Host genetics and microbiome associations from the lens of genome wide association studies

Omer Weissbrod$^{1,2}$, Daphna Rothschild$^{1,2}$, Elad Barkan$^{1,2}$ and Eran Segal$^{1,2,*}$

$^1$Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel
$^2$Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 7610001, Israel
*Correspondence: eran.segal@weizmann.ac.il

Abstract
Recent studies indicate that the gut microbiome is partially heritable, motivating the need to investigate microbiome-host genome associations via microbial genome-wide association studies (mGWAS). Existing mGWAS demonstrate that microbiome-host genotypes associations are typically weak and are spread across multiple variants, similar to associations often observed in genome-wide association studies (GWAS) of complex traits. Here we reconsider mGWAS by viewing them through the lens of GWAS, and demonstrate that there are striking similarities between the challenges and pitfalls faced by the two study designs. We further advocate the mGWAS community to adopt three key lessons learned over the history of GWAS: (a) Adopting uniform data and reporting formats to facilitate replication and meta-analysis efforts; (b) enforcing stringent statistical criteria to reduce the number of false positive findings; and (c) considering the microbiome and the host genome as distinct entities, rather than studying different taxa and single nucleotide polymorphism (SNPs) separately. Finally, we anticipate that mGWAS sample sizes will have to increase by orders of magnitude to reproducibly associate the host genome with the gut microbiome.
Introduction

In recent years the importance of the gut microbiome in human metabolism and health is increasingly gaining recognition [1–12]. Recent studies have associated the microbiome with various health parameters including obesity, diabetes mellitus, cancer, and inflammatory, metabolic and neurodegenerative disorders [13–17].

A fundamental question is how strongly the microbiome is genetically inherited as opposed to being shaped by the environment. The microbiome evolves during childhood, and then becomes relatively stable and robust to perturbations [18–20]. This apparent host-adaptation evokes the classic question of ‘nature versus nurture’: Does the microbiome adapt to its host due to shared early environmental exposure, or are certain microbiome compositions inherently more suitable to specific host genomes?

The recent advent of 16S rRNA gene sequencing and metagenomic sequencing technologies enable carrying out gut microbiome studies with thousands of individuals [21]. Recent studies employing these technologies have uncovered evidence for both environmental and host genetic association with the microbiome composition [8,11,20,22–37]. However, to date there is no consensus regarding how and to what extent does host genetics shapes the gut microbiome, as compared to environmental factors.

In this article, we first review recent studies of environment and host genome associations with the human gut microbiome. We show that existing evidence suggests that the gut microbiome is predominantly shaped by environmental factors, and that host genotype-microbiome associations are weak, spread across multiple sites across the host genome, and together explain a relatively small fraction of the microbiome configuration of individuals. We then draw parallels between existing mGWAS and early GWAS, and use these to demonstrate how some of the pitfalls encountered in early GWAS, and their respective solutions, could be applied to mGWAS.

The microbiome is predominantly shaped by non-genetic factors

Recent studies have provided strong evidence that environmental factors play a much greater role than host genetics in shaping the gut microbiome. It can be difficult to tease apart environmental from genetic inheritance in humans, since children typically live with their parents. However, twin studies can tease these factors apart by comparing microbiome similarity among monozygotic (MZ) and dizygotic (DZ) twins, under the assumption that significant differences in the degree of similarity are attributed solely to genetic effects. Recent twin studies have demonstrated that the level of similarity is almost the same for MZ and DZ twins [24], that both MZ and DZ twins have extremely similar gut microbiomes compared with non-twin pairs [24], and that this similarity decreases when twins start living apart [32]. These results indicate that environment overshadows host genetics in shaping the gut microbiome.

Recent non-twin studies provide additional support for the dominant role of environment in shaping the gut microbiome. First, there is an excessive bacterial similarity among individuals sharing a household, but no such similarity was observed across family members without a history of household sharing [8,20,22,35]. Second, over 20% of gut microbiome β-diversity variance can be inferred via several measured environmental factors, such as answers to food frequency and drug use questionnaires [8,33,34], whereas no statistically significant result was obtained when applying a similar methodology to
genetic variants [8]. These results further demonstrate that the gut microbiome is predominantly shaped by environmental factors.

