Towards an information theory of quantitative genetics

David J. Galas*, James Kunert-Graf, Lisa Uechi[+], and Nikita A. Sakhanenko

Pacific Northwest Research Institute

Seattle, Washington USA

Abstract

Quantitative genetics has evolved dramatically in the past century, and the proliferation of genetic data enables the characterization of complex interactions beyond the scope of its theoretical foundations. In this paper, we lay the foundations of an alternative formulation of quantitative genetics based on information theory. Information theory can provide sensitive measures of statistical dependencies among variables, and provides a natural mathematical language for an alternative view of quantitative genetics. In previous work we examined the information content of discrete functions and applied this formalism to the analysis of genetic data. We present here a set of relationships that both unifies the information measures for the set of discrete functions, and uses them to express key quantitative genetic relationships. Information theory measures of variable interdependency are used to identify significant interactions, and a general approach is described for inferring functional relationships within genotype and phenotype data. We present information-based measures of the genetic quantities: penetrance, heritability and degrees of statistical epistasis. Our scope here includes the consideration of three variable dependencies and independently segregating variants, which captures two locus effects, genetic interactions, and two phenotype pleiotropy. However, this formalism and general theory naturally applies to multi-variable interactions and higher-order complex dependencies, and can be adapted to account for population structure, linkage and non-randomly segregating markers. This paper therefore lays the initial groundwork for a full formulation of quantitative genetics based in information theory.

*Communication to: dgalas@pnri.org Contacts: nsakhanenko@pnri.org, jkunert@pnri.org, luechi@pnri.org, dgalas@pnri.org + Current address: University of California Los Angeles,

1. Introduction

The critical questions for understanding a genetic system, its functions, structure and complexity, lie in the actual dependencies among the system's variables, both the phenotypes and genotypes as well as external factors. Of course, the phenotypes can range from highly specific cell or molecular measures to broader, organismal level phenotypes. The information architecture of the genetic system's variables is at the heart of the dependency problem, and the difficulty of determining this architecture from data is significant for truly complex systems, which well describes many genetic problems. These problems are inherent in the challenges of the past concerning the genetic explanation of complex traits, the notion of missing heritability and the effects of gene interaction. Quantitative genetics has evolved substantially over the 100 years since Fisher and Wright laid its foundations, in these papers [1,2], for example. It has been pointed out repeatedly, however, that while their methods were powerful and innovative, there are some problems with the general approach. It is not that the methods are not correct and powerful, but rather that the proliferation of data types calls for additional approaches and mathematical descriptions. Nelson, Petterson and Carlborg, for example, have argued effectively that the Fisherian paradigm has reached its limits in the ability to deal with complex traits and modern genetic data [3]. Their summary, "… many of the current tools are adaptations of methods designed during the early days of quantitative genetics. The present analysis paradigm in quantitative genetics is at its limits in regard to unraveling complex traits and it is necessary to re-evaluate the direction that genetic research is taking for the field to realize its full potential," is a clear call for new quantitative approaches. It is also true that, in spite of their innovative statistical approach in 1918, the Fisherian methods have often been misused in present quantitative genetics. Huang and MacKay have clearly made the case, for example, that the genetic architecture of quantitative traits cannot be inferred from variance component analysis as it has been applied in many studies over the years [4]. It is clear that genetic interactions, called epistasis in one use of this term, have been implicated as components essential for understanding complex traits [5-9], though it has also been challenged as being unimportant in evolution [10]. Recent results have strongly supported the importance of interactions in understanding complex traits [7, 24-27], including those in humans. In addition, quantitative inference of interacting loci will likely be important for polygenic risk scores which are currently being generated using non-interacting models.

Here we propose that information theory can provide the foundations of a new approach to quantitative genetics which is different and exhibits useful properties, and we begin the process of building that foundation with this paper. It is not our position that present methods are faulty, but rather that it is likely that establishing a new approach and formulation will reveal new insights and provide methodologies because of the fundamentally different viewpoint. For example, the ability to detect two locus dependencies without significant single locus dependence extends the analysis power beyond the GWAS method. This extension is a natural feature of the information theory formulation.

Information theory, while originally directed at the problem of understanding communications quantitatively, has been very effective outside of this original domain and has been applied widely to physical, biological, and chemical problems, and to other fields [13]. In almost all scientific domains the problem of inferring the quantitative dependencies among measurable variables, and even causal relationships, is the central problem, and the information measures, functionals of probability distributions, have been shown to be important tools in these problems of inference. We have previously shown how information theory methods can be used to analyze complex data, and also shown how genetic data is amenable to some such applications [11-13]. Here we extend both the formulation of the relationships and methods and their interpretation, and recast the theory into a more comprehensive description of quantitative genetics. While we take only the first few steps here towards a full information theory of complex genetics, we show how this approach forms a fruitful way to describe the complex genetic architecture of a system. Specifically, we describe familiar concepts like gene interaction, pleiotropy, penetrance, degree of epistasis and heritability in terms of information theory.

Our concept of the information architecture of a system derives primarily from the idea of using information measures to define the levels of dependencies among variables. Information theory, being model-free, is broadly applied to extracting statistical properties from the data, which are in turn determined by the joint probability distributions of the variables. The measures have the advantage of being completely agnostic of any models or prior assumptions affecting dependencies, and this allows the data to fully drive the conclusions, and reduces the sensitivity of

the measures to small variations in the data, and to limitations of samples numbers. Thus, we argue that the application of information theory to genetics can provide a powerful approach to deciphering the structure of complex systems and to extracting their information architecture, which is distinct from the genetic, or model architecture. This paper advances our previous work in which we defined an information landscape [11], and illustrated the use of discrete functions and noise on this landscape to analyze genetic data. Here we focus on specifically elucidating the relationships of three variable dependencies, and complete this picture by providing a way to extract the specific functional nature of dependencies for variables whose dependency has been detected and measured.

There is always "noise" in the data, deriving both from unknown variables and from stochastic processes inherent to the biology and data acquisition processes. We enclose the word in quotes here to emphasize the composite nature of these several factors in determining "noise". This quantity can only be inferred from data when we explicitly define the character and degree of the dependencies we are including. For example, if we only consider direct, pairwise effects from each of two loci on a phenotype, then the interaction between these loci affecting a phenotype (what we will call a three-way dependency) as well as any other more complex interactions will contribute to the "noise". Likewise, if we only include the effects of the genetic variants that we have ascertained, the loci not included will contribute "noise". All unknown genetic variants and other unknown factors potentially contribute to the "noise". In this way we both more tightly define the quantitative nature of genetic penetrance and provide a well-defined method for a data driven estimate of the quantities. We envision two fundamental steps in a general method for the inference of the relevant dependencies: first, the detection of levels of dependencies using the information theory measures, followed by the inference of the functional nature of dependence and the "noise" level.

General as it is, the application of information theory to any specific area carries with it certain assumptions and premises which need be made explicit. The idea that the statistical inferences from the data reflect the subtle features of variable dependency assumes that the sampling issues and density of data represent these features in sufficient detail for information methods to make reliable estimates. In actual use this is often a rough approximation only and the approximation

must be explicitly quantitated and its meaning acknowledged. We discuss this question later in the paper and for the purpose of explication initially assume simply that the data set is large enough to be fully reflective of the underlying relationships. It is also clear that by its nature information theory is inadequate to fully represent some distinctions among certain distributions. There are distinct distributions with identical information measures. The mapping of probability distributions of variables into information measures is decidedly many-to-one. There are therefore many models and architectures that have the same sets of measures. In addition, the question of how many variables participate in synergistic dependencies in a complex system must be truncated in any practical application because of statistical and computational limits [14]. While the method is entirely general we limit ourselves here to two and three variable dependencies. Nonetheless, the power of the three variable method is amply demonstrated.

The symmetries of the relationships among the information functionals are surprisingly simple, but also subtle. The multiple measures of information theory have strikingly symmetric relations [15-17], and have a number of symmetries that we have previously reported [18]. The symmetries all derive from the fact that all information measures are specific linear combinations of joint entropies, organized by lattices whose partial order is determined by inclusion of variable subsets. In addition, there are a number of problems that can be fully analyzed for discrete functions, which are the most common manifestations of the variables we deal with in data analysis. By this we mean that the dependent variables in a complex system can be viewed as functions of one another, and the discrete values of the data can therefore be viewed as reflecting these discrete functions. While real genetic data has various levels of probabilistic determinants and "noise", much of the character of the dependency can be represented by multi-valued discrete functions, which are mixed with various levels of "noise" to describe the realistic intervariable dependencies. This gives us a distinct mathematical advantage since, in principle, we can characterize the properties of all possible discrete functions with finite alphabets. We examine here the properties of discrete functions and their information architecture and relationships, show in detail how functions can be classified, and examine the extension of this analysis to include probability density functions that result from adding "noise" or subtracting determinism from the discrete functions.

2. Overview of Formalism

The complexity of genetics arises not only from the interactive functions encoded in the genome, and the range and complexity of phenotypes, but also from the structures of study populations and inheritance patterns in complex pedigrees. In this paper, while recognizing the important effects of population structures on quantitative genetics measures, we defer addressing these important issues. They will be considered in a later paper, and we restrict our considerations here to large, randomly mating populations, described as *panmictic*, recognizing that no natural population is fully panmictic, and few artificial, experimental populations in practice.

The basic components of the formalism presented here are summarized in these five points:

1. The information measure we call the *symmetric delta*, symmetrized conditional interaction information [13], is used to detect the dependence of subsets of loci with phenotypes in the data. In this paper, we consider pairwise and three-way dependencies.
2. The general relation between genetic loci and phenotypes is embodied in discrete valued *loci-phenotype tensors*: $f(X_1, X_2, \dots X_n)$, where $\{X_i\}$ is the set of *n* genetic loci and the function determines the phenotype. This is identical in two dimensions to what geneticists often embody in a matrix connecting three variables, called a "gene-phenotype table". We limit ourselves to one or two genetic variables (loci) here. Without loss of generality we could include multiple phenotype variables as well.
3. The essential "noise" distributions, when added to these tensors, form the genotype-phenotype tensors (GPT) which describe the phenotype in terms of loci, noise and penetrance

$$G(X_1, X_2, \dots X_n; p) = pf(X_1, X_2, X_3 \dots X_n) + (1-p)\varepsilon$$

   where ε is a "noise" function, and (1-*p*) is the noise level (*p* is the penetrance.) The noise can be assumed to follow a particular structure (*e.g.*, uniform random noise).
4. The tensor functions *f* and *G* for tuples of variables with significant dependence are inferred from the data using relatively simple algorithms.
5. These tensors are then used to calculate penetrance, heritability, gene interactions and pleiotropy.

This paper is organized as follows. We first present the basic discrete function expression of genetics, gene-phenotype tables which assumes full genetic dependence (no "noise"), then review the basics of the information measures previously introduced [11-14]. We then describe some specifics of three variable dependencies and the symmetries their information measures exhibit [18]. We review the information landscape notion previously proposed [11] and extend it to a more general form. Introducing a formal tensor structure for extending the information landscape allows us to systematically handle all probability distributions, which are essential for the introduction of "noise," for arbitrary size alphabets. This formulation shows that the information content of the discrete functions is strongly dependent on both the alphabet size and the symmetries of the functions. This rich area is only partially explored here, but provides us some initial insights and a flexible set of theoretical tools with which to characterize complex genetic systems. We then define a set of transformations that maps the three variable functions into a two-variable function space, and allows us to greatly simplify the identification of the functional structure of the inferred dependencies.

We discuss the implications of these results and tools for the analysis of genetic data using information-based methods, and describe, in addition to penetrance, the genetic notions of gene interaction, pleiotropy, and heritability in terms of information theory measures. Finally, we apply our methods to some real yeast data and discuss the analysis of complex genetic data [23].

