Opportunities and limits of combining microbiome and genome data for complex trait prediction

Miguel Pérez-Enciso[1,2,4], Laura M. Zingaretti[2,4], Yuliaxis Ramayo-Caldas[3], Gustavo de los Campos[4]

1 ICREA, Passeig de Lluís Companys 23, 08010 Barcelona, Spain

2 Centre for Research in Agricultural Genomics (CRAG), CSIC-IRTA-UAB-UB, 08193 Bellaterra, Barcelona, Spain

3 Animal Breeding and Genetics Program, Institute for Research and Technology in Food and Agriculture (IRTA), Torre Marimon, 08140 Caldes de Montbui, Barcelona, Spain

4 Michigan State University, Dept. of Epidemiology & Biostatistics, and Dept. of Statistics & Probability, East Lansing, MI 48824, USA.

Short title:

Combining microbiome and genome data for complex trait prediction

Correspondence:

M. Pérez-Enciso

CRAG

Campus UAB

08193 Bellaterra, Spain

miguel.perez@uab.es

Abstract

The analysis and prediction of complex traits using microbiome data combined with host genomic information is a topic of utmost interest. However, numerous questions remain to be answered: How useful can the microbiome be for complex trait prediction? Are microbiability estimates reliable? Can the underlying biological links between the host's genome, microbiome, and the phenome be recovered? Here, we address these issues by (i) developing a novel simulation strategy that uses real microbiome and genotype data as input, and (ii) proposing a variance-component approach which, in the spirit of mediation analyses, quantifies the proportion of phenotypic variance explained by genome and microbiome, and dissects it into direct and indirect effects. The proposed simulation approach can mimic a genetic link between the microbiome and SNP data via a permutation procedure that retains the distributional properties of the data. Results suggest that microbiome data could significantly improve phenotype prediction accuracy, irrespective of whether some abundances are under direct genetic control by the host or not. Overall, random-effects linear methods appear robust for variance components estimation, despite the highly leptokurtic distribution of microbiota abundances. Nevertheless, we observed that accuracy depends in part on the number of microorganisms' taxa influencing the trait of interest. While we conclude that overall genome-microbiome-links can be characterized via variance components, we are less optimistic about the possibility of identifying the causative effects, i.e., individual SNPs affecting abundances; power at this level would require much larger sample sizes than the ones typically available for genome-microbiome-phenome data.

Author summary

The microbiome consists of the microorganisms that live in a particular environment, including those in our organism. There is consistent evidence that these communities play an important role in numerous traits of relevance, including disease susceptibility or feed efficiency. Moreover, it has been shown that the microbiome can be relatively stable throughout an individual's life and that is affected by the host genome. These reasons have prompted numerous studies to determine whether and how the microbiome can be used for prediction of complex phenotypes, either using microbiome alone or in combination with host's genome data. However, numerous questions remain to be answered such as the reliability of parameter estimates, or which is the underlying relationship between microbiome, genome, and phenotype. The few available empirical

studies do not provide a clear answer to these problems. Here we address these issues by developing a novel simulation strategy and we show that, although the microbiome can significantly help in prediction, it will be difficult to retrieve the actual biological basis of interactions between the microbiome and the trait.

Introduction

The relevance of microbial ecosystems associated with humans and animals in health and production is now widely recognized, e.g., [1–5]. To quantify its influence, the fraction of variance of a given trait explained by the microbiome has been named 'microbiability' (b^2) [6], in symmetry with the classical 'heritability' (h^2) concept [7]. Previously, the term "hologenome" had been coined to describe the joint action of genome and microbiome in explaining an observed phenotype [8].

A consequence of microbiability being typically larger than zero is that it can be used to predict complex phenotypes, be it a disease or productive traits. This is an important issue since the use of microbiome data has the potential to alter how medical diagnosis in humans or breeding decisions agricultural species are performed. Several studies have demonstrated the potential value of microbiome data for complex-trait prediction. For example, Rothschild et al. [9] showed that microbiome can be used to improve accuracy in the prediction of obesity and many other phenotypes in humans. Likewise, Lloyd-Price et al. showed that microbiome-data was predicted if future outbursts of bowel disease [10]. In cattle, various studies have shown the predictive power of microbiome for methane emission from rumen microbiome [4,11], feed efficiency and carcass traits in pigs [12,13], and various plant phenotypes (e.g., crop yield and diseases predicted from the microbiota data from the rhizosphere, [14]). On the other hand, since the groundbreaking study of Meuwissen, Hayes and Goddard [15], the prediction of complex traits using genome information has been embraced in both plant [16] and animal breeding [17] as well as in human genetics [18]. Therefore, a natural step further is combining host's genome and microbiome information to improve complex-trait prediction, a topic that is currently receiving much attention [12,19].

Importantly, microbiome composition can be affected by the host's genome. For instance, Wang et al. [20] argue that it is evolutionarily justified that the microbiome is under partial host genetic control since a non-negligible fraction of cells in an adult body is

made up of microbes, especially in the gut. Beginning with the seminal work by Pomp's team [21], several studies have confirmed the relationship between host's genotype and microbiome composition, e.g., [20,22,23]. These microbiome genome-wide association studies (mGWAS) suggest that microbiome abundances can be treated as any other complex trait in humans or livestock [22]. For instance, Crespo-Piazuelo et al. [24] or Ramayo-Caldas *et al.,* [25,26] identified several quantitative trait loci (QTL) that modulate gut bacterial and eukaryotic communities. In general, although the 'heritability' of each genera or OTU (Operational Taxonomic Unit) is typically weak, considering the whole microbiome simultaneously should increase power [27].

Large scale studies in humans suggest a predominant role of the environment in shaping the gut microbiome [9]. However, regardless of the relative importance of genetic and environmental factors in shaping the microbiota, microbiome composition *per se* can have predictive value. Yet, the use of microbiota for prediction of future phenotypes/disease outcomes, require some level of stability of the microbiome throughout time. In the case of the gastrointestinal tract, microbiota colonization starts at birth, where vertical transmission through the mother's birth canal occurs. Afterward, microbiota diversity and richness tend to increase as the host ages and reaches stability at adulthood [28,29]. In ruminants, populations inhabiting the rumen progressively appear after birth and partly persists throughout life [30].

As noted, the genome-microbiome-phenome is a complex system; understanding the links between host-genome, microbiota, and phenotypes is an important step towards the effective use of microbiome data for complex trait prediction. In all, despite published reports, we still lack detailed guidelines on the joint usage of microbiome and genome information for complex trait prediction, and on the reliability of parameter inferences. We are ignorant of the number of genes affecting microorganism abundance that can be confidently identified, or on how many microorganism taxa can influence a given phenotype. With this work, we aim to contribute to this important topic focusing on three inter-related questions:

1. How useful can the microbiome be for complex trait prediction?
2. Are microbiability estimates reliable?

3. Can the underlying biological genome-microbiome-links be inferred at a system-level? In a more refined level, can microbiome groups (e.g., OTUs, genera) with sizable causal effects on phenotypes be identified with the typical size of microbiome data sets?

In this study, we address the questions mentioned above via a novel simulation strategy that uses real microbiome and genotype data as input and proposing a variance-component approach which, in the spirit of mediation analyses, quantifies the proportion of phenotypic variance explained by genome and microbiome, and dissects it into direct and indirect effects. Importantly, the approach allows simulating a partial genetic control of host's genome on the microbiome. This is accomplished using a partial permutation approach that preserves the distribution of the genome and microbiome. We use Bayesian variable selection models to estimate parameters which contemplate the possibility that some or all the features available in the genome and/or the microbiome, have no effects on the trait of interests. We investigate the questions presented above across diverse scenarios regarding the links between host genomes and microbiomes, and of their relations with a complex trait.

Results and Discussion

The exact nature of the links between genome (**G**), microbiome (**B**), and phenotype (**y**) are largely unknown and will likely vary from case to case. However, we will use the six generic causal models ('scenarios') depicted in Fig 1 to shed light on the nature of the genome-microbiome-phenome links. In the 'Null' scenario, there is no link between any of the data-layers; while this is unlikely, it serves as an 'overall null hypothesis' and it is useful to assess potential biases in parameter estimates. Model 'Genome' assumes that **G** only affects the phenotype. In turn, only **B** has a direct effect on phenotype in 'Microbiome' and 'Indirect' scenarios. The Indirect scenario, however, allows for some of the causative abundances to be controlled genetically. This would be similar to a scenario where a phenotype is directly controlled by gene expression levels and expression in turn is controlled genetically [31,32]. The 'Joint' scenario is the simplest configuration for a trait under the influence of both genes and microbiome. It assumes microbiome and genome are independent and that their effects on the phenotype are also independent. The Joint model is the most widely assumed, implicitly, or explicitly, in the

literature, e.g., [4,9,12]. The 'Recursive' model is similar to the Joint model; however, the Recursive model contemplates the possibility that some causative OTU may be under partial genetic control by the host. Therefore, in this case, the genome has both direct and indirect (microbiome-mediated) effects on phenotypes. Note the Recursive model does not assume that the same loci have simultaneously direct and indirect effects, neither it assumes that all OTU abundances are under genetic control.

