PNAS RESEARCH ARTICLE | MICROBIOLOGY

Microbiome-associated human genetic variants impact phenome-wide disease risk

Robert H. George Markowitz[a,b], Abigail Leavitt LaBella[b], Mingjian Shi[c], Antonis Rokas[b], John A. Capra[d,e], Jane F. Ferguson[a,f], Jonathan D. Mosley[c,f], and Seth R. Bordenstein[a,b,g,h,1]

Edited by Jeffrey Gordon, Washington University in St. Louis School of Medicine, St. Louis, MO; received January 11, 2022; accepted April 29, 2022

Human genetic variation associates with the composition of the gut microbiome, yet its influence on clinical traits remains largely unknown. We analyzed the consequences of nearly a thousand gut microbiome-associated variants (MAVs) on phenotypes reported in electronic health records from tens of thousands of individuals. We discovered and replicated associations of MAVs with neurological, metabolic, digestive, and circulatory diseases. Five significant MAVs in these categories correlate with the relative abundance of microbes down to the strain level. We also demonstrate that these relationships are independently observed and concordant with microbe by disease associations reported in case–control studies. Moreover, a selective sweep and population differentiation impacted some disease-linked MAVs. Combined, these findings establish triad relationships among the human genome, microbiome, and disease. Consequently, human genetic influences may offer opportunities for precision diagnostics of microbiome-associated diseases but also highlight the relevance of genetic background for microbiome modulation and therapeutics.

gut microbiome | host–microbe interactions | PheWAS | microbiome-associated disease

Significance

Following the sequencing of the human genome and gut microbiome, microbiome-associated variants (MAVs) were discovered that link human genetic variation with either microbiome compositions or the relative abundance of particular microbes in the gut. Reciprocally, human gut microbes have been associated with a growing number of diseases. Using an integrative framework of case–control studies in humans and two of the largest repositories of electronic health records globally, we assembled a collection of MAVs from 11 microbiome by human genome studies and tested their associations with tissue-wide gene expression patterns, clinical diseases, and evolutionary patterns. These results reveal the origins and diversity of MAVs and their linkages with disease risk.

The human gut microbiome associates with health and disease due to wide-ranging roles in immune system training and maintenance, metabolite production and conversion, and homeostatic signaling (1–4). While many factors (e.g., diet, environment, social exposures, etc.) explain degrees of gut microbiome variation (5–8), studies indicate that human genomic variation associates with microbiome variation in the gut, skin, vagina, and mouth, with the gut microbiome being the most deeply characterized to date (9–15). For example, human microbiome by genome-wide association studies (mbGWAS) on individuals without chronic disease reveal hundreds of associations between human gut microbiome-associated variants (MAVs) and microbiome traits, including community diversity and taxon relative abundance (16–23). The most consistent and recurring gene-microbe associations are between the lactose digestion *LCT/MCM6* genomic region and the gut genus *Bifidobacterium* (17, 23–26); however, variation in this genomic region is not associated with preferential expansion of any one species of this taxon (27). In contrast, associations between the blood antigen genomic region *ABO* and several gut microbes are inconsistently detected (22, 23, 26, 28–31), indicating that in addition to sample size, other biological and technical factors may belie the discovery of gene–microbe relationships. Determining whether host genetics simultaneously associates with differential microbial abundance and disease risk is a central challenge to resolve with substantive potential for personalized diagnostics and/or treatments for disease. If select taxa are adapted for certain human physiological or metabolic niches, their modulation through behavioral, dietary, or medical intervention may prevent or contribute to deleterious outcomes. However, how MAVs interrelate with disease risk is largely uncatalogued.

While methods to assess microbiome–disease relationships often rely on direct observation in case–control studies or model organism experimentation, the full spectrum of microbiome-linked disease phenotypes is likely unknown and requires an unbiased discovery effort in large diverse populations. Moreover, through the assemblage of MAVs from large and geographically diverse populations of individuals free from chronic disease, there is an opportunity to assess how human genetic variation links with microbial variation for both patterns and targets of human gene regulation and the evolutionary histories of MAVs. There is also a prospect to test if sites in the human genome that associate with specific microbes in healthy individuals link with disease in the collective medical records of diseased individuals. Finally, if MAVs associate with diseases from case–control microbiome studies, is the presence or relative abundance of the predicted microbe concordant with that measured in healthy individuals with a given allele?

MAVs occur across all 22 pairs of autosomes and in coding and noncoding regions of the genome, including as expression quantitative trait loci (eQTL), wherein variation at a

Author contributions: R.H.G.M. and S.R.B. designed research; R.H.G.M. and A.L.L. performed research; R.H.G.M. and A.L.L. analyzed data; and R.H.G.M., A.L.L., M.S., A.R., J.A.C., J.F.F., J.D.M., and S.R.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: s.bordenstein@psu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2200551119/-/DCSupplemental.

Published June 24, 2022.

nucleotide is associated with the differential expression of a target gene (32). Environmental exposures affect gene regulation through gene by environment eQTLs with the type and magnitude of response depending on the variation at an eQTL (33, 34). The modulation of the gut microbiome directly in mice and indirectly in humans suggests that similar microbiome by gene regulation mechanisms are present (35–37). In fact, recent work connecting host gene regulation across four hominid species, including humans, indicates that microbiome by gene interactions are mostly conserved at this level; among divergent gene responses, traits for inflammation and apoptosis are enriched (38). Microbiome associations are evident in many chronic diseases, such as obesity and metabolic disease, inflammatory bowel disease, diabetes, and gastrointestinal cancer, with changes preceding the onset or worsening of disease (39–42). Here, we integrate gene expression data from the Genotype-Tissue Expression Consortium (GTEx) database to investigate the genes and pathways that connect human genetic variation to the microbiome (32), and we delineate these expression patterns through simulation across 28 tissues types (Fig. 1).

Population-level approaches are in their infancy in the microbiome field and have thus far addressed a small number of MAVs (16, 19, 22, 43–45). Although an invaluable data source, electronic health repositories (EHRs) are sparse, and each differs in their patient and phenotype distribution as well as access to researchers, with preceding works leveraging an EHR-based biobank outside of the United States. We connect Vanderbilt's EHR-based biobank of ∼90,000 individuals to a comprehensive set of MAVs assembled across 11 studies to assess the clinical traits and disease phenotypes of significant MAVs identified to date (Fig. 1) (46). To determine the replicability of MAV associations, we compare disease-linked variants with phenome-wide association study (PheWAS) data in the UK Biobank, the largest collection of clinical health records globally. By extension with microbiome by genome-wide association data and gene expression patterns, this framework may indirectly connect gut microbes with a plurality of human phenotypes that covary at shared genomic loci. These unique resources coupled with PheWAS open an unbiased lens to broadly characterize the triad between human genomes, gut microbiomes, and human diseases.

Results and Discussion

Assembling a Catalog of Gut Microbiome-Associated Human Genetic Variants. To assemble a set of significant and unbiased gut MAVs, we used 11 large mbGWAS to investigate single-nucleotide polymorphisms (SNPs) that associate with gut microbial taxon relative abundance and beta diversity (Table 1). Binary microbiome traits (presence/absence) were excluded in this analysis. Genotypes and microbiome traits were measured in individuals

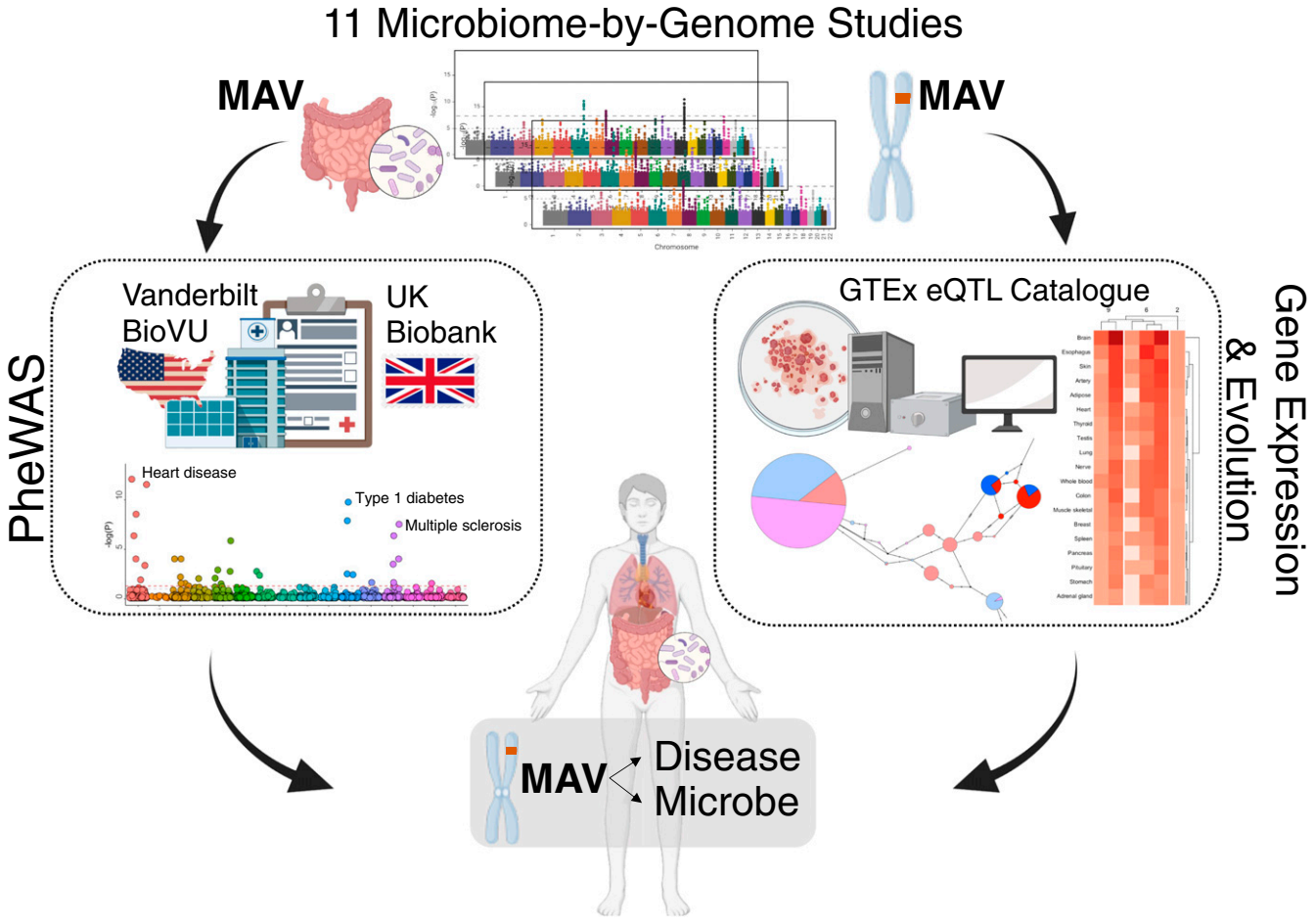

Fig. 1. An integrative framework to identify associations between human genetics, gene expression, disease phenotypes, and evolutionary patterns. Collating significant associations from 11 of the largest microbiome by genome-wide studies to date, we created an expansive set of nearly a thousand human genetic MAVs. This collection of genetic variants associated with microbiome traits is then annotated to identify the composition and location of MAVs and their relationship with human gene expression. In a complementary phenome-wide scan of hundreds of thousands of medical records from two DNA-linked health record repositories, we connected these MAVs to a catalog of disease phenotypes. Taken together, MAV-linked diseases coupled with MAV-linked microbiome traits create triads, in which the human genome, microbiome, and disease risk are linked. These previously undescribed associations establish hypotheses that have the potential to unravel microbiome-associated disease.

