PopGen, FoCus and SHIP cohorts, all variants were tested for an association. If a variant showed a nominal significant association ($P < 0.05$) in at least one of the cohorts, this variant was tested in the KORA cohort. If a variant was then nominally significant in at least two of these four cohorts, it was also tested in the SHIP-TREND cohort.

Meta-analysis. Genomic inflation (λ_{GC}) was assessed for all cohorts and features, and all showed values below the proposed threshold of 1.05. Results from the separate cohorts were combined using a meta-analysis framework. Prevalence- and abundance-based results were submitted to an inverse-variance-based strategy, calculating effects based on effect size and variance of the respective cohorts. For the β diversity meta-analysis, we chose a weighting based on the sample size of the respective cohorts. Both approaches were adapted from the METAL software package for GWAS meta-analysis[54]. Criteria for the reporting of a significant association were a genome-wide significant meta-analysis $P < 5 \times 10^{-8}$, and nominal significance in at least two cohorts for the single-feature tests and at least three cohorts for the β diversity analysis. As the univariate microbial features can be correlated across the different taxonomic levels in the analysis, the matSpDlite algorithm was used to estimate the effective number of independent (effective) variables across all levels, based on the variance of eigenvalues of the univariate abundances and presence–absence patterns[55,56]. This yielded 141 and 127 effective variables for the abundance-based and the prevalence-based analysis, respectively. From this, we defined a study-wide significance threshold of $P < 5 \times 10^{-8}/268 = 1.866 \times 10^{-10}$. Heterogeneity statistics—Cochran's *Q* and variation across studies due to heterogeneity I^2—for individual variants from the presence–absence and relative abundance association analyses were calculated as described in Deeks et al.[57].

Analysis of influence of blood groups and secretor status. Hurdle models were used to investigate prevalence and abundance patterns in connection with ABO blood group and secretor status. Nine models were used for analysis. Models 1–4 analyzed the effects of the individual's counts of A alleles, B alleles, the sum of A and B alleles, and the binary status O versus non-O, respectively. Models 5–8 investigated the same factors, although in interaction with FUT2 secretor status, so only taking nonzero values when assigned as a 'secretor'. The last model investigated only the effects of the binary secretor status. All models included the covariates age, sex, BMI, and the first ten genetic PCs, in analogy to the GWAS. Inverse-variance-weighted (IVW) meta-analysis was used to combine the results into a composite result per taxon and model.

Mendelian randomization. MR analysis was performed using the TwoSampleMR package (v.0.4.25)[25] for R. Using the MR-Base database (mrbase.org), 41 binary traits from the subcategories 'Anthropometric', 'Autoimmune/Inflammatory', 'Bone', 'Cancer', 'Cardiovascular', 'Diabetes', 'Kidney', 'Pediatric disease' and 'Psychiatric/neurological' were selected for analysis of directional effect of microbial features on these outcomes. A full list of the selection criteria, used outcome traits and the used database IDs can be found in the Supplementary Note. To ensure power and suitability of the instruments used for MR, only variants with $P < 10^{-5}$ and *F*-statistics[58] >10 were included as exposure/instrument variables in the analysis. The remaining instruments were LD clumped to include only independent signals. Using the power_prune() function, the best set of instrumental variables for each trait was selected using instrument strength and sample size as selection criteria (method 2). Primary MR analysis was performed for sets with multiple instrument variables and single instrument variables using the IVW analysis and Wald's test, respectively. Additional sensitivity analyses using weighted median, weighted mode and Egger regression were performed for analyses with more than two instrument variables available. Per-microbial trait, a suggestive threshold was defined as $P < 0.05/41 = 1.220 \times 10^{-3}$. For study-wide significance, *P* values were adjusted using Benjamini–Hochberg FDR correction; for the resulting *q* value the threshold was set to 0.05. For β diversity analysis, no MR was performed, because the nonparametric test used for analysis did not include a β value for effect size needed for MR.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Cohort-level summaries of microbial feature abundances are available in Supplementary Table 1. Complete summary statistics of all tested microbial features are available via the NHGRI-EBI GWAS catalog (https://www.ebi.ac.uk/gwas), GCP ID: GCP000068; study accession nos. GCST90011301–GCST90011730. The German mGWAS browser application is available for local query of results from Dockerhub: https://hub.docker.com/r/mruehlemann/german_mgwas_browser_app. Due to the informed consent obtained from the participants, phenotypes, as well as genotyping and not all 16S rRNA gene-sequencing data, can be deposited

