Published on Sep 27, 2023 by Arcadia Science DOI: 10.57844/arcadia-53f4-da1a

Applying information theory to genetics can better explain biological phenomena

Genetic models of complex traits often rely on incorrect assumptions that drivers of trait variation are additive and independent. An information theoretic framework for analyzing trait variation can better capture phenomena like allelic dominance and gene-gene interaction.

Contributors (A-Z)

Feridun Mert Celebi, Megan L. Hochstrasser, David Q. Matus, David G. Mets, Austin H. Patton, Taylor Reiter, Ryan York

Version 1 · Mar 31, 2025

Purpose

Genetic analysis has been one of the most powerful tools for biological discovery, providing insight into almost every aspect of biology, ranging from identifying mechanisms supporting the cell cycle **[1][2]**, to guiding selective breeding for agriculture **[3]**, and identifying targets for disease treatment **[4]**. While phenotypes can be simple (e.g., a single gene can cause differences in pea color or lead to a genetic disease) the vast majority are subject to more elaborate causal mechanisms involving many genetic and non-genetic factors. Researchers studying these phenotypes (often called "complex" phenotypes) have relied on assumptions of additivity and independence among the elements driving individual-to-individual phenotypic variation. It’s widely appreciated that, in real data, these assumptions are often violated, potentially limiting the utility of and accuracy of some analyses **[5]**. However, this broad framework is retained for both historical and practical reasons **[6][7]**. Here, we explore a different and complementary mathematical framework that makes no assumptions about the drivers of phenotypic variation — we apply information theory to genetic questions with the objective of conducting system-wide analysis of large sets of genes, phenotypes, and environmental data.

This pub is intended as a regularly updated document covering how we are applying information theory to broad questions in genetics. As time progresses, and we release empirical studies of different topics, we will add sections here covering the information theory relevant to those studies. This work should be of interest to both geneticists and information theorists, but is primarily intended to formalize an information theoretic approach to genetic problems and make that approach available to geneticists. Accordingly, the first section after the introduction is a primer on major concepts in information theory intended for geneticists. The subsequent sections contain information theoretic definitions for genetic concepts and demonstrations of how these definitions provide insight into genetic processes.

- This pub is part of the **platform effort**, "Genetics: Decoding evolutionary drivers across biology." Visit the platform narrative for more background and context.

Historical background

Contemporary quantitative genetics treats genetic influences on phenotypes as additive and independent of one another, and, generally any one phenotype is assumed to be separate from others **[6]**. The reasons for this are both historical and practical. Just prior to the turn of the last century, the study of human phenotypic variation (biometrics) was at its peak. Early biometric studies observed that phenotypic distributions among humans were often continuous and, across generations, appeared to vary gradually and not in jumps (e.g., **[8]**). Therefore, the field assumed that drivers of this continuous variation were themselves continuous, a model consistent with the then-new theory of evolution — phenotypes were expected to change gradually across generations. The tools and principles developed during the period (e.g. the mixture model, the t-distribution, the chi-square distribution) reflect these assumptions and, ultimately, came to form much of the theoretical backing of modern statistical genetics **[9]**.

Around the same time von Tischermark, de Vries, Spillman, and Correns "rediscovered" the work of Gregor Mendel **[10]**. Mendel's observations contradicted the dogma of continuity developed by biometrics. Through now-famous sets of experiments, Mendel found that phenotypes can in fact vary discretely within populations and across generations. For example, the hybridization of a yellow and a green pea plant could produce offspring that were either yellow or green, but not a combination of the two. Thus, some of the inherited drivers of phenotypic differences were discrete and not continuous. Subsequent experimental work in a variety of different organisms has strongly reinforced this view **[10]** and ultimately led to the generation of the term "gene" to describe the indivisible unit of heritable variation **[11]**.

