Challenges and future directions for studying effects of host genetics on the gut microbiome

Sanna, Serena; Kurilshikov, Alexander; van der Graaf, Adriaan; Fu, Jingyuan; Zhernakova, Alexandra

Published in:
Nature genetics

DOI:
10.1038/s41588-021-00983-z

Publication date:
2022

nature genetics

PERSPECTIVE

https://doi.org/10.1038/s41588-021-00983-z

Challenges and future directions for studying effects of host genetics on the gut microbiome

Serena Sanna[1,2]✉, Alexander Kurilshikov[2], Adriaan van der Graaf[2,3], Jingyuan Fu[2,4] and Alexandra Zhernakova[2]✉

The human gut microbiome is a complex ecosystem that is involved in its host's metabolism, immunity and health. Although interindividual variations in gut microbial composition are mainly driven by environmental factors, some gut microorganisms are heritable and thus can be influenced by host genetics. In the past 5 years, 12 microbial genome-wide association studies (mbGWAS) with >1,000 participants have been published, yet only a few genetic loci have been consistently confirmed across multiple studies. Here we discuss the state of the art for mbGWAS, focusing on current challenges such as the heterogeneity of microbiome measurements and power issues, and we elaborate on potential future directions for genetic analysis of the microbiome.

Recent progress in genetics has been driven by the development of inexpensive wide-scale genotyping methods and multiplex sequencing technologies. Since 2007, genome-wide association studies (GWAS) have become a routine analysis in population genetics and genetic epidemiology, identifying thousands of genetic variants associated with complex traits. In parallel, a new generation of sequencing technologies has changed microbiology research, allowing us to move beyond the study of single microorganisms to studying whole microbial communities, including holobiont systems—that is, microbiomes and their hosts. The composition of the microbial community in the human gut (the gut microbiome) can now be efficiently quantified using second-generation and third-generation sequencing technologies. Currently, two major methods are widely used for microbial quantification: 16S rRNA gene sequencing (16S) and metagenomic sequencing (MGS). The 16S method is focused on targeted sequencing of highly polymorphic domains of the bacterial 16S rRNA gene, and the technique allows for reasonably accurate annotation of bacteria and archaea at the level of genera. By contrast, MGS sequences all genetic material present in the sample, allowing for annotation of microbial taxa down to the species and strain level, enabling analysis of other taxonomic groups including viruses, fungi and protozoa. MGS also enables abundance estimates of individual microbial genes and gene families, which can be merged to quantify functional pathways and clusters, including biosynthesis pathways of specific xenobiotics and clusters of antibiotic resistance and virulence genes.

In the past 5–10 years, the human gut microbiome, the largest and most diverse microbial community in the human body, has been at the center of microbiome research, and hundreds of microbial species have now been identified. The gut microbiome is known to show remarkable interindividual variation: only a small number of genera or species (fewer than 20) are shared by >95% of individuals[1–3], a shared subset called the core microbiome. With the analysis of larger cohorts, the number of shared bacteria has continued to decrease as more new, rare bacteria are identified. In recent examples, a metagenomic study of 8,208 Dutch individuals identified 733 bacterial species[4] (not yet peer reviewed), an analysis of 5,959 Finnish individuals identified 1,123 bacterial species[5] and an analysis of 9,428 metagenomes from different populations and body sites (mostly from stool) identified 4,930 species-level taxa[6].

Many studies have demonstrated that interindividual variation of the gut microbiome is mainly determined by environmental factors such as diet, medications, smoking, the presence of pets and other factors[1,7,8]. However, there is also evidence from twin, family and population studies[4,9–14] for the heritability of some gut microorganisms, and the underlying genetic components remain of interest for understanding the mutualistic host–microorganism relationship and coevolution. In this Perspective, we summarize current knowledge on the effect of host genetics on the gut microbiome from GWAS (mbGWAS), estimate the detection power of genetic analysis for microorganisms with various abundances and discuss future perspectives for genetic studies of microbiome communities.

Analysis of the effect of host genetics on the human microbiome started in 2014–2015 with several small studies (<100 individuals)[15,16]. In 2016, heritability analysis and mbGWAS were performed on a set of 1,126 twins from the TwinsUK cohort[10]. These analyses established that a proportion of gut microorganisms showed substantial heritability: 90 out of 945 reported taxa (9.5%) showed heritability (h^2) greater than 0.20. These heritability estimates were later confirmed in Canadian[11] and Dutch family studies[3,4] and are in the range of the heritability reported for many other human complex traits, such as fasting glucose levels ($h^2=0.31$), insulin levels ($h^2=0.25$) and blood pressure ($h^2=0.15$)[17]. In the TwinsUK study, the strongest heritability was observed for the *Christensenellaceae* family and related taxa, which were also linked to metabolic parameters in the individual. Other heritable bacteria included bifidobacteria, the abundance of which was related to the functional genetic variant near the lactase gene (*LCT*), a finding that was recently confirmed by other studies[2,3,5,18]. In 2016, three mbGWAS in Dutch, Canadian and German populations reported associations of dozens of loci with the abundance of various bacteria and other microbiome traits (β-diversity, microbial pathways and bacterial

