Computational framework for statistical epistasis supports XOR penetrance function in a living system

Sandra Batista[1◕*], Vered Senderovich Madar[2◕], Philip J. Freda[1◕], Priyanka Bhandary[1], Attri Ghosh[1], Apurva S. Chitre[3], Abraham A. Palmer[3,4‡], Jason H. Moore[1‡*],

1 Department of Computational Biomedicine, Cedars-Sinai Medical Center, 700 N. San Vicente Blvd., Pacific Design Center, Suite G540, West Hollywood, CA, 90069, U.S.A.
2 Independent Researcher, Chapel Hill, NC, U.S.A
3 Department of Psychiatry, University of California San Diego, 9500 Gilman Dr., Mail Code: 0667, La Jolla, CA, 92093-0667, U.S.A.
4 Institute for Genomic Medicine, University of California San Diego, 9500 Gilman Drive, Mail Code: 0667, La Jolla, CA, 92093-0667, U.S.A.

◕These authors contributed equally to this work.
‡These authors also contributed equally to this work.
* Sandra.Batista@cshs.org, Jason.Moore@csmc.edu

Abstract

Statistical epistasis has been studied extensively because of its potential to provide evidence for genetic interactions for phenotypes, but there have been methodological limitations to its exhaustive, widespread application. We present new algorithms for the interaction coefficients for standard regression models for epistasis that permit many varied encodings for the interaction terms for loci and efficient memory usage. The algorithms are given for two-way and three-way epistasis and may be generalized to higher order epistasis. Statistical tests for the interaction coefficients are also provided. We also present a new efficient matrix based algorithm for permutation testing for two-way epistasis. We offer a proof and experimental evidence that methods that look for epistasis only at loci that have main effects may not be justified. Given the computational efficiency of the algorithm, we applied the method to a rat data set and mouse data set, with at least 10000 loci and 1000 samples each, using the standard Cartesian encoding and the XOR penetrance function for the interactions, to test for evidence of statistical epistasis for the phenotype of body mass index. This study revealed that the XOR penetrance function found greater evidence for statistical epistasis in many more pairs of loci in both data sets and in the rat data set, those pairs of loci found using the XOR penetrance function are enriched for biologically relevant pathways.

Author summary

Epistasis, the interaction between two or more genes, is likely integral to the study of genetics and present throughout nature. Yet, it is seldom fully explored as most approaches primarily focus on single-locus effects (such as GWAS), partly because analyzing all pairwise and higher-order interactions requires significant computational resources. Many current methods for epistasis detection only consider a Cartesian encoding for interaction terms. This is likely limiting as epistatic interactions can evolve

to produce varied relationships between genes, some non-linear. In this work we describe computationally efficient algorithms for the detection of statistical epistasis that allow for varied interaction encodings for modeling epistasis. Our methodology efficiently detects pairwise and three-way epistatic interactions in two closely related species (rat and mouse) under both Cartesian and XOR interaction encodings. Our results in both species show that many biologically relevant epistatic relationships would have been undetected if only one interaction encoding was applied providing evidence that more varied models for interaction may need to be applied to describe epistasis that occurs in living systems.

Introduction

Epistasis is challenging to detect yet likely widespread and integral in biology. Evidence for epistasis has been discovered in a host of biological systems and phenotypes including mandible size in mice [1], cardiovascular disease susceptibility [2], coronary artery restenosis [3], cystic fibrosis [4, 5], and sporadic breast cancer [6] in humans, and most recently and robustly in two studies investigating non-additive genetic effects in yeast [7, 8]. The yeast studies have collectively identified thousands of epistatic two-way and three-way interactions that vary across lineages and growth conditions. Additionally, they have identified large epistatic hubs that are involved in most interactions detected. In the most recent study in yeast, non-additive effects accounted for one-third of the broad-sense heritability [8]. These studies provide strong evidence that epistasis accounts for a large portion of the non-additive genetic variation in natural systems.

Since loci may contribute to phenotypes through nonlinear interactions, epistasis may account for genetic variation not explained by single-locus approaches. Indeed, it has been shown that the main effect of one SNP can significantly change when the allele frequencies in a second SNP are altered [9]. Examples of where epistasis may have a crucial role include biomolecular interactions in gene regulation, signal transduction, and biochemical networks [10–12]. Thus, many phenotypes can be viewed as the result of vast interconnected biological networks and systems [13–15]. These biological networks likely arise to form compensatory networks that aid in buffering against genetic and environmental change (i.e., canalization) [16–19]. It is likely that, at the core of these networks, interactions among polymorphisms from multiple pathways exist and are integral to organismal development, homeostasis, and survival [17].

Robust methodologies aimed at detecting and describing statistical epistasis are required to investigate genotype-phenotype associations and disease susceptibility [11]. Since experiments that could biologically validate detected statistical epistatic interactions are rare [13–15], the development of these methodologies will assist in initiating further scientific exploration. In this work, we examine how to efficiently compute linear regression models for epistasis that permit varied encodings of the interactions of loci and provide statistical evidence for epistasis. Recent computational and theoretical work has presented a new way to calculate each of the coefficients of a linear regression model [20]. In this work, we aim to demonstrate the usefulness and practical significance of the closed forms in general and more specifically in genetic studies and in particular, epistasis. We present algorithms for providing statistical evidence for two-way and three-way epistasis using standard models for epistasis using these closed forms that may be used efficiently on subsets of loci or genome-wide, that are entirely parallelizable, provide statistical tests, and permit flexibility in encoding the interactions of loci. Many methods for interaction terms only consider the Cartesian encoding that essentially multiplies the genotypes at two loci together. In our method, any function may be applied to interacting genotypes and we demonstrate this using the exclusive or (XOR) encoding penetrance function. We also discuss how these algorithms

may be generalized for higher order interactions. Permitting many types of encodings for interactions allows for nonlinearity to be included in the traditional statistical models for epistasis and for biologists to use a variety of penetrance functions to model epistasis.