3. Discrete functions and genetics

The classic genotype-phenotype table for two loci can be usefully considered as a discrete function where the phenotype variable, Z, is expressed as a function of the two genotype variables, X and Y. Diploid binary variants for X and Y are, of course, three-valued, haploid binary variants are two-valued, while the phenotype alphabet can be any size, in principle. The alphabet can certainly be expanded to include more than binary allele variants, but we do not consider these in this paper. Often a two-valued phenotype is sufficient to effectively describe a phenotype, but quantitative phenotypes require a larger alphabet. These tables are similar to the Punnett squares in classical genetics. Consider the discrete functions where all three variables, X, Y, and Z are 3-valued, and $Z = f(X,Y)$, with X and Y independent. Each of the functional relationships can be represented by a 3-by-3 table. Table 1, for example, shows three functions that can be seen as tables for logical

AND, logical XOR, and equality (EQ) functions, extended to three variables. In these tables the genotypes are encoded as: 0 = homozygous major alleles, 1 = heterozygote, and 2 = homozygous minor allele.

		X		
	AND	0	1	2
Y	0	0	0	0
	1	0	1	1
	2	0	1	2

		X		
	XOR	0	1	2
Y	0	0	1	2
	1	1	2	0
	2	2	0	1

		X		
	EQ	0	1	2
Y	0	0	1	2
	1	0	1	2
	2	0	1	2

Table 1. Examples of extended (3-by-3) tables defining 3-valued versions of genetic functions. From left to right these correspond to: logical AND, XOR, and EQ functions. These functions can also be represented in a linear notation, as 000011012, 012120201, and 000111222, respectively.

The EQ function is a degenerate case for which *Z* is only a function of *X*. This general scheme can also be thought of as implementing a three-valued logic. We call the function defined by the leftmost in the table an extended AND, because the lower value of the two arguments dominates, as in the binary version. These discrete functions describe the phenotype as a function of the two genetic loci. Since there may be phenotype determinants other than these two loci, as described above, we will generally need a more complex function to describe the phenotype in a population. A three-variable tensor, which we consider in detail later, can embody this complexity. The discrete function is modified by the random effects of both unknown and stochastic effects, the "noise" represented by a random function, and the penetrance, the degree of determination of the phenotype by the genetic variables.

The diploid case is, of course, most commonly encountered in genetics of mammals, but the haploid case is not unknown in genetic data, for example in the case of recombinant inbred populations. We apply our methods to an example of a haploid case in data from *Saccharomyces cerevisiae*. In the case of haploid genetics the alphabet of values for the genetic variables is binary so that the genotype-phenotype table is 2x2. This simplification can be very useful in practical calculations, as we discuss later.

4. Elements of the Theory

4.1 Genetic Dependence

We begin by reviewing some definitions and previous results, and then introduce extensions of these relations. The first important point is that mutual information, an inherently pair-wise measure, is unable by itself to capture the full information in dependencies. Full representation requires many variable subsets, but even for three variable tuples considered here mutual information is insufficient, and require additional measures to fully characterize the dependence among three variables. As has been pointed out before, a clear example is the exclusive OR relationship (XOR) for any size of alphabet [11]. For the binary alphabet, three-variable case it is evident that the mutual informations between all pairs of variables for this function vanish. We have demonstrated that the ternary XOR-like functions with larger alphabets (Table 1) also exhibit this property [11]. It is also true for any size alphabet and is reflective of the symmetry of the dependencies.

Even the interdependency of two variables has a surprising level of complexity in the ways it can be expressed. Mutual information has several equivalent mathematical expressions. The most common form is as a difference of entropies

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \qquad (1)$$

In terms of the conditional entropies we also have these symmetric expressions for mutual information. $I(X,Y) = H(X) - H(X|Y)$ or $I(X,Y) = H(Y) - H(Y|X)$.

An important information measure is a generalization of mutual information for multiple variables, called the interaction information, or co-information [16,17]. For *n* variables this is defined by the recursion relation

$$I(X_1, X_2, X_3 \dots X_n) = I(X_1, X_2, X_3 \dots X_{n-1}) - I(X_1, X_2, X_3 \dots X_{n-1}|X_n) \qquad (2)$$

This measure is also given by the sum with the Möbius function for the lattice of subsets, τ, of the full set of variables $\nu = \{X_1, \dots X_n\}$:

$$I(\nu) = \sum_{\tau \subset \nu} (-1)^{|\tau|+1} H(\tau) \tag{3}$$

For three-variables the differential information is simply expressed in terms of the entropies.

$$I(X_1, X_2, X_3) = H(X_1) + H(X_1) + H(X_1) - H(X_1, X_2) - H(X_1, X_3) - H(X_2, X_3) + H(X_1, X_2, X_3) \tag{4}$$

We define the *differential* interaction information as the change in the interaction information that occurs when we add another variable to the set. We have in general, where $X_m \notin \nu_m$ and $\nu_m \cup X_m = \nu$, the differential interaction information defined as

$$\Delta(\nu_m; X_m) = I(\nu) - I(\nu_m) = \sum_{\tau_m \subseteq V | X_m \in \tau_m} (-1)^{|\tau_m|+1} H(\tau_m) \tag{5}$$

The general measure of the fully collective dependence among all variables, the symmetric Delta, $\bar{\Delta}$, is defined as the product

$$\bar{\Delta}_m = \bar{\Delta}(\nu_m) \equiv \prod_{i=1}^{m} \Delta(\nu_i; X_i) \tag{6}$$

This is the measure we have proposed and used for measuring collective dependence of a set of variables. For three variables the differential interaction informations (the "deltas") are obtained by permutation of the variables.

$$\Delta(X, Y; Z) = I(X, Y) - I(X, Y, Z) = I(X, Y|Z) = -H(Z) + H(X,\ Z) + H(Y, Z) - H(X, Y, Z)$$

and the symmetric delta for three variables is

$$\bar{\Delta}(X, Y, Z) = \Delta(X, Y; Z) \cdot \Delta(Z, Y; X) \cdot \Delta(X, Z; Y) \tag{7a}$$

If we want to focus on the dependence of one class of variables on another, as we do in genetics, we can limit the symmetric product to two factors. If *X* and *Y* are genetic variants and *Z* is a phenotype, for example, our interest lies in the variants-phenotype dependence, not on the variant-variant dependence, which is actually linkage disequilibrium. Thus, for these three variables we can use the product of $\Delta(X, Z; Y) = I(X, Z) - I(X, Y, Z)$, and $\Delta(Z, Y; X) = I(Z, Y) - I(X, Y, Z)$.

$$\tilde{\Delta}(X, Y, Z) = \Delta(Z, Y; X) \cdot \Delta(X, Z; Y) \tag{7b}$$

While we do not make use of this further in this paper, this dependence measure will allow us to factor out a strong linkage disequilibrium effect that can otherwise amplify weak dependencies and distort genetic dependencies. We can also usefully expand the *n* variable joint entropy $H(X_1, \dots X_n)$ into a sum of terms each of which depends on the number of variables, using the Möbius inversion [14]. This gives us an expression for the entropy as a sum over interaction-informations over all possible subsets of variables. Truncation at a specific limit to the number of variables that simultaneously depend on each other generates consistent approximations, based on the degree of dependencies considered. This approach generates a series of approximations in the number of variables considered, and represents a systematic way forward in generalizing the genetic formalism practically for more than three total variables, in that it provides the appropriate approximation for each limiting assumption. We will make use of this in future work.

4.1.1 Multi-information as Total Dependence

Another important information measure that we will use in several ways is the multi-information for *n* variables (originally defined and called "total correlation", by Watanabe [19] and discussed by others [20,21]). It is defined as

$$\Omega(X_1, \dots X_n) \equiv \sum_{i=1}^{n} H(X_i) - H(X_1, \dots X_n) \tag{8}$$

the difference between the sum of entropies of each variable separately and the joint entropy of all the variables together. The multi-information is the collective measure of all dependencies among the *n* variables; that is, the sum of all possible subset dependencies. It is zero only when all the variables are independent. This stands in contrast to the symmetric delta, which is the measure of the full, synergistic dependency of all the *n* variables together. It is zero when any one of the variables is independent of the others. Since the multi-information deals with dependence of all possible subsets, while the symmetric delta deals with dependence of the entire set, they are bookends of the dependency measures. As shown in the next section the multi-information is a key element in the quantitative relationships we use in this formalism.

4.1.2 Three-Variable Dependencies

While the restriction to pairwise dependency analysis is equivalent to classical association studies in genetics, which are sufficient for some problems, the use of even three-variable dependencies can add much to the power of the analysis and is essential for any genetic system that involves pleiotropy or gene interaction. We focus in this section on understanding systems at the three-variable level. The relations among the three variable information measures are simple, but subtle, and illustrate the strong symmetries inherent to the information measures. Furthermore, it is useful to examine carefully the bounds on their values. First, a few more preliminaries.

From here on we will use a further simplified notation, where the three variables are labeled by integers: $X \to 1, Y \to 2, Z \to 3$. Wherever the meaning is clear, we will abbreviate using these labels within a subscript; for example, $\Omega(\mathrm{X,Y,Z}) \to \Omega_{123}$. The relations between the mutual informations and the multi-information, and the deltas (where we define $\Delta_1 \equiv I(2,3|1)$, $\Delta_2 \equiv I(1,3|2)$, $\Delta_3 \equiv I(1,2|3)$) are related by these equations:

$$\begin{aligned} \Delta_1 &= \Omega_{123} - I_{12} - I_{13} \\ \Delta_2 &= \Omega_{123} - I_{12} - I_{23} \\ \Delta_3 &= \Omega_{123} - I_{13} - I_{23} \end{aligned} \tag{9a}$$

For two variables only, of course, $\Omega = I_{ij}$, the mutual information. It can easily be seen that these equations are symmetric in the variables, and the only asymmetry arises from the differences among these terms.

For genetic data, where X and Y are independently segregating genetic loci (Z is the phenotype variable here), the three mutual informations in (9a) become two since $I(X,Y) = I_{12} = 0$ [11]. The assumption of independently segregating variants is essentially equivalent to assuming linkage equilibrium. In this case there are only three relevant measures in the set of relations (9a), Ω_{123}, I_{13}, and I_{23} , and the relationship is significantly simplified.

$$\begin{aligned} \Delta_1 &= \Omega_{123} - I_{13} \\ \Delta_2 &= \Omega_{123} - I_{23} \\ \Delta_3 &= \Omega_{123} - I_{13} - I_{23} \end{aligned} \tag{9b}$$

We can normalize the Equations 9b by dividing through by Ω, as long as there is some dependency so that $\Omega > 0$. We get the normalized delta coordinates (only for the case of $I_{12} = 0$) which were the coordinates used in [11] to define the geometry of the information landscape. The coordinates of the information landscape then are these:

$$\begin{aligned} \delta_1 &= 1 - \mu_{13} \\ \delta_2 &= 1 - \mu_{23} \\ \delta_3 &= 1 - \mu_{13} - \mu_{23} \end{aligned} \tag{10}$$

where the $\delta's$ are the normalized Δ's, and the $\mu's$ are the normalized mutual informations.
We can rearrange the above equations into a simple relation for δ_3 as a function of δ_1 and δ_2:

$$\delta_3 = \delta_1 + \delta_2 - 1 \tag{11}$$

The condition for δ_3 to be non-zero then is $\delta_1 + \delta_2 > 1$. This is one side of the line defined by $\delta_1 + \delta_2 = 1$. Let us look more closely at the constraints on δ_1 and δ_2 imposed by $I_{12} = 0$. If we look at the 3-D space defined by the three δ's, which is what we call the information landscape, we can see that we have three coordinates and one linear constraint that thereby defines a two-dimensional plane. One natural question is whether negative coordinates are possible. The answer is that they are not.

Proposition 1: $\Omega_{123} \geq I_{13}$, $\Omega_{123} \geq I_{12}$ and $\Omega_{123} \geq I_{23}$ (12)

Proof:

We first prove that $\Omega_{123} \geq I_{13}$.

$$\Omega_{123} = H_1 + H_2 + H_3 - H_{123}$$
$$I_{13} = H_1 + H_3 - H_{13}$$

Subtract the lower from the upper, to get $\Omega_{123} - I_{13} = H_2 - H_{123} + H_{13}$. Since $\Omega_{123} \geq 0$, and $\mathrm{H}_i \geq 0$, and since the sum of the entropies of any subset of variables is greater than or equal to the joint entropy, it is clear that $H_2 + H_{13} \geq H_{123}$, and the proposition is proven. Starting with I_{23} or I_{12} yields the parallel inequalities. There are, however, tighter bounds.

