In this document, we provide supplementary materials for the background, methods, and results sections.

1 Background

1.1 The Term Epistasis

Epistasis has been used to denote two contextually distinct phenomena. The first, referred to as either biological (Moore and Williams, 2005), physical (Brodie III, 2000), or functional (Phillips, 2008) epistasis, is the result of physical interactions among biomolecules within gene regulatory networks and biochemical pathways. The second, referred to as statistical (Moore and Williams, 2005; Phillips, 2008), or populational (Brodie III, 2000) epistasis, or more classically as episty (Fisher et al., 1918), has been defined as a deviation from additivity in a mathematical model summarizing the relationship between multi-locus genotypes and phenotypic variation in a population.

More precise definitions of what makes a model epistatic have been based on a deviation from independence (Frankel and Schork, 1996; Wade et al., 2001; Cordell, 2002), additivity (Fisher et al., 1918; Hallgrímsson and Yuster, 2008), or multiplicativity (Morris and Whittaker, 1999; Wade et al., 2001; Wilson, 2001).

1.1.1 More on Impure Epistasis

In quantitative trait genetics, impure epistatic models have been decomposed into their main effect and epistatic components (Cockerham, 1954; Cheverud and Routman, 1995; Hallgrímsson and Yuster, 2008). Considering these decompositions we view impure models of epistasis as populating the space between pure epistasis and non-epistatic models, with varying underlying magnitudes of additive, multiplicative, and epistatic effect.

2 Methods

Calculations of marginal penetrances for Table 1 are presented as follows;
Table 1: A 2-locus purely epistatic penetrance function.

| SNP 1 Genotype | SNP 2 | Marginal Penetrance |
|----------------|-------|---------------------|
|                | BB(.25) | Bb (.5) | bb(.25) |                |
| AA(.36)        | .266    | .764     | .664    | .614            |
| Aa (.48)       | .928    | .398     | .733    | .614            |
| aa(.16)        | .456    | .927     | .147    | .614            |
| Marginal       | .614    | .614     | .614    | K = .614        |

\[ \langle .25, .5, .25 \rangle \cdot \langle .266, .764, .664 \rangle \]
\[ = \langle .25, .5, .25 \rangle \cdot \langle .928, .398, .733 \rangle \]
\[ = \langle .25, .5, .25 \rangle \cdot \langle .456, .927, .147 \rangle \]
\[ = \langle .36, .48, .16 \rangle \cdot \langle .266, .928, .456 \rangle \]
\[ = \langle .36, .48, .16 \rangle \cdot \langle .764, .398, .927 \rangle \]
\[ = \langle .36, .48, .16 \rangle \cdot \langle .664, .733, .147 \rangle \approx .614 \]

2.1 Describing n-locus models

Recall that n SNPs are strictly and purely epistatic if together they are predictive of disease, but no proper subset of them is. Some consequences of this definition are discussed after a preliminary paragraph on terminology.

We assume that n SNPs in any individual are in one of three states (1, 2 or 3) corresponding to the genotypes AA, Aa and aa. Let \( p_{ij} \) be the probability that SNP \( i \) of a random individual is in state \( j \), and let \( \vec{P}_i = \langle p_{i1}, p_{i2}, p_{i3} \rangle \). The vector \( \vec{P}_i \) arranges the genotype frequencies of SNP \( i \). In table 1, \( \vec{P}_1 = \langle .25, .5, .25 \rangle \) and \( \vec{P}_2 = \langle .36, .48, .16 \rangle \). These vectors are used in equations 1 through 6.

