An extension of the Walsh-Hadamard transform to calculate and model epistasis in genetic landscapes of arbitrary shape and complexity

Andre J. Faure[1*], Ben Lehner[1,2,3,4], Verónica Miró Pina[1], Claudia Serrano Colome[1], and Donate Weghorn[1,2]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain.
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[3]Institució Catalana de Recerca i estudis Avançats (ICREA), Barcelona, Spain
[4]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK
[*]Corresponding author: andre.faure@crg.eu

Abstract

Inferring accurate models describing the relationship between genotype and phenotype is necessary in order to understand and predict how mutations to biological sequences affect the fitness and evolution of living organisms. The apparent abundance of epistasis (genetic interactions), both between and within genes, complicates this task and how to build mechanistic models that incorporate epistatic coefficients (genetic interaction terms) is an open question. The Walsh-Hadamard transform underlies a number of related definitions of epistasis. One of its main limitations is that it can only accommodate two alleles (amino acid or nucleotide states) per sequence position. In this paper we provide an extension of the Walsh-Hadamard transform that allows the calculation and modeling of background-averaged epistasis (also known as ensemble or statistical epistasis) in genetic landscapes with an arbitrary number of states per position (20 for amino acids, 4 for nucleotides, etc.). We also provide a recursive formula for the inverse matrix and then derive formulae to directly extract any element of either matrix without having to rely on the computationally intensive task of constructing or inverting large matrices. Finally, we demonstrate the utility of our theory by using it to model epistasis within a combinatorially complete multiallelic genetic landscape of a tRNA, revealing that both pairwise and higher-order genetic interactions are enriched between physically interacting positions.

Introduction

A fundamental challenge in biology is to understand and predict how changes (or mutations) to biological sequences (DNA, RNA, proteins) affect their molecular function and ultimately the phenotype of living

organisms. The phenomenon of 'epistasis' (genetic interactions) – broadly defined as the dependence of mutational effects on the genetic context in which they occur [1, 2, 3] – is widespread in biological systems, yet knowledge of the underlying mechanisms remains limited. Defining the extent of epistasis and better understanding of its origins has relevance in fields ranging from genetic prediction, molecular evolution, infectious disease and cancer drug development, to biomolecular structure determination and protein engineering [3].

Evolutionarily related sequences, natural genetic variation within populations, and more recently results of techniques such as deep mutational scanning (DMS) [4] – also known as massively parallel reporter assays (MPRAs) and multiplex assays of variant effect (MAVEs) – represent valuable sources of data to study epistasis [5, 1]. In particular, DMS enables the systematic measurement of phenotypic effects across entire combinatorially complete genetic landscapes [5, 6, 7, 8, 9, 10, 11, 12, 13]. However, precisely how to extract the most biologically relevant pairwise and higher-order epistatic coefficients (genetic interaction terms) from this type of data is an unresolved problem.

Quantitative definitions of epistasis vary among fields, but it has been argued that one particular formulation termed 'background-averaged' epistasis, also known as 'ensemble' or 'statistical' epistasis [1, 12], may provide the most useful information on the epistatic structure of biological systems [2]. The underlying rationale is that by averaging the effects of mutations across many different genetic backgrounds (contexts), the method is robust to local idiosyncrasies in the relationship between genotype and phenotype.

The current mathematical formalism of background-averaged epistasis is based on the Walsh-Hadamard transform [2]. Interestingly, although widely used in physics and engineering, the Walsh-Hadamard transform was first applied to non-biological fitness landscapes in the field of genetic algorithms (GA) [14], subsequently being proposed as the basis of a population genetic framework for the computation of higher-order epistasis [15]. However, the Walsh-Hadamard transform can only accommodate two alleles (amino acid or nucleotide states) per sequence position, with no extension to multialleleic landscapes (cardinality greater than two) yet made, as confirmed by multiple recent reports [2, 16, 17, 18]. Alternative implementations for multiallelic landscapes either rely on 'one-hot encoding' elements of larger alphabets as biallelic sequences – requiring the manipulation of prohibitively large Walsh-Hadamard matrices – or constructing graph Fourier bases [17], which is mathematically complex and provides no straightforward way to interpret epistatic coefficients. The result is that the application of background-averaged epistasis has been severely limited and its properties remain largely unexplored in more biologically realistic scenarios.

In this work we provide an extension of the Walsh-Hadamard transform that allows the calculation and modeling of background-averaged epistasis in genetic landscapes with an arbitrary number of states (20 for amino acids, 4 for nucleotides, etc.). We also provide a recursive formula for the inverse matrix, which is required to infer epistatic coefficients using regression. Furthermore, we derive convenient formulae to directly extract any element of either matrix without having to rely on the computationally intensive task of constructing or inverting large matrices. Lastly, we apply these formulae to the analysis of a multiallelic DMS dataset, demonstrating that sparse models inferred from the background-averaged representation (embedding) of the underlying genetic landscape more regularly include epistatic terms corresponding to direct physical interactions.

1 Extension of the Walsh-Hadamard transform to multiallelic landscapes

In this work, a genotype sequence is represented as a one-dimensional ordering of monomers, each of which can take on s possible states (or alleles), for example $s = 4$ for nucleotide sequences or $s = 20$ for amino acid sequences. Without loss of generality, the s states can be labelled $0, 1, 2, \dots, s-1$, where 0 denotes the wild-type allele. We are going to consider genotype sequences of length $n \in \mathbb{N}$, i.e. sequences taking values in $\mathcal{S}^n$, where $\mathcal{S} := \{0, 1, \dots, s-1\}$.

Each genotype $\vec{i} \in \mathcal{S}^n$ is associated with its phenotype $y_{\vec{i}}$. The phenotypic effect of a genotype $\vec{i}$ is measured with respect to the wild-type, i.e. it is given by $y_{\vec{i}} - y_{(0,\dots,0)}$.

If the phenotypic effects of individual mutations were independent, they would be additive, meaning that the phenotypic effect of $\vec{i} = (i_1, \dots, i_n)$ would be the sum of the phenotypic effects of the single mutants $(i_1, 0, \dots, 0), \dots, (0, \dots, 0, i_n)$. The epistatic coefficient quantifies how much the observed phenotypic effect of $\vec{i}$ deviates from this assumption. In the case of background-averaged epistasis, we quantify the interactions between a set of mutations by averaging over all possible genotypes for the remaining positions in the sequence. For example, if $n = 3$ and $s = 2$, the pairwise epistatic coefficient involving the mutations at positions 2 and 3 is calculated by averaging over all states (backgrounds) for the remaining positions, in this case given by the two states of the first position ($*$ denotes the positions at which the averaging is performed), i.e.

$$\begin{aligned}
\varepsilon_{(*,1,1)} &= \frac{1}{2}\Big(\big[(y_{(1,1,1)} - y_{(1,0,0)}) - (y_{(1,1,0)} - y_{(1,0,0)}) - (y_{(1,0,1)} - y_{(1,0,0)})\big] + \\
&\qquad \big[(y_{(0,1,1)} - y_{(0,0,0)}) - (y_{(0,1,0)} - y_{(0,0,0)}) - (y_{(0,0,1)} - y_{(0,0,0)})\big]\Big) \\
&= \frac{1}{2}\Big(\big[y_{(1,1,1)} - y_{(1,1,0)} - y_{(1,0,1)} + y_{(1,0,0)}\big] + \big[y_{(0,1,1)} - y_{(0,1,0)} - y_{(0,0,1)} + y_{(0,0,0)}\big]\Big).
\end{aligned}$$

More generally, in [2] it is shown that for $s = 2$ and any sequence length n, phenotypic effects can be decomposed into background-averaged epistatic coefficients with

$$\bar{\varepsilon}_n = \hat{V}_n \cdot \hat{H}_n \cdot \bar{y}_n,$$

where $\bar{y}_n$ is the vector $(y_{\vec{i}}, \vec{i} \in [0,1]^n)$, $\bar{\varepsilon}_n$ is the vector $(\varepsilon_{\vec{j}}, j \in [*,1]^n)$ and $\hat{H}_n$ and $\hat{V}_n$ are $2^n \times 2^n$ matrices defined recursively as follows:

$$\hat{H}_{n+1} = \begin{pmatrix} \hat{H}_n & \hat{H}_n \\ \hat{H}_n & -\hat{H}_n \end{pmatrix} \qquad \hat{H}_0 = 1,$$

$$\hat{V}_{n+1} = \begin{pmatrix} \frac{1}{2}\hat{V}_n & 0 \\ 0 & -\hat{V}_n \end{pmatrix} \qquad \hat{V}_0 = 1.$$

The matrix $\hat{H}$ is known as the Walsh-Hadamard transform [19, 20] and $\hat{V}$ is a diagonal weighting (or nor-

malisation) matrix to correct the sign and account for averaging over different numbers of backgrounds as a function of epistatic order [2].

In this work, we provide an extension of this theory to describe background-averaged epistasis for sequences with an arbitrary number of states s. Before writing a general formula, we consider the simplest possible multi-state (multiallelic) landscape i.e. a sequence of length $n = 1$ with $s = 3$,

$$\begin{pmatrix} \varepsilon_{(*)} \\ \varepsilon_{(1)} \\ \varepsilon_{(2)} \end{pmatrix} = \begin{pmatrix} 1/3 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} y_{(0)} \\ y_{(1)} \\ y_{(2)} \end{pmatrix} := V_1 \cdot H_1 \cdot \bar{y}_1.$$

Consistent with the definition of background-averaged epistasis for biallelic landscapes [2], the zeroth-order epistatic coefficient $\varepsilon_{(*)}$ is the mean phenotypic value across all genotypes and the first-order epistatic coefficients $\varepsilon_{(1)}$ and $\varepsilon_{(2)}$ are simply the respective individual phenotypic effects of genotypes $y_{(1)}$ and $y_{(2)}$ with respect to the wild-type. However, the key feature of H_1 for multiallelic landscapes – and where it departs from the canonical Walsh-Hadamard transform – is the introduction of zero elements to exclude phenotypes that are irrelevant for the calculation of a given epistatic coefficient. In other words, these phenotypes are excluded because they correspond neither to relevant intermediate genotypes nor alternative genetic backgrounds.