As specific examples, we apply our algorithms to detect statistical evidence for two-way and three-way interactions using the Cartesian and XOR penetrance functions on the phenotype of body mass index (BMI) in real data sets from rats (*Rattus norvegicus*) from a GWAS study [21–23] investigating obesity-related traits and from mice (*Mus musculus*) from Wellcome Trust [24, 25]. For both data sets, we use approximately 10,000 SNPs for two-way epistasis detection using both penetrance functions. The implementations of these algorithms given in Python are offered via GitHub.

In the following sections we survey computational challenges for methods providing statistical evidence for epistasis and penetrance functions. In our methods section we review a regression model for epistasis, provide statistical tests for the interactions and present our algorithms. We also discuss consequential concerns detecting higher order epistasis. We then present results of applying our algorithms on the rat and mouse data sets with exploratory analysis providing statistical support for non-additive genetic effects for BMI in two model systems.

Computational Challenges in Detecting Epistasis

Methods to explore genotype-phenotype associations and detect epistatic interactions are often computationally intensive and may completely ignore non-additive effects. Single-locus analyses like genome-wide associations studies (GWAS) can detect strong main effects, but face difficulties when applied to combinations of variables for many reasons [14, 15]. The first is that multi-locus genotype (MLG) combinations have smaller representative samples compared to the original data set. Second, in most approaches that attempt to model interactions using linear models, interactions are only considered when significant main effects of variables are identified [14, 15]. Although it is tempting to expect loci with significant main effects to also be involved in interactions, there is no statistical justification for this. Third, linear models have greater power to detect main effects compared to interactions [26–28]. Fourth, many linear model approaches model interactions using the Cartesian product for ease of computation whereas other models for interactions may also be plausible. We expand more on this in the following section. Finally, when considering higher order interactions, as the number of loci in k-wise combinations increases, the number of variables in the standard regression models increases exponentially and the total number of sets of loci of size k to consider of all n possible loci increases polynomially as more loci are considered. This exhaustive search space creates issues with computational tractability as investigating pairwise and higher-order interactions becomes extremely difficult to achieve efficiently. Given these challenges, some of which are inherent in exhaustively considering all possible subsets for k-wise interactions in a set of n loci, many methods apply various techniques to reduce this search space or to exploit parallelism in underlying matrix libraries for computational efficiency.

There are many techniques to reduce the search space for epistasis algorithms. One example, the multifactor dimensionality reduction (MDR) technique finds multilocus genotypes that have high or low association with disease and defines new variables that explain the relationship of both loci [6, 29, 30]. MDR can be combined with other machine learning methods and has been extended to handle population structure [31]. In an approach that bypasses the conventional search space of epistasis algorithms, Crawford et al. developed the "MArginal ePIstasis Test" (MAPIT) method that measures the marginal epistatic effect of a single loci against all other loci all at

once [32]. MAPIT can be used as a way to screen for all loci that have significant marginal epistastic effects for subsequent tests to see which pairs of loci may be involved in epistasis. More comprehensive recent surveys on epistasis are offered by Ogbunugafor and Scarpino on higher order epistasis [33], Niel et al. on statistical and computational challenges of varied approaches [34], and Russ et al. for performance comparison of many varied epistasis detection methods [35].

For exhaustive methods for statistical epistasis for discrete (case/control) phenotypes, a common approach to make the problem more computationally tractable is to use contingency tables for approximating pairwise epistasis. One of the most commonly used software packages for applying the standard pairwise epistasis regression models is PLINK [36]. In PLINK pairs of loci are scanned using an approximate method that evaluates the Z-scores for the odds ratio at the loci between cases and controls or BOOST that uses bitwise operations for calculating the likelihood ratio test [37]. It is then possible to apply the entire regression to the pairs of loci that pass the screening stage. Zhang et al. use minimum spanning trees to update contingency tables for test statistics for evaluating pairs of loci for epistasis in the TEAM algorithm [38]. These aforementioned approaches work for pairwise epistasis. Bayat et al. offer the BitEpi method for handling up to four-way epistasis using bit efficient counting for contingency tables, an entropy metric to calculate the interaction effect, and permutations for p-value calculations [39].

Another approach to ease the computational burden for exhaustive methods for statistical epistasis is to exploit the parallelism from hardware architectures or underlying matrix operations. Schupbach et al. present FastEpistasis as an extension to PLINK that for quantitative phenotypes calculates the interaction term for a pair of loci by using the QR decomposition to solve the ordinary least squares [40]. The efficiency from their method is from the parallelism of the calculations and the hardware architectures. Zhu and Fang offer the method MatrixEpistasis that evaluates the standard regression model for epistasis for the interaction term by computing the residuals of the linear regression model between the phenotype and loci and the residuals for the regression model between the interaction and loci [41]. Like our method some of the efficiency of MatrixEpistasis is achieved by only focusing on the interaction term and corresponding test statistic, but the efficiency of their method also relies on the efficiency of underlying matrix operations and accordingly the method can only handle Cartesian encodings of the interaction term. The work of Zhu and Fang also considered covariate adjustment, but that may be done before applying our algorithm as is essentially done in their implementation and commonly done in practice as well. Unlike these other methods, our method permits many different encodings for the interaction, thus potentially introducing some nonlinearity into the model and explicitly deals with higher order epistasis subsets. The method we present is the result of new computational and algorithmic insight for solving the ordinary least squares that expresses the estimates in closed forms. As a result our method is also very efficient in terms of memory and entirely parallelizable for testing each subset of loci for epistasis.