[1]Institute for Genetic and Biomedical Research (IRGB), National Research Council (CNR), Monserrato, Cagliari, Italy. [2]Department of Genetics, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. [3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [4]Department of Pediatrics, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. ✉e-mail: serena.sanna@irgb.cnr.it; sasha.zhernakova@gmail.com

Table 1 | Summary of published mbQTL studies

No.	Authors	Population	*n* samples	Method	*n* loci, $P < 5 \times 10^{-8}$
1	Goodrich et al.[10]	UK	1,126 twin pairs	16S	22 taxa
2	Wang et al.[20]	German	1,812	16S	40 taxa 42 BD
3	Turpin et al.[11]	Canadian	1,561	16S	55 taxa
4	Bonder et al.[19]	Dutch	1,514	MGS	9 taxa 21 PW 12 GO
5	Rothshild et al.[7]	Israeli	1,046 (814[a])	16S	43 taxa 2 BD
6	Hughes et al.[12]	Belgian, German	3,890	16S	13 taxa
7	Xu et al.[13]	Chinese	1,475 + 199[b]	16S	10 taxa 1 BD
8	Liu et al.[14]	Chinese	1,295	MGS	36 taxa 8 PW 4 BD
9	Rühlemann et al.[18]	German	8,956	16S	34 taxa 4 BD
10	Kurilshikov et al.[2]	Many populations	18,340	16S	31 taxa
11	Qin et al.[5]	Finnish	5,959	MGS	422 taxa
12	Lopera et al.[3]	Dutch	7,738	MGS	6 taxa 14 PW

All 12 mbQTL studies with >1,000 independent samples published before May 2021. Taxa, analysis of taxonomic abundance or presence or absence; BD, β-diversity; PW, MetaCyc pathways; GO, gene ontology terms. Studies are listed in chronological order. Genome-wide significant SNPs reported by each study and associated microbial traits (taxa, BD, PW, GO) are listed in Supplementary Table 1. The number and name of studies reporting each locus can be found in Supplementary Table 2. [a]In the study by Rothshild et al.[7], 1,046 individuals were available, but only 814 individuals were included in the mbQTL analysis. [b]In the study by Xu et al.[13], a replication cohort (199 individuals) was genotyped for only part of the SNPs identified in the discovery cohort.

presence)[11,19,20]. Findings from these studies included several relevant genes, such as genes coding for C-type lectins[19], which are known to regulate microbiota composition in shrimps and mosquitoes[21], and the vitamin D receptor gene[20]. However, with the exception of the *LCT* locus, none of these results have been replicated across mbGWAS. Over the past few years, several other studies have explored the effect of genetics on microbiome composition. To date, there have been 12 mbGWAS published that included close to or more than 1,000 participants (Table 1 and Fig. 1). All 12 of these studies reported dozens of hits at the genome-wide significance level ($P < 5 \times 10^{-8}$) but results for only two loci, *LCT* and *ABO*, were consistently replicated across at least three studies (Supplementary Tables 1 and 2 and Supplementary Note).

The *LCT* locus

Genetic variants at or near the *LCT* gene were independently reported to be associated with the clade of Actinobacteria, genus *Bifidobacterium* and its related species. This association has been reported at genome-wide significance in UK[10], Dutch[3], Canadian[11] and Finnish populations[5] and in the meta-analysis of the MiBioGen consortium[2]. Other cohorts showed this association at lower significance levels[7,18,20], making this locus the most validated finding of microbial quantitative trait locus (mbQTL) studies thus far. Indeed, bifidobacteria are among the most heritable taxa reported in UK[9,10] and Dutch[4] populations. The strongest association with bifidobacteria was observed for the functional variant rs4988235 (NC_000002.12:g.135,851,076G > A) or its proxies located near the *LCT* gene. *LCT* encodes the lactase enzyme that cleaves the milk sugar lactose into glucose and galactose. The rs4988235*G/G genotype corresponds to the lactase non-persistence phenotype, which is a decreased ability to metabolize lactose after weaning. These alleles have been under selective pressure due to animal domestication and consumption of milk after weaning, which has driven an increase in the frequency of lactase-persistence alleles[22]. Interestingly, the lactase non-persistence genotype is associated with higher gut abundance of bifidobacteria, which have the ability to degrade lactose[23], and this association is dependent on milk consumption[19].

The *ABO* locus and its interaction with *FUT2*

Microbiome associations with the *ABO* locus have been reported in German, Dutch and Finnish cohorts[3,5,18]. Interestingly, the underlying associations and reported taxa are not the same across studies. In the German cohort, two independent SNPs in the *ABO* locus were associated with the abundance of *Faecalibacterium* and *Bacteroides*. In the Finnish cohort, variants in partial linkage disequilibrium (LD) near *ABO* were associated with the abundance of *Faecalicatena lactaris* and *Collinsella*. In the Dutch population, the same and other variants in LD are associated with *Bifidobacterium* abundance, the lactose-degradation pathway and *Collinsella* abundance. The association of the *ABO* locus with the gut microbiome was also reported in pigs, in which a common deletion that inactivates the *ABO* gene is associated with the abundance of the family *Erysipelotrichaceae*[24].