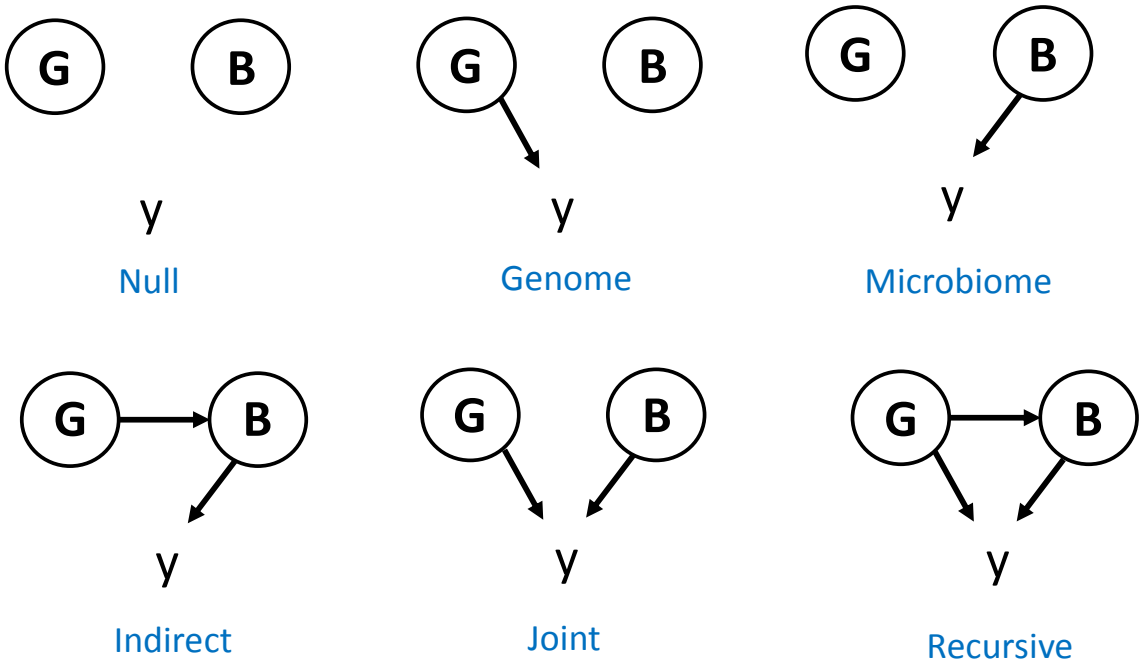


Fig 1. *Representation of the scenarios evaluated:* ***G****, genome, typically comprises marker data;* ***B****, microbiome;* ***y****, phenotype of interest; arrows indicate causality. An arrow from* ***G*** *to* ***y*** *indicates that there is a subset of* ***G*** *elements (causative SNPs) that influence* ***y****; an arrow from* ***G*** *to* ***B*** *indicates there exists a subset of* ***G*** *that influences a subset of abundances in* ***B*** *which, in turn, may also influence* ***y****. An arrow departing from* ***B*** *indicates there is a subset of microbial abundances (the causative abundances) that influence* ***y****. The SNPs affecting* ***B*** *need not necessarily be the same SNPs affecting* ***y*** *directly in the Recursive scenario. Note* ***B*** *can contain one or more sets of abundances such as archaea and bacteria communities, or different time or site sampling points. Without loss in generality, we assume* ***B*** *is a single community.*

We use the causal models depicted in Fig 1 to simulate genome-microbiome-phenotype data using different configurations regarding the number of causative loci (QTN) and the number of OTUs with effects on phenotypes, as well as the number of OTUs that were

187 affected by host's genome. Table 1 summarizes the simulation models and parameter
188 values.
189
190 **Table 1**. Definition of scenarios evaluated and parameters chosen: **G**, genome; **B**,
191 microbiome; **y**, phenotype of interest; N_{QTN}, number of SNPs with a direct causal effect
192 on **y**; N_{OTU}, number of OTUs with a direct effect on **y**; $N_{OTU(g)}$, number of OTUs with a
193 direct effect on **y** that are genetically determined, i.e., they are a subset of N_{OTU}; h^2 is
194 heritability, b^2 is microbiability, and $r^2 = h^2 + b^2$.
195

Scenario	**Abbreviation**	N_{QTN}	N_{OTU}	$N_{OTU(g)}$	r^2	h^2	b^2
Null	0	-	-	-	0	0	0
Joint	J	100	25	0	0.25	0.125	0.125
					0.50	0.25	0.25
Genome	G	100	0	0	0.25	0.25	0.00
					0.50	0.50	0.00
Microbiome	M	0	25	0	0.25	0.00	0.25
					0.50	0.00	0.25
Recursive	R	100	25	25	0.25	0.125	0.125
					0.50	0.25	0.25
Indirect	I	0	25	25	0.25	0.00	0.25
					0.50	0.00	0.50

196
197 **Table 2**. Scenarios used to evaluate sensitivity to the number of causative OTUs. Symbols
198 as in Table 1.
199

Scenario	**Abbreviation**	N_{QTN}	N_{OTU}	$N_{OTU(g)}$	r^2	h^2	b^2
Joint	J10	100	10	0	0.50	0.25	0.25
	J100	100	100	0	0.50	0.25	0.25
	J250	100	250	0	0.50	0.25	0.25
Recursive	R10	100	10	5	0.50	0.25	0.25
	R100	100	100	50	0.50	0.25	0.25
	R250	100	250	125	0.50	0.25	0.25

200

A novel data-driven strategy to generate microbiome-genome-phenotype experiments

Two facts make it the simulation of scenarios in Fig 1 challenging: (i) microbiome data follow zero-inflated highly leptokurtic multivariate distributions [33,34], it is not obvious how to sample from these distributions *conditionally* on genome data as required in the Recursive and Indirect scenarios; and (ii) it is difficult to obtain accurate estimates of key parameters, such as microbiability values, in the absence of large scale published – and public – datasets. To circumvent, or at least to alleviate, these constraints we use real data for both **G** and **B**. Specifically, we used publicly available data from two of the largest microbiome studies in livestock, genome data were downloaded from [11] and OTU abundances from [4].

Fig 2 recapitulates the simulation strategy. Full details are given in Material and Methods section, and R code to replicate the analyses are in https://github.com/miguelperezenciso/simubiome). We assume the effects of the causative microbiome abundances are additive on the log scale. Simulation under the Joint scenario is straightforward, since **G** and **B** act independently: sample a list of causative SNPs and abundances, simulate their effects, and apply Eqn. 1 (Material and Methods) to generate phenotype values given observed genotypes and abundances. The case of Recursive and Indirect scenarios is not that obvious because causative abundances are under genetic control and a link must exist between **G** and **B** (Eqn. 2 in Material and Methods). We solved this issue by rearranging abundances within individuals such that the desired correlation between abundance and individual's genotypes is attained (see Algorithm in Box 1 and R-code in https://github.com/miguelperezenciso/Simubiome/blob/master/sortCor.R). This strategy has the important advantage that the distribution of abundances is not changed.

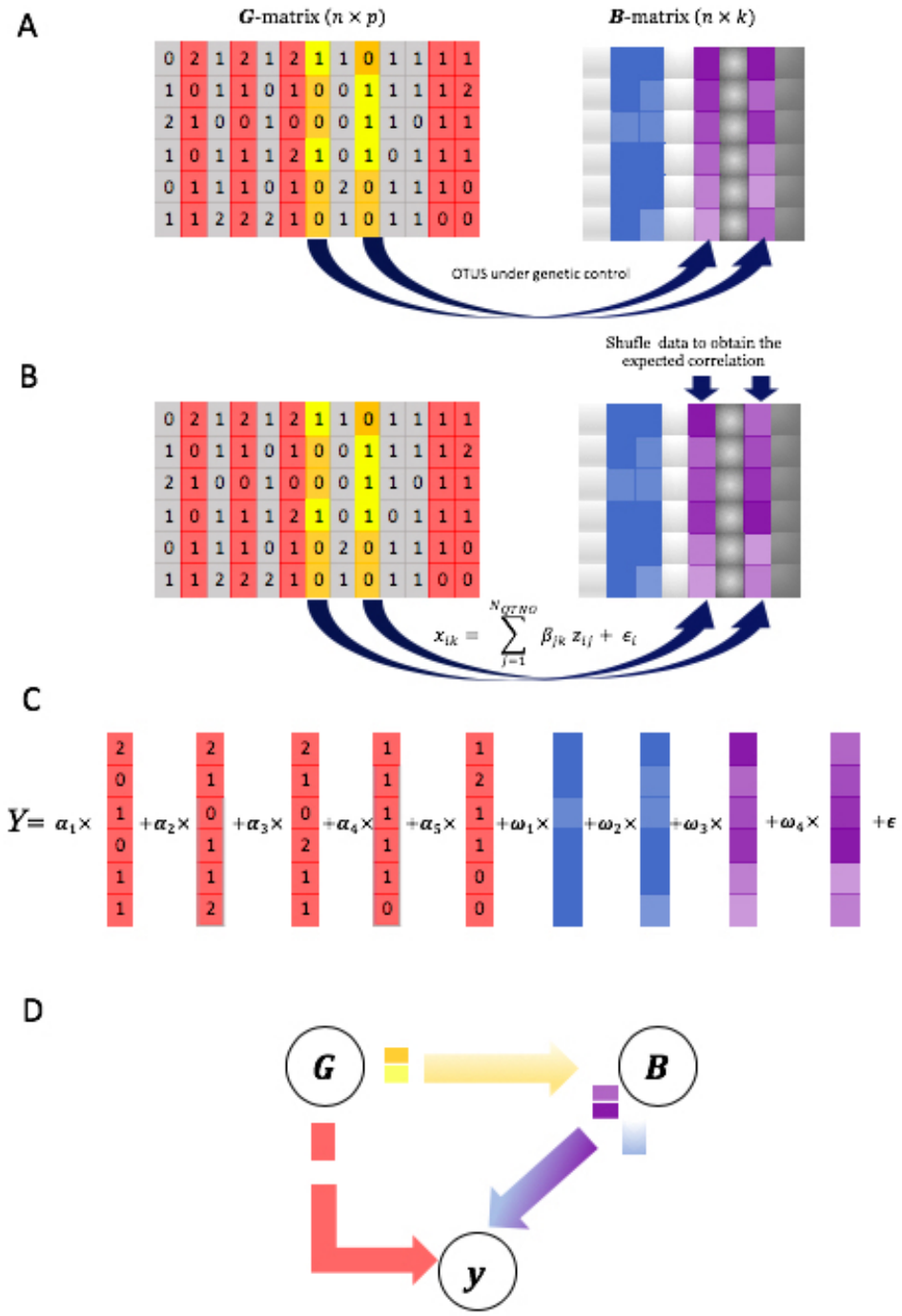


Fig 2: *Simulation scheme for the Recursive scenario, i.e., the most complex scenario (Fig 1). **A)** Real input data comprises p genotypes (**G** matrix) and k taxa abundances (**B** matrix). SNPs in grey are neutral, those in red act directly on the phenotype **y**, and those in yellow/orange influence some OTU abundances (marked in magenta color in **B** matrix); abundances in blue are not genetically controlled. **B)** Given simulated effects, a genotypic value controlling the abundances is obtained via Eqn. 2. To fulfill the required heritability, abundances in magenta are reordered; high abundances (represented by a darker color) are associated with genotype '1' just to simplify visualization. A single SNP is shown as causative for each of the two OTUs but there is no limit in practice. **C)** The phenotype is simulated by adding the genome and the microbiome contributions plus a residual. **D)** The general causal diagram is shown.*

How useful can microbiome be for complex trait prediction?