If we now consider a sequence of length $n = 2$ with $s = 3$, then the H_2 and V_2 matrices become 9×9 ($s^n \times s^n$) and can be constructed from recurring to the case $n = 1$ above, giving

$$\begin{pmatrix} \varepsilon_{(*,*)} \\ \varepsilon_{(*,1)} \\ \varepsilon_{(*,2)} \\ \varepsilon_{(1,*)} \\ \varepsilon_{(1,1)} \\ \varepsilon_{(1,2)} \\ \varepsilon_{(2,*)} \\ \varepsilon_{(2,1)} \\ \varepsilon_{(2,2)} \end{pmatrix} = \begin{pmatrix} \frac{1}{9} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} y_{(0,0)} \\ y_{(0,1)} \\ y_{(0,2)} \\ y_{(1,0)} \\ y_{(1,1)} \\ y_{(1,2)} \\ y_{(2,0)} \\ y_{(2,1)} \\ y_{(2,2)} \end{pmatrix}$$

$$:= V_2 \cdot H_2 \cdot \bar{y}_2,$$

where the colors highlight the block structure of the matrices. In V_2, the red square corresponds to $\frac{1}{s}V_1$ and the light red squares to $-V_1$. In H_2, the gray squares correspond to H_1 and the blue squares to $-H_1$. In Table 1 we show the results of background-averaged epistatic coefficients calculated by applying the above formula to an empirical multiallelic landscape with $n = 2$ and $s = 3$ [6].

Nucleic acid sequence	**Base $s = 3$ representation**	**Phenotypic effect** $\bar{y}_2$	**Epistatic term** $\bar{\varepsilon} = V_2 \cdot H_2 \cdot \bar{y}_2$
GC	(0,0)	0	-0.17
GA	(0,1)	-0.14	-0.21
GT	(0,2)	-0.07	0.02
AC	(1,0)	-0.13	-0.24
AA	(1,1)	-0.8	-0.53
AT	(1,2)	-0.01	0.19
TC	(2,0)	-0.19	-0.05
TA	(2,1)	0	0.33
TT	(2,2)	-0.18	0.08

Table 1: Interaction terms based on background-averaged epistasis ($\bar{\varepsilon}$) for an empirical multiallelic genotype-phenotype landscape consisting of all combinations of two mutations each at positions 6 and 66 in the tRNA-Arg(CCU) [6], i.e. $n = 2$ and $s = 3$. The first two columns indicate nucleic acid sequences and their base 3 representations. Here the 'GC' reference (wild-type) genotype corresponds to that of *S. cerevisiae*, denoted by $(0, 0)$. The second two columns show the measured phenotypic effects and corresponding background-averaged epistatic coefficients. See Results for a regression analysis of the entire dataset.

More generally, for any value of s, when $n = 1$,

$$\begin{pmatrix} \varepsilon_{(*)} \\ \varepsilon_{(1)} \\ \varepsilon_{(2)} \\ \vdots \\ \varepsilon_{(s-1)} \end{pmatrix} = \begin{pmatrix} 1/s & 0 & 0 & \dots & 0 \\ 0 & -1 & 0 & \dots & 0 \\ 0 & 0 & -1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \dots & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} y_{(0)} \\ y_{(1)} \\ y_{(2)} \\ \vdots \\ y_{(s-1)} \end{pmatrix} := V_1 \cdot H_1 \cdot \bar{y}_1,$$

where $\varepsilon_{(*)}$ corresponds to averaging phenotypes over all possible genotypes and the remaining coefficients simply correspond to the phenotypic effects of each mutation.

For $n = 2$, we have to consider different combinations of mutations in both positions. In this case, the phenotypes can be written as

$$y_{(0,0)}, y_{(0,1)}, \dots, y_{(0,(s-1))}, y_{(1,0)}, \dots, y_{(1,(s-1))}, \dots, y_{((s-1),0)}, \dots, y_{((s-1),(s-1))} \,.$$

A natural ordering of the phenotypes is given by interpreting genotype $\vec{i}$ as the base s representation of an integer (see Table 1). From this, we can see how the first s genotypes correspond to combining the wild-type allele at the first position with a state from the case $n = 1$, i.e. to genotypes that can be written $0 \frown \vec{i} := (0, \vec{i})$, with $\vec{i} \in \mathcal{S}^1$. The next s genotypes correspond to the first mutated allele at the first position combined with all the genotypes of $n = 1$, i.e. $1 \frown \vec{i}, \vec{i} \in \mathcal{S}^1$, and so on. Therefore, we can write the matrices H and V following a block structure. In the case $n = 2$ and any given s, we would then have

$$H_2 = \begin{pmatrix} H_1 & H_1 & H_1 & \dots & H_1 \\ H_1 & -H_1 & 0 & \dots & 0 \\ H_1 & 0 & -H_1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ H_1 & 0 & \dots & 0 & -H_1 \end{pmatrix},$$

where the number of H_1 blocks corresponds to the number of states of the first position, so s. Moreover, each of these blocks must be normalized to yield the corresponding background-averaged epistatic terms. Therefore V_2 can also be expressed as a function of V_1 as follows:

$$V_2 = \begin{pmatrix} \frac{1}{s}V_1 & 0 & \dots & 0 \\ 0 & -V_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -V_1 \end{pmatrix}.$$

Given these two matrices, we can write the background-averaged epistatic coefficients for the case of $n = 2$ and s different states per position as $\bar{\varepsilon}_2 = V_2 \cdot H_2 \cdot \bar{y}_2$. More generally, the decomposition of phenotypic effects into background-averaged epistatic coefficients is given by

$$\bar{\varepsilon}_n = V_n \cdot H_n \cdot \bar{y}_n, \tag{1}$$

where H_n and V_n can be defined recursively as

$$H_{n+1} = \begin{pmatrix} H_n & H_n & H_n & \dots & H_n \\ H_n & -H_n & 0 & \dots & 0 \\ H_n & 0 & -H_n & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ H_n & 0 & \dots & 0 & -H_n \end{pmatrix} \qquad H_0 = 1 \qquad \text{and} \qquad H_n \text{ is } s^n \times s^n, \tag{2}$$

$$V_{n+1} = \begin{pmatrix} \frac{1}{s}V_n & 0 & 0 & \dots & 0 \\ 0 & -V_n & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & -V_n \end{pmatrix} \qquad V_0 = 1 \qquad \text{and} \qquad V_n \text{ is } s^n \times s^n. \tag{3}$$

2 Recursive inverse matrix

Equation (1) defines the vector of epistatic coefficients, $\bar{\varepsilon}_n$, as a function of the vector of phenotypes, $\bar{y}_n$, which in general is the quantity that is measured experimentally. However, usually phenotypic measurements are

only available for a subset of genotypes. An alternative is therefore to estimate the epistatic coefficients $\bar{\varepsilon}_n$ by regression,

$$\bar{y}_n = H_n^{-1} \cdot V_n^{-1} \cdot \bar{\varepsilon}_n \,, \tag{4}$$

where the product $H_n^{-1} \cdot V_n^{-1}$ represents a matrix of sequence features. This is analogous to the more widely used one-hot encoding strategy, which implicitly relies on a 'background-relative' (or 'biochemical') view of epistasis when regressing to full order [2]. We discuss other advantages of estimating background-averaged epistatic coefficients using regression at the end of this manuscript.

Since V_n is a diagonal matrix, its inverse is also a diagonal matrix whose elements are the inverse of the elements of V_n.

The inverse of H_n is the matrix A_n which can be defined recursively as

$$A_{n+1} = \frac{1}{s}\begin{pmatrix} A_n & A_n & A_n & \dots & A_n \\ A_n & (1-s)A_n & A_n & \dots & A_n \\ A_n & A_n & (1-s)A_n & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & A_n \\ A_n & A_n & \dots & A_n & (1-s)A_n \end{pmatrix} \qquad A_0 = 1 \qquad \text{and} \qquad A_n \text{ is } s^n \times s^n. \tag{5}$$

See Proposition 1 in the Supplementary Methods for a proof of this result. This is the most efficient method to determine the full matrix A_n (see Results) and, to the best of our knowledge, the first reported recursive definition of the inverse Walsh-Hadamard transform.

3 Formulae to obtain elements of the matrices

When regressing phenotypes on genotypes, a common goal is to determine whether epistatic coefficients up to the rth order (where $r < n$) are sufficient to describe the complexity of the biological system. Furthermore, as mentioned above, some fraction of phenotype values within combinatorially complete genetic landscapes are typically unavailable, representing missing data. Restricting the epistatic order and missing phenotypes respectively correspond to omitting rows and columns from H_n (and vice versa from A_n). Formulae to directly obtain elements of the matrices in equations (1) and (4) would therefore be convenient.

In order to write the matrix element $(H_n)_{ij}$, we need to compare the genotype sequences $\vec{i}, \vec{j} \in \mathcal{S}^n$,

$$\vec{i} = (i_1, i_2, \dots, i_n)$$
$$\vec{j} = (j_1, j_2, \dots, j_n)\,,$$

where $\vec{i}$ denotes the ith element in $\mathcal{S}^n$, $\mathcal{S} = \{0, 1, \dots, s-1\}$, and the elements of $\mathcal{S}^n$ are ordered by the base s representation of integers. For instance, for any value of n, we will denote the wild-type state with index

$i = 1$ and write $\vec{i} = \vec{1} = (0, \dots, 0)$. The element denoted with index $i = 2$ would be $\vec{i} = \vec{2} = (0, \dots, 0, 1)$ and so on.