The XOR Penetrance Function

It is not fully understood how epistatic interactions evolve, are maintained, or are structured in biological systems considering that methodologies specifically designed to systematically identify non-linear interactions are not common [13–15]. As an example of the possible complexity of epistasis, Li and Reich propose over a hundred full penetrance interaction models, some of which are not linearly separable [42]. The possible types of interactions become infinite when we consider incomplete penetrance. To test our methodology's capability of supporting multiple interaction encodings when detecting epistasis, we applied the exclusive or (XOR) penetrance function encoding in

addition to Cartesian encoding (File S2). The XOR function is described by Li and Reich as model M170 [42]. We chose XOR because of its extreme difference compared to a Cartesian encoding. Under the full penetrance XOR model, the phenotype is entirely dependent on the multi-locus genotype. Thus, in it's purest form, XOR is not linearly separable or detectable using any single-locus analyses like GWAS. Due to these aspects, the XOR model is commonly considered to not be biologically plausible. Despite being an extreme example of epistasis, models that could exist in a biological system and are consistent with the XOR model have been postulated [43, 44]. If we detect statistically significant epistatic interactions that are consistent with the XOR model in a biological system, it may justify describing and investigating more complex genetic architectures in nature. Furthermore, XOR detection may justify validation experiments to detect biological epistasis following the XOR model in living systems. We demonstrate that the XOR encoding detects statistical epistatic interactions not identifiable by a Cartesian model.

Methods

Models for Epistasis

At the center of modeling epistasis is regression and we may now consider whether locus i and locus j have an epistatic interaction, $I_{i,j}$, affecting the phenotype, p, by considering their interaction effect term, β_3 :

$$p = \beta_0 + \beta_1 g_i + \beta_2 g_j + \beta_3 I_{i,j}$$

We may remove the intercept and simplify the computation of the model by subtracting off the means of the variables, i.e. $\tilde{g}_i = g_i - \bar{g}_i$, $\tilde{p} = p - \bar{p}$ and $\widetilde{I_{i,j}} = I_{i,j} - \bar{I}_{i,j}$:

$$\tilde{p} = \beta_1 \tilde{g}_i + \beta_2 \tilde{g}_j + \beta_3 \widetilde{I_{i,j}}$$

We are only concerned with β_3, the coefficient for the interaction term, $\widetilde{I_{i,j}}$, in the model because it represents the partial correlation between the phenotype and the interaction term given both genes or loci. It is important to note that there is flexibility in this model in how the interaction is encoded. Although it is common to multiply values of the genes or loci for each sample as is commonly done for the Cartesian encoding for ease of computation, many different encodings may be used as we demonstrate for the XOR penetrance function (File S2). While the Fisher t-statistics for hypothesis testing can be applied (e.g. [41]), a direct T test for a regression coefficient may also be applied instead.

The model for epistasis may be generalized to higher order interactions. For $k-$order interactions between a set of k genes or loci, g_i for $1 \leq i \leq k$, we must consider the interactions of all subsets of the genes or loci from the empty set to the entire set of all k in a manner reminiscent of applying the binomial theorem where the coefficient is the interaction rather than bionomial coefficient. For the epistatic interaction of a subset, we will use the variable I_s where s is the subset. An appropriate function, f, for the encoding, such as XOR or Cartesian, may be applied to the sets as well as the interaction term of the entire set of k loci. All variables may then also be centered by subtracting off the means. The model for the epistatic interactions affecting phenotype p may be expressed as follows:

$$\widetilde{p} = \left[\sum_{i=1}^{k-1} \sum_{s \subset \{g_1,\ldots,g_k\}, |s|=i} \beta_s f(\widetilde{X_s}) \right] + \beta_{\{g_1,\ldots,g_k\}} \widetilde{I_{\{g_1,\ldots,g_k\}}}$$

The model that we will consider for 3-way epistasis for loci g_2, g_2, g_3 is accordingly after centering the variables:

$$\tilde{p} = \beta_1 \tilde{g_1} + \beta_2 \tilde{g_2} + \beta_3 \tilde{g_3} + \beta_{\{1,2\}} \tilde{I}_{\{1,2\}} + \beta_{\{1,3\}} \tilde{I}_{\{1,3\}} + \beta_{\{2,3\}} \tilde{I}_{\{2,3\}} + \beta_{\{1,2,3\}} \tilde{I}_{\{1,2,3\}}$$

For detecting epistasis the interaction term of interest is $\beta_{\{1,2,3\}}$ for 3-way epistasis and $\beta_{\{g_1,\ldots,g_k\}}$ for k-way epistasis. It is important to note that for $k-$way epistasis the regression must include $O(2^k)$ variables for each set of k loci and that to check for all possible $k-$wise interactions between n loci, there are $O(n^k)$ sets that must be checked for the given model. For a given set of k loci, we explore how it is not necessary to find the interaction coefficients for all $2^k - 1$ variables and may focus only on the final coefficient of interaction along with more efficient and powerful statistical tests of interaction. We give algorithms explicitly for $k = 2, 3$ and consider how these may be generalized.

Estimation and Statistical tests for interaction

The conditional association between the phenotype, p, and the interaction term, I, while conditioning upon the main effects of the loci, $\mathbf{Z} = \{g_i, g_j\}$, is the epistasis interaction. This may be computed in multiple ways. The first way is to compute the partial correlation coefficient between p and I, as the dot product between two standardized residuals: the residuals for the regression model between the phenotype and the loci and the residuals for the regression model between the interaction and loci. This approach requires the computation of two vectors of residuals estimated by two linear regression models. The first linear regression model is between the phenotype and loci, $p \sim \beta_0 + \mathbf{Z}$ and the second model is between the interaction and loci, $I \sim \beta_0 + \mathbf{Z}$, as was done in the work of Zhu and Zhang [41]. The standardized residual of the linear regression model between the phenotype and loci, $p \sim \beta_0 + \mathbf{Z}$, is $resid_p = p - \hat{\beta}_0^* - \hat{\beta}_1^* g_i - \hat{\beta}_2^* g_j$ where the $\hat{\beta}^*$ are the estimates from the model. The standardized residual of the regression between the interaction and loci, $I \sim \beta_0 + \mathbf{Z}$, is $resid_I = I - \hat{\beta}_0^{**} - \hat{\beta}_1^{**} g_i - \hat{\beta}_2^{**} g_j$ where again the $\hat{\beta}^{**}$ are the estimates from the model. This means two separate regression models are applied as was done in the work of Zhu and Zhang [41]. The partial correlation coefficient between p and I is thus r :

$$r = \left\langle \frac{resid_p}{\sqrt{< resid_p, resid_p >}}, \frac{resid_I}{\sqrt{< resid_I, resid_I >}} \right\rangle$$