This depends on how much phenotypic variance is jointly explained by the genome (h^2) and the microbiome (b^2), but also on how efficiently methods capture the relationship between the microbiome and the phenotype, and on how stable the microbiome is. Note prediction accuracy is conditionally independent of whether the microbiome itself is heritable or not. This means that, *given* observed abundances **B** and observed genotypes **G**, it does not matter whether the biological process generating **B** is affected by **G**. In other words, prediction should not be affected by whether the Joint or Recursive scenarios hold, for a constant $r^2 = h^2 + b^2$. The implications for breeding, however, could be dramatically different. Breeding schemes targeting the microbiome could be designed provided the Recursive scenario holds but make no sense under the Joint scenario.

We compared predictive performance of Bayes C [15] when both genome and microbiome are employed in the model (*Bayes Cgb*) only genome (*Bayes Cg*), or only microbiome data (*Bayes Cb*). First, we verified the null model resulted in no false predictive accuracies (Fig S1A). Fig 3 shows simulated predictive accuracies for the two r^2 values considered (0.25 and 0.50) and for each causative scenario (Fig 1). Predictive accuracies using *Bayes Cgb* were consistently the best. As expected, this was especially the case when both h^2 and b^2 are larger than zero, that is, when Joint or Recursive scenario hold. In these scenarios, using both sources of variation clearly improved prediction compared to using only genome (*Bayes Cg*) or microbiome data (*Bayes Cb*). Importantly, predictive accuracy was somewhat lower in Joint and Recursive scenarios than in Microbiome or Genome scenarios. This indicates that predictive accuracy does not depend only on total r^2, but also on how this variance is split between genome and microbiome. Although this likely occurs because of the larger noise in Recursive or Joint scenarios than in Microbiome or Genome scenarios, it also suggests that our analysis strategy may not be optimum. There is room to develop more efficient tools, especially when the Recursive scenario holds. Note that variance of prediction was larger in the Recursive than in the Joint scenario, i.e., the fact that some abundances are inherited is an additional source of noise.

It is noticeable that predictions were better when only the microbiome influenced the phenotype than when the genome was the only source of variation, a phenomenon also

observed with real data [11,12,19]. In this simulation, this occurs likely because the number of causative effects and of input variables (SNPs vs. OTUS) is smaller in the Microbiome or Indirect scenarios than in the Genome scenario. In fact, we do observe a consistent negative correlation between the number of causative OTUs and predictive accuracy in both Joint and Recursive scenarios (Fig 4A).

In all, our results suggest that predictive accuracy could be increased by ~ 50 % when considering microbiome data, provided microbiability is of the same order as heritability (Fig 3). We speculate that this is probably an upper limit, since it will be difficult to have microbiome data collected homogeneously across time and in different locations. While individuals can be genotyped at birth, the microbiome in early life is not representative of adult or later stages. Maltecca et al., for instance, show that early life microbiota is not a good proxy for carcass composition in pigs [35].

We observed, roughly, a two-fold increase in predictive accuracy when doubling heritability for Genome, Joint and Recursive scenarios, and a 50% increase for Microbiome or Indirect scenarios (Fig 3A vs. 3B).

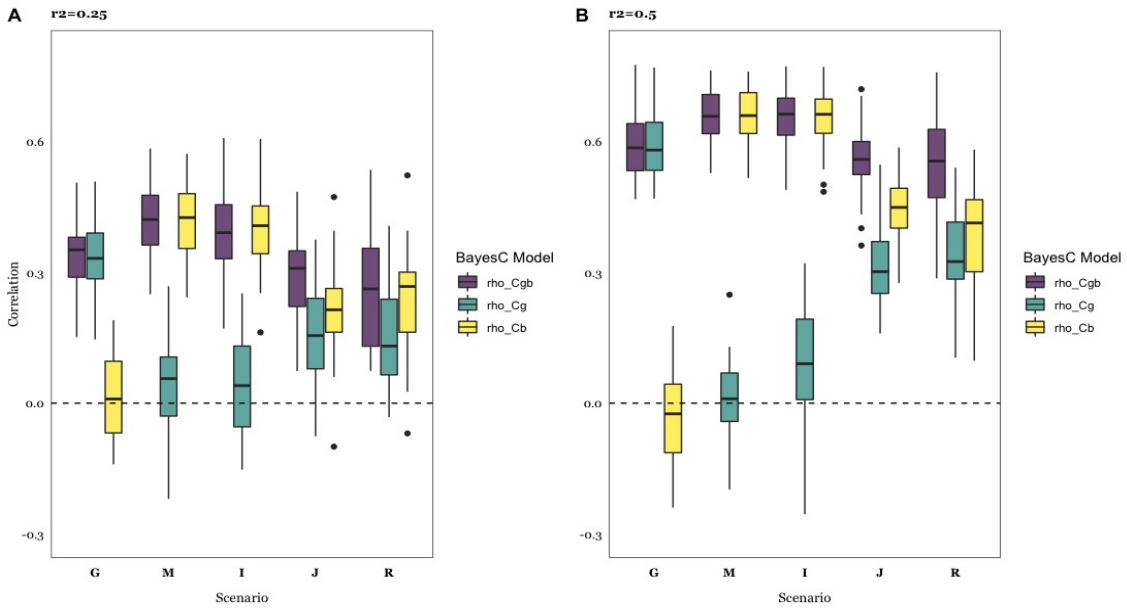


Fig 3. *Predictive accuracy, computed as correlation between predicted and observed phenotypes across causal scenarios (Fig 1), for each of the Bayes C models: Cgb considers microbiome and genome; Cg includes genome data only, and Cb includes microbiome data only.* ***A****:* $r^2 = 0.25$*;* ***B****:* $r^2 = 0.50$*. Details of scenarios are in Table 1: G, Genome; M, Microbiome; I, Indirect; J, Joint; R, Recursive. Results are average of 30 replicates per case.*

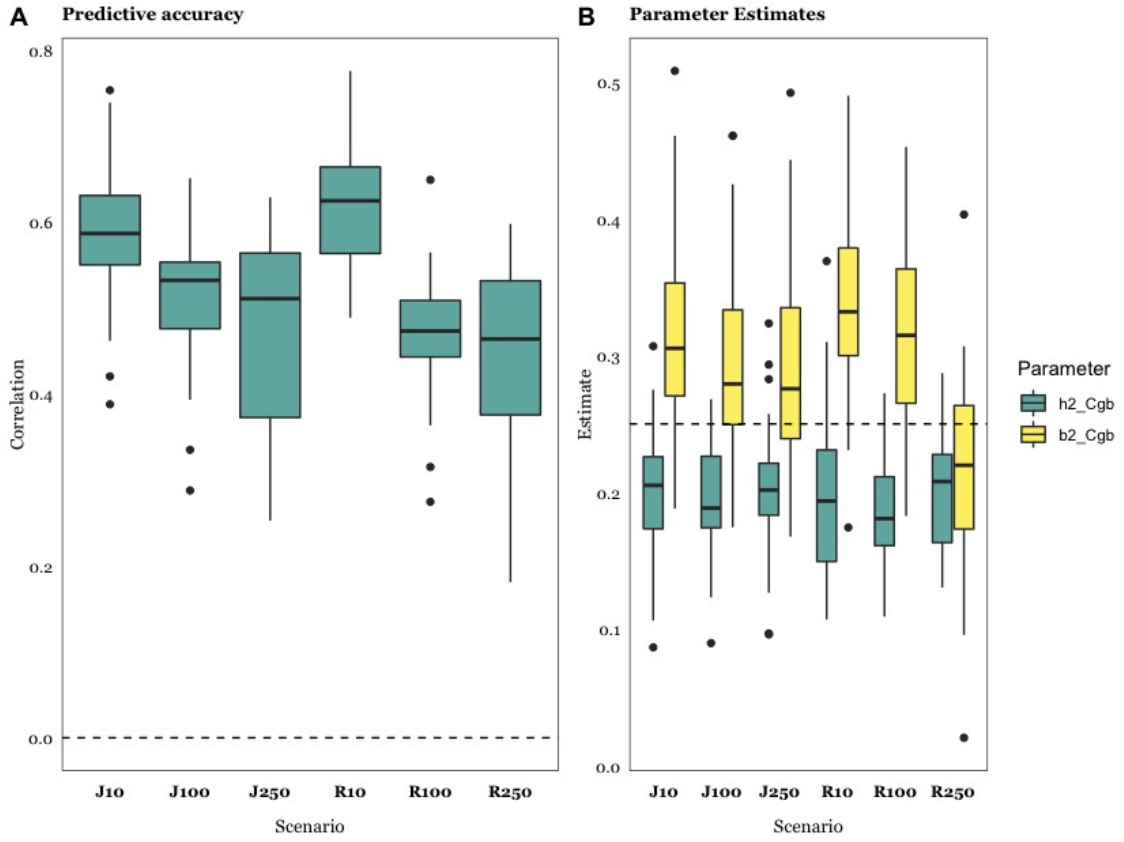


Fig 4. *Effect of varying number of causative OTUs, r^2 = 0.5.* ***A****: Predictive accuracy, computed as correlation between predicted and observed phenotypes, using Bayes Cgb.* ***B:*** *Heritability and microbiability estimates using Bayes Cgb. Details of scenarios are in Table 2 and diagrams in Fig 1: Jx, Joint scenario; Rx, Recursive scenario, with x being the number of causative OTUs (x = 10, 100, 250), half of them under genetic control. Results shown are the average of 30 replicates.*

Are microbiability estimates reliable?

Reliable parameter estimates are needed to optimize the design of breeding schemes or microbiome wide association studies (MWAS) [36]. They are also needed for understanding the biology behind the interaction of microbiome and complex phenotypes. Thus far, microbiability has been usually estimated using 'standard' linear methods, e.g., [4,9,27], much as we have done here. It is of interest then to know how accurate these estimates could be.

Fig 4 shows estimates of variance components for each of the scenarios in Table 1. *Bayes Cgb* allows us to assess whether h^2 and/or b^2 are different from zero: microbiability estimate is near zero when the data are simulated according to the Genome scenario and heritability is zero when the Indirect or Microbiome scenarios hold, as it should.