publicly; however, all data are available upon request from the respective biobanks (see Supplementary Note for details). PopGen and Focus: 16S rRNA-sequencing data are available at the National Center for Biotechnology Information Sequence Read Archive, accession no. PRJNA673102; http://www.uksh.de/p2n/Information+for+Researchers.html. KORA FF4: https://epi.helmholtz-muenchen.de. SHIP and SHIP-TREND: https://www.fvcm.med.uni-greifswald.de/dd_service/data_use_intro.php (German website; English-speaking assistance for the application process can be requested via: transfer@uni-greifswald.de).

Code availability

Microbiome data pre-processing, GWAS analysis and post-processing code are available via github: https://github.com/mruehlemann/german_mgwas_code.

References

41. Frost, F. et al. Impaired exocrine pancreatic function associates with changes in intestinal microbiota composition and diversity. *Gastroenterology* **156**, 1010–1015 (2019).
42. Paré, G. et al. Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. *PLoS Genet*. **4**, e1000118 (2008).
43. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).
44. Callahan, B. J. et al. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
45. Cole, J. R. et al. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–D642 (2014).
46. McMurdie, P. J. & Holmes, S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
47. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).
48. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *Peer J.* **4**, e2584 (2016).
49. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
50. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
51. Oksanen, J. et al. The vegan package. *Community Ecol. Package* **10**, 631–637 (2007).
52. Zhou, J. & Ning, D. Stochastic community assembly: does it matter in microbial ecology? *Microbiol. Mol. Biol. Rev.* **81**, e00002–e00017 (2017).
53. Rühlemann, M. C. et al. Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in SLC9A8 (NHE8) and 3 other loci. *Gut Microbes* **9**, 68–75 (2017).
54. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
55. Qin, Y. et al. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. Preprint at *medRxiv* https://doi.org/10.1101/2020.09.12.20193045 (2020).
56. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221–227 (2005).
57. Deeks, J. J., Higgins, J. P. & Altman, D. G. in *Cochrane Handbook for Systematic Reviews of Interventions* (eds. Higgins, J. P. T. & Green, S.) 243–296 (John Wiley & Sons, 2008).
58. Yarmolinsky, J. et al. Circulating selenium and prostate cancer risk: a Mendelian randomization analysis. *J. Natl Cancer Inst.* **110**, 1035–1038 (2018).

Acknowledgements

We thank T. Hauptmann, I. Urbach and I. Wulf of the IKMB Microbiome Lab for excellent technical assistance. We thank K. Wade for her valuable input on the MR analysis and M. Schulzky for support in figure design. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) Collaborative Research Center 1182 'Origin and Function of Metaorganisms' (grant no. SFB1182, Project A2 to A.F.) and the DFG Cluster of Excellence 2167 'Precision Medicine in Chronic Inflammation (PMI)' (grant no. EXC2167 to A.F.). The SHIP part of the study was supported by the PePPP-project (ESF/14-BM-A55_0045/16 to M.M.L.) and the RESPONSE-project (BMBF grant no. 03ZZ0921E to M.M.L.). The SHIP is part of the Research Network Community Medicine of the University Medicine Greifswald, which is supported by the German Federal State of Mecklenburg-West Pomerania.