The elements of H_n can be written as

$$(H_n)_{ij} = \begin{cases} (-1)^{(E_n)_{ij}} & \text{if } (M_n)_{ij} = n \\ 0 & \text{otherwise,} \end{cases}$$

where M and E are $s^n \times s^n$ matrices whose elements are

$$(E_n)_{ij} = \sum_{\substack{k=1 \\ i_k \cdot j_k > 0}}^{n} \delta_{i_k j_k} \tag{6}$$

$$(M_n)_{ij} = \sum_{\substack{k=1 \\ i_k \cdot j_k > 0}}^{n} \delta_{i_k j_k} + \sum_{\substack{k=1 \\ i_k \cdot j_k = 0}}^{n} 1 = (E_n)_{ij} + \sum_{\substack{k=1 \\ i_k \cdot j_k = 0}}^{n} 1,$$

where δ_{ij} denotes the Kronecker delta of i, j, which is equal to 1 when $i = j$ and 0 if $i \neq j$. In words, $(E_n)_{ij}$ counts the number of positions at which the genotype sequences $\vec{i}$ and $\vec{j}$ carry the same mutated allele and $(M_n)_{ij}$ is equal to $(E_n)_{ij}$ plus the number of positions where $\vec{i}$ or $\vec{j}$ carry the wild-type allele. See Proposition 2 in the Supplementary Methods for a proof of this result.

Furthermore, the elements of A_n can be written as

$$(A_n)_{ij} = \frac{1}{s^n}(1 - s)^{(E_n)_{ij}}, \tag{7}$$

where E_n is defined as in (6). See Proposition 3 in the Supplementary Methods for a proof of this result.

Finally, the matrices V_n and V_n^{-1} are diagonal matrices whose diagonal elements can be written as

$$(V_n)_{ii} = (-1)^{n - W_n(\vec{i})} \frac{1}{s^{W_n(\vec{i})}} \tag{8}$$

and

$$(V_n^{-1})_{ii} = (-1)^{n - W_n(\vec{i})} s^{W_n(\vec{i})}, \tag{9}$$

where

$$W_n(\vec{i}) := \sum_{k=1}^{n} w_k, \text{ with } w_k := \delta_{i_k 0}$$

and $\vec{i}$ again denotes the ith element in S^n when ordered by the base s representation of integers. In words, $w_k = 1$ if the genotype sequence $\vec{i}$ carries the wild-type allele at position k and $W_n(\vec{i})$ counts the number of positions in $\vec{i}$ carrying the wild-type allele. We prove this result in Proposition 4 in the Supplementary Methods.

4 Generalization to different numbers of states per position

We can generalize the formulae described in the previous section further by considering that each position can have different numbers of states. In this case, we can denote s_k the number of possible states at position k. For $n = 1$, this corresponds to exactly the same matrix as in the previous case but with $s = s_1$, which is the number of possible states in this position. For $n = 2$, the matrix changes because now the new position can have a different number of possible states, s_2. Following the recursive definition of H_n, we can construct H_2 by repeating H_1 s_2 times, with the structure stated in (2). Therefore, we have

$$H_2 = \underbrace{\begin{pmatrix} \overbrace{H_1}^{s_1} & H_1 & H_1 & \dots & H_1 \\ H_1 & -H_1 & 0 & \dots & 0 \\ H_1 & 0 & -H_1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ H_1 & 0 & \dots & 0 & -H_1 \end{pmatrix}}_{s_2 \text{ blocks of size } s_1 \Longrightarrow s_2 s_1}.$$

So the structure is exactly the same but the size of the matrix for each n varies according to the number of possible states of the new position. The definition of H_n is the same as in (2) but the dimensions of the matrix are $\prod_{k=1}^{n} s_k \times \prod_{k=1}^{n} s_k$. Similarly, the inverse matrix A_{n+1} can be written recursively as

$$A_{n+1} = \frac{1}{s_{n+1}} \begin{pmatrix} A_n & A_n & A_n & \dots & A_n \\ A_n & (1-s_{n+1})A_n & A_n & \dots & A_n \\ A_n & A_n & (1-s_{n+1})A_n & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & A_n \\ A_n & A_n & \dots & A_n & (1-s_{n+1})A_n \end{pmatrix}, \quad A_0 = 1 \quad \text{and} \quad A_n \text{ is } \prod_{k=1}^{n} s_k \times \prod_{k=1}^{n} s_k. \tag{10}$$

The matrix A_n defined in (10) is the inverse of the matrix H_n in the general case where each position can have a different number of states.

In this general case, the elements of H_n and A_n can be written as

$$(H_n)_{ij} = \begin{cases} (-1)^{(E_n)_{ij}} & \text{if } (M_n)_{ij} = n \\ 0 & \text{otherwise} \end{cases}$$

$$(A_n)_{ij} = \frac{\prod_{k=1}^{n} \left(1 - s_k\right)^{e_k}}{\prod_{k=1}^{n} s_k},$$

where E_n and M_n are defined as in (6) and $e_k = \begin{cases} 1 & \text{if } i_k = j_k \neq 1 \\ 0 & \text{otherwise} \end{cases}$.

The matrices V_n and V_n^{-1} are diagonal matrices whose diagonal elements can be written as

$$(V_n)_{ii} = (-1)^{n-W_n(\vec{i})} \prod_{k=1}^{n} \left(\frac{1}{s_k}\right)^{w_k}$$

and

$$(V_n^{-1})_{ii} = (-1)^{n-W_n(\vec{i})} \prod_{k=1}^{n} {s_k}^{w_k},$$

where

$$W_n(\vec{i}) := \sum_{k=1}^{n} w_k, \text{ with } w_k := \delta_{i_k 0}.$$

We prove the results in this section in Propositions 5, 6 and 7 in the Supplementary Methods.

The above formulae permit the calculation and modeling of background-averaged epistasis in arbitrarily-shaped genetic landscapes, i.e. with any number of alleles (states) per position, as well as the direct construction of sub-matrices for regression to any desired epistatic order and/or in the presence of missing data. In the following sections we report benchmarking results comparing the performance of alternative methods to obtain H_n and A_n, as well as results from the application of our theory extension to an empirical multiallelic genotype-phenotype landscape.

5 Results

5.1 Benchmarking

Figure 1a-d provides a visualization of the matrices H_n and A_n for different values of n and s, clearly showing a fractal pattern in all cases due to their recursive nature.

In this paper, we provide different methods to construct $A_n = H_n^{-1}$. First, H_n can be numerically inverted using standard matrix inversion algorithms (here we use the `linalg.inv` function from the SciPy library in Python), referred to as "Recursive H_n inverse" in Figure 1e,f. Alternatively, the recursive definition of the inverse given by equation (5) can be used, which we refer to as "Recursive A_n". As can be seen in Figure 1e, this method is faster than numerically inverting H_n.

Finally, we also provide a convenient formula for extracting specific individual elements of A_n (Proposition 3), referred to as "All elements A_n" in Figure 1e,f. This method is more computationally intensive than the previously described methods, due to the formula relying on the computation of $(E_n)_{ij}$, which equates to counting the number of sequence positions that are identically mutated in vectors $\vec{i}$ and $\vec{j}$, each of size n. However, in situations where subsets of elements (or sub-matrices) – rather than full matrices – are desired, Proposition 3 provides a method that can be faster and more memory efficient (see "10 elements A_n" in Figure 1e,f).