Table 1. Total collection of gut MAVs across studies and population

Ref.	Year	Population	Samples	MAVs	Method
Wang et al. (21)	2016	Germany	1,812	95	16S
Turpin et al. (105)	2016	Canada	1,098	51	16S
Bonder et al. (17)	2016	Holland	984	83	WGS
Scepanovic et al. (18)	2019	Western Europe	858	285	16S
Hughes et al. (19)	2020	Germany	3,900	13	16S
Xu et al. (45)	2020	China	1,475	17	16S
Rühlemann et al. (22)	2021	Germany	8,956	38	16S
Liu et al. (20)	2021	China	1,295	36	WGS
Kurilshikov et al. (16)	2021	Multiple	18,340	30	16S
Lopera-Maya et al. (26)	2022	Holland	7,738	18	WGS
Qin et al. (23)	2022	Finnish	5,959	554	WGS

without chronic illness from multiple continents and biogeographical populations, and we aggregated variants below a genome-wide significance of $P < 5\times10^{-8}$ in these studies through a review of the published genome-wide association study (GWAS) summary statistics. This produced a dataset of 1,220 significant MAVs and 925 unique MAVs following the removal of duplicates that were significant within and between studies (Dataset S1). All measured gut microbiome traits and differential relative abundances derive from subject fecal samples using either 16S ribosomal RNA (16S rRNA) or shotgun metagenomic sequencing (whole-genome shotgun sequencing [WGS]). MAVs were annotated to the exact level of taxonomic resolution from their study of discovery down to the strain level where available.

For each study, the geographic population is noted. Multiple populations were sampled in one large study spanning 11 nations, including the United States, Canada, the United Kingdom, Israel, South Korea, Germany, Denmark, the Netherlands, Belgium, Sweden, and Finland. The sequencing method for each study is listed as 16S rRNA amplicon sequencing or WGS.

MAVs Are Enriched for Gene Expression in 15 Tissues and Functional Pathways Relevant to Gut Microbiome Biology and Immune System Function. Of the 925 MAVs, 908 had annotations based on the Ensembl genome database using Variant Effect Predictor and SNPNexus (Dataset S2) (47–49). Annotated MAVs span chromosomes 1 to 22, with a global minor allele frequency (MAF) range of 0.02 to 49.8% (mean = 15.5%) (*SI Appendix*, Fig. S1). Just 4 of 908 MAVs are protein coding, representing two synonymous and two nonsynonymous variants. Among the remaining 904 MAVs, 437 are intergenic with nearest genes noted, 415 are intronic, and 45 are variants in the 3′ untranslated region (UTR; 18) or the 5′ UTR (27) (Fig. 2*A*). Twenty-five percent of intergenic variants were identified between two long intronic noncoding RNA or two processed pseudogene regions. In total, two MAVs and nine MAV-containing genes replicated between two studies (*AC005833.1*, *BANK1*, *CDH13*, *FHIT*, *FNDC3B*, *MAP4K4*, *R3HDM1*, *RAB3-GAP1*, *SPRY4-AS1*), while one MAV and two MAV-containing genes (*MCM6* and *PTPRD*) replicated between three studies (Dataset S2). Notably, recurring associations between the microbiome and the lactose digestion *LCT/MCM6* genomic region have been well described (17, 23–26), while the significance of microbiome-linked variants in or near *PTPRD* has not previously been investigated.

Given that MAVs are nearly all intergenic and intronic noncoding variants, we annotated their associations with differences in gene expression and eQTL across 48 tissues in the GTEx database (32); 373 MAVs were annotated as eQTLs associated with differential expression of 688 genes in 28 tissues (FDR < 0.05) (Dataset S3 and *SI Appendix*, Fig. S2). To determine if the tissue-specific frequency of MAV eQTLs differed from a simulated set of matched SNPs, 200 SNP sets were simulated based on MAV allele frequency, the number of SNPs in linkage disequilibrium, the distance to the nearest gene, and gene density (50). We were able to simulate SNP panels for 790 of 908 annotated MAVs, and z scores and P values were computed from each tissue-specific distribution (Dataset S4). MAV eQTLs were significantly enriched in 15 of 28 tissues groups when compared with simulated SNP sets in each tissue (FDR < 0.05) (Fig. 2*B*). Collectively among these significantly enriched tissues, *MCM6* (of the *LCT/MCM6* genomic region) is reproduced as the most targeted gene by MAV eQTLs because of cumulative representation across 10 tissues. In contrast, *LINGO2* (51) and *R11-123G.1* (52) (an interim gene identifier) are enriched in a single tissue only (Fig. 2*C*).

Since MAV eQTLs are enriched in sites related to gut microbiome functions, including the gastrointestinal and cardiovascular systems, we next explored the biological pathways of genes correlated with MAV eQTLs from the identified enriched tissues. Using the human Reactome database (53), we performed overrepresentation analysis by first narrowing down the collection of MAVs to those from the 15 tissues that were significantly enriched. Next, we analyzed MAV eQTL target genes individually in the following tissues (number of target genes): skin [223], esophagus [202], thyroid [189], nerve [179], artery [179], adipose [178], whole blood [148], testis [143], muscle skeletal [132], lung, colon, heart, pancreas, spleen, and pituitary. We identified significant overrepresentation shared among all 15 enriched tissues, including tissues of the colon, heart, and lung (FDR < 0.05) (Fig. 2*D*, Dataset S5, and *SI Appendix*, Fig. S3). Many of these common pathways, including interferon, T cell receptor (TCR), and Programmed Cell Death Ligand 1 (PD-1) signaling pathways, collectively reflect T cell interactions as well as other immune system functions. This indicates that both generalized and specific pathways are ubiquitously enriched by target genes from MAV eQTL. The more generalized "adaptive immune system" pathway term was enriched in 10 of 15 tissues (Fig. 2*D*). Interestingly, no single tissue had MAV eQTL target genes associated with a unique overrepresented pathway. Collectively, pathway representation corresponds with expected host–microbiome immunological interactions as well as inflammatory and autoimmune disorders that emerged during the PheWAS analyses described below.