Author contributions

M.C.R. performed microbiome sample preparation, data generation and curation, implemented ABO blood group inference, implemented statistical models, performed the (meta-)analysis, curated and interpreted results, and wrote the manuscript draft. B.M.H. curated and interpreted results and wrote the manuscript draft. C.B. performed microbiome sample preparation, data generation, curated and interpreted results, and advised in the writing of the draft manuscript. S.D. implemented statistical models, performed the (meta-)analysis and wrote the manuscript draft. L.M.-S. and L.B.T. curated and interpreted results. F.F. and F.D. performed data QC and curation. M.W. implemented ABO blood group inference. J.K. implemented statistical models and performed the (meta-)analysis. F.U.W. performed microbiome sample preparation, data generation and curation. A.P., U.V., S.W. H.G., M.L. and W.L. performed genotype and phenotype data generation and collection. K.H. performed microbiome sample preparation, data generation and curation. H.V. performed genotype and phenotype data generation and collection and data QC and curation. G.H. performed genotype and phenotype data collection. D.H. and M.M.L. designed the experiment. J.F.B. and A.F. designed the experiment and advised on the writing of the draft manuscript. All authors reviewed, edited and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-00747-1.

Correspondence and requests for materials should be addressed to A.F.

Peer review information *Nature Genetics* thanks Tao Zhang, Jonathan Braun and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Prof. Dr. Andre Franke

Last updated by author(s): Oct 22, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	All statistical analyses and plots were carried out in R (v3.5.1; https://www.r-project.org/); code for microbiome data processing and GWAS analysis is available via GitHub: https://github.com/mruehlemann/german_mgwas_code; Results browser was built using Shiny in R (https://shiny.rstudio.com/) and can be downloaded for local access and query at Dockerhub: https://hub.docker.com/r/mruehlemann/german_mgwas_browser_app; Genotype data filtering and preparation was performed using Plink (https://www.cog-genomics.org/plink2), imputaiton was performed using the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html) and EAGEL v2.3 for phasing (https://hpc.nih.gov/apps/Eagle.html). FlashPCA (https://github.com/gabraham/flashpca) was used for calculation of genetic principle components.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Cohort-level summaries of microbial feature abundances are available in Supplementary Table 1. Complete summary statistics of all tested microbial features are available via the NHGRI-EBI GWAS catalog (https://www.ebi.ac.uk/gwas/), accession ID: GCP000068. The German mGWAS Browser application is available for local query of results from Dockerhub: https://hub.docker.com/r/mruehlemann/german_mgwas_browser_app. Due to the informed consent obtained from the participants, phenotypes, as well as genotyping and 16S rRNA gene sequencing data cannot be deposited publicly, however all data are available upon request from

the respective biobanks:
- PopGen and Focus: http://www.uksh.de/p2n/Information+for+Researchers.html
- KORA FF4: https://epi.helmholtz-muenchen.de/
- SHIP and SHIP-TREND: https://www.fvcm.med.uni-greifswald.de/dd_service/data_use_intro.php

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Previous studies attempting association-analysis of genetics and microbiome data ranged around max ~ 2,000 individuals. Using five independent German cohorts, we increase this size by > 4-fold. GWAS analyses of complex binary traits reach sizes of > 100,000 individuals, however, these samples sizes are currently not possible when using microbial traits.
Data exclusions	Individuals for which either phenotypic data used as covariates (age, BMI, sex), genotype data or 16S amplicon sequencing data was missing were excluded. The individual univariate analyses excluded individuals dynamically in the abundance analysis either when the respective microbial feature was not present in the individual, or if its abundance deviated >5 inter-quartile ranges from the median abundance of that feature in the respective cohort.
Replication	The study uses a meta-analysis setup of five independent cohorts and reports genome-wide significant results in the meta analysis only when they replicate in at least two of these cohorts.
Randomization	16S rRNA gene amplicon sequencing was performed per study cohort in random batches of up to 384 samples. Genotyping was performed randomly, as all individuals included are population controls. As all samples are population controls and the assessed traits are based on the outcome of the amplicon sequencing and respective features, a confounding is unlikely in this large scale dataset.
Blinding	All individuals included are population controls. True identities are managed by the respective biobanks and pseudonymized data is used in the analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Individuals in f ive independent cohorts from three geographic regions in Germany are included in the study. Cohort: PopGen. Type: General population. The PopGen cohort consists of 724 individuals with genotype and 16S rRNA gene amplicon sequencing microbiome data available. Mean age is 61.5 years (SD=12.6 years), mean BMI is 27.3 kg/m² (SD=4.52 kg/m²) and 45.3% of individuals are female. Cohort: FoCus. Type: General population. The FoCus cohort consists of 957 individuals with genotype and 16S rRNA gene amplicon sequencing microbiome data available.