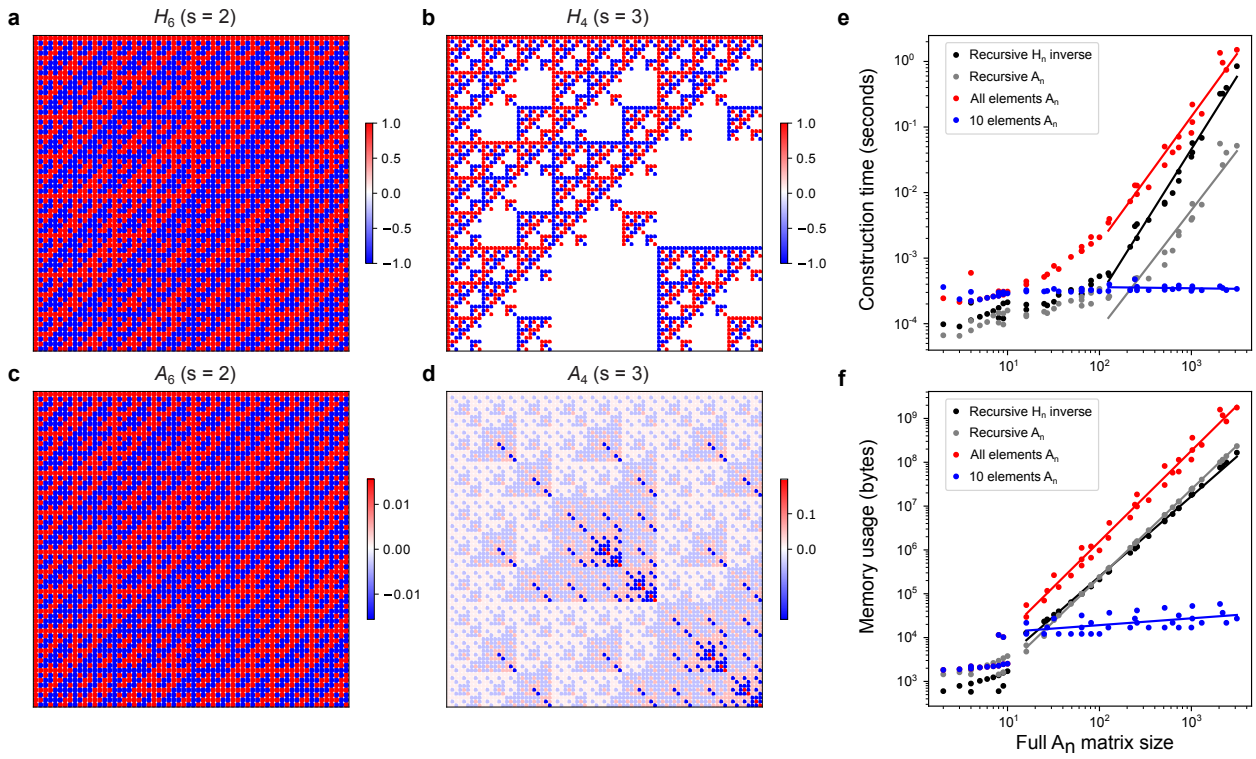

Figure 1: Benchmarking results and heat map representations of matrices corresponding to the binary (biallelic) and multi-state (multiallelic) extension of the Walsh-Hadamard transform, and their corresponding inverses. **a**, H_6 Walsh-Hadamard transform. **b**, H_4 multi-state extension of the Walsh-Hadamard transform for $s = 3$. **c**, A_6 Inverse Walsh-Hadamard transform. **d**, A_4 multi-state extension of the inverse Walsh-Hadamard transform for $s = 3$. **e**, Computational time on a MacBook Pro (13-inch, 2.017, 2.3GHz dual-core Intel Core i5) for extracting elements of A_n matrices of various dimensions and numbers of states (alleles) per position ($s \in [2, 10]$). Comparisons are shown between numerically inverting the recursively constructed H_n (using `scipy.linalg.inv`), i.e. "Recursive H_n inverse", using the recursive formula for A_n, using the formula to extract all elements of A_n and extracting 10 random elements of A_n (see legend). The mean across 10 replicates is depicted. Linear regression lines were fit to data from matrices with at least 100 elements. **f**, Similar to **e** but indicating memory usage. Linear regression lines were fit to data from matrices with at least 10 elements.

For example, in the case of a 10-mer DNA sequence, constructing the full inverse transform A_{10} with $s = 4$ would require $> 10^{23}$ bytes (100 million petabytes) of memory in the best-case scenario ("Recursive H_n inverse" in Figure 1f, log-linear extrapolation). Similarly, the full inverse transform for a 4-mer amino acid sequence (A_4 with $s = 20$) would impose a memory footprint $> 10^{20}$ bytes. On the other hand, calculating the subset of elements from these matrices required for the prediction of a single phenotype using epistatic coefficients up to third order (three-way genetic interaction terms) is feasible in both situations using Proposition 3 (3,675 and 29,678 elements; 2.5 GB and 192 GB of memory; 1.8 and 99 seconds, respectively). This memory footprint can easily be diminished further using data chunking, which is a unique benefit of this method.

5.2 Application to a multiallelic genotype-phenotype landscape

In order to demonstrate the utility of our theory, we used it to model epistasis within a combinatorially complete multiallelic genetic landscape of a tRNA. Figure 2a-c summarises the model system and DMS experimental strategy employed in [6]. Briefly, a budding yeast strain was used in which the single-copy arginine-CCU tRNA (tRNA-Arg(CCU)) gene is conditionally required for growth. A library of variants of this gene was designed to cover all 5,184 ($2^6 \times 3^4$) combinations of the 14 nucleotide substitutions observed in ten positions in post-whole-genome duplication yeast species (Figure 2a,b). The library was transformed into *S. cerevisiae*, expressed under restrictive conditions and the enrichment of each genotype in the culture was quantified by deep sequencing before and after selection (Figure 2c). After reprocessing of the raw data, we retained high quality fitness estimates for 3,847 variants (74.2%).

Although the findings in [6] were based on the application of background-averaged epistasis theory, the prior limitation of the Walsh-Hadamard transform to only two alleles per sequence position required the authors to adopt an *ad hoc* strategy that involved performing separate analyses on combinatorially complete biallelic sub-landscapes.

However, with the extensions provided in this work, we were able model background-averaged epistasis in this multiallelic landscape using all available data simultaneously. We trained Lasso regression models of the form in equation (4) to predict variant fitness from nucleotide sequences, where the inferred model parameters correspond to background-averaged epistatic coefficients up to eighth order (Figure 2d; see Methods). To determine the effect of data sparsity on the results, we sub-sampled the original data to obtain training dataset sizes ranging from 64% to 1% of all variants with high quality fitness estimates. The resulting models incorporate many higher-order epistatic coefficients (Figure 2e, 'Background-averaged models') yet exhibit extreme sparsity, with the median number of non-zero coefficients of any order ranging from 19 to 60 i.e. approximately 1% of all possible coefficients of eighth order or less (Figure 2f, Supplementary Figure S1d). Figure 2e indicates model performance on held out test data, with all models except those fit using the most severely subsampled data (1%) tending to explain more than 50% of the total explainable variance.

For comparison, we used the same procedure to fit Lasso regression models of the form $\bar{y} = G^{-1} \cdot \bar{\varepsilon}$, where G^{-1} represents a matrix of one-hot encoded sequence features i.e. the presence or absence of a given mutation

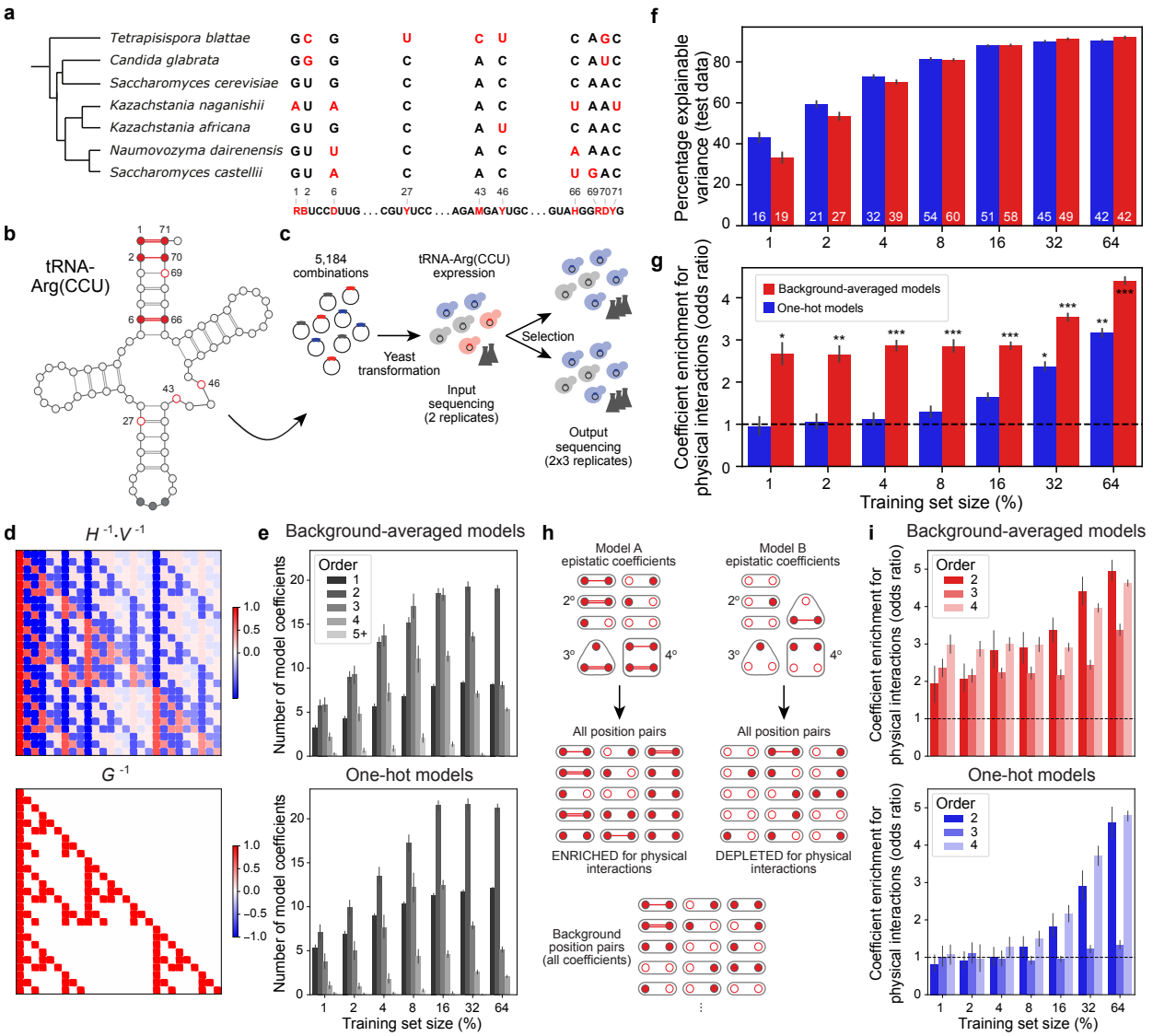

Figure 2: Learning sparse models from the near combinatorially complete fitness landscape of a tRNA. **a**, Species phylogenetic tree and multiple sequence alignment of the tRNA-Arg(CCU) orthologues indicating variable positions across the seven yeast species and the synthesized library below: R (A or G); B (C, G or T); D (A, G or T); Y (C or T); M (A or C); H (A, C or T). **b**, Secondary structure of *S. cerevisiae* tRNA-Arg(CCU) indicating variable positions (open and closed red circles) and three Watson–Crick base pairing (WCBP) interactions between pairs of variable positions i.e. [1,71], [2,70] and [6,66] (red lines and closed red circles). **c**, DMS experiment to quantify the phenotypic effects of all variants in the combinatorially complete genetic landscape. See [6] for details. **d**, Cartoon depiction of alternative feature matrices for inferring epistatic coefficients by linear regression. G^{-1} in the lower panel indicates the matrix of one-hot encoded sequence features – or embeddings – typically used when fitting models of genotype-phenotype landscapes [2]. The upper panel represents the matrix of sequence features used to infer background-averaged epistatic coefficients, as in equation (4). **e**, Numbers of non-zero espistatic coefficients of different orders in Lasso regression models inferred using different random fractions of the DMS data indicated in panels a-c. **f**, Performance of Lasso regression models. The median number of model coefficients is indicated. Colour scale as in panel g. **g**, Enrichment of direct physical interactions (red lines in panel b) in non-zero epistatic coefficients (see panel h). *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. **h**, Strategy for testing enrichment of direct physical interactions in Lasso regression model coefficients. **i**, Same as panel g, except enrichment test results are shown separately for epistatic coefficients of different orders. Error bars indicate nonparametric bootstrap 95% confidence intervals of the mean in all panels.