The second way is to compute a single regression coefficient, the corresponding regression coefficient, $\hat{\beta}_3$ in the model between the phenotype, loci, and interaction term: $p \sim \beta_0 + \beta_2 g_1 + \beta_2 g_2 + \beta_3 I$. Both measures are variants of each other, as a computed $r = 0$ will necessarily imply a computed $\hat{\beta}_3 = 0$ and vice versa. This is due to Yule's equivalence formula [45],

$$\hat{\beta}_3 = r \cdot \frac{\sqrt{< resid_p, resid_p >}}{\sqrt{< resid_I, resid_I >}} \quad \text{or the equivalent form} \quad \hat{\beta}_3 = \frac{< resid_p, resid_I >}{< resid_I, resid_I >}$$

We know now that $< resid_p, resid_I > = < \mathbf{p}, resid_I >$ and know a direct way to estimate $\hat{\beta}_3$ and each of the other coefficients that is the crux of our algorithm [20]. Consider the estimation of the model assuming all variables are mean centered:

$$\widetilde{\mathbf{p}} = \beta_1 \widetilde{\mathbf{g}}_i + \beta_2 \widetilde{\mathbf{g}}_j + \beta_3 \widetilde{\mathbf{I}}$$

We have by considering the residuals of each loci with the interaction term an expression for the interaction effect term, β_3:

$$\hat{\beta}_3 = \frac{\left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{I}} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle - \frac{\left\langle \widetilde{\mathbf{I}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}} \left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle}{\left\langle \widetilde{\mathbf{I}}, \widetilde{\mathbf{I}} \right\rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} - \frac{\left(\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{I}} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}} \rangle \right)^2}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}}}$$

In particular, we may write the vector of partial residuals of the interaction term with each loci, $resid_I$, directly as

$$resid_I = \widetilde{\mathbf{I}} - \frac{\left\langle \widetilde{\mathbf{I}}, \widetilde{\mathbf{g}_i} \right\rangle}{\langle \mathbf{g}_i, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} - \frac{\left\langle \widetilde{\mathbf{I}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}} \left(\widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right)$$

while in a similar way we may write the vector of partial residuals of the phenotype with each loci, $resid_p$, directly as

$$resid_p = \widetilde{\mathbf{p}} - \frac{\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{g}_i} \rangle}{\langle \mathbf{g}_i, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} - \frac{\left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}} \left(\widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right)$$

As a result we may rewrite the interaction effect term, $\hat{\beta}_3$, it may be expressed in terms of the ratio of two dot products of the dependent variable and the partial residuals and the independent variable I as follows:

$$\hat{\beta}_3 = \frac{\langle \widetilde{\mathbf{p}}, resid_I \rangle}{\left\langle \widetilde{\mathbf{I}}, resid_I \right\rangle}$$

As for hypotheses test, we test the null hypotheses

$$H_0: \text{ The conditional association between } I \text{ and } p = 0$$

against the alternative

$$H_1: \text{ The conditional association between } I \text{ and } p \neq 0$$

This calls for multiple testing correction to protect against the inflated Type I error and we use the Benjamini Hochberg false discovery rate (FDR) controlling procedure for this. The statistical test will be a T test with $n - v - 1$ as degrees of freedom where n is the number of samples and v is the number of parameters in the model.

Zhu and Zhang address the partial correlation between the interaction term I and the phenotype p, and compute it using the dot product of the residuals [41]. As for statistical test to test for null value for the association, the Fisher approximate test is then used [46, pp. 26]:

$$T_{Fisher} = \sqrt{n - v - 1} \frac{r}{\sqrt{1 - r^2}}$$

The T test statistic for testing the null hypothesis against the alternative will be in terms of successive residuals: the partial residual, $resid_p$ aforementioned, and the global residual, $resid = p - \hat{\beta}_0 - \hat{\beta}_1 g_i - \hat{\beta}_2 g_j - \hat{\beta}_3 I$, for $p \sim \beta_0 + \beta_2 g_1 + \beta_2 g_2 + \beta_3 I$ after

adding I to the model. We observe that we may calculate the mean sum of squares or the mean square error, MSE, using the global residual, that is,

$$MSE = \frac{< resid, resid >}{(n-v-1)} = \frac{< \widetilde{\mathbf{p}}, resid >}{(n-v-1)}$$

As a result we may calculate the test statistic as well:

$$T(\beta_3) = \hat{\beta}_3 \cdot \sqrt{\frac{< \widetilde{\mathbf{I}}, resid_I >}{MSE}}$$

$$T(\beta_3) = \sqrt{n-v-1} \frac{< \widetilde{\mathbf{p}}, resid_I >}{\sqrt{< \widetilde{\mathbf{I}}, resid_I > \cdot < \widetilde{\mathbf{p}}, resid >}}$$

Both of the T tests above require the computation of the $< \widetilde{\mathbf{p}}, resid_p >$ $(=< resid_p, resid_p >)$ or the MSE (for modeling p) or where it is possible to verify that

$$MSE \cdot (n-v-1) =< \widetilde{\mathbf{p}}, resid >=< \widetilde{\mathbf{p}}, resid_p > -\hat{\beta}_3 \cdot < \widetilde{\mathbf{p}}, \widetilde{\mathbf{I}} > .$$

