Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

Phenotype Heritability in holobionts: An Evolutionary Model

Saúl Huitzil, Santiago Sandoval-Motta, Alejandro Frank, Maximino Aldana

Abstract. Many complex diseases are expressed with high incidence only in certain populations. Genealogy studies determine that these diseases are inherited with a high probability. However, genetic studies have been unable to identify the genomic signatures responsible for such heritability, as identifying the genetic variants that make a population prone to a given disease is not enough to explain its high occurrence within the population. This gap is known as the missing heritability problem. We know that the microbiota plays a very important role in determining many important phenotypic characteristics of its host, in particular, the complex diseases for which the missing heritability occurs. Therefore, when computing the heritability of a phenotype it is important to consider not only the genetic variation in the host but also in its microbiota. Here we test this hypothesis by studying an evolutionary model based on gene regulatory networks. Our results show that the holobiont (the host plus its microbiota) is capable of generating a much larger variability than the host alone, greatly reducing the missing heritability of the phenotype. This result strongly suggests that a considerably large part of the missing heritability can be attributed to the microbiome.

S. Huitzil

Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, Avenida Universidad
s/n, Colonia Chamilpa, Cuernavaca, Morelos, Mexico
Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City,
Mexico
e-mail: saulhuitzil@gmail.com

S. Sandoval-Motta

Instituto Nacional de Medicina Genómica, Mexico City, Mexico
Consejo Nacional de Ciencia y Tecnología, Cátedras CONACyT, Mexico City, Mexico
Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City,
Mexico
e-mail: santiago.sandoval.m@gmail.com

A. Frank

Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, Mexico City, Mexico
Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City,
Mexico
e-mail: frank@nucleares.unam.mx

M. Aldana

Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México, Avenida Universidad
s/n, Colonia Chamilpa, Cuernavaca, Morelos, Mexico
Centro de Ciencias de la Complejidad, Universidad Nacional Autónoma de México, Mexico City,
Mexico
e-mail: max@icf.unam.mx

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

1 Introduction

One of the main objectives of the Human Genome Project, the HapMap Project, and later of the GWAS studies, has been to understand the genetic architecture Responsible for the emergence of complex diseases [6,7,34]. For a long time, through genealogy studies, it has been estimated that such complex diseases have a strong heritability determined by genetic components [23]. However, this type of study has the inconvenience that it is limited to a small number of individuals. By contrast, thanks to high-throughput techniques such as GWAS, the size of the population under study could be considerably increased because the degree of kinship between individuals can be determined through their genetic variants (SNPs).

The first GWAS study was performed in 2002 [25]. Since then, this method has been very successful in associating thousands of genetic variants with different phenotypic traits [4]. However, the effect of these genetic variants on a particular phenotype is very small [22] and, consequently, they can explain only a small fraction of the heritability of the associated phenotype. This is because the fraction of phenotypic variance that can be attributed to additive genetic factors is considerably smaller than expected when we consider the genetic variants associated with a given phenotype by GWAS [13].

The gap between the heritability observed through genealogy studies and the one measured from GWAS is known as the *missing heritability*, and for some phenotypes can be as large as 60% [33]. A typical example is human height. On the one hand, familial studies have estimated that this phenotype has a heritability larger than 80%. On the other hand, through GWAS about 700 SNPs have been associated with human height. However, these SNPs can explain only 20% of its heritability [37]. The most used strategy to fill up the missing heritability gap has been to find more and more genetic variants associated with a particular phenotype in order to increase its estimated heritability [13]. In 2010, Yang et al. computed the heritability produced by 29,4831 SNPs, which allowed them to estimate the heritability of human height around 45% [38]. Nonetheless, if the missing heritability problem were to be solved by increasing the number of SNPs, then we would have to consider so many genetic variants that we would still be unable to understand the genetic architecture behind the emergence of complex diseases and phenotypes. Several solutions have been proposed, such as taking into account epigenetic and epistatic effects, the effect of non-coding RNAs in gene regulation, or an overestimation of the heritability itself [22, 23, 29, 41, 14, 10]. However, there is no consensus yet in the scientific community as to how to solve this problem.

During the last decade, numerous studies have shown the importance of the microbiome in determining important phenotypic traits of the host organism [3]. The microbiota composition and functionality have been correlated with many complex diseases, allergies, neurological disorders, metabolism of antibiotics and other drugs, to mention just a few [5, 17, 30, 19, 36]. Additionally, there is evidence suggesting that phenotypic traits can be transmitted (or inherited) from the parents to their offspring through the microbiota [18]. For instance, the phylogenetic congruence (phylosymbiosis) between the host organism and its microbiota suggests that a set of microorganisms have been inherited throughout evolution [21, 24]. Furthermore, it is known that

the metabolic response of several host organisms to the microbiota is conserved throughout evolution across different vertebrates, including the zebrafish, mice, and humans [26]. The microbiota participates in practically all important biological tasks of the host [5]. Given the strong symbiotic interactions between the host and its microbiota, and the evidence showing that microorganisms can be transmitted both vertically and horizontally from parents to offspring, Rosenberg, and his collaborators formulated the hypothesis of the holobiont as a unit of selection in evolution [39]. This hypothesis states that in order to understand the evolution of the phenotypic traits of a given organism one has to consider the hologenome (the genome of the host and the genomes of all of its microbes) as a unit of selection [39]. This means that changes in the phenotype of a host organism can be produced by mutations in the host's genome, in the genes of its microbes (the host's microbiome), or in both, and to understand these phenotypic changes one has to consider genetic variation in the entire hologenome. This may be particularly true for complex phenotypes for which missing heritability exists. Therefore, it is reasonable to assume that, when computing the heritability of a phenotype, the genetic variants of the entire hologenome have to be taken into account. This hypothesis was proposed in previous works [28, 32], but to our knowledge, no measurements have been performed so far validating or refuting this idea. Here we present an evolutionary model, based on Boolean networks, showing quantitatively that a large part of the missing heritability is recovered when the genetic variants are computed using the entire hologenome rather than just the genome of the host.

2 Boolean Network model

We choose to work with Boolean networks as representations of the genomes of both the host and its microbes. These networks were proposed by S. Kauffman in 1969 as models for gene regulation, and have proven to reproduce the gene expression patterns observed experimentally for several organisms [2,9]. A Boolean network consists of N nodes representing the genes in the genome. The state of expression of these genes is encoded in a set of variables $\{g_1, g_2, \dots, g_N\}$ such that $g_n = 1$ if the n^{th} gene is expressed, or $g_n = 0$ if it is not expressed. The value of each g_n is determined by a particular set of k_n other genes in the network, which we represent as $\{g_{n_1}, g_{n_2}, \dots, g_{n_{k_n}}\}$ (see Fig.1a). We will refer to this set as the *inputs* or *regulators* of g_n , and to k_n as its *input connectivity* which can be different from one gene to another. Once each gene in the networks has been assigned with a set of regulators, the network dynamics is determined by the simultaneous update of the state of all the genes according to the equation

$$g_n(t + \Delta t) = F_n(g_{n_1}(t), g_{n_2}(t), \dots, g_{n_{k_n}}(t)), \quad (1)$$

where $F_n(g_{n_1}, g_{n_2}, \dots, g_{n_{k_n}})$ is a Boolean function of k_n variables and Δt is the average time it takes for a gene to respond to changes in its regulators. For networks of real organisms, the set of regulators of each gene g_n is determined from experimental observations and its Boolean function is constructed by hand according to the activating or inhibitory nature of its regulators (see Fig. 1b). Here we are interested in the general evolutionary properties of holobionts and not on any particular organism. Therefore, we are going to work with Boolean networks constructed in the following way: (a) the number k_n of inputs of each gene g_n will be randomly chosen from a

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

predefined probability function $P_I(k)$, which has average K and variance σ_k^2 ; (b) the k_n inputs of each gene are chosen randomly from anywhere in the network with uniform probability; (c) each Boolean function F_n will be constructed randomly such that for each of the 2^{k_n} configurations of its k_n arguments, $F_n = 1$ with probability p and $F_n = 0$ with probability $1 - p$.