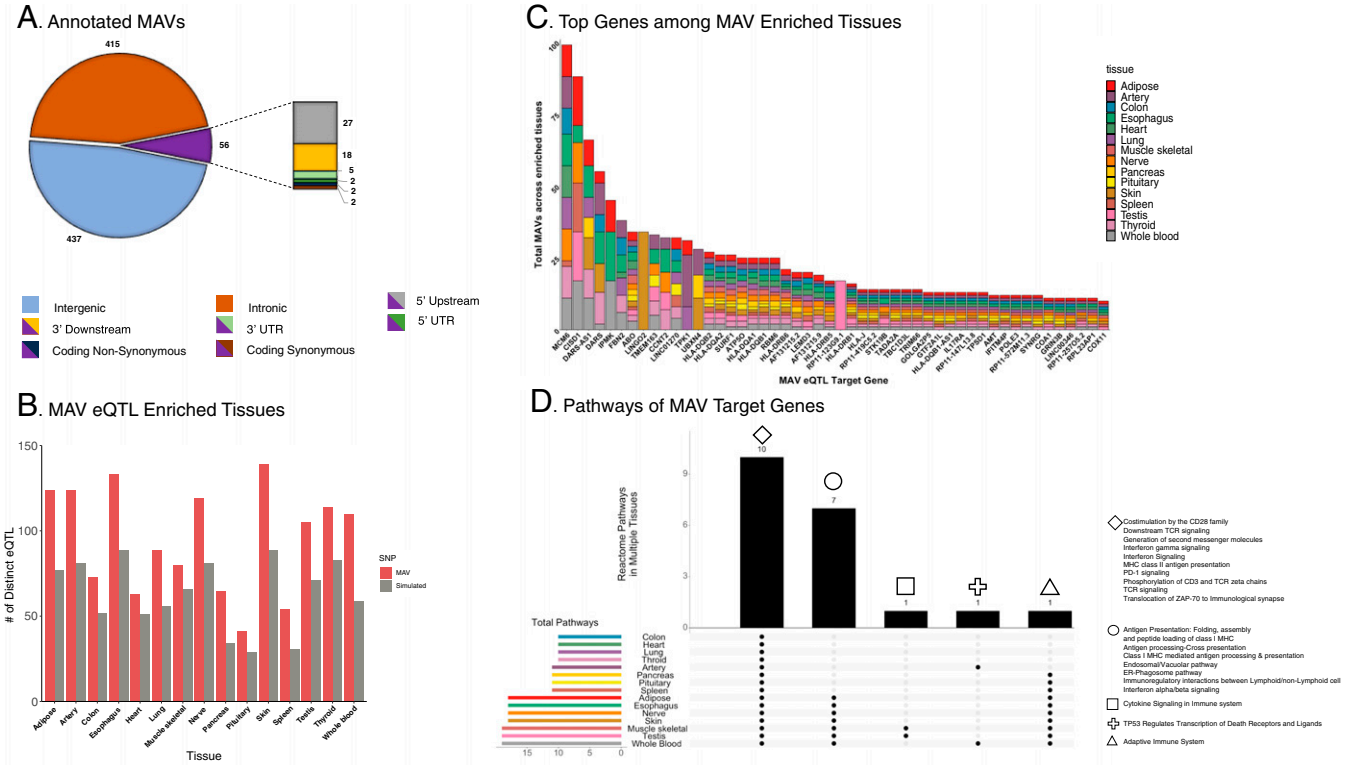

Fig. 2. Gut MAVs are predominantly intronic/intergenic and enriched in tissues and functional pathways involved in the nervous system, circulatory system, gastrointestinal tract, and immune system. (*A*) The 908 annotated variants were genetically classified by Ensembl annotation and summarized according to frequency. Boxes containing both purple and another color indicate they are elements of the smaller subset of nonintergenic or nonintronic variants. (*B*) Among 908 MAVs, simulated-matched SNP panels for 790 MAVs could be matched for allele frequency, number of SNPs in linkage disequilibrium (LD), distance to the nearest gene, and gene density. MAV eQTLs are significantly enriched in 15 of 28 tissue groups (False Discovery Rate (FDR) < 0.05) (Dataset S4). (*C*) Within the 15 enriched tissues identified (indicated in the color key), gene targets of distinct MAV eQTLs within each tissue were summed to represent the top 50 targets genes, with *MCM6* (of the *LCT/MCM6* genomic region) as the most frequently occurring. (*D*) Reactome overrepresentation analysis of MAV eQTL gene targets from enriched tissues reveals a strong association with adaptive and innate immune system functions common across all 15 MAV-enriched tissues, including TCR, interferon, PD-1 signaling, major histocompatibility complex class (MCH) II antigen presentation, and endoplasmic reticulum (ER) -phagosome interactions, among others (FDR < 0.05) (Dataset S5). Colors of tissues correspond to the color key in *C*. Numbers above each set of shared pathways correspond to numbers of shared pathways. Symbols above vertical bars correspond to pathways in the pathway legend. Individual plots of pathway overrepresentation in each of the 15 tissues are presented in *SI Appendix* (*SI Appendix*, Fig. S3).

PheWAS of Gut MAVs Produce Cross-Validating Diseases in Two Large Biobanks. To identify which, if any, MAVs are associated with disease, we performed a PheWAS using the 908 annotated MAVs in populations stratified by European ancestry (EA) and African ancestry (AA), the two largest ancestral sample populations in Vanderbilt's biobank-linked EHR repository. In its simplest form, PheWAS is a regression analysis that can determine if a genetic variant is associated with a disease based on the medical records of large populations (with and without the phenotype) that have all been genotyped at a given locus. Briefly, the PheWAS analysis was performed as a series of individual logistic regressions (e.g., MAV by phenotype), which included variables for age, sex, and the first five principal components of genetic ancestry. Analytical methods for EHR biobank data are chosen based on the characteristics of the sampled population, including size, sequencing methodologies, scope of collection criteria, and medical phenotyping definitions. For example, the PheWAS method used for BioVU data removes related individuals prior to the analysis, while the method used for UK Biobank adjusts for relatedness in each regression model. In the BioVU discovery cohort, we applied a conservative cutoff of at least 200 cases of a phenotype to be considered and required that a medical code in an individual's EHR appear twice or more to avoid spurious conditions. At the time of analysis, the Vanderbilt BioVU dataset for the 908 MAVs contained 201 to 25,467 cases and 24,935 to 56,748 controls spanning 815 EHR-derived phenotypes.

This discovery PheWAS identified 31 clinical traits that significantly associated with 10 MAVs in the EA cohort (FDR < 0.05) (Fig. 3*A*). There were no significant PheWAS associations in the AA cohort at the study-wide significance threshold (discussed further below). No disease-linked MAVs were associated with multiple taxa in their initial GWAS discovery cohorts, which simplifies the MAV–microbe interpretation in this study. In the EA population, genetic variants were distributed on chromosomes 2, 3, 6, 9, 15, and 18 (Fig. 3*B*), with the largest number of associations stemming from three MAVs on chromosome 6 in the human leukocyte antigen (HLA) region. Clinical traits identified in the EA cohort include the following categories (number of MAVs): circulatory system (seven), neurological (six), dermatologic (five), endocrine/metabolic (five), musculoskeletal (four), hematopoietic diseases (two), digestive (one), neoplasms (one), and sense organs (one). Most notably, 6 of 10 MAVs having associations with neurological, hematological, dermatological, and metabolic phenotypes also were corresponding eQTLs in brain, vascular, skin, and gastrointestinal (GI) tract tissues, respectively (*SI Appendix*, Fig. S4).

We tested the replicability of these 10 MAV associations in the UK Biobank PheWAS catalog among 1,419 EHR-derived phenotypes from more than 400,000 individuals of EA. The UK

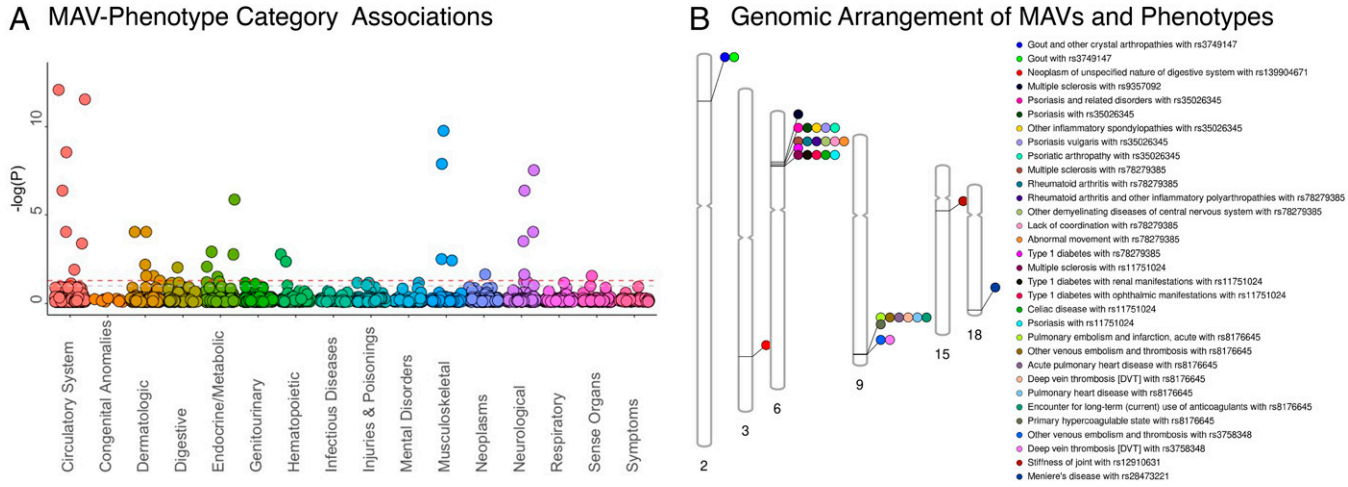

Fig. 3. Gut MAVs in the discovery cohort associate with a range of diseases, including neurological, metabolic, and hematological disorders. (*A*) PheWAS identified 31 clinical trait associations that are colored by disease category among 10 MAVs and nine disease categories in BioVU. The significance threshold is in red as FDR < 0.05 corrected over all MAVs and phenotypes. (*B*) A phenogram plot shows the 31 phenotype associations among 10 MAVs from the PheWAS analysis (Datasets S6 and S7). The list of colored dots corresponds to all significant clinical traits associated with the MAVs.

Biobank dataset contains 51 to 77,714 cases and 167,467 to 407,136 controls tested and is roughly five times larger than BioVU. The UK Biobank PheWEB catalog was precomputed using the Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) method (54) to account for the size and imbalance of cases to controls. As with the BioVU data, logistic regression results were computed for each MAV and phenotype and then corrected for multiple testing. In this targeted replication analysis, we restricted our analysis to only the 10 previously identified MAVs from the BioVU discovery analysis and the full set of phenotypes for each MAV present in the UK Biobank PheWAS catalog. We replicated 13 phenotypes with five MAVs that were further grouped into six diseases/disorders and eight MAV–disease associations (e.g., multiple sclerosis [MS]) (Table 2 and Dataset S8). These replicated associations agreed in the size and direction of disease risk of those identified in the discovery cohort. Notably, all replicated MAVs are eQTLs.

Following identification of 31 PheWAS phenotype associations in the EA cohort in Vanderbilt BioVU, 13 of 31 significant phenotype associations from five MAVs were reidentified in the UK Biobank PheWEB database (Dataset S8). A deep vein thrombosis (DVT) PheCode (452.2) is not represented in the UK Biobank; instead, an association with (451.2) phlebitis and thrombophlebitis of lower extremities as a closest match is used. Odds ratios (ORs) shown are for the disease risk allele; OR > 1 indicates an increased likelihood of the phenotype with the indicated allele (FDR < 0.05).

Triad Relationships among Genome, Microbiome, and Disease Span Autoimmune, Cardiovascular, and Metabolic Diseases. An open question is whether the elements of a MAV–microbe–disease triad align, wherein a microbial relative abundance controlled by a MAV also links with the disease that the MAV associates with by PheWAS. If the triads occur, human genetic influences on both microbial relative abundance and disease may offer opportunities for precision diagnostics as well as potential challenges for probiotic therapies that target microbial relative abundance changes. To this point, we hypothesized that the links between MAV–microbe and MAV–disease may connect and predict the differential relative abundance of a microbe in the replicated phenotypes we detected in BioVU and the UK Biobank. We informally explored this through a post hoc analysis in the absence of a human phenotype–microbe simulation approach. For each of the eight replicated phenotypes, we sought peer-reviewed support in human case–control studies for a specific disease and significant change in the relative abundance of the previously associated microbe. Interestingly, we identify a supported triad in five of eight associations in which the human genotype links with both disease risk and a microbial relative abundance direction independently connected to disease in case–control studies, thus completing the triadic cycle of genome influences phenome, and microbiome and microbiome influences or is influenced by disease (Fig. 4 and Dataset S9). We cautiously present these observations as hypotheses for future research as the data are derived from multiple independent studies, and interestingly, half of these triads involve the core gut family Lachnospiraceae that exists in humans from infancy to adulthood (56). Lachnospiraceae are enriched in pathways degrading diet-derived polysaccharides and frequently associated with inflammatory conditions, depressive syndromes, and MS that link with gut microbial relative abundance changes. The implications and specific associations of Lachnospiraceae with phenotypes are explored below.