The presence of discrete units of inheritance (genes) and, in some settings, dramatic phenotypic change across generations led to a "non-gradualistic" view of inheritance (e.g., **[12]** and **[13]**). The "gradualists" and the "non-gradualists" were divided by a fundamental problem: how could phenotypes — often continuous and only gradually changing — be caused by discrete units of inheritance? Ronald Fisher provided a reconciliation in 1918. Through groundbreaking theoretical work, Fisher demonstrated that many discrete, additive, independent units of inheritance of small effect could generate continuously varying phenotypes within a population **[14]**. Furthermore, these assumptions were consistent with Mendel's results. Fisher suggested that each trait (and the factors influencing that trait) could segregate independently following mating. By elegantly providing a resolution to the continuous/discrete paradox, Fisher thus forged the fundamental assumptions for genetic analysis that we still rely on today **[7]**.

However, in the following decades, extensive work on the function and inheritance of genes established clear violations of additivity and independence **[10]**. Instead, modern biology has demonstrated that genes and their products are highly interactive and involved in complicated, nonlinear processes such as physical complexes, regulatory circuits, and metabolic circuits. Furthermore, these complex interactions may drive phenotypic variation across individuals via dependent and non-additive relationships between genes.

A clear example of such nonlinear relationships is epistasis **[15]**, in which the effect of one gene can mask or modify the phenotypic impact of another. Epistasis is a common feature of genetic systems and is so prevalent that researchers began to use it to identify functionally related genes **[10]**. Genes that, when combined, caused no different phenotype than the individual genes alone were called "epistasis groups." For example, in *Saccharomyces cerevisiae*, the members of the RAD52 epistasis group were all individually sensitive to irradiation, and when combined, were no more sensitive than any one mutant. This suggests a functional relationship between the individual genes; if a mutation in any one of the genes disrupts the "functional unit," then further mutations in other members of that unit will not change the phenotype **[16]**. Many epistasis groups were identified through mutagenesis, but naturally occurring epistasis is prevalent and important for evolution **[17]**. Fisher's initial reconciliation assumed no epistasis, an assumption that largely remains in contemporary models **[7]**. Given the complexity of biological systems, the resulting potential for phenomena like epistasis, and empirical evidence that such phenomena exist, a modeling framework that does not include gene-gene interactions (as is common in quantitative genetics) will likely fail to account for key aspects of the genotype-phenotype map. Indeed, in recent years many studies have explicitly demonstrated this problem **[18]**.

To date, the solution has not been obvious. If we use the same statistical framework that's been applied historically, capturing nonlinear relationships among genes would require data from an enormous number of individuals. Including interactions in traditional linear models (e.g., genome-wide association studies) would require the number of model parameters to scale with the square of the number of genetic or environmental factors. It's common to conduct human genetic analysis using hundreds of thousands of genetic loci. Capturing interactions between even 100,000 loci would require a model with 10 billion parameters. Fitting such a model would require data from more humans than exist. As a result, despite increasing computational power, the utility of these models to effectively capture nonlinearity will always be limited by the available data.

We suggest using information theory to quantify the drivers of trait variation. Information theory was originally developed to formalize thinking about encoding schemes for communication **[19]**, and to provide answers to questions like, "What's the minimal amount of information required to encode a message?" or "How many bits of information are required to store this text document?" Since its inception, information theory has become very broad. Importantly for genetic analysis, we can use it to partition and quantify the impact of factors driving variation in a set of data. This allows us to answer questions like, "How much better can I predict the phenotype of an individual if I know that individual’s genotype?" or "How much information does genetic data contain about disease state?" In contrast to methods traditionally used in quantitative genetics, it makes no assumptions about the nature of factors impacting variation, so it may enable new, tractable, analyses capturing nonlinear relationships and lead to better mappings between genotypes and phenotypes.

Entropy, divergence, and mutual information

In this section, we review some fundamental components of information theory and provide examples of how we might apply them to genetic data. In subsequent sections, we’ll expand on these examples and contrast genetically relevant information theoretic measures to similar measures from classical statistical genetics.

Entropy

Entropy, H, is the average amount of information necessary to unambiguously encode an event from a given "source" (defined by a probability distribution) and serves as a measure of the "randomness" of the event and the source that generated the event. In the context of genetics, the "source" could be a specific pair of parents or a specific population of individuals and the "events" would be the offspring of the cross or the members of that population. Across a given population, you could interpret the entropy of a phenotype as its predictability (e.g., "How reliably can you guess the phenotype of any given individual?"). Both genetic information (e.g., allelic state at a given locus) or phenotypic information (e.g., disease state) could define a random variable. Here, we provide the definition of entropy and examples of entropy calculations, first in the simple context of coin flips and then in the context of genes and phenotypes.