Proposition 2: If Ω_{123} is the multi-information for three variables, then

$$\Omega_{123} \geq I_{13} + I_{23},\ \Omega_{123} \geq I_{12} + I_{23},\ \Omega_{123} \geq I_{13} + I_{12} \qquad (13)$$

This result is intuitive and elementary, but still not entirely obvious. For completeness we provide a proof in Appendix A.

The interaction information, I_{123} , is defined [16,17] in terms of the entropies as

$$I_{123} = H_1 + H_2 + H_3 - H_{12} - H_{13} - H_{23} + H_{123}$$

and by the definitions of mutual information we have

$$\Omega_{123} = I_{12} + I_{13} + I_{23} - I_{123} \qquad (14)$$

Notice that if $I_{12} = 0$, by proposition 2 this expression implies that $I_{123} \leq 0$.

A few more points about dependencies among genetic variables are in order here. Equation 9b applies when I_{12} is strictly zero, however I_{12} may as well be nonzero in real data, because of disequilibrium or noise in the data, including sampling induced fluctuations. We will deal with the linkage disequilibrium (LD) issue in a future paper, but it is important to note that even in the presence of LD the symmetric delta represents the full interaction score for any triplet, including the contribution due to LD. A significant problem to be dealt with in a future publication is that it is more difficult in this case to extract the quantitative score for the strictly three-way component (we call this the epistatic component). The potential entanglement of epistasis and linkage disequilibrium, often overlooked, is at the heart of this issue.

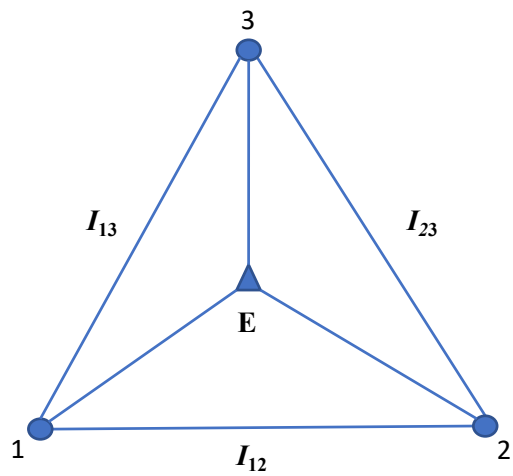


Figure 1. Three-variable dependencies that make up the multi-information or total correlation (we adopt the convention here that X is 1, Y is 2 and Z is 3). The lines represent the components of dependence among the variables (small circles) as in the above equation, where the epistatic component is represented by the lines emanating from the triangle. The epistatic component is $E = -I_{123} + S$.

There are many ways of expressing the set of relationships described above for three variables. For example, Equation 9 leads directly to the expression for the multi-information as in Equation 14. Since it is clear that if the dependencies are pairwise, and $I_{12} = 0$, then the mutual informations contain all the dependence, in which case $\Omega = I_{13} + I_{23}$. Thus, we can ascribe the three-way or epistatic component to the value of $-I_{123}$ if $I_{12} = 0$ (the minus sign comes from our sign convention above). This is a useful way to decompose the multi-information. This relation for the triplet dependencies is illustrated in Figure 1. Note that the epistatic component is $-I_{123}$ only when $I_{12} = 0$. In the general case $-I_{123}$ is equal to the epistatic component minus the information shared by 1 and 2 affecting 3, which we denote as S. The above equations allow us to define several important limiting conditions. This is further illustrated in Figure 2. We summarize these conditions and their implications in Table 2.

Constraint	Consequence
$I_{12} = 0$	Independent segregation of loci, implies no linkage disequilibrium
$\Delta_3 = 0$	Pairwise dependencies only, $\Omega = I_{13} + I_{23}$; $\Delta_1 = I_{23}$; $\Delta_2 = I_{13}$
$I_{13} = I_{23} = 0$	Three-way dependence only, $\Delta_3 = \Omega$; $\Delta_1 = \Delta_2 = \Omega$

Table 2. Several limiting constraints on the information relations, with their interpretations or consequences. Keep in mind that these rules apply strictly only to the discrete functions without noise.

4.1.3 Genetic Dependencies

The genetic architecture of a phenotype is determined by the dependencies among the genetic variables and the phenotype variable. The application of the information formalism can, however, be rather subtle. In this section we define the problem in a bit more detail and make the specific connections between information theory quantities and genetic quantities. The dependencies of phenotypes on more than one genetic locus define what we mean by genetic interactions and consists of a wide, but finite, range of possible forms of interactions. These effects have been recognized for 110 years when William Bateson proposed this as an explanation for deviations from simple Mendelian ratios. If a phenotype is dependent on two loci, each exclusively in a pairwise fashion, we call this effect additive, and distinguish it from what in the usual terminology is called an epistatic effect or interaction. Fisher called this statistical epistasis *epistacy*, and attributed the deviations from additivity to his linear statistical model. Many modern authors have argued and provided evidence that gene-gene interactions are rather common [9]. The most common way to deal with these interactions quantitatively, however, has been to use regression methods [30], and more recently other machine learning tools. In these cases, however, the starting loci are most often those identified by GWAS or some pairwise method, which will then miss those loci invisible to pairwise methods.

Quantitating "gene interaction", that is, measuring the amount of the phenotype that depends on the combined markers, can be done naturally with the measures defined here. We need to be precise, however, in defining what we mean by gene interaction, and we need to distinguish additive effects from epistatic interactions, the former being strictly pairwise, the latter not including any pairwise effects. Again, we are here assuming independently segregating variants and $I_{12} = 0$. This simplifies the problem considerably, but deviations from this constraint are important and will be considered elsewhere.

If the genetic variant variables are X and Y, and the phenotype variable is Z, we consider all possible three variable dependencies, as in Figure 1. In this general, three-variable case we can quantitate the information contribution of X and Y to the determination of Z by the mutual information between Z and the joint X,Y variables, $I(Z,(X,Y))$. Using the mutual information chain rule

$$I(Z,(X,Y)) = I(Z,Y) + I(Z,X|Y) \tag{15}$$

and identifying $I(Z,X|Y) = \Delta_Y$ and using Equation 12a we have simply

$$I(Z,(X,Y)) = \Omega - I(X,Y) \tag{16}$$

In the case of independent segregation of markers, where $I(X,Y) = I_{12} = 0$ this becomes $I(Z,(X,Y)) = \Omega$, as expected since in the absence of shared information between X and Y the mutual information $I(Z,(X,Y))$ describes the full extant dependence. As shown in the previous section the decomposition of the information contributions becomes simple in this case. In Figure 2 we illustrate the nature of the dependencies.

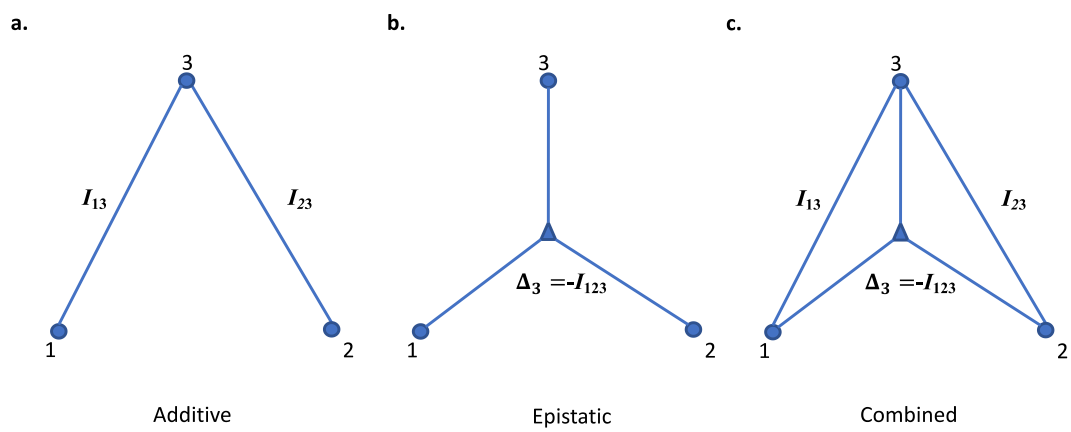


Figure 2. Independent segregation interaction relationships. The genetic contributions, (1 & 2) to the phenotype, 3, illustrating the distinction between the additive and epistatic effects.

We wish to emphasize that the relationships present here permit the decomposition of the information structure, the dependencies, of the variables. It is important that the dependency can be decomposed and we can determine what fraction of the dependence is pairwise and what fraction is three-way dependent (synergistic or epistatic). These fractions can be derived simply from the equation $\Delta_3 = \Omega_{123} - I_{13} - I_{23}$ provided that $I_{12} = 0$. Since the pairwise dependence of the phenotype on the two loci is the sum of the mutual informations, I_{13} and I_{23}, and the total dependence is Ω_{123}, the three-way dependence is given by their difference, Δ_3. The fractional dependencies are then simply the ratios

$$\text{Pairwise fraction} = F_p = \frac{I_{13}+I_{23}}{\Omega_{123}}, \text{ Epistatic fraction} = F_e = \frac{\Delta_3}{\Omega_{123}} \qquad (17)$$

If the genetic variables are 1 and 2, and the phenotype is 3, then these fractions represent the pairwise additive contributions of 1 & 2 to the phenotype, F_p, and the non-additive, epistatic contribution, F_e.

The Equations 14 and 15 apply in this case and the separation of the additive and non-additive, or epistatic effects is clear.[1] We will address the case of non-zero linkage disequilibrium and similar effects in a future paper.

The epistatic interaction in the case of no disequilibrium is measured entirely by Δ_3. This is also intuitive since the multi-information, Ω_{123}, quantitates the total dependence and the mutual information quantitates the pairwise, non-redundant pairwise dependencies between each variant and the phenotype. Thus, their difference measures epistatic gene interaction.

$$Epistasis\ measure = \Delta_3 = \Omega_{123} - I_{13} - I_{23} \qquad (18)$$

There is another three-variable dependence that is important in genetics. A single locus affecting two phenotypes, which is called *pleiotropy*, can be described by the general equations, but the limiting constraint of independent segregation which makes the mutual information vanish does not apply in this case. In this case, as for gene interaction in linkage disequilibrium, the independence assumptions do not apply. The intuition here is simply that, unlike for two genetic variables, we cannot decompose the information contributions of one genetic variant and another phenotype on a phenotypic variable. The potential complexities are significant and will be considered in future work.

[1] If $I_{12} \neq 0$ the situation is more complex, and distinctions need to be made between the redundant information provided by X and Y, the unique information provided, and the synergistic information and the quantities indicated in figure 2 do not apply. It is worth noting that the negative co-information in Equation 14, $-I_{123}$, is actually the difference between the synergistic and the shared information. These are respectively the epistatic interaction information and the dependence information between the markers (as caused by linkage equilibrium.) This is essentially the information decomposition problem, which has no universally accepted method of computation [29].

4.2 Information theoretic relations and properties

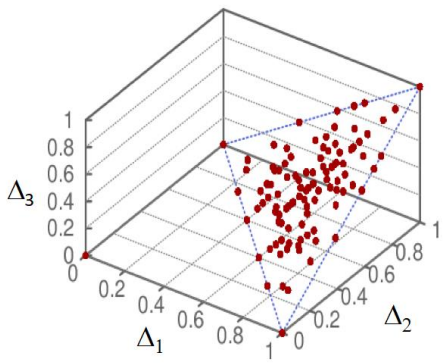
When two loci (X ,Y) are involved in determining a phenotype, Z, we can represent the relation as a genotype-phenotype matrix. These three-variable matrices have discrete values and thus are discrete functions of two variables, $Z(X,Y)$. We have shown that the three-dimensional information landscape, defined by the three normalized deltas from Equation 9, defines a plane when $I_{12} = 0$, and under this condition all discrete functions lie on this plane.