Two MAVs had significant independent associations with MS and sequelae including lack of balance and coordination, and these MAVs are not in linkage disequilibrium. First, MAV rs9357092 (G) is identified with increased risk of MS and reduced relative abundance of Coriobacteriaceae (family), a commensal microbe of the oral, gut, and genital microbiome (57). These bacteria are depleted in the guts of untreated MS patients (58), supporting a triadic connection between microbiome by human genotypeand microbiome–MS association studies. This MAV is located within a zinc ribbon domain containing the pseudogene *ZNRD1ASP* that is proximal to the HLA complex. It suggests that it may be linked to an effect allele related to immune function. The associated taxon, Coriobacteriaceae, is also a recurrently identified ethnicity-associated taxa differing between Asian, Hispanic, and European populations (59). The other risk allele, rs11751024 (C), for MS correlates with the decreased relative abundance of the bacterial family Lachnospiraceae (coabundant gene group 882 [CAG-882]) (23). A decreased relative abundance of Lachnospiraceae occurs in relapsing–remitting MS (60) in adults and is associated with MS relapse among pediatric cases (61). This intergenic variant falls within the *HLA* genomic complex between

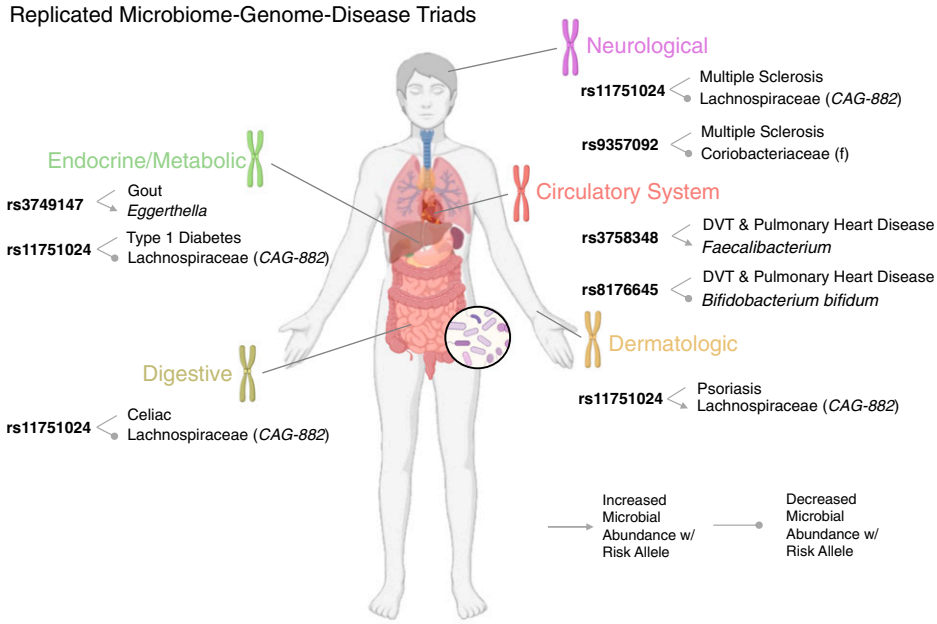

Fig. 4. Microbiome–genome–disease triads replicate across two biobanks. Five MAVs across eight diseases replicated in BioVU and the UK Biobank. Colored phenotype categories correspond to those shown in Fig. 3*A* and Table 2. Triad relationships between MAV, microbe, and disease are grouped by disease category, with a line connecting category and the effected site. MAVs and associated diseases are connected by a line. MAVs and the relative abundance of linked microbes are illustrated with an arrowed line denoting the increased relative abundance of a microbe or a line with a rounded end denoting the decreased abundance of a microbe. All triads are oriented toward the disease risk allele of a MAV such that the relative abundance between a microbe and MAV is always related to the increased risk for the associated disease. Microbial taxonomy is at the genus level unless indicated by taxonomic designation, such as family (f). CAG taxonomy is provided in parentheses for taxa annotated using Genome Taxonomy Database–aligned taxonomy (55).

HLA-DQA1/HLA-DRB1 and *HLA-DRB5/HLA-DRB9*. Taken together with the prior Reactome analysis of immune pathways enriched by MAVs in the *HLA* complex, these results support a triad model that immunogenetic variation and microbial balance may together predispose or cause neuroimmunological dysregulation (62–64).

MAV rs11751024 has four additional clinical trait associations spanning psoriasis, cardiovascular disease (CD), and type 1 diabetes (T1D), and all four associations replicate in the UK Biobank population (Table 2). Effect sizes for these phenotypes were all in the same direction between BioVU and UK Biobank populations. At this MAV, an increased disease risk for CD and T1D links with the decreased relative abundance of Lachnospiraceae (CAG-882) (23). Accordingly, Lachnospiraceae is depleted in active CD (65) and consistently abundant in genetically predisposed children leading up to onset (66). Also consistent with this MAV–T1D connection, we cross-validate depletion of Lachnospiraceae in infants and genetically predisposed children with the disease (67, 68). In contrast to other associations at the C allele of rs11751024, the A allele associates with psoriasis risk and the relative decreased relative abundance of Lachnospiraceae. Interestingly, intergenic MAV rs35026345 (T) in proximity to the *HLA* complex also associates with risk of psoriasis (and psoriatic arthropathy and psoriasis vulgaris) in conjunction with the reduced relative abundance of Lachnospiraceae (CAG-81) (23). Lachnospiraceae is a taxon of interest in psoriasis patients with both concordant (69, 70) and discordant (71) risk associations matching the bidirectional relative abundance traits identified here among CAGs. Given their inter- and intraspecies diversity and notable prevalence in psoriasis patients, these results indicate that the observed variation of Lachnospiraceae is associated with two distinct triads (72).

Recurrent observations of a relationship between MAV rs11751024 and Lachnospiraceae are of particular interest. This MAV is a significant eQTL among all 25 tissues with 27 eQTL gene targets that recapitulate the broad set of immunological pathways and signaling outlined in Fig. 2*C*, including Major Histocompatibility Complex (MHC) class II antigen presentation; generation of second messenger molecules; and interferon gamma, TCR, and PD-1 signaling (Dataset S5). It is tempting to hypothesize how this eQTL differentially alters the relative abundance of Lachnospiraceae. Lachnospiraceae's role as a prominent short-chain fatty acid (SCFA) producer in the gut suggests that in addition to human genetic variation, the timing and concentration of SCFAs produced or delivered could play roles in preventing and driving inflammation (73). Similar to work demonstrating that inhibitory innate immune sensor *NLRP12* preserved Lachnospiraceae relative abundance in high-fat diet–induced obesity (74), variation in innate regulation of the above-mentioned signaling pathways could also explain the depletion of this taxa. Alternatively, the abundance of this taxa and its SCFAs could be compensatory as well, since SCFA may reduce epithelia inflammation and limit autoimmune responses (75).

In contrast to the aforementioned autoimmune phenotypes, we identified gut MAV–phenotype associations with hematological and cardiovascular traits for pulmonary embolism and infarction, venous embolism and thrombosis, DVT, and pulmonary heart disease. These interrelated phenotypes associate with intronic MAVs rs8176645 (A) and rs3758348 (C) of *ABO* and *SURF4*, respectively. The *ABO/SURF4* region is one of a few recurring genomic regions linked to gut microbiome composition and associates with risk for cardiovascular disease (76). rs8176645(A) (*ABO* risk allele) links with reduced relative abundance of *Bifidobacterium bifidum* (26) (Bifidobacteriaceae [family]), whereas rs3758348 (C) (*SURF4* risk allele) links with the increased relative abundance of *Faecalibacterium* (Oscillospiraceae [family]) (22). Interestingly, *Bifidobacterium* is the most recurrently identified genus across all mbGWAS (10, 17, 23–26), and it has numerous beneficial health associations,

Table 2. Microbiome-linked diseases discovered in the BioVU replicate in the UK Biobank

Disease/disorder and aligned phenotype	MAV	Vanderbilt BioVU		UK Biobank	
		OR	FDR	OR	FDR
Psoriasis					
Psoriasis	rs11751024 (A)	1.67	2.73E-02	1.32	1.15E-18
Celiac disease					
Celiac disease	rs11751024 (C)	2.89	9.02E-03	1.68	2.08E-49
Type 1 diabetes					
Type 1 diabetes with renal manifestations	rs11751024 (C)	3.78	1.36E-06	1.88	4.30E-03
Type 1 diabetes with ophthalmic manifestations	rs11751024 (C)	3.16	1.69E-03	1.80	3.71E-18
MS					
MS	rs11751024 (C)	1.92	2.94E-08	1.40	2.60E-17
MS	rs9357092 (G)	1.79	3.12E-04	1.36	6.36E-12
Gout					
Gout and other crystal arthropathies	rs3749147 (A)	1.52	1.21E-03	1.11	1.89E-04
Gout	rs3749147 (A)	1.45	8.16E-03	1.14	1.29E-05
DVT and pulmonary heart disease					
Acute pulmonary heart disease	rs8176645 (A)	1.95	2.82E-09	1.30	6.83E-30
Pulmonary heart disease	rs8176645 (A)	1.36	4.19E-04	1.30	6.83E-30
DVT	rs8176645 (A)	1.65	4.21E-07	1.46	2.08E-49
Other venous embolism and thrombosis	rs8176645 (A)	1.60	2.82E-12	1.27	2.17E-04
DVT	rs3758348 (C)	1.64	1.21E-02	1.39	1.53E-20

including a reduction in cardiovascular disease (77–79). The *ABO* risk allele phenotypes replicate in the UK Biobank (Table 2). The *SURF4* risk allele has enhancer activity, increasing gene expression by ∼50%, and significantly associates with blood protein levels of platelet endothelial cell adhesion molecule-1, which is independently associated with thrombosis (80–82). This MAV also links with an increased thrombosis risk for DVT and pulmonary heart disease. As such, it is interesting to speculate that a MAV-induced increase in the relative abundance of *Faecalibacterium*, whose butyrate metabolites have antithrombotic properties (83), is compensatory to MAV–cardiovascular disease links. Indeed, *Faecalibacterium* is depleted in older patients with coronary artery disease and heart failure (84). Human fecal microbiota transplant studies demonstrate that healthy microbiome supplementation can suppress pathological coagulation in patients with metabolic syndrome (85, 86).

A single association for metabolic disease was identified between gout and MAV rs3749147 (A), which is linked with the increased relative abundance of *Eggerthella* (Eggerthellaceae [family]) (20). Gout is a common metabolic disease in which purine metabolism and urate transport dysfunction lead to the formation of urate crystals, causing joint and renal damage (87, 88). The gout-linked MAV is an eQTL that targets, among several genes, *GCKR*, a glucokinase inhibitor with significant gout associations (89, 90). The associated taxon for this MAV, *Eggerthella* (20), is differentially abundant in urine microbiome samples of gout patients and has not been associated with gut microbiomes in gout patients (91).