Despite variations in the associated taxa, all three human studies identified an interaction between *ABO* and *FUT2* variants and bacterial abundance: a nonsense mutation in the *FUT2* gene (rs601338, NC_000019.10:g.48,703,417G > A) determines the expression of ABO antigens on mucosal cells. Specifically, individuals homozygous for the G allele (rs601338*G/G genotype) do not express or expose the A or B antigens of ABO on their mucosa, including their gut mucosa. These individuals are called non-secretors. In all studies, the effect of *ABO* on associated bacteria was dependent on the host's secretor status. Associations between secretor status for the *FUT2* gene and the gut microbiome are expected but were only observed at the genome-wide significant level in one study[2], although the large German cohort[18] also reported a suggestive association of this locus as did earlier smaller studies of the colonic

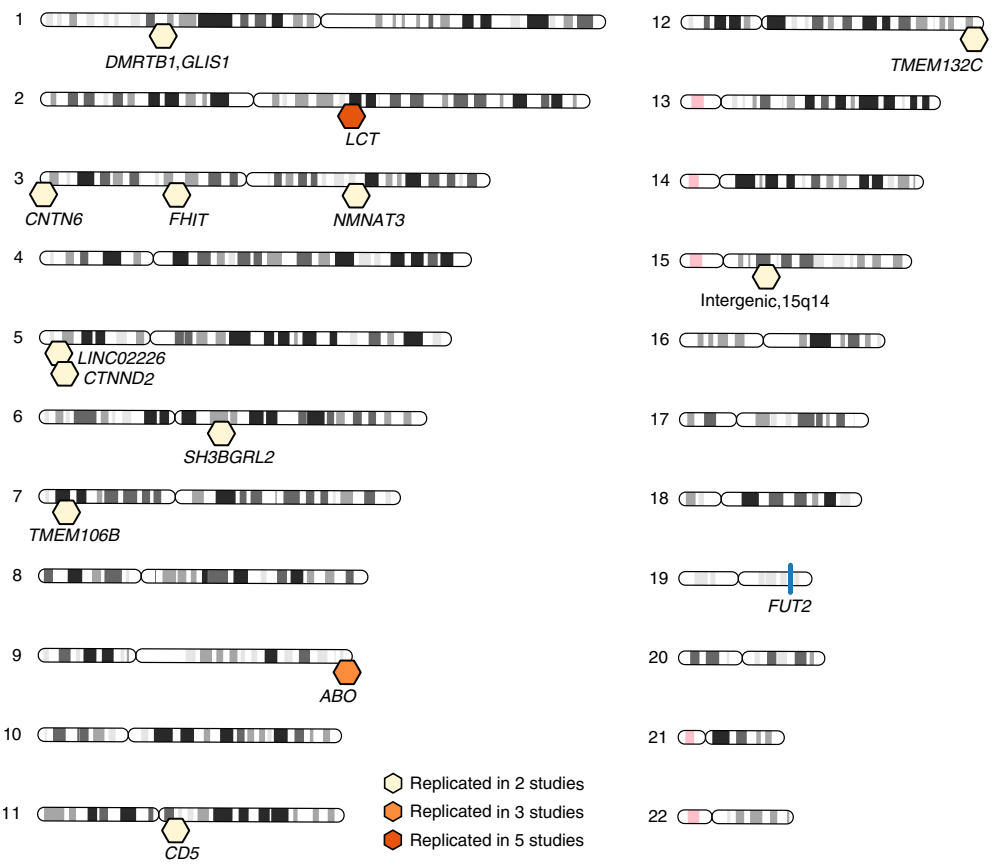

Fig. 1 | Genomic loci reported at genome-wide significance in 12 microbiome GWAS. The figure shows chromosome ideograms annotated with colored hexagons. Each hexagon represents a genomic region found to be associated at the genome-wide significant level by at least two of the 12 studies discussed in this article. Hexagons are colored according to the number of studies reporting signals in the same genomic region, and genes located in these loci are indicated. The location of the *FUT2* gene is also highlighted. Chromosome lengths and bands are depicted according to the National Center for Biotechnology Information genome decoration page and refer to genome build GRCh37/hg19. The number and names of studies reporting each locus as well as loci reported in only one study can be found in Supplementary Table 2. Of note, none of the loci reported by $n=2$ studies were reported to be associated with bacteria from the same family in the two studies that reported them, and their corresponding top hits were in weak LD ($r^2<0.1$ for Europeans in 1000 Genomes, with the exception of the *NMNAT3* locus, where r^2 was 0.34).

microbiome in patients with Crohn's disease[25], the bile microbiome in patients with primary sclerosis cholangitis[26] and others[27]. These observations and the consistent results of gene–gene interaction analysis of *ABO* and *FUT2* loci suggest that this locus will continue to be found in future studies with increased sample size.

Other potentially interesting loci

Beyond the *LCT* and *ABO* loci, another 546 loci have been reported at $P<5\times10^{-8}$ in at least one of the 12 studies (Fig. 1 and Supplementary Table 1). Of these, 11 loci were reported by two studies and point to potentially interesting candidate genes. Examples include *CD5*, which plays a role in T cell proliferation and survival and other immune functions[28], and *RBP1*, the protein product of which is involved in transport of retinol (vitamin A alcohol) from the liver to peripheral tissues. However, it is unclear whether these 11 loci represent signals that were truly replicated across studies. In fact, none of these loci were reported to be associated with bacteria from the same family in the two studies that reported them, and their corresponding top hits were in weak LD ($r^2<0.35$ in Europeans in 1000 Genomes).