Algorithms for Epistasis Detection

We may now give an algorithm for computing the interaction coefficient and test statistic for epistasis for two genes or loci, g_i and g_j. We may encode their interaction, $I_{i,j}$ using any type of epistasis that may be suspected and often by default the Cartesian encoding (or product) is used. We will use the centered variables for our algorithm where the means have been subtracted, i.e. $\tilde{g}_i = g_i - \bar{g}_i$, $\tilde{p} = p - \bar{p}$ and $\tilde{I_{i,j}} = I_{i,j} - \bar{I}_{i,j}$. The algorithm residualizes the first locus from the second and then each loci from their interaction in order. The algorithm does the same in residualizing each loci from the phenotype in order. Finally to calculate the residuals of the entire model we residualize the interaction from the phenotype. In the case of pairwise epistasis, the algorithm only requires the centered variables to calculate several vector dot products and vector additions (subtractions). In the exposition given below we calculate the direct t-test statistic for the regression coefficient, but it is straightforward to calculate any of the other recommended test statistics. Let the number of samples for each loci and the phenotype be m.

Algorithm 1 Interaction Coefficient for Pairwise Epistasis

Given $\tilde{g}_i, \tilde{g}_j, \tilde{I_{i,j}}, \tilde{p}, m$
$v = \tilde{I_{i,j}} - \frac{<\tilde{g}_i, \tilde{I_{i,j}}>}{<\tilde{g}_i, \tilde{g}_i>} \tilde{g}_i$
$resid = \tilde{p} - \frac{<\tilde{g}_i, \tilde{p}>}{<\tilde{g}_i, \tilde{g}_i>} \tilde{g}_i$
$q_2 = \tilde{g}_j - \frac{<\tilde{g}_i, \tilde{g}_j>}{<\tilde{g}_i, \tilde{g}_i>} \tilde{g}_i$
$v = v - \frac{<\tilde{I}, q_2>}{<q_2, q_2>} q_2$
$resid = resid - \frac{<\tilde{p}, q_2>}{<q_2, q_2>} q_2$
$\beta_3 = \frac{<v, \tilde{p}>}{<v, v>}$
$resid = resid - \beta_3 * v$
$t = \sqrt{(m-4) < v, v >} \beta_3$
$t = t / \sqrt{< resid, resid >}$
return β_3 as the interaction coefficient and t as the t-test statistic

The algorithm is linear in the number of samples m. However, if the aim is to test all possible pairwise epistasis for n loci, this will still be an $O(n^2m)$ algorithm. The flexibility here is to test any pair of loci for epistasis with a variety of test statistics and a variety of possible interaction encodings. There is also the practical concern regarding floating point representation in implementations if any of the vector norms or norms squared, such as the norm of the residuals (in line 10), are close to zero. In some implementations this could raise an error because of division by zero, for example, because of limitations in floating point representations. In this case we recommended that any such errors that arise, while not errors in the algorithm or implementation themselves, but merely limitations of floating point representations, be logged for further inspection as we have done in our experiments.

We will now consider the algorithm for 3-way epistasis for a set of loci. While technically the algorithm is still linear in the number of samples since $k = 3$ is a constant, we begin to see more of the effects on the complexity of the number of variables in the model. For the case of 3-way epistasis, there are only 7 variables, but more generally there are $O(2^k)$ variables in models for $k-$way epistasis and the algorithm is quadratic in the number of variables while linear in the number of samples, so the more general case has complexity $O(2^{2k}m)$ for a single set of k loci. The models themselves introduce a number of variables exponential in k and for all possible sets of k loci from n the complexity is $O(n^k 2^{2k} m)$, so there is a need to prune the space of loci to check and variables in the models.

However, first we present the algorithm for 3-way epistasis detection and for the sake of exposition mention how it can be generalized to $k-$wise epistasis with the caveats we have given for its time complexity. For the encoding of all interactions we assume that a Cartesian encoding (or product of the loci) are used. We assume that we have three loci, g_1, g_2, g_3 and the phenotype, p as well as their encodings and they are all centered (with their means subtracted from them). First we construct an 8 by m matrix Q that will be used to residualize each variable against each other in the order given. For ease of exposition, we will index the columns of Q for each variable starting at 1 where 1 will be the index for $\tilde{g_1}$, 7 for $\tilde{I_{\{1,2,3\}}}$and 8 of $\tilde{p}$. It is important to note that we are overwriting Q. This is important to mention since it clarifies not only the correctness of the algorithm, but highlights the space efficiency of these algorithms. Once the intermediate matrix for the 8 (or more generally 2^k) variables for the model is constructed for which only read only access of the entire data matrix is needed, no additional access to the data matrix is needed and only additional constant space for several dot products and to hold the interaction coefficient and test statistic are needed.

To generalize the 3-way epistasis algorithm to be a $k-$wise epistasis algorithm, we need to first enumerate the 2^k variables of the model and center them by subtracting off their means to construct $\mathbf{Q}$. Then it is necessary to residualize the variables by changing 8 to the value 2^k on line 3. To get the interaction coefficient for all k loci, $\beta_{\{1,\ldots,k\}}$, on line 7 check that $i == 2^k$ and $j = 2^k - 1$. To calculate the t-test statistic, on line 13 set $v = \mathbf{Q}[:, 2^k - 1]$ and on line 14 $resid = \mathbf{Q}[:, 2^k]$ and adjust the degrees of freedom on line 15. If we have a set of loci of interest, this algorithm and its generalization can calculate their interaction coefficient and test statistic efficiently. However, as we noted if we do not have a known set of loci a priori, then there is work to be done to prune what would be the exhaustive search space as we will consider in the next section.