4.2.1 Discrete Functions

There are several possible ways of defining the information content of discrete functions, and since discrete functions in are useful in characterizing quantitative genetic relations, this information is key. Here we define the inherent information as the measures calculated from the probabilities inferred directly from the function. Note that all discrete functions map into distributions, but not all distributions are discrete functions. Information measures (all are linear combinations of joint entropies) are functionals of distributions. However, the finite size of the set of discrete functions (for a given set of alphabet sizes) and the (infinite) size of the set of all possible distributions are incommensurate. There are a finite number of discrete functions for any finite number of variables and alphabets, but there is an infinity of distributions for any finite number of variables and alphabets. The addition of "noise" to the discrete functions, then generates an infinite range of distributions. As we will see this is a key consideration in quantitative genetics. The "noise" determines the penetrance of the genetic dependence on the discrete function.

As shown in [11], we can map all the discrete functions onto the information plane (for example, there are 19,683 functions on this plane for the 3x3 case). When the information measures are calculated for the 3x3 functions and plotted in the plane they form simple rectilinear patterns for each value of Ω. The positions of all function families (those functions with identical normalized delta coordinates) are shown in Figure 3a.

(a)



(b)

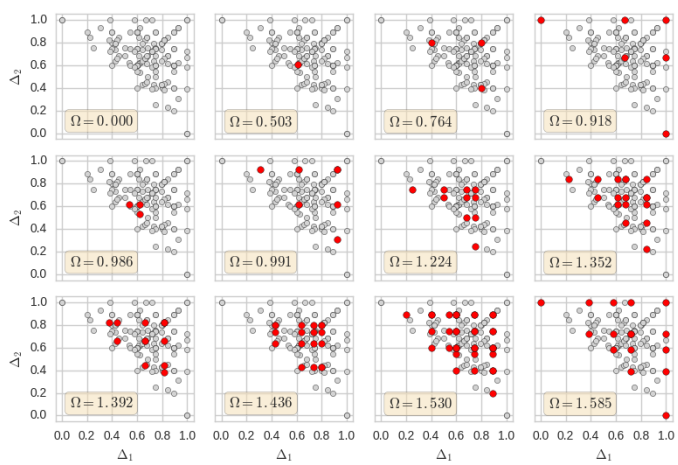


Figure 3. Function classes (3x3) on the landscape. Each spot in both panels represents a function class, or family. **a**) The information landscape shows the orientation of the plane with respect to the 3D landscape. **b**) A set of 12 panels, one each for the complete set of possible values of the multi-information, Ω, for the 3X3 functions. The plane is the projected diagonal plane of the three-dimensional landscape, the grey spots are the same for each panel and show the positions of <u>all</u> of the families of functions. The red spots are the families specific for each specific value of Ω. The upper left panel has no function as the information content of the uniform functions are zero, and all Δ's are zero.

In Figure 3b the families are shown according to whether they contain functions with the possible values of Ω, total dependence. Even though all functions in a family have the same delta coordinates, not all of the functions in a family need have the same value of Ω. Notice the symmetry in the triangular plane that results from the exchange of *X* and *Y*.

In the case of haploid genetics the information plane for three variables shows a similar geometric symmetry, but with many fewer functions. Many published yeast genetic data sets are haploid,

including the data we have analyzed here to demonstrate our methods [22]. Haploid genetic state variants are binary and since there are N^4 discrete functions, where N is the alphabet size for the phenotype, Z, this can lead to a significant simplification of the information landscape for binary phenotypes. For $N = 2$ there are only 16 functions in all, but as N increases from 3 to 5 the number of functions grows rapidly, and there are 8 families of functions, each family having identical information coordinates. Surprisingly, as N increases past 5 the number of families stays the same even as the number of functions grows rapidly. In Figure 4 the information plane and the families are shown.

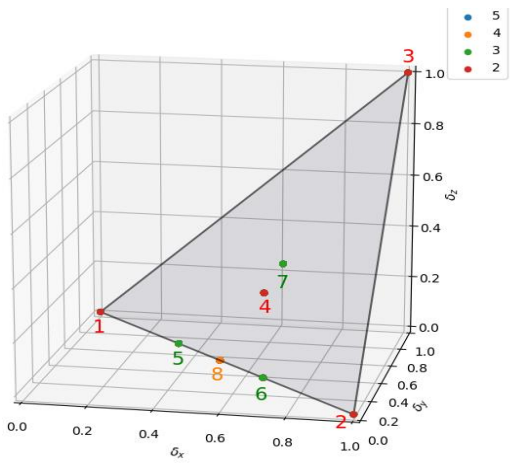


Figure 4. The information plane for haploid genetics, binary genetic variables. The color-coded points show the locations of the function families corresponding to the alphabet size of the phenotype. Families 1-4 correspond to a binary phenotype alphabet. Families 5-7 are added for a three-letter alphabet, and family 8 is added for a four-letter alphabet. The blue dot in the legend is not seen since does not correspond to any specific family. The five-letter alphabet functions all fall into the previous eight families. While the limit is eight families, as the alphabet size increases the number of functions in every family grows. The families 1, 2, 5, 6 and 8 are functions with only pairwise interactions ($\delta_Z = 0$).

5. Analysis of Data and the Role of "Noise"

5.1 Probabilistic Tensor Formalism

The discrete functions of three variables, interpreted as distributions are illustrated in the above landscapes (see Figures 3 and 4), where the information measures are calculated from these functions. Since any phenotype is not fully determined by genetic functions, "noise" is recognized as an important factor in quantitative genetics, as we emphasized above. What we mean specifically by noise, however, includes unknown sources of effect, as well as truly stochastic

factors, both biological and technical. The mathematical noise function we use here, ε, thus arises from several sources, particularly including these six:

1. Measurement errors in any of the variables, both phenotypes and genotypes
2. Environmental influences on the phenotypes
3. Epigenetic effects
4. Stochastic developmental & physiological effects
5. The effects of uncharacterized genetic variants (rare SNPs, CNV's, etc. - anything not included in our genetic variables), interactions that involve more than 2 genetic loci, which we do not include here, and weak effects from other two and three-variable tuples that are below a statistical threshold for consideration. These collectively are what are usually called genetic background effects.
6. Sampling noise (allele frequencies *vs* subjects, etc.), purely statistical fluctuations.

The "noise" as defined here, of course, is actually not noise in the usual sense of the word, but the composite of all unknown influences as well as truly stochastic inputs.

The discrete functions represent a vanishingly small fraction of all possible information functions. However, they can be used to describe real genetic effects, and generalized, by adding a noise function that modifies the probability of occurrence of each possible alphabet value of the phenotype. This allows us to flexibly represent general distributions for any specific alphabet size, and thereby defines the "noise" in our functions as described above. It is clear that the locations of these general functions on the information landscape are continuously distributed, as illustrated in [11] where we introduced random noise into the discrete functions. Here we introduce a systematic formulation, combining discrete functions with noise, by defining tensor functions.

In general terms, the relation between genetic loci and phenotypes is embodied in this formalism in discrete valued *loci-phenotype tensors*: $f(X_1, X_2, \ldots X_n; \Phi)$, where $\{X_i\}$ is the set of *n* genetic loci and Φ is the phenotype. For two genetic loci this is identical to a "gene-phenotype table". When we take account of "noise" distributions we add a uniform distribution to these tensors and form the genotype-phenotype tensors (GPT). The relation between these for *n* variables is simply

$$G(X_1, X_2, \dots X_n; p) = pf(X_1, X_2, X_3 \dots X_n) + (1-p)\varepsilon \qquad (19)$$

where ε is a uniform random "noise" function, and (1-p) is the "noise" level. The parameter p is the *penetrance*.

Since in the three-letter alphabet the discrete functions determine a third variable as a function of two others, the functions can be represented by 3x3x3 tensors, G_{ijk} where the indices i and j specify the genetic variables, and the index k specifies the alphabet of the phenotype. The entries in the tensor for a specific function defines the probability for a position to have a given letter of the alphabet. Let us define G_{ijk} such that G_{ijk}= 1 when $f_{ij} = k$, for the indices ranging over the alphabet, which are considered as non-negative integers here. Using this representation we can add noise to functions by entering probability values in the tensor that are other than 0 or 1. We use the uniform distribution, for which all letters are equally likely, as it is the only entirely unbiased representation of the "noise" component of the tensor.

For the genotype-phenotype tensor the fractional balance between the "noise" and the discrete function is a variable factor we call penetrance, in keeping with the usual use in genetics. If the penetrance is 1 there is no confounding noise, and if is small the genetic function plays that correspondingly small role in determining the phenotype. Note that if the *penetrance* is small the significance of the genetic effect is also small. Thus, there is a clear relation between penetrance and the p-value of the effect. This relation will be explicated further elsewhere.

We assume here that a full penetrance effect can be described by a single discrete function. It is possible that some linear combinations of discrete functions could be useful in some cases. We do not consider this more complex extension further in this paper.

It is important to see what happens to the coordinates as the penetrance decreases ("noise" increases). To see what the delta coordinates in the information landscape for tensor functions with low penetrance we examine the limiting ratios of the information functions. Since we cannot calculate deltas for the uniform distribution consider distributions infinitesimally close to the uniform distribution or to zero penetrance.

$$\left[\mathrm{P}_{\mathrm{ijk}}^{\mathrm{rand}}\right] = \frac{1}{27}\left[\begin{bmatrix} 1+\epsilon & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1-\epsilon & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\right], \tag{20}$$

and calculate the delta coordinates in the limit $\epsilon \to 0$. The above tensor can be used to calculate the corresponding joint entropy, for example, for the 3x3 case.

$$H_{123} = -\frac{25}{27}\log\frac{1}{27} - \left(\frac{1+\epsilon}{27}\right)\log\left(\frac{1+\epsilon}{27}\right) - \left(\frac{1-\epsilon}{27}\right)\log\left(\frac{1-\epsilon}{27}\right)$$

The delta-coordinates can be calculated from these entropies. The first delta-coordinate is

$$\delta_1 = \frac{-H_1 + H_{12} + H_{13} - H_{123}}{\sum_i H_i - H_{123}}$$

However, both the numerator and denominator of this expression go to zero in the limit when $\epsilon \to 0$, so we take the limit using L'Hospital's rule. The first derivatives are also each zero, but the second derivatives yield the limit:

$$\lim_{\epsilon \to 0} \delta_1 = \frac{8/81}{32/243} = \frac{3}{4}$$

This is in agreement with previous numerical results [11]. This location, $(\frac{3}{4}, \frac{3}{4}, \frac{3}{4})$, on the information plane has very particular properties that need to be carefully considered. When the tensor becomes entirely uniform the "noise" completely swamps out the information content of the functions, and the genetic information has no effect on the phenotype. This point corresponds to a value of the penetrance, *p*, of zero. It is the location on the information plane that we called the "black hole" previously [11]. As noise increases the functions all move on the landscape, eventually converging on this spot.

5.2 Algorithm for inferring Genotype-Phenotype Tensors

Since the relation between genetic loci and phenotypes is described by discrete valued *loci-phenotype tensors*, (see Equation 19), once we have used the information measures to determine that there is a significant dependence for a given set of variables, we need to infer the function itself. However, since the tensor is not described by a discrete function alone, we also need to infer the level of the essential "noise" distribution. As described above, together these

components, the discrete function and the "noise" level, or penetrance, form the genotype-phenotype tensors (GPT), where the function ε is a uniform random "noise" function, and (1-p) is the noise level, and the parameter p is the penetrance. We will henceforth write the tensors using indices that range over the variables and the alphabets. Thus $f(X_1, X_2, X_3 \ldots X_n, k)$ is written as $f_{i_1 i_2 i_3 \ldots i_n\, k}$, where the $\{i_1\}$ are genetic variant indices, and k is an alphabet index.

$$G(X_1, X_2, \ldots X_n;\ p) = p f_{i_1 i_2 i_3 \ldots i_n k} + (1-p)\varepsilon \qquad (21)$$

Given a data set and a significant tuple of variables (the dependence) there is a simple way to infer the function and p. The problem can be posed as follows for a three-way dependence. Let us represent the data by the data frequency tensor for a phenotype as a function of two genetic variants

$$D_{ijk} = \left\{ \begin{bmatrix} d_{111} & d_{121} & d_{131} \\ d_{211} & d_{221} & d_{231} \\ d_{311} & d_{321} & d_{331} \end{bmatrix}, \begin{bmatrix} d_{112} & d_{122} & d_{132} \\ d_{212} & d_{222} & d_{232} \\ d_{312} & d_{322} & d_{332} \end{bmatrix}, \begin{bmatrix} d_{113} & d_{123} & d_{133} \\ d_{213} & d_{223} & d_{233} \\ d_{313} & d_{323} & d_{333} \end{bmatrix} \right\} \qquad (22)$$

where all variables range over three-letter alphabets. This frequency tensor is defined to be normalized so that the sum of all components is 1.