We expect that other genetic signals exist for these and other taxa. Both the MiBioGen consortium and a Dutch population study identified a positive correlation between the number of significant and suggestive hits and the heritability of microorganisms, suggesting that a larger sample size (and thus higher power) is required for identification of additional genetic loci[2,3].

Power remains a major issue of mbGWAS

The poor replicability of most genome-wide associations detected by mbGWAS is the result of biological realities and ongoing methodological issues. The microbiome community comprises hundreds of species, but only a few are present in almost all samples and many are present in only a subset. According to estimates from population cohorts[4,5], the majority of taxa will be present in less than 50% of the samples, leading to a halving of the effective sample size for genetic analyses[3,5].

Population heterogeneity can also interfere with replication and induce false positive or false negative associations as was first noted in classical GWAS. For example, an association of the *FTO* locus with type 2 diabetes (T2D) showed unexpectedly low replication rates[29–31]. It was later found that *FTO* variants actually exert their effect on body mass index rather than on T2D directly; thus, in body mass index-matched case–control cohorts, the effects of *FTO* on T2D were not detected[32]. Considering that the microbiome is heavily influenced by diet and environment, population heterogeneity may also play a major role in mbGWAS, with unmatched cohorts likely to show low cross-study replication rates. In addition to these biological aspects, multiple methodological issues related to

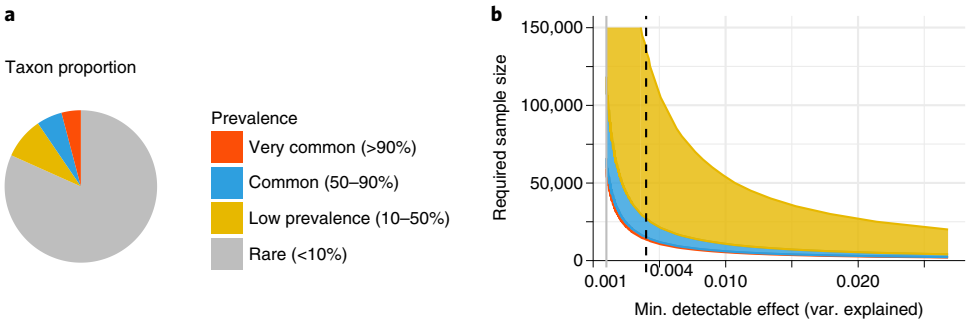

Fig. 2 | Power analysis for different taxon prevalence. a, Pie chart showing the proportion of taxa with different prevalences that would be detectable in a study cohort, according to observations from the Dutch Microbiome Project[4]. **b**, The total sample size (*y* axis) required to detect a genetic effect (variance (var.) explained (*x* axis)), with 80% power and *P* value $<1\times10^{-10}$, for different taxon prevalences. Taxon prevalences follow the color key legend in **a**). For a taxon with prevalence *q*, the required total sample size (n_{tot}) for the cohort to be studied is calculated as $n_{tot}=n_{eff}+n_{eff}\times(1-q)\times q^{-1}$, where n_{eff} is the estimated effective sample size needed to detect that genetic effect. The genetic effect is represented here by variance explained, which accounts for variations in both the additive effect size and minor allele frequency. Shaded areas (using the colors and prevalences indicated in **a**) represent estimates for a given range of prevalences. The dashed vertical line indicates an effect of 0.4%, that is, half of the effect observed by Lopera et al. for variants at the *LCT* locus on *B. adolescentis*[3]. Power estimates were derived in RStudio (version 1.03.136) as described by Lopera et al.[3]. Min., minimum.

sample processing and metagenomic data processing might lead to low replication rates, as we will discuss below.

Furthermore, the classical genome-wide significance threshold of $P<5\times10^{-8}$ is probably too lenient when analyzing hundreds or thousands of microbiome traits, and a study-wide threshold that accounts for multiple tests should be considered instead. In fact, the two loci that consistently replicate across cohorts, *LCT* and *ABO*, are the only ones passing this more stringent threshold in the larger mbGWAS[2,3,5,18], with the only exception to this being a signal near the *MED13L* gene found in the Finnish population. However, this variant is very rare in non-Finnish populations (minor allele frequency of 0.0003 in the Genome Aggregation Database in non-Finnish Europeans) and thus was not tested in non-Finnish cohorts.

The requirement for a more stringent threshold underlines the need for larger sample sizes to detect robust associations and additional loci. This concept may seem obvious, as it is what we have learned after 15 years of GWAS. After initial discoveries of associated loci with large effects made in only a few hundred samples[33,34], GWAS rapidly moved to using datasets of several thousands of samples, and some GWAS now comprise more than a million individuals[29,35–40]. Compared to human quantitative phenotypes, the need for very large samples is even more pressing in mbGWAS because most taxa are present in <10% of samples. However, there is no absolute estimate of the minimum number of samples necessary to detect new mbGWAS loci because their effect sizes are not known beforehand. A good guideline in this situation is to estimate a minimum detectable effect at a given sample size, which will depend on the genetic architecture of the trait and not strictly on its heritability. For example, even for highly heritable traits such as height ($h^2\approx0.8$–0.9)[17,41], individual common variants (such as those located at the *HMGA2* and *GDF5* loci[42,43]) still only account for 0.3–0.7% of total variance, while, for others traits with lower heritability estimates, such as fetal hemoglobin ($h^2\approx0.6$)[17], very large effect sizes have been discovered (common variants in the *BCL11A* gene explain 8–14% of the variance of fetal hemoglobin levels)[44,45].

