Peer Community Journal

Research article

Correspondence
solene.pety@inrae.fr

Simulating transgenerational hologenomes under selection with RITHMS

Solène Pety [iD],[1,2], Ingrid David [iD],[3], Andrea Rau [iD],[2], and Mahendra Mariadassou [iD],[1]

Abstract

A holobiont is made up of a host organism together with its microbiota. In the context of animal breeding, the holobiont can be viewed as the single unit upon which selection operates. Therefore, integrating microbiota data into genomic prediction models may be a promising approach to improve predictions of phenotypic and genetic values. Nevertheless, there is a paucity of hologenomic transgenerational data to address this hypothesis, and thus to fill this gap, we propose a new simulation framework. Our approach, an R Implementation of a Transgenerational Hologenomic Model-based Simulator (RITHMS) is an open-source package. It builds upon simulated transgenerational genotypes from the Modular Breeding Program Simulator (MoBPS) package and incorporates distinctive characteristics of the microbiota, notably vertical and horizontal transmission as well as modulation due to the environment and host genetics. In addition, RITHMS can account for a variety of selection strategies and is adaptable to different genetic architectures. We simulated transgenerational hologenomic data using RITHMS under a wide variety of scenarios, varying heritability, microbiability, and microbiota transmissibility. We found that simulated data accurately preserved key characteristics across generations, notably microbial diversity metrics, exhibited the expected behavior in terms of correlation between taxa and of modulation of vertical and horizontal transmission, response to environmental effects and the evolution of phenotypic values depending on selection strategy. Our results support the relevance of our simulation framework and illustrate its possible use for building a selection index balancing genetic gain and microbial diversity and for evaluating the impact of partially observed microbiota data. RITHMS is an advanced, flexible tool for generating transgenerational hologenomes under selection that incorporate the complex interplay between genetics, microbiota and environment.

[1]Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France,   [2]Université Paris-Saclay, INRAE, GABI, 78350, Jouy-en-Josas, France,   [3]Université de Toulouse, INRAE, ENVT, GenPhySE, 31326, Castanet-Tolosan, France

CENTRE MERSENNE

Introduction

An individual host organism and its associated microbiota together constitute a holobiont, a complex biological system characterized by symbiotic interactions (Ghataora and Ellis, 2025; Guerrero et al., 2013). The concept of the holobiont was introduced several decades ago (Margulis and Fester, 1991) and, although sometimes debated (Theis et al., 2016), it has gained increasing recognition as a relevant biological entity on its own, integrating both the host and its microbial partners. The holobiont concept has notably garnered attention in recent years in studies focused on the sustainable production of livestock species within a One Health framework (Evans and Leighton, 2014). In particular, the complex dependencies between host genetics and microbiota, together with other environmental factors, have been shown to impact agriculturally important phenotypes such as digestive and feed efficiency (Déru et al., 2022; Li et al., 2019; Niu et al., 2015), milk composition (Brulin et al., 2024; Martinez Boggio et al., 2022) and mastitis triggering (Derakhshani et al., 2018).

The co-evolution of the host genome and its microbiome under selective pressure suggest the potential interest of using the holobiont as a selection unit for animal breeding (Zilber-Rosenberg and Rosenberg, 2008). Since the end of the Second World War, selection has led to substantial gains in performance traits for livestock (Goldberg and Deb, 1991), further strengthened in the past two decades by the widespread use of genomic data to estimate breeding values. These genomic prediction methods are most often based on linear models assuming additive genetic effects, allowing a decomposition of phenotypic variance into genetic and environmental components (Leal, 2001; Legarra et al., 2018). In this quantitative genetics framework, a phenotype is simply the measurable trait of interest (e.g. milk yield, growth, or disease resistance), often considered on a continuous scale. Thanks to the relatively low cost of sequencing technologies, it is now straightforward to assay the microbiota using metabarcoding data in addition to host genomic data. However, jointly considering host genotypes and microbiota for selection is complicated in part by the multiple potential modes of transmission of the microbiota and in part by their interplay on the phenotype. For example, in the first few moments of existence for mammals, maternal contact during delivery and nursing play a crucial role in establishing the initial microbiota through vertical transmission (Bruijning et al., 2022; Cortes-Macías et al., 2021; Rutayisire et al., 2016). For non-mammalian vertebrates, such as chickens (Shterzer et al., 2023), the maternal contribution is likely to be considerably weaker than in mammals. In stark contrast to genotypes, a fraction of the microbiota is acquired from the environment through horizontal transmission, and the microbiota continues to evolve throughout a host's life. In addition, both host genes and environmental factors influence the colonization, development, and function of the microbiota, which in turn contributes to host phenotypes.

Several recent studies have sought to evaluate the added value of hologenomic as compared to genomic selection (Martinez Boggio et al., 2024; Ross et al., 2013; Weishaar et al., 2020). In large-scale studies that collect microbiome data on a large number of individuals, as is done in genomic studies, the microbiota is typically measured at a single time point (or at most a few). Consequently, what is analysed is not the entire temporal dynamic of microbiota, but the resulting community at a specific and relatively stable stage. Increasing the scale and scope of such studies raises a number of statistical and computational challenges with respect to the simultaneous integration of host genotypes and microbiota. Moreover, benchmarking studies to evaluate

predictive hologenomic models require sufficiently large and fully paired transgenerational genomic and microbiota data, notably for the comparison of predicted breeding values to observed offspring phenotypes. Such experimental data are costly to acquire and can be impacted by biases (*e.g.*, fluctuating environmental conditions).

Simulation therefore offers an efficient and cost-effective approach to assess the relevance of hologenomic prediction strategies based on host genotypes and a stable snapshot of the microbiota. Several tools for simulating transgenerational genotypes are well known and implemented in user-friendly software such as the Modular Breeding Program Simulator (MoBPS) (Pook et al., 2020) or AlphaSim (Faux et al., 2016). These tools provide flexible and efficient implementations that allow for a wide range of breeding schemes under a variety of scenarios (*e.g.* heritability) but do not integrate microbiota data. With respect to hologenomic data, other simulation approaches have focused on modeling the structure and dynamics of the microbiota, both for exploring breeding strategies, as well as integrating different types of data to account for complex microbiota-host genome interactions (Pérez-Enciso et al., 2021; Wirbel et al., 2024). These methods all focus on the potential added value of the microbiota for selection, and thus generally focus on generating a stable snapshot rather than attempting to model and reproduce its dynamic throughout the lifetime of the host, even though this may result in a lack of power. This focus on a single, relatively stable time point is in line with current large-scale microbiome studies, which typically measure the microbiota at only one (or a few) sampling time. Concepts such as transmissibility have also emerged, taking into account the transmissibility of non-genetic information and thus broadening the vision of inclusive heritability (David and Ricard, 2019). However, none of these hologenomic simulation approaches incorporate a transgenerational aspect.

One recent exception for simulating a co-evolving genome and microbiota under selection is HoloSimR (Casto-Rebollo et al., 2025), which generates a fully synthetic set of genomic and microbiota data over multiple generations based on a user-provided population demographic history and pre-defined model of species abundance distribution using AlphaSimR (Gaynor et al., 2021) and mobsim (May et al., 2018), respectively. Although it constitutes a solid approach for simulating transgenerational hologenomic data, HoloSimR does have several limitations. First, real hologenomic datasets cannot be used to initialize simulations, and the base population is directly artificially generated under assumptions that may oversimplify the complex patterns within and between microbiota and genomic data. Second, it uses a single heritability value for all taxa and thus cannot reproduce the distribution of heritability values inherent to real holobionts, with some taxa being selected for and highly heritable whereas other are selected against and less heritable. Third, the impact of short- and long-term environmental perturbations on the microbiota, such as antibiotic treatments or diet changes, cannot be incorporated in the simulations.

To address these limitations, we introduce an R Implementation of a Transgenerational Hologenomic Model-based Simulator (RITHMS). RITHMS is a flexible framework for simulating transgenerational hologenomic data that accounts for the specificities of microbiota transmission and covers the same range of breeding schemes as MoBPS, but under additional and more complex scenarios (heritability, microbiability, microbiota transmissibility, etc). Real genomic and microbiota data are used to construct a base population from which subsequent generations are derived. This work describes the general framework and strategy used for simulations, and shows

that simulated data preserve key characteristics of real data. Finally, it demonstrates the useful-ness of transgenerational hologenomic data simulated with RITHMS through a case study on mixed-objective selection, incorporating both phenotype values and microbial diversity, as well as a case study focusing on partially unobserved microbiota data.

Material and methods

RITHMS directly incorporates the complexity of transgenerational hologenomic data in sev-eral ways (Figure 1): (1) it uses user-provided paired genomic and microbiota data to create a realistic base population from which successive non-overlapping generations are generated, (2) it takes into account the particularities of microbiota transmission (vertical and horizontal) and genetic modulation, (3) it leverages the functionalities provided in MoBPS (Pook et al., 2020) to define complex genetic architectures and breeding selection steps using indices based on breed-ing values, microbial descriptors, or a combination of the two, and (4) it facilitates simulations under a variety of scenarios.

RITHMS works with an initialization step of the base population followed by repeated steps (once per generation) of simulation for subsequent generations, summarized in Figure 1 and discussed in greater detail in the following. Key simulation parameters and notations are detailed in Table 1.

Formatting of the base population

The base population corresponds to user-provided paired genotype $\mathbf{G}^{(0)}$ and microbiota $\mathbf{M}^{(0)}$ data for N individuals, with respectively n_g SNPs and n_b taxa. For the base population alone (G0), random matings among all individuals are used to generate the following generation (G1) using MoBPS. Exactly n_{ind} individuals are generated according to the user-specified sex ratio (0.5 by default).

Genotype data. Genotype data $\mathbf{G}^{(0)}$ should be coded as the number of alternative alleles at each variant for each individual (a $n_g \times N$ matrix). Genotypes are provided by the user and can there-fore correspond to real data, ensuring realistic linkage disequilibrium (LD) and allelic frequency features in subsequent simulated generations. Sex chromosomes and sample meta-data beyond individual identifiers are not used for the simulations and are ignored if present. Each individual in the population is assigned as female or male according to a sex ratio parameter (set to 0.5 by default).