Let us assume for the moment that the allele frequencies are equal. We will modify the resulting simple algorithm for non-equal allele frequencies later (note that this is moot for the haploid case we analyze in the last section). A greedy algorithm for finding the most likely function, f_{ijk}, simply identifies the maximum d_{ijk} for each letter of the alphabet k, and assigns a probability of one to that k and zeros to the other two for all i and j: $\forall i,j:\ f_{ijk} = \begin{cases} 1, & if\ d_{ijk} = \max_t(d_{ijt}) \\ 0, & otherwise \end{cases}$

The estimate of p is then the average frequency of the tensor elements not assigned a value of 1 in f_{ijk}. If the expectation is taken over all tensor elements then since there are 9 non-zero entries for f_{ijk}, we can write the expression for the penetrance, p, as

$$p \cong 1 - \tfrac{3}{2}\langle d_{ijk}(1 - f_{ijk})\rangle \qquad (23)$$

The algorithm yields the genotype-phenotype tensor

$$G_{ijk} = pf_{ijk} + (1-p)\varepsilon$$

Measuring how well the data is described by such an inferred tensor function, there are many ways of characterizing the resulting fit. In the spirit of the current formalism we can calculate the Kullback-Leibler divergence between D_{ijk} and G_{ijk}, but a chi-squared test also works. These tensors are normalized so that they can treated directly as distributions.

5.3 Simplification: Transformations of 3-functions into 2-functions

Each of the discrete functions of two variables, $Z(X,Y)$, can be specifically transformed into a pairwise function without loss of dependency information. By this we mean expressing it as a function of a single variable that maps to values of the pair X and Y. For example, the 3x3 function we call an XNOR-like function is represented by this matrix that defines Z values (the columns are X, the rows are Y).

$$\begin{bmatrix} 2 & 0 & 1 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

A transformation that maps pairs of (X,Y) values into another variable, call it W, can represent the essential information in the function: $\{(X_i, Y_j)\} \Rightarrow W$. What this means is that the matrix can be fully reconstructed by the mapping

$$\{(1,0),(2,1),(0,2)\} \Rightarrow 0$$
$$\{(2,0),(0,1),(1,2)\} \Rightarrow 1$$
$$\{(0,0),(1,1),(2,2)\} \Rightarrow 2$$

since this transformation simply yields values of the variable W such that W=Z. Every function can be mapped similarly. The resulting two-variable functions have the properties that the mutual information between W and Z are maximal. These transforms take these three-way functions with a variety of mixes between two and three variable dependencies into a space of pairwise-functions with only pairwise dependencies. An important question, however, is whether the mapping of the three variable-function space into a pairwise-function space of the same size exhibits symmetries and redundancies that reduce the complexity of the one-to-one transformations. In other words,

are there a set of "basis" transforms that can distinguish each of the 3-functions from one another when mapped into the pairwise-function space?

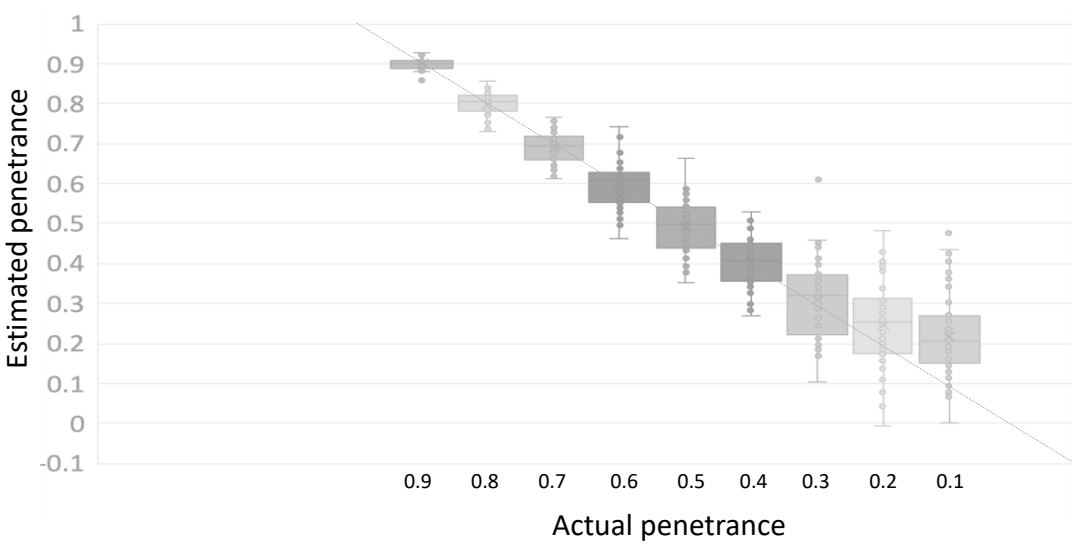
The information measures described here are sensitive to dependencies, but do not define the functional nature of the dependencies, the functions themselves. In other words, the mapping of functions into delta coordinates is many to one, as is evident from the multiple functions in a family of functions with the same coordinates. The geometry of the information landscape, however, limits the possibilities since the delta values define a location in the landscape and thereby restrict the functions that could be generating the observed dependence. The detection of dependence, localization on the information landscape, followed by the identification of the actual function that then leads to transformations from 3-function space to 2-function space are the steps in the process of complete characterization of genetic phenomena with two-loci functions affecting phenotypes. The next question then, is how can this paradigm be implemented?

5.4 Inferring the Function and Penetrance in Simulated Data

To test the effectiveness of the simple algorithm described in the previous section we simulated a data set of 100 subjects. As an example, we used a specific discrete function and a uniform "noise" function. The discrete function of three variables chosen in the case described here is shown in Figure 5, which exhibits both pairwise and three-way dependence.

We generated the genotypes randomly, making the simplifying assumption of equal allele frequencies, and used this function to determine the phenotypes, then added uniform noise to the tensor using the relation of Equation 19 to generate the data set for specific values for the penetrance. Correction for allele frequencies is a simple linear transformation. The algorithm was used to infer the discrete function, and estimate the penetrance. The results, both for this function and others not shown, showed that the algorithm works well to infer the exactly correct discrete function for all values of the penetrance greater than about 0.24. Penetrance levels less than this value lead to some incorrect entries, as shown in Figure 5.

a)



b)

Function								
0	1	0	1	0	0	0	0	1
1	1	0	0	0	1	0	0	0
0	0	1	0	1	0	1	0	0

Algorithm Result for p=0.1								
0	1	0	1	0	0	0	0	1
1	1	0	0	0	1	0	0	0
1	1	1	0	0	0	0	0	0

Algorithm Result for p=0.1								
0	0	0	1	0	1	0	1	0
1	1	0	0	0	1	0	0	0
0	0	0	0	1	0	1	0	1

Figure 5. Analysis of simulated data. a) These are values of the penetrance calculated for 50 simulated data sets each for 9 values of penetrance, *p*. For all of these values, except those shown in red (p =0.2 and p =0.1), the greedy algorithm returned the exact function. b) For these two there were a few errors in the function (top panel is correct function), as shown in these examples for two cases of p =0.1 simulations (errors are highlighted.)

It is clear that the simple algorithm provides a reasonably robust method for inferring a complex discrete function from data as well as estimating the penetrance. For larger data sets, of course, the thresholds for inference errors will be smaller than seen here (p = 0.24).

5.5 Genetic Heritability

The quantitation of heritability has been an important, long-standing problem in quantitative genetics. The ideas of broad and narrow sense heritability have their roots in the Fisher paradigm [1,2]. In the classical model the components of the trait value (phenotype) are commonly the population mean, the genetic effect and the "error term," that determine the population mean. The assumption of normally distributed components, with no covariance means that the variances add. The broad sense heritability is simply defined as the ratio of the genetic variance to the population average. Genomic heritability is the fraction of the genetic variance that can be explained by regression on the markers, and will only be quantitatively accurate when all causal variants are genotyped. See [22] for a nice discussion of heritability for panmictic populations.

When there is a way to determine the additive component of the genetic variance (requiring separate experimental data) the ratio of additive component to the population average is the "narrow sense" heritability. Classical methods use analysis of variance of full and half-sibling families, and using maximum-likelihood methods for relatives with different degrees of relatedness.

In the information theory formalism heritability can be reduced to a single quantity and is actually rather simple to state. Since the total dependence of a set of loci and a phenotype can be quantitated by the multi-information, we can use this quantity effectively to define heritability. For a given phenotype then we propose to define the heritability as the ratio of the total of all multi-informations over all subsets of dependent variables affecting the phenotype, divided by the "maximum possible" dependence for this set of loci. The maximum possible dependence is defined as the multi-information if there were full penetrance for all dependencies, no effective "noise". Therefore the heritability, $\mathcal{H}$ can be given by the expression

$$\mathcal{H}_{\phi} = \frac{\sum_{\tau} p_{\tau} \Omega_{\tau}(\phi)}{\sum_{\tau} \Omega_{\tau}(\phi)} \tag{24a}$$

where τ is any subset of variables containing the phenotype variable; that is, subsets of the set of all genetic loci and the specific phenotype, ϕ. The sum is over all possible such subsets, τ. If the penetrance is full for all determinants of the phenotype the heritability becomes one. This would mean that there are no environmental effects or other sources of "noise". One might think that since the sum over dependent tuples may not be disjoint, having some overlaps, that the dependency could be overcounted. This is not the case, however, since the measures are weighted by the penetrance and normalized by the total sum, including any possible overlaps. Note that in this definition we need not assume linkage equilibrium. In fact, Equation 24a is completely general.

Practically we must limit the sum to those subsets of variables whose dependencies are significant, so the criteria for significance enters the determination of heritability. This is because the calculation of p_τ can only be accurate if the dependence is significant. This definition is different from the classical form then in another way. We are calculating the heritability of all of trait based on all the variants considered in the analysis, while the variance form purports to includes all genetic effects. If the dependencies for a phenotype were all single locus then the heritability would be a function of the mutual informations between these loci and the phenotype, ϕ:

$$\mathcal{H}_\phi(single\ locus) = \frac{\sum_i p_i I(i,\phi)}{\sum_i I(i,\phi)} \qquad (24b)$$

where $I(i,\phi)$ is the mutual information between the loci and the phenotype ϕ.

In this paper we also restrict the sum to triplets, subsets of three variables, two loci and the phenotype of interest, although Equation 24a is certainly valid for any size of subset τ, and any number of genetic loci. Since the composition of dependence for each triplet can be clearly separated into the components due to single locus and two loci dependence, as long as the two loci are independently segregating ($I(i,j) = 0$) we can also then separate the heritability into two components by separating the sum in the numerator into two parts, the pairwise or additive effect and the three-way effect. From Equation 9 it is clear in this case that for the heritability limited to two locus epistatic interactions, $\mathcal{H}_\phi(triple)$, we have (where τ indicates all the triple dependencies)

$$\mathcal{H}_{\phi}(triple), = \frac{\sum_{\tau} p_{\tau} \Delta_3(\tau)}{\sum_{\tau} \Omega_{\tau}} \tag{25}$$

where τ indicates all triple dependencies. This formulation provides a rigorous and complete description of heritability given the division between the genetic determinants and the unknowns, the "noise". It also provides a practical way to calculate the heritability under specific assumptions. Contrast this with Fisher's heritability, *in the broad sense*, which is the ratio of the variances of phenotype to genotype in the population. Narrow sense heritability is more important in the sense that it quantitates the proportion of the phenotypic variation that is transmitted from parents to offspring [22]. The argument for this interpretation that ignores epistatic effects, which are frequently disrupted by segregation, is plausible, but incomplete.