An n-locus penetrance function consists of \( 3^n \) penetrance values, one corresponding to each possible MLG. Entries are typically denoted by \( f_{i_1 \ldots i_n} \). Here the subscript \( i_1 \ldots i_n \) indicates the MLG where SNP 1 has state \( i_1 \), SNP 2 state \( i_2 \), up to SNP \( n \), which has state \( i_n \). \( f_{i_1 \ldots i_n} = P(s|i_1 \ldots i_n) \) is the probability that an individual has disease if their MLG is \( i_1 \ldots i_n \). For example, \( f_{213} \) indicates the penetrance value within a 3-locus model corresponding to the MLG where SNP 1 is in state Aa, SNP 2 is in state BB and SNP3 is in state cc.

We also assume that the states of different SNPs are independent. Thus the product \( p_{i_1 i_2 \ldots p_{i_n}} \) of genotype frequencies is the probability that an individual has MLG \( i_1 \ldots i_n \). For example, with the genotype frequencies in Table 1, the probability that genotype AAbb occurs is \( p_{11}p_{23} = .36 \cdot .25 = .09 \).

2.2 Computing population prevalence \( K \) and heritability \( h^2 \)

We abbreviate the notation for multivariate genotypes to \( G \). The population prevalence for any n-locus model can be computed by

\[ K = \sum_G P(G)f_G . \]
2.4 Generating Pure, Strict Epistatic Products with Common Value

are given in equations (2) through (6).

pure epistasis which is equivalent to ours. In the example of Table 1, the 2
have the same value. The equality of these dot products gives an alternative definition of strict,
to $K$.

It is convenient here to work with what we call pre-penetrance functions. These are, by definition,
and the dot products discussed in §2.3 are required to be
$K$.

Similarly, for $j = 1, \ldots, n$, the dot products of $\vec{P}_j$ with all the $j$-stiffs are necessarily equal
to $K$. Thus for a strict, pure, epistatic, $n$-locus penetrance function these $3^{n-1}$ dot products
have the same value. The equality of these dot products gives an alternative definition of strict,
epistasis which is equivalent to ours. In the example of Table 1, the $2 \cdot 3^{2-1} = 6$ dot
products with common value are given in equations (2) through (6).

2.3 Defining Strict, Pure, $n$-locus Epistatic Models

We use the term $j$-staff to refer to any of the $3^{n-1}$ vectors of the form

\[(f_{i_1 \ldots i_{j-1}, j, i_{j+1} \ldots i_n}, f_{i_1 \ldots i_{j-1}, j, i_{j+1} \ldots i_n}).\]

Note that here the $j$th subscript of the three components are 1, 2 and 3 (indicating the genotype
of the $i$th SNP), while the other subscripts are identical in the three components. As an example,
the vector $(f_{13}, f_{23}, f_{33}) = (0.664, 0.733, 1.147)$ making up the third column of the penetrance
function in Table 1 is a 1-staff. A staff is any $j$-staff, where $j \in \{1, \ldots, n\}$. Staffs generalize
the notion of rows and columns of 2-locus penetrance functions to $n$-locus penetrance functions.

Now assume that the MLG of an individual is known except for the state of the $j$th SNP
($J$); say his genotype is $i_1 \ldots i_{j-1}, J, i_{j+1} \ldots i_n$. Then, as in equations (2) through (6),
the probability the individual has disease is the dot product of $\vec{P}_j$ with the $j$-staff displayed above.
If the $n$ SNPs are strictly and purely epistatic, this probability must be the prevalence $K$ since the
 genotype $i_1 \ldots i_{j-1}, J, i_{j+1} \ldots i_n$ offers no predictive value beyond $K$.

Similarly, for $j = 1, \ldots, n$, the dot products of $\vec{P}_j$ with all the $j$-stiffs are necessarily equal
to $K$. Thus for a strict, pure, epistatic, $n$-locus penetrance function these $n3^{n-1}$ dot products
have the same value. The equality of these dot products gives an alternative definition of strict,
epistasis which is equivalent to ours. In the example of Table 1, the $2 \cdot 3^{2-1} = 6$ dot
products with common value are given in equations (2) through (6).