Similarly, both h^2 and b^2 estimates are near zero when the null scenario holds (Fig S1B). An overestimation of b^2 is nevertheless evident in Fig 4, and it does not vanish at higher r^2. This upward bias in b^2 estimate is accompanied by an underestimation of h^2, indicating that variance estimates are confounded when using *Bayes Cgb* model. This bias decreases though when the number of causative OTUs increases. For instance, the bias in b^2 estimate is ~ 40% when $N_{OTU} = 10$ but is reduced to ~10% with $N_{OTU} = 250$ (Fig 4B). Therefore, it is likely that the presence of a few causative OTUs, but of large effect, combined with the presence of highly leptokurtic abundance distributions, may result in biased parameter estimates. This should be considered when interpreting microbiability estimates in real experiments. For instance, Difford et al. [4] report estimates $h^2 = 0.21$ and $b^2 = 0.13$ (N = 750), finding **G** and **B** to behave independently. Assuming the number of causative OTUs is small compared to that of SNPs with an effect on abundances (QTNs), we can presume Difford's estimate of b^2 to be inflated. This means that the actual microbiome contribution may be too small to improve prediction over that obtained from using marker data exclusively. Although authors focused on inference and not so much in prediction, Difford et al reported that no bacteria genera were significantly associated with methane emissions [4]. Other authors in turn have reported polymicrobial associations, including members of bacterial, archaeal, fungal, and protozoan communities, with methane emissions, e.g., [11,25,37–39].

For comparison, Fig 5B,D show the estimates obtained with *Bayes Cg*, when only h^2 is estimated, or *Bayes Cb,* only b^2 is estimated. The most noticeable aspect is that bias in b^2 estimates is somewhat reduced relative to that found with *Bayes Cgb*, signaling again some confounding between b^2 and h^2. Bias was reduced overall at higher r^2 but did not vanish.

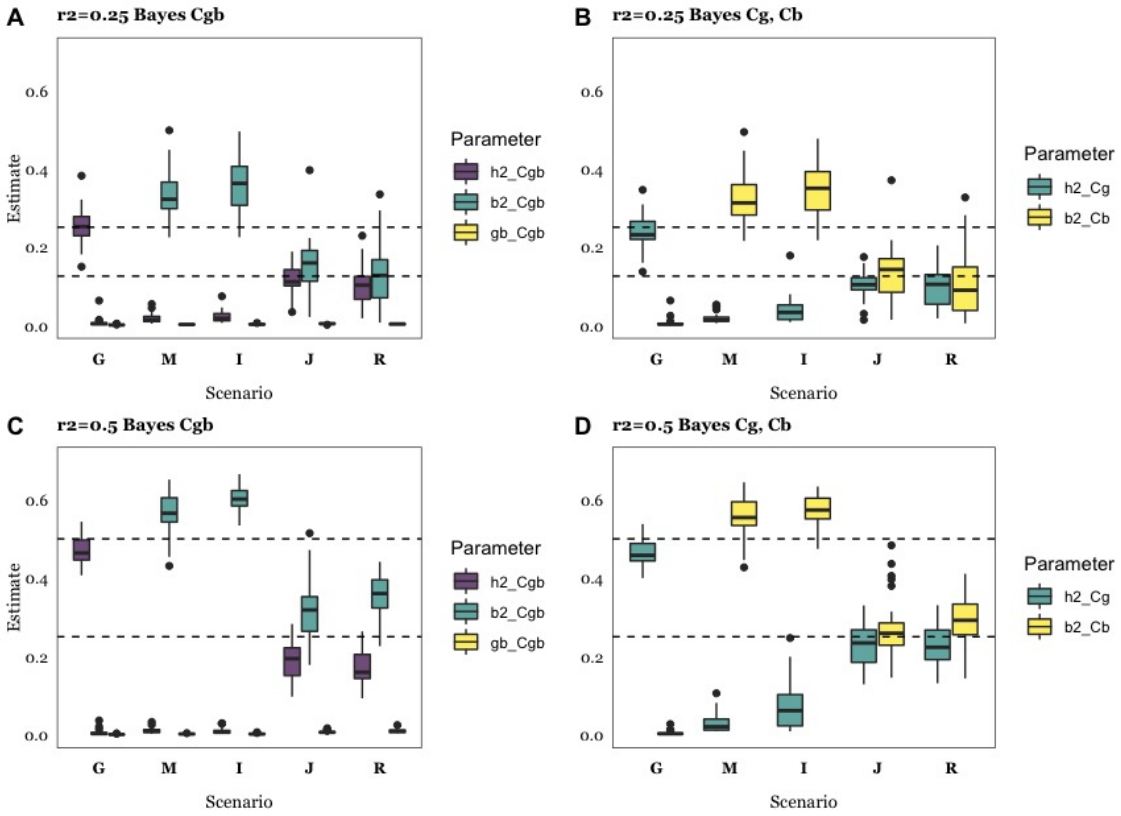


Fig 5: *Estimates of heritability (h2), microbiability (b2), and correlation between genome and microbiome (gb) for each of the three Bayes C analysis models: Cgb includes microbiome and genome in the model (left panels); Cg includes genome only, and Cb includes microbiome data only (right panels). Upper rows correspond to r^2 = 0.25 and lower rows to r^2 = 0.50. Details of simulation scenarios are in Table 1: G, Genome; M, Microbiome; I, Indirect; J, Joint; R, Recursive. Horizontal dashed lines indicate true h^2 or b^2 parameter values (0.125, 0.25, 0.5 depending on the scenario and on r^2). Results are average of 30 replicates.* ***A****: r^2=0.25, Bayes Cgb estimates (h^2, b^2 and gb);* ***B****: r^2 = 0.25, Bayes Cg (h^2) and Cb (b^2) estimates;* ***C****: r^2 = 0.50, Bayes Cgb estimates;* ***D****: r^2 = 0.50, Bayes Cg and Cb estimates. Data are average of 30 replicates per case.*

Can the underlying biological scenario be recovered? Can causative OTUs be identified?

An important goal of many experiments is to dissect the biological basis of microbiome and genome interactions, even if this is not strictly needed for prediction. So far, our simulations suggest that standard statistical methods can be used to quantify – with some bias – microbiability contribution to phenotypic variance. It also seems feasible to distinguish whether the Microbiome or Genome scenario fits real data best. Similarly, it seems plausible to assess when **G** and **B** contribute to the phenotypic variance, i.e. when Recursive or Joint scenarios are plausible.

Could Joint vs. Recursive scenarios be distinguished? Can data point to which of Indirect or Microbiome scenarios is more plausible, if any? Further, can causative OTUs be identified? These are far more difficult questions to answer than assessing prediction performance or estimating microbiability. Compare variance component estimates obtained under the Joint or Recursive scenarios (Fig 5): they are nearly identical for the same r^2. The two scenarios differ in that at least some causative OTUs abundances can be under partial genetic control in the Recursive scenario. The Recursive scenario should result in a covariance between **G** and **B**. We conjectured that the two scenarios could be distinguished by analyzing the covariance $Cov(\boldsymbol{u}^{(i)}, \boldsymbol{v}^{(i)})$ / $Var(\boldsymbol{y})$ (see methods). Unfortunately, these estimates are close to zero irrespective of the true scenario (Fig 5A, C). The likely reason is that the actual fraction of phenotypic variance explained by indirect effects is *conditionally* negligible. Note there can be a genetic effect of **G** on **B** but, for our purposes, we are interested only in those genes that affect causative OTUs (i.e., those that affect the phenotype) and not on the whole microbial system.

An alternative approach to infer whether the Recursive causative scenario holds or not is to run a genome-wide association study (GWAS) for each of the OTU abundances on each SNP, where the SNP P-values can indicate a genetic basis for some of the abundances. If we identify significant SNPs for OTUs likely influencing **y**, we could conclude that the Recursive scenario is plausible. Unfortunately, this analysis can be doomed by the large number of tests to be realized, i.e., N_{OTU} x N_{SNP}. To illustrate the caveats of GWAS on abundances, Fig 6A shows the distribution of -log10 P-values of neutral SNPs vs. SNPs with an effect on abundances. Assume we take the 5% empirical threshold of the neutral P-value distribution as indicative of association. Simulations suggest that only ~3% of causative SNP P-values will be above that threshold, i.e., approximately what is expected by chance. These P-values depend of course on the actual number of causative SNPs and on abundance heritabilities, but most evidence so far points to a weak relationship between genome and microbiome [22]. We warn it is going to be very difficult to identify abundance causative SNPs using GWAS information alone [9,20].

Another question of interest is how many of the OTUs affecting the phenotype can we expect to discover. One option is to count the frequency of a given OTU entering into the Bayes C model during sampling. Fig 6B shows the probability of including a causative

OTU in the Bayes C sampling chain, which varied between ~5% (b^2 = 0.125) to ~20% (b^2 = 0.25). About 50% (b^2 = 0.25) or 30% (b^2 = 0.125) of causative OTUs were among the 5% most frequently included OTUs in the Bayes C chain, on average. Since the number of causative OTUs was 25, the rate of false positives was high nevertheless. We can conjecture that only a few causative OTUs are likely to be identified in medium-sized experiments, such as this one.

An alternative approach is a Microbiome Wide Association Study (MWAS), i.e., to perform a linear regression of the phenotype on each of OTU abundances and then select the significant results as potential causative OTUs [4]. Fig 6C shows the average power, defined as the percentage of true causative OTUs within the 5% most significant results. Power was ~15% and ~20% for b^2 = 0.125 and 0.25, respectively, in the Recursive scenario. Again, this is not too satisfactory, as we expect a high fraction of false positives. In this particular scenario, it is perhaps more useful to consider probabilities of inclusion in the Bayes C chain rather than at P-values since the former are the result of a joint analysis of all OTUs and can be used directly for prediction.

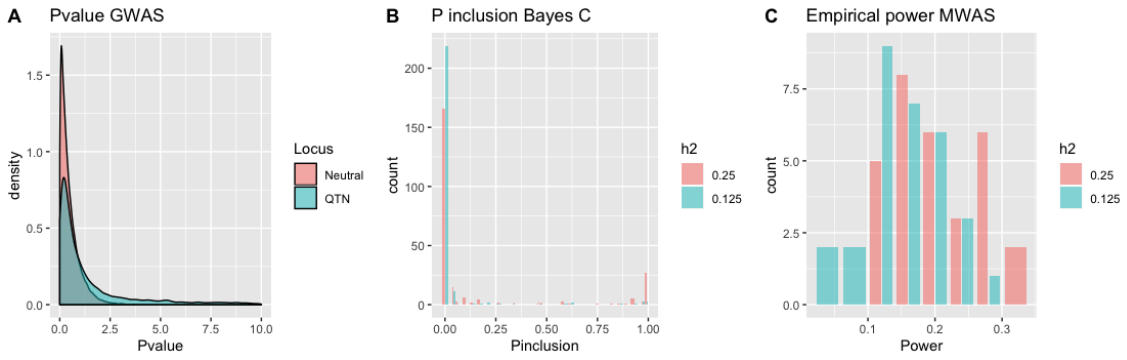


Fig 6: ***A)*** *Distribution of -log10 P-values of a GWAS of abundances on SNP data;* ***B)*** *Probability of inclusion in the Bayes Cgb model of causative OTUs for the two levels of microbiability considered.* ***C)*** *Power of identifying a causative OTU computed as the probability of exceeding the 95% threshold of the empirical distribution of P-values in an MWAS for the Recursive scenario.*

Finally, we investigated the pattern of abundance heritabilities. Fig 7A shows the simulated heritabilities for the causative, inherited OTUs, which approximately follows a gamma distribution, together with estimated heritabilities for the causative OTUs in the Recursive scenario. We observe that both distributions are rather similar although estimates are somewhat shrunk towards zero, a consequence of using a REML-like prior.