Microbiota data. As for the base population genotypes, $\mathbf{M}^{(0)}$ can be based on real data, as long as they can be summarized as count table. Our motivating example and illustration are based on 16S rRNA metabarcoding data as they are the most common in breeding studies (Goodrich et al., 2016) but it could be equally applied to WGS data. A $n_b \times N$ count matrix is expected, with the same N individuals as those from $\mathbf{G}^{(0)}$. No modulation is applied to this initial micro-biota as it is already considered to be under genetic influence. The provided raw abundances of taxa are used to estimate compositions (*i.e.* vectors of relative abundances) using an empiri-cal Bayes approach to avoid zeroes, as a data-driven alternative to pseudocounts. To remove the compositionality constraint when incorporating genetic and environmental modulations, relative abundances are subsequently transformed using the centered log-ratio (CLR) from the *composi-tions* R package (Van den Boogaart and Tolosana-Delgado, 2008), $\text{clr}(\mathbf{x}) = \left(\ln x_i - \frac{1}{n_b} \sum_{j=1}^{n_b} \ln x_j \right)_i$. This transformation formally includes a centering term such that the resulting values sum to zero

Table 1 – Table of key parameters and notations for RITHMS

Symbol	Definition	Dimensions	Default initialization	RITHMS parameter
N	Number of individuals in the base population	1×1	(*)	
n_{gen}	Number of generations after the base population	1×1	5	n_gen
n_{ind}	Number of individuals per generation	1×1	N	n_ind
Genotype parameters and notations				
n_{g}	Number of SNPs	1×1	(*)	
v	Number of genetic clusters	1×1	(*)	
$\mathbf{G}^{(0)}$	Base population genotypes encoded as 0,1,2	$n_{\mathrm{g}} \times N$	(*)	
$\mathbf{G}^{(t)}$	Genotypes encoded as 0,1,2	$n_{\mathrm{g}} \times n_{\mathrm{ind}}$		
$\mathrm{QTL}_{\mathrm{y}}$	Number of causative QTL on the phenotype	1×1	100	qtn_y
$\mathrm{QTL}_{\mathrm{o}}$	Number of causative QTL on taxa abundances (per taxon)	1×1	(*) $n_{\mathrm{g}} * 0.2/v$	
Microbiota parameters and notations				
n_{b}	Number of taxa	1×1	(*)	
λ	Proportion of vertical transmission	1×1	0.5	lambda
β	QTL effects on taxa abundances	$n_{\mathrm{b}} \times n_{\mathrm{g}}$	$\sim \mathcal{N}(0, \sigma_\beta^2)$	
σ_β	Standard deviation for non-null cluster- and taxon-specific genetic effects on taxa abundances	1×1	0.1	effect_size
σ_m	Standard deviation value for microbiota noise	1×1	0.1	noise.microbiome
$\mathbf{M}^{(0)}$	Base population taxa counts	$n_{\mathrm{b}} \times N$	(*)	
$\mathbf{M}^{(t)}$	Taxa abundances	$n_{\mathrm{b}} \times n_{\mathrm{ind}}$		
$\mathbf{B}^{(t)}$	CLR-transformed relative abundance values for taxa of all individuals at generation t, $\mathrm{CLR}(\mathbf{M}^{(t)})$	$n_{\mathrm{b}} \times n_{\mathrm{ind}}$		
k	Number of fixed effects	1×1		
θ	Environmental effects on taxa abundances	$n_{\mathrm{b}} \times k$		
$\mathbf{E}^{(t)}$	Environmental factors	$n_{\mathrm{ind}} \times k$		
$\mathrm{OTU}_{\mathrm{g}}$	Percentage of taxa under genetic control	1×1	5%	otu_g
$\mathrm{OTU}_{\mathrm{y}}$	Percentage of causative taxa on phenotype	1×1	otu_g	
Phenotype parameters and notations				
$\mathbf{y}^{(t)}$	Phenotype, $\alpha^T\mathbf{G}^{(t)} + \omega^T\mathbf{B}^{(t)} + \epsilon_y^{(t)}$	$n_{\mathrm{ind}} \times 1$		
ω	Taxa effects on phenotype	$1 \times n_{\mathrm{b}}$	$\sim \Gamma(1.4, 3.8)$	
α	QTL effects on phenotype	$1 \times n_{\mathrm{g}}$	$\sim \Gamma(0.4, 5)$	
h_{d}^2	Direct heritability, $\mathrm{var}(\alpha^T\mathbf{G}^{(t)})/\mathrm{var}(\mathbf{y}^{(t)})$	1×1	(**)	h2
$\omega^T\mathbf{B}^{(t)}$	Microbiota effect	$n_{\mathrm{ind}} \times 1$		
b^2	Microbiability, $\mathrm{var}(\omega^T\mathbf{B}^{(t)})/\mathrm{var}(\mathbf{y}^{(t)})$	1×1	(**)	b2
h^2	Total heritability, $[\mathrm{var}(\alpha^T\mathbf{G}^{(t)}) + \mathrm{var}(\omega^T\beta\mathbf{G}^{(t)})]/\mathrm{var}(\mathbf{y}^{(t)})$	1×1		
$\mathbf{BV}_t^{(t)}$	Total breeding value, all genetic effects, $\mathbf{BV}_t^{(t)} = \mathbf{BV}_m^{(t)} + \mathbf{BV}_d^{(t)}$	$n_{\mathrm{ind}} \times 1$		
$\mathbf{BV}_m^{(t)}$	Microbiota-mediated breeding value, $\mathbf{BV}_m^{(t)} = \omega^T\beta\mathbf{G}^{(t)}$	$n_{\mathrm{ind}} \times 1$		
$\mathbf{BV}_d^{(t)}$	Direct breeding value, $\mathbf{BV}_d^{(t)} = \alpha^T\mathbf{G}^{(t)}$	$n_{\mathrm{ind}} \times 1$		

(*) = Calibrated from input data, (**) = Required parameters, $^{(t)}$ indicates quantities pertaining to generation t
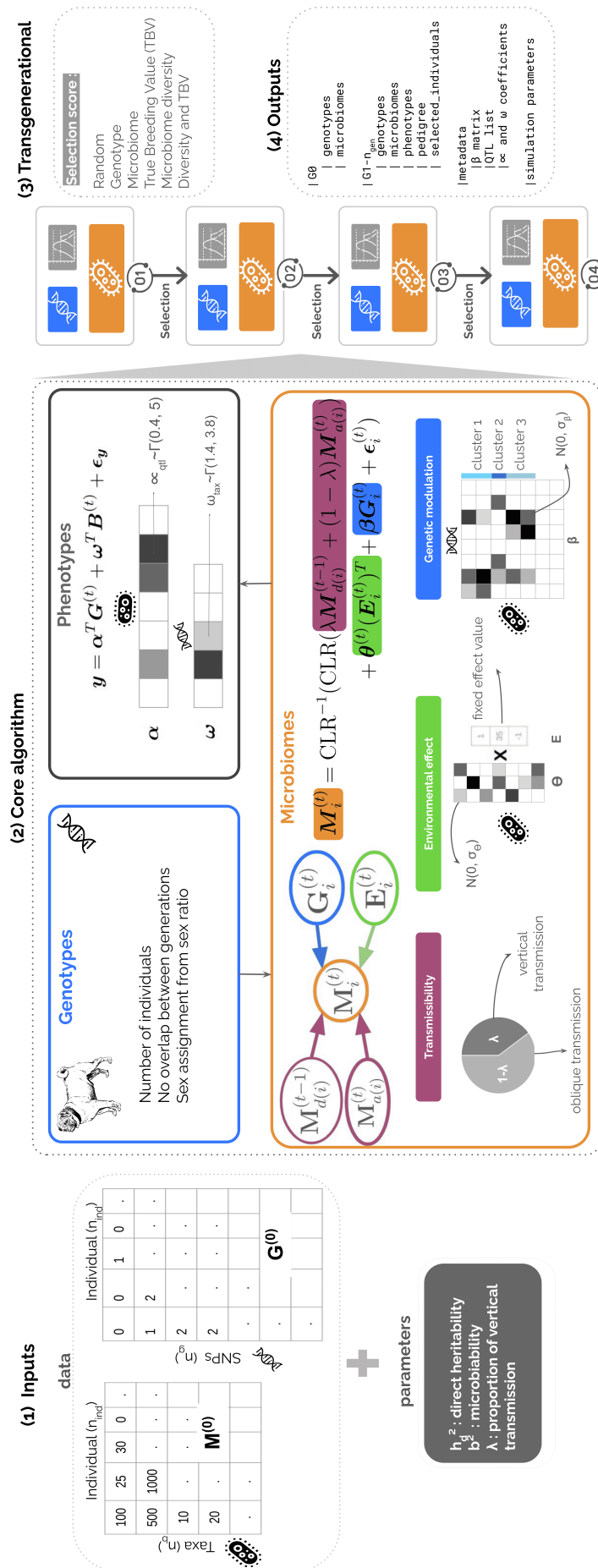
Figure 1 – Overview of RITHMS. (1) User-provided inputs include paired microbiota abundances and genotypes (encoded as 0/1/2) and the following required parameters: direct heritability h_d^2, microbiability b^2 and λ, which modulates the vertical versus horizontal transmission ratio. (2) For each simulated generation t, the genotypes and pedigree are generated using the MoBPS package (Pook et al., 2020). Microbiota are then constructed by first combining maternal and ambient microbiota in proportions λ and $1 - \lambda$ respectively, and subsequently applying genetic and possibly environmental modulation. Genotypes and microbiota are then integrated to simulate the phenotypes of the generation using the recursive model of Pérez-Enciso et al. (2021). (3) To proceed with the next generation, 30% of the males and 30% of the females are selected, either randomly or based on a selection index chosen by the user. (4) The output of the package consists of a nested list including the simulation parameters and, for each generation t, the microbiota $\mathbf{M}^{(t)}$ and genotype $\mathbf{G}^{(t)}$ tables.
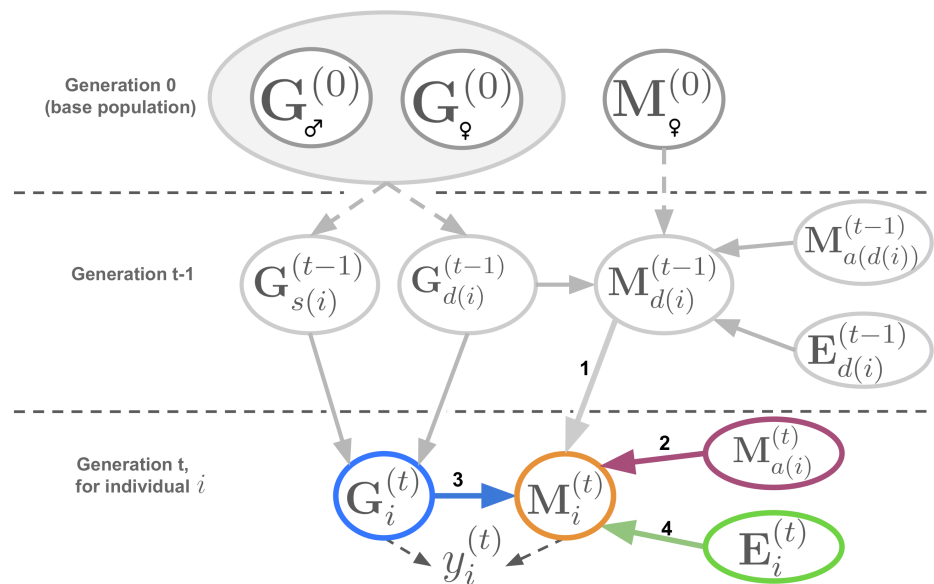
Figure 2 – Schematic illustration of transgenerational hologenomic simulations with RITHMS. The base generation is calibrated on user-provided data for sire genotypes $\mathbf{G}_{\sigma}^{(0)}$, dam genotypes $\mathbf{G}_{\varphi}^{(0)}$, and microbiota data from dams $\mathbf{M}_{\varphi}^{(0)}$. Genotypes are simulated using MoBPS (Pook et al., 2020). The sources contributing to the variability of taxa abundances of individual i at generation t are as follows: (1) vertical transmission from the individual's mother $\mathbf{M}_{d(i)}^{(t-1)}$, for example during delivery; (2) horizontal transmission of the individual-specific ambient microbiota $\mathbf{M}_{a(i)}^{(t)}$; (3) the host selective filter, through which the host's genotype $\mathbf{G}_i^{(t)}$ facilitates the colonization and establishment of certain microorganisms; and (4) individual-specific environmental effects $\mathbf{E}_i^{(t)}$, such as diet or treatment effects, modulating the microbiota composition. Phenotypes $\mathbf{y}_i^{(t)}$ are simulated according to a linear model, where the microbiota has a direct effect and the genome has both direct and microbiota-mediated effects (Pérez-Enciso et al., 2021).