The extension of the algorithms for permutation tests can be easily done by permutations of the dependent variable $\mathbf{p}$ while computing $resid_I$, $resid_p$ and in the case of a pairwise epistasis, $\hat{\beta}_3$ as

$$\hat{\beta}_3 = \frac{\langle \widetilde{\mathbf{p}}, resid_I \rangle}{\left\langle \widetilde{\mathbf{I}}, resid_I \right\rangle}$$

Algorithm 2 Interaction Coefficient for 3-way Epistasis

Given $\tilde{g_1}, \tilde{g_2}, \tilde{g_3}, \tilde{I_{\{1,2\}}}, \tilde{I_{\{1,3\}}}, \tilde{I_{\{2,3\}}}, \tilde{I_{\{1,2,3\}}}, \tilde{p}, m$
$\mathbf{Q} = \tilde{g_1}|\tilde{g_2}|\tilde{g_3}|\tilde{I_{\{1,2\}}}|\tilde{I_{\{1,3\}}}|\tilde{I_{\{2,3\}}}|\tilde{I_{\{1,2,3\}}}|\tilde{p}$
for $2 \leq i \leq 8$ **do**
    $x = \mathbf{Q}[:, i]$
    for $i - 1 \leq j \leq 1$ **do**
        $y = \mathbf{Q}[:, j]$
        if $i == 8$ and $j == 7$ **then**
            $\beta_{\{1,2,3\}} = \frac{<x,y>}{<y,y>}$
        end if
        $\mathbf{Q}[:, i] = \mathbf{Q}[:, i] - \frac{<x,y>}{<y,y>} y$
    end for
end for
$v = \mathbf{Q}[:, 7]$
$resid = \mathbf{Q}[:, 8]$
$t = \sqrt{m - 8}\beta_{\{1,2,3\}}\sqrt{< v, v >}$
$t = t/\sqrt{< resid, resid >}$
return $\beta_{\{1,2,3\}}$ as the interaction coefficient and t as the t-test statistic

$$resid_I = \widetilde{\mathbf{I}} - \frac{\left\langle \widetilde{\mathbf{I}}, \widetilde{\mathbf{g}_i} \right\rangle}{\langle \mathbf{g}_i, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} - \frac{\left\langle \widetilde{\mathbf{I}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}} \left(\widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right)$$

and

$$resid_p = \widetilde{\mathbf{p}} - \frac{\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{g}_i} \rangle}{\langle \mathbf{g}_i, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} - \frac{\left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}} \left(\widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right)$$

They may be used in order to obtain the test statistic

$$T(\beta_3) = \sqrt{n-4} \frac{< \widetilde{\mathbf{p}}, resid_I >}{\sqrt{< \widetilde{\mathbf{I}}, resid_I > \cdot < \widetilde{\mathbf{p}}, resid >}}$$

In the case of pairwise epistasis

$$\langle \widetilde{\mathbf{p}}, resid \rangle = \langle \widetilde{\mathbf{p}}, resid_p \rangle - \hat{\beta}_3 \cdot \left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{I}} \right\rangle = \langle \widetilde{\mathbf{p}}, resid_p \rangle - \frac{\langle \widetilde{\mathbf{p}}, resid_I \rangle}{\left\langle \widetilde{\mathbf{I}}, resid_I \right\rangle} \cdot \left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{I}} \right\rangle.$$

Note that $resid_p$ and the dot product of the permuted phenotype and interaction term needs to be updated for each permutation. The interaction term and its residual may be reused for all permutations. The extension of the pairwise epistasis algorithm with permutation testing follows in Algorithm 3. We present the algorithm using matrix calculations on the permuted phenotype matrix, P, an m by K matrix, where m is the number of samples and K is the number of permutations and each column of P is a permutation of the original phenotype vector, $\widetilde{\mathbf{p}}$. There are several advantages to this approach in that calculating P once for all permutation tests permits efficient parallelization of the permutation testing. The permuted phenotype matrix may be divided into submatrices of columns of permutations if need be especially for memory although calculation of p-value would be modified to be done across the divisions, i.e. return the number of tests at least as large original t-test in absolute value and calculate p-value after all computations are done. Second using matrix operations

permits implementations of the algorithm to exploit underlying parallelization and efficiency of matrix libraries such as BLAS and LAPACK used by NumPy in Python.

Algorithm 3 Permutation Test P-value for Pairwise Epistasis

Given $\widetilde{g_i}, \widetilde{g_j}, \widetilde{I_{i,j}}, \widetilde{\mathbf{p}}, m$, $\mathbf{P}$, the permuted phenotype matrix, and K the number of permutations
Compute $resid_I$, $resid_p$, $\hat{\beta}_3$ and the T statistic t, $dot_I =< v, v >$ and $q_2 = resid_{g_j}$ using Algorithm 1
$perm_{g_i} = \frac{P^T \widetilde{g_i}}{<\widetilde{g_i}, \widetilde{g_i}>}$
$resid = P - \widetilde{g_i} * perm_{g_i}^T$
$perm_{g_j} = \frac{P^T q_2}{<q_2, q_2>}$
$resid = resid - \widetilde{g_j} * perm_{g_j}^T$
$perm_I = \frac{P^T resid_I}{dot_I}$
$resid = resid - \widetilde{I_{i,j}} * perm_I^T$
$test_\beta = diag(P^T resid)$
$test_\beta = \sqrt{dot_I * test_\beta}$
$test_\beta = \sqrt{m-4} \frac{perm_I}{\sqrt{dot_I * test_\beta}}$
return $\frac{1}{K} \sum \{|test_\beta| \geq |t|\}$ as p-value

It should be noted that permutation tests offer a robust alternative statistical test which is preferred in case there is suspicion that the required linear regression assumptions of nonlinearity or nonnormality fail to hold. Permutation tests are also known to control better the family wise error rate in the scope of the specific model under testing. As each permutation test is performed under a specific epistasis model, when considered altogether, for all pairs, the control of the global inflation of the Type I error is not assured.