A problem of course is that we do not know which OTUs are inherited and which are not, and the true distribution of OTU heritability estimates will be a mixture. Fig 7B illustrates the heritability distributions of neutral (non-inherited) and causative (inherited) OTUs. In Fig 7B, we mixed 1.7 neutral OTU per causative OTU. This is completely arbitrary since we do not know the actual number of OTUs under genetic control, but we did so because the resulting mixture is similar to the distribution of heritabilities observed by Difford et al. (Fig 7C). If distributions in Fig 7B were representative of the true state of nature, this would suggest that about 1/(1+1.7) ~ 40% rumen OTUs could show some genetic additive variance in the experiment reported by Difford et al.[4].

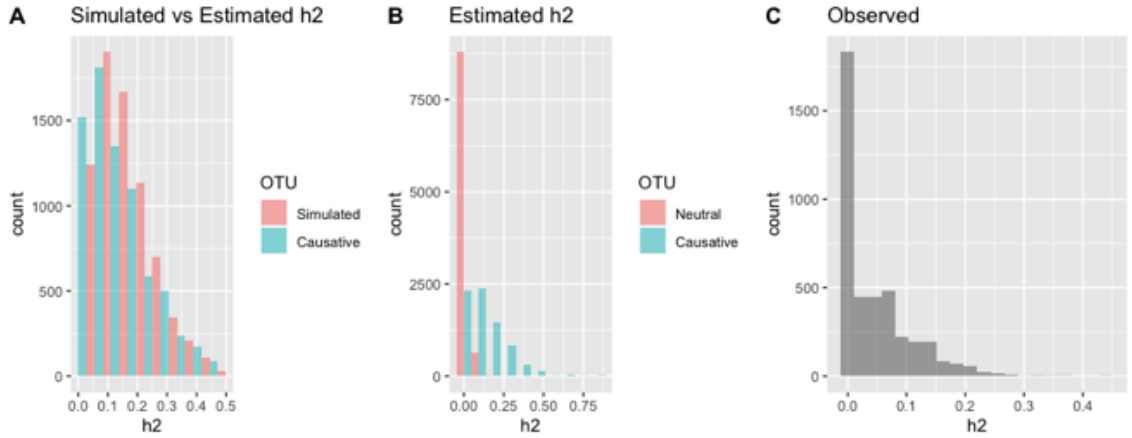


Fig 7: A: *'True' (simulated) and GBLUP estimated distribution of abundance heritabilities for causative OTUs in the Recursive scenario.* ***B:*** *GBLUP estimated distribution of abundance heritabilities for neutral and causative OTUs in the Recursive scenario.* ***C:*** *Actual distribution of OTU abundance heritabilities reported by Difford et al.*[4].

Discussion

Fig 1 represents but highly simplified relationships between the genome, microbiome, and phenotype. These scenarios are nevertheless important to interpret empirical data and can help to identify limiting factors in prediction. Further, provided a good fit is found, they will help in designing experiments that combine microbiome and genetic data. We chose parameter combinations that represent extreme case scenarios and we found that results were, qualitatively, robust to parameter choice such as r^2. A parameter that can be relevant though is the number of causative microbiome taxa, i.e., those with an effect on the phenotype. This number seems to affect the bias of microbiability estimates (Fig 4).

Here, we have proposed a new simulation procedure that addresses some important challenges. First, the algorithm avoids the need for actual phenotype simulation by using real genotype and abundance data. Although we concede that this procedure may limit the generality of the study, e.g., in terms of data size, we believe the advantages of using real data are numerous, since no simulation procedure can accommodate all known and unknown subtleties of the highly dimensional distributions at hand. Second, we develop an ingenious permutation procedure (Box 1) that allows linking previously uncorrelated data to fit a desired genetic hypothesis. By also permuting all OTUs within a given cluster, we minimize disruption of the whole covariance structure (Fig S2).

Numerous studies have reported microbiability values for economically important traits e.g. [4,12,25,39], but their actual reliability is not known. Estimates may be affected by the estimation procedure. There are numerous alternatives to estimate b^2, among them Bayes C [15], GBLUP [40], Bayesian RKHS regression using either Bray–Curtis dissimilarities as relationship matrix [25] or with the variance-covariance from the log-transformed OTUs as kinship matrix[25,41]. Our results (Fig 5) indicate that BayesC estimates may be biased upwards, especially when b^2 is higher than 0.25 and the number of causative OTUs is small. However, we found that estimates of b^2 derived with Bayes C were very close to zero in the null scenario (Fig S1B); therefore, we conclude that models using priors from the Spike-Slab family, which contemplate a priori the possibility of null effects, can be used to test whether heritability or microbiability is substantial. Ramayo-Caldas et al. [25] report that estimates using Bray-Curtis based kernels are higher than those using the log-transformed covariance matrix. The behavior of estimation methods for microbiability merits further research.

One conclusion from this work is that it is going to be difficult to distinguish between some underlying scenarios or to identify the causative OTUs and SNPs, at least using standard linear models as was done here. The distinction between Joint and Recursive scenarios is of special relevance for breeding. The latter assumes partial genetic control of some causative OTUs. Yet, we found both scenarios result in very similar patterns (Figs 3, 4, 5). Perhaps, a more powerful approach would be to use structural equation models (SEM), which allow including a variable both as independent and dependent. Saborio-Montero et al. [42] compared a linear bivariate (one OTU and the phenotype) model with a SEM but found few differences. One restriction of their approach is that one

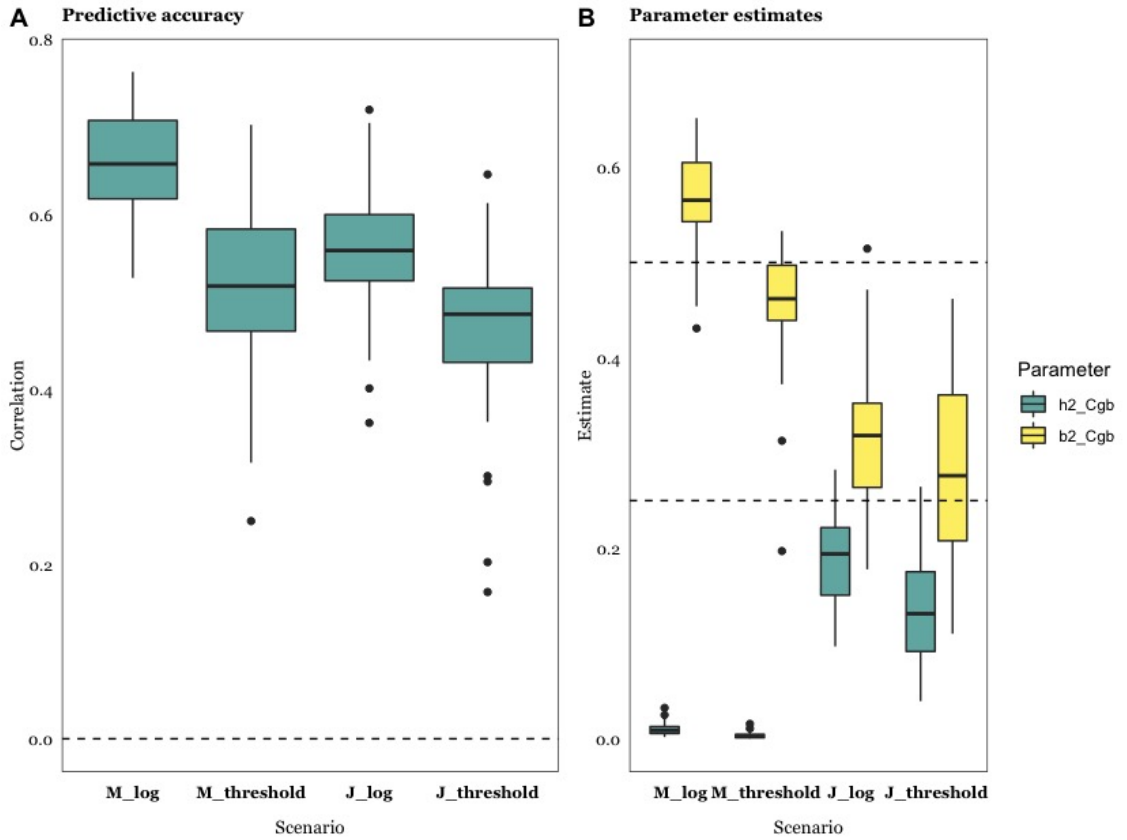
SEM was fitted for each abundance. A whole-genome approach seems in principle more adequate; however, modeling recursive effects in this context is both statistically and computationally challenging because of the large number of SNP-OTU combinations that would need to be considered.

A line of research that we have not considered involves possible microbiome-DNA interactions. Although the number of possible interactions to consider can be huge when the number of SNPs and the number of OTUs is large, interactions between features in two high-dimensional sets can be modeled in a Gaussian context using co-variance functions. These functions are the Hadamard product of set-specific similarity matrices such as the Hadamard product of a SNP-derived and an OTU-derived 'relationship' matrix. Such an approach has been used before to model, e.g. interactions between SNPs or between SNPs and environmental covariates (e.g., 43).

The usefulness of microbiome in prediction depends crucially on its stability in time and space. For instance, although measures of gastrointestinal microbiome abundances are known to be repeatable, it cannot be expected to remain stable throughout an individual's life span. After weaning and under standard management conditions, e.g., constant diet and absence of antibiotic treatment, the diversity of monogastric gut microbiota increases with host age until its composition remains stable. Rumen microbial communities are highly resilient and host-specific [44,45] but change in early life. The transition towards a more stable an adult-like ruminal ecosystem occurs between weaning and one year of age [46]. Therefore, for prediction purposes, we recommend the inclusion of microbial information at least after weaning, preferably at adulthood. This may limit the usefulness of microbiota for prediction in breeding schemes as compared to genomic data in livestock.