(Gloor et al., 2017). However, backward transformation to relative abundances is performed using the inverse CLR function (`clrinv`), which is invariant to translation, such as centering. We can therefore omit it from our equations without loss of generality. The empirical Bayes approach uses a Dirichlet prior $\mathcal{D}(S\mathbf{p})$ with scale parameter S and mean parameter $\mathbf{p}$. The latter is set to the population-level composition, estimated as the average of relative abundances across all individuals. In practice, the composition of an individual thus corresponds to a weighted average between its empirical composition and that of the mean population with proportions π and $1 - \pi$ respectively. The default value $\pi = 0.75$ corresponds to a scale parameter S set to a third of the sample total count.

Real hologenomic data used as a base population. To illustrate the functionality of RITHMS, a set of hologenomic data from a single line of $N = 750$ pigs fed a conventional diet (Déru et al., 2020) were used as a base population in this work. Individual pigs were genotyped using a 70K SNP GeneSeek GGP Porcine HD chip, and microbiota composition was analysed using the V3-V4 region of 16S rRNA gene (see Déru et al. (2022) for additional details on data acquisition and processing). We focused here on a subset of the first 5000 SNPs from the chip manifest as well as the 1845 taxa with a prevalence higher than 5% after rarefaction (rarefied depth = 4100 reads).

Simulation of subsequent generations

Simulation of genotypes. We simulate pedigrees and successive generations of genotype data using MoBPS based on the genotype data provided as input. By default, $n_{\text{gen}} = 5$ non-overlapping generations are simulated, each with $n_{\text{ind}} = 500$ individuals and a sex ratio of 0.5. Regardless of the simulation strategy employed, 30% of females and 30% of males at each generation are chosen to reproduce for the following generation, with selection either performed randomly or based on a user-specific selection score (see the *Selection* section).

Simulation of microbiota. As our goal is to evaluate the potential added value of hologenomic data in breeding schemes, we aim to simulate a stable snapshot of microbiota data that incorporates both genetic and environmental modulations. Our simulation framework is based on the idea that the initial composition of an individual's microbiota can be partly inherited from its mother through vertical transmission as well as from the direct environment and is subsequently modulated by the host genotype and environmental factors prior to the snapshot (Figure 2). Therefore, we propose the following model for the microbiota abundances of individual i at generation t ($t = 1, \ldots, n_{\text{gen}}$):

$$\mathbf{M}_i^{(t)} = \text{CLR}^{-1}\left(\text{CLR}\left(\lambda\mathbf{M}_{d(i)}^{(t-1)} + (1-\lambda)\mathbf{M}_{a(i)}^{(t)}\right) + \theta(\mathbf{E}_i^{(t)})^T + \beta\mathbf{G}_i^{(t)} + \epsilon_i^{(t)}\right)$$

based on the following matrices, with their dimensions :

- $\mathbf{M}_i^{(t)}$: taxa abundances in individual i ($n_{\text{b}} \times 1$)
- λ : proportion of microbiota inherited via vertical transmission from the mother before modulation by selective filtering and random perturbations
- $\mathbf{M}_{d(i)}^{(t-1)}$: taxa abundances of the dam of individual i ($n_{\text{b}} \times 1$)
- $\mathbf{M}_{a(i)}^{(t)}$: Ambient taxa abundances for individual i ($n_{\text{b}} \times 1$)
- θ : environmental effects on taxa abundances ($n_{\text{b}} \times k$)
- $\mathbf{E}_i^{(t)}$: environmental factors for individual i ($1 \times k$)
- β : multiplicative effect of genotype on taxa abundances ($n_{\text{b}} \times n_{\text{g}}$)
- $\mathbf{G}_i^{(t)}$: genotype of individual i ($n_{\text{g}} \times 1$)
- $\epsilon_i^{(t)} \sim \mathcal{N}(0, \sigma_m^2 \mathbf{I}_{n_{\text{b}}})$: multivariate Gaussian white noise.

Ambient microbiota for each individual. We assume that no herd structure is considered here, that is, that individuals from the same generation live in similar conditions and are therefore exposed to the same sources of microorganisms. As is common in breeding simulations, we further assume non-overlapping generations. To account for these constraints and avoid confounding horizontal and direct non-maternal transmission, we propose a horizontal transmission that is conceptually similar to an oblique transmission between generations. Specifically, we propose a two-step process based on (1) an ambient microbiota that evolves slowly across generations, and (2) an individual-specific incorporation of this ambient microbiota.

First, to simulate a slowly-evolving ambient microbiota at generation t, we use the average composition of the previous generation, $\overline{\mathbf{M}}^{(t-1)}$. Second, we allow each individual to integrate these potential new communities in a different way, leading to the need to include inter-individual variability in the ambient microbiota composition while preserving the structure of the average composition of the previous generation. To introduce inter-individual variability, a random composition is sampled from a Dirichlet prior $\mathbf{M}_{r(i)}^{(t)} \sim \text{Dir}(\eta\overline{\mathbf{M}}^{(t-1)})$, where $\eta > 0$ (set to

25 by default) is the dispersion parameter calibrated to mimic the dispersion in the base population. Here, η was set to 25 based on visual inspection and minimal differentiation of real and simulated compositions according to a PERMANOVA test (results not shown). Similar values of η were found with other datasets (Chaillou et al., 2015; Pérez-Enciso et al., 2021). Although this sampling results in compositions centered around $\overline{\mathbf{M}}^{(t-1)}$, extremely low abundances, corresponding to very large CLR-transformed values may occur, thus overwhelming any possible modulation by noise or genetics. To regularize this sampling, in particular when $\lambda = 0$ (*i.e.* no vertical transmission), we compute a weighted average between the sampled composition and the average composition of the previous generation $\overline{\mathbf{M}}^{(t-1)}$:

$$\mathbf{M}_{a(i)}^{(t)} = \pi \mathbf{M}_{r(i)}^{(t)} + (1-\pi)\overline{\mathbf{M}}^{(t-1)},$$

with the same weights $(\pi, 1-\pi)$ as those used to compute the base population microbiota.

Environmental fixed effects. Taxa abundances can be modulated by environmental fixed effects, which may come from different sources and must be modeled accordingly. In particular, some covariates may not impact all taxa, individuals in a generation, or generations in the same way. To incorporate such effects, RITHMS incorporates the term $\theta(\mathbf{E}_i^{(t)})^T$ to indicate the taxa, individuals, and generations for which environmental effects are applied from G1 onwards. In practice, we recommend that nonzero θ values be drawn from a standard normal distribution to ensure they have the same scale as other parameters.

Genetic modulation. It has been shown that some taxa are correlated, regardless of whether or not a taxonomic link exists between them (Wu et al., 2021). It is thus reasonable to assume that the genetic modulation of taxa has a clustered structure so as to mimic existing correlations between taxa. In particular, simulations should yield strongly positive correlations between taxa in the same cluster, and weak or even negative correlations between taxa in different clusters. These clusters are identified from the rarefied taxa counts of the base population microbiota data $\mathbf{M}^{(0)}$, using hierarchical clustering based on Bray–Curtis dissimilarities, as previously applied for taxa clustering (Aguilar Marín et al., 2020; Pang et al., 2023); note that other distance measures could be considered. By default, 100 clusters are generated. To choose the OTU_g taxa (by default 5% of all taxa) under genetic control, we randomly and iteratively select clusters of size 10 to 25 taxa until a threshold of OTU_g taxa is reached.