2.4 Generating Pure, Strict Epistatic $n$-Way Models

It is convenient here to work with what we call pre-penetrance functions. These are, by definition,
like strict, pure, epistatic penetrance functions, but their entries are not required to lie
in the interval $[0, 1]$, and all of the $n3^{n-1}$ dot products discussed in §2.3 are required to be
zero. In §2.4.5 we will apply a linear function to all the entries of a pre-penetrance function
to convert it to a penetrance function. This requirement on the dot products assures that the
converted table is strictly and purely epistatic.

The entry of a pre-penetrance function corresponding to the MLG $i_1 \ldots i_n$ is denoted by
$g_{i_1 \ldots i_n}$. We usually think of each $g_{i_1 \ldots i_n}$ as a variable since its value depends on the particular
pre-penetrance function. We show in §2.4.1 and later sections that assigning values to certain choices
of $2^n$ of these variables determines a unique pre-penetrance function. The $2^n$ variables
making up any such choice are referred to as parameters. In other words, parameters are
independent variables selected in order to generate a pre-penetrance function. Indeed, assigning
any values to them uniquely determines the values of the remaining $3^n - 2^n$ variables, the
dependent ones, of a pre-penetrance function.

GAMETES usually works with the variable $p_{11}, \ldots, p_{n1}, g_{i_1 \ldots i_n}$ rather than $g_{i_1 \ldots i_n}$. Incor-porating the probabilities $p_{11}, \ldots, p_{n1}$ into the notation in this way simplifies expressions
for dot products considerably, cf. equations (9), (10) and (11) below. We use $a_{i_1...i_n}$ to denote $p_{i_1} \cdots p_{i_n} \cdot g_{i_1...i_n}$ if the corresponding entry $g_{i_1...i_n}$ is a parameter and $b_{i_1...i_n}$ if it is a dependent variable.

Our strategy for generating pure, strict epistatic models involves the following steps; (1) Generate random parameters; choose $2^n$ entries of an $n$-locus table as independent parameters and specify all remaining entries (i.e. dependent variables) in terms of these, (2) Determine pre-penetrance functions; randomly seed values for the independent parameters and compute remaining dependent variables, (3) Determine penetrance functions; linearly scale the entries of the pre-penetrance function to lie in the penetrance value interval $[0, 1]$, and (4) transform the penetrance function to the desired heritability and prevalence.

We considered four strategies for generating random parameters; the Standard Parameter Method, the Point Method, the Random Parameter Method, and the Sudoku Method.

2.4.1 Standard Parameter Method

This method always selects the same $2^n$ variables as parameters making it the simplest but also the most biased method. For example, heritabilities of the resulting penetrance functions tend to fall in a limited range when this method is used.

The standard parameters are, by definition, the variables $a_{i_1...i_n}$ with none of the $n$ subscripts equal to 1. This leaves two choices, namely 2 and 3, for each of the subscripts. Thus there are $2^n$ standard parameters. Our choice of standard parameters is an arbitrary decision.

Next, we show that if values are assigned to the standard parameters, then there is a unique pre-penetrance function with these values. In other words, the standard parameters uniquely determine the entire pre-penetrance function.

Given values for the standard parameters, the values of the dependent variables are computed as follows. First, the value of each variable $b_{i_1...i_n}$ with exactly one subscript equal to 1 is computed using the unique staff containing it and two standard parameters. For example, in the 3-locus case the value of $b_{i,1,1}$, where $i \neq 1 \neq j$, is computed using

$$b_{i,j} + a_{i,2,j} + a_{i,3,j} = 0.)

Then the value of each variable $b_{i_1...i_n}$ with exactly two subscripts equal to 1 is computed using either of the two staffs containing it and the two variables previously computed with exactly one subscript equal to 1. In the 3-locus case $b_{i,1,1}$, where $i \neq 1$, can be computed using either

$$b_{i,1,1} + b_{i,2,1} + b_{i,3,1} = 0 \quad \text{or} \quad b_{i,1,1} + b_{i,1,2} + b_{i,1,3} = 0.)