At present, modeling the influence of microbiome abundances on complex phenotypes is an open area of research. Here we have presumed that the effects on abundances are additive in the log scale. Similar models are widely used in a diversity of scenarios, e.g., multiplicative models are used to accommodate fitness effects in evolutionary genetics [47] or to deal with highly leptokurtic distributions such as raw abundances, an effect that is smoothed with the log transformation. In addition to the log-transformation, a widely popular choice in genetics is the threshold model [7], which assumes the presence of a

528 continuous liability (here abundances) with an effect value '0' below a given threshold
529 and '1' otherwise. This model has the advantage of being independent on whether
530 abundances are log-transformed or not and is also biologically sound since it is
531 conceivable that a minimum microorganism abundance is required to trigger a particular
532 effect. To test the robustness of the log-transformation, we simulated phenotypes such
533 that 25% of causative abundance observations were above the threshold and the analysis
534 was performed on the log transformed abundances as before. As could be expected, using
535 a 'wrong' model for the analyses was detrimental to prediction but not dramatically (Fig
536 8A). Parameter estimates were affected downwards compared to the multiplicative model
537 (Fig 8B). We suggest that major conclusions from this work should hold even if the
538 relationship between variables and phenotype is not strictly multiplicative.
539



540
541 ***Fig 8***. *Comparison of multiplicative (log) and threshold Microbiome (M) and Joint (J)*
542 *scenarios (r^2 = 0.5).* ***A****: Predictive accuracy, computed as correlation between predicted*
543 *and observed phenotypes, using Bayes Cgb.* ***B:*** *Heritability (h^2) and microbiability (b^2)*
544 *estimates using Bayes Cgb. Results are average of 30 replicates. Scenarios M and J as*
545 *specified in Table 1; the log transformation results are shown for completeness and are*
546 *the same as in Figs 3 and 5. Data are average of 30 replicates.*
547

548 **Conclusion**

This study suggests that microbiome data can significantly improve the prediction of complex phenotypes, irrespective of whether some abundances are under direct genetic control or not. For this strategy to be successful, though, medium to large-sized experiments are required, the microbiome should be relatively stable and should be available before the phenotype is collected. This limits the usefulness of microbiome for prediction in breeding schemes as compared to genome data, which can be collected at birth and remains unchanged. Important potential applications remain nevertheless, such as predicting methane emission in cattle, feed efficiency, disease predisposition, or crop production using soil metagenome. Overall, we can be rather confident that standard linear methods can be used despite the highly leptokurtic distributions observed in OTU abundances. There is room for specific theoretical developments though, perhaps along the lines proposed by Saborio-Montero et al. [42], but these should be based on a better understanding of the relation between microbiome and phenotype. It seems critical to quantify, even approximately, the number of taxa affecting the phenotype and to characterize the distribution of their effects. We are far less optimistic in what regards the identification of causative OTUs, and in particular of the putative QTNs affecting relative abundances.

Materials and Methods

Simulation Strategy

There is ample literature and software available on the simulation of 'standard' complex phenotypes, e.g., [48–51]. These algorithms, however, are not suited for some of the scenarios posed in Fig 1. Here we propose simulating the joint influence of genome and microbiome on a quantitative trait by adding their contributions plus a random noise:

$$y_i = \sum_{j=1}^{N_{QTN}} \alpha_j z_{ij} + \sum_{k=1}^{N_{OTU}} \omega_k x_{ik} + \varepsilon_i, \quad (1)$$

where y_i is the i-th individual record, α_j is the genetic effect of j-th causal SNP (QTN), with $j = 1, N_{QTN}$, the number of QTNs, z_{ij} is the genotype of the i-th individual for j-th SNP coded say -1, 0 and 1 (strict additivity was assumed for all QTN), ω_k is the linear effect of the k-th OTU abundance (x_{ik}), with $k = 1, N_{OTU}$, the number of abundances that

influence the phenotype and ε is a normally distributed residual. The OTU's coefficient can be interpreted as the expected change in phenotype per OTU's abundance unit increase. Since abundances are in the log scale, this is equivalent to a multiplicative effect model. Equation (1) is valid for all scenarios in Fig 1, except that the term involving markers $\sum_{j=1}^{N_{QTN}} \alpha_j z_{ij}$ is removed in the Microbiome and Indirect scenarios whereas the term $\sum_{k=1}^{N_{OTU}} \omega_k x_{ik}$ is removed in the Genome scenario.

For the Indirect and Recursive scenarios, we also need to model the variation in abundances ($\boldsymbol{x}$) that is explained by the genome (Fig 1). Again, we can resort to a linear model where the abundance itself is treated as a standard complex phenotype:

$$x_{ik} = \sum_{j=1}^{N_{QTNk}} \beta_{jk} z_{ij} + \epsilon_i, \qquad (2)$$

where x_{ik} is the abundance level of the k-th OTU that is under partial genetic control for i-th individual, β_j is the genetic effect of j-th QTN on abundance, and z_{ij} is the genotype of the i-th individual for j-th SNP. The j-th sum is across the QTNs influencing k-th abundance, j = 1, $N_{QTN(k)}$. Note abundances x_{ik} in Eqn. (2) are a subset of those in (1). There may be other non-causative abundances under genetic control, but this is irrelevant for our purposes. A phenotype following the Recursive scenario can then be simulated via a two-step procedure: first, simulate abundances (x) using Eqn. (2) followed by phenotype simulation using (1) given the abundances obtained.

We used real genome and microbiome data as input for the simulation procedure. We downloaded the rumen abundance table of 4,018 OTUs from dairy cattle rumen (N = 750, [4]). A pseudo-count equal to one was added to zero abundances, which were next total-sum scaled and log-transformed. This results in much less leptokurtic and less asymmetric distributions than original raw abundances. In Eqns. 1 and 2, x_{ik} represent the already log-transformed abundances. As for genotypes, high-density array genotypes from 750 dairy cows among the total available were downloaded from [11]. To prune SNPs and facilitate computation, 35% of all genotypes with a minimum allele frequency of 0.01 and a maximum missing percentage of 1% were retained. A total of 32,204 autosomal SNPs

was finally retained. The few missing values were simply imputed with the mean. Thirty simulation replicates per scenario were simulated.

Under the Joint scenario, which assumes independence between **G** and **B**, we can simply sample the list of causative SNPs and abundances, simulate their effects, and apply Eqn. 1 to generate phenotype values given observed genotypes and abundances. The case of Recursive and Indirect scenarios is not that obvious because we need to sample abundances that are under genetic control and a link must exist between **G** and **B** (Eqn. 2). We solved this issue by rearranging abundances of a given OTU between individuals such that the desired correlation between abundance and individual's genotypes is attained. This strategy has the important advantage that the distribution of abundances is not changed. Suppose $\boldsymbol{\gamma_{ik}} = \sum_{j=1}^{N_{QTNO}} \beta_{jk}\, z_{ij}$ is the simulated genetic effect of the i-th individual for k-th abundance (Eqn. 2) and that the desired heritability for that abundance is h_k^2. The algorithm (Box 1) is based on the simple observation that, given any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, correlation is maximum ($\rho \sim 1$) when observations in both vectors are sorted and ρ is ~ zero when they are shuffled. Therefore, there must be some order $\boldsymbol{y_{sort}}$ then that fulfills, approximately, the constraint $\text{cor}(\boldsymbol{x}, \boldsymbol{y_{sort}}) = \rho$. For our purposes, we need to rearrange the observed abundances $\boldsymbol{x_k}$ such that the correlation between rearranged $\boldsymbol{x_k}$ and $\boldsymbol{\gamma_k}$ is h_k, the square root of heritability for k-th abundance. The algorithm is detailed in the Box.

A drawback of this algorithm is that it locally breaks the covariance between abundances of different OTUs. To alleviate this, we permuted all abundances that fell within the same OTU cluster. We clustered abundances using R function hclust(dist(.), method="ward.D2") and cut the tree in K = 500 clusters. We chose K = 500 because the first quartile of intra-cluster average correlation was above the third quartile of the average correlation between random abundances, that is, clusters were made up of highly correlated abundances compared to average. We also explored K = 200 but we did not find any difference neither in predictive accuracy nor in heritability estimates. To verify that the shuffling algorithm did not alter the whole structure of the data, we show the principal component analysis of the original and a few shuffled microbiome sets in Fig S3.

646

Algorithm 1: Find a permutation of vectors x and y such that the correlation between permuted vectors is a predetermined value ρ

Take $\mathbf{x},\boldsymbol{y},\rho$, where $\boldsymbol{x}$ and $\mathbf{y}$ are arbitrary, uncorrelated vectors in R^n and $0 \leq \rho \leq 1$ is the desired correlation. The aim is to find a permutation of $\boldsymbol{y}$ such that correlation $\text{cor}(\boldsymbol{x}, \boldsymbol{y}_{sort}) = \rho$, approximately. The algorithm can be equally applied when $\boldsymbol{x}$ and / or $\boldsymbol{y}$ are integer numbers, normality is not required either. Performance of the algorithm improves as n increases and when normality holds.

1. Sort the values of $\boldsymbol{x}$ and $\boldsymbol{y}$ in increasing or decreasing order. The correlation $cor(\boldsymbol{x}_{sort},\boldsymbol{y}_{sort}) \cong 1$.
2. Generate a dummy variable $\boldsymbol{z} = \boldsymbol{y}_{sort} + \boldsymbol{e}$ where $\boldsymbol{e}$ values are sampled from $\boldsymbol{e} \sim N\left(0, S_y^2 \frac{1-\rho^2}{\rho^2}\right)$, S_y^2 is the sample variance of $\boldsymbol{y}$. The correlation $\text{cor}(\boldsymbol{x}_{sort}, \boldsymbol{z}) \sim \rho$.
3. Create an index variable $\boldsymbol{iy}$ which indicates how $\boldsymbol{y}_{sort}$ should be reordered according to $\boldsymbol{z}$ order. This dummy index $\boldsymbol{iy}$ = *order(y)[order(z)]* contains the order of $\boldsymbol{y}$ when values are back-sorted according to the order of $\boldsymbol{z}$.
4. Reorder $\boldsymbol{iy} = \boldsymbol{iy}[\text{rank}(\boldsymbol{x})]$ to match the index with positions $\mathbf{y}_{\text{sort}}$ in the original vector $\boldsymbol{x}$. This is needed since $\boldsymbol{x}$ remains unchanged and only $\boldsymbol{y}$ is permuted.
5. The correlation $cor(\boldsymbol{x},\boldsymbol{y}[\boldsymbol{iy}]) \cong \rho$.