The challenge then lies in constructing a sparse matrix of QTL effects on taxa CLR abundances, β. The term $\beta\mathbf{G}^{(t)}$ corresponds to the cumulative effect of the QTL_o causative SNPs on the taxa. Here, by default we set QTL_o to 20% of the total number of SNPs (n_g) divided by the number of clusters under genetic control. To reach a given intra-cluster level of genetic correlation, causative SNPs are sampled randomly for each cluster but common to all taxa within a cluster. The non-null coefficient β_{sg} for genetically modulated taxon s from cluster $c(s)$ with QTL g is then set to $\beta_{sg} = \tilde{\beta}_{c(s)g} + \tilde{\beta}_{sg}$, where $\tilde{\beta}_{c(s)g} \sim \mathcal{N}(0, \sigma_\beta^2)$ and $\tilde{\beta}_{sg} \sim \mathcal{N}(0, \sigma_\beta^2)$. Note that $\tilde{\beta}_{c(s)g}$ depends only on the cluster, ensuring within-cluster correlation, whereas $\tilde{\beta}_{sg}$ is specific to each taxon under genetic control. In this way, as there may be some overlap in QTLs selected for different clusters, both intra- and inter-cluster genetic correlations are induced. The strength of this correlation is mainly limited by the number of clusters and the level of overlap of causative SNPs between clusters. The direction of correlation between taxa in two clusters is given by the

sign of $\tilde{\beta}_{c(s)g} \times \tilde{\beta}_{c(s')g}$, summed over the common SNPs between the two clusters , as illustrated in the schematic diagram in Figure 1. Finally, for each taxon we center $\beta \mathbf{G}^{(t)}$ to ensure both positive and negative genetic modulation, rather than systematic enrichment or depletion within the population.

We have provided a function within RITHMS, `calibrate_gen_effect()`, to help users evaluate the impact of σ_β on the distribution of taxa heritabilities. In practice, it is often reasonable to expect the majority of taxa to have heritabilities on the order of 0.1, with maximum values of no more than 0.5 (Zang et al., 2022).

Quantifying microbiota diversity. A variety of metrics exist to quantify the α-diversity (intra-sample diversity) from microbiota data. For a composition $\mathbf{p} = (p_1, \ldots, p_{n_b})$, where $\sum_1^{n_b} p_j = 1$, we consider the Shannon index $H^1(\mathbf{p})$ defined as $H^1(\mathbf{p}) = -\sum_{j=1}^{n_b} p_j \log(p_j)$ with the convention that $0 \log(0) = 0$. When computed directly from sequencing data, this index is based on species counts transformed to relative abundances and thus suffers from potential undersampling of rare species. In order to obtain counts from the relative abundances and mimick this sampling step, n_{ind} multinomial samplings $M(10000, (\mathbf{p_{i,1}}, \ldots \mathbf{p_{i,n_b}}))$ are performed, with $(\mathbf{p_{i,1}}, \ldots, \mathbf{p_{i,n_b}})$ the relative abundances of taxa for individual i, equivalent to the cutoffs on sequencing depth used in the dataset analyses described above. Note that 10 000 corresponds to the average sequecing depth in our motivating example but can be replaced with any value, *e.g.* the mean depth in the base population, or vector of values, *e.g.* samples from a sequencing depth distribution, whether synthetic (Gaussian, negative binomial, etc) or non-parametric (values in the base population). Numerous diversity indices have been proposed in the literature (Cazzolla Gatti et al., 2020; Chao et al., 2014), each capturing a distinct ecological specificity of the microbiota structure. While we focus here on the Shannon index as a popular index to account for differences in abundances, alternative metrics could be considered.

Transgenerational simulation of phenotypes. Phenotypes at generation (t) are simulated as the result of the combined effects of the microbiota and direct genetic effects following the model developed by Pérez-Enciso et al. (2021):

(1) $$\mathbf{y}^{(t)} = \boldsymbol{\alpha}^T \mathbf{G}^{(t)} + \boldsymbol{\omega}^T \mathbf{B}^{(t)} + \epsilon_y^{(t)},$$

with:

- $\boldsymbol{\alpha}$ the regression coefficients corresponding to the QTL effects on the phenotype ($1 \times n_g$),
- $\mathbf{G}^{(t)}$ the genotype values of all individuals at generation t ($n_g \times n_{\mathrm{ind}}$),
- $\boldsymbol{\omega}$ the regression coefficients corresponding to taxa effects on the phenotype ($1 \times n_b$),
- $\mathbf{B}^{(t)} = \mathrm{CLR}(\mathbf{M}^{(t)})$, the CLR-transformed relative abundance values for taxa of all individuals at generation t ($n_b \times n_{\mathrm{ind}}$),
- $\epsilon_y^{(t)} \sim \mathcal{N}(0, 1)$, univariate Gaussian noise.

Note that the variance of the Gaussian noise is set to 1 to ensure that changes in mean phenotypic values are expressed in units of standard deviations. In our simulation settings, we assume that all heritable taxa also have an effect on the phenotype; as such, the microbiota effect, corresponding to $\boldsymbol{\omega}^T \mathbf{B}^{(t)}$, also includes an indirect genetic effect.

Breeding values and heritability. Under this formulation, we define the Direct Breeding Value as $\mathbf{BV}_d^{(t)} = \alpha^T \mathbf{G}^{(t)}$, the Microbiota-mediated Breeding Value as $\mathbf{BV}_m^{(t)} = \omega^T \beta \mathbf{G}^{(t)}$ and the Total Breeding Value $\mathbf{BV}_t^{(t)} = \mathbf{BV}_d^{(t)} + \mathbf{BV}_m^{(t)}$. This $\mathbf{BV}_t^{(t)}$, containing all of the genetic effects on the phenotype, takes into account both the direct genetic effect ($\alpha^T \mathbf{G}^{(t)}$) due to the transmission of the genotype and the indirect microbiota-mediated ones ($\omega^T \beta \mathbf{G}^{(t)}$), due to the fraction of the microbiota that has an effect on the phenotype and is under genetic control.

From these quantities, it is possible to define a few quantities of interest: (1) the total heritability $h^2 = \left[\mathrm{var}\left(\alpha^T \mathbf{G}^{(t)} + \omega^T \beta \mathbf{G}^{(t)} \right) \right] / \mathrm{var}(\mathbf{y}^{(t)})$, (2) the direct heritability $h_d^2 = \mathrm{var}(\alpha^T \mathbf{G}^{(t)}) / \mathrm{var}(\mathbf{y}^{(t)})$, and (3) the microbiability $b^2 = \mathrm{var}(\omega^T \mathbf{B}^{(t)}) / \mathrm{var}(\mathbf{y}^{(t)})$.

Parameter calibration. The regression vectors α and ω are fixed across generations and calibrated on the base population. The calibration consists in rescaling ω based on α in order to reach user-specified values for the direct heritability h_d^2 and the microbiability b^2. If h_d^2 is set to 0, then all α coefficients are set to zero and the calibration only affects ω. Initial values $\tilde{\alpha}$ for the non-zero coefficients of α are sampled from a $\Gamma(0.4, 5)$ distribution and $\tilde{\omega}$ for the non-zero coefficients of ω are sampled from a $\Gamma(1.4, 3.8)$, as done in the Simubiome method (Pérez-Enciso et al., 2021), before rescaling takes place.

Selection. To select the individuals that will make up the breeding stock for the next generation, by default 30% of the males and 30% of the females are selected at each generation. This proportion reflects common practice in many breeding programs, as typical selection rates generally range between 20 % and 40 %, corresponding to a moderate selection intensity. These fractions can be modified via the parameters `size_selection_F` and `size_selection_M`. If no selection is specified, individuals are chosen at random. Otherwise, a user-specified selection criterion is used to rank individuals, and only a fraction (specified above) of the top performers are retained to reproduce and contribute to the next generation. The available selection criteria are:

- the microbiota effect, $\omega^T \mathbf{B}^{(t)}$, measuring the contribution of the microbiota to the phenotype at generation t,
- the Microbiota-mediated Breeding value, $\mathbf{BV}_m^{(t)}$, capturing the portion of the host's genetic effects that modulate the phenotype via their impact on microbiota composition,
- the Direct Breeding Value, $\mathbf{BV}_d^{(t)}$, quantifying the direct effect of the host genotype on the phenotype,
- the Total Breeding Value, $\mathbf{BV}t^{(t)}$, corresponding to the overall genetic expectation of the phenotype, combining both direct and microbiota-mediated genetic effects (in the absence of environmental or transmission effects), corresponding to classic genomic selection,
- the microbiota diversity, $\delta^{(t)}$, computed as the Shannon diversity index,
- a weighted index combining microbiota diversity and total breeding value, $\mathrm{wdiv}\delta^{(t)} + (1 - \mathrm{wdiv})\mathbf{BV}t^{(t)}$, with wdiv set by the user.

These quantities represent different components of the potential of an individual for transmitting traits to the next generation. The genetic breeding value captures the contribution of host genetics, whereas the microbial diversity value reflects the influence of the microbiota on phenotype. The combined metric integrates both sources of variation, taking into account the idea that host performance may rely on both genetics and microbial composition.

Results

In this section, we will explore a wide range of simulation scenarios to demonstrate the capabilities and features of RITHMS, from sanity checks to our initial use cases. These results were obtained on the dataset described in the *Real hologenomic data used as a base population* section. Unless otherwise specified, all scenarios use the following simulation parameters: $h_d^2 = 0.25$, $b^2 = 0.25$, $\sigma_\beta \times \sqrt{\text{QTL}_o} = 0.3$, $\sigma_m = 0.6$, $n_{\text{ind}} = 500$, $\lambda = 0.5$ and $n_{\text{gen}} = 5$. Other parameters are set to default values as described in the package documentation.

Simulated microbiota reflect expected structure

We first evaluate whether the simulated microbiota exhibit expected characteristics. The pairwise correlation matrix of simulated abundances (Figure 3A) shows that RITHMS successfully produces both strong intra-cluster genetic correlations as well as more modest inter-cluster anti-correlations, thanks to the set of partially overlapping QTLs between clusters. Likewise, increasing QTL effect sizes on taxa increases the heritability of taxa abundances, as expected (Figure 3B). The density plots are produced by `calibrate_gen_effect()` and are intended to guide the user in choosing an appropriate effect size to achieve a target distribution of taxa heritabilities. In this setting, a reasonable distribution of taxa heritabilities appears to roughly correspond to a value of $\sigma_\beta \times \sqrt{\text{QTL}_o} = 0.3$. We also confirm the impact of λ in modulating the relative importance of vertical and horizontal transmission (Figure 3C). When $\lambda = 0$, corresponding to no vertical transmission, offspring α-diversity is strongly correlated with that of the ambient microbiota (values averaged over 10 simulated datasets). As λ increases, so does the correlation between maternal and offspring microbiota α-diversity. In constrast, the correlation between paternal and offspring microbiota diversity is low for all values of λ. This is expected, as the sire microbiota does not directly contribute to that of its offspring. Finally, in the absence of selection or environmental filters, the distribution of α-diversity remains stable across generations (Figure 3D), as expected for communities evolving in a neutral framework.