Results from previous and recent mbGWAS can be used to estimate an upper bound for the largest effects for microbiome traits, thereby allowing us to speculate on the sample size needed to detect additional loci. For example, an association with *Bifidobacterium adolescentis* at the *LCT* locus explains 0.8% of variance, which means that at least 20,000 samples would be needed to detect association at a locus that explains an effect half that size (0.4%). *B. adolescentis* is a common taxon (present in >80% of samples). For less common taxa (prevalence of 10–50%), a dataset of ~30,000–135,000 samples would be needed to detect a similar genetic effect (Fig. 2). However, associated genetic variants with larger effect sizes could exist for other microbial taxa. For example, for *Bifidobacterium bifidum*, another bacterial species of the *Bifidobacteriaceae* family present in only 26.3% of samples, genetic variants at the *ABO* locus were found to explain 2.7% of variance. While this estimate could be inflated due to the winner's curse phenomenon[46] or overestimated by unaccounted-for gene–environment interactions, it does indicate that the polygenic architecture that underlies microbial traits is complex and likely to differ substantially among microbial taxa.

Other challenges in mbQTL studies

In addition to the need for larger sample sizes, several other aspects are important for the success of future mbGWAS. Technical variations, such as the use of different DNA isolation methods, the choice of 16S domain or the use of different pipelines and reference databases for 16S amplicon data can greatly influence the results of microbial identification and abundance. For example, in the MiBioGen consortium, archaea were present in 25–35% of individuals from various cohorts sequenced for the variable regions V3 and V4 but were not detected at all in cohorts that used the variable regions V1 and V2[2]. MGS methods should in principle provide very high resolution that can identify microbial taxa down to the strain level; however, widely used reference-based methods (that is, MetaPhlAn or Kraken) do not currently provide complete and accurate taxonomic annotation. These technical differences can partly explain low replication rates of heritability analyses. For example, *Christensenellaceae* was identified as the most heritable bacteria in twin studies, but it is not present in MetaPhlAn2 and was therefore not investigated in many MGS studies. De novo metagenomic assembly pipelines should in theory allow for identification of many new species, but they require enormously high per-sample read coverage to assemble low-abundance microbial genomes. Further precision in microbiome characterization can also be gained through the use of better measurement methodology, such as DNA extraction kits that lead to less bias in the microbial composition obtained, quantification of microbial cells in the sample and longer sequencing reads as well as other methods[47,48].

Where will the field go?

Combining studies and traits to gain power. As we have discussed, sample sizes of tens of thousands of microbiome samples are

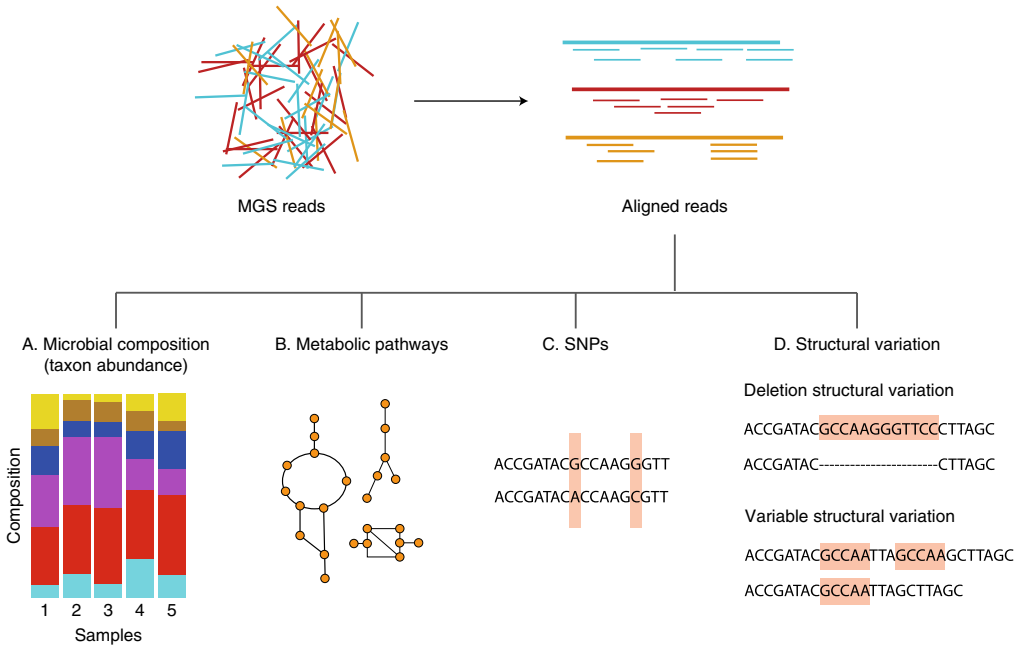

Fig. 3 | Characterizing microbial composition and genetic landscape from MGS. Schematic representation of the information that can be derived from MGS data regarding microbiome composition, functional pathways and genetic variations of bacterial genomes.

necessary to identify sufficient genetic locations and move beyond the currently established loci. Currently available and soon-to-be available middle-sized MGS cohorts could be meta-analyzed to analyze approximately 20,000–30,000 samples, provided that taxonomic classification is harmonized across studies. Furthermore, multi-trait GWAS approaches that leverage intertrait correlations to reduce phenotype variability could also be used to gain power[49].