Screening for higher order epistasis

We consider whether there is any mathematical justification from the models for epistasis to prune the search space of higher order interactions from lower order interactions or vice versa. For example, a common heuristic is to search the main effects for the pairwise epistasis and pairwise epistasis for 3-way epistasis. A variant of an approach that uses stages like this is used for example by Laurie et al [47].

Without loss of generality we are considering $k+1$ loci, $g_1, \ldots g_{k+1}$ and their interactions are all encoded using Cartesian (product) encoding. If such justification exists we may be able to show claims such that if there is a significant $k+1$-way interaction among $k+1$ loci, then all k interactions in the set of $k+1$ loci are significant. The contrapositive of this claim is more practically useful: If any k-wise interaction is not significant from the set of loci, then $k+1$-wise interaction is also not significant. The converse (i.e. if for a set of $k+1$ loci, all $k-$wise epistasis exist, then the loci have $k+1-$wise epistasis) is less plausible still. There is empirical evidence in the rat data set in this study (and many others) against both claims in considering only main effects and pairwise epistasis. Namely, there are main effects loci that are not in pairwise epistasis with each other; there are main effect loci in pairwise epistasis with a loci that is not a main effect and there exists pairwise epistasis between loci that are not main effects.

Nevertheless we examine why these claims do not hold for the models used in epistasis. To do so we consider two loci and their staged models for main effects and pairwise epistasis. Here we express the models simultaneously assuming the variables

are centered. First we consider main effects:

$$\widetilde{p} = \beta_{1i}\widetilde{g_i} \qquad \text{or} \qquad \widetilde{p} = \beta_{1j}\widetilde{g_j}$$

and pairwise epistasis

$$\widetilde{p} = \beta_{2i}\widetilde{g_i} + \beta_{2j}\widetilde{g_j} + \beta_{ij}\widetilde{I_{i,j}}$$

In terms of the coefficients in the models the first claim is that at least one loci not being a main effect implies no pairwise epistasis: if $\hat{\beta}_{1i} = 0$ or $\hat{\beta}_{1j} = 0$, then $\hat{\beta}_{ij} = 0$. The concern here is really the inference needed regarding the coefficients of the pairwise model given knowledge of the coefficients in the main effects model. While having evidence $\hat{\beta}_{ij} = 0$ does not give us implications about $\hat{\beta}_{2j}$ or β_{2i}, knowing that either is 0 would give an implication regarding $\hat{\beta}_{ij} = 0$. In general, without loss of generality, our closed forms (cite OLS paper) give extensions to the formula of Yule [45],

$$\hat{\beta}_{1i} = \frac{< \widetilde{\mathbf{p}}, \widetilde{\mathbf{g}_i} >}{< \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} >} \qquad \text{and} \qquad \hat{\beta}_{1j} = \frac{< \widetilde{\mathbf{p}}, \widetilde{\mathbf{g}_j} >}{< \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} >}$$

and for the pairwise epistasis model

$$\hat{\beta}_{ij} = \frac{\left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{I}_{ij}} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}_{ij}} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle - \frac{\left\langle \widetilde{\mathbf{I}_{ij}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}} \left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle}{\left\langle \widetilde{\mathbf{I}_{ij}}, \widetilde{\mathbf{I}_{ij}} \right\rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}_{ij}} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} - \frac{\left(\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{I}_{ij}} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}_{ij}} \rangle \right)^2}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}}}$$

In the numerator for $\hat{\beta}_{ij}$, we note there are components that may be expressed in terms of the main effect coefficients, $\hat{\beta}_{1i}$ and $\hat{\beta}_{1j}$. These will be the following two components:

$\left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle = \langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle \hat{\beta}_{1j} - \langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle \hat{\beta}_{1i}$ and

$\left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{I}_{ij}} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}_{ij}} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle = \left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{I}_{ij}} \right\rangle - \left\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}_{ij}} \right\rangle \beta_{1i}$

As a result, in terms of the main effect coefficients, $\hat{\beta}_{1i}$ and $\hat{\beta}_{1j}$, we may write the pairwise epistasis interaction term as

$$\hat{\beta}_{ij} = \frac{\left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{I}_{ij}} \right\rangle - \left\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}_{ij}} \right\rangle \beta_{1i} - \frac{\left\langle \widetilde{\mathbf{I}_{ij}}, \widetilde{\mathbf{g}_j} - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \widetilde{\mathbf{g}_i} \right\rangle}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}} \left(\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle \hat{\beta}_{1j} - \langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle \hat{\beta}_{1i} \right)}{\left\langle \widetilde{\mathbf{I}_{ij}}, \widetilde{\mathbf{I}_{ij}} \right\rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}_{ij}} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} - \frac{\left(\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{I}_{ij}} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle} \langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{I}_{ij}} \rangle \right)^2}{\langle \widetilde{\mathbf{g}_j}, \widetilde{\mathbf{g}_j} \rangle - \frac{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_j} \rangle^2}{\langle \widetilde{\mathbf{g}_i}, \widetilde{\mathbf{g}_i} \rangle}}}$$

From this we observe that the numerator of $\hat{\beta}_{ij}$ is determined upon the numerical value of the product $\left\langle \widetilde{\mathbf{p}}, \widetilde{\mathbf{I}_{ij}} \right\rangle$ and zero values for $\hat{\beta}_{1i}$ and $\hat{\beta}_{1j}$ do not imply $\hat{\beta}_{ij} = 0$. So insignificant interaction term, or value of $\hat{\beta}_{ij} = 0$ does not imply $\hat{\beta}_{1i} = 0$ or $\hat{\beta}_{1j} = 0$ and on the other hand $\hat{\beta}_{1i} = 0$ or $\hat{\beta}_{1j} = 0$ does not imply $\hat{\beta}_{2i} = 0$, $\hat{\beta}_{2j} = 0$ or $\hat{\beta}_{ij} = 0$. This may be considered a warning that we may not be justified only searching main effects for pairwise epistasis. Rather a biological reason for pruning loci, such as pruning of SNPs in LD, for epistasis may be more justified. For example, in the rat data set we applied the epistasis algorithms to loci screened from a previous GWAS study because of biological interest in those loci for the phenotype of BMI [21–23]. In a recent study in yeast, Ang et al. observed that epistasis most often occurred between loci with minor allele frequency between five and ten percent [48], so pruning loci for epistasis based on minor allele frequency may be another strategy for pruning the search space for epistasis detection for biological reasons.