Introduction of sporadic or sustained environmental effects

In breeding and selection programs, it is essential to account for fixed environmental effects, given their strong role in modulating an individual's phenotype. It is therefore important to verify that simulated transgenerational hologenomic data can correctly integrate such factors under a variety of plausible scenarios, such as short-term treatments or long-term diet effects. For the microbiota, as fixed environmental effects can be cumulated with varying effects on each taxa, RITHMS allows users to specify a (potentially sparse) θ matrix, corresponding to the environmental effect sizes on CLR-transformed taxa abundances. To illustrate this, we consider two scenarios introducing either a sporadic (Figure 4A-B) or sustained (Figure 4C-D) environmental effect, as would respectively be the case if a subset of individuals in one generation were administered antibiotics or if individuals in each generation were randomly assigned to different diet groups.

In the first case, we assume that half of the individuals in G1 are administered an antibiotic, provoking significant abundance changes across all taxa. The values for this effect were sampled from a normal distribution $\mathcal{N}(0, 5)$. This one-time environmental effect leads to a strong separation into two groups with very distinct microbiota compositions (Figure 4A) and constrasted
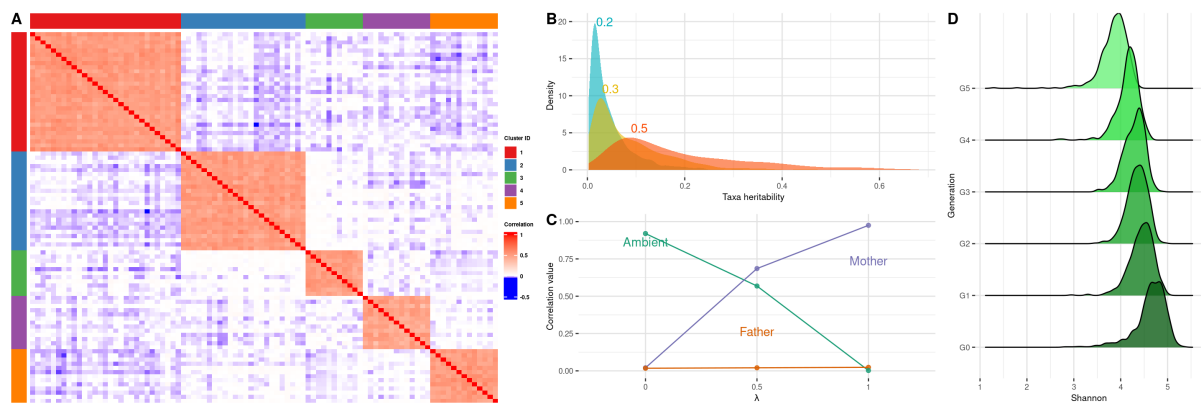
Figure 3 – Key characteristics of microbiota data simulated with RITHMS. (A) Pairwise correlation matrix of taxa abundances. Abundances were simulated assuming all taxa are under genetic control and distributed in five clusters (shown with color bars in the margins). Taxa are sorted based on the cluster they belong to. (B) Density plot of the distribution of taxa heritability for increasing genetic effect sizes ($\sigma_\beta \times \sqrt{\mathrm{QTL_o}}$), shown above each curve. (C) Correlation between offspring α-diversity (from G2) and that of its mother (purple), father (orange) or ambient microbiota (green) for increasing values of λ. Correlations are computed from a population of 500 offsprings and averaged over 10 repetitions. (D) Density plots of the distribution of α-diversity values in the base population (G0) and five consecutive generations (G1 to G5), in the absence of selection and environmental filters.

α-diversity, as evidenced by the bimodal distribution of α-diversity values in generation G1 (Figure 4B). In the absence of continued antibiotic intake after G1, the lower diversity observed for the antibiotic group is progressively attenuated in the following generations due to random mating, and the bimodality disappears, although the α-diversity is reduced on average compared to the base generation (*e.g.*, when comparing G3 and G0 in Figure 4B). Likewise, the strong group structure in microbiota compositions induced by the treatment progressively disappears in following generations, but the diversity of the overall population shifts towards that of the antibiotic-treated microbiota, suggesting long-lasting changes of the treatment. Similar trends were observed when varying the λ parameter, with stronger and longer-lasting antibiotic effects on diversity for higher values of λ (results not shown).

In the second case, we assume that individuals from each generation following the base population are randomly assigned to one of two diets, one of which favors abundances in 2 randomly chosen taxa clusters. To simulate a relatively modest effect on the CLR-scale, non-zero values of θ were drawn from a normal distribution with smaller variance than that of the previous case, $\mathcal{N}(0, 2)$. This sustained environmmental effect induces a progressive separation of the diet groups that becomes particularly marked at G3 (Figure 4C). As two taxa clusters are preferentially favored in one of the diet groups, with the effect accumulating across generations, we remark the emergence of a group with an increasingly large drop in diversity (Figure 4D).

Impact of genomic, microbiota and hologenomic selection strategies

In the previous sections, we showed that the microbiota simulated by RITHMS reflect expected characteristics in terms of inter- and intra-cluster genetic correlations among taxa, taxa heritability, vertical or horizontal transmission, as well as microbiota diversity across generations, in the presence or absence of environmental effects. We now turn our attention to phenotypes simulated from the transgenerational hologenomic data under the model in Equation (1). Two
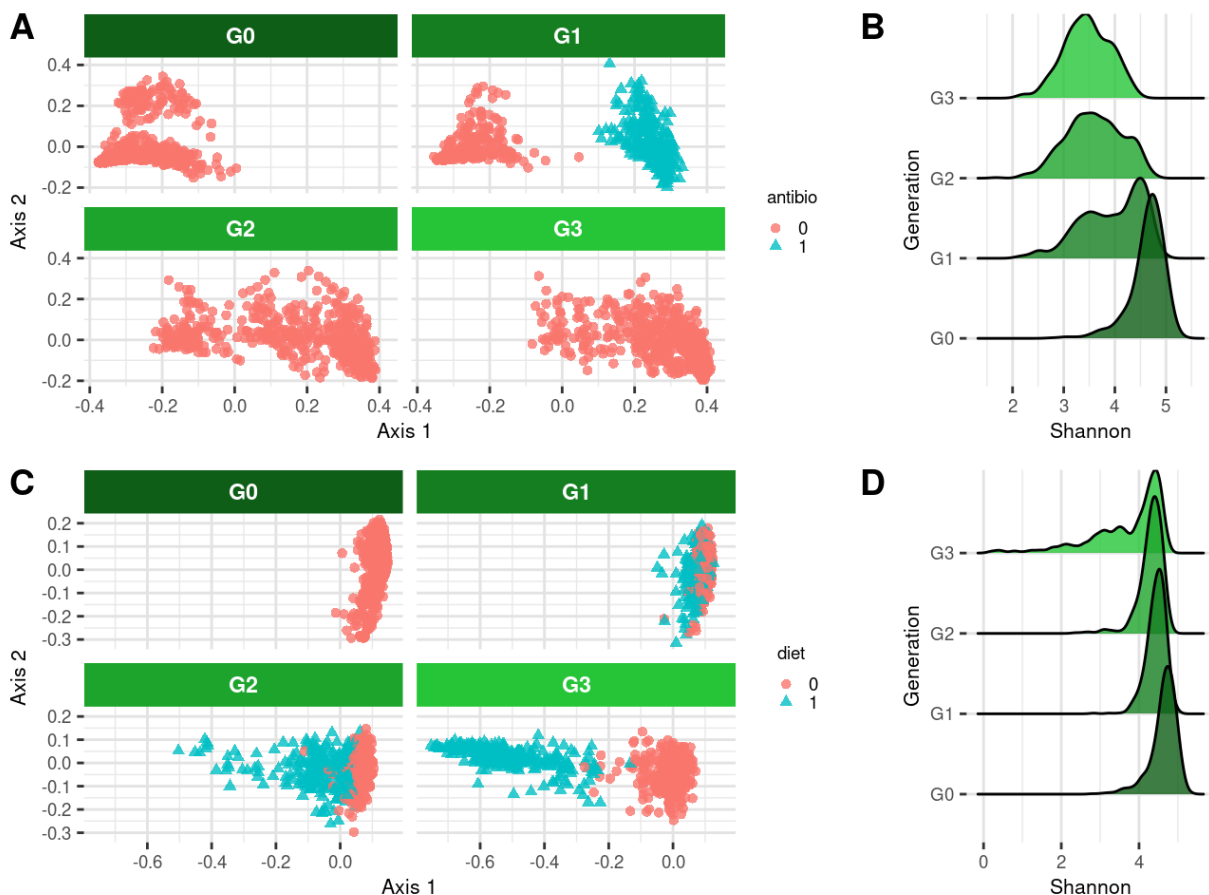
Figure 4 – Simulation of sporadic (top) and sustained (bottom) environmental effects in RITHMS. (A) Multidimensional scaling (MDS) of microbial abundance data (Bray-Curtis distances). Half the individuals at G1 (blue triangles) are subject to a sporadic antibiotic treatment. (B) Density plots of α-diversity values before (G0), during (G1) and after (G2 to G3) sporadic antibiotic treatment. (C) Multidimensional scaling (MDS) of microbial abundance data (Bray-Curtis distances). Starting from G1, half the individuals at each generation (blue triangles) are subject to a diet favoring two clusters of taxa.(D) Density plots of α-diversity values before (G0) and during (G1 to G3) sustained diet intervention.

critical user-provided parameters for RITHMS simulations are the direct heritability h_d^2 and microbiability b^2. In the absence of selection, we next sought to verify that the target values are reached and maintained across generations in the case of $h_d^2 = b^2 = 0.25$ (Figure 5A), corresponding to intermediate values and similar to those used in Pérez-Enciso et al. (2021). h_d^2 and b^2 were computed using the true values of α and ω and the simulated values of $\mathbf{G}^{(t)}$ and $\mathbf{M}^{(t)}$ at each generation. As α and ω are calibrated using the base population to achieve target heritability and microbiability (See the *Parameter calibration* section), it is no surprise that h_d^2 and b^2 are exactly at 0.25 for G0. In subsequent generations, the direct heritability varies only slightly around its target value, and we remark that the observed microbiability tends to be slightly lower than its target value.