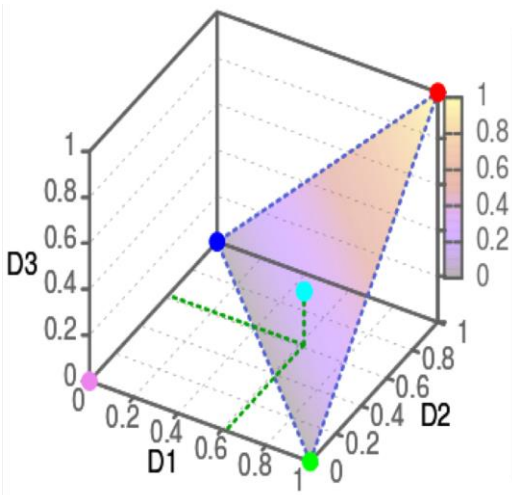
5.6 Protective Alleles

The interaction of two loci, of course, means that each locus may modify the effect of the other in some way. To make this more precise, and illustrate this kind of interaction in our formalism, we look at a specific, concrete case and examine the instance of protective alleles. The hallmark of protective effects is easily described in terms of the genotype-phenotype table. For simplicity let us consider a binary phenotype where 1 is a negative phenotype, a disease state, and 0 is normal. The notion of a protective allele then simply means that one variant of gene A has the effect of reversing the disease effect of gene B and making the phenotype normal. An example of a model that shows protective effects is illustrated in Table 3.

B\A	**0**	**1**	**2**
0	0	0	0
1	1	0	0
2	1	1	0

Table 3. A genotype-phenotype table (100% penetrance) illustrating a protective effect of gene A alleles on the disease-causing alleles of gene B. Phenotype 1 indicates diseased, 0 is normal.

The minor allele of gene B causes the pathology except in the presence of the minor allele of gene A, the protective allele. There are, of course other functions that exhibit such effects.



Func	D1	D2	D3	Logic	!
0000	0	0	0	Z=0	0
1111	0	0	0	Z=1	0
0110	1	1	1	Z=XOR(X,Y)	.69
1001	1	1	1	Z=NOT(XOR(X,Y))	.69
0001	0.616	0.616	0.233	Z=AND(X,Y)	.56
0010	0.616	0.616	0.233	Z=AND(NOT(X),Y)	.56
0100	0.616	0.616	0.233	Z=AND(X,NOT(Y))	.56
0111	0.616	0.616	0.233	Z=OR(X,Y)	.56
1000	0.616	0.616	0.233	Z=NOT(OR(X,Y))	.56
1011	0.616	0.616	0.233	Z=OR(NOT(X),Y)	.56
1101	0.616	0.616	0.233	Z=OR(X,NOT(Y))	.56
1110	0.616	0.616	0.233	Z=NOT(AND(X,Y))	.56
0101	1	0	0	Z=X	.69
1010	1	0	0	Z=NOT(X)	.69
0011	0	1	0	Z=Y	.69
1100	0	1	0	Z=NOT(Y)	.69

Protective Allele models

Figure 6. Four of the 16 2x2 genetic models show protective effects. The functions are shown in linear form and color coded according to the families as marked on the information plane.

In order to illustrate the systematic effect in a simple case we consider the haploid genetics case with binary phenotype, where there are only 16 possible genetic models (Figure 6). Only four of the 16 possible 2x2 genetic models exhibit protective effects.

6. Analyzing a Yeast Genetic Data set

To illustrate the application of the information theory approach to quantitative genetics we analyze a data set of haploid data from a large yeast cross generated by Kruglyak and colleagues [23]**.** The data consist of 4,390 haploid strains resulting from the cross of a wild, vineyard strain, and a widely used laboratory strain of *Saccharomyces cerevisiae.* This is an F2 cross, so that the recombinations between the two parental chromosomes occur in a single meiosis event for each of the resulting strains. The resulting haploid strains are essentially the gametes from the hybrid F1 strains. The data includes genotypes of all 4,390 strains, at 28,820 SNP positions, and 20 phenotypes, average growth rates under different conditions and in the presence of different compounds. We have restricted our use of the data to those phenotypes that showed a relatively high reproducibility in replicates. We used only those phenotypes whose replicates exhibited highly consistent correlation coefficients. This criteria, a replicate correlation coefficient above

0.8, selected 4 of the 20 phenotypes reported by Bloom *et al*. [23]. We report the analysis of two of these four phenotypes: growth in the presence of Neomycin (correlation coefficient 0.86), and Copper sulfate (0.82).

We calculated the pairwise effects, mutual information, between single genetic variants and the phenotype, and the measure of three-way effects, using a representative set of 100 variant markers across the genome. To calculate the three-way interactions accurately we wanted independently segregating markers, so we selected a set of 100 markers that were isolated by iteratively eliminating one of each pair of markers that had a mutual information of more than 0.05. The markers were widely spread and we calculated the recombination frequencies between each pair of neighboring markers to assess the statistics of segregation. The results for pairwise and three-way genetic dependencies for some of the phenotypes are shown in Figures 7-10. For our significance calculations, we permuted the phenotype variables 100 times and derived *p*-values from the resulting distribution of dependency measures.

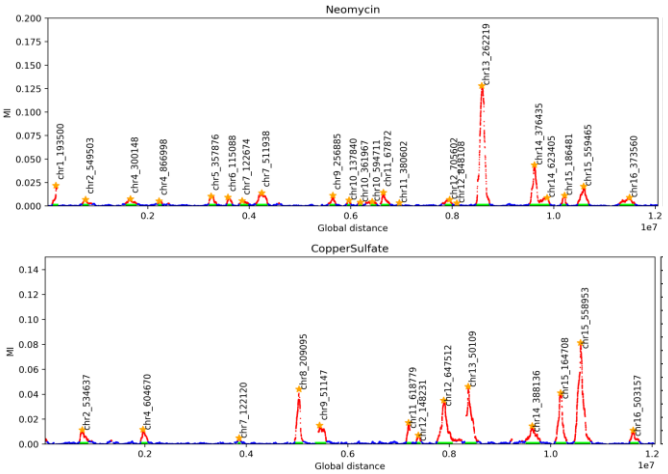


Figure 7. The pairwise peaks for the genetic determinants of two phenotypes. The locations indicated are the chromosomal coordinates of the highest scoring marker in the peak.

The two panels in Figure 7 show the pairwise peaks resulting from plotting the mutual information for the entire set of 28,820 genetic markers for each of the four phenotypes. The tables at the right of each panel show the location of the highpoint of each peak, the standard deviation and the width. The width of the peak is defined by the outermost boundary of the peak determined by the locations of the last significant marker by mutual information on each side of the peak. These are the same

peaks indicated. The peak widths in this case are due to the co-segregation of contiguous blocks in the F2 meiosis.

Neomycin

Peak	STD	Width
chr1_193500	22923.9	87228
chr2_549503	32012.6	123030
chr4_300148	55234.1	220149
chr4_866998	22360.5	83612
chr5_357876	38574.9	148433
chr6_115088	33540.9	102487
chr7_122674	41580.8	166285
chr7_511938	60005.7	226469
chr9_256885	31559.4	117476
chr10_137840	18979.8	68372
chr10_361967	18761.3	66082
chr10_594711	30922.9	113917
chr11_67872	48640.4	170876
chr11_380602	1230.0	4248
chr12_705602	39973.0	139637
chr12_848108	2503.8	8462
chr13_262219	80355.8	263928
chr14_376435	47175.2	190403
chr14_623405	48656.3	204340
chr15_186481	24329.6	88475
chr15_559465	60872.8	216317
chr16_373560	87976.6	296871

Copper sulfate

Peak	STD	Width
chr2_534637	48233.6	169846
chr4_604670	36277.6	140271
chr7_122120	15259.7	52115
chr8_209095	36374.4	159134
chr9_51147	44126.4	151924
chr11_618779	19211.5	69362
chr12_148231	21742.2	83786
chr12_647512	87081.0	309232
chr13_50109	50420.5	203113
chr14_388136	82244.5	286873
chr15_164708	56742.5	212833
chr15_558953	82886.9	311335
chr16_503157	45362.8	178043

Table 4. Pairwise peaks: global location, standard deviation and widths. The coordinates here are the global coordinates in base pairs, beginning with chromosome I. The widths and the standard deviations are in base pairs. The width is the distance between the two extreme, significantly scoring markers.

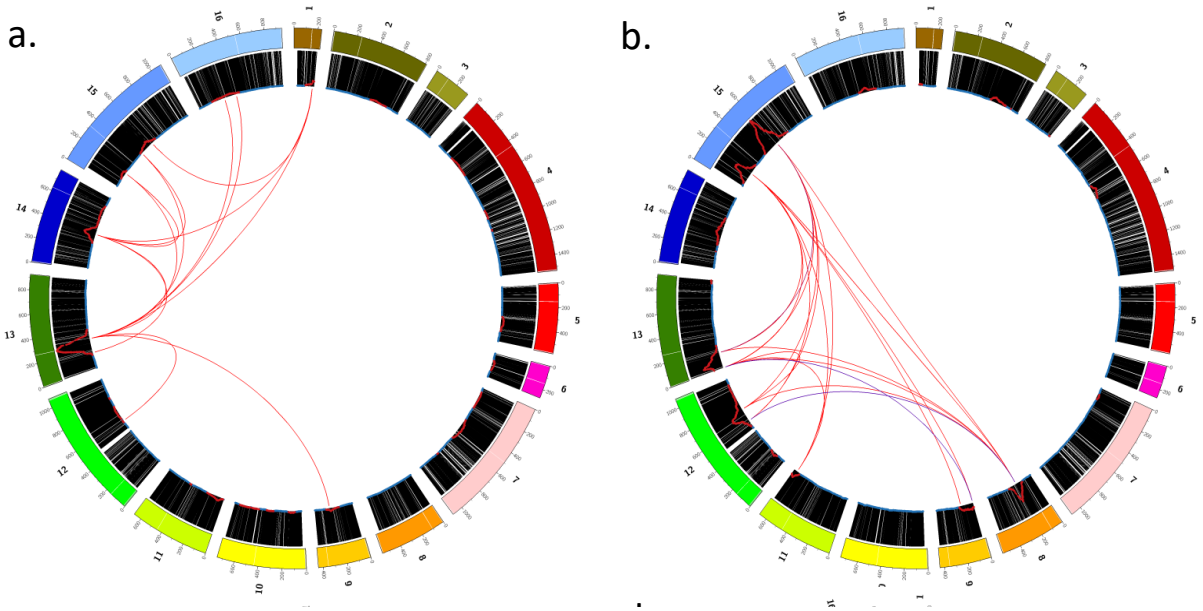


Figure 8. Genetic effects for two growth phenotypes. The full genome is shown with the single and two locus effects. The pairwise peaks for these phenotypes, as shown in Figure 7, are indicated as the red curves in the black band (using all 28,820 markers.) The variant pair interaction effects for these phenotypes are indicated by the internal red lines (all pairs are shown in the supplementary tables). The red lines indicate the significant three-way dependencies between the two markers at the ends of the line and the phenotype, indicating genetic loci interacting. **a.** Genetics of Growth on Neomycin, **b.** Growth on Copper sulfate.

In supplementary Tables S5a and S5b we indicate the quantitative results for the inter-chromosomal interactions shown in Figure 8. We have focused on these interactions because in this cross, with only one meiotic recombination the correlated blocks of markers are significant and even the widely spaced markers used here have some residual correlation. The inter-chromosomal pairs have no systematic correlations and the epistatic effect calculations are most accurate for these, as explained above. The fractions of the interactions that are attributed to epistatic and additive effects are indicated in the final two columns. There are variable levels of epistasis, but they are all below 10% of the effect.

The well-known metallothionein gene, CUP1, which is in the peak on chromosome 8, participates in several interactions. Note for all the interactions described here that since the markers used for this analysis have relatively low resolution, two neighboring markers may well indicate the same epistatic interaction. This is particularly likely when these markers are within the same mutual information peak, but we do not separate these effects here.