– or mutation combination (interaction) – with respect to the reference (wild-type) genotype is denoted by a '1' or '0' respectively (Figure 2d-f, 'One-hot models'). The definition of G and its relationship to the biochemical (or background-relative) view of epistasis is explained in [2]. The sparsity of one-hot models is similar to that of background-averaged models regardless of training set size (Figure 2e,f). However, the latter tend to incorporate greater numbers of higher-order epistatic terms, particularly with larger training set sizes (Figure 2e, orders 3,4,5+), whereas the former tend to perform slightly better with very small training set sizes (Figure 2f).

To evaluate whether the inferred models report on biologically relevant features of the underlying genetic landscapes, we tested whether sparse model coefficients were more likely to comprise genetic interactions (or modulators thereof) involving known physically interacting positions in the wild-type tRNA secondary structure (Figure 2b,h). Regardless of data sparsity, background-averaged model coefficients tend to be significantly enriched for physical interactions (Figure 2g, Supplementary Figure S1a). On the other hand, in the case of even moderate sub-sampling of training data (16%), one-hot model coefficients show no such enrichment (Figure 2g). Importantly, repeating a similar enrichment analysis using randomly selected model coefficients of identical number and distribution over coefficient orders speaks to the validity of the Fisher's Exact Test null hypothesis with only minor inflation of the corresponding test statistic (Supplementary Figure S1b,c). Restricting the enrichment analysis to epistatic coefficients of specific orders shows qualitatively similar results, with background-averaged model coefficients up to fourth order significantly enriched for physical interacting position pairs, even at the most severe sub-sampling fractions (Figure 2i, Supplementary Figure S1e,f).

6 Discussion

We have provided an extension to the most rigorous computational framework available for describing and modeling biological genotype-phenotype mappings. Beyond the study of background-averaged epistasis with respect to mutations in the primary sequence, this also permits the inclusion of 'epimutations' (changes in the epigenetic state of DNA), amino acid post-translational modifications or even particular environmental/experimental conditions.

In the simplest application, background-averaged epistatic coefficients (genetic interaction terms) can be directly computed from phenotypic measurements via the decomposition in equation (1). However, estimating epistatic coefficients by regression – as in equation (4) – is a more natural choice in the presence of missing data, when data for multiple related phenotypes is available [21] and/or in the presence of global epistasis [22, 23]. Our mathematical results provide three alternative methods to compute the multi-state (multiallelic) extension of the inverse Walsh-Hadamard transform A_n, one of which allows the direct extraction of specific elements or sub-matrices. In which situations might this capability be desirable?

First, constructing full A_n matrices – particularly by numerical inversion – is impractical for large genetic landscapes. Second, the result of the product $H_n^{-1} \cdot V_n^{-1}$ represents a matrix of sequence features when setting up the inference of epistatic (model) coefficients $\bar{\varepsilon}_n$ from phenotypic measurements $\bar{y}_n$ as a regression

task [21, 22, 24, 25]. The ability to construct this feature matrix in batches (of rows) allows computational resource-efficient iteration over large datasets when using frameworks such as TensorFlow or PyTorch.

Third, there are currently no methods to comprehensively map empirical genotype-phenotype landscapes with size greater than the low millions of genotypes. Therefore, assaying landscapes of this size or larger will typically involve experimental measurement of a (random) sub-sample of genotypes, corresponding to distinct rows in A_n. In other words, it is usually unnecessary to construct full A_n matrices when modeling real experimental data. Finally, there is evidence of extreme sparsity in the epistatic architecture of biomolecules where only a small fraction of theoretically possible genetic interactions are non-zero [7]. The feasibility of sampling very large background-averaged epistatic coefficient spaces may improve methods to infer accurate genotype-phenotype models.

Using results from the analysis of a near combinatorially complete multiallelic fitness landscape of a tRNA, we have shown that sparse regression models relying on a background-averaged definition of epistasis can efficiently capture salient features of the underlying biological system – namely direct physical interactions – even in situations of sparse sampling of phenotypes. This behaviour, which we speculate is due to a richer representation of the sequence feature space compared to one-hot models (i.e. higher level of constraint during model fitting; Figure 2d), is particularly desirable in the case of very large genetic landscapes where comprehensive phenotyping is infeasible. However, more work is needed to determine whether this result holds more generally. One difficulty in such comparisons between approaches is the requirement for a set of interactions or landscape features that are known to be critical for biomolecular function. Here we rely on Watson–Crick base pairing interactions whose importance for RNA secondary structure and function is well-established.

More broadly, this work opens the door to investigations of the biological properties of background-averaged epistasis in genetic landscapes of arbitrary shape and complexity. Beyond applications within the field of DMS, we believe our theory extensions have the potential to influence the research of evolutionary geneticists, complex trait analysis – or QTL mapping – the GWAS community and others. In future it will be important to compare the performance and properties of models relying on this definition of epistasis to those of other recently proposed models that incorporate higher-order genetic interactions for phenotypic prediction [26, 27].

7 Methods

Raw sequencing (FASTQ) files obtained from the tRNA-Arg(CCU) DMS experiment in [6] were re-processed with DiMSum v1.3 [28] using default parameters with minor adjustments. We obtained fitness estimates for 5,059 out of a total of 5,184 possible variants (97.6%) in the combinatorially complete genetic landscape. We restricted the data to a high quality subset by requiring fitness estimates in all six biological replicates as well as at least 10 input read counts in all input samples. This resulted in a total of 3,847 retained variants (74.2%) for downstream analysis.

We trained Lasso regression models to predict variant fitness estimates from nucleotide sequences using the 'scikit-learn' Python package. Training data comprised random subsets of 1, 2, 4, 8, 16, 32 and 64% of retained variants of all mutation orders. All remaining held out variants comprised the 'test' data which was unseen during model training in each case.

To train models inferring background-averaged epistatic coefficients we used feature matrices of the form $H_n^{-1} \cdot V_n^{-1}$ (see equation (4)). For comparison, one-hot encoded matrices of sequence features were used. Linear regression was performed using 10-fold cross validation to determine the optimal value of the L1 regularization parameter λ in the range [0.005, 0.5] ('LassoCV' and 'RepeatedKFold' functions). Final models were fit to all training data. In order to estimate model-related statistics and performance results we fit 100 models to different random subsets of the training data for each model type and training data fraction. In Figure 2 and Supplementary Figure S1 we plot the mean or median of the indicated measures over all models, where 95% confidence intervals were obtained using a nonparametric bootstrap approach and 1000 bootstrap samples. For performance estimates in Figure 2f we estimated the maximum explainable variance by taking the square of the mean Pearson correlation between replicate fitness estimates over all 15 pairwise combinations.

To test enrichment of physical interactions in Lasso model coefficients we used the strategy illustrated in Figure 2h. For each model, all position pairs represented in non-zero epistatic coefficients of at least second order were determined. The number of position pairs corresponding to direct physical interactions was counted and an associated enrichment score (odds ratio) and P-value calculated using Fisher's Exact Test. The background set consisted of all position pairs in all possible epistatic coefficients. To test the appropriateness of the null hypothesis we also repeated enrichment analyses using random models i.e. randomly chosen sets of epistatic coefficients matching the numbers of non-zero coefficients in Lasso models and their distribution over different epistatic orders.

8 Data availability

Source code for all computational analyses is available at https://github.com/lehner-lab/whmatrixextms.

9 Acknowledgments

We thank all members of the Lehner and Weghorn Labs for helpful discussions and suggestions.

10 Funding

This work was funded by European Research Council (ERC) Advanced grant (883742), the Spanish Ministry of Science and Innovation (PID2020-118723GB-I00, EMBL Partnership, Severo Ochoa Centre of Ex-

cellence), the la Caixa Research Foundation (LCF/PR/HR21/52410004), the Bettencourt Schueller Foundation, the AXA Research Fund, Agencia de Gestio d'Ajuts Universitaris i de Recerca (AGAUR, 2017 SGR 1322), and the CERCA Program/Generalitat de Catalunya. A.J.F. was funded by a Ramón y Cajal fellowship (RYC2021-033375-I) financed by the Spanish Ministry of Science and Innovation (MCIN/AEI /10.13039/501100011033) and the European Union (NextGenerationEU/PRTR).

11 Conflicts of interest

The authors declare no competing interests.