Fig. 1 (a) Illustrative example of a network with 4 genes, each one having $K = 2$ inputs. The arrows indicate the regulatory interactions. (b) Specific example of the Boolean functions associated with each of the four genes in the network. (c) Structure of a Boolean function. The first column represents the index of the input configuration, ranging from $i = 1$ to $i = 4$. The second and third columns show the configurations of the inputs. The last column is the value of the function acquired on the i^{th} configuration of the inputs. (d) The Boolean DNA is constructed by concatenating the values of all the Boolean functions over all of their configurations. The genotype G of the network is the particular sequence the Boolean DNA consists of.

Once the inputs and Boolean functions have been assigned to every gene in the network, they do not change throughout time. What changes, is the state of expression of each gene according to Eq.(1). In other words, the inputs to each gene and its Boolean function are randomly assigned at the beginning of the simulation, when the network is constructed. After that, the topological structure of the network (the specific regulators of each gene) and the Boolean functions do not change in time. This is important since, as the network structure does not change throughout time, we can define a “Boolean DNA sequence” (BDNA) that characterizes the network. For that, let us note that a Boolean function with K inputs (arguments) has 2^K values, one for each configuration of its inputs (Fig. 1b). These 2^K configurations can be indexed according to their Boolean value, as Fig.1c shows. Thus, for instance, a function with $K = 3$ inputs has $2^3 = 8$ configurations, ranging from configuration $\{000\}$ (the three inputs are off) to configuration $\{111\}$ (the three inputs are on), with all the intermediate configurations in between. As Fig.1c shows,

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

these configurations can be indexed from $i = 1$ to $i = 2^K$. Let us define f_n^i as the value of the n^{th} Boolean function F_n on the i^{th} configuration of its inputs (see Fig.1c). Then, the Boolean DNA of the network is defined as the concatenation of the values of all the Boolean functions in the network for all of their input configurations (see Fig. 1d):

$$BDNA = \{f_1^1, f_1^2, f_1^3, \dots, f_1^{2^{k_1}}, f_2^1, f_2^2, f_2^3, \dots, f_2^{2^{k_2}}, \dots, f_N^1, f_N^2, f_N^3, \dots, f_N^{2^{k_N}}\}. \quad (2)$$

A network with N genes, with the n^{th} gene having k_n inputs, has a Boolean DNA with $\Omega = \sum_{n=1}^N 2^{k_n}$ digits or loci. If k_n is the same for all the genes, let's say $k_n = K$ for every $n = 1, 2, \dots, N$, then $\Omega = N \times 2^K$. The dynamics of the network are determined by both its structural topology and its Boolean DNA. However, given a particular network structure, the Boolean DNA completely determines the dynamics of the network. Mutations in the Boolean DNA can change the network dynamics and its gene expression pattern. From now on we will indistinctly refer to the Boolean DNA of the network defined in Eq. 2 also as the *genome* of the network, and to a particular Boolean sequence as its *genotype* (see Fig. 1d).

3 Genetic variants in a population

Since we want to study the heritability of phenotypes in a population, we have to define the phenotype in a quantitative way as well as the variants of that phenotype. To do this, let us consider a network with N genes connected in a specific way and with a specific Boolean DNA (see Fig.2). This network, which we will denote as H_0 , represents the genome (or part of the genome) of a given organism. Under some environmental conditions a subset of N_S genes of this network, which we will refer to as the *transduction nodes* and represent by $S = \{g_1^S, g_2^S, g_3^S, \dots, g_{N_S}^S\}$, have an expression pattern determined by the function $R_0(t)$, defined over a time interval $0 \leq t \leq T$. A gene expression pattern is always associated with a phenotypic trait. Therefore, we will refer to $R_0(t)$ as the *phenotype* of the network. This function may characterize the response of the transduction genes $S = \{g_1^S, g_2^S, g_3^S, \dots, g_{N_S}^S\}$ to external metabolites, stressors or environmental conditions (see Fig. 2). It can be for instance the phenotype needed to metabolize sugar or fatty acids. Now imagine that the environment permanently changes, and for the organism to adapt to the new environmental challenge, the subset of genes S has to change its expression pattern to a new function $F(t)$. This new function $F(t)$ is the *desired phenotype* for the organism to properly survive in the new environment. Therefore, the *actual* phenotype, encoded in the gene expression pattern $R_0(t)$, has to transform into the desired phenotype $F(t)$. This change will not happen immediately after the environmental change has occurred. Instead, it will be through a series of mutations and partial adaptations that the actual phenotype $R_0(t)$ will transform into the desired phenotype $F(t)$. Therefore, we will have to train the network H_0 to perform the desired function $F(t)$. To do this, we will implement a standard evolutionary algorithm in which a population of networks evolves through mutations and selection so that at each generation, only the networks that better approach the target function $F(t)$ are the ones that survive and pass to the next generations.

Fig. 2 (a) Under external signals, a subset of the genes in a cell has to change its expression pattern in order to adapt to the environmental conditions. **(b)** A network representing the genome (or part of the genome) of a cell. The square nodes (in blue) are the transduction nodes that will respond to the external signal (a metabolite, stressor, etc.). These nodes generate a pattern of expression determined by the function $R(t)$, which in this work is just the simple sum of the state of the nodes. **(c)** The desired phenotype $F(t)$ is an arbitrary function defined over an interval of $0 < t \leq T$. In general, for a randomly constructed network, its actual phenotype $R(t)$ is quite different from the desired phenotype $F(t)$.

The function $R_0(t)$ is the phenotype of the network H_0 , and is determined by both its structure and its Boolean DNA (i.e., its genotype). The phenotype $R_0(t)$ has to transform into the desired phenotype $F(t)$. To implement this transformation we will produce mutations in the Boolean DNA

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

by keeping the structural topology of the network fixed. We measure the adaptation of the network to the new environmental conditions through the mean squared error ξ_0 between the actual phenotype $R_0(t)$ and the desired phenotype $F(t)$:

$$\xi_0 = \frac{1}{T} \sum_{t=1}^T (R_0(t) - F(t))^2. \quad (3)$$

If $\xi_0 = 0$, the network is perfectly adapted to the new task, as the actual phenotype $R_0(t)$ of the network is identical to the desired one $F(t)$. By contrast, large values of ξ_0 correspond to a poor adaptation, and the larger the value of ξ_0 , the poorer the adaptation.