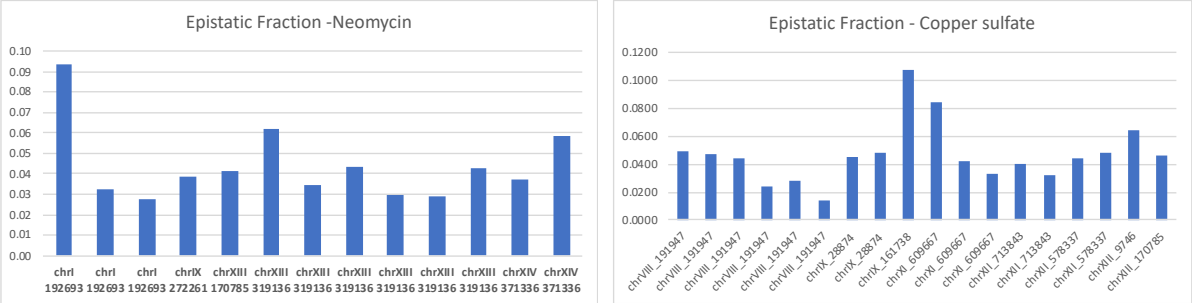


Figure 9. Epistatic fractions compared. The bars represent the fractional epistatic effect for the interactions in the order listed in supplementary Tables S5a and S5b. The marker indicated below each bar is the one listed in the left-hand column of the tables, representing the pair.

There are several notable features in the data of Figure 9. First, the epistatic fractions vary several-fold, the variation extends to those interactions with the same marker. The left-most section of the copper sulfate panel (six interactions) indicates the interaction of the marker on chromosome VIII near the CUP1 locus.

Bloom *et al*. [23] reported that the Zeocin phenotype exhibited the most non-additive interactions, which is also consistent with our results (not shown here). The Magnesium sulfate phenotype (not shown here) exhibits a very different pattern of interchromosomal interactions in that one locus on

chromosome 7 participates in fifteen different interactions, and there are no other pairs that are significant. The chromosome 7 locus is also the strongest pairwise locus.

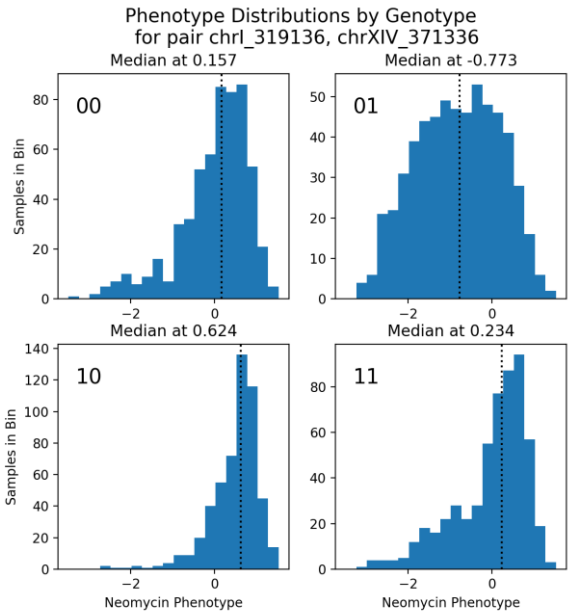


Figure 10. Phenotype distributions by genotype. We examined the tuple with the highest value of Ω for the Neomycin phenotype: loci chrI_319136 and chrXIV_371336. The panels show the phenotype distribution for each genotype (*e.g.*, plot 01 shows samples with chrI_319136=0 and chrXIV_371336=1).

To see an example of the detected epistatic interactions in the data, we analyzed the tuple which had the largest detected multi-information, Ω , with the Neomycin phenotype: loci chrI_319136 and chrXIV_371336. As shown in Table S5a, this tuple has a multi-information of $\Omega = 0.1116$, and an epistatic fraction of only 0.062 (*i.e.*, an additive fraction of 0.938). The phenotype distribution for samples of each unique genotype is shown in Figure 10. Even with this large additive fraction, it is clear from the phenotype distributions that we cannot consider this interaction to be entirely additive. Genotype 10 has the highest median phenotype at 0.624. Flipping either locus to 00 or 11 results in medians of 0.157 and 0.234, respectively (decreases of 0.467 and 0.39). If these effects simply added, we would expect a decrease of about 0.86. Genotype 01, however, has its median decreased by nearly 1.4, well beyond that expected from an additive effect. This is corroborated by a qualitative assessment of each distribution: the

distribution of genotype 01 is markedly different from that of any other genotype, implying that even the relatively small epistatic interactions which we detect have a real and noticeable consequence.

7. Discussion

With the formulation presented here we have begun to build a structure based on information theory for describing and making calculations in quantitative genetics. The advantages of an information theory approach to quantitative genetics are multiple and include these: a model-free, agnostic approach to multivariable dependence, the simplicity of the formalism, and the fundamental separation of dependence detection from the ascertainment of the functional nature of the dependency. These differences permit us to formulate the basic quantities in a direct, simple and calculable way. This contrasts with the classical formulation which focuses on the variances of phenotypes and genetic variance, and the, sometimes tacit, dependence on Gaussian assumptions and regression. As Huang and MacKay [4] have pointed out: "The crux of the problem is the undesirable features of the classical model as well as the alternative parameterizations that there is not a one-to-one correspondence between gene action at underlying quantitative trait loci and the partitioning of variance components except under very specific and restrictive circumstances". Many researchers have pointed to the need for an alternative approach to the classical model [3,4] and this paper begins the construction of such a formulation using information theory. We recognize that the formulation of this paper, limited to panmictic populations provides only the first steps and that more development, which we plan for future publications, is needed. It has been remarked that no natural population is panmictic, and that linkage disequilibrium, relatedness and population structure must be described in the information theory formulation. These situations are amenable to information formulation, albeit somewhat more complex, and will be treated in a future publication.

Since the formulation presented here uses discrete functions extensively it is important to understand the advantages and limitations of this underlying structure. While there are a finite number of discrete functions for any finite number of variables and alphabets, there are an infinite number of possible distributions for any finite number of variables and alphabets. For example, there are 3^9 =19,683 3x3 discrete functions, (3 variables and three-letter alphabets). This obvious

distinction is important, and clear that the function-to-distribution mapping is not one-to-one – there are vastly more distributions than discrete functions. Since we have assumed no linkage disequilibrium in this paper the "information landscape" for two variants was confined to a plane. When the genetic variants are not independent variables, and linkage is present, the "information landscape" is no longer a plane.

It is clear that the addition of "noise" to the discrete functions (see [11] and section 4.6) generates distributions around the discrete functions at varying distances in the information landscape. The point where the uniform noise distribution fully dominates and masks all information content, the "black hole" of information [11], represents the single distribution with uniform probabilities – it has maximal entropies. It may seem extraordinary, however, that this point has finite and distinct coordinates in the information landscape, as we describe in section 4.6. This points to the fact that the landscape includes functions with a wide range of penetrance and heritability, and that they are not at all distributed continuously on the landscape.

What is important here is to clearly define the distinctions and classifications that information theory can provide, and what these measures do and do not distinguish. This is an important specific question. One striking example is that the distribution of the "black hole" [11] and the discrete function that is XOR-like (with no non-zero pairwise measures) have distinct and unique coordinates in the information space for three variables, but as the alphabet size increases without bound (allowing more and more precise definitions of the variable values) they converge in the normalized space coordinates.

The application of this information-based formalism to genetics brings forward a number of interesting and important relations. The quantitation of the penetrance, for example, is simple and direct, but depends on a very clear division between what variables are being included to make the inferences and what is being ignored and considered to contribute to the "noise". We characterize the non-included variables with the "noise," even if they include more complex genetic interaction effects that are ignored. This is obvious and commonplace, but our formulation forces it to be explicit. Likewise, the information theory expression for heritability is both direct and intuitive, but also depends, as it must, on the precise assumptions made. Here we have ignored any variant

interactions among more than two genetic loci, and effects that involve three or more loci can therefore contribute noise directly to the penetrance and heritability calculations. We are forced to be explicit about what is meant by genetic background effects, which also include any effects of variants not considered as variables in the analysis.

Since the yeast two-strain cross involves only a single meiosis per strain to produce the collection analyzed, individual marker segregation within chromosomes is not a good assumption since blocks of markers will necessarily segregate together, while individual chromosome segregation, on the other hand, is generally an excellent assumption, and is supported by the data. In future work we will consider the more complex case, as in human data, where the possibility of non-independent segregation, and non-uniform population structures, which can have similar effects, are included. This extension will be an important part of the full formalism.

Bloom *et al*. [23] argue that the additive effects are much greater than the epistatic effects on the quantitative traits they measured using the variance component method, estimating a 9% overall interaction effect. We agree with this qualitative conclusion, but we calculate the interactions specifically for each pair of inter-chromosomal markers, and found that the levels of significant interactions among variants are highly variable, but in this general range. Huang and MacKay [4] specifically point out that these authors did not use the variance calculations correctly, and in fact, the variance method cannot be used to make this estimate. It is therefore not surprising to find quantitative disagreement with our results. Results in similar crosses in mice also show small epistatic effects as expected [29]. These authors use logistic regression methods to look at both pleiotropic and epistatic effects of pairwise identified loci.

We argue that the ability to quantitate the pairwise, additive and epistatic effects is definitely a significant step forward, and provides a distinctly different and practical alternative to the classical model. Comparison of our specific results with those of Bloom *et al*., show close agreement for the pairwise effects, but there is no way to make the full comparison for multi-loci effects effect for the overall estimate of additivity. The major difference with [23] is that we accurately calculate the specific fraction of additive and epistatic effects for each pair of loci.

The present contribution, a first step towards a full information theory of quantitative genetics, leaves a number of important problems unaddressed. These include the characterization of population structure, relatedness and linkage disequilibrium, and their incorporation into the formalism, so they are accounted for in the genetic inferences. Genes with significant linkage may also interact and the information theory formalism method must be extended to treat these cases accurately. While large data set sizes are required to accurately assess more interacting loci this situation can also be addressed in the formalism, and may be important in future as data sets of genetic information grow rapidly. Future work will focus on addressing all these issues and incorporating extensions into a formalism that will provide a broadly applicable set of descriptive and analytic tools.

Acknowledgements

This work was partially supported by the NSF National Science Foundation, EAGER grant (IIS-1340619), by the Pacific Northwest Research Institute, and the Bill and Melinda Gates Foundation. In the early stages of this work the Luxembourg Centre for Systems Biomedicine (LCSB) of the University of Luxembourg also provided significant partial support. We thank Joe Nadeau, Aimee Dudley and Greg Carter for stimulating discussions and for useful comments on the manuscript, and Elaine Skeffington for expert copy editing of the manuscript. None of the views expressed here should be attributed to the funding organizations. Neither did they play any role in the decision to publish or process of publication.

Author contributions: DG conceived of the project, DG, NS and JK-G provided and developed key ideas for specific parts of the formalism and paper sections. LU and JK-G carried out calculations on the yeast data. DG wrote the first draft of the paper and NS and JK-G further edited the manuscript.

References

1. Fisher, R.A., "The correlation between relatives on the supposition of Mendelian inheritance." Trans R Soc Edinburgh, 52(2):399-433. doi: 10.1017/S0080456800012163 (1918).
2. Wright, S.G., "A frequency curve adapted to variation in percentage occurrence." J Am Stat Assoc, 21(154):162-178 (1926).
3. Nelson, R.M., Petterson, M.E., Carlborg,O., "A century after Fisher: time for a new paradigm in quantitative genetics." Trends Genet, 29(12):669-676 (2013).
4. Huang, W. and Mackay, T.F., "The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis." PLoS Genet, 12(11). doi:10.1371 (2016).
5. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., Nadeau, J.H., "Missing heritability and strategies for finding the underlying causes of complex disease." Nat Rev Genet, 11(6):446–450 (2010).
6. Mackay, T.F., "Epistasis and Quantitative Traits: Using Model Organisms to Study Gene-Gene Interactions." Nat Rev Genet, 15(1):22-33. doi: 10.1038/nrg3627 (2014).
7. Phillips, P.C., "Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems." Nat Rev Genet, 9(11):855-867. doi: 10.1038/nrg2452 (2008).
8. Hill, W.G., Goddard, M.E., Visscher, P.M., "Data and theory point to mainly additive genetic variance for complex traits". PLoS Genet; 4(2):e1000008 (2014).
9. Gilbert-Diamond, D. and Moore, J.H., "Analysis of Gene-gene Interactions." Curr Protoc Hum Genet, 70(1) doi:10.1002/0471142905.hg0114s70 (2011).
10. Crow, J.F., "On epistasis: why it is unimportant in polygenic directional selection." Philos Trans R Soc Lond B Biol Sci, 365(1544):12411244 (2010).
11. Sakhanenko, N.A., Kunert-Graf, J. and Galas, D.J., "The Information Content of Discrete Functions and Their Application to Genetic Data Analysis." J Comput Biol, 24(12):1153-1178 (2017).
12. Sakhanenko, N.A. and Galas, D.J., "Biological data analysis as an information theory problem: Multivariable dependence measures and the Shadows algorithm." J Comput Biol, 22(11):1005-24 (2015).