In $n$-locus terms each step involves computing the $b_{i_1...i_n}$ with exactly $u$ ones in their subscripts using the precomputed variables having subscripts containing $u - 1$ ones. This can be done using any of the $u$ staffs containing the variable and two precomputed variables having $u - 1$ ones in their subscripts. To see that the choice of the staff does not matter or, equivalently, that all $u$ dot products involving the staffs containing the variable equal zero, note that

$$b_{i_1...i_n} = (-1)^{u} \sum_{i_1 \in \{2,3\}} b_{i_1,1,1,\ldots, u} \cdots$$

$$= (-1)^{2} \sum_{i_1, i_2 \in \{2,3\}} b_{i_1, i_2,1,\ldots, u}$$
(11) \[ \sum_{i_1, \ldots, i_u \in \{2,3\}} a_{i_1, \ldots, i_u+1, \ldots, u_n} \times (-1)^u. \]

Here we have first used the staff containing \( b_1, \ldots, 1_{i_u+1}, \ldots, n \), \( b_2, 1_{i_u+1}, \ldots, i_u \), and \( b_3, \ldots, 1_{i_u+1}, \ldots, n \). Then, using another two staffs, these last two variables are expressed in terms of variables having \( u - 2 \) subscripts equal to one. The process continues until \( b_1, \ldots, 1_{i_u+1}, \ldots, n \) is expressed in terms of the standard parameters. While we have made choices for the staffs used, the last line of the above equation shows that these do not effect the value of \( b_1, \ldots, 1_{i_u+1}, \ldots, n \). The same conclusion holds regardless of where the ones are positioned in the subscripts. We point out that equation (11) expresses the dependent variables in terms of the standard parameters.

Note that the number of variables with \( u \) ones in their subscripts is \( \binom{n}{u} \). The number of staffs containing each such variable and two other variables having one fewer one in their subscripts is \( u \). Thus the total number of distinct dot products shown to be zero is

\[ \sum_{u=1}^{n} u \binom{n}{u} 2^{n-u} = n3^{n-1} \]

where the last equality comes from differentiating the binomial expansion for \((2 + x)^n\) and then setting \( x = 1 \). This shows that the standard parameters determine a unique pre-penetration function.

2.4.2 The Point Method

Like the standard parameter method, the Point method always finds parameters. These vary somewhat with each iteration making the method less biased than the standard parameter method, but more biased than the two methods described in §2.4.3 and §2.4.4.

The Point method consists of picking one variable, a reference point, \( b_{i_1, \ldots, i_n} \) at random from all \( 3^n \) variables and then using as parameters all the \( a_{j_1, \ldots, j_n} \) with \( j_\ell \neq i_\ell \) for \( \ell = 1, \ldots, n \). Dependent variables are then computed in terms of these parameters as they are with the standard parameter method.

2.4.3 Random Parameter Method

Our discussion of the standard parameter method showed that the set of all \( n \)-locus pre-penetration functions is parametrized by \( 2^n \) variables. Some choices of \( 2^n \) of the \( 3^n \) variables are parameters, some are not. For example, any choice with three variables in one staff does not give parameters. Nor does a choice with four variables in two staffs which share one variable, as exemplified by the four light grey boxes in each of A-D in Figure 2 of the main paper. Less obvious examples of variables which are not parameters arise when more than two loci are considered. See Figure 1.