MAVs Lack Associations in Individuals of AA. Using the same PheWAS parameters as above, we did not identify significant MAV associations in the AA cohort (*SI Appendix*, Fig. S5). Additionally, we performed a secondary PheWAS on a subset of MAVs in the AA cohort for comparison using a 100-case cutoff, MAF greater than 25% (which included 18 MAVs from the significant EA results), and a control–case ratio greater than 4:1 to account for limitations in the size of the AA cohort and the quality of International Classification of Diseases, Ninth and Tenth Revision (ICD-9/10) codes. Prior PheWAS power analysis simulations using these parameters estimated 75% power to detect allele disease penetrance above 20% in binary traits (92). This PheWAS did not detect any associations between MAV and disease phenotypes. We reason that the lack of identified associations in the AA cohort could be attributed to two factors. Mainly, AA sample sizes at large and in comparable phenotypes lack the power to observe an association in many EA-positive phenotypes in the AA cohort (Dataset S10). Additionally, differences among population genetic architectures may contribute to diluted EA effect sizes in non-EA populations (93, 94). We explored the extent to which differentiation between populations could contribute to the variation of MAVs below.

Disease-Linked MAVs Have Different Evolutionary Histories. To examine the evolutionary processes acting on the regions identified in the PheWAS, we ran the 10 significant MAVs from our discovery cohort through an evolutionary analysis pipeline designed to detect enrichment in evolutionary signals using a variety of sequence-based metrics of population differentiation and selection (95). Of the 10 MAVs identified in the BioVU discovery cohort with disease phenotype associations, 4 MAVs with phenotype associations, including increased risks for cardiovascular, sensory, and musculoskeletal traits, produced sufficient background control datasets to enable testing and interpretation following simulation for null hypothesis testing (Fig. 5*A* and Dataset S11). The interpretation of six MAVs that did not produce suitable simulation data for comparison testing is described in *Methods*.

MAV rs28473221 associates with Meniere's disease and the increased relative abundance of Erysipelotrichaceae (family) (105). It has a high F_{ST} between AA and other ancestry groups (F_{ST} between East Asian ancestry and AA groups: z score = 2.679, P = 0.028; F_{ST} between EA and AA groups: z score = 1.65, P = 0.08). This is further reflected in the differing MAFs between ancestries (AA: 48%, EA: 11%, East Asian ancestry: 1%). The variant is also low between AA and East Asian ancestry in XP-EHH (cross-population extended haplotype homozygosity), which measures population-specific positive selection

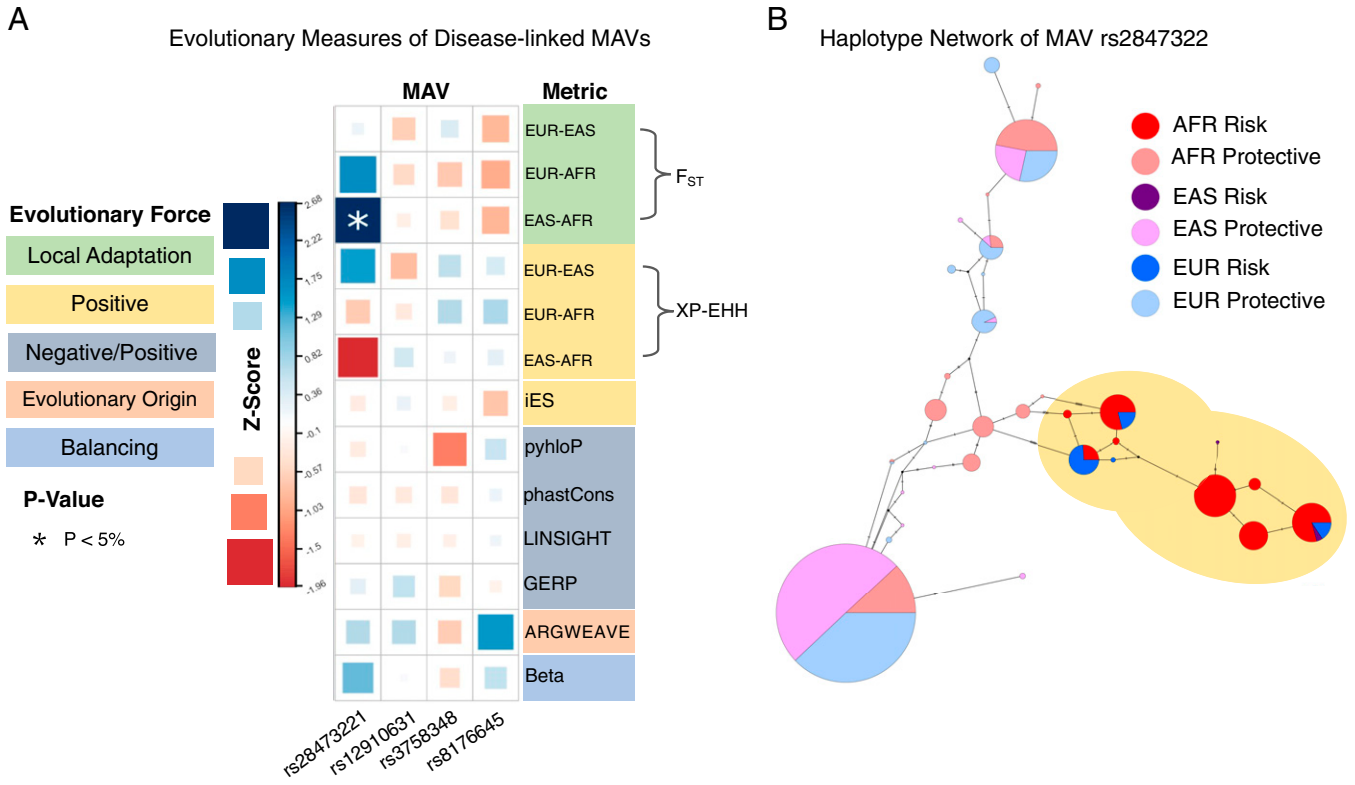

Fig. 5. Disease-linked MAVs have a variety of evolutionary histories. (*A*) Quantification of evolutionary pressures on four MAVs by an evolutionary analysis pipeline (*y* axis) (95). The size and color of the squares represent the magnitude and direction of *z* scores of a range of evolutionary test metrics on each MAV compared with a background control set matched for LD, MAF, and gene proximity. Significant enrichment ($P < 0.05$) is indicated with a white star. The range of evolutionary signal enrichments across these four MAVs can be seen in the variety of directions, with red indicating negative (decreased relative to the control set for this metric) and blue indicating positive (increased relative to the control set for this metric) *z*-score values. The sequence-based evolutionary metrics used are as follows. 1) F_{ST} measures the variance of allele frequency among populations to infer population differentiation (96). 2) XP-EHH measures haplotype homozygosity to detect population-specific positive selection (97). 3) integrated EHH (iES) detects haplotype homozygosity to infer positive selection over a chromosomal region (98). 4) Phylogenetic p-values (PhyloP) calculates substitution rate to infer positive and negative selection based on expected neutral drift (99). 5) The phastCons 100 score is the probability that each nucleotide belongs to a conserved element as calculated from the multiz alignment of 100 vertebrate species (100). 6) The LINSIGHT score measures the probability of negative selection on noncoding sites (101). 7) The genomic evolutionary rate profiling (GERP) score measures the reduction in the quantity of substitutions in the multispecies sequence alignment as compared with the neutral expectation (102). 8) The ARGWEAVE score measures the TMRCA as computed by the ARGweaver pipeline (103). 9) Beta measures balanced polymorphisms to infer balancing selection (104). (*B*) The haplotype network of the nonrecombining block containing the MAV rs28473221 obtained from the 1000 Genomes dataset. The haplotypes are colored by ancestry and presence of the protective (*A*) allele. The majority of individuals in the East Asian ancestry (EAS) and EA (EUR) group have one of two main haplotypes (large circles) containing the protective allele, while the protective allele is less common in AA (AFR) groups, where the risk allele is prominent (golden highlight).

(*z* score of −1.96, $P = 0.07$) (97). The low MAF (G) in East Asian ancestry together with the low XP-EHH score supports a selective sweep acting on the major allele (A) (Fig. 5*B*). Although current clinical observations do not reveal equal incidence or prevalence of Meniere's disease among these ancestries, it may be underdiagnosed among Africans (106). Further attention is warranted to determine if Meniere's disease represents an unrecognized health disparity or if genetic susceptibility is a driver of disease prevalence.

Two MAVs with replicated phenotypes in size and direction in the UK Biobank also produced viable evolutionary scores for simulation with the evolutionary analysis pipeline. First, MAV rs8176645, linking *B. bifidum*, hypercoagulable traits, and pulmonary heart disease, has a high ARGWEAVE (103) score (*z* score = 1.551), low F_{ST} values, and similar MAFs across ancestries (43% African, 38% East Asian, and 40% European). ARGWEAVE estimates the time to the most recent common ancestor (TMRCA) to be 115,179 y for this variation, which is much higher than in the control set (62,855 y). That suggests this MAV arose in ancient humans. The low F_{ST} and similar MAFs between ancestries collectively indicate that the allele is under balancing selection in concordance with an old TMRCA. Notably, both alleles (T and A) of this MAV are observed in Denisovans supporting the observed TMRCA estimate (107, 108). Second, MAV rs3758348, linked with *Faecalibacterium* and also associated with hypercoagulable traits and pulmonary heart disease, stands out for its low PhyloP100 score (*z* score = −1.36), which measures evolutionary conservation (99). The ancestral G allele is the major allele at this position and is conserved throughout old world monkeys and apes. The negative value at this location is most consistent with human-specific mutational acceleration. In contrast to the previous MAV, which shared hypercoagulability and pulmonary heart disease traits and is under balancing selection, this MAV demonstrates faster evolution than expected under neutral drift. Finally, MAV rs12910631, associated with microbiome beta diversity and joint stiffness, does not exhibit any *z* scores for evolutionary metrics greater or less than one, which suggests that the region is likely evolving neutrally.