Algorithm available at https://github.com/miguelperezenciso/Simubiome, see sortCorr function.

647

648 **Parameter fitting**

649 Little is known neither on the number of OTUs influencing a given phenotype nor on how
650 many of those are partly inherited. For that reason, we chose some extreme, yet 'educated'
651 values for each of the five scenarios depicted in Fig 1. We considered $r^2 = h_g^2 + h_b^2 =$
652 0.25 and 0.50; $r^2 = 0.25$ is grossly the value reported by Difford et al. 2019 with N = 750,
653 whereas values closer to $r^2 = 0.50$ were reported by Wallace et al. in some farms. Overall,
654 augmenting r^2 values tries to mimic the effect of increasing sample size. We assumed h_g^2
655 $= h_b^2$ for Joint and Recursive scenarios, as also reported by Difford et al or Camarinha-
656 Silva et al. approximately. The number of QTNs was fixed to 100. This figure is
657 somewhat arbitrary, but the specific number of loci would not affect much the results.
658 Barton et al. [52] showed theoretically that most properties of the infinitesimal model

converge as fast as the inverse of the number of loci, or ~ 1% deviance with $N_{QTN} = 100$. In general, genomic prediction is known to be relatively insensitive to the number of QTNs [53]. As for individual genetic effects α, numerous empirical and theoretical works show that they are not uniformly distributed and can be approximated by a gamma-like distribution [54,55]. Here we sampled genetic effects $\alpha \sim \Gamma(\text{shape} = 0.2, \text{scale} = 5)$, as suggested by Caballero et al. [56], and also used previously by us [57].

Much less is known on the number of causative OTUs (N_{OTU}), although we can presume that N_{OTU} should be smaller than the number of QTNs. For instance, Duvallet et al. [36] found in a large meta-analysis that the human diseases studied were affected on average by 10 - 15 changes in abundances at the genus level. Here we considered $N_{OTU} = 25$ (0.6% of all OTUs), although we also evaluated N_{OTU} = 10, 100 and 250. Similarly, for the Recursive and Indirect scenarios, we took the extreme scenario where all causative OTUs are genetically determined, i.e., $N_{OTU} = N_{OTU(g)}$. The genetic effects β on abundances (Eqn. 2) were sampled from the same distribution $\beta \sim \Gamma(\text{shape} = 0.2, \text{scale} = 5)$ as direct genetic effects α. We are much more ignorant regarding the distribution of abundances' effects ω on the phenotype (Eqn. 1). We took as proxy the regression coefficients of methane emission on abundances published by Difford et al. [4], in their supplementary information S4, which can be approximated by a $\Gamma(\text{shape}=1.4, \text{scale}=3.8)$. Fig S3 compares both gamma distributions and the fit to empirical data. This model predicts that the variance of OTUs' effects is wider and of larger individual effect on average than that of SNPs. Although this is speculative at this point, it is sensible to assume that only a few taxa do have a sizeable influence on the phenotype, say methane emission.

Analysis

We used Bayes C algorithm [15] as implemented in BGLR [58] to assess prediction performance and reliability of parameter estimates. We also tested Bayesian RKHS regression, equivalent to GBLUP [40], but results were similar or worse and are not presented. Three models were used to analyze the data:

Bayes Cgb: $\mathbf{y} = \mathbf{Z}\,\mathbf{a} + \mathbf{b}\,\mathbf{W} + \mathbf{e}$ (3a)

Bayes Cg: $\mathbf{y} = \mathbf{Z}\,\mathbf{a} + \mathbf{e}$ (3b)

692

693 Bayes Cb: $\mathbf{y} = \mathbf{W}\,\mathbf{b} + \mathbf{e}$ (3c)

694

695 where **y** is the vector containing the simulated phenotypes, **a** contains the marker effect
696 estimates, **Z** contains the observed genotypes for the 33k markers, **b** contains the OTU
697 abundance effects, **W** is a matrix with all 4,018 abundances in the 750 individuals, and **e**
698 is the residual. Prior to the analyses, phenotypes, abundances, and genotypic values were
699 standardized to mean zero and SD = 1. As priors π for SNPs or abundances probabilities
700 to enter into the model, we used $\pi \sim \text{Beta}(p_0 = 5, \pi_0 = 0.001)$, which has expectation π_0
701 and variance $\pi_0 (1- \pi_0) / (p_0+1)$. We also considered a much more liberal, flat prior for π
702 ~ $\text{Beta}(p_0 = 2, \pi_0 = 0.01)$, but we did not observe strong differences. A total of 50k
703 iterations were run per Bayes C chain, a plot of the residual variances along iterations
704 indicated convergence was attained with this number of iterations. To assess predictive
705 accuracy, 75 (10% of N) phenotypes were randomly removed and predicted with the fitted
706 model. Correlation between observed and predicted phenotypes was used as measure of
707 predictive accuracy.

708

709 The 'heritability' is not explicitly defined in a Bayes C framework, and here we used the
710 proposal by [58] (https://github.com/gdlc/BGLR-R/blob/master/inst/md/heritability.md).
711 In short, at each iteration *i*, the algorithm samples SNPs and OTUs effects:

712

713 $\boldsymbol{u}^{(i)} = \boldsymbol{Z}\,\hat{a}^{(i)}$

714 $\boldsymbol{v}^{(i)} = \boldsymbol{W}\,\hat{b}^{(i)}$

715

716 where $\boldsymbol{u}^{(i)}$ and $\boldsymbol{v}^{(i)}$ are genome and microbiome effects at i-the iteration for the set of
717 individuals, respectively, $\hat{a}^{(i)}$ and $\hat{b}^{(i)}$ are current SNP and OTU abundances solutions;
718 therefore, $Var(\boldsymbol{u}^{(i)})$ / *Var(**y**)* and $Var(\boldsymbol{v}^{(i)})$ / *Var(**y**)* are i-th iterate heritability and
719 microbiability estimates wherefrom posterior means can be estimated by averaging over
720 iterations. For Bayes Cgb, we also sampled the absolute covariance between **u** and ***v***, i.e.,
721 $|Cov(\boldsymbol{u}^{(i)}, \boldsymbol{v}^{(i)})|$ / *Var(**y**)*.

722

723 To assess how likely is to identify causative OTUs, we retained the probability of a given
724 OTU entering into the model, averaged over Gibbs sampling iterations. We run a GWAS

of abundances ($\boldsymbol{x_k}$, k=1, N_{OTU}) on SNP genotypes ($\boldsymbol{z_j}$, j=1, N_{SNP}) using R function lm($\boldsymbol{x_k} \sim \boldsymbol{z_j}$) and we computed the P-value of both causative QTNs, i.e., affecting abundances, and neutral SNPs. This was done in the Recursive scenario only. In this scenario, we also computed the heritabilities of all abundance levels using GBLUP via a RKHS strategy (https://github.com/gdlc/BGLR-R/blob/master/inst/md/GBLUP.md#RKHS) using BGLR. Weakly informative priors for variances were used to mimic a REML-like estimator.

Author contributions

MPE, GDLC and LMZ conceived research. MPE and LMZ performed research. All authors discussed research. MPE wrote the manuscript with help from the rest of authors.

Acknowledgments

This work was developed while MPE and LMZ spent a research stay at Michigan State University kindly funded by GDLC. LMZ is supported by a Ph.D. grant from the Ministry of Economy and Science (MINECO, Spain), MPE is funded by MINECO grants AGL2016-78709-R and PID2019-108829RB-I00, from the EU through BFU2016-77236-P (MINECO/AEI/FEDER, EU) and "Centro de Excelencia Severo Ochoa 2016-2019" award SEV-2015-0533. YRC was funded by Marie Skłodowska-Curie grant (P-Sphere) agreement No 6655919 (EU).

References

1. Ruff WE, Greiling TM, Kriegel MA. Host–microbiota interactions in immune-mediated diseases [Internet]. Nature Reviews Microbiology. Nature Research; 2020. pp. 521–538. doi:10.1038/s41579-020-0367-2
2. Zhang Q, Difford G, Sahana G, Løvendahl P, Lassen J, Lund MS, et al. Bayesian modelling reveals host genetics associated with rumen microbiota jointly influence methane emission in dairy cows. ISME J. 2020;14: 2019–2033. doi:10.1038/s41396-020-0663-x
3. Maltecca C, Bergamaschi M, Tiezzi F. The interaction between microbiome and pig efficiency: A review. J Anim Breed Genet. Blackwell Publishing Ltd; 2020;137: 4–13. doi:10.1111/jbg.12443
4. Difford GF, Plichta DR, Løvendahl P, Lassen J, Noel SJ, Højberg O, et al. Host genetics and the rumen microbiome jointly associate with methane emissions in

dairy cows. PLoS Genet. Public Library of Science; 2018;14: e1007580. doi:10.1371/journal.pgen.1007580

5. Kundu P, Blacher E, Elinav E, Pettersson S. Our Gut Microbiome: The Evolving Inner Self [Internet]. Cell. Cell Press; 2017. pp. 1481–1493. doi:10.1016/j.cell.2017.11.024

6. Difford GF, Lassen J, Løvendahl P. Genes and microbes, the next step in dairy cattle breeding. EAAP—67th Annual Meeting. 2016.