The random parameter method consists of randomly picking \( 2^n \) variables repeatedly until parameters are obtained. Determining if chosen variables are parameters can be done by expressing each of the chosen variables in terms of the standard parameters using equation (11), and then determining if the resulting system of linear equations can be solved for the standard parameters in terms of the chosen variables. If they can or, equivalently, if the matrix \( M \) of coefficients of this linear system is invertible, then the variables are parameters. For example, the matrix \( M \) corresponding to the nine variables shown in Figure 1 has rank seven. Removing any row from \( M \) does not increase its rank, so any eight of these nine variables are not parameters. Alternatively, the matrix corresponding to the eight variables marked with a shown in Figure 2 is invertible, so these variables are parameters.
Figure 1: Example of failed starting positions for a 3-locus interaction. Each of the 27 cells corresponds to a different MLG (combining 3 SNPs), where each 3x3 table corresponds to one genotype of one of these SNPs. Any eight of the nine indicated positions may be chosen as starting positions but will not yield a pre-penetrance function.

Figure 2: An invertible 3-locus interaction matrix for which the Sudoku method fails to complete a pre-penetrance function, but still yields valid parameters. Starting points are indicated with a’s, and the only three positions which could be filled in using Sudoku are marked as b’s. This occurs regardless of the order in which the eight positions are chosen.

The random parameter method is the least biased method of generating pre-penetrance functions we discuss. However, it is also the least efficient since determining if $2^n$ variables are parameters requires attempting to invert a matrix, which often fails. That is, many choices of variables do not give parameters. For instance, in the 2-locus case, of the $\binom{9}{4} = 126$ ways to pick four variables, 45 do not give parameters, and so only about 64% iterations of the method produce pre-penetrance functions.

2.4.4 The Sudoku Method

This method differs from the others in that the parameters are not all selected initially. Rather, they are chosen sequentially with table entries computed after each choice. This provides a much more efficient method of finding parameters than the Random Parameter Method. However, it is also more biased since a few possibilities for $2^n$ variables which are parameters can not be chosen (see Figure 2).

This method starts by randomly picking two variables. If they happen to be in the same staff, the third entry of this staff is filled. By filling an entry we mean expressing it in terms of chosen variables. A third variable is then chosen from among the non-filled entries and any entries which can now be filled, are. At this step (and later ones), filled entries are naturally expressed in terms of chosen variables and previously filled entries. But the latter, in turn, can be expressed in terms of chosen variables. In $n$-locus terms, for each step of the method, a variable is chosen at random from the non-filled positions and then any entries which can be filled by completing staffs are filled. If all entries of an $n$-locus table are filled after picking $2^n$ variables, we say that the method succeeds. Otherwise, it fails.

We show below that if the Sudoku Method succeeds then

(a) the $2^n$ chosen variables are parameters and
(b) the pre-penetrance function determined by these parameters is that given by the Sudoku method.
Unfortunately, the Sudoku method sometimes fails. This can occur both when the $2^n$ chosen variables are parameters, as exemplified in Figure 2, and when they are not, as when any eight of the nine variables shown in Figure 1 are chosen. Which case holds can be determined with a matrix inversion, as in §2.4.3.

Proof of (a). Form the matrix $M$ of the $2^n$ linear equations which express the $2^n$ chosen variables in terms of the standard parameters. For each variable in a staff which is computed by the Sudoku Method, add a row to $M$ which corresponds to the equation expressing the variable in terms of the standard parameters. The resulting matrix $M^*$ has the same rank as $M$ since its rows are linear combinations of those of $M$. If the Sudoku method completes the table then the matrix $M^*$ has $3^n$ rows. The $2^n$ rows corresponding to the standard parameters form the identity matrix (up to permuting rows) together with a column whose entries are the standard parameters expressed in terms of the chosen variables. Thus $M^*$, and so also $M$, has maximal rank. Hence the variables are parameters.

Proof of (b). Assume now, that for a given choice of $2^n$ variables, the Sudoku method fills all entries. Then, as just proved, the matrix $M$ used above has maximum rank, so is invertible. Write $C = MA$, where $A$ and $C$ are the column matrices with entries the standard parameters and, respectively, the $2^n$ chosen variables, denoted by $c_I$. Let $F$ be the pre-penetration function determined by equation (11) with the values of the standard parameters given by $A = M^{-1}C$. Then the chosen variables of $F$ are just the $c_I$. The other entries of $F$ also agree with those given by the Sudoku method since whenever two entries of a staff in two pre-penetration functions agree the third entry also agrees.