Given that the direct heritability and microbiability appear to be reasonably maintained near their target values in the absence of selection, we next evaluate trends in phenotypic improvement as a function of four different selection strategies for varying values of h^2 and b^2 (Figure 5B): selection of 30% of males and 30% of females based on (i) no criterion (random), (ii) the total breeding value ($\mathbf{BV}_t^{(t)}$), (iii) the direct breeding value ($\mathbf{BV}_d^{(t)}$), or (iv) the microbiota effect value. We

observe that the phenotypic change is up to twice as large for higher values of direct heritability and microbiability ($h_d^2 = b^2 = 0.4$) as compared to lower values ($h_d^2 = b^2 = 0.05$). Microbiota selection outperforms the other modes of selection only when microbiability is large compared to the direct heritability ($b^2 = 0.4$ and h_d^2=0.05). Generally speaking, given the modest contribution of vertical transmission used here ($\lambda = 0.1$) that affects only the female path of transmission, hologenomic selection appears to act as a form of mass selection and provides little selection gain compared to genomic selection alone. As an indication, these results were obtained based on a total of 1800 simulated datasets (4 selection modes × 9 pairs of h_d^2 and b^2 values × 50 repetitions for each), using the pig hologenomic data described above as a base population, corresponding to 770 minutes of computational time with a maximum memory usage of around 1GB RAM on a laptop with 16 GB RAM (Intel(R) Core(TM) i5-1135G7 CPU @ 2.40GHz x 8). An implementation for parallelizing repeated simulations is available and demonstrated in the package vignette.
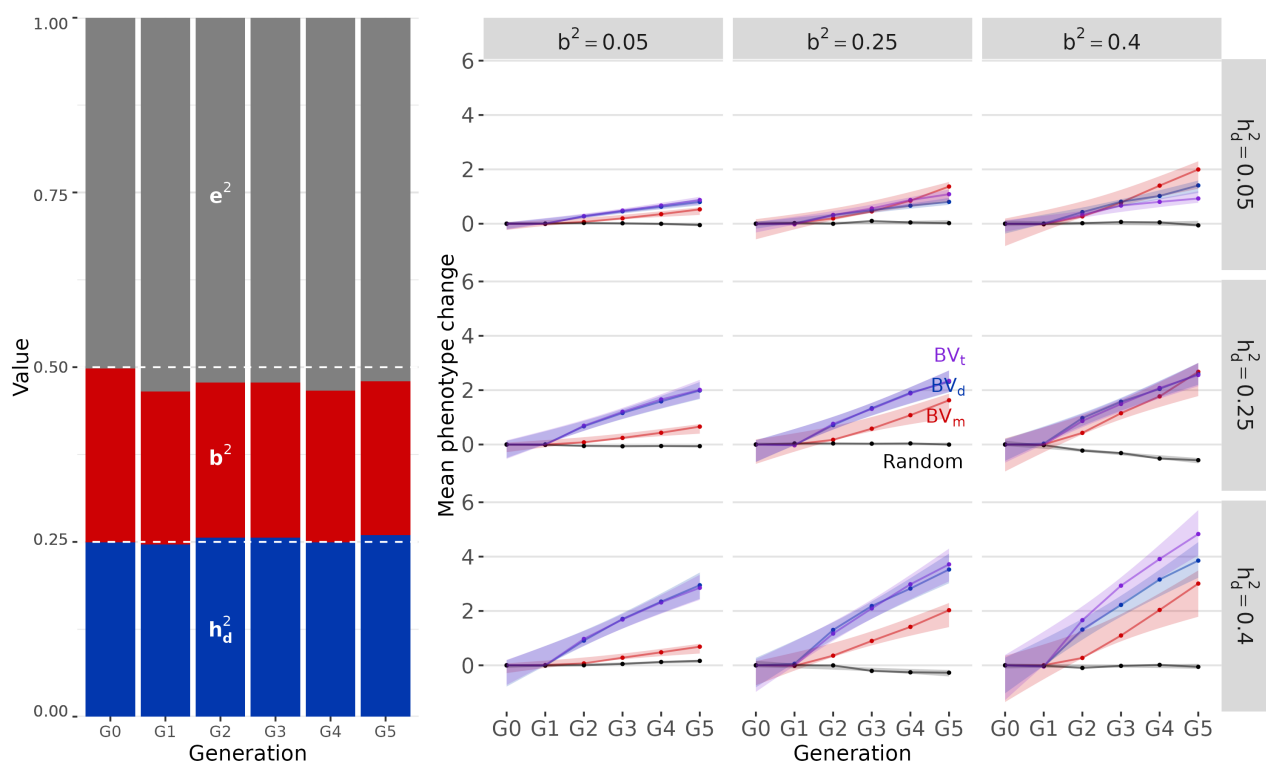

Figure 5 – Direct heritability and microbiability of RITHMS simulations under various selection strategies. (A) Observed direct heritability h_d^2 and microbiability b^2 (averaged over 50 simulated datasets) in a scenario with random selection and target values $h_d^2 = b^2 = 0.25$.(B) Mean phenotypic change across five generations, (averaged over 50 simulated datasets, shaded regions correspond to 95% confidence intervals) with $\lambda = 0.1$, according to various values of direct heritability (rows) and microbiability (columns) and different selection strategies: $\mathbf{BV}_d^{(t)}$ (direct breeding values, blue line), $\mathbf{BV}_m^{(t)}$ (microbiota breeding values, red line), $\mathbf{BV}_t^{(t)}$ (total breeding values, purple line), random selection of parents for the next generation (black line).

Constructing an index for mixed-objective selection

To demonstrate the flexibility and usefulness of RITHMS, we consider a practical case study of a simplified breeding program with a multi-trait objective: maximizing phenotypic change,

based on a quantitative trait of interest $\mathbf{y}^{(t)}$, while preserving microbial α-diversity. One way to achieve this is to use a selection score that combines both objectives into a single value. With access to hologenomic data at each generation, such a score can be constructed as a weighted combination of phenotypic change and diversity. Formally, we define our selection index as $\mathrm{w_{div}} \cdot \delta^{(t)} + (1 - \mathrm{w_{div}}) \cdot \mathbf{BV}_t^{(t)}$ (see the *Selection* section) as a linear combination of the microbial diversity $\delta^{(t)}$ and the total breeding value $\mathbf{BV}_t^{(t)}$, with weight $\mathrm{w_{div}} \in [0, 1]$. Note that $\mathrm{w_{div}} = 0$ corresponds to classic genomic selection. This index is used to identify the 30% of males and 30% of females constituting the breeding stock for the next generation.

Here, we leverage RITHMS to construct a simulation study to identify an optimal weight to achieve gains on both components in a reasonable number of generations. In particular, we simulated data over five generations to evaluate the impact of $\mathrm{w_{div}} \in \{0, 0.1, \dots, 1\}$ on changes in microbial diversity and phenotypic change, with direct heritability $h_d^2 = 0.25$, microbiability $b^2 = 0.25$, vertical transmission $\lambda = 0.5$ and $n_{\mathrm{ind}} = 500$ individuals per generation (Figure 6). Although there is considerable variability among simulated datasets, we remark that there is a tradeoff between mean phenotypic change and microbial diversity (*i.e.*, one comes at the expense of the other), which varies with $\mathrm{w_{div}}$. Larger weights ($\mathrm{w_{div}} = 0.8$ or 0.9) simultaneously achieve phenotypic improvement and increased microbial diversity after five generations. However, for these scenarios, phenotypic change is more modest than for scenarios that increasingly mimic classic genomic selection ($\mathrm{w_{div}} < 0.8$). These results suggest that a value of $\mathrm{w_{div}} = 0.6$ achieves phenotypic change comparable to classic genomic selection in this case study, while drastically limiting the loss of microbial diversity.

Impact of data subsampling on variance component estimation

To illustrate another possible use of RITHMS, we conducted a study of the estimation of variance components under different data availability scenarios. Such a study is of particular interest to evaluate the impact of different sampling strategies, for example to account for financial or experimental constraints. First, we simulated a population of 2500 hologenotyped and phenotyped individuals over 5 generations (500 per generation), with $\lambda = 0.5$. Then we estimated the direct heritability (h_d^2) and microbiability (b^2) with a GMBLUP model using the *BGLR* R package (Pérez and de los Campos, 2014) and linear scaled kernels of genotypes and microbiota, based on various subsets of individuals: (a) the full dataset (2500 individuals), (b) a random subset of 30% of individuals (750 individuals), (c) only breeding parents (750 individuals), and (d) only mothers (375 individuals). Figure 7 shows the distribution of the estimated variance components across 500 independent iterations for each subset.

As expected, the most accurate estimates were obtained when all individuals were hologenotyped. Random subsampling led to a systematic underestimation of both h_d^2 and b^2, reflecting the loss of information about the underlying genetic and microbial covariance structures. When restricting the analysis to mothers only, estimates remained close to those obtained with the full dataset, even though a substantial part (85%) of the information was omitted. This reflects the key role played by maternal microbiomes for phenotypic variance decomposition, notably due to the vertical transmission of the microbiota through mothers only. These results highlight the importance of a meaningful subsampling strategy for hologenomic selection, e.g. focusing on reproducing individuals, to mitigate the loss of accuracy associated with incomplete data collection. An interesting follow-up study could investigate cases where genotypes are collected on a large number of individuals and microbiota data on only a targeted subset; using all available

genotypes, even when microbiome data are missing, would represent a more realistic setting and powerful setup for future applications.
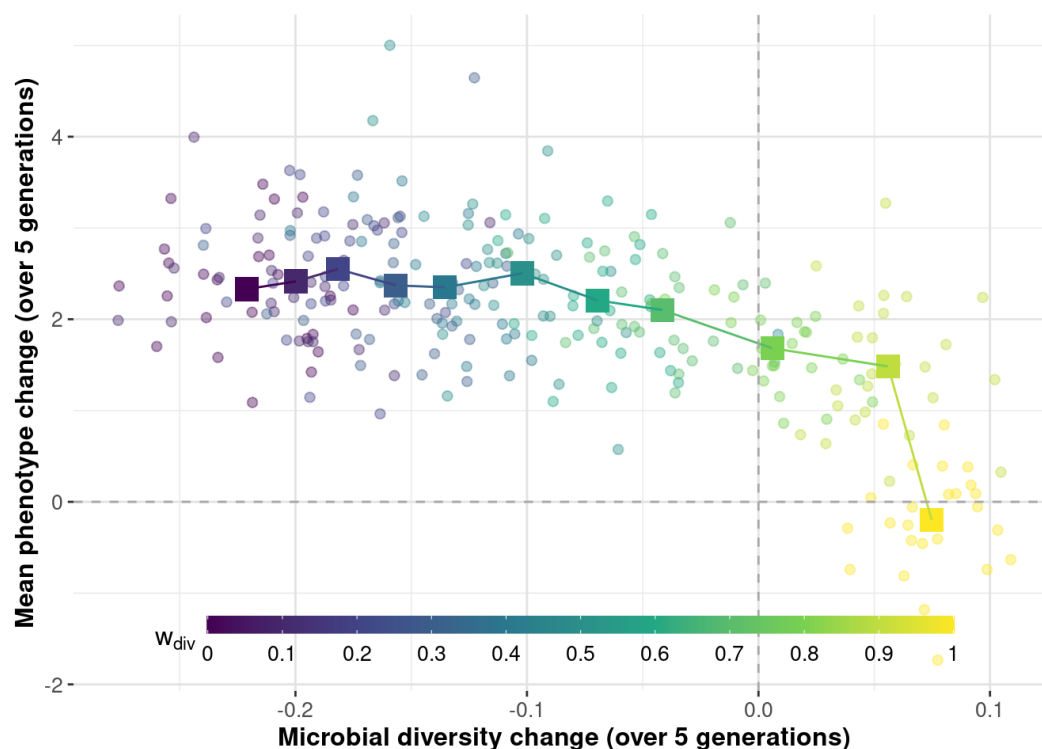