Conclusion

Understanding the precise relationships between the human genome, microbiome, and clinical traits remains limited yet central for research and efforts aimed at developing new therapies and personalized diagnostics of diseases with unclear etiologies. Here, using an unbiased framework integrating genetic,

transcription, microbiome, and evolutionary approaches, we leveraged 11 mbGWAS and case–control studies with a phenome-wide scan of clinical health records from two independent health record repositories to uncover the genetic footprints of disease that interrelate with gut microbiome variation. We report six key results. 1) 10 MAVs associate with 31 BioVU clinical traits spanning neurological, autoimmune, metabolic, dermatological, and hematological diseases in EA but not AA individuals. 2) Five of these MAVs replicate the size and direction in the UK Biobank. 3) MAVs with clinical traits relevant to gut microbiome biology, including digestive and neurological diseases as well as dermatological traits, tend to exhibit complete triads in which the human genotype links with both disease risk and a microbial relative abundance direction connected to disease. Conversely, clinical traits that are not regularly linked with gut microbiome biology, including sensory and musculoskeletal functions, do not exhibit complete triads with microbial relative abundance. 4) Gut MAV eQTLs are enriched in tissue-specific gene expression profiles related to gut microbiome functions, including the gut–lung axis, gastrointestinal and cardiovascular tissues, and metabolic traits. 5) Pathways of MAV eQTLs are overrepresented in all enriched tissues for immune signaling and antigen processing pathways, while disease MAVs were most enriched in diseases of immunological dysfunction, supporting a model that immunogenetic variation and microbial relative abundance collectively influence dysregulation. 6) Finally, evolutionary analyses demonstrate significant population differentiation of a MAV-linked disease risk allele and a haplotype pattern, suggesting a selective sweep on a disease risk allele in EA and East Asian ancestry populations. Together, these results establish relationships between the genome, microbiome, and human diseases, with opportunities to account for these triads in targeted case–control disease studies while generating hypotheses for disease–microbiome associations.

As the discovery of genome-wide MAVs has predominantly been biased in populations with EA, we highlight the need to unbias microbiome studies for more inclusive and diverse research as none of the PheWAS results replicate across populations with AA. The reasons are multifaceted but are potentially related to sample sizes and disease incidence differences that may relate to health inequities. Indeed, time to treatment; access to care and health resources; and social factors, such as racial discrimination and bias, can exacerbate disparities in health and therefore, contribute to heterogeneity in population health records. While we hypothesized that some shared associations would be detected between populations, we recognize that this does not rule out the discovery of shared MAVs in the future as GWAS and microbiome sampling become larger and more representative. Connections between MAVs and ancestry are especially relevant in light of work identifying racially and ethnically varying taxa overlapping known heritable taxa (59, 109–114). While social and environmental exposures contribute to modulating microbiome compositions (115–117), how these influences converge with variation in population structure and shape disease risk and/or predisposition remain important goals for the future.

In summary, our results establish a set of relationships between the genome, microbiome, and human diseases, with opportunities to account for triadic relationships in targeted case–control disease studies. However, genetic influences on microbiome variation associated with clinical traits could also pose operational hurdles for therapies that target microbial compositions, which are presumed to be malleable but may be less so due to underlying host genetic factors. As such, human genetic influences may offer opportunities for precision diagnostics, especially as the relationship between the relative abundance of specific microbes and disease risk is further elucidated.

Methods

Collection and Annotation of Microbiome-Linked Alleles. Large-scale microbiome by genotype studies testing gut microbiome traits in fecal samples were reviewed. Studies were not considered if they did not report sufficient GWAS summary statistics for further analysis or relied on related individuals. Preference was given to studies of 1,000 or more individuals during the literature review. A significance threshold of $P < 5\times10^{-8}$ was used for inclusion of SNPs in the creation of a SNP panel for further analysis. In studies that found associations between multiple microbiome traits and an individual SNP , the microbiome trait with the lowest P value was retained as the presumptive leading microbial trait. Variants were annotated and accessed using a web interface for Variant Effect Predictor (48) and SNPNexus (49), both utilizing the Ensembl genome database (47) for reference. All computational analyses were performed using the R statistical programming language (version 3.6.2) in RStudio (version 1.4.1103).

SNPsnap Enrichment Testing and eQTL Analysis. Simulated background sets were generated using SNPsnap (default parameters, HLA not excluded) (50), which matches input SNPs with randomly drawn sets of SNPs based on allele frequency, the number of SNPs in LD, the distance to the nearest gene, and gene density. Two hundred SNPsnap sets were generated to calibrate background expectations. Quantitative trait loci annotation was performed using the using the R package Qtlizer (118) (Qtlizer::get_qtls) to retrieve GTEx (version 8) tissue-specific gene expression results (119). SNP sets were processed using custom R scripts to calculate the number of distinct SNP eQTLs in each tissue and tissue group. Tissue grouping was applied uniformly in both SNP and MAV groups to condense related tissues of a generalizable tissue type. MAV z score and one-sided P value (FDR < 0.05) were computed in each tissue and then, FDR corrected across all tissues (FDR < 0.05). Only significant tissues are shown in Fig. 2*B*. All summary statistics are shown in Dataset S4.

Reactome Pathway Analysis. Pathway analysis was performed using the Reactome Knowledgebase web platform (53). Following eQTL analysis, MAV eQTL target genes were filtered by tissue, selecting only the 15 tissues that were shown to be significantly enriched. Target genes within each tissue were then used to perform pathway overrepresentation analysis on the gene list with the Reactome web-based platform with the selected options for "Project to human" and "include interactors." P values and FDR-adjusted P values were computed under a hypergeometric model to determine if the number of selected entities (eQTL target genes) associated with a Reactome pathway is larger than expected. The entities ratio (the total entities in the pathway divided by the total entities for the entire species) and the reaction ratio (the total reactions in the pathway divided by the total reactions for the entire species) are additionally reported for each tissue.

BioVU Study Populations. The PheWAS populations were derived from the Vanderbilt University Medical Center (VUMC) BioVU repository, in which deidentified EHR data are linked to a DNA biobank extracted from discarded patient blood samples. BioVU subjects genotyped using the Illumina Infinium Multi-Ethnic Genotyping Array platform were stratified by genetic ancestry using principle component analysis (PCA) in conjunction with HapMapreference populations to define AA and EA populations. This resulted in 15,862 AA individuals and 75,408 EA individuals in total. This population was further filtered to include only individuals over 18 y old during PheWAS. Case and control sizes for individual BioVU phenotypes are listed in Datasets S6 and S9.

Phenotyping. ICD-9/10 codes from VUMC BioVU EHRs were converted to 1,815 PheCodes version 1.2. Cases were defined by individuals with two or more occurrences of an ICD code on different dates in their medical records. Controls for each phenotype were defined using PheWAS case–control definitions. PheCodes were analyzed at two case count/MAF criteria levels with 200 cases and MAF ≥ 0.01 as default parameters and 100 cases and MAF ≥ 0.25 MAF to detect associations in the smaller AA population. These levels are

supported by prior simulation to retain statistical power ≥75% when cases ≥ 100, MAF ≥ 0.25, case–control ratio > 4, and disease penetrance ≥ 0.15 (92).

Vanderbilt BioVU PheWAS Discovery Analysis. The phenome-wide association analysis was performed in R with the PheWAS package (120) using multivariable logistic regression on binary traits (phenotype present or phenotype absent) with each stratified population separately. The code for this analysis is available in GitHub (*Data Availability*). Briefly, the first 10 principle components (PCs) calculated within each ancestry were retained with the first five covariables used in the regression model within each stratified genetic ancestry group. Additionally, terms for age and sex were included. To adjust for multiple testing, *P* values were adjusted using the false discovery rate Benjamini–Hochberg method. FDR < 0.05 was considered significant.

A PheWAS association between rs139904671 and "neoplasm of unspecified nature of digestive system" was observed; however, due to the very low allele frequency, lack of extant variant annotation, and unrealistic effect size, we omit this result from downstream analysis (Dataset S6).

The UK Biobank Replication Analysis. PheWAS replication was performed using the TOPMed-imputed (121) UK Biobank PheWeb browser (122) to identify summary statistics for variants of interest among 1,419 EHR-derived phenotypes in >400,000 White British individuals (cases: 51 to 78,000; controls: 167,000 to 407,000). Analyses on binary outcomes were computed using the SAIGE method (54) to account for size and case–control imbalances. Covariables were included for genetic relatedness, sex, birth year and the first four principal components. Associated *P* values were multiple test corrected in R using "p.adjust, method = fdr."

Detecting and Annotating Differences in Evolutionary Signatures. To examine the evolutionary history of the regions identified in the PheWAS, we ran the significant variants through an evolutionary analysis pipeline (more details are in ref. 95). This pipeline is designed to detect enrichment in evolutionary signals using a variety of sequence-based metrics of selection. The value for these metrics observed for each variant is compared against 5,000 control variants matched on linkage disequilibrium structure, MAF, and proximity to genes. The pipeline produces both a *P* value and *z* score for each metric. The *z* score represents the amount of enrichment or depletion for the test variant as compared with the background distribution. The *P* value represents the statistical value of that enrichment or depilation based on the background distribution.

The sequence-based evolutionary metrics used are as follows. 1) F_{ST} (96) measures the variance of allele frequency among populations to infer population differentiation. 2) XP-EHH (97) measures haplotype homozygosity by comparing integrated extended haplotype homozygosity profiles to detect cross-population positive selection at the same SNP. 3) iES (98) is based on the iintegrated extended haplotype homozygosity (123) measure and detects haplotype homozygosity to infer positive selection over a chromosomal region. 4) PhyloP (99) calculates a substitution rate to infer positive and negative selection based on expected neutral drift. 5) The phastCons(100)100 score is the probability that each nucleotide belongs to a conserved element as calculated from the multiz alignment of 100 vertebrate species (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/phastCons100way/). 6) The LINSIGHT (101) score measures the probability of negative selection on noncoding sites. 7) The GERP (102) score measures the reduction in the quantity of substitutions in the multispecies sequence alignment as compared with the neutral expectation. 8) The ARGWEAVE (103) score measures the TMRCA as computed by the ARGweaver pipeline. 9) Beta (104) measures balanced polymorphisms to infer balancing selection.

The haplotype analysis was conducted using the 1000 Genomes Phase3 data (124). Haplotype blocks were extracted using PLINK (125) version 1.9b_5.2 using the standard "blocks" command. These haplotype blocks were then formatted into a nexus file for analysis in POPART (126). Median joining networks were then constructed and color coded. Analysis of allele conservation was performed using the "Multiz Alignment of 100 Vertebrates" track of the University of California, Santa Cruz (UCSC) genome browser (127, 128).

The interpretation of nonviable MAV variants following the simulation for null hypothesis testing with the evolutionary analysis pipeline is as follows. The four variants located on chromosome 6 (rs35026345, rs78279385, rs9357092, rs11751024) are in highly variable regions of the human genome, making sequence-based evolutionary metrics difficult to deploy. The variant rs139904671 was not present in the database used for identifying linked variants and is very rare among modern human populations. The variant rs3749147 did not produce viable results due to an unusual linkage and MAF structure that produced only 230 of the 5,000 required matched control regions.