13. Galas, D. J., Sakhanenko, N. A., Skupin, A., and Ignac, T., "Describing the complexity of systems: multi-variable "set complexity" and the information basis of systems biology." J Comput Biol, 21(2):118-40 (2014).
14. Galas, D.J., Dewey, Kunert-Graf, J., Sakhanenko, N.A., "Expansion of the Kullback-Leibler Divergence, and a new class of information metrics." Axioms, 6(2), 8; doi:10.3390/axioms6020008 (2017).
15. Galas, D.J., Nykter, M., Carter, G.W., Price, N.D., and Shmulevich, I., "Biological information as set based complexity." IEEE Trans Inf Theory, 56(2):667 – 677 (2010).s
16. McGill, W. J., "Multivariate information transmission." Psychometrika, 19(2):97–116 (1954).
17. Bell, A. J., "The co-information lattice", In *ICA 2003*, Nara, Japan, April (2003).
18. Sakhanenko, N.A., Galas, D.J., "Symmetries among multivariate information measures explored using Möbius operators." Entropy, 21(1), 88 (2019).
19. Watanabe, S., "Information theoretic analysis of multivariate correlation", IBM J Res Dev, 4: 66-82 (1960).
20. Hu Kuo Tin, "On the Amount of Information", Theory Probab Appl, 7(4):439-444 (1962).
21. Han, T.S., "Multiple Mutual Information and multiple interactions in frequency data." Information and Control, 46:26-45; doi: 10.1016/s0019-9958(80)90478-7 (1980).
22. Lstiburek, M., Bittner, V., Hodge, G.R., Picek, J. and Mackay, T.F.C., "Estimating realized heritability in panmictic populations." Genetics, 208(1):89-95 (2018).
23. Bloom, J.S., Kotenko, I., Sadhu, M.J., Treusch, S., Albert, F.W., and Kruglyak, L., "Genetic interactions contribute less than additive effects to quantitative trait variation in yeast." Nat Commun, 6, 8712 (2015).
24. Gregersen, J.W., *et al*., "Functional epistasis on a common MHC haplotype associated with multiple sclerosis." Nature, 443(7111):574-577 (2006).
25. Coutinho *et al*., "Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in determination of platelet serotonin levels." Hu Genet, 121:243-256 (2007).
26. Philips, P.C., Johnson,N.A., "The population genetics of synthetic lethals." Genetics, 150(1):449-458 (1998).
27. Wiltshire, S. *et al*., "Epistasis between type 2 diabetes susceptibility loci on chromosomes 2q21-25 and 10q23-26 in Northern Europeans." Ann Hum Genet, 70:726-737 (2006).

28. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J., “Shared Information: new insights and problems in decomposing information in complex systems.” *Proceedings of the ECCS, 2012*, Springer, 251-269, (2012).
29. Tyler, A.L., Donahue, L.R., Churchill, G.A., Carter, G.W., “Weak Epistasis Generally Stabilizes Phenotypes in a Mouse Intercross.” PLoS Genet, 12(2): e1005805s (2016).

Supplementary Materials

Appendix A. Proof of Proposition 2

By definition $\Omega = H_1 + H_2 + H_3 - H_{123}$ and the mutual informations, and interaction information, I_{123}, are

$$I_{123} = H_1 + H_2 + H_3 - H_{12} - H_{13} - H_{23} + H_{123}$$

$$I_{12} = H_1 + H_2 - H_{12} \text{ etc…}$$

So we have

$$\Omega = I_{12} + I_{13} + I_{23} - I_{123}$$

Since the recursion relations for the interaction information are

$$I_{12} - I_{123} = I_{12|3}$$

$$I_{13} - I_{123} = I_{13|2}$$

$$I_{23} - I_{123} = I_{23|1}$$

we can write

$$\Omega = I_{13} + I_{23} + I_{12|3}$$

$$\Omega = I_{12} + I_{23} + I_{13|2}$$

$$\Omega = I_{13} + I_{12} + I_{23|1}$$

Since by the definition of the mutual information $I_{12|3} = H_{1|3} + H_{2|3} - H_{12|3}$, and we have the identity $H_{2|3} - H_{12|3} = H_{1|23}$ by the basic probability identities, so $I_{12|3} = H_{1|3} - H_{1|23}$.

Now since $H_{1|3} \geq H_{1|23}$, $I_{12|3} \geq 0$, the above three equations imply,

$$\Omega \geq I_{13} + I_{23}$$

$$\Omega \geq I_{12} + I_{23}$$

$$\Omega \geq I_{13} + I_{12}$$

And the theorem is proved.

Appendix B. Maximizing the Symmetric Delta finds any XOR-like dependence first

Equation 9a implies that if the mutual informations are all zero or very small, for a triplet, the deltas must all be equal, or close to it. Furthermore, they must all be equal (or close to it for small mutual informations) to the Ω. The product of the deltas, the symmetric delta, is maximized when the factors are equal given a constant sum. This means that maximizing the symmetric delta will tend toward picking out the functions with deltas that are closest, given the total correlation or multi-informations, Ω, are constant. The XOR-like functions therefore maximize the symmetric delta for given Ω. An easy calculation shows the maximum of the symmetric delta, $\overline{\Delta}$, occurs when all factors are the same. The variable measures defined as

$$\overline{\Delta} = \Omega^3 \left(1 - \frac{x_1}{\Omega}\right)\left(1 - \frac{x_2}{\Omega}\right)\left(1 - \frac{x_3}{\Omega}\right) = \Omega^3 \prod_k \left(1 - \frac{x_k}{\Omega}\right) \tag{B1}$$

where $x_1 = \frac{I_{12}+I_{13}}{\Omega}$, $x_2 = \frac{I_{12}+I_{23}}{\Omega}$, $x_3 = \frac{I_{13}+I_{23}}{\Omega}$

Then for all triplets, indicated by *i* and *j*, $\overline{\Delta}_i > \overline{\Delta}_j$ only if

$$\Omega_i{}^3 \prod_k \left(1 - \frac{x_{ki}}{\Omega_i}\right) > \Omega_j{}^3 \prod_k \left(1 - \frac{x_{kj}}{\Omega_j}\right) \tag{B2}$$

It is clear that for similar values of Ω the largest is the one with all x's equal to zero, which is the XOR-like case. Thus, if there is an XOR-like dependence present in the data it will be at the top of the symmetric delta scoring list.

Appendix C. A matrix formulation of the 3-way dependence relations

The set of equations in 9a can be expressed simply in matrix form by defining the vectors:

$$\vec{\Omega} \equiv \begin{bmatrix} \Omega \\ \Omega \\ \Omega \end{bmatrix}, \quad \vec{\Delta} \equiv \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \end{bmatrix}, \quad \vec{\mu} \equiv \begin{bmatrix} \mathrm{I}_{12} \\ \mathrm{I}_{13} \\ \mathrm{I}_{23} \end{bmatrix} \tag{C1}$$

All of the components of these vectors are non-negative, and for the moment we are not assuming that $I_{12} = 0$, so this holds for the general case. Then we have a matrix relation,

$$\vec{\Omega} = \vec{\Delta} + \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \vec{\mu} \tag{C2}$$

The functions are thus confined to the landscape in only the non-negative sector (all coordinates are ≥ 0.

The matrix $\bar{\bar{M}} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$, has the inverse, $\bar{\bar{M}}^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \end{bmatrix}$, and writing the expressions for the other two vectors gives these simple equations

$$\vec{\mu} = \bar{\bar{M}}^{-1}\left(\vec{\Omega} - \vec{\Delta}\right) \tag{C3}$$

Thus, the 3-variable relations reduce to a simple equation in matrix form.

Appendix D. Interactions in the yeast data [24]

Marker1	Marker2	Omega	p-value	Epistatic Fraction	Additive fraction
chrI 192693	chrXV_605280	0.0345	2.020E-06	0.0936	0.9064
chrI 192693	chrXIII_319136	0.0856	2.020E-06	0.0324	0.9676
chrI 192693	chrXIV_371336	0.0633	2.020E-06	0.0275	0.9725
chrIX 272261	chrXIII_319136	0.0718	2.020E-06	0.0385	0.9615
chrXIII 170785	chrXIV_371336	0.0594	2.020E-06	0.0415	0.9585
chrXIII 319136	chrXIV_371336	0.1116	2.020E-06	0.0620	0.9380
chrXIII 319136	chrXV_189231	0.0757	2.020E-06	0.0346	0.9654
chrXIII 319136	chrXVI_433340	0.0693	2.020E-06	0.0435	0.9565
chrXIII 319136	chrXV_455183	0.0696	2.020E-06	0.0297	0.9703
chrXIII 319136	chrXVI_304984	0.0711	2.020E-06	0.0290	0.9710
chrXIII 319136	chrXII_578337	0.0686	2.020E-06	0.0429	0.9571
chrXIV 371336	chrXV_189231	0.0539	2.020E-06	0.0373	0.9626
chrXIV 371336	chrXV_455183	0.0491	2.020E-06	0.0582	0.9418

Table S5a. The inter-chromosomal interactions for the Neomycin phenotype, exhibiting the markers (identical markers are color-coded for better visualization.). These interactions all have <5% FDR. The p-values, calculated from the permutation of the data, are approximately the same.

Marker1	Marker2	Omega	p-val	Epistatic Fraction	Additive Fraction
chrVIII_191947	chrXIII_9746	0.0575	2.02E-06	0.0493	0.9507
chrVIII_191947	chrXV_189231	0.0655	2.02E-06	0.0468	0.9532
chrVIII_191947	chrXIII_170785	0.0444	2.02E-06	0.0438	0.9562
chrVIII_191947	chrXII_578337	0.0449	8.08E-06	0.0243	0.9757
chrVIII_191947	chrXII_713843	0.0484	2.02E-06	0.0279	0.9721
chrVIII_191947	chrXV_605280	0.0724	2.02E-06	0.0137	0.9863
chrIX_28874	chrXV_189231	0.0423	2.02E-06	0.0454	0.9546
chrIX_28874	chrXIII_9746	0.0343	8.08E-06	0.0486	0.9514
chrIX_161738	chrXV_189231	0.0436	2.02E-06	0.1076	0.8924
chrXI_609667	chrXIII_9746	0.0392	2.02E-06	0.0842	0.9158
chrXI_609667	chrXV_605280	0.0550	2.02E-06	0.0419	0.9581
chrXI_609667	chrXV_189231	0.0451	2.02E-06	0.0327	0.9673
chrXII_713843	chrXIII_9746	0.0392	2.02E-06	0.0397	0.9603
chrXII_713843	chrXV_605280	0.0561	2.02E-06	0.0318	0.9682
chrXII_578337	chrXV_605280	0.0534	2.02E-06	0.0437	0.9563
chrXII_578337	chrXV_189231	0.0442	2.02E-06	0.0484	0.9516
chrXIII_9746	chrXV_189231	0.0566	2.02E-06	0.0638	0.9362
chrXIII_170785	chrXV_189231	0.0428	2.02E-06	0.0466	0.9534
chrXIII_170785	chrXV_605280	-0.0011	0.0508	4.04E-06	0.0215
chrXIII_9746	chrXV_605280	-0.0014	0.0633	2.02E-06	0.0214
chrXV_189231	chrXV_605280	-0.0016	0.0713	2.02E-06	0.0224

Table S5b. The inter-chromosomal interactions for the Copper sulfate phenotype, exhibiting the markers (identical markers are color-coded for better visualization.). These interactions all have <5% FDR. The p-values, calculated from the permutation of the data, are approximately the same.