For random variable X that can take values of the alphabet $\mathcal{X}$ and is distributed according to $p(x) =$ Probability{$X = x$} for all $x \in \mathcal{X}$, the entropy, $\mathrm{H}(X)$, of the discrete random variable X is

$$\mathrm{H}(X) := -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

$\mathrm{H}(X)$ is the average (calculated above as the weighted sum) uncertainty of the values of X. By convention $0 \log 0 = 0$, so values of x with probability zero contribute no entropy. The selection of base for the logarithm determines the units of information. Here and for the rest of this work we use base 2, which results in information measured in bits. For reference, one bit is the amount of information that can be encoded by a binary digit.

Example 1: Coin tosses

Consider two coins: one fair, Pr{heads = 0.5}, and one biased, Pr{heads = 0.9}. The degree of uncertainty about the outcome of a coin toss is higher for the fair coin as compared to the biased coin. A toss of the fair coin is equally likely to result in heads or tails. The biased coin is more likely to turn up heads. Entropy captures this intuition. The entropy for the fair coin is

$$\begin{aligned}\mathrm{H}(X) &= -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \\ &= -\sum_{i=1}^{2} 0.5 \log_2 0.5 \\ &= -\sum_{i=1}^{2} 0.5 \cdot -1 \\ &= 0.5 + 0.5 = 1\end{aligned}$$

Whereas the entropy of the biased coin is

$$\begin{aligned}\mathrm{H}(X) &= -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \\ &= -0.9 \cdot \log_2(0.9) - 0.1 \cdot \log_2(0.1) \\ &\approx 0.137 + 0.332 \approx 0.469\end{aligned}$$

Thus, entropy is lower for the more predictable (biased) coin than for that of the less predictable (fair) coin. Indeed, the fair coin, with equivalent probability for all states, has the maximum entropy (1 bit) for a random variable with two states. For any random variable, a probability distribution that is uniform across states results in the maximal entropy.

Example 2: Allelic state at a single locus

Now consider two different genes, A and B, with variation in allelic state across a population of diploid organisms. One gene A has two alleles A and a, resulting in three allelic states, AA, Aa, and aa, for any individual in this population. Similarly, gene B has two alleles and three allelic states, BB, Bb, and bb. The allelic states of gene A are distributed uniformly across the population such that 1/3 individuals are AA, 1/3 are Aa, and 1/3 are aa. In contrast, gene B is distributed such that 8/10 individuals are BB, 1/10 are Bb, and 1/10 are bb. The entropy of the allelic state of gene A is

$$\begin{aligned}
\mathrm{H}(A) &= -\sum_{i=1}^{n} p(a_i)\log_2 p(a_i) \\
&= -\sum_{i=1}^{3} \frac{1}{3}\log_2 \frac{1}{3} \\
&= -\sum_{i=1}^{3} \frac{1}{3} \cdot -1.58 \\
&= -p(a_{AA})\log_2 p(a_{AA}) - p(a_{Aa})\log_2 p(a_{Aa}) - p(a_{aa})\log_2 p(a_{aa} \\
&\approx 0.528 + 0.528 + 0.528 \approx 1.58
\end{aligned}$$

As compared to the fair coin, with only two possible outcomes, the "fair" (equal probability of each allelic state across individuals) gene, with three possible states, has an increase in entropy: 1 bit vs ~ 1.58 bits. This is consistent with an increase in uncertainty for variables with more possible states. The entropy of B, with non-uniform probability of allelic states, is

$$\begin{aligned}\mathrm{H}(B) &= -\sum_{i=1}^{n} p(b_i)\log_2 p(b_i) \\ &= -p(b_{BB})\log_2 p(b_{BB}) - p(b_{Bb})\log_2 p(b_{Bb}) - p(b_{bb})\log_2 p(b_{bb}) \\ &= -0.8\cdot\log_2(0.8) - 0.1\cdot\log_2(0.1) - 0.1\cdot\log_2(0.1) \\ &\approx 0.258 + 0.332 + 0.332 \approx 0.922\end{aligned}$$

Thus, the difference between $\mathrm{H}(B)$ and $\mathrm{H}(A)$ is the difference in randomness between those two variables. As with the coin example, the gene with a uniform probability distribution over possible states has more entropy (is more random) than the gene with a non-uniform probability distribution over states.