**Twin studies identify significantly heritable gut microbiome taxa**

Despite the strong role of non-genetic factors in shaping the gut microbiome, recent twin studies identified 33 significantly heritable bacterial taxa (most notably the family Christensenellaceae [36]). The estimated heritability of these taxa was typically 10%-30%, which is substantially lower than several well-known human complex traits, such as height, body mass index (BMI), and even education attainment [38]. A recent re-analysis of the largest reported twin study to date (2,252 twins) found that the average heritability of gut bacterial taxa likely lies between 1.9% and 8% [8]. Taken together, these results indicate that there are several genetically heritable bacterial taxa, but that the overall gut microbiome heritability is relatively small.

**Limited evidence for gut microbiome-host genotype associations from non-twins data**

A potential shortcoming of twin studies is the difficulty of assembling large cohorts. Genotyping of unrelated individuals with a relatively common environment facilitates the assembly of much larger cohorts. These cohorts enable directly associating the gut microbiome with the host genotype, by searching for a greater co-presence of bacterial taxa among genetically closer individuals. However, the results from such studies have been inconclusive and mostly failed to replicate.

One of the first studies to employ the above approach identified a significant correlation between the top microbiome principal coordinate (PCo) and top host-genome principal component (PC), based on human DNA residues extracted from stool samples [27]. An analysis of 127 Hutterites reported several heritable taxa [39], but the statistical significance of these results after multiple testing correction has not been reported. Additionally, several recent studies have identified a significant heritability of bacterial α-diversity [28,30,39]. In contrast, a recent analysis of 1,046 Israeli individuals from different ancestral origins but a relatively shared environment did not replicate any of the above results, and did not identify statistically significant host-genomics associations with either the overall microbiome composition or individual taxa [8]. Another recent study identified significant co-occurrence of bacterial taxa among 270 family members [28], and several other studies identified a significantly different microbiome composition between individuals from different populations [20,40], but it is not possible to tease apart the roles of genetics and environment in such studies. Overall, these inconclusive results again suggest that the heritable component of the gut microbiome is small.

**Limited power of microbiome genome wide association studies**

Microbiome association studies attempt to not only identify heritable taxa, but also to pinpoint the host genetic variants that underlie this heritability [11,36,37]. The first such studies in humans focused on specific genes and pathways, and have identified several significant microbiome-associated variants [41–45]. However, a potential shortcoming of the above studies is that they require previous knowledge of associated genes, and thus cannot discover new associations. Thus, recent studies have performed
unbiased microbiome-genome wide association studies (mGWAS) spanning 93-1,812 individuals [8,27–31].

A substantial difficulty of mGWAS is the high number of tested hypotheses, which is equal to the number of genetic variants multiplied by the number of tested taxa, genes and pathways. This leads to a severe multiple testing correction and to reduced power (Figure 1). Consequently, most mGWAS findings are not statistically significant after multiple testing correction. A recent analysis demonstrated that there is almost no overlap between the loci reported in different studies, even when allowing SNPs up to 1Mb apart and associations with different bacterial taxa to be considered as overlapping [8]. This lack of consistency could originate either from differences in the underlying analysis methods or from lack of reproducibility, necessitating further investigation of the reported associations. The only genetic variants consistently shown to be microbiome-associated in multiple mGWAS are located in close proximity to the LCT gene, which is associated with lactase persistence [8,27,30,31,46]. However, while important, this association may be confounded by lactose consumption [46].

Several recent studies alleviated the multiple testing burden by testing for association with the entire microbiome composition rather than individual taxa, and identified genetic variants located in the vitamin D receptor gene and in several genes associated with health disorders [8,29,30]. A recent study further argued that a small number of genetic variants can infer over 10% of the microbiome β-diversity composition [29]. However, the results found in one study could not be replicated in others, with the exception of LCT related variants [8].

Other than the LCT variants, the most consistently reported host-microbiome associations involve immunity-related variants, although no two studies reported an association with the same variant (see ref. [37] and references therein for a comprehensive review). It has also been observed that many mGWAS hits are found near genes associated with complex diseases [24,27–29,31,39,41,47–50], and that multiple studies have implicated variants residing in the same genes, though the exact loci differed between studies [11,36,41].

The above results demonstrate that certain bacterial taxa are clearly heritable, but that the variants underpinning this heritability have not been reliably identified. This contradiction suggests that the heritability of bacterial taxa arises due to the aggregated effects of multiple genetic variants, each having an individually weak effect that cannot be reliably identified with existing sample sizes. This property has long been recognized as being common to most complex human traits, and has been extensively studied in GWAS, as elaborated below.