2.4.5 From Parameters to Random Penetrance Functions

Throughout the rest of this section, we use the notation of 2-locus interactions for convenience, but all the arguments hold for any number of interacting SNPs.

Once parameters have been obtained for $n$-locus pre-penetration functions using any of the four methods discussed above, all $n$-locus pre-penetration functions are easily described. Indeed, there is a one-to-one correspondence between all values of the parameters and all pre-penetration functions. Using this correspondence, we pick a random direction in the space of all $n$-locus pre-penetration functions by picking a random direction, i.e. a non-zero vector, in $\mathbb{R}^{2^n}$, the parameter space (see Figure 3). The choice of this vector is made using G. W. Brown’s algorithm discussed in (Knuth, 1968).

Linearly scaling the entries of any of the pre-penetration functions in the chosen direction gives a random class of penetrance functions. We identify the penetrance function in this class with maximum heritability. It may be viewed as a random penetrance function among all those of highest heritability in each class of penetrance functions. Discussed in the next section, the unique (if any) penetrance function in the class having a specified heritability and prevalence can now be identified. It may be viewed as a random penetrance function among all those with the specified genetic constraints.

The linear scaling function $S$ we use to convert pre-penetration functions to penetrance functions is defined by: the $ij$th entry of $S(G)$, $G$ being a pre-penetration function, is given by $S(G)_{ij} = g_{ij} - m M - m$. Here $g_{ij}$ is the $ij$th entry of $G$, and $M$ and $m$ are the maximum and minimum, respectively, of the entries of $G$. The entries of $S(G)$ lie in the interval $[0, 1]$, with the minimum entry 0 and the maximum 1.

An equivalence relation is defined on the set of $n$-locus epistatic penetrance functions by $P_1 \sim P_2$ if applying a linear function $f(x) = mx + b$ to each entry of $P_1$ gives the corresponding entry of $P_2$. We note each equivalence class is two dimensional because of the two parameters.
Figure 3: A schematic depiction of our method for picking random penetrance functions. Points in the parameter space, $\mathbb{R}^{2n}$, correspond to potential pre-penetrance functions, and points in the shaded region correspond to potential penetrance functions. The vector $\vec{V}$ indicates a random direction in the set of pre-penetrance functions. $\vec{V}$ determines a random class of penetrance functions, which is indicated by the line, passing through the shaded region. The dot on the boundary of this region (A), indicates one (of the two) penetrance functions in this class with maximum heritability. If a random penetrance function with specific values of prevalence and heritability is desired, then from this class, we pick the unique (if any) penetrance function (B) with these values.

$m$ and $b$. Thus the equivalence classes from which we pick random elements form a space of dimension $2^n - 2$.

Let $\tilde{S}(G)$ be the equivalence class containing $S(G)$. If $G$ is generated by the Sudoku method, then $\tilde{S}(G)$ is a random equivalence class in the sense that $G$ is random and $\tilde{S}(G)$ consists of all penetrance functions that can be obtained from $G$ by applying a linear function to each entry of $G$.

The penetrance function $S(G)$ has maximum heritability among all those in $\tilde{S}(G)$. To see this, note that if a penetrance function with entries $f_{ij}$ has prevalence $K \leq \frac{1}{2}$ then

$$h^2 \left( \frac{f_{ij} - \min \{ f_{ij} \}}{\max \{ f_{ij} - \min \{ f_{ij} \} \}} \right) \geq h^2 (f_{ij} - \min \{ f_{ij} \}) \geq h^2 (f_{ij}) \quad (13)$$