Example 3: Single phenotype

Similar to allelic state, we can calculate the entropy of a phenotype in a population. Unlike allelic state, phenotypes are often continuous (e.g., height) and not discrete (e.g., disease state). Throughout this work, for simplicity of exposition, we will only examine equations for discrete phenotypes. However, there are tools for estimating the information theoretic values we describe for continuous variables as well. Consider a disease trait T that can have two conditions, sick t and healthy T, and T is distributed according to probability mass function $p(t)$. Across the population, 1/10 individuals are sick and 9/10 individuals are healthy. The entropy of T is

$$\begin{aligned}\mathrm{H}(T) &= -\sum_{i=1}^{n} p(t_i)\log_2 p(t_i) \\ &= -p(d_T)\log_2 p(t_T) - p(t_t)\log_2 p(t_t) \\ &\approx 0.137 + 0.332 \approx 0.469\end{aligned}$$

Joint Entropy

We can extend the definition of entropy stated above to more than one random variable. Given genes A and B with a joint distribution over allelic states of $p(a, b)$ their joint entropy is

$$\mathrm{H}(A, B) := -\sum_{a\in A}\sum_{b\in B} p(a, b)\log_2 p(a, b)$$

where the joint entropy is less than or equal to the maximal entropy of A and B, $\mathrm{H}(A,B) \leq \mathrm{H}(A) + \mathrm{H}(B)$, with equality, $\mathrm{H}(A,B) = \mathrm{H}(A) + \mathrm{H}(B)$, if and only if A and B are independent. Two examples of "independent" genes would be genes that are unlinked (e.g. two genes on different chromosomes) in a family or genes that have no correlated structure in a more complex population. The joint entropy of these genes would simply be the sum of their individual entropies. A corollary is that genes that are linked or genes that are correlated in a larger population will have a joint entropy that is less than the sum of their individual entropies.

As we will discuss later, the comparison between the maximal entropy and the joint entropy of a set of variables (such as phenotypes) is the decrease in randomness caused by relatedness among those variables. For a pair of traits, T_1 and T_2, $\mathrm{H}(T_1) + \mathrm{H}(T_2) - \mathrm{H}(T_1, T_2)$ is the decrease in randomness in the set of variables caused by knowing their joint distribution. Similarly, for a gene, G, and a disease, T, that is partially caused by that gene, the distribution of G and the distribution of T are not independent. Therefore $\mathrm{H}(G) + H(T) - \mathrm{H}(G,T)$ will be positive and, if there is no other population structure, is a measure of the amount of variation in disease state that is caused by the gene, G.

Conditional entropy

For two variables A and B, conditional entropy is the remaining randomness of A if B is known and is defined as

$$\mathrm{H}(A|B) := -\sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a,b) \log_2 p(a|b)$$

If A and B are genes whose allelic state is evenly distributed across a population and are completely linked, then knowing the allelic state of B would tell you the allelic state of A and $\mathrm{H}(A|B) = 0$. In contrast, in a similar population, if A and B are completely unlinked then $\mathrm{H}(A|B) = \mathrm{H}(A)$; knowing the allelic state of B tells you nothing about the allelic state of A. Here is a less deterministic example: for a gene, G, and a disease, T that is partially caused by that gene, $\mathrm{H}(T|G)$ is the amount of variation in disease state that is caused by factors other than G.

Furthermore, $\mathrm{H}(A|B) \neq H(B|A)$. In the context of genetics, if gene A has three allelic states in a population and gene B has two allelic states, but A and B are completely linked, then $\mathrm{H}(A) > \mathrm{H}(B)$. If you know the allelic state of A, you know the allelic state of B ($\mathrm{H}(B|A) = 0$), but, knowing the allelic state of B does not completely specify the allelic state of A; $\mathrm{H}(A|B) > 0$.