	Loci																Phenotype
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
G_0	0	1	1	0	0	0	0	1	0	1	1	1	1	0	0	1	ξ_0
$G(v_1)$	1	1	1	0	0	0	0	1	0	1	1	1	1	0	0	1	$\xi(v_1)$
$G(v_2)$	0	0	1	0	0	0	0	1	0	1	1	1	1	0	0	1	$\xi(v_2)$
$G(v_3)$	0	1	0	0	0	0	0	1	0	1	1	1	1	0	0	1	$\xi(v_3)$
$G(v_4)$	0	1	1	1	0	0	0	1	0	1	1	1	1	0	0	1	$\xi(v_4)$
$G(v_5)$	0	1	1	0	1	0	0	1	0	1	1	1	1	0	0	1	$\xi(v_5)$
$G(v_6)$	0	1	1	0	0	1	0	1	0	1	1	1	1	0	0	1	$\xi(v_6)$
$G(v_7)$	0	1	1	0	0	0	1	1	0	1	1	1	1	0	0	1	$\xi(v_7)$
$G(v_8)$	0	1	1	0	0	0	0	0	1	1	1	1	1	0	0	1	$\xi(v_8)$
$G(v_9)$	0	1	1	0	0	0	0	1	1	1	1	1	1	0	0	1	$\xi(v_9)$
$G(v_{10})$	0	1	1	0	0	0	0	1	0	0	1	1	1	0	0	1	$\xi(v_{10})$
$G(v_{11})$	0	1	1	0	0	0	0	1	0	1	0	1	1	0	0	1	$\xi(v_{11})$
$G(v_{12})$	0	1	1	0	0	0	0	1	0	1	1	0	1	0	0	1	$\xi(v_{12})$
$G(v_{13})$	0	1	1	0	0	0	0	1	0	1	1	1	0	0	0	1	$\xi(v_{13})$
$G(v_{14})$	0	1	1	0	0	0	0	1	0	1	1	1	1	1	0	1	$\xi(v_{14})$
$G(v_{15})$	0	1	1	0	0	0	0	1	0	1	1	1	1	0	1	1	$\xi(v_{15})$
$G(v_{16})$	0	1	1	0	0	0	0	1	0	1	1	1	1	0	0	0	$\xi(v_{16})$
$G(v_3, v_6, v_{12})$	0	1	0	0	0	1	0	1	0	1	1	0	1	0	0	1	$\xi(v_3, v_6, v_{12})$

Fig. 3 Illustrative example of the genetic variants for the genome of the network shown in Fig.1. This genome has $\Omega = 16$ loci, indicated at the top of the matrix. The next row shows the genotype G_0 of the original network H_0 with respect to which the variants are defined. The subsequent rows show the particular genotypes $G(v_i)$ containing the variants v_i , which are indicated in blue. Variant v_i consists of a single mutation (SNP) occurring at the i^{th} position (locus) of the genome G_0 of the original network H_0 . The very last row shows a genotype $G(v_3, v_6, v_{12})$ that contains the three variants v_3, v_6 and v_{12} . The column to the far right just illustrates that each variant v_i corresponds to a particular error (phenotype) ξ_i .

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

Since the network H_0 was constructed randomly, its phenotype $R_0(t)$ will, in general, be quite different from the desired phenotype $F(t)$ (see Fig. 2c). For $R_0(t)$ to approach $F(t)$ we have to mutate the genome of the network. This allows us to define the *variants* of the network as follows. Let us define G_0 as the genotype of the network H_0 , namely, the particular Boolean sequence that characterizes the genome of H_0 (like the particular Boolean sequence in Fig.1d). The variant v_i is a single mutation of G_0 occurring at its i^{th} digit (locus). The variant v_i is thus a Single Nucleotide Polymorphism (SNP) occurring at the i^{th} locus of the genome. We will denote as $G(v_i)$ the genotype that contains the variant v_i . In other words, $G(v_i)$ is a Boolean sequence almost identical to the genotype G_0 of the original network H_0 except that it has a SNP at the i^{th} locus of the genome (see Fig. 3). Clearly, for a network with N genes, each having k_n input connections, there are $\Omega = \sum_{n=1}^N 2^{k_n}$ of such variants, $v_1, v_2, \dots, v_\Omega$ (see Fig.3). In this section and the following one, we will work with networks with $N = 50$ nodes, each with $K = 2$ regulators, which produces $\Omega = 200$ loci in the Boolean DNA and the same number of variants.

We have pointed out before that changes in the Boolean DNA of the network may change its dynamics. Therefore, even one mutation in the genotype of the network can change its phenotype $R_0(t)$. Therefore, to each variant v_i there corresponds a particular phenotype $R_i(t)$. Let us denote as $\xi(v_i)$ the error associated with the variant v_i , which is defined in an analogous way as in Eq. 3:

$$\xi(v_i) = \frac{1}{T} \sum_{t=1}^T (R_i(t) - F(t))^2. \quad (4)$$

It should be noted that $\xi(v_i)$ is a quantitative measure of how well the phenotype $R_i(t)$ approximates the desired phenotype $F(t)$. For instance, if $F(t)$ were the phenotype required to metabolize sugar, then $\xi(v_i) \approx 0$ would mean that the organism with the genotype $G(v_i)$ is healthy, whereas a large value of $\xi(v_i)$ would mean that variant v_i is associated with diabetes. Different variants may produce different values of the error $\xi(v_i)$, which could be interpreted as variability in the diabetes phenotype. Therefore, we can consider $\xi(v_i)$ as the quantitative measure of the phenotype $R_i(t)$ that corresponds to the genotype $G(v_i)$, which ultimately is associated with variant v_i . Since the error function $\xi(v)$ is a quantitative measure of the phenotype, (ot how well the network performs the desired phenotype), we will indistinctly refer to $R(t)$ and $\xi(v)$ as the phenotype of the network.

Analogously, we will represent as $G(v_{i_1}, v_{i_2}, \dots, v_{i_n})$ the genotype obtained from G_0 by simultaneously implementing the n variants $v_{i_1}, v_{i_2}, \dots, v_{i_n}$, which are SNPs occurring at positions $i_1, i_2, \dots, i_n$ of the genome (see Fig. 3). The phenotype corresponding to this genotype will be denoted as $R_{i_1, i_2, \dots, i_n}(t)$, and its corresponding error as $\xi(v_{i_1}, v_{i_2}, \dots, v_{i_n})$, which is computed in a similar way as in Eq. 4.

Some of the variants $v_1, v_2, \dots, v_\Omega$ will have an associated error larger than ξ_0 (the error of the original network H_0), while some other variants will have an error smaller than ξ_0 . To measure the

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

effect of the variant v_i on the adaptation of the network to the desired phenotype $F(t)$, we define the error difference $\delta\xi_i$ as

$$\delta\xi_i = \xi(v_i) - \xi_0. \quad (5)$$

If $\delta\xi_i < 0$ then the network containing the variant v_i is better adapted to the desired phenotype than the original network H_0 , whereas the opposite happens when $\delta\xi_i > 0$. When $\delta\xi_i = 0$ then variant v_i is neutral. Analogously, we can define the effect of the genotype $G(v_{i_1}, v_{i_2}, \dots, v_{i_n})$ (with n variants) on the adaptation of the network as $\delta\xi_{i_1, i_2, \dots, i_n} = \xi(v_{i_1}, v_{i_2}, \dots, v_{i_n}) - \xi_0$. Fig.4a shows a plot of the effects $\delta\xi_i$ for all the variants $v_1, v_2, \dots, v_{200}$ for a network with $N = 50$ and connectivity $K = 2$, while Fig. 4b shows the same data ordered in the increasing order of $\delta\xi_i$. Note that, while most of the variants are neutral, some variants produce considerably large effects (up to 50%), both positive and negative. This means that only one SNP can bring the network very close to the desired phenotype $F(t)$, or very far from it.

One of the main arguments trying to explain the missing heritability of phenotypes is that the effects of the different variants on the phenotype are not additive [40, 11, 12]. Nonlinear interactions between these variants, known as epistatic effects, could be hiding the heritability of phenotypes. While this may be the case, here we show that nonlinear interactions between the variants are not enough to explain the missing heritability, for even when these epistatic interactions are negligible, the missing heritability persists.

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

Fig. 4 (a) Plot of the normalized effect $\delta\xi_i/\xi_0$ of the variants v_i as a function of the variant index i , with $1 \leq i \leq 200$, for a network with $N = 50$ and connectivity $K = 2$. **(b)** Same data as in the previous panel but ordered in increasing order. Note that some variants can have relatively large effects, ranging from -33% to almost 50% of the original error ξ_0 . Note also that most of the variants are neutral, with $\delta\xi_i \approx 0$.