Mean age is 51.4 years (SD=14.6 years), mean BMI is 26.4 kg/m² (SD=5.26 kg/m²) and 57.8% of individuals are female.

Cohort: KORA FF4.
Type: General population.
The KORA FF4 cohort consists of 1915 individuals with genotype and 16S rRNA gene amplicon sequencing microbiome data available. Mean age is 60.6 years (SD=12.3 years), mean BMI is 27.9 kg/m² (SD=5.01 kg/m²) and 50.9% of individuals are female.

Cohort: SHIP.
Type: General population.
The SHIP cohort consists of 2029 individuals with genotype and 16S rRNA gene amplicon sequencing microbiome data available. Mean age is 57.5 years (SD=13.5 years), mean BMI is 28.3 kg/m² (SD=4.89 kg/m²) and 52.9% of individuals are female.

Cohort: SHIP-TREND.
Type: General population.
The SHIP-TREND cohort consists of 3382 individuals with genotype and 16S rRNA gene amplicon sequencing microbiome data available. Mean age is 51.3 years (SD=14.9 years), mean BMI is 28.0 kg/m² (SD=5.14 kg/m²) and 51.7% of individuals are female.

Recruitment

Cohort: PopGen
Type: General population
The PopGen cohort is a population registry based cohort from the area around Kiel, Schleswig-Holstein, Germany. First recruitment took place between 2005 and 2007, follow-up between 2010-2012 included sampling of the participants feces. The cohort is managed by the Institue of Epidemiology, Kiel University, Kiel, Germany and part of the P2N: PopGen 2.0 Network (https://www.uksh.de/p2n/).
Cohort: FoCus
Type: General population
The FoCus cohort (Food Chain Plus) is a population registry based control cohort from the area around Kiel, Schleswig-Holstein, Germany. Recruitment of participants took place in 2011-2013. The cohort is managed by the Institue of Epidemiology, Kiel University, Kiel, Germany and part of the P2N: PopGen 2.0 Network (https://www.uksh.de/p2n/).
Cohort: KORA FF4
Type: General population
The KORA cohort (Kooperative Gesundheitsforschung in der Region Augsburg) is cohort incepted in 1984 in the Region of Augsburg, Bavaria, Germany. The fourth follow-up (FF4) took place in 2013/2014 including sampling of the participants feces. The cohort is manged by the Helmholtz Center Munich - German Research Center for Environmental Health (https://www.helmholtz-muenchen.de/epi/research/cohorts/kora-cohort/objectives/index.html).
Cohort: SHIP
Type: General population
The SHIP cohort (Study of Health in Pomerania) is a population based cohort of individuals from the area around Greifswald, Mecklenburg-West Pomerania, Germany. First recruitment started in 1997 to 2001, at the second follow up between 2008 and 2012 fecal samples were collected. The cohort is managed by the department "Community Medicine" of the University Medicine Greifswald, Greifwald, Germany (http://www2.medizin.uni-greifswald.de/cm/fv/ship/studienbeschreibung/).
Cohort: SHIP-TREND
Type: General population
The SHIP-TREND cohort (Study of Health in Pomerania) is a second, independent population based cohort of individuals from the area around Greifswald, Mecklenburg-West Pomerania, Germany. First recruitment took place in parallel to the second follow-up of the original SHIP cohort between 2008 and 2012. The cohort is managed by the department "Community Medicine" of the University Medicine Greifswald, Greifwald, Germany (http://www2.medizin.uni-greifswald.de/cm/fv/ship/studienbeschreibung/).

Ethics oversight

Individual written informed consent was granted by all participants and ethical vote for all experiments, data generation and analyses was obtained from the local ethics boards of the respective biobanks in accordance to the declaration of Helsinki.

Note that full information on the approval of the study protocol must also be provided in the manuscript.