Data Visualization. Data parsing and visualization were performed using Tidyverse version 4 (129) and ComplexHeatmap version 2.12 (130) packages. The Manhattan plot of PheWAS results by category was created in R using custom scripts. The phenogram plot was created using the Phenogram tool (131). Figs. 1 and 4 were created using illustrations from BioRender.com with full publishing rights secured.

Data Availability. PheWAS summary statistics, eQTL data, SNPsnap data, and Reactome data have been deposited in FigShare under the project "Microbiome-associated human genetic variants impact phenome-wide disease risk" (https://figshare.com/projects/Microbiome-associated_human_genetic_variants_impact_phenome-wide_disease_risk/125482) (132). Code to generate results and figures has been deposited in GitHub (https://github.com/BordensteinLaboratory/VMI_MAV) (133). Due to privacy issues concerning the human genotype data and EHRs, access is restricted to authorized Vanderbilt researchers; other researchers may contact The Vanderbilt Institute for Clinical and Translational Research (research.support.services@vumc.org) to make inquiries about data access.

ACKNOWLEDGMENTS. This work was supported by NIH NLM T32LM012412 (to R.H.G.M.), R35 GM127087 (to J.A.C.), and R01 HL142856 (to J.F.F. and J.D.M.); NIH Vanderbilt-Ingram Cancer Center Support Grant P30CA068485 (to S.R.B.); and the Vanderbilt Microbiome Innovation Center. Data for this project were obtained using BioVU and the Synthetic Derivative at VUMC, which is supported by multiple institutional, private, and federal grant sources and by Clinical and Translational Science Awards Grant ULTR000445 from National Center for Advancing Translational Sciences /NIH. This research was conducted using resources from the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN.

Author affiliations: [a]Vanderbilt Microbiome Innovation Center, Vanderbilt University, Nashville, TN 37232; [b]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37232; [c]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232; [d]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143; [e]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA 94143; [f]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232; [g]Vanderbilt Institute for Infection, Immunology and Inflammation, Vanderbilt University Medical Center, Nashville, TN 37232; and [h]Department of Pathology, Microbiology, and Immunology, School of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232

1. C. Huttenhower *et al.*; Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
2. D. Zheng, T. Liwinski, E. Elinav, Interaction between microbiota and immunity in health and disease. *Cell Res.* **30**, 492–506 (2020).
3. A. M. Martin, E. W. Sun, G. B. Rogers, D. J. Keating, The influence of the gut microbiome on host metabolism through the regulation of gut hormone release. *Front. Physiol.* **10**, 428 (2019).
4. G. Sharon *et al.*, Specialized metabolites from the microbiome in health and disease. *Cell Metab.* **20**, 719–730 (2014).
5. V. K. Gupta, S. Paul, C. Dutta, Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Front. Microbiol.* **8**, 1162 (2017).
6. R. C. E. Bowyer *et al.*, Socioeconomic status and the gut microbiome: A TwinsUK Cohort Study. *Microorganisms* **7**, 17 (2019).
7. P. Herd *et al.*, The influence of social conditions across the life course on the human gut microbiota: A pilot project with the Wisconsin longitudinal study. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **73**, 124–133 (2017).
8. A. Sharma *et al.*, Longitudinal homogenization of the microbiome between both occupants and the built environment in a cohort of United States Air Force Cadets. *Microbiome* **7**, 70 (2019).
9. H. Xie *et al.*, Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584.e3 (2016).
10. J. K. J. Goodrich *et al.*, Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
11. A. Gomez *et al.*, Host genetic control of the oral microbiome in health and disease. *Cell Host Microbe* **22**, 269–278.e3 (2017).
12. E. R. Davenport *et al.*, Genome-wide association studies of the human gut microbiota. *PLoS One* **10**, e0140301 (2015).

13. D. Awany, E. R. Chimusa, Heritability jointly explained by host genotype and microbiome: Will improve traits prediction? *Brief. Bioinform.* **22**, bbaa175 (2021).
14. J. Si, S. Lee, J. M. Park, J. Sung, G. Ko, Genetic associations and shared environmental effects on the skin microbiome of Korean twins. *BMC Genomics* **16**, 992 (2015).
15. W. Fan *et al.*, Association between human genetic variants and the vaginal bacteriome of pregnant women. *mSystems* **6**, e0015821 (2021).
16. A. Kurilshikov *et al.*, Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **53**, 156–165 (2021).
17. M. J. Bonder *et al.*, The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
18. P. Scepanovic *et al.*; Milieu Intérieur Consortium, A comprehensive assessment of demographic, environmental, and host genetic associations with gut microbiome diversity in healthy individuals. *Microbiome* **7**, 130 (2019).
19. D. A. Hughes *et al.*, Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat. Microbiol.* **5**, 1079–1087 (2020).
20. X. Liu *et al.*, A genome-wide association study for gut metagenome in Chinese adults illuminates complex diseases. *Cell Discov.* **7**, 9 (2021).
21. J. Wang *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).
22. M. C. Rühlemann *et al.*, Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. *Nat. Genet.* **53**, 147–155 (2021).
23. Y. Qin *et al.*, Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *Nat. Genet.* **54**, 134–142 (2022).
24. R. Blekhman *et al.*, Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).
25. R. Kolde *et al.*, Host genetic variation and its microbiome interactions within the Human Microbiome Project. *Genome Med.* **10**, 6 (2018).
26. E. A. Lopera-Maya *et al.*; LifeLines Cohort Study, Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. *Nat. Genet.* **54**, 143–151 (2022).
27. V. Schmidt, H. Enav, T. D. Spector, N. D. Youngblut, R. E. Ley, Strain-level analysis of *Bifidobacterium* spp. from gut microbiomes of adults with differing lactase persistence genotypes. *mSystems* **5**, e00911-20 (2020).
28. W. Turpin *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).
29. H. Mäkivuokko *et al.*, Association between the ABO blood group and the human intestinal microbiota composition. *BMC Microbiol.* **12**, 94 (2012).
30. P. Wacklin *et al.*, Secretor genotype (FUT2 gene) is strongly associated with the composition of Bifidobacteria in the human intestine. *PLoS One* **6**, e20113 (2011).
31. E. R. Davenport *et al.*, ABO antigen and secretor statuses are not associated with gut microbiota composition in 1,500 twins. *BMC Genomics* **17**, 941 (2016).
32. J. Lonsdale *et al.*; GTEx Consortium, The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
33. D. A. Knowles *et al.*, Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* **14**, 699–702 (2017).
34. E. D. Flynn *et al.*, Transcription factor regulation of eQTL activity across individuals and tissues. *PLoS Genet.* **18**, e1009719 (2022).
35. Y. Furusawa *et al.*, Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446–450 (2013).
36. E. Larsson *et al.*, Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. *Gut* **61**, 1124–1131 (2012).
37. J. G. Camp *et al.*, Microbiota modulate transcription in the intestinal epithelium without remodeling the accessible chromatin landscape. *Genome Res.* **24**, 1504–1516 (2014).
38. A. L. Muehlbauer *et al.*, Interspecies variation in hominid gut microbiota controls host gene regulation. *Cell Rep.* **37**, 110057 (2021).
39. A. A. Hibberd *et al.*, Intestinal microbiota is altered in patients with colon cancer and modified by probiotic intervention. *BMJ Open Gastroenterol.* **4**, e000145 (2017).
40. K. Dabke, G. Hendrick, S. Devkota, The gut microbiome and metabolic syndrome. *J. Clin. Invest.* **129**, 4050–4057 (2019).
41. Ö. Aydin, M. Nieuwdorp, V. Gerdes, The gut microbiome as a target for the treatment of type 2 diabetes. *Curr. Diab. Rep.* **18**, 55 (2018).
42. E. A. Franzosa *et al.*, Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
43. H. E. Groot *et al.*, Human genetic determinants of the gut microbiome and their associations with health and disease: A phenome-wide association study. *Sci. Rep.* **10**, 14771 (2020).
44. S. Sanna *et al.*, Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 600–605 (2019).
45. F. Xu *et al.*, The interplay between host genetics and the gut microbiome reveals common and distinct microbiome features for complex human diseases. *Microbiome* **8**, 145 (2020).
46. I. Danciu *et al.*, Secondary use of clinical data: The Vanderbilt approach. *J. Biomed. Inform.* **52**, 28–35 (2014).
47. K. L. Howe *et al.*, Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
48. W. McLaren *et al.*, The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
49. J. Oscanoa *et al.*, SNPnexus: A web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res.* **48**, W185–W192 (2020).
50. T. H. Pers, P. Timshel, J. N. Hirschhorn, SNPsnap: A Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* **31**, 418–420 (2015).
51. N. M. Belle *et al.*, TFF3 interacts with LINGO2 to regulate EGFR activation for protection against colitis and gastrointestinal helminths. *Nat. Commun.* **10**, 4408 (2019).
52. J. L. Ashurst *et al.*, The vertebrate genome annotation (vega) database. *Nucleic Acids Res.* **33**, D459–D465 (2005).
53. M. Gillespie *et al.*, The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* **50** (D1), D687–D692 (2022).
54. W. Zhou *et al.*, Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).
55. D. H. Parks *et al.*, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
56. M. Vacca *et al.*, The controversial role of human gut lachnospiraceae. *Microorganisms* **8**, 573 (2020).
57. T. Clavel, P. Lepage, C. Charrier, "The family *Coriobacteriaceae*" in *The Prokaryotes*, E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, F. Thompson, Eds. (Springer, Berlin, Germany, 2014), pp. 201–238.
58. S. Jangi *et al.*, Alterations of the human gut microbiome in multiple sclerosis. *Nat. Commun.* **7**, 12015 (2016).
59. A. W. Brooks, S. Priya, R. Blekhman, S. R. Bordenstein, Gut microbiota diversity across ethnicities in the United States. *PLoS Biol.* **16**, e2006842 (2018).
60. J. Chen *et al.*, Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. *Sci. Rep.* **6**, 28484 (2016).
61. H. Tremlett *et al.*; US Network of Pediatric MS Centers, Gut microbiota in early pediatric multiple sclerosis: A case-control study. *Eur. J. Neurol.* **23**, 1308–1321 (2016).
62. J. T. Russell *et al.*, Genetic risk for autoimmunity is associated with distinct changes in the human gut microbiome. *Nat. Commun.* **10**, 3621 (2019).
63. P. R. Sternes *et al.*, HLA-A alleles including HLA-A29 affect the composition of the gut microbiome: A potential clue to the pathogenesis of birdshot retinochoroidopathy. *Sci. Rep.* **10**, 17636 (2020).
64. L. M. Cox, H. L. Weiner, The microbiome requires a genetically susceptible host to induce central nervous system autoimmunity. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 27764–27766 (2020).
65. G. Serena, C. Davies, M. Cetinbas, R. I. Sadreyev, A. Fasano, Analysis of blood and fecal microbiome profile in patients with celiac disease. *Hum. Microbiome J.* **11**, 100049 (2019).
66. M. M. Leonard *et al.*; CD-GEMM Team, Microbiome signatures of progression toward celiac disease onset in at-risk children in a longitudinal prospective cohort study. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2020322118 (2021).
67. A. D. Kostic *et al.*; DIABIMMUNE Study Group, The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).
68. I. Leiva-Gea *et al.*, Gut microbiota differs in composition and functionality between children with type 1 diabetes and MODY2 and healthy control subjects: A case-control study. *Diabetes Care* **41**, 2385–2395 (2018).
69. M. Sikora *et al.*, Gut microbiome in psoriasis: An updated review. *Pathogens* **9**, 463 (2020).
70. S. Yegorov *et al.*, Psoriasis is associated with elevated gut IL-1α and intestinal microbiome alterations. *Front. Immunol.* **11**, 571319 (2020).
71. X. Zhang, L. Shi, T. Sun, K. Guo, S. Geng, Dysbiosis of gut microbiota and its correlation with dysregulation of cytokines in psoriasis patients. *BMC Microbiol.* **21**, 78 (2021).
72. M. T. Sorbara *et al.*, Functional and genomic variation between human-derived isolates of lachnospiraceae reveals inter- and intra-species diversity. *Cell Host Microbe* **28**, 134–146.e4 (2020).
73. M. Mizuno, D. Noto, N. Kaga, A. Chiba, S. Miyake, The dual role of short fatty acid chains in the pathogenesis of autoimmune disease models. *PLoS One* **12**, e0173032 (2017).
74. A. D. Truax *et al.*, The inhibitory innate immune sensor NLRP12 maintains a threshold against obesity by regulating gut microbiota homeostasis. *Cell Host Microbe* **24**, 364–378.e6 (2018).
75. D. Takahashi *et al.*, Microbiota-derived butyrate limits the autoimmune response by promoting the differentiation of follicular regulatory T cells. *EBioMedicine* **58**, 102913 (2020).
76. H. E. Groot *et al.*, Genetically determined ABO blood group and its associations with health and disease. *Arterioscler. Thromb. Vasc. Biol.* **40**, 830–838 (2020).
77. F. Turroni *et al.*, *Bifidobacterium bifidum*: A key member of the early human gut microbiota. *Microorganisms* **7**, 544 (2019).
78. A. O'Callaghan, D. van Sinderen, Bifidobacteria and their role as members of the human gut microbiota. *Front. Microbiol.* **7**, 925 (2016).
79. M. Trøseid, G. Ø. Andersen, K. Broch, J. R. Hov, The gut microbiome in coronary artery disease and heart failure: Current knowledge and future directions. *EBioMedicine* **52**, 102649 (2020).
80. X. Wang *et al.*, Receptor-mediated ER export of lipoproteins controls lipid homeostasis in mice and humans. *Cell Metab.* **33**, 350–366.e7 (2021).
81. A. Buniello *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
82. S. Falati *et al.*, Platelet PECAM-1 inhibits thrombus formation in vivo. *Blood* **107**, 535–541 (2006).
83. I. Robles-Vera *et al.*, Protective effects of short-chain fatty acids on endothelial dysfunction induced by angiotensin II. *Front. Physiol.* **11**, 277 (2020).
84. M. Kummen *et al.*, Gut microbiota signature in heart failure defined from profiling of 2 independent cohorts. *J. Am. Coll. Cardiol.* **71**, 1184–1186 (2018).
85. R. A. Hasan, A. Y. Koh, A. Zia, The gut microbiome and thromboembolism. *Thromb. Res.* **189**, 77–87 (2020).
86. Y. Mohammed *et al.*, The intestinal microbiome potentially affects thrombin generation in human subjects. *J. Thromb. Haemost.* **18**, 642–650 (2020).
87. B. N. Cronstein, R. Terkeltaub, The inflammatory process of gout and its treatment. *Arthritis Res. Ther.* **8** (suppl. 1), S3 (2006).
88. A. G. Stack *et al.*, Gout and the risk of advanced chronic kidney disease in the UK health system: A national cohort study. *BMJ Open* **9**, e031550 (2019).
89. C. Li *et al.*, Genome-wide association analysis identifies three new risk loci for gout arthritis in Han Chinese. *Nat. Commun.* **6**, 7041 (2015).
90. G. Sandoval-Plata, K. Morgan, A. Abhishek, Variants in urate transporters, ADH1B, GCKR and MEPE genes associate with transition from asymptomatic hyperuricaemia to gout: Results of the first gout versus asymptomatic hyperuricaemia GWAS in Caucasians using data from the UK Biobank. *Ann. Rheum. Dis.* **80**, 1220–1226 (2021).
91. Y. Ning *et al.*, Characteristics of the urinary microbiome from patients with gout: A prospective study. *Front. Endocrinol. (Lausanne)* **11**, 272 (2020).
92. A. Verma *et al.*, A simulation study investigating power estimates in phenome-wide association studies. *BMC Bioinformatics* **19**, 120 (2018).
93. M. L. Benton *et al.*, The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* **22**, 269–283 (2021).
94. C. S. Carlson *et al.*; PAGE Consortium, Generalization and dilution of association results from European GWAS in populations of non-European ancestry: The PAGE study. *PLoS Biol.* **11**, e1001661 (2013).
95. A. L. LaBella *et al.*, Accounting for diverse evolutionary forces reveals mosaic patterns of selection on human preterm birth loci. *Nat. Commun.* **11**, 3731 (2020).
96. K. E. Holsinger, B. S. Weir, Genetics in geographically structured populations: Defining, estimating and interpreting F(ST). *Nat. Rev. Genet.* **10**, 639–650 (2009).
97. P. C. Sabeti *et al.*; International HapMap Consortium, Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