4 Nonlinear (epistatic) effects

In the model proposed by Yang et al. in 2010 (Ref. [38]; see also the article by Eskin in Ref. [11]), to measure the heritability of a quantitative phenotype, linear regression is performed based on the following equation

$$\xi_j = \bar{\xi} + \sum_{i=1}^{\Omega} g_{ij}\beta_i + e_j, \quad (6)$$

where ξ_j represents the value of the phenotype of the j^{th} individual in the population, $\bar{\xi}$ is the average value of the phenotype over the entire population, β_i is the size of the effect of the i^{th} variant v_i on the phenotype, and g_{ij} is a matrix that contains the variants of each individual, so

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

its entries acquire the values 0, 1 or 2 depending on whether none of the alleles of the j^{th} individual have the variant v_i , only one allele has it, or both alleles have the variant. Finally, e_j is an error inherent in the measurement of the phenotype that is associated with the effect of the environment. It is clear that Eq. 6 assumes that the effects of the different variants are additive. In this section, we test whether or not the additivity hypothesis is true for the network model we are analyzing.

Fig. 5 Distribution $P(n)$ of the number of variants n per genome in the population for a variant probability threshold $\delta_v = 0.025$ (panel **a**) and $\delta_v = 0.1$ (panel **c**). In the first case, the average number of variants per individual is $\bar{n} \approx 2.5$, whereas in the second case it is $\bar{n} \approx 10$. Panels **b** and **d** show the distribution $P(\Delta_{NL})$ of nonlinear effects Δ_{NL} for the corresponding populations in the previous two panels. Note that in both cases $P(\Delta_{NL})$ has a sharp maximum at $\Delta_{NL} = 0$, which shows that in our model, the epistatic (nonlinear) interactions between individual variants are not important.

We construct a population of networks by making $N_p = 10,000$ identical copies of H_0 , which is a random network with $N = 50$ nodes, each one having $K = 2$ regulators. Therefore, the Boolean DNA has $\Omega = N \times 2^K = 200$ loci. Then, with probability q_i we mutate the i^{th} locus of the genome of each network in the population ($1 \leq i \leq \Omega$). Thus, a mutation in the first position of the genome

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

(variant v_1) will occur in $N_P \times q_1$ networks in the population, a mutation in the second position (variant v_2) will occur in $N_P \times q_2$ networks, etc. The probabilities $q_1, q_2, \dots, q_\Omega$ are chosen randomly in the interval $0 < q_i \leq \delta_v$, where δ_v is the upper bound for the mutation probabilities. By increasing the threshold δ_v more mutations will be produced in the genomes of the networks. The reason for choosing different probabilities for the different loci in the Boolean DNA is to allow the different variants $v_1, v_2, \dots, v_\Omega$ to occur with different frequencies in the population. Note that the genome of one network in the population can accumulate several mutations, i.e., it can contain several variants $v_{i_1}, v_{i_2}, \dots, v_{i_n}$. Figs. 5a and 5b show the probability $f(n)$ that the genotype of a randomly chosen network in the population has n variants, for $\delta_v = 0.025$ and $\delta_v = 0.1$, respectively. In the first case the average number of variants in each genome is $\bar{n} = 2.5$ whereas in the second case it is $\bar{n} = 10$.

To determine the nonlinear interaction of the variants we consider a genotype $G(v_{i_1}, v_{i_2}, \dots, v_{i_n})$ that has n variants. We know the error $\xi(v_{i_1}, v_{i_2}, \dots, v_{i_n})$ associated with this genotype because we can measure it. Analogously, we know the individual errors $\xi(v_{i_1}), \xi(v_{i_2}), \dots, \xi(v_{i_n})$ associated with each of the individual variants $v_{i_1}, v_{i_2}, \dots, v_{i_n}$. To determine whether these individual variants contribute linearly or not to the phenotype $R_{i_1, i_2, \dots, i_n}(t)$ that corresponds to the genotype $G(v_{i_1}, v_{i_2}, \dots, v_{i_n})$, we define the *linear contribution* $\xi_{i_1, i_2, \dots, i_n}^L$ of the individual variants $v_{i_1}, v_{i_2}, \dots, v_{i_n}$ as

$$\xi_{i_1, i_2, \dots, i_n}^L = \xi_0 + \sum_{m=1}^n \delta \xi_{i_m}. \quad (7)$$

The linear contribution $\xi_{i_1, i_2, \dots, i_n}^L$ is an approximation to the real phenotype $\xi(v_{i_1}, v_{i_2}, \dots, v_{i_n})$ produced by the genotype $G(v_{i_1}, v_{i_2}, \dots, v_{i_n})$. If the nonlinear interactions between the individual variants $v_{i_1}, v_{i_2}, \dots, v_{i_n}$ are indeed very strong, then $\xi_{i_1, i_2, \dots, i_n}^L$ will be very different from $\xi(v_{i_1}, v_{i_2}, \dots, v_{i_n})$, whereas if these nonlinear interactions are not so strong, then $\xi_{i_1, i_2, \dots, i_n}^L \approx \xi(v_{i_1}, v_{i_2}, \dots, v_{i_n})$. Therefore, a measure of the nonlinearity of the interactions between different variants is determined by the quantity

$$\Delta_{NL} = \xi(v_{i_1}, v_{i_2}, \dots, v_{i_n}) - \xi_{i_1, i_2, \dots, i_n}^L. \quad (8)$$

Figs. 5b and 5d show the probability distribution function $P(\Delta_{NL})$ for the same populations used to generate the data in Figs. 5a and 5c, respectively. Note that in both cases $P(\Delta_{NL})$ has a sharp maximum centered at $\Delta_{NL} = 0$, which means that the nonlinear interactions between individual variants are not quite important in our model. This is particularly true for the case in which there is just a small number of SNPs per genome, as in Fig. 5a, with an average of $\bar{n} = 2.5$ variants in each Boolean DNA. This represents a SNP occurrence probability of 0.0125 per locus (2.5/200). In the human genome this probability is even lower by at least one order of magnitude, as the SNP occurrence probability in the human genome is 0.0007 per nucleotide (on average 7 SNPs

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

per 10 Kb [1, 27]). We do not claim that the results reported in Fig.5 can be extrapolated to the human genome (or to any other organism). In our model, and just in our model, the effect of most of the variants can be very well reproduced by a linear contribution of the effects of the corresponding individual variants. This is a remarkable result given that the network dynamics determined by Eq. 1 are highly nonlinear.

5 GWAS

For the sake of clarity, it is useful to keep in mind the analogy of the desired phenotype $F(t)$ as the capability to metabolize sugars. Taking into account that the error function $\xi(v_{i_1}, v_{i_2}, \dots, v_{i_n})$ is a quantitative measure of how well the organism with the genotype $G(v_{i_1}, v_{i_2}, \dots, v_{i_n})$ expresses the desired phenotype, then $\xi = 0$ would correspond to healthy organisms whereas $\xi \neq 0$ would correspond to diabetic ones. In this analogy, we would be interested in the variants associated with this particular disease, namely, the variants that most increase the value of the phenotype ξ . (From Fig.4 it is clear that there is a symmetric behaviour between the positive and negative values of $\delta\xi$; therefore, equivalent results would be obtained if we were looking for the variants that most decrease the value of ξ). In our model, we can exactly measure the value of the phenotype $\xi(v_i)$ corresponding to each variant v_i , as well as the phenotype $\xi(v_{i_1}, v_{i_2}, \dots, v_{i_n})$ associated to any combination of variants $v_{i_1}, v_{i_2}, \dots, v_{i_n}$. However, in real life we cannot know the phenotype $\xi(v_i)$ that each individual variant v_i produces because such variants generally do not occur isolated in the genome. Therefore, one has to estimate the contribution $\xi(v_i)$ that each individual variant v_i has on the phenotype under analysis. This allows us to identify the variants that are relevant to the phenotype in question. Those are variants that appear with a significantly large frequency among the population that presents the phenotype. This is where the Genome Wide Association Studies (GWAS) enters into play. The main objective of GWAS is to determine the variants that are significantly associated with the phenotype of interest. The main result of GWAS is to associate a p -value $p(v_i)$ to each variant v_i . The variants with the smallest p -values will be the ones that are significantly associated with the phenotype. The natural question arises: how small the p -value has to be for the corresponding variant to be associated with the phenotype? As a rule of thumb, people in the community always choose a threshold $p_T = 0.05$ for the p -value to be significant. If $p(v_i) < p_T$, then the corresponding variant v_i is accepted. Otherwise, it is rejected. Another fact to keep in mind is that, in practice, to perform a GWAS analysis not every individual in the population is tested. It is only a small fraction of the population whose DNA variants are analyzed. The size of the sampled population may have an important influence in the GWAS results, a phenomenon known as the *sampling size effect* [12, 11]. Some authors claim that by increasing the sampling size (the number of tested individuals), the results GWAS yields would be more trusty.