References

[1] Patrick C Phillips. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11):855–867, 2008.

[2] Frank J Poelwijk, Vinod Krishna, and Rama Ranganathan. The Context-Dependence of Mutations: A Linkage of Formalisms. *PLoS Computational Biology*, 12(6):e1004771, June 2016.

[3] Júlia Domingo, Pablo Baeza-Centurion, and Ben Lehner. The causes and consequences of genetic interactions (epistasis). *Annu. Rev. Genomics Hum. Genet.*, 20:433–460, August 2019.

[4] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nat. Methods*, 11(8):801–807, August 2014.

[5] J Arjan GM de Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15(7):480–490, 2014.

[6] Júlia Domingo, Guillaume Diss, and Ben Lehner. Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature*, 558(7708):117–121, June 2018.

[7] Frank J Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature communications*, 10(1):1–11, September 2019.

[8] Pablo Baeza-Centurion, Belén Miñana, Jörn M Schmiedel, Juan Valcárcel, and Ben Lehner. Combinatorial genetics reveals a scaling law for the effects of mutations on splicing. *Cell*, 176(3):549–563, 2019.

[9] Victoria O Pokusaeva, Dinara R Usmanova, Ekaterina V Putintseva, Lorena Espinar, Karen S Sarkisyan, Alexander S Mishin, Natalya S Bogatyreva, Dmitry N Ivankov, Arseniy V Akopyan, Sergey Ya Avvakumov, et al. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS genetics*, 15(4):e1008079, 2019.

[10] Devin P Bendixsen, James Collet, Bjørn Østman, and Eric J Hayden. Genotype network intersections promote evolutionary innovation. *PLoS biology*, 17(5):e3000300, 2019.

[11] Valerie WC Soo, Jacob B Swadling, Andre J Faure, and Tobias Warnecke. Fitness landscape of a dynamic rna structure. *PLoS genetics*, 17(2):e1009353, 2021.

[12] Alief Moulana, Thomas Dupic, Angela M Phillips, Jeffrey Chang, Serafina Nieves, Anne A Roffler, Allison J Greaney, Tyler N Starr, Jesse D Bloom, and Michael M Desai. Compensatory epistasis maintains ace2 affinity in sars-cov-2 omicron ba. 1. *Nature Communications*, 13(1):1–11, 2022.

[13] Rachapun Rotrattanadumrong and Yohei Yokobayashi. Experimental exploration of a ribozyme neutral network using evolutionary algorithm and deep learning. *Nature communications*, 13(1):1–14, 2022.

[14] David E Goldberg. Genetic algorithms and walsh functions: Part i, a genetle introduction. *Complex systems*, 3:129–152, 1989.

[15] Daniel M Weinreich, Yinghong Lan, C Scott Wylie, and Robert B Heckendorn. Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics & development*, 23(6):700–707, 2013.

[16] Frank J. Poelwijk and Rama Ranganathan. The relation between alignment covariance and background-averaged epistasis. *arXiv*, 10.48550/ARXIV.1703.10996, 2017.

[17] David H Brookes, Amirali Aghazadeh, and Jennifer Listgarten. On the sparsity of fitness functions and implications for learning. *Proc. Natl. Acad. Sci. U. S. A.*, 119(1), January 2022.

[18] C. Brandon Ogbunugafor. The mutation effect reaction norm (mu-rn) highlights environmentally dependent mutation effects and epistatic interactions. *Evolution*, 76(s1):37–48, 02 2022.

[19] Tom Beer. Walsh transforms. *American Journal of Physics*, 49(5):466–472, 1981.

[20] David S Stoffer. Walsh-fourier analysis and its statistical applications. *Journal of the American Statistical Association*, 86(414):461–479, 1991.

[21] Andre J Faure, Júlia Domingo, Jörn M Schmiedel, Cristina Hidalgo-Carcedo, Guillaume Diss, and Ben Lehner. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature*, 604(7904):175–183, 2022.

[22] Ammar Tareen, Mahdi Kooshkbaghi, Anna Posfai, William T Ireland, David M McCandlish, and Justin B Kinney. Mave-nn: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome biology*, 23(1):1–27, 2022.

[23] Jakub Otwinowski, David M McCandlish, and Joshua B Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115(32):E7550–E7558, 2018.

[24] Talitha L Forcier, Andalus Ayaz, Manraj S Gill, Daniel Jones, Rob Phillips, and Justin B Kinney. Measuring cis-regulatory energetics in living cells using allelic manifolds. *Elife*, 7:e40618, 2018.

[25] Justin B Kinney, Anand Murugan, Curtis G Callan Jr, and Edward C Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, 2010.

[26] Juannan Zhou, Mandy S Wong, Wei-Chia Chen, Adrian R Krainer, Justin B Kinney, and David M McCandlish. Higher-order epistasis and phenotypic prediction. *Proceedings of the National Academy of Sciences*, 119(39):e2204233119, 2022.

[27] Juannan Zhou and David M McCandlish. Minimum epistasis interpolation for sequence-function relationships. *Nature communications*, 11(1):1–14, 2020.

[28] Andre J Faure, Jörn M Schmiedel, Pablo Baeza-Centurion, and Ben Lehner. Dimsum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biology*, 21(1):1–23, 2020.

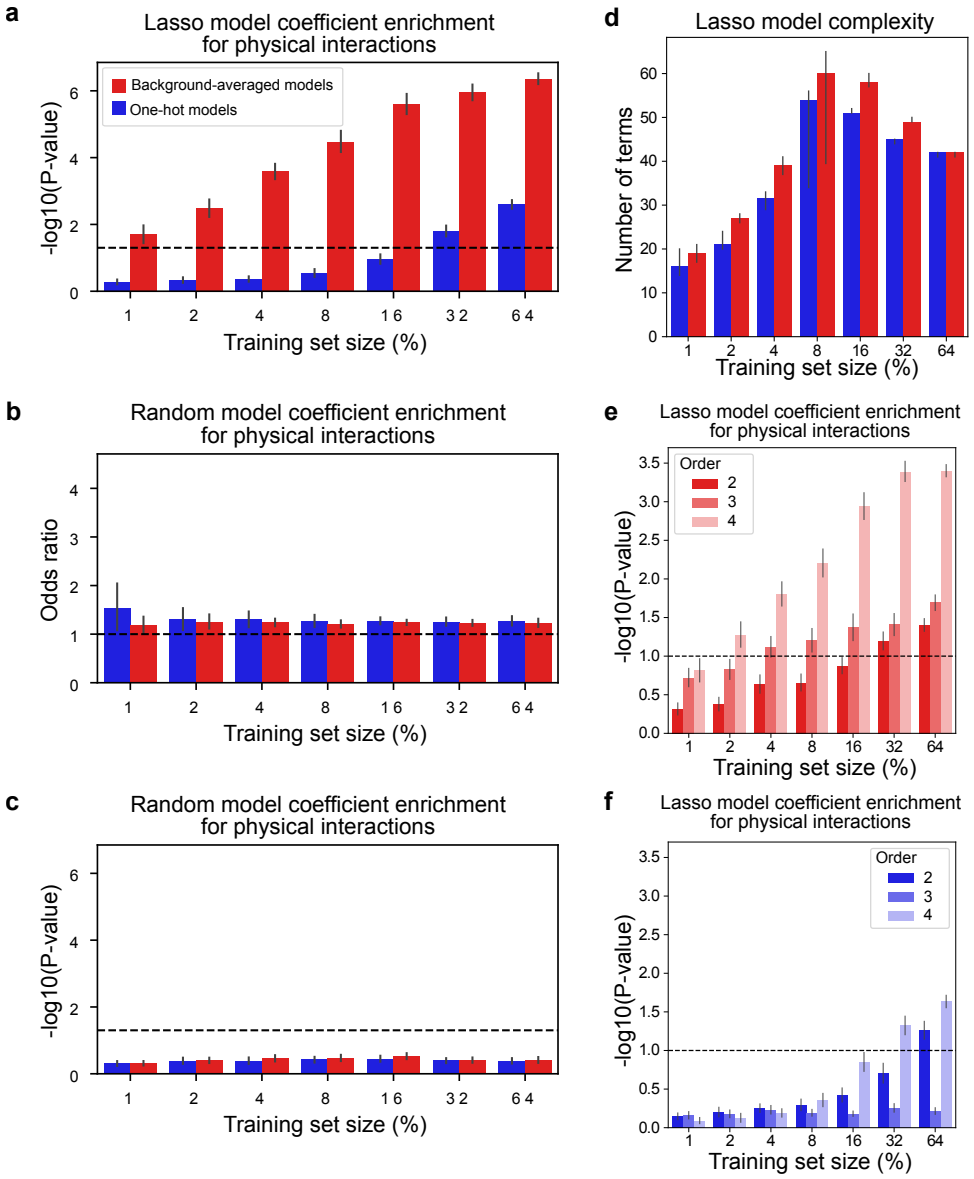

Figure S1: Supplementary figure related to Figure 2. **a**, P-value from Fisher's Exact Test for enrichment of direct physical interactions in non-zero epistatic coefficients (related to Figure 2g). **b**, Enrichment of direct physical interactions in non-zero epistatic coefficients of random models with matching numbers of epistatic coefficients of different orders (related to Figure 2g). P-value from Fisher's Exact Test for enrichment of direct physical interactions in non-zero epistatic coefficients of random models with matching numbers of epistatic coefficients of different order (related to panel b). **d**, Median number of epistatic terms in Lasso models (related to Figure 2f). **f**, P-value from Fisher's Exact Test for enrichment of direct physical interactions in non-zero epistatic shown separately for epistatic coefficients of different orders (realted to Figure 2i). Error bars indicate nonparametric bootstrap 95% confidence intervals of the mean in all panels, except in panel d where these correspond to the median.