**Contrasting microbiome and traditional genome-wide association studies**

It is beneficial to reflect on the current state of mGWAS by drawing parallels with the history of GWAS [51–53]. The key idea behind GWAS is to associate genetic variants with traits of interest using large cohorts of unrelated individuals. Since 2005, over 3,200 GWAS with unique PubMed IDs have been reported in the GWAS Catalog [54], compared with 6 published mGWAS [8,27–31]. The initial motivation for GWAS arose due to the observation that common traits, such as height or BMI, are associated with a large number of genetic variants with small effect sizes, thus requiring large cohorts to be identified.
reliably [51]. The small effect sizes reported in existing mGWAS suggests that the same pattern holds for host genome-microbiome associations.

The very first GWAS, which became possible thanks to the advent of low cost genotyping arrays, were met with high hopes. However, it soon became apparent that most reported associations failed to replicate [55]. The GWAS community consequently took actions to encourage reproducibility [56], chief among which was the adoption of stringent requirements for reporting associations. The same process seems to occur in current mGWAS, which also became possible due to the declining costs of the required technologies, and whose reported associations typically fail to replicate. Unfortunately, the number of hypotheses tested in a typical mGWAS is orders of magnitude larger than in a typical GWAS, suggesting that even more stringent statistical criteria need to be enforced.

A second important development in the history of GWAS was the adoption of common data formats, data processing techniques, analysis workflows and reporting guidelines [57]. These developments helped streamline, reduce the technical burden, and facilitate replication efforts in GWAS. Notably, the adoption of the common plink format [58] helped method developers to release software that could be used across different research groups in a unified manner. Future mGWAS would greatly benefit from such a standardization effort, as existing mGWAS were carried out using vastly different statistical methods, which hinders replication efforts. Some of the technical mGWAS aspects deserving additional investigation include the taxonomic levels of taxa to test, whether to also test for associations with the functional composition of the microbiome (e.g. bacterial genes), and statistical modeling of zero-inflation.

As more and more GWASs were being published, several common threads began to emerge. First, it became apparent that virtually all common traits are extremely polygenic, to the degree that most studied traits have to date been associated with dozens or hundreds of genetic variants with small effect sizes [54]. Second, most associated loci do not reside within coding regions, and there is often an excellent correlation between the number of associations on a chromosome and the chromosome length [59], suggesting that associated variants are spread uniformly throughout the genome. Third, pleiotropy was found to be extremely common, as a great number of loci were independently implicated with multiple traits. These observations suggested that the genetic architecture of common traits was far more complicated than initially thought, with some researchers hypothesizing that almost all genetic variants are associated with every trait [60]. As the gut microbiome is a highly complex organism [61], we believe that it is quite likely for the same patterns to emerge in mGWAS.

In response to the perceived complexity of common traits, GWAS gradually became larger and larger. While typical GWAS in 2007 spanned 3,000 individuals, many GWAS today are two orders of magnitude larger (Figure 2), with the recently released UK Biobank spanning approximately 500,000 individuals [62], and with plans underway to genotype 1,000,000 individuals for the Million Veteran Program [63]. These developments suggest that mGWAS sample sizes will similarly have to increase by at least two orders of magnitude to uncover the underlying biology behind gut microbiome and host genome interactions. In recent years many GWAS began releasing publicly available summary statistics that describe the association of each genetic variant to the studied trait, which enable combining results across multiple studies without the logistic and legal complications required to access private genetic data [64]. The mGWAS community would likely benefit from such data sharing ethics as well. Other potential approaches to increase power include oversampling of individuals with extreme microbiome-associated phenotypes, such as obesity [65,66], and restricting the analysis to taxa previously established to be heritable.
A second response of the GWAS community to the perceived complexity of common traits was phrasing of new research questions, which treat the genome in a global manner rather than a local one. Instead of trying to find genetic associations with tiny effects, many researchers began investigating the genetic architecture of common traits as a distinct entity [67,68]. This line of research arguably began with genotype-based heritability estimation [69], and has since opened the door to many new research questions, which have provided us with a new understanding of common traits. For example, recent developments enable us to study the degree of polygenicity of a given trait [70–72], the genetic similarity between pairs of traits [73–75], and the degree to which different traits are affected by various functional elements in the genome [76–80]. Such global approaches have arguably provided more insights into the underlying biology of common traits than direct genetic associations. We anticipate that similar global approaches could be carried over to mGWAS, as elaborated below.