Microbial genetic variants (SNPs and structural variations) as alternative phenotypes. To date, mbGWAS have mostly focused on microbial composition, that is, the abundance of microbial taxa or the abundance of microbial functionality, for example, microbial pathways. However, we also need to recognize that the functionality of the gut microbiome may not only be reflected by the abundance of certain species or metabolic pathways but also by genetic variants in microbial genomes. The gut microbiome contains 100–1,000 times more genes than the human genome, and its genetic landscape can dynamically adapt to environmental exposures and changes via mutagenesis. There is evidence that suggests that the gut microbiome has co-adapted to the environment with the human genome during the expansion of the human population across the globe[22]. Mutations can occur in bacterial genomes, but the host may maintain and pass bacterial genetic variants to the next generation through selective pressure. Moreover, co-occurrence of genetic variants in the human genome and the gut microbiome can also be expected due to the selection via the same environmental exposures.

Association analysis of genetic variants in the microbiome faces several challenges. First, the genetic landscape of the gut microbiome itself remains largely unexplored, even though this is technically feasible with MGS data (Fig. 3). Major attempts include the study by Schloissnig et al. that revealed 10.3 million SNPs and many other types of genetic variations in 252 fecal samples[50] and the study by Xie et al. that revealed profiles of >8 million bacterial SNPs and showed a higher similarity between twins that decreased slowly after decades of living apart[51]. More recently, Zeevi et al. developed the SVfinder pipeline and reported over 7,000 structural variations in the gut microbiome[52]. These studies laid the foundation to explore bacterial SNP and structural variation profiles. In addition, several bioinformatic tools have been developed to call SNPs from MGS reads, for example, metaSNV[53] and inStrain[54]. Moreover, general probability-based SNP-calling tools such as HaplotypeCaller[55] can also be used to call bacterial SNPs with excellent accuracy and sensitivity[56]. In future, long-read sequencing, single-cell microbial sequencing and deep sequencing of microbial isolates should provide more accuracy and resolution for identification of microbial genetic variants. Second, the mutation rate of bacterial genomes is typically around 0.001 per generation but varies greatly between different species[57]. Analysis of the interaction of host genetics with bacterial genetic variation should therefore focus on genetic variants that are temporally stable. A recent study assessed the genetic stability of the gut microbiome over 4 years in 338 individuals and identified several species that show individual specificity and temporal stability in their genetic makeup[58]. The individual specificity of microbial genetic makeup can be attributed not only to individual environmental exposures but also to host genetics. Interestingly, bacterial species that colonize the gut in early life via mother-to-baby transmission are genetically stable over time and have high heritability, for example, *Bifidobacterium* species. Third, association analysis of microbial genetic variants, rather than of abundance or presence or absence, brings both advantages and challenges. The advantage is the definition of more specific microbial 'phenotypes' with higher resolution, similar to the analysis of endophenotypes of diseases, which could enhance power in smaller samples[59,60]. The downside is the increase in the number of statistical tests that need to be performed, which will have a negative effect on detection power, already the biggest issue in mbGWAS. The number of genetic variants may be at the scale of millions or even trillions, which is much larger than the number of identified common taxa and pathways. We envisage that, with sample sizes achievable in the near term, use of dimensionality-reduction approaches on microbial genomes (for example, focusing on haplotypes or on coding SNPs) would be an interesting avenue to gain early insights into human genetic–microbial genetic associations.

Bacterial metabolites and other omic layers. Notably, omic approaches have also been extended to the microbiome field. For instance, metatranscriptomic and metaproteomic data have been informative about bacterial pathway activity relevant to inflammatory bowel disease and colon cancer[61–63]. Fecal metabolomics can be considered as a functional readout of the gut microbiome, which is dominantly determined by the gut microbiome rather than by host genetics[64]. By contrast, the plasma metabolome is often seen as the outcome of host–microorganism interactions[65]. Therefore, omic readouts of the microbiome can be treated as bacterial endophenotypes to assess the impact of host genetics on microbial activities and functionality. When using metabolomic data as a microbial readout, it is important to recognize that different scenarios face different challenges. First, it will be important to distinguish metabolites that are only produced by the microbiome from those that are under the control of both host genetics and the microbiome. For instance, short-chain fatty acids, some vitamin families and essential amino acids can only be obtained from diet or produced by the microbiome. However, synthesis of some metabolites, such as secondary bile acids (BA) and trimethylamine *N*-oxide, requires enzymes for which production is controlled by both the human genome and the gut microbiome. For these metabolites, it is important to disentangle the impact of genetics and microbiome. In our recent study assessing the genetic and microbial effect on BA metabolism, for instance, we used the ratio of secondary BA to primary BA to correct for the activity of human enzymes[66]. Second, even for microbiome-driven metabolites, their abundance levels in feces or blood are still the outcome of both the human genome and gut microbiome and depend on microbial activity, host absorption, transportation and elimination. For instance, 95% of short-chain fatty acids are absorbed and used as energy sources by colonocytes, while only 5% are secreted in feces. Genetic associations with these metabolites may point to not only genes involved in microbial activities but also those involved in absorption, usage, transportation and elimination of metabolites. Third, diet is an important confounding factor when we consider metabolism as a microbial readout. The production of trimethylamine *N*-oxide and short-chain fatty acids is largely dependent on meat and fiber intake, but correcting for diet is often difficult in mbGWAS analysis, as precise measurement of diet components is still unfeasible in large cohorts.