Mutual information

Mutual information, I, is the amount of information shared between two random variables. $\mathrm{I}(A; B)$ between two random variables A and B is the decrease in randomness in A if you know B, or B if you know A.

For two random variables A and B, which can take values from alphabet $\mathcal{A}$ and $\mathcal{B}$ respectively, and are distributed according to $p(a) =$ Probability {$A = a$} for all $a \in \mathcal{A}$ and $p(b) =$ Probability {$B = b$} for all $b \in \mathcal{B}$, the mutual information between A and B is

$$\mathrm{I}(A; B) := -\sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} p(a, b) \log_2 \frac{p(a, b)}{p(a)p(b)}$$

$\mathrm{I}(A; B)$ is always positive, or is zero if and only if A and B are independent, and $\mathrm{I}(A; B) = \mathrm{I}(B; A)$. An alternative definition is

$$\mathrm{I}(A; B) := \mathrm{H}(A) + \mathrm{H}(B) - \mathrm{H}(A, B)$$

In other words, it is the degree to which dependency between A and B reduces the joint entropy, $\mathrm{H}(A, B)$, below the maximum possible joint entropy. For two completely linked genes, A and B, with the same number of alleles that are evenly distributed in a population, $\mathrm{I}(A; B) = \mathrm{H}(A) = \mathrm{H}(B)$. For similar but unlinked genes, $\mathrm{I}(A; B) = 0$. In the context of a disease, T, and a gene, G, $\mathrm{I}(T; G)$ is the decrease of uncertainty about disease state because you know the allelic state of G.

Conditional mutual information

For three random variables, A, B, and C, we can define conditional mutual information as the shared information between A and B if we also know C.

$$\mathrm{I}(A; B|C) := \mathrm{H}(A|C) - \mathrm{H}(A|B, C)$$

$\mathrm{I}(A;B|C) \geq 0$ with equality if and only if A and B are independent if you know C. The conditional mutual information is the reduction in the uncertainty of A with knowledge of C if we then add knowledge about B. For example, we have a population where two genes, G_1 and G_2, and a trait, T, are segregating. The distribution of allelic state of G_1 is unrelated to the distribution of allelic state of G_2 (i.e., $\mathrm{I}(G_1;G_2) = 0$), but variation in G_1 combined with allelic variation at G_2 *causes* all of the variation in T. In this case, even though G_1 tells you nothing about G_2 on its own, if conditioned on knowledge of T, G_1 can tell you something about G_2. In other words, $\mathrm{I}(G_1;G_2|T) \geq 0$ even though $\mathrm{I}(G_1;G_2) = 0$. Furthermore, conditional mutual information provides an extension to more than two variables, a property we will take advantage of later.

Kullback-Leibler divergence

Kullback-Leibler divergence, D_{kl}, (also called relative entropy) is a quantification of the difference between two probability distributions. The D_{kl} between distributions p and q using the same alphabet $\mathcal{A}$ is the extra information needed to encode a set of data distributed according to p using q. It is defined as

$$\mathrm{D}_{kl}(p||q) := \sum_{a\in\mathcal{A}} p(a)\log_2\frac{p(a)}{q(a)}$$

$\mathrm{D}_{kl}(p||q)$ is always positive, and zero if and only if $p = q$. It is a critical component of information theory and is used (in addition to the highly related cross-entropy) extensively in machine learning when the goal is to approximate an unknown probability distribution. We include it here because examining the equivalency below can provide intuition not only about D_{kl}, but also mutual information. An alternate definition for mutual information is

$$\mathrm{I}(A;B) = \mathrm{D}_{kl}(p(a,b)||p(a)p(b))$$

In other words, the mutual information between A and B is the information lost by assuming that A and B are distributed independently when, in fact, they are not.