12 Supplementary Methods

This section contains the proofs of the mathematical results shown in the main text.

Proposition 1. *Let us define the matrices A_n recursively as*

$$A_{n+1} = \frac{1}{s}\begin{pmatrix} A_n & A_n & A_n & \dots & A_n \\ A_n & (1-s)A_n & A_n & \dots & A_n \\ A_n & A_n & (1-s)A_n & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & A_n \\ A_n & A_n & \dots & A_n & (1-s)A_n \end{pmatrix} \qquad A_0 = 1 \qquad \textit{and} \qquad A_n \ \textit{is} \ s^n \times s^n. \tag{5}$$

For $n \in \mathbb{N}$, A_n is the inverse of the matrix H_n defined in (2).

Proof. Let us prove this by induction. For $n = 1$ we have

$$H_1 = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & -1 & 0 & \dots & 0 \\ 1 & 0 & -1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 1 & 0 & \dots & 0 & -1 \end{pmatrix} \tag{11}$$

$$A_1 = \frac{1}{s}\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1-s & 1 & \dots & 1 \\ 1 & 1 & 1-s & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 1 & 1 & \dots & 1 & 1-s \end{pmatrix}. \tag{12}$$

The rows and columns of these two matrices can be described as follows:

$$(H_1)_{i\cdot} = \begin{cases} (1, 1, \dots, 1) & \text{if } i = 1 \\ (1, 0, \dots, 0, h_i, 0, \dots, 0) & \text{if } i \neq 1, \end{cases}$$

$$(A_1)_{\cdot j} = \frac{1}{s}\begin{cases} (1, 1, \dots, 1)^T & \text{if } j = 1 \\ (1, 1, \dots, 1, a_j, 1, \dots, 1)^T & \text{if } j \neq 1, \end{cases}$$

where $h_i := (H_1)_{ii} = -1 \ \forall i > 1$ and $a_j := (A_1)_{jj} = 1 - s \ \forall j > 1$.

Therefore,

$$(H_1 \cdot A_1)_{ij} = (H_1)_{i\cdot} \cdot (A_1)_{\cdot j} = \frac{1}{s}\begin{cases} s & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

$$H_1 \cdot A_1 = \frac{1}{s}\begin{pmatrix} s & 0 & \dots & 0 \\ 0 & s & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & s \end{pmatrix} = I_{s\times s},$$

where $I_{s\times s}$ is the identity matrix of size $s \times s$. Since both H_1 and A_1 are symmetric, it is also true that $A_1 \cdot H_1 = I_{s\times s}$. Therefore, A_1 is the inverse of H_1.

Assume that the hypothesis is true for a fixed $n \in \mathbb{N}$. Let us now prove that it is also true for $n+1$. Following the recursive definitions of H_{n+1} and A_{n+1} in equations (2) and (5), we can write the blocks of these matrices as follows:

$$(H_{n+1})_{[i][\cdot]} = \begin{cases} (H_n, H_n, \dots, H_n) & \text{if } i = 1 \\ (H_n, 0, \dots, 0, \tilde{h}_i, 0, \dots, 0) & \text{if } i \neq 1, \end{cases} \tag{13}$$

where $(H_{n+1})_{[i][j]}$ denotes the block at position i, j in H_{n+1} and $\tilde{h}_i := (H_{n+1})_{[i][i]} = -H_n \ \forall i > 1$;

$$(A_{n+1})_{[\cdot][j]} = \frac{1}{s}\begin{cases} (A_n, A_n, \dots, A_n)^T & \text{if } j = 1 \\ (A_n, A_n, \dots, A_n, \tilde{a}_j, A_n, \dots, A_n)^T & \text{if } j \neq 1, \end{cases} \tag{14}$$

where $(A_{n+1})_{[i][j]}$ denotes the block at position i, j in A_{n+1} and $\tilde{a}_j := s(A_{n+1})_{[j][j]} = (1-s)A_n \ \forall j > 1$. We can therefore write the block at position i, j of the product of these matrices as follows:

$$(H_{n+1} \cdot A_{n+1})_{[i][j]} = (H_{n+1})_{[i][\cdot]} \cdot (A_{n+1})_{[\cdot][j]} = \frac{1}{s}\begin{cases} sH_n \cdot A_n & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

According to the induction hypothesis we know that H_n and A_n are inverse matrices i.e. $H_n \cdot A_n = I_{s^n\times s^n}$. Therefore, the blocks on the diagonal are identity matrices and the blocks outside the diagonal are zeros. This means that $H_{n+1} \cdot A_{n+1} = I_{s^{n+1}\times s^{n+1}}$. Similarly, due to the symmetry of the matrices we can also prove that $A_{n+1} \cdot H_{n+1} = I_{s^{n+1}\times s^{n+1}}$.

We can then conclude that $A_n = H_n^{-1}$ for every $n \in \mathbb{N}$. □

Proposition 2. *The elements of H_n can be written as*

$$(H_n)_{ij} = \begin{cases} (-1)^{(E_n)_{ij}} & \textit{if } (M_n)_{ij} = n \\ 0 & \textit{otherwise,} \end{cases}$$

where M and E are $s^n \times s^n$ matrices whose elements are

$$(E_n)_{ij} = \sum_{\substack{k=1 \\ i_k \cdot j_k > 0}}^{n} \delta_{i_k j_k} \tag{6}$$

$$(M_n)_{ij} = \sum_{\substack{k=1 \\ i_k \cdot j_k > 0}}^{n} \delta_{i_k j_k} + \sum_{\substack{k=1 \\ i_k \cdot j_k = 0}}^{n} 1 = (E_n)_{ij} + \sum_{\substack{k=1 \\ i_k \cdot j_k = 0}}^{n} 1,$$

where δ_{ij} denotes the Kronecker delta of i, j.

Proof. Let us prove the formula by induction. For $n = 1$ and any given s, H_n is given by (11). Therefore, we can write $(E_1)_{ij}$ and $(M_1)_{ij}$ as follows:

$$(M_1)_{ij} = \begin{cases} 1 & \text{if } i = 1 \text{ or } j = 1 \\ 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \neq 1 \end{cases}$$

$$(E_1)_{ij} = \begin{cases} 1 & \text{if } i = j \neq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Therefore, since $n = 1$, $(M_1)_{ij} = n = 1$ only when either $i = 1$, $j = 1$ or $i = j$. In the rest of the cases $(M_1)_{ij} \neq n = 1$ and, according to the formula, the elements of the matrix will be 0. Now, for the cases where $(M_1)_{ij} = n = 1$, we need to check the value of $(E_1)_{ij}$. We can see how $(E_1)_{ij} = 1$ only when $i = j$ and they are different from 1. This means that all the elements of the diagonal of H_1, except the first one, will be $(-1)^1 = -1$ and the first row and first columns will have $(-1)^0 = 1$. The rest of the elements correspond to $(M_1)_{ij} \neq n = 1$ so they will be filled with zeros. Putting all this together, we find the expression as H_1 from equation (11).

Assume now that the expression is true for a fixed $n \in \mathbb{N}$ and let us prove it for $n + 1$. In this case, the matrix of H_{n+1} is defined by blocks (2). We first define the indices $P \in \{1, \dots, s\}$ and $Q \in \{1, \dots, s\}$ for row and column blocks, respectively. The first matrix block of H_{n+1} corresponds to $P = 1$ and $Q = 1$. The corresponding blocks of the matrices M_{n+1} and E_{n+1}, which are necessary for the derivation of the block in H_{n+1}, are computed by comparing the genotype sequences $\vec{p}, \vec{q} \in S^{n+1}$ for which $\vec{p} = (0, i_1, \dots, i_n) := 0 \frown \vec{i}$

for $\vec{i} \in \mathcal{S}^n$ and $\vec{q} = 0 \frown \vec{j}, \vec{j} \in \mathcal{S}^n$. More generally, the block in the Pth position with respect to the rows and Q^{th} position with respect to the columns can be obtained by comparing the genotype sequences $\vec{p}, \vec{q} \in \mathcal{S}^{n+1}$ for which $\vec{p} = (P-1) \frown \vec{i}, \vec{i} \in \mathcal{S}^n$ and $\vec{q} = (Q-1) \frown \vec{j}, \vec{j} \in \mathcal{S}^n$. See Figure S2 for a visual description of the notation.

From these observations, it can easily be deduced that for any $\vec{p} = (P-1) \frown \vec{i}, \vec{q} = (Q-1) \frown \vec{j}$ with $\vec{i}, \vec{j} \in \mathcal{S}^n$, $P-1, Q-1 \in \mathcal{S}$,

$$(M_{n+1})_{pq} = \begin{cases} (M_n)_{ij} + 1 & \text{if } P = 1 \text{ or } Q = 1 \\ (M_n)_{ij} + 1 & \text{if } P = Q \\ (M_n)_{ij} & \text{if } P \neq Q \neq 1 \end{cases} \tag{16}$$

$$(E_{n+1})_{pq} = \begin{cases} (E_n)_{ij} + 1 & \text{if } P = Q \neq 1 \\ (E_n)_{ij} & \text{otherwise,} \end{cases} \tag{17}$$

where $i = p - s^n(P-1)$, $j = q - s^n(Q-1)$, $P = \lceil p/s^n \rceil$ and $Q = \lceil q/s^n \rceil$, with $\lceil . \rceil$ denoting the ceiling function.

$$X_{n+1} = \begin{array}{c} \\ \text{row block } P \\ \\ \\ \end{array} \begin{array}{c} \overbrace{\phantom{X_n^{1Q}}}^{\text{column block } Q} \\ \begin{pmatrix} X_n^{11} & \dots & X_n^{1Q} & \dots & X_n^{1s} \\ \vdots & \ddots & \vdots & & \vdots \\ X_n^{P1} & \dots & X_n^{PQ} & \dots & X_n^{Ps} \\ \vdots & & \vdots & \ddots & \vdots \\ X_n^{s1} & \dots & X_n^{sQ} & \dots & X_n^{ss} \end{pmatrix} \\ \uparrow \\ \text{column } q \end{array} \begin{array}{c} \\ \\ \leftarrow \text{row } p \\ \\ \\ \end{array}$$

Figure S2: Description of the notation of the block structure of the matrices. Here X_{n+1} denotes any generic matrix following the structure of the matrices in (2), (3), (5), (16) and (17).