Here we use $h^2$ followed by an expression in parentheses for the heritability of the penetrance function whose $ij^{th}$ entry is the parenthetical expression. The right-hand inequality holds since, when a positive constant is subtracted from each entry $f_{ij}$, the numerator of $h^2(f_{ij}) = \frac{\sum p_{1i} p_{2j} (f_{ij} - K)^2}{K(1-K)}$ is unchanged, while $K$ decreases in turn decreasing the denominator. The left-hand one holds since, for any constant $c$,

$$h^2(cf_{ij}) = \frac{c^2 \sum p_{1i} p_{2j} (f_{ij} - K)^2}{cK(1-cK)} = \frac{c(1-K)}{1-cK} h^2(f_{ij}) \quad (14)$$

which, as a function of $c$, increases on the interval $[1, \frac{1}{K})$.

In the case where $\frac{1}{2} < K \leq 1$, the argument just above applied to $1 - f_{ij}$ also gives the inequality (13). Note that if the maximum and minimum entries differ from 0 and 1, then at
least one of the two inequalities in (13) is strict. So applying this with \( f_{ij} = L(g_{ij}) \), where \( L \) is any linear function, shows that \( L(m) \) and \( L(M) \) must be 0 or 1, in either order. Thus the penetrance functions equivalent to \( S(G) \) with maximum heritability are \( S(G) \) itself and \( 1 - S(G) \) since a linear function is determined by its values at two points, in this case \( m \) and \( M \).

2.5 Estimating Maximum Heritability

![Figure 4: Maximum heritability estimates for pure, 2-locus models. Each point represents the maximum heritability estimate for a respective combination of genetic constraints. These points illustrate the estimated boundary of genetic constraint combinations, above which penetrance functions cannot be generated.](image)

Our method for estimating maximum heritabilities makes use of the many choices of parameters for penetrance functions. The problem of finding an epistatic 2-locus penetrance function of maximum heritability with specified penetrance and MAFs amounts to finding the maximum value of a quadratic function over a 4-dimensional set. This set, called the parameter space, varies with the choice of parameters. To construct a penetrance function of maximum heritability with specified MAFs and prevalence, we assign values to appropriately chosen parameters which, we estimate, yield penetrance functions with maximal heritabilities. These values depend on the specified constraints. In other words, we choose a suitable parameter space and points in the space determined by the specified constraints. We do not prove that the corresponding penetrance functions have maximum heritabilities, but found that our estimates were comparable to those given in Table 2 of (Culverhouse et al., 2002). They agree
for all 56 choices of MAFs and prevalence values, except in two cases where we found larger estimates of maximum heritability.

2.6 Dataset/Sample Generation

Given a genetic model, random dataset replicates with a specific sample size and number of SNPs were generated in the following manner: (1) Genotypes for SNPs specified in the genetic model are probabilistically generated for each sample in the dataset based on their specified MAFs. (2) Affection status (i.e. case/control status) for each sample is determined probabilistically based on the penetrance value (given by the model's penetrance function) of the genotypes specified for that sample step 1. (3) Genotypes for all other "noise" SNPs, not specified by the model are generated probabilistically based on a randomly assigned uniform distribution of MAFs ranging from 0.05 to 0.5.

2.7 Implementing the n-way Strategy

Implementation of this proposed strategy for the generation of epistatic models relies on constructing a multidimensional penetrance function to store the penetrance values. This is somewhat complicated because the data structure needs to accommodate arbitrary numbers of dimensions, and a variably sized penetrance function. In order to efficiently realize this high dimensionality structure, our implementation indexes into a one dimensional array. each index is recorded as a digit in a ternary representation of the location in the array, and this number is then used to dereference the stored data. For example, with a 2-way function, the nine locations would be stored as ((0,0),(0,1),(0,2),(1,0), (1,1),(1,2), (2,0),(2,1),(2,2)). This is equivalent to counting zero to eight in base 3, with each digit representing a dimension.