Figure 6 – Simulation-guided exploration of mixed selection index. Mean phenotype and microbial diversity changes from the base population (G0) to G5 as a function of w_{div}. The simulation is repeated 25 times for each value of w_{div}. Each simulation is shown as semi-transparent dots whereas square dots correspond to the mean computed over the 25 repetitions.

Discussion

In this work, we introduced a novel algorithm for simulating transgenerational hologenomic data, implemented in the R package RITHMS. Our tool expands the scope of existing genomic simulation methods (Gaynor et al., 2021; Pook et al., 2020) by adding a microbiota compartment and of existing hologenomic simulation methods (Pérez-Enciso et al., 2021) by enabling the simulation of multiple generations. In contrast to the only other transgenerational hologenomic simulator currently available, HoloSimR (Casto-Rebollo et al., 2025), RITHMS uses real data as input, structures the microbiota into taxa clusters and incorporates potential environmental covariates. RITHMS directly accounts for the structure and characteristics of the microbiota as well as its complex transmission mechanisms (from both the mother and the ambient environment, with filters linked to host genetics) and the impact of sporadic or sustained environmental covariates. It is possible to calibrate both (i) the size of genetic effects on the microbiota to obtain a realistic distribution of taxa heritability and (ii) the direct genetic and microbial effects to achieve target values of direct heritability and microbiability. Complex breeding schemes using the genome, the microbiota or the hologenome combined with different selection scores or sampling strategies were used to showcase the flexibility and usefulness of RITHMS. RITHMS is available as an R
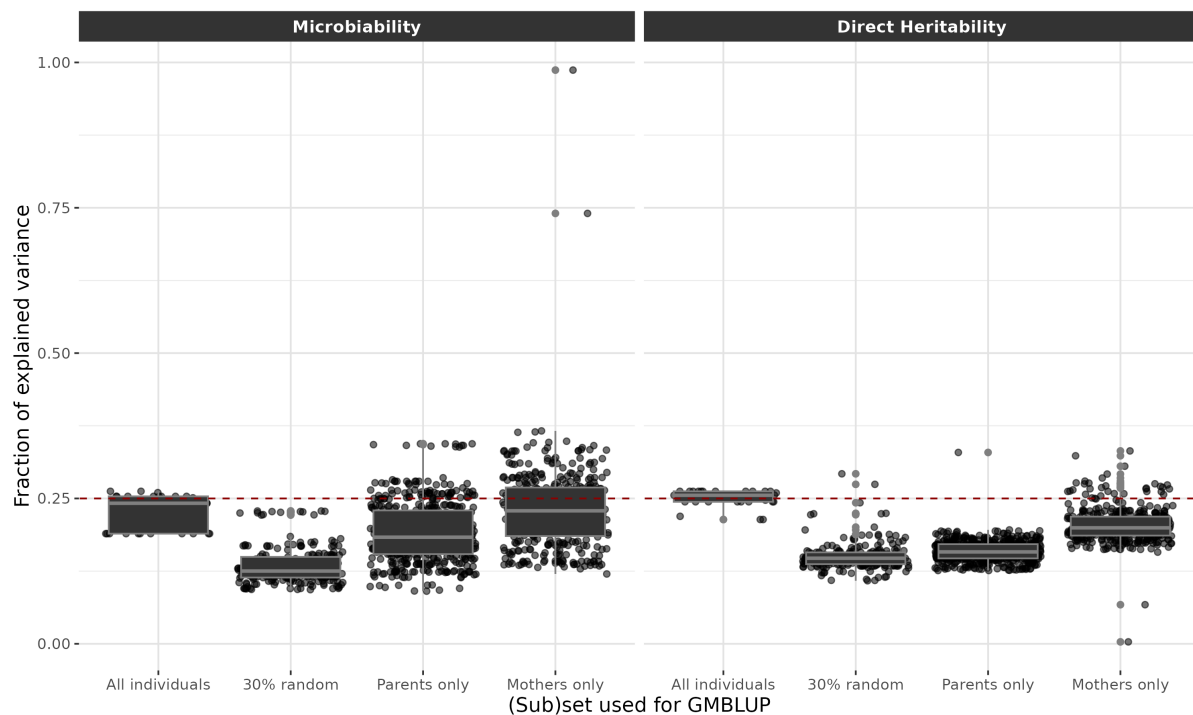
Figure 7 – Estimated variance components for microbiability (b^2) and direct heritability (h_d^2) obtained using GMBLUP over 500 simulation replicates under different subsampling scenarios: all individuals (2500 individuals), 30% random (750 individuals), parents only (750 individuals), and mothers only (375 individuals). The red dotted line corresponds to real simulated values ($b^2 = h_d^2 = 0.25$).

package, runs on a commercial laptop and is able to generate transgenerational hologenomic data ($n_g = 5000$, $n_b = 2000$ taxa, $n_{ind} = 500$ individuals) for five generations in a few seconds.

Our approach presents several limits and opportunities for future improvements. First, we remark on the slight negative bias we observed between the simulated and target values for microbiability b^2 (Figure 5A, from generation G1 onwards). Since taxa effects on the phenotype ω are calibrated on G0, we hypothesize that this bias originates from a small loss of α-diversity between G0 and G1, as the model cannot reproduce fully the complexity of the base population. Second, our simulation framework is based on a linear model, which has the advantage of being both interpretable and computationally tractable; however, it would be of interest to explore alternatives such as neural networks to introduce non-linearity into RITHMS. Third, our simulated microbiota correspond to snapshots in the lifetime of an animal that are intended for use in predictive models of hologenomic breeding values. However, the microbiota corresponds to a highly dynamic measure that evolves throughout an animal's life, and future work could consider a dynamic model to simulate the microbiota at different time points.

Likewise, it would be interesting to extend the RITHMS model to (i) account for microbial interactions with a non-diagonal covariance matrix for the noise component σ_m of the taxa abundances, (ii) allow for the inclusion of more complex environmental effects, and (iii) allow for the use of semi-complete, rather than fully paired, genomic and microbiota data to create the base population, which would enable RITHMS simulations to be calibrated on a datasets for which some samples lack genomic or microbiota data.

Finally, in future work we plan to extend the use of RITHMS to alternative hologenomic datasets, notably for a variety of species and experimental designs, and additional use cases for the evaluation of complex breeding schemes.

Acknowledgements

Fundings

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Data, script and code availability

RITHMS is an open-source package available on GitHub (https://github.com/SolenePety/RITHMS) and all data and code used for this study are available in Zenodo (https://doi.org/10.5281/zenodo.17629188; Pety, 2025). A subset of the original dataset is available and can be used to reproduce the figures of the paper in a vignette.

References

Aguilar Marín S, Betancur C, Mesa H, Isaza G, Jovel J (2020). Lower Methane Emissions Were Associated with Higher Abundance of Ruminal Prevotella in a Cohort of Colombian Buffalos. *BMC Microbiology* **20**, 16. https://doi.org/10.1186/s12866-020-02037-6.

Bruijning M, Henry LP, Forsberg SKG, Metcalf CJE, Ayroles JF (2022). Natural Selection for Imprecise Vertical Transmission in Host–Microbiota Systems. *Nature ecology & evolution* **6**, 77–87. https://doi.org/10.1038/s41559-021-01593-y.

Brulin L, Ducrocq S, Even G, Sanchez MP, Martel S, Merlin S, Audebert C, Croiseau P, Estellé J (2024). Short Communication: *Bifidobacterium* Abundance in the Faecal Microbiota Is Strongly Associated with Milk Traits in Dairy Cattle. *animal* **18(8)**, 101243. https://doi.org/10.1016/j.animal.2024.101243.

Casto-Rebollo C, Pocrnic I, Gorjanc G, Ibáñez-Escriche N (2025). HoloSimR: A Comprehensive Framework for Simulating Breeding Programs Including the Hologenome Scenario. *Journal of Animal Science*, Epub ahead of print. https://doi.org/10.1093/jas/skaf371.

Cazzolla Gatti R, Amoroso N, Monaco A (2020). Estimating and Comparing Biodiversity with a Single Universal Metric. *Ecological Modelling* **424**, 109020. https://doi.org/10.1016/j.ecolmodel.2020.109020.

Chaillou S, Chaulot-Talmon A, Caekebeke H, Cardinal M, Christieans S, Denis C, Hélène Desmonts M, Dousset X, Feurer C, Hamon E, Joffraud JJ, La Carbona S, Leroi F, Leroy S, Lorre S, Macé S, Pilet MF, Prévost H, Rivollier M, Roux D, et al. (2015). Origin and Ecological Selection of Core and Food-Specific Bacterial Communities Associated with Meat and Seafood Spoilage. *The ISME Journal* **9**, 1105–1118. https://doi.org/10.1038/ismej.2014.202.

Chao A, Chiu CH, Jost L (2014). Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics* **45**, 297–324. https://doi.org/10.1146/annurev-ecolsys-120213-091540.