To simulate a GWAS analysis in our model, we consider a population consisting of $N_p = 10,000$ networks, each with $N = 50$ and connectivity $K = 2$. Variant v_i has been implemented with probability q_i in the genotype of the networks, with $0 < q_i < \delta_v = 0.025$ and $1 \leq i \leq 200$. This would result in a population similar to the one used to generate the data reported in Figs. 5a and 5c. To perform GWAS on this population, we extract a subpopulation with $N_{sub} = 500$ networks,

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

which are the ones that have the largest phenotypes ξ (in our analogy, these networks would correspond to clear cases of diabetes). In this subpopulation, the frequency $\hat{q}_i$ corresponding to the variant v_i may be different from the frequency q_i of this same variant in the original population. If $\hat{q}_i$ is significantly larger than q_i , then one can think that v_i is a variant associated with the phenotype. By contrast, if $\hat{q}_i$ is significantly smaller than q_i , then one can conclude that v_i is not associated with the phenotype. GWAS provides, through the p -value, the level of significance of the over-representation (or under-representation), of the variant v_i in the subpopulation that exhibits the phenotype.

Fig. 6a shows the normalized phenotype $\delta\xi_i/\xi_0 = (\xi(v_i) - \xi_0)/\xi_0$ in increasing order for a subpopulation $N_{sub} = 500$ networks and $\Omega = 200$ variants, while Fig.6b shows the p -values corresponding to these variants. Clearly, the GWAS analysis in our model is detecting the variants v_i with the highest values of the phenotype $\xi(v_i)$ (the larger the value of $\xi(v_i)$, the smaller the corresponding p -value).

The results presented so far indicate two important aspects of our model: (1) epistatic effects (nonlinear interactions) are negligible when the number of variants in the genome is small, and (2) GWAS effectively detects the variants significantly associated with the phenotype under consideration (in this case, large values of the error $\xi(v_i)$). In the next section we will show that the missing heritability problem is not necessarily a consequence of nonlinear effects nor of undersampling in the GWAS analysis. Instead, it may be a consequence of not taking into account the microbiota of the organism in the computation of the heritability, as the microbiota can be fundamental for the occurrence of some phenotypes.

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

Fig. 6 (a) Effect size $\delta \xi_i$ for the 200 variants v_i in a population of $N_p = 10,000$ networks, each with $N = 50$ nodes and connectivity $K = 2$. The data are plotted in the increasing order of $\delta \xi_i$. **(b)** A subpopulation of $N_{sub} = 500$ networks (5%) is extracted. These networks are the ones with the largest effect sizes $\delta \xi_i$. The graph shows the p -value, computed through a GWAS analysis, that correlates the occurrence frequency $\hat{q}_i$ of variant v_i in the subpopulation with the corresponding effect size $\delta \xi_i$. It can be seen that the lowest p -values correspond to the largest effect sizes, which indicates that GWAS is correctly detecting the variants that contributed the most to the phenotype.

6 Holobiont evolution and the missing heritability problem

So far we have not implemented any evolutionary algorithm to train the network H_0 to perform the desired task $F(t)$. We have just generated genetic and phenotypic variability in a population of networks (initially all of them identical to H_0) by implementing by hand the different variants $v_1, v_2, \dots, v_\Omega$ with some given probabilities. In this section, in order to compute the heritability of the phenotype, we implement an evolutionary algorithm to actually train (evolve) the network H_0 to perform the desired task $F(t)$. In what follows we will refer to H_0 as the *host network*, in analogy with the host organism in a holobiont.

Fig. 7 Schematic representation of a holobiont in our model. The holobiont consists of a host network and a microbial network, which can interact through regulatory connections between them (broken lines). In this example, the host network receives two regulatory connections from the microbial network, whereas the microbial network receives three from the host network. The host network still has to evolve to acquire the desired phenotype $F(t)$, as in Fig. 2. The difference now is that the microbial network will “help” in the evolution of the host network.

Following Huitzil et al. [16], the training of the host network H_0 will be assisted by another network, M , which we will call the *microbial network*. This means that during the evolutionary process, the host network can acquire regulatory connections from the microbial network and vice versa (see Fig. 7). The main difference between the host and microbial networks are their mutation rates, with a mutation rate $\mu_h = 0.001$ for the host network and $\mu_m = 0.01$ for the microbial network. The reason for this is that microbes can generate variability at a rate that is at least ten times larger than for the cells in a host organism (plants or animals). In our model, the host network H_0 represents the host organism, whereas the microbial network M represents its microbiota. The regulatory connections between H_0 and M represent the metabolic and genetic interactions between the host and its microbiota. Therefore, we will refer to the entire network made up of H_0 and M as the *holobiont* [39]. It is important to mention that, although the host network is interacting with the microbial network, the target function $F(t)$ is defined only for the host network and therefore it is just the phenotype $R_0(t)$ of the host network that will be used to train it. Therefore, the microbial network can be considered only as an auxiliary mechanism to help the host network reach its goal. The details of the evolutionary algorithm can be found in [16]. The important thing to mention here is that we have shown that the training of the host network is faster and more efficient with the help of the microbial network than without it and that most of the variability of the holobiont resides in the microbial network (due in part to its increased mutation rate).

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

The main idea in this section is to compute the heritability of the phenotype $R(t)$ (which is quantitatively measured through the error function ξ) in two different ways: (i) by computing the genetic variability of the host network only, and (ii) by computing the genetic variability of the whole holobiont (host and microbial networks). As we will see, the missing heritability is considerably larger in the first case than in the second one.

The evolutionary model consists of a population of $N_p = 1000$ holobionts. Initially, the host and microbial networks in each holobiont have $N = 50$ nodes, each node with $K = 2$ regulatory connections. At each generation, each node in each network is mutated with probability μ_h for the host network and μ_m for the microbial network. Once a node has been chosen for mutation, the mutations consist of the following: (i) Adding a new regulatory interaction (input connection); (ii) rewiring an existing regulatory interaction; (iii) changing the value of one entry of the Boolean function. Mutations (i) and (ii) can occur between nodes within the same network, or between one node in the host network and the other node in the microbial network. This last possibility is what makes the two networks develop regulatory interactions between them (see Fig. 7).