98. K. Tang, K. R. Thornton, M. Stoneking, A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* **5**, e171 (2007).
99. K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, A. Siepel, Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
100. A. Siepel *et al.*, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
101. Y. F. Huang, B. Gulko, A. Siepel, Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
102. G. M. Cooper *et al.*; NISC Comparative Sequencing Program, Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
103. M. D. Rasmussen, M. J. Hubisz, I. Gronau, A. Siepel, Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
104. K. M. Siewert, B. F. Voight, Detecting long-term balancing selection using allele frequency correlation. *Mol. Biol. Evol.* **34**, 2996–3005 (2017).
105. W. Turpin *et al.*; GEM Project Research Consortium, Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
106. T. S. Ibekwe, G. T. A. Ijaduola, Meniere's disease: Rare or underdiagnosed among Africans. *Eur. Arch. Otorhinolaryngol.* **264**, 1399–1403 (2007).
107. D. Reich *et al.*, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
108. M. Meyer *et al.*, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
109. K. Liu *et al.*, Ethnic Differences Shape the Alpha but Not Beta Diversity of Gut Microbiota from School Children in the Absence of Environmental Differences. *Microorganisms* **8**, 254 (2020).
110. M. Deschasaux *et al.*, Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nat. Med.* **24**, 1526–1531 (2018).
111. S. Jayaraman, Of ethnicity, environment, and microbiota. *Cell. Mol. Immunol.* **16**, 106–108 (2019).
112. M. R. Mason, H. N. Nagaraja, T. Camerlengo, V. Joshi, P. S. Kumar, Deep sequencing identifies ethnicity-specific bacterial signatures in the oral microbiome. *PLoS One* **8**, e77287 (2013).
113. J. M. Fettweis *et al.*; The Vaginal Microbiome Consortium, Differences in vaginal microbiome in African American women versus women of European ancestry. *Microbiology (Reading)* **160**, 2272–2282 (2014).
114. J. C. Stearns *et al.*; NutriGen Alliance, Ethnic and diet-related differences in the healthy infant microbiome. *Genome Med.* **9**, 32 (2017).
115. A. L. Dunlop *et al.*, Stability of the vaginal, oral, and gut microbiota across pregnancy among African American women: The effect of socioeconomic status and antibiotic exposure. *PeerJ* **7**, e8004–e8004 (2019).
116. K. Findley, D. R. Williams, E. A. Grice, V. L. Bonham, Health disparities and the microbiome. *Trends Microbiol.* **24**, 847–850 (2016).
117. J. Xu *et al.*, Ethnic diversity in infant gut microbiota is apparent before the introduction of complementary diets. *Gut Microbes* **11**, 1362–1373 (2020).
118. M. Munz *et al.*, Qtlizer: Comprehensive QTL annotation of GWAS results. *Sci. Rep.* **10**, 20417 (2020).
119. A. Battle, C. D. Brown, B. E. Engelhardt, S. B. Montgomery; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)–Analysis Working Group; Statistical Methods groups–Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site–NDRI; Biospecimen Collection Source Site–RPCI; Biospecimen Core Resource–VARI; Brain Bank Repository–University of Miami Brain Endowment Bank; Leidos Biomedical–Project Management; ELSI Study; Genome Browser Data Integration &Visualization–EBI; Genome Browser Data Integration &Visualization–UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
120. R. J. Carroll, L. Bastarache, J. C. Denny, R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376 (2014).
121. D. J. Burgess, The TOPMed genomic resource for human health. *Nat. Rev. Genet.* **22**, 200 (2021).
122. S. A. Gagliano Taliun *et al.*, Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).
123. P. C. Sabeti *et al.*, Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
124. A. Auton *et al.*; 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
125. S. Purcell *et al.*, PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
126. J. W. Leigh, D. Bryant, popart: Full-feature software for haplotype network construction. *Methods Ecol. Evol.* **6**, 1110–1116 (2015).
127. M. Blanchette *et al.*, Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
128. W. J. Kent *et al.*, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
129. H. Wickham *et al.*, Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
130. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
131. D. Wolfe, S. Dudek, M. D. Ritchie, S. A. Pendergrass, Visualizing genomic information across chromosomes with PhenoGram. *BioData Min.* **6**, 18 (2013).
132. R. H. G. Markowitz *et al.*, PheWAS and genomic data associated with "Microbiome-associated human genetic variants impact phenome-wide disease risk." Figshare. https://figshare.com/projects/Microbiome-associated_human_genetic_variants_impact_phenome-wide_disease_risk/125482. Deposited 14 March 2022.
133. R. H. G. Markowitz *et al.*, BordensteinLaboratory/MAV. GitHub. https://github.com/BordensteinLaboratory/MAV. Deposited 3 November 2021.