Now, similar to the case $n = 1$, we have that $(M_{n+1})_{pq} = n + 1$ only when $(M_n)_{ij} = n$ and either $P = 1$, $Q = 1$ or $P = Q$, which corresponds to the newly added state being $p_1 = P - 1 = 0$, $q_1 = Q - 1 = 0$ or $p_1 = q_1$. In the rest of the cases $(M_{n+1})_{pq} \neq n + 1$ and the elements of the matrix H_{n+1} will be 0. Now, for the cases where $(M_{n+1})_{pq} = n+1$, we will have that $(E_{n+1})_{pq}$ has either the same value of the corresponding entry in E_n or it will be increased by 1. This means, that when $P = Q \neq 1$, the sign of the entry in H_{n+1} will be inverted. Otherwise, the sign of the element of the matrix stays the same. With this we prove the formula for H_{n+1}, since we have shown how we can find the same block structure as in (2). By induction, we can conclude that the formula holds for every $n \in \mathbb{N}$. □

Proposition 3. *The elements of A_n can be written as*

$$(A_n)_{ij} = \frac{1}{s^n}(1-s)^{(E_n)_{ij}}, \tag{7}$$

where E_n is defined as in (6).

Proof. Let us prove the formula by induction.

For any given s, A_1 is defined in (12). Its diagonal elements are equal to $(1-s)/s$ for $i > 1$, $1/s$ for $i = 1$ and its off-diagonal elements are equal to $1/s$. It can easily be observed from (15) that the RHS of (7) is equal to $(1-s)/s$ for $i = j \neq 1$ and to $1/s$ otherwise, so (7) is true for $n = 1$.

Assume now that (7) is true for a fixed $n \in \mathbb{N}$ and let us prove it for $n+1$. We use the recursive definition of A_{n+1} in (5) and the block representation of the genotype sequences as in the proof of 2.

Let us start with the first block in the diagonal of A_{n+1}, i.e. $P = 1$ and $Q = 1$, where the entries of the corresponding block in E_{n+1} are derived from comparisons of pairs of genotype sequences of the form $\vec{p} = 0 \frown \vec{i}, \vec{q} = 0 \frown \vec{j}$ for $\vec{i}, \vec{j} \in \mathcal{S}^n$. From (5), this block is equal to $\frac{1}{s}A_n$, so writing $i = p \bmod s^n$ and $j = q \bmod s^n$, we have

$$(A_{n+1})_{pq} = \frac{1}{s}(A_n)_{ij} = \frac{1}{s^{n+1}}(1-s)^{(E_n)_{ij}}.$$

From (17), $(E_{n+1})_{pq} = (E_n)_{ij}$, which yields the desired result.

Now let us consider the elements in the other diagonal blocks of A_{n+1}, where the entries correspond to pairs of genotype sequences of the form $\vec{p} = (P-1) \frown \vec{i}, \vec{q} = (Q-1) \frown \vec{j}$ with $P = Q \neq 1$ and $\vec{i}, \vec{j} \in \mathcal{S}^n$. From (5), this block is equal to $\frac{1}{s}(1-s)A_n$, i.e.

$$(A_{n+1})_{pq} = \frac{1}{s}(1-s)(A_n)_{ij} = \frac{1}{s^{n+1}}(1-s)^{(E_n)_{ij}+1}.$$

From (17), $(E_{n+1})_{pq} = (E_n)_{ij} + 1$, which yields the desired result.

Finally, let us consider the elements in the off-diagonal blocks of A_{n+1}, where the entries correspond to pairs of genotype sequences of the form $\vec{p} = (P-1) \frown \vec{i}, \vec{q} = (Q-1) \frown \vec{j}$ with $P \neq Q$ and $\vec{i}, \vec{j} \in \mathcal{S}^n$. From (5), this block is equal to $\frac{1}{s}A_n$, so we have

$$(A_{n+1})_{pq} = \frac{1}{s}(A_n)_{ij} = \frac{1}{s^{n+1}}(1-s)^{(E_n)_{ij}}.$$

From (17), $(E_{n+1})_{pq} = (E_n)_{ij}$, which completes the proof. □

Proposition 4. *The matrices V_n and V_n^{-1} are diagonal matrices whose diagonal elements can be written as*

$$(V_n)_{ii} = (-1)^{n-W_n(\vec{i})}\frac{1}{s^{W_n(\vec{i})}} \tag{8}$$

and

$$(V_n^{-1})_{ii} = (-1)^{n-W_n(\vec{i})} s^{W_n(\vec{i})}, \tag{9}$$

where

$$W_n(\vec{i}) := \sum_{k=1}^{n} w_k, \text{ with } w_k := \delta_{i_k 0}$$

and $\vec{i}$ again denotes the ith element in $\mathcal{S}^n$ when ordered by the base s representation of integers.

Proof. Let us prove (8) by induction, (9) follows directly.

One can easily check from (3) that the formula holds for $n = 1$. Let us now assume that (8) is true for a fixed $n \in \mathbb{N}$ and let us prove it for $n + 1$. We use the recursive definition of V_{n+1} in (3). Let us consider the element $(V_{n+1})_{pp}$. If $\vec{p} = 0 \frown \vec{i}$, this corresponds to the first block of V_{n+1}, i.e. $P = 1$, where the elements are multiplied by $1/s$ and $W_{n+1}(\vec{p}) = W_n(\vec{i}) + 1$, so

$$(V_{n+1})_{pp} = \frac{1}{s}(V_n)_{ii} = (-1)^{n-W_n(\vec{i})} \frac{1}{s^{W_n(\vec{i})+1}} = (-1)^{n+1-W_{n+1}(\vec{p})} \frac{1}{s^{W_{n+1}(\vec{p})}}.$$

Similarly, if $\vec{p} = (P-1) \frown \vec{i}$ and $P > 1$, i.e. for the other diagonal blocks, from the recursive formula the elements are multiplied by -1 and $W_{n+1}(\vec{p}) = W_n(\vec{i})$, so by writing again $i = p - s^n(P-1)$, where $P = \lceil p/s^n \rceil$, we have

$$(V_{n+1})_{pp} = -(V_n)_{ii} = (-1)^{1+n-W_n(\vec{i})} \frac{1}{s^{W_n(\vec{i})}} = (-1)^{n+1-W_{n+1}(\vec{p})} \frac{1}{s^{W_{n+1}(\vec{p})}},$$

which completes the proof. □

Proposition 5. *The matrix A_n defined in* (10) *is the inverse of the matrix H_n in the general case where each position can have a different number of states.*

Proof. Let us prove by induction that $H_n \cdot A_n = I$ where I is the identity matrix of the corresponding size. Since H_n and A_n are symmetric, this would imply that $A_n \cdot H_n = I$ as well, and therefore, $A_n = H_n^{-1}$.

The case $n = 1$ corresponds exactly to the case $n = 1$ of the proof of Proposition 1 by setting $s = s_1$. Therefore, $H_1 A_1 = I_{s_1 \times s_1}$, and A_1 is the inverse of H_1.

Now, assume the hypothesis is true for a fixed $n \in \mathbb{N}$ and let us prove that this is also true for $n + 1$. We can write the rows and columns of the matrices H_{n+1} and A_{n+1} as (13) and (14), respectively. The only difference is that we need to replace s by s_{n+1} and the size of the matrices is different. Following exactly the same derivation as in Proposition 1 we can conclude that $H_{n+1} A_{n+1} = I$ and this proves by induction that $A_n = H_n^{-1}$. □

Proposition 6. *In this general case, the elements of H_n and A_n can be written as*

$$(H_n)_{ij} = \begin{cases} (-1)^{(E_n)_{ij}} & \textit{if } (M_n)_{ij} = n \\ 0 & \textit{otherwise} \end{cases}$$

$$(A_n)_{ij} = \frac{\prod_{k=1}^{n} \left(1 - s_k\right)^{e_k}}{\prod_{k=1}^{n} s_k},$$

where E_n *and* M_n *are defined as in* (6) *and* $e_k = \begin{cases} 1 & \textit{if } i_k = j_k \neq 1 \\ 0 & \textit{otherwise} \end{cases}$.

Proof. The proof follows directly from the proofs of Propositions 2 and 3. The only difference in the induction step is that s is replaced by s_{n+1}. □

Proposition 7. *The matrices* V_n *and* V_n^{-1} *are diagonal matrices whose diagonal elements can be written as*

$$(V_n)_{ii} = (-1)^{n-W_n(\vec{i})} \prod_{k=1}^{n} \left(\frac{1}{s_k}\right)^{w_k}$$

and

$$(V_n^{-1})_{ii} = (-1)^{n-W_n(\vec{i})} \prod_{k=1}^{n} {s_k}^{w_k},$$

where

$$W_n(\vec{i}) := \sum_{k=1}^{n} w_k, \textit{ with } w_k := \delta_{i_k 0}.$$

Proof. The proof follows the same steps as the proof of Proposition 4. The only difference in the induction step is that s is replaced by s_{n+1}. □