Let us denote as $R_n(t)$ and ξ_n the phenotype of the n^{th} holobiont in the population and its corresponding error, respectively. At each generation, we choose the best 100 holobionts in the population (the ones with the lowest values of ξ_n) to pass to the next generation. Then, we replicate these holobionts by making 10 copies of each one in order to restore the population to its original size $N_p = 1000$. Then we repeat the entire process (mutation, selection, replication, etc.). We do this for several generations until the average population error $\bar{\xi} = \frac{1}{N_p} \sum_{n=1}^{N_p} \xi_n$ becomes smaller than a threshold $\varepsilon = 1$. This means that on average most of the holobionts in the population are well adapted to the desired phenotype $F(t)$. We stop the simulation as soon as the condition $\bar{\xi} < \varepsilon$ is satisfied. However, this algorithm eventually produces a population with almost no variability, namely, a population in which almost all the holobionts are identical, a phenomenon known as *purifying selection* [8, 20, 15]. Therefore, in order to generate variability in the population, we proceed as follows. After the condition $\bar{\xi} < \varepsilon$ has been fulfilled, we continue the evolutionary process for 5 more generations. But now, in each one of these 5 generations we mutate only the Boolean DNA of the networks, which makes their corresponding errors to change: some errors will increase while some others will decrease. Then, at each one of these 5 generation, we allow the n^{th} holobiont to pass to the next generation with a probability $p_n = C/\xi_n$ (C is a proportionality constant). Thus, holobionts with a large error (poorly adapted) still have a probability, although small, to continue through the next generation. In this way, the final population will have more variability than if we were always to choose the best holobionts and replicate them.

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

Fig. 8 (a) Illustrative example of five Boolean genotypes G_1, G_2, G_3, G_4 and G_5 corresponding to five individuals in a population. Each genome has 10 loci. The consensus sequence, which is shown at the bottom of the table, is constructed using a simple majority rule at each locus. The variants (or SNPs) of each individual are those loci whose value is different from the corresponding one in the consensus sequence. These values are highlighted in blue in the figure. Since the genotypes of the 5 individuals are different from each other, their phenotypes ξ_i will usually be different from each other as well. (b) Table showing the occurrence frequency q_i of each variant v_i in the population. Less frequent variants are highlighted in red. (c) Knowing the consensus sequence of the population, all the information in (a) can be summarized by just indicating the variants of each individual.

At the end of this process all the holobionts in the population have the same structural topology but different genotypes, all of the same length. (The genome of a holobiont is the concatenation of the genome of its host network and the genome of its microbial network). Since through the evolutionary process regulatory connections within and between the host and microbial networks were added or rewired, different genes in the final networks will have a different number of input connections. We will denote as Ω_h, Ω_m and Ω_{HL} the length of the genomes of the host network, the microbial network, and the entire holobiont, respectively, with $\Omega_{HL} = \Omega_h + \Omega_m$. Since all the holobionts in the final population have the same structural topology but different Boolean DNAs,

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

it is possible to align their genotypes and determine a consensus sequence in the same way as one does with the genetic sequences of real organisms (see Fig. 8). Once the consensus sequence of the population has been determined, the variants are defined as mutations (SNPs) of the consensus sequence. There are Ω_h variants for the host network, Ω_m variants for the microbial network, and of course $\Omega_{HL} = \Omega_h + \Omega_m$ variants for the holobiont, which we will denote as $\{v_1, v_2, \dots, v_{\Omega_h}, v_{\Omega_h+1}, v_{\Omega_h+2}, \dots, v_{\Omega_{HL}}\}$. In the final population each variant v_i ($i = 1, \dots, \Omega_{HL}$) occurs with a frequency q_i . Rare variants in the population are defined as those that satisfy the condition $q_i < \delta_f$, where δ_f is a parameter chosen in the interval $[0.001, 0.01]$. We vary δ_f in this interval to analyze the effect of taking into account (or not) the less common variants in the computation of the heritability, finding no significant changes in the results when δ_f varies in the interval mentioned above.

The strict sense heritability h^2 of the phenotype is defined as [11,38]

$$h^2 = \sigma_g^2 / \sigma_P^2, \quad (9)$$

where σ_g^2 and σ_P^2 are the genotypic and phenotypic variances (or variabilities) of the population, respectively. The phenotypic variance σ_P^2 in our model is easily computed, as it is just the variance of the phenotype ξ over all the holobionts in the population:

$$\sigma_P^2 = \frac{1}{N_P} \sum_{n=1}^{N_P} (\xi_n - \bar{\xi})^2. \quad (10)$$

The genotypic variance σ_g^2 is more difficult to compute, as it has to do with the variants that are associated with the desired phenotype. Therefore, one has to determine first, though GWAS, the variants that correlate with the phenotype and then compute the variability of those variants throughout the population. Nonetheless, we can see that in the genetic variance σ_g^2 of our holobiont model there are two contributions, one coming from the genome of the host network and the other other from the genome of the microbial network. We will denote these contributions as $\sigma_{g,h}^2$ and $\sigma_{g,m}^2$, respectively. Therefore, $\sigma_g^2 = \sigma_{g,h}^2 + \sigma_{g,m}^2$ and the heritability in Eq.9 can be written as

$$h^2 = (\sigma_{g,h}^2 + \sigma_{g,m}^2) / \sigma_P^2. \quad (11)$$

Defining the host heritability and the microbial heritability as $h_h^2 = \sigma_{g,h}^2 / \sigma_P^2$ and $h_m^2 = \sigma_{g,m}^2 / \sigma_P^2$, respectively, then the heritability of the holobiont is $h^2 = h_h^2 + h_m^2$. Since both h_h^2 and h_m^2 are positive, it is clear that the two following conditions are satisfied: $h_h^2 \leq h^2$ and $h_m^2 \leq h^2$. If when computing the genetic variance σ_g^2 only the genome of the host organism is considered, as it is usually the case, then we will obtain a value h_h^2 which could be considerably smaller than the real value h^2 , unless the microbial heritability h_m^2 is negligibly small compared to the host heritability h_h^2 . But this is not a sensible hypothesis, in view that the microbiota is strongly correlated with the

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

emergence of some phenotypes (like diabetes). In our network model, the microbial network M strongly participates in the training of the host network H_0 to achieve its desired phenotype $F(t)$. The influence of the microbial network M on the training of the host network H_0 is stronger the more regulatory connections occur between them. We will denote as K_{ex} the number of regulatory connections from the microbial network to the host network, and refer to them as the *external host regulations*.

Fig. 9 Host and microbial heritabilities, h_h^2 and h_m^2 respectively, as functions of the number K_{ex} of host external regulations. Only for small values of K_{ex} (weak host-microbe interaction) the heritability of the host network is larger than the heritability of the microbial network. However, as K_{ex} increases (strong interaction), h_m^2 increases and very rapidly becomes larger than h_h^2 .

The details of the algorithm to compute the heritability h^2 in our model is presented in the Appendix. It is based in Eq. 6 and closely follows the method reported in Eskin 2015 and Yang 2010 [38, 11]. The results of this computation are presented in Fig.9, which shows the host and microbial heritabilities, h_h^2 and h_m^2 respectively, as functions of the number of external host regulations, K_{ex} . These results were computed for the final population of holobionts obtained from the evolutionary algorithm described in this section. Note from Fig.9 that when $K_{ex} = 0$, i.e. when there is no interaction between the host and microbial networks, almost all the contribution to the heritability h^2 of the phenotype comes from the heritability h_h^2 of the host network, as should be expected. However, when K_{ex} increases the interaction between the host and microbial networks becomes stronger. In this case, the heritability h_h^2 of the host network decreases while the heritability h_m^2 of the microbial network increases. Even for $K_{ex} \approx 2.5$, which corresponds to a relatively weak interaction between the host and microbial networks, the microbial heritability h_m^2 starts to surpass the host heritability h_h^2 . For $K_{ex} = 10$ (a strong interaction) almost all the

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

contribution to the heritability h^2 comes from the microbial network. In this strong-interaction case h_h^2 is very small with the result that $h^2 \approx h_m^2$.