Non-bacterial members of the gut. In addition to the routes described above, that is, increasing study sample sizes, performing quantitative analysis of taxa and focusing on microbial genetics, additional insights could be gained by studying the largely unexplored non-bacterial communities of the gut ecosystem, including viruses, fungi and protozoa. Although the number of viruses in the gut is similar to the number of bacteria, the low coverage of virome sequencing and lack of universal markers for virome analysis make viruses challenging to study. Virome-specific isolation protocols can be applied, but they are time consuming. Moreover, as viruses are even more individual specific than bacteria, these analyses demand even larger sample sizes than those required for mbGWAS.

Conclusions

The genetics of the gut microbiome is still a field in its infancy and shares many similarities with early GWAS on complex human traits. We foresee that lessons learned from GWAS over the past 15 years, such as the need for larger sample sizes and the use of endophenotypes, will help us to move rapidly toward new discoveries. Data sharing and collaboration to combine GWAS in meta-analysis were major factors that contributed to advances in human genetic studies of complex traits. We therefore encourage researchers working in the field of microbiome genetics to embrace these scientific practices.

Even with very large sample sizes, it is conceivable that the overall effect of host genetics on the gut microbiome that can be explained by GWAS will remain moderate, in the range of 1–10% of variance. Nevertheless, we believe that identifying additional host genetic factors that influence the gut microbiome, even those with small effects, will provide important insights into complex host–microbiome interactions and could inform therapies and personalized treatments. For example, the effect of the functional variant in the *LCT* locus is rather small, ~0.8% of the variance in bifidobacterial species abundance, yet bifidobacterial species are associated with many conditions, including response to anti-cancer therapy[67,68]. Coupling genetic and microbiome screening could therefore help to improve personalized treatments and predict drug response. We expect that larger studies will provide sufficient power to investigate the role of rare variants and thus identify other functional variants, which may provide clues for unraveling drug-response heterogeneity or for developing drugs with the aim to modulate specific microorganisms.

Finally, we expect that application of systems genetics (multi-omic) approaches to both the human genome and the gut microbiome, including the integration of expression, proteomics, metabolomics and other -omic layers, will be a necessary future step for better understanding this complex multi-kingdom ecosystem.

Received: 21 June 2021; Accepted: 2 November 2021;
Published online: 03 February 2022

References

1. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
2. Kurilshikov, A. et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **53**, 156–165 (2021).
3. Lopera-Maya, E. A. et al. Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. *Nat. Genet.* https://doi.org/10.1038/s41588-021-00992-y (2022).
4. Gacesa, R. et al. The Dutch Microbiome Project defines factors that shape the healthy gut microbiome. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.27.401125 (2020).
5. Qin, Y. et al. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *Nat. Genet.* https://doi.org/10.1038/s41588-021-00991-z (2022).
6. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2019).
7. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
8. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
9. Goodrich, J. K. et al. Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
10. Goodrich, J. K. et al. Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**, 731–743 (2016).
11. Turpin, W. et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **48**, 1413–1417 (2016).
12. Hughes, D. A. et al. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat. Microbiol.* **5**, 1079–1087 (2020).
13. Xu, F. 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).
14. Liu, X. et al. A genome-wide association study for gut metagenome in Chinese adults illuminates complex diseases. *Cell Discov.* **7**, 9 (2021).
15. Blekhman, R. et al. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).
16. Davenport, E. R. et al. Genome-wide association studies of the human gut microbiota. *PLoS ONE* **10**, e0140301 (2015).
17. Pilia, G. et al. Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* **2**, e132 (2006).
18. Rühlemann, M. C. 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).