Equivalencies

We note here a series of useful equivalencies. Throughout the rest of this pub, we will use G to refer to genes and T to refer to traits or phenotypes.

$$\begin{aligned}
&\mathrm{I}(G;T) = \mathrm{H}(G) + \mathrm{H}(T) - \mathrm{H}(G,T) \\
&\mathrm{I}(G;T) = \mathrm{I}(T;G) \\
&\mathrm{I}(G;T) = \mathrm{H}(G) - \mathrm{H}(G|T) \\
&\mathrm{I}(G;T) = \mathrm{H}(T) - \mathrm{H}(T|G)
\end{aligned}$$

Extension to multiple genes and multiple phenotypes

Thus far we have mostly discussed individual random variables (e.g., single genes or phenotypes), but we can extend entropy, mutual information, and Kullback-Leibler divergence to cover the joint distribution of many variables, like a set of genetic loci or phenotypes. This results from the chain rule for probability and is most readily seen for entropy, where we have already defined joint and conditional entropy.

Chain rule for entropy

The joint entropy of A and B can be written as

$$\mathrm{H}(A,B) = \mathrm{H}(A) + \mathrm{H}(B|A)$$

Or, the joint entropy of A and B is the entropy of A plus the residual entropy in B if you know A. Repeated application of this method provides

$$\begin{gathered}
\mathrm{H}(A,B,C) = \mathrm{H}(A) + \mathrm{H}(B|A) + \mathrm{H}(C|B,A) \\
\vdots \\
\mathrm{H}(A_1, A_2 \ldots, A_n) = \sum_{i=1}^{n} \mathrm{H}(A_i|A_{i-1}, \ldots, A_1)
\end{gathered}$$

In other words, the joint entropy of a set of variables is the sum of their conditional entropies. For A, B, and C, or any other set of variables that are independent, their joint entropy is equal to the sum of their individual entropies. Or,

$$\mathrm{H}(A, B, C) = \mathrm{H}(A) + \mathrm{H}(B) + \mathrm{H}(C)$$

if A, B, and C are independent.

Chain rule for mutual information

We can apply a similar chain rule for mutual information, letting us extend to multiple random variables. We will not expand on this here, but, essentially, the variable expansion done previously to define conditional mutual information (jump to that equation) can be repeatedly applied to show that

$$\mathrm{I}(A_1, A_2 \ldots, A_n; B) = \sum_{i=1}^{n} \mathrm{I}(A_i; B | A_{i-1}, \ldots, A_1)$$

Essentially, the mutual information between a set of variables and another set of variables is the sum of the conditional mutual information values.

Given the ability to extend these measures to an arbitrary number of variables, we will indicate sets of variables with a sub bar. For example, we will denote sets of genes, phenotypes (or traits), and environments as $\underline{G}$, $\underline{T}$, and $\underline{E}$, respectively.

Applying information theory to genetics

Having established some of the fundamental measures in information theory and examples of their application, we now expand on these definitions and apply them to broader genetic questions. Where appropriate, we compare the information theory-based assessments with classical statistical genetic measures.

Polyphenotypic analysis

Genetic analysis has most often focused on individual phenotypes, e.g., "How tall are the members of a population?", or, "Do cells pause at a particular stage of the cell cycle?" But considering multiple phenotypes simultaneously may provide more insight into overall organismal features than focusing on any one phenotype. For example, an organism's height is likely linked to other organismal features (e.g., mass and metabolic rate) both causally and otherwise, so studying both height and metabolic rate together may enable more accurate predictions than studying height alone. However, the quantitative genetic infrastructure for simultaneous analysis of multiple phenotypes is poorly developed.

In a companion pub **[20]**, we argue that examining multiple phenotypes simultaneously can provide better insight into the nature of individual phenotypes. Across a population, phenotypes are often correlated. That correlation could result from shared, causal, genetic variation, or from non-causal correlation like genetic drift or migration. We've shown that incorporating the correlational relationships between phenotypes into predictive models can increase prediction accuracy. We further showed empirically that increasing pleiotropy among a fixed set of genes ($\underline{G}$) and phenotypes ($\underline{T}$) decreases the joint phenotypic entropy. If we measure the total phenotypic entropy as $\mathrm{H}(\underline{T})$, then the joint entropy must be less than or equal to the maximum entropy