**Global versus Local Approaches**

We encourage future mGWAS to adopt a holistic approach towards studying microbiome and host genome interactions. Specifically, we distinguish between local approaches, which consider the microbiome as a collection of taxa and the host genome as a collection of variants, and global approaches, which consider them as distinct entities (Table 1).

Global approaches are arguably more suitable for microbiome analyses because they can capture complex dynamics involving several taxa, in line with the view of the microbiome as a complex organism. In addition, global approaches can be substantially more powerful because they involve fewer tested hypotheses, leading to a less severe multiple testing correction, and because they aggregate multiple effects that may individually be too weak to be noticed. Finally, global approaches can jointly investigate the dynamics of the microbiome, the host genome, and additional factors, such as dietary habits.

Global approaches in statistical genetics are often carried out using linear mixed models [69,81–83], whereas in statistical ecology they are typically performed via ordination methods [84], Mantel tests [84] or multivariate analysis of variance [85]. To date, relatively little work has been done on combining the two frameworks together. One example is the recently proposed microbiome-association index, which measures the association of a host phenotype with the entire genomes of the microbiome and of the host in a single analysis [8]. We believe that this emerging field is a fertile ground for future developments.

**Concluding remarks**

The first mGWAS made many interesting discoveries, but have largely raised interesting questions rather than providing conclusive findings. It is our view that mGWAS would greatly benefit from adopting the lessons learned by the GWAS community over the several years. We specifically advocate adopting stringent statistical criteria, standard data formats, and a holistic approach towards studying microbiome and host genome interactions. Such approaches will require the development of new statistical methods, that will likely combine state of the art techniques from statistical genetics and statistical ecology. We anticipate that the combination of such approaches, along with larger sample sizes and with the integration of an increasing number of lifestyle and diet related factors, will lead to exciting new discoveries.
Acknowledgements

We thank the Segal group members for fruitful discussions. E.S. is supported by the Crown Human Genome Center; the Else Kroener Fresenius Foundation; Donald L. Schwarz, Sherman Oaks, CA; Jack N. Halpern, New York, NY; Leesa Steinberg, Canada; and grants funded by the European Research Council and the Israel Science Foundation. D.R. received a Levi Eshkol PhD Scholarship for Personalized Medicine by the Israeli Ministry of Science.
References


58. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al.: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007, **81**:559–75.


66. Sham PC, Purcell SM: Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* 2014, **15**:335–346.


Table 1: Overview of **global and local analyses**. Shown are common types of local and global analyses of the microbiome and of the host genome. Also shown is the approximate number of tests required for every type of analysis, under the assumption that there are one million host genetic variants and several hundred taxa.

**Microbiome**

<table>
<thead>
<tr>
<th>Host Genome</th>
<th>Local</th>
<th>Global</th>
</tr>
</thead>
</table>
| **Local**   | SNP-taxon association test: 
Associating a specific host SNP with a specific bacterial taxon. 
Requires hundreds of millions of tests |
| **Global**  | SNP-microbiome association tests: 
Associating a specific host SNP with microbiome β-diversity. 
Requires one million tests |
| **Local**   | Microbiome-host association tests 
Measuring the correspondence between host-genome similarity and microbiome β-diversity (e.g. via a Mantel test); 
Requires hundreds of tests |
| **Global**  | Microbiome-association index: 
Measuring the fraction of variance of a host phenotype that can be jointly inferred by the gut microbiome and host genome contents 
Requires a single test |
Figure 1: Plot of the fraction of variance a tested SNP needs to explain in order to be identified with 80% power, as a function of the number of tested hypotheses, for various sample sizes (based on standard derivations [52]). Increasing the number of hypotheses leads to reduced power to identify variants with small effect sizes, due to the severe multiple testing correction. Variants previously implicated in GWAS often explain less than 0.1% of the trait variance [52].

Figure 2: A bar chart depicting the largest GWAS performed up to every year between 2005-2017, as reported in the GWAS catalog [54]. GWAS sizes increased by almost 500-fold over 12 years.