Cortes-Macías E, Selma-Royo M, García-Mantrana I, Calatayud M, González S, Martínez-Costa C, Collado MC (2021). Maternal Diet Shapes the Breast Milk Microbiota Composition and Diversity: Impact of Mode of Delivery and Antibiotic Exposure. *The Journal of nutrition* **151**, 330–340. https://doi.org/10.1093/jn/nxaa310.

David I, Ricard A (2019). A Unified Model for Inclusive Inheritance in Livestock Species. *Genetics* **212**, 1075–1099. https://doi.org/10.1534/genetics.119.302375.

Derakhshani H, Fehr KB, Sepehri S, Francoz D, De Buck J, Barkema HW, Plaizier JC, Khafipour E (2018). Invited Review: Microbiota of the Bovine Udder: Contributing Factors and Potential Implications for Udder Health and Mastitis Susceptibility. *Journal of Dairy Science* **101**, 10605–10625. https://doi.org/10.3168/jds.2018-14860.

Déru V, Bouquet A, Hassenfratz C, Blanchet B, Carillier-Jacquin C, Gilbert H (2020). Impact of a High-Fibre Diet on Genetic Parameters of Production Traits in Growing Pigs. *Animal : an international journal of animal bioscience* **14**, 2236–2245. https://doi.org/10.1017/S1751731120001275.

Déru V, Bouquet A, Zemb O, Blanchet B, De Almeida ML, Cauquil L, Carillier-Jacquin C, Gilbert H (2022). Genetic Relationships between Efficiency Traits and Gut Microbiota Traits in Growing Pigs Being Fed with a Conventional or a High-Fiber Diet. *Journal of animal science* **100**. https://doi.org/10.1093/jas/skac183.

Evans B, Leighton F (2014). A History of One Health. *Revue Scientifique et Technique de l'OIE* **33**, 413–420. https://doi.org/10.20506/rst.33.2.2298.

Faux AM, Gorjanc G, Gaynor RC, Battagin M, Edwards SM, Wilson DL, Hearne SJ, Gonen S, Hickey JM (2016). AlphaSim: Software for Breeding Program Simulation. *The plant genome* **9(3)**. https://doi.org/10.3835/plantgenome2016.02.0013.

Galiez C (2025). A Tool to Help the Design of Breeding Procedures Accounting for Holobiont Host-Microbiome Interplay. *Peer Community in Mathematical and Computational Biology* **1**, 100417. https://doi.org/10.24072/pci.mcb.100417.

Gaynor RC, Gorjanc G, Hickey JM (2021). AlphaSimR: An R Package for Breeding Program Simulations. *G3 Genes|Genomes|Genetics* **11**, jkaa017. https://doi.org/10.1093/g3journal/jkaa017.

Ghataora JS, Ellis T (2025). Rewiring Holobiont Systems with Synthetic Biology. *Trends in Biotechnology*, Articles in press. https://doi.org/10.1016/j.tibtech.2025.09.017.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8**. https://doi.org/10.3389/fmicb.2017.02224.

Goldberg DE, Deb K (1991). A Comparative Analysis of Selection Schemes Used in Genetic Algorithms. In: *Foundations of Genetic Algorithms*. Ed. by GREGORY J. E. Rawlins. Vol. 1, pp. 69–93. https://doi.org/10.1016/B978-0-08-050684-5.50008-2.

Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE (2016). Genetic Determinants of the Gut Microbiome in UK Twins. *Cell host & microbe* **19**, 731–743. https://doi.org/10.1016/j.chom.2016.04.017.

Guerrero R, Margulis L, Berlanga M (2013). Symbiogenesis: The Holobiont as a Unit of Evolution. *International Microbiology: The Official Journal of the Spanish Society for Microbiology* **16**, 133–143. https://doi.org/10.2436/20.1501.01.188.

Leal SM (2001). Genetics and Analysis of Quantitative Traits, *American Journal of Human Genetics* **68**, 548–549. https://doi.org/10.1086/318209.

Legarra A, Lourenco DAL, Vitezica ZG (2018). Bases for Genomic Prediction. URL: https://alegarra.github.io/GSIP.pdf.

Li F, Li C, Chen Y, Liu J, Zhang C, Irving B, Fitzsimmons C, Plastow G, Guan LL (2019). Host Genetics Influence the Rumen Microbiota and Heritable Rumen Microbial Features Associate with Feed Efficiency in Cattle. *Microbiome* **7**, 92. https://doi.org/10.1186/s40168-019-0699-1.

Margulis L, Fester R (1991). *Symbiosis as a Source of Evolutionary Innovation: Speciation and Morphogenesis*. Ed. by MIT Press. Cambridge, MA, ISBN: 9780262519908. URL: https://mitpress.mit.edu/9780262519908/symbiosis-as-a-source-of-evolutionary-innovation/.

Martinez Boggio G, Meynadier A, Buitenhuis AJ, Marie-Etancelin C (2022). Host Genetic Control on Rumen Microbiota and Its Impact on Dairy Traits in Sheep. *Genetics Selection Evolution* **54**, 77. https://doi.org/10.1186/s12711-022-00769-9.

Martinez Boggio G, Monteiro HF, Lima FS, Figueiredo CC, Bisinotto RS, Santos JEP, Mion B, Schenkel FS, Ribeiro ES, Weigel KA, Peñagaricano F (2024). Host and Rumen Microbiome Contributions to Feed Efficiency Traits in Holstein Cows. *Journal of Dairy Science* **107**, 3090–3103. https://doi.org/10.3168/jds.2023-23869.

May F, Gerstner K, McGlinn DJ, Xiao X, Chase JM (2018). Mobsim: An r Package for the Simulation and Measurement of Biodiversity across Spatial Scales. *Methods in Ecology and Evolution* **9**, 1401–1408. https://doi.org/10.1111/2041-210X.12986.

Niu Q, Li P, Hao S, Zhang Y, Kim SW, Li H, Ma X, Gao S, He L, Wu W, Huang X, Hua J, Zhou B, Huang R (2015). Dynamic Distribution of the Gut Microbiota and the Relationship with Apparent Crude Fiber Digestibility and Growth Stages in Pigs. *Scientific Reports* **5**, 9938. https://doi.org/10.1038/srep09938.

Pang SEH, Slik JWF, Zurell D, Webb EL (2023). The Clustering of Spatially Associated Species Unravels Patterns in Tropical Tree Species Distributions. *Ecosphere* **14**, e4589. https://doi.org/10.1002/ecs2.4589.

Pérez P, de los Campos G (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics* **198**, 483–495. https://doi.org/10.1534/genetics.114.164442.

Pérez-Enciso M, Zingaretti LM, Ramayo-Caldas Y, de los Campos G (2021). Opportunities and Limits of Combining Microbiome and Genome Data for Complex Trait Prediction. *Genetics Selection Evolution* **53**, 65. https://doi.org/10.1186/s12711-021-00658-7.

Pety S (2025). RITHMS v1.0.0 - Stable Release Supporting Manuscript Version. Zenodo. https://doi.org/10.5281/zenodo.17629188.

Pook T, Schlather M, Simianer H (2020). MoBPS - Modular Breeding Program Simulator. *G3 (Bethesda, Md.)* **10**, 1915–1918. https://doi.org/10.1534/g3.120.401193.

Ross EM, Moate PJ, Marett LC, Cocks BG, Hayes BJ (2013). Metagenomic Predictions: From Microbiome to Complex Health and Environmental Phenotypes in Humans and Cattle. *PLOS ONE* **8**, e73056. https://doi.org/10.1371/journal.pone.0073056.

Rutayisire E, Huang K, Liu Y, Tao F (2016). The Mode of Delivery Affects the Diversity and Colonization Pattern of the Gut Microbiota during the First Year of Infants' Life: A Systematic Review. *BMC gastroenterology* **16**, 86. https://doi.org/10.1186/s12876-016-0498-0.

Shterzer N, Rothschild N, Sbehat Y, Dayan J, Eytan D, Uni Z, Mills E (2023). Vertical Transmission of Gut Bacteria in Commercial Chickens Is Limited. *Animal microbiome* **5**, 50. https://doi.org/10.1186/s42523-023-00272-6.

Theis KR, Dheilly NM, Klassen JL, Brucker RM, Baines JF, Bosch TCG, Cryan JF, Gilbert SF, Goodnight CJ, Lloyd EA, Sapp J, Vandenkoornhuyse P, Zilber-Rosenberg I, Rosenberg E, Bordenstein SR (2016). Getting the Hologenome Concept Right: An Eco-Evolutionary Framework for Hosts and Their Microbiomes. *mSystems* **1**, e00028–16. https://doi.org/10.1128/mSystems.00028-16.

Van den Boogaart KG, Tolosana-Delgado R (2008). "Compositions": A Unified R Package to Analyze Compositional Data. *Computers & Geosciences* **34**, 320–338. https://doi.org/10.1016/j.cageo.2006.11.017.

Weishaar R, Wellmann R, Camarinha-Silva A, Rodehutscord M, Bennewitz J (2020). Selecting the Hologenome to Breed for an Improved Feed Efficiency in Pigs—A Novel Selection Index. *Journal of Animal Breeding and Genetics* **137**, 14–22. https://doi.org/10.1111/jbg.12447.

Wirbel J, Essex M, Forslund SK, Zeller G (2024). A Realistic Benchmark for Differential Abundance Testing and Confounder Adjustment in Human Microbiome Studies. *Genome Biology* **25**, 247. https://doi.org/10.1186/s13059-024-03390-9.

Wu G, Zhao N, Zhang C, Lam YY, Zhao L (2021). Guild-Based Analysis for Understanding Gut Microbiome in Human Health and Diseases. *Genome Medicine* **13**, 22. https://doi.org/10.1186/s13073-021-00840-y.

Zang XW, Sun HZ, Xue MY, Zhang Z, Plastow G, Yang T, Guan LL, Liu JX (2022). Heritable and Nonheritable Rumen Bacteria Are Associated with Different Characters of Lactation Performance of Dairy Cows. *mSystems* **7**, e00422–22. https://doi.org/10.1128/msystems.00422-22.

Zilber-Rosenberg I, Rosenberg E (2008). Role of Microorganisms in the Evolution of Animals and Plants: The Hologenome Theory of Evolution. *FEMS Microbiology Reviews* **32**, 723–735. https://doi.org/10.1111/j.1574-6976.2008.00123.x.