$$\mathrm{H}(\underline{T}) \leq \sum_{i=1}^{n} \mathrm{H}(T_i)$$

with equality if, and only if, all phenotypes are independent of one another. Thus, the difference between the maximal phenotypic entropy and the total joint phenotypic entropy is the reduction in uncertainty caused by correlations (additive or otherwise) across phenotypes. In other words, we can quantify the amount of phenotype-phenotype structure by estimating the difference between the joint entropy and the maximal entropy. Importantly, this quantification provides examination of the relatedness (or lack thereof) among phenotypes without genetic or environmental information. Phenotypes with maximal entropy share no common cause or non-causal drivers of correlation. Thus, absent environmental variation or phenotypic correlations that are created by population structure, pairs of phenotypes with less than maximum entropy share a cause and those causes are, to some degree, epistatic.

Examination of many phenotypes likely provides information about any one phenotype

Given dependence among phenotypes, examining one phenotype should provide information about other phenotypes. In other words, conditioning the entropy of one set of phenotypes, $\underline{T}$, on another phenotype, T_i, will reduce the entropy (except in the case of independence).

Theorem:

$$\mathrm{H}(\underline{T}|T_i) \leq \mathrm{H}(\underline{T})$$

Proof:

$$\begin{aligned}
&\mathrm{I}(\underline{T};T_i) \geq 0 \\
&\mathrm{H}(\underline{T}) - \mathrm{H}(\underline{T}|T_i) \geq 0 \\
&\mathrm{H}(\underline{T}) \geq \mathrm{H}(\underline{T}|T_i) \\
&\mathrm{H}(\underline{T}|T_i) \leq \mathrm{H}(\underline{T})
\end{aligned}$$

This shows that, given some correlated structure among traits, examining many phenotypes will be useful in predicting any one phenotype; something we have empirically demonstrated in our companion pub **[20]**. Furthermore, in the same pub we show that examining increasing numbers of phenotypes doesn't reduce the amount of information about any one phenotype. However, we often estimate information theoretic values using numerical methods and, as a result, there is a limit to the number of phenotypes it is practical to examine.

Pleiotropy decreases total trait entropy

Pleiotropy is the observation that allelic state at any one genomic location impacts multiple phenotypes. Intuitively, for any fixed set of phenotypes and genes impacting those phenotypes, increasing pleiotropy will increase co-variation among phenotypes and thus decrease the total trait entropy. For traits T_1 and T_2 and gene G, we can define the pleiotropy as

$$Pleio(T_1, T_2, G) = \mathrm{I}(T_1; T_2) - \mathrm{I}(T_1; T_2|G)$$

This is the amount of information shared between T_1 and T_2 that can be accounted for if G is known. This is an extension of mutual information to multiple variables, known as interaction information. Unlike mutual information, interaction information can be negative. However, if T_1, T_2, and G form a Markov chain such that T_1 and T_2 are independent, conditional on G, then $\mathrm{I}(T_1, T_2|G) = 0$ and this reduces to $\mathrm{I}(T_1, T_2)$. With this definition of pleiotropy, we can show that the presence of pleiotropy will decrease the joint phenotypic entropy.

Theorem:

If T_1, T_2, and G form a Markov chain such that T_1 and T_2 are independent conditional on G, then increasing pleiotropy will lead to decreased joint trait entropy.

Proof:

$$\begin{aligned}
&Pleio(T_1, T_2, G) > 0 \\
&\mathrm{I}(T_1; T_2) - \mathrm{I}(T_1; T_2|G) > 0 \\
&\mathrm{I}(T_1; T_2) > 0 \\
&\mathrm{H}(T_1) + \mathrm{H}(T_2) - \mathrm{H}(T_1, T_2) > 0 \\
&\mathrm{H}(T_1) + \mathrm{H}(T_2) > \mathrm{H}(T_1, T_2)
\end{aligned}$$

where $\mathrm{H}(T_1) + \mathrm{H}(T_2)$ is the maximum possible entropy if T_1 and T_2 are totally independent and $\mathrm{H}(T_1, T_2)$ is the joint entropy of T_1 and T_2.