7 Discussion and conclusions

The missing heritability problem has been a matter of intense debate for more than one decade. The fact that many phenotypes that are transmitted across generations cannot be significantly correlated with a particular set of genetic variants has generated many possible explanations. Among those explanations three stand out [12]. First, it has been proposed that epistatic (nonlinear) effects emerging from the interactions between these genetic variants may be hindering the identification of significant correlations between the genetic variants and the phenotype under consideration. Second, there is the undersampling problem, which consists in that rare variants that have a strong effect on the phenotype are not being detected due to the combination of two factors: these variants occur with very low frequency in the population and the size of the sampled subpopulation is too small. Third, there are many common variants whose combined effect on the phenotype is very strong, but whose individual effects are too small to be detected individually. While these answers to the missing heritability problem might be true, there is one aspect that has not been considered when measuring the heritability of the phenotype of a given organism, which is the genetic variability of its microbiota.

More than one decade ago, the pioneering work by Turnbaugh and his coworkers [31] showed that the microbiota can have a strong influence on the phenotypic traits of its host organism. This is particularly true for many of the phenotypes that researchers have tried to correlate with genetic variants, such as obesity, diabetes, cancer, metabolic syndrome, etc. It has also been shown that the microbiota can be transmitted across generations. Furthermore, there is evidence that the microbiota can evolve together with its host organism, strongly influencing (or even substituting) some of the genetic and metabolic functions of the host organism. Therefore, it is natural to assume that, when computing the heritability of some phenotypic trait, (particularly one that is strongly influenced by the microbiota), one does not only have to take into account the genotypic variance of the host organism, but also the genotypic variance of its microbiota. This is the approach that we have adopted in this work.

We have presented an evolutionary algorithm based on the Boolean gene regulatory network model proposed by S. Kauffman in 1969, which has proven to accurately reproduce the gene expression patterns experimentally observed for several organisms. The objective is to train a population of networks to perform a predefined task $F(t)$, which represents the phenotype of the networks. To do this training, we mutate the Boolean functions of the networks (their genotypes), which introduces genetic variability in the population (all the networks have the same structural topology but slightly different Boolean functions). The main advantage of working with this model is that we can exactly measure the effect that each variant has on the phenotype, which in turn allows us to simulate a GWAS analysis in order to compute the correlations between the genetic variants and the desired phenotype. From this analysis, three important results are obtained. First, most of the variants have a very small effect on the phenotype, but there are a few variants that

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

have a strong effect (Fig. 4). Second, when several variants are present in the genotype of one individual, their effects on the phenotype are mostly additive (Fig. 5). Therefore, epistatic (nonlinear) effects can be ignored when computing the contribution of several variants to the phenotype. This is an interesting result given that the dynamics determined by Eq. 1 are highly nonlinear. And third, GWAS effectively reveals the variants that have a strong effect on the phenotype (Fig. 6). These results indicate that all the conditions are met in our model to obtain a good estimate of the heritability h^2 of the phenotype based on the variability of the Boolean DNA of the networks.

To compute this heritability we evolved a population of holobionts, where each holobiont consists of a host network and a microbial network that can interact through regulatory connections between them (Fig. 7). The goal of the evolutionary process is to train the host network to acquire the desired phenotype $F(t)$. It is important to stress that the phenotype is defined only on the host network. However, since the host and microbial networks can interact, the adaptation of the host network to the desired phenotype also depends on the microbial network, and this dependence is greater the more regulatory interactions exist between the host and microbial networks. Therefore, in order to compute the heritability h^2 of the phenotype, one has to take into account the genetic variance of both the host and the microbial networks. The heritability h^2 has two contributions and can thus be written as $h^2 = h_h^2 + h_m^2$, where h_h^2 and h_m^2 are the heritabilities computed by taking into account the genetic variance of only the host and only the microbial networks, respectively. It has happened that when the heritability of a particular phenotypic trait, (such as height, cancer or diabetes), is computed, only the genetic variance of the human genome is taken into account. Therefore, researchers all over the world have been computing h_h^2 instead of h^2 . From Fig. 9 it can be seen that $h^2 \approx h_h^2$ (i.e. h_h^2 is a good estimate of h^2) only when there are no interactions whatsoever between the host and the microbial networks (this would correspond to an organism with no microbiota). However, as soon as the number of interactions between these two networks starts to increase, very quickly one has $h_h^2 < h^2$. The difference $h_m^2 = h^2 - h_h^2$ is the missing heritability, and is completely attributed to the genetic variability of the microbiota. As Fig. 9 shows, this missing heritability becomes larger the stronger the interaction between the host and the microbial networks.

The most important conclusion of the work presented here is the following: even when epistatic (nonlinear) effects between genetic variants can be neglected, and even when GWAS can efficiently detect the most important variants that contribute to a phenotype, if the genetic variance of the microbiota is not taken into account in the computation of the heritability h^2 of some particular phenotype, then this heritability will always be underestimated by a large amount. This is particularly true for those phenotypes that are strongly determined by (or correlated with) the microbiota composition of the host organism. This problem has recently started to be addressed experimentally through Metagenome Wide Association Studies (MWAS), through which the genetic variance of both the host and its microbiota can be measured [35]. Our results strongly suggest that MWAS will be essential to fill the missing heritability gap.

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

Acknowledgements S.H. thanks CONACyT for a Ph.D. scholarship and the C3-UNAM for economic support. This work was partly supported by PAPIT-UNAM grant IN226917.

Appendix.

Following Yang 2010 and Eskin 2015 [38, 11], to measure the heritability h^2 of a given phenotype, we start with Eq. 6 of the main text, which determines the value ξ_j of the phenotype of the j^{th} individual in the population as a linear contribution of the effects of the different genetic variants v_i ($i = 1, 2, \dots, \Omega_{HL}$):

$$\xi_j = \bar{\xi} + \sum_{i=1}^{\Omega_{HL}} g_{ij} \beta_i + e_j. \quad (12)$$

In this equation, $\bar{\xi}$ is the average value of the phenotype in the population; β_i is the contribution of variant v_i to the phenotype; e_j is an error that has to do with the unknown effect of the environment on the phenotype; and g_{ij} is a matrix whose entries acquire the values 1 and 0 depending on whether the j^{th} individual contains variant v_i or not, respectively. For simplicity in the calculation, it is convenient to perform the change of variable

$$x_{ij} = (g_{ij} - q_i) / \sigma_{\xi},$$

where q_i is the occurrence frequency of variant v_i in the population, and $\sigma_{\xi} = \sqrt{q_i(1 - q_i)}$ is the standard deviation of this quantity. With this change of variable, Eq.(12) can be written in matrix form as

$$\boldsymbol{\xi} = \bar{\xi} \cdot \mathbf{I} + \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{e},$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_{N_P})$ is the vector containing the phenotypes of the individuals in the population (analogously for the vectors $\boldsymbol{\beta}$ and $\mathbf{e}$), and $\mathbf{I}$ is the identity matrix. In our model, we know exactly the effect β_i that variant v_i has on the phenotype (it is the error difference $\delta\xi_i$ defined in Eq. 5 and reported in Fig.4). However, in a real situation this effect cannot be known accurately. Instead, it has to be estimated as $\hat{\beta}_i = \frac{1}{N_P} \mathbf{X}_i^T \boldsymbol{\xi}$, where $\mathbf{X}_i$ is the i^{th} row of matrix $\mathbf{X}$ and N_P is the number of individuals in the population.

To determine whether the effect of variant v_i is or not correlated with the phenotype, we compute the p -value of the effect of this variant on the phenotype. The p -value $p(v_i)$ corresponding to the variant v_i is computed as

$$p(v_i) = 2(1 - \Phi(1 \leq |t_0(v_i)|)),$$

where $t_0(v_i) = \hat{\beta}_i / \sqrt{C}$. In this expression, $C = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}$ is the covariance of the linear regression with $\hat{\sigma}^2 = \mathbf{e}^T \mathbf{e} / (N_P - \Omega_{HL})$ being the error of the estimation (Ω_{HL} is the number of different variants occurring in the population). The function $\Phi(x)$ is the (cumulative) normal distribution function.