3 Results

Figure 5 presents the success rate of MDR to detect the correct underlying 2-locus model within replicate datasets. In this evaluation, MDR demonstrates the highest detection within datasets of higher samples size and heritability. This is an example of how GAMETES simulated models and respective datasets may be used to set up an algorithmic evaluation. We evaluate models of lower heritabilities in order to identify which constraint values MDR can significantly detect. The results of MDR analysis on 3 to 5-locus models are available in Figures 6, 7, and 8, respectively.

References

Brodie III, E. (2000). Why evolutionary genetics does not always add up. *Epistasis and the evolutionary process*, pages 3–19.

Cheverud, J. and Routman, E. (1995). Epistasis and its contribution to genetic variance components. *Genetics*, 139(3), 1455.

Cockerham, C. (1954). An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. *Genetics*, 39(6), 859.

Cordell, H. (2002). Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20), 2463.
Culverhouse, R., Suarez, B., Lin, J., and Reich, T. (2002). A perspective on epistasis: limits of models displaying no main effect. *The American Journal of Human Genetics, 70*(2), 461–471.

Fisher, R. *et al.* (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh, 52*, 399–433.

Frankel, W. and Schork, N. (1996). Who’s afraid of epistasis? *Nature genetics, 14*(4), 371–373.

Hallgrímsdóttir, I. and Yuster, D. (2008). A complete classification of epistatic two-locus models. *BMC genetics, 9*(1), 17.

Knuth, D. (1968). *The Art of Computer Programming 1: Fundamental Algorithms 2: Seminumerical Algorithms 3: Sorting and Searching*. Addison-Wesley.

Moore, J. and Williams, S. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays, 27*(6), 637–646.

Morris, A. and Whittaker, J. (1999). Generalization of the extended transmission disequilibrium test to two unlinked disease loci. *Genetic epidemiology, 17*, S661.

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

Wade, M., Winther, R., Agrawal, A., and Goodnight, C. (2001). Alternative definitions of epistasis: dependence and interaction. *Trends in Ecology & Evolution*, 16(9), 498–504.

Wilson, S. (2001). Epistasis and its possible effects on transmission disequilibrium tests. *Annals of human genetics, 65*(6), 565–575.
Figure 5: Algorithmic Evaluation of MDR Detection: All datasets in this figure were generated from 2-locus models. Each bar represents the successful detection frequency of MDR over 100 random replicate datasets. Each group of 5 differentially shaded bars corresponds to the 5 random models generated for each constraint combination. Each sub-plot corresponds to a specific combination of heritability and sample size. The values of 0.2 and 0.4 under each sub-plot refers to the MAF of that set of models.
Figure 6: Algorithmic Evaluation of MDR Detection: All datasets in this figure were generated from 3-locus models. Each bar represents the successful detection frequency of MDR over 100 random replicate datasets. Each group of 5 differentially shaded bars corresponds to the 5 random models generated for each constraint combination. Each sub-plot corresponds to a specific combination of heritability and sample size. The values of 0.2 and 0.4 under each sub-plot refers to the MAF of that set of models.
Figure 7: Algorithmic Evaluation of MDR Detection: All datasets in this figure were generated from 4-locus models. Each bar represents the successful detection frequency of MDR over 100 random replicate datasets. Each group of 5 differentially shaded bars corresponds to the 5 random models generated for each constraint combination. Each sub-plot corresponds to a specific combination of heritability and sample size. The values of 0.2 and 0.4 under each sub-plot refers to the MAF of that set of models. Constraint combinations for which models could not be found are labelled by “model not found” (MNF).
Figure 8: Algorithmic Evaluation of MDR Detection: All datasets in this figure were generated from 5-locus models. Each bar represents the successful detection frequency of MDR over 100 random replicate datasets. Each group of 5 differentially shaded bars corresponds to the 5 random models generated for each constraint combination. Each sub-plot corresponds to a specific combination of heritability and sample size. The values of 0.2 and 0.4 under each sub-plot refers to the MAF of that set of models. Constraint combinations for which models could not be found are labelled by “model not found” (MNF).