In this section, we've shown several ways in which we can apply information theory to the analysis of multiple phenotypes. First, we showed that the deviation between the maximal phenotypic entropy and the joint phenotypic entropy provides a quantification of the relational structure of a set of phenotypes, which may result from shared causes. Importantly, we can use this to show that some phenotypes are unrelated from others, a situation that would only result if there was no shared causation among those phenotypes. Second, we show that increasing the number of phenotypes in an analysis should increase our understanding of other phenotypes. And finally, we provide a mathematical definition of pleiotropy and show that increasing pleiotropy should, in some circumstances, decrease overall phenotypic entropy. While also demonstrating these findings empirically in a companion pub **[20]**, these formalisms provide certain guarantees about such analyses.

Key takeaways

- We provide formalisms for the analysis of cohorts of phenotypes ("polyphenotypes") using information theory.
- Analysis of individual phenotypes will benefit from examining a polyphenotype.
- Polyphenotypic analysis does not require genetic or other causal information.
- We can identify sets of phenotypes that are causally independent.

What's next?

We've presented a few examples of information theory applied to genetic questions. We view this as a work in progress and will, along with empirical and numerical studies in other pubs, expand these ideas into other areas of genetics and genetic analysis as our work progresses.

References

[1] Hartwell LH, Culotti J, Reid B. (1970). Genetic Control of the Cell-Division Cycle in Yeast, I. Detection of Mutants. https://doi.org/10.1073/pnas.66.2.352

[2] Reid BJ, Culotti JG, Nash RS, Pringle JR. (2015). Forty-five years of cell-cycle genetics. https://doi.org/10.1091/mbc.e14-10-1484

[3] Cole JB, Dürr JW, Nicolazzi EL. (2021). Invited review: The future of selection decisions and breeding programs: What are we breeding for, and who decides? https://doi.org/10.3168/jds.2020-19777

4 Chakravarti A. (2011). Genomic contributions to Mendelian disease. https://doi.org/10.1101/gr.123554.111

5 Mackay TFC. (2013). Epistasis and quantitative traits: using model organisms to study gene–gene interactions. https://doi.org/10.1038/nrg3627

6 Hill WG, Mackay TFC. (2004). D. S. Falconer and Introduction to Quantitative Genetics. https://doi.org/10.1093/genetics/167.4.1529

7 Visscher PM, Goddard ME. (2019). From R.A. Fisher's 1918 Paper to GWAS a Century Later. https://doi.org/10.1534/genetics.118.301594

8 Galton F. (1886). Regression Towards Mediocrity in Hereditary Stature. https://doi.org/10.2307/2841583

9 Sturtevant AH. (2001). A History of Genetics. https://www.google.com/books/edition/A_History_of_Genetics/wDIisw1ZqAMC

10 Johannsen W. (1909). Elemente der exakten Erblichkeitslehre. https://books.google.com/books?id=SkdEAQAAMAAJ

11 Punnett RC. (1915). Mimicry in Butterflies. https://books.google.com/books?id=XBIxoBBJoVEC

12 Committee RS (Great BE. (1910). Reports to the Evolution Committee of the Royal Society: Reports I-V. 1902–09. https://books.google.com/books?id=anFJb0zdq5kC

13 Fisher RA. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. https://doi.org/10.1017/s0080456800012163

14 Bateson W, Mendel G. (1909). Mendel's Principles of Heredity. https://books.google.com/books?id=8U7WwbwyB8oC

15 Game JC, Mortimer RK. (1974). A genetic study of X-ray sensitive mutants in yeast. https://doi.org/10.1016/0027-5107(74)90176-6

16 Phillips PC. (2008). Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. https://doi.org/10.1038/nrg2452

17 Huang W, Mackay TFC. (2016). The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. https://doi.org/10.1371/journal.pgen.1006421

18 Shannon CE. (1948). A Mathematical Theory of Communication. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

19 Avasthi P, Mets DG, York R. (2024). Harnessing genotype-phenotype nonlinearity to accelerate biological prediction. https://doi.org/10.57844/ARCADIA-5953-995F