19. Bonder, M. J. et al. The effect of host genetics on the gut microbiome. *Nat. Genet.* **48**, 1407–1412 (2016).
20. 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).
21. Pang, X. et al. Mosquito C-type lectins maintain gut microbiome homeostasis. *Nat. Microbiol.* **1**, 16023 (2016).
22. Suzuki, T. A. & Ley, R. E. The role of the microbiota in human genetic adaptation. *Science* **370**, eaaz6827 (2020).
23. Hove, H., Nørgaard, H. & Mortensen, P. B. Lactic acid bacteria and the human gastrointestinal tract. *Eur. J. Clin. Nutr.* **53**, 339–350 (1999).
24. Yang, H. et al. An ancient deletion in the *ABO* gene affects the composition of the porcine microbiome by altering intestinal *N*-acetyl-galactosamine concentrations. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.16.206219 (2020).
25. 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).
26. Folseraas, T. et al. Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. *J. Hepatol.* **57**, 366–375 (2012).
27. Tong, M. et al. Reprograming of gut microbiome energy metabolism by the *FUT2* Crohn's disease risk polymorphism. *ISME J.* **8**, 2193–2206 (2014).
28. Burgueño-Bucio, E., Mier-Aguilar, C. A. & Soldevila, G. The multiple faces of CD5. *J. Leukoc. Biol.* **105**, 891–904 (2019).
29. Burton, P. R. et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
30. Scott, L. J. et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
31. Saxena, R. et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
32. Timpson, N. J. et al. Adiposity-related heterogeneity in patterns of type 2 diabetes susceptibility observed in genome-wide association data. *Diabetes* **58**, 505–510 (2009).
33. Klein, R. J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
34. Menzel, S. et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.* **39**, 1197–1199 (2007).
35. Uda, M. et al. Genome-wide association study shows *BCL11A* associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia. *Proc. Natl Acad. Sci. USA* **105**, 1620–1625 (2008).
36. Sinnott-Armstrong, N. et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).
37. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
38. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
39. Karlsson Linnér, R. et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* **51**, 245–257 (2019).
40. Zheng, T. et al. Genome-wide analysis of 944 133 individuals provides insights into the etiology of haemorrhoidal disease. *Gut* **70**, 1538–1549 (2021).
41. Perola, M. et al. Combined genome scans for body stature in 6,602 European twins: evidence for common Caucasian loci. *PLoS Genet.* **3**, e97 (2007).
42. Weedon, M. N. et al. A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nat. Genet.* **39**, 1245–1250 (2007).
43. Sanna, S. et al. Common variants in the *GDF5–UQCC* region are associated with variation in human height. *Nat. Genet.* **40**, 198–203 (2008).
44. Galarneau, G. et al. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nat. Genet.* **42**, 1049–1051 (2010).
45. Danjou, F. et al. Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nat. Genet.* **47**, 1264–1271 (2015).
46. Palmer, C. & Pe'er, I. Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet.* **13**, e1006916 (2017).
47. Lloréns-Rico, V., Vieira-Silva, S., Gonçalves, P. J., Falony, G. & Raes, J. Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. *Nat. Commun.* **12**, 3562 (2021).
48. Vandeputte, D. et al. Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
49. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
50. Schloissnig, S. et al. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45–50 (2013).
51. Xie, H. et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584 (2016).
52. Zeevi, D. et al. Structural variation in the gut microbiome associates with host health. *Nature* **568**, 43–48 (2019).
53. Costea, P. I. et al. metaSNV: a tool for metagenomic strain level analysis. *PLoS ONE* **12**, e0182392 (2017).
54. Olm, M. R. et al. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nat. Biotechnol.* **39**, 727–736 (2021).
55. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
56. Andreu-Sánchez, S. et al. A benchmark of genetic variant calling pipelines using metagenomic short-read sequencing. *Front. Genet.* **12**, 537 (2021).
57. Ramiro, R. S., Durão, P., Bank, C. & Gordo, I. Low mutational load and high mutation rate variation in gut commensal bacteria. *PLoS Biol.* **18**, e3000617 (2020).
58. Chen, L. et al. The long-term genetic stability and individual specificity of the human gut microbiome. *Cell* **184**, 2302–2315 (2021).
59. Steri, M. et al. Overexpression of the cytokine BAFF and autoimmunity risk. *N. Engl. J. Med.* **376**, 1615–1626 (2017).
60. Plomin, R., Haworth, C. M. A. & Davis, O. S. P. Common disorders are quantitative traits. *Nat. Rev. Genet.* **10**, 872–878 (2009).
61. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
62. Schirmer, M. et al. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat. Microbiol.* **3**, 337–346 (2018).
63. Long, S. et al. Metaproteomics characterizes human gut microbiome function in colorectal cancer. *npj Biofilms Microbiomes* **6**, 14 (2020).
64. Zierer, J. et al. The fecal metabolome as a functional readout of the gut microbiome. *Nat. Genet.* **50**, 790–795 (2018).
65. Bar, N. et al. A reference map of potential determinants for the human serum metabolome. *Nature* **588**, 135–140 (2020).
66. Chen, S. et al. Runx2^{+} niche cells maintain incisor mesenchymal tissue homeostasis through IGF signaling. *Cell Rep.* **32**, 108007 (2020).
67. Sivan, A. et al. Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti–PD-L1 efficacy. *Science* **350**, 1084–1089 (2015).
68. Matson, V. et al. The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104–108 (2018).

Acknowledgements

We thank K. McIntyre for help developing the manuscript. A.Z. is supported by European Research Council Starting grant 715772, Netherlands Organization for Scientific Research (NWO) VIDI grant 016.178.056, CVON grant 806 2018-27 and NWO Gravitation grant ExposomeNL 024.004.017. J.F. is supported by CVON grant 2018-27, European Research Council Consolidator grant 101001678 and NWO VICI grant VI.C.202.022.

Author contributions

S.S., A.K., A.v.d.G. and A.Z. performed data analyses; S.S., J.F. and A.Z. wrote the manuscript draft; A.v.d.G. and A.K. provided critical revisions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00983-z.

Correspondence should be addressed to Serena Sanna or Alexandra Zhernakova.

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