7. Falconer D, Mackay T. Introduction to Quantitative Genetics. essex: Longman Publishing Group; 1996.

8. Zilber-Rosenberg I, Rosenberg E. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. FEMS Microbiol Rev. Oxford Academic; 2008;32: 723–735. doi:10.1111/j.1574-6976.2008.00123.x

9. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. Nature. Nature Publishing Group; 2018;555: 210–215. doi:10.1038/nature25973

10. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. Nature Publishing Group; 2019;569: 655–662. doi:10.1038/s41586-019-1237-9

11. Wallace JR, Sasson G, Garnsworthy PC, Tapio I, Gregson E, Bani P, et al. A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. Sci Adv. American Association for the Advancement of Science; 2019;5: eaav8391. doi:10.1126/sciadv.aav8391

12. Camarinha-Silva A, Maushammer M, Wellmann R, Vital M, Preuss S, Bennewitz J. Host genome influence on gut microbial composition and microbial prediction of complex traits in pigs. Genetics. Genetics Society of America; 2017;206: 1637–1644. doi:10.1534/genetics.117.200782

13. Khanal P, Maltecca C, Schwab C, Fix J, Tiezzi F. Microbiability of meat quality and carcass composition traits in swine. bioRxiv. Cold Spring Harbor Laboratory; 2019; 833731. doi:10.1101/833731

14. Pereyra MA, Creus CM. Modifying the Rhizosphere of Agricultural Crops to Improve Yield and Sustainability: Azospirillum as a Model Rhizotroph. Rhizotrophs: Plant Growth Promotion to Bioremediation. Springer Singapore; 2017. pp. 15–37. doi:10.1007/978-981-10-4862-3_2

15. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157: 1819–1829. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1461589&tool=pmcentrez&rendertype=abstract

16. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, et al. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives [Internet]. Trends in Plant Science. Elsevier Current Trends; 2017. pp. 961–975. doi:10.1016/j.tplants.2017.08.011

17. Meuwissen T, Hayes B, Goddard M. Accelerating Improvement of Livestock with Genomic Selection. Annu Rev Anim Biosci. 2013;1: 221–237. doi:10.1146/annurev-animal-031412-103705

18. De Los Campos G, Gianola D, Allison DB. Predicting genetic predisposition in humans: The promise of whole-genome markers. Nat Rev Genet. Nature Publishing Group; 2010;11: 880–886. doi:10.1038/nrg2898

19. Tierney BT, He Y, Church GM, Segal E, Kostic AD, Patel CJ. The predictive power of the microbiome exceeds that of genome-wide association studies in the discrimination of complex human disease. BioRxiv. Cold Spring Harbor Laboratory; 2020; 2019.12.31.891978. doi:10.1101/2019.12.31.891978

20. Wang J, Chen L, Zhao N, Xu X, Xu Y, Zhu B. Of genes and microbes: solving the intricacies in host genomes. Protein Cell. Higher Education Press; 2018;9: 446–461. doi:10.1007/s13238-018-0532-9

21. Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J, et al. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc Natl Acad Sci U S A. National Academy of Sciences; 2010;107: 18933–18938. doi:10.1073/pnas.1007028107

22. Goodrich JK, Davenport ER, Clark AG, Ley RE. The Relationship Between the Human Genome and Microbiome Comes into View. Annu Rev Genet. Annual Reviews; 2017;51: 413–433. doi:10.1146/annurev-genet-110711-155532

23. Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, et al. Host genetic variation impacts microbiome composition across human body sites. Genome Biol. BioMed Central Ltd.; 2015;16: 191. doi:10.1186/s13059-015-0759-1

24. Crespo-Piazuelo D, Migura-Garcia L, Estellé J, Criado-Mesas L, Revilla M, Castelló A, et al. Association between the pig genome and its gut microbiota

composition. Sci Rep. Nature Publishing Group; 2019;9: 8791. doi:10.1038/s41598-019-45066-6

25. Ramayo-Caldas Y, Zingaretti L, Popova M, Estellé J, Bernard A, Pons N, et al. Identification of rumen microbial biomarkers linked to methane emission in Holstein dairy cows. J Anim Breed Genet. 2019; 49–59. doi:10.1111/jbg.12427

26. Ramayo-Caldas Y, Prenafeta-Boldú F, Zingaretti LM, Gonzalez-Rodriguez O, Dalmau A, Quintanilla R, et al. Gut eukaryotic communities in pigs: diversity, composition and host genetics contribution. Anim Microbiome. BioMed Central; 2020;2: 18. doi:10.1186/s42523-020-00038-4

27. Weissbrod O, Rothschild D, Barkan E, Segal E. Host genetics and microbiome associations through the lens of genome wide association studies. Current Opinion in Microbiology. Elsevier Ltd; 2018. pp. 9–19. doi:10.1016/j.mib.2018.05.003

28. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe. Cell Press; 2015;17: 690–703. doi:10.1016/j.chom.2015.04.004

29. Jakobsson HE, Abrahamsson TR, Jenmalm MC, Harris K, Quince C, Jernberg C, et al. Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by Caesarean section. Gut. BMJ Publishing Group; 2014;63: 559–566. doi:10.1136/gutjnl-2012-303249

30. Furman O, Shenhav L, Sasson G, Kokou F, Honig H, Jacoby S, et al. Stochasticity constrained by deterministic effects of diet and age drive rumen microbiome assembly dynamics. Nat Commun. Nature Research; 2020;11: 1–13. doi:10.1038/s41467-020-15652-8

31. Gamazon ER, Segrè A V., Van De Bunt M, Wen X, Xi HS, Hormozdiari F, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet. Nature Publishing Group; 2018;50: 956–967. doi:10.1038/s41588-018-0154-4

32. Hormozdiari F, van de Bunt M, Segrè A V., Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. Cell Press; 2016;99: 1245–1260. doi:10.1016/j.ajhg.2016.10.003

33. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol.

Frontiers; 2017;8: 2224. doi:10.3389/fmicb.2017.02224

34. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. BioMed Central; 2017;5: 27. doi:10.1186/s40168-017-0237-y

35. Maltecca C, Lu D, Schillebeeckx C, McNulty NP, Schwab C, Shull C, et al. Predicting Growth and Carcass Traits in Swine Using Microbiome Data and Machine Learning Algorithms. Sci Rep. Nature Publishing Group; 2019;9: 1–15. doi:10.1038/s41598-019-43031-x

36. Power RA, Parkhill J, De Oliviera T. Microbial Genome-Wide Association Studies: Lessons from Human GWAS. doi:10.1101/093211

37. Roehe R, Dewhurst RJ, Duthie CA, Rooke JA, McKain N, Ross DW, et al. Bovine Host Genetic Variation Influences Rumen Microbial Methane Production with Best Selection Criterion for Low Methane Emitting and Efficiently Feed Converting Hosts Based on Metagenomic Gene Abundance. Leeb T, editor. PLoS Genet. 2016;12: e1005846. doi:10.1371/journal.pgen.1005846

38. Huws SA, Creevey CJ, Oyama LB, Mizrahi I, Denman SE, Popova M, et al. Addressing global ruminant agricultural challenges through understanding the rumen microbiome: Past, present, and future. Frontiers in Microbiology. Frontiers Media S.A.; 2018. p. 2161. doi:10.3389/fmicb.2018.02161

39. Wallace RJ, Rooke JA, McKain N, Duthie CA, Hyslop JJ, Ross DW, et al. The rumen microbial metagenome associated with high methane production in cattle. BMC Genomics. BioMed Central Ltd.; 2015;16: 839. doi:10.1186/s12864-015-2032-0

40. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. Elsevier; 2008;91: 4414–4423. doi:10.3168/jds.2007-0980

41. Ross EM, Moate PJ, Marett LC, Cocks BG, Hayes BJ. Metagenomic predictions: from microbiome to complex health and environmental phenotypes in humans and cattle. White BA, editor. PLoS One. Public Library of Science; 2013;8: e73056. doi:10.1371/journal.pone.0073056

42. Saborío-Montero A, Gutiérrez-Rivas M, García-Rodríguez A, Atxaerandio R, Goiri I, López de Maturana E, et al. Structural equation models to disentangle the biological relationship between microbiota and complex traits: Methane production in dairy cattle as a case of study. J Anim Breed Genet. Blackwell

Publishing Ltd; 2020;137: 36–48. doi:10.1111/jbg.12444

43. Muñoz PR, Resende MFR, Gezan SA, Resende MDV, de los Campos G, Kirst M, et al. Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics. Genetics Society of America; 2014;198: 1759–1768. doi:10.1534/genetics.114.171322

44. Cole NA. Effects of animal-to-animal exchange of ruminal contents on the feed intake and ruminal characteristics of fed and fasted lambs. J Anim Sci. Oxford Academic; 1991;69: 1795. doi:10.2527/1991.6941795x

45. Weimer PJ. Redundancy, resilience, and host specificity of the ruminal microbiota: Implications for engineering improved ruminal fermentations. Frontiers in Microbiology. Frontiers Media S.A.; 2015. p. 296. doi:10.3389/fmicb.2015.00296

46. Dill-Mcfarland KA, Breaker JD, Suen G. Microbial succession in the gastrointestinal tract of dairy cows from 2 weeks to first lactation. Sci Rep. Nature Publishing Group; 2017;7: 1–12. doi:10.1038/srep40864

47. Crow J, Kimura M. An Introduction to Population Genetics Theory. The Blackburn Press; 1970.

48. Peng B, Kimmel M. simuPOP: A forward-time population genetics simulation environment. Bioinformatics. 2005;21: 3686–3687. doi:10.1093/bioinformatics/bti584

49. Messer PW. SLiM: simulating evolution with selection and linkage. Genetics. 2013;194: 1037–1039. doi:10.1534/genetics.113.152181

50. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. The American Society of Human Genetics; 2011;88: 76–82. doi:10.1016/j.ajhg.2010.11.011

51. Pérez-Enciso M, Ramírez-Ayala LC, Zingaretti LM. SeqBreed: A python tool to evaluate genomic prediction in complex scenarios. Genet Sel Evol. BioMed Central Ltd.; 2020;52. doi:10.1186/s12711-020-0530-2

52. Barton NH, Etheridge AM, Véber A. The infinitesimal model: Definition, derivation, and implications. Theor Popul Biol. Elsevier Inc.; 2017;118: 50–73. doi:10.1016/j.tpb.2017.06.001

53. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The impact of genetic architecture on genome-wide evaluation methods. Genetics. 2010;185: 1021–1031. doi:10.1534/genetics.110.116855

929 54. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new
930 mutations. Nat Rev Genet. 2007;8: 610–618. doi:10.1038/nrg2146
931 55. Barton NH, Keightley PD. Understanding Quantitative Genetic Variation. Nat
932 Rev Genet. 2002;3: 11–21. doi:10.1038/nrg700
933 56. Caballero A, Tenesa A, Keightley PD. The nature of genetic variation for
934 complex traits revealed by GWAS and Regional Heritability Mapping analyses.
935 Genetics. 2015;201: 1601–1613.
936 57. Pérez-Enciso M, Forneris N, de Los Campos G, Legarra A. Evaluating Sequence-
937 Based Genomic Prediction with an Efficient New Simulator. Genetics. Genetics;
938 2017;205: 939–953. doi:10.1534/genetics.116.194878
939 58. Pérez P, de Los Campos G. Genome-Wide Regression & Prediction with the
940 BGLR Statistical Package. Genetics. 2014;198: 483–95.
941 doi:10.1534/genetics.114.164442
942