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

Once the effect $\hat{\beta}_i$ of each variant v_i has been estimated, the variants whose effect on the phenotype is significantly small are discarded. Let us denote as S_V the set of variants that are not discarded and remain in the analysis, namely, the set of variants that have a strong effect on the phenotype. These variants can occur in the genome of both the host network and microbial networks. Therefore, the set S_V can be partitioned into two disjoint subsets, S_H and S_M such that $S_V = S_H \cup S_M$, where S_H is the set of relevant variants that occur in the genome of the host network and S_M is the set of relevant variants occurring in the microbial network. The genetic variance of the genotype is then estimated as

$$\sigma_g^2 = \sum_{v_i \in S_V} (\hat{\beta}_i)^2 = \sum_{v_i \in S_H} (\hat{\beta}_i)^2 + \sum_{v_i \in S_M} (\hat{\beta}_i)^2.$$

The heritability h^2 is then computed as

$$h^2 = \sigma_g^2 / \sigma_p^2,$$

where σ_p^2 is the phenotypic variability computed as

$$\sigma_p^2 = \frac{1}{N_p} \sum_{i=1}^{N_p} (\xi_i - \bar{\xi})^2.$$

References

1. Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, and David R. et al. Bentley. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.
2. Stefan Bornholdt. Boolean network models of cellular regulation: prospects and limitations. *Journal of the Royal Society Interface*, 5(suppl_1):S85–S94, 2008.
3. Andrew W Brooks, Kevin D Kohl, Robert M Brucker, Edward J van Opstal, and Seth R Bordenstein. Phyllosymbiosis: Relationships and Functional Effects of Microbial Communities across Host Evolutionary History. *PLoS Biology*, 14(11):e2000225, 2016.
4. Annalisa Buniello, JacquelineAL MacArthur, Maria Cerezo, LauraWHarris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2018.
5. Ilseung Cho and Martin J Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260, 2012.

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

6. Francis S Collins, Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, et al. New goals for the us human genome project: 1998-2003. *Science*, 282(5389):682–689, 1998.
7. International HapMap Consortium et al. The international HapMap project. *Nature*, 426(6968):789, 2003.
8. Ivana Cvijovic, Benjamin H. Good, and Michael M. Desai. The effect of strong purifying selection on genetic diversity. *Genetics*, 209(4):1235–1278, 2018.
9. Barbara Drossel. Random Boolean networks. *Reviews of nonlinear dynamics and complexity*, 1:69–110, 2008.
10. Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450, 2010.
11. Eleazar Eskin. Discovering genes involved in disease and the mystery of missing heritability. *Communications of the ACM*, 58(10):80–87, 2015.
12. Emmanuelle Génin. Missing heritability of complex diseases: case solved? *Human Genetics*, 139(1):103–113, 2019.
13. Greg Gibson. Population genetics and gwas: a primer. *PLoS biology*, 16(3):e2005485, 2018.
14. Valérie Grandjean, Danielle A Badro, and Jafar Kiani. RNA: a possible contributor to the "missing heritability". *Basic and clinical andrology*, 23(1):9, 2013.
15. Shi Huang. New thoughts on an old riddle: What determines genetic diversity within and between species? *Genomics*, 108(1):3–10, 2016.
16. Saúl Huitzil, Santiago Sandoval-Motta, Alejandro Frank, and Maximino Aldana. Modeling the role of the microbiome in evolution. *Frontiers in physiology*, 9:1836, 2018.
17. Yuzaburo Inoue and Naoki Shimojo. Microbiome/microbiota and allergies. In *Seminars in immunopathology*, volume 37, pages 57–64. Springer, 2015.
18. Pinreddy Karunakar, Ameek Bhalla, and Abhay Sharma. Transgenerational inheritance of cold temperature response in drosophila. *FEBS letters*, 593(6):594–600, 2019.
19. Dong-Hyun Kim. Gut microbiota-mediated drug-antibiotic interactions. *Drug Metabolism and Disposition*, 43(10):1581–1589, 2015.

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

20. Ellen M. Leffler, Kevin Bullaughey, Daniel R. Matute, Wynn K. Meyer, Laure Ségurel, Aarti Venkat, Peter Andolfatto, and Molly Przeworski. Revisiting an Old Riddle: What Determines Genetic Diversity Levels within Species? *PLoS Biology*, 10(9), 2012.
21. Ruth E Ley, Catherine A Lozupone, Micah Hamady, Rob Knight, and Jeffrey I Gordon. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology*, 6(10):776, 2008.
22. Brendan Maher. Personal genomes: The case of the missing heritability. *Nature News*, 456(7218):18–21, 2008.
23. Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin MRamos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747, 2009.
24. Howard Ochman, Michael Worobey, Chih-Horng Kuo, Jean-Bosco N. Ndjongo, Martine Peeters, Beatrice H. Hahn, and Philip Hugenholtz. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLOS Biology*, 8(11):1–8, 11 2010.
25. Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, et al. Functional SNPs in the lymphotoxin-gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*, 32(4):650, 2002.
26. John F Rawls, Buck S Samuel, and Jeffrey I Gordon. Gnotobiotic zebrafish reveal evolutionarily conserved responses to the gut microbiota. *Proceedings of the National Academy of Sciences*, 101(13):4596–4601, 2004.
27. Ravi Sachidanandam, David Weissman, Steven C Schmidt, Jerzy M Kakol, Lincoln D Stein, Gabor Marth, Steve Sherry, James C Mullikin, Beverley J Mortimore, David L Willey, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–934, 2001.
28. Santiago Sandoval-Motta, Maximino Aldana, Esperanza Martínez-Romero, and Alejandro Frank. The human microbiome and the missing heritability problem. *Frontiers in genetics*, 8:80, 2017.
29. Montgomery Slatkin. Epigenetic inheritance and the missing heritability problem. *Genetics*, 182(3):845–850, 2009.
30. Helen Tremlett, Kylynda C Bauer, Silke Appel-Cresswell, Brett B Finlay, and Emmanuelle Waubant. The gut microbiome in human neurological disease: a review. *Annals of neurology*, 81(3):369–382, 2017.

Huitzil, Sandoval-Motta, Frank and Aldana. Phenotype heritability in holobionts

31. P. Turnbaugh, R. Ley, and M. et al. Mahowald. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444:1027–1031, 2006.
32. Sophie Veigl, Javier Suárez, and Adrian Stencel. Does inheritance need a rethink? Conceptual tools to extend inheritance beyond the DNA. *Extended Evolutionary Synthesis*, November 5, 2019. <https://extendedevolutionarysynthesis.com/inheritance-rethink/>
33. Peter M Visscher. Sizing up human height variation. *Nature Genetics*, 40(5):489–490, 2008.
34. Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.
35. JunWang and Huijue Jia. Metagenome-wide association studies: Fine-mining the microbiome. *Nature Reviews Microbiology*, 14(8):508–522, 2016.
36. Ian D Wilson and Jeremy K Nicholson. Gut microbiome interactions with drug metabolism, efficacy, and toxicity. *Translational Research*, 179:204–222, 2017.
37. Andrew R Wood, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, 46(11):1173, 2014.
38. Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, GrantWMontgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565, 2010.
39. Ilana Zilber-Rosenberg and Eugene Rosenberg. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. *FEMS Microbiology Reviews*, 32(5):723–735, 2008.
40. Or Zuk, Eliana Hechter, Shamil R. Sunyaev, and Eric S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.
41. Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.