Application to Body Mass Index Data

We have applied our algorithms to a dataset of an outbred, related rat (*Rattus norvegicus*) population of males and females derived from eight inbred founders (Heterogenous Stock, [49]) in order to detect possible epistatic interactions associated with Body Mass Index (BMI). We also used the rat data to showcase the extension of our algorithm to higher-order epistasis by investigating three-way interactions between putative QTLs identified via GWAS ([21–23]). Additionally, we have applied our algorithm to a dataset of an inbred population of mice (*Mus musculus*) derived from 17 mouse strains ([24, 25] to compare epistatic interactions, both Cartesian and XOR, in a closely related species to rat.

Detecting Epistasis

The GWAS dataset of body mass index (BMI) residuals (corrected for sex and location) in 5,566 rats (*Rattus norvegicus*) containing approximately 129,000 autosomal SNPs was used for our rat analysis ([21–23]. We selected the top 10,000 SNPs (SNPs with the 10,000 lowest p-values) from the original GWAS for further analysis as related work has illustrated positive correlations between main effects of a loci and instances of detected epistasis ([7, 8]. For mice, we used the genotype/phenotype information from the Wellcome Trust Mouse Genomes Project ([24, 25]) found in the BGLR package ([50] in R ([51]). We extracted the genotype and BMI phenotype data available for 10,347 SNPs for 1,814 mice.

To account for the relatedness of the samples and population structures, the genetic relatedness matrices (GRM) were calculated for the rat data using the method of Yang et al. using the GCTA software tool [52] and the mouse data using the method of Sul et al [53]. We used the GRMs to calculate the variance component analysis for the phenotype of BMI for each data set using the methods of Joo et al. [54] and Kang et al. ([55, 56]) as implemented in the mmer function with method EMMA of the sommer package version 3.2 in R [57]. From the variance component analyses we obtained the inverses of the covariance matrices for both data sets as returned by the sommer package and then used the the sqrtm function from the expm package version 0.999-7 in R to obtain their square roots [58]. These inverse square root covariance matrices were then used to correct for population structure and essentially solve a generalized linear model by performing a weighted least squares by multiplying each variable in the model by the half inverse matrix before applying our algorithms as a pre-processing step. This "mixed model trick" is summarized by Suh et al [53].

Both Algorithms 1 and 2 were implemented in Python 3.9 for our experiments. To calculate the p-values for the two-sided T tests for the test statistics returned from our algorithms, the t.sf function from the scipy.stats package was used [59]. For two-way epistasis, Algorithm 1 was applied to all possible pairs of 10,000 SNPs in the rat data set and all possible pairs of 10,347 SNPs in the mouse data set. This was done as two separate experiments for each data set, once for the XOR and another for the Cartesian product. In addition for the XOR penetrance function for all 10,000 SNPs for the rat data set, p-values were also calculated using the permutation testing algorithm with 1000 permutations. (This was computationally intensive since we calculate nearly fifty billion tests and was done in parallel.) To account for multiple testing, FDR as implemented in the fdrcorrection function in the statsmodels package in Python [60] using a p-value threshold of 0.05 was applied. For three-way epistasis for the rat data set, we applied Algorithm 2 to the 18 putative main effect QTLS from the rat GWAS study [21–23] to see if any were involved in significant three-way epistatic interactions.

Pruning Redundant Epistatic Events

For mice, we mapped all SNPs to their respective genomic location (assembly GRCm39) and retained those that mapped to the 19 autosomes in mice. This left us with 9,525 SNPs for further analyses. Although GWAS SNPs were LD-pruned, it is likely that many detected pairs are redundant in that one or both epistatic partners of a particular pair are in LD with others in close genomic proximity. We chose a conservative threshold of 10Mb upstream or downstream to prune redundant pairs in both species. All pruning steps were performed in R ([51]). Under both Cartesian and XOR encodings, to prune interchromosomal pairs, four conditions had to be met: 1.) locus one in pair one and locus one in pair two are on the same chromosome, 2.) locus two in pair one and locus two in pair two are on the same chromosome, 3.) the absolute value of the difference in chromosomal position (in bp) between locus one in pair one and locus one in pair two is less than 10Mb, and 4.) the absolute value of difference in chromosomal position between locus two in pair one and locus two in pair two is less than 10Mb. If all four conditions are true, the epistatic pair with the lower FDR-corrected p-value is retained while the other is omitted. We then checked for mirror redundancies where two pairs technically meet the above criteria, but the chromosomal combination is reversed. The pair that is identified first is retained while the other is omitted. To prune intrachromosomal pairs, we check if two pairs exist on the same chromosome and then use the same bp criteria as outlined above for interchromosomal pairs. We also take an additional pruning step for intrachromosomal pairs where the absolute value of the bp difference between locus one and locus two is less than 10Mb. This step removes close-acting cis-regulatory epistatic events as a byproduct of controlling for high LD.

There are nine total multi-locus genotypes (MLGs) possible under pair-wise epistasis of two biallelic SNPs (File S2). We limit our results to only those pairs where all nine MLGs occur. Additionally, we remove any significant epistatic pair, both in Cartesian and XOR, that do not have at least 10 or 5 observations of all nine MLGs in the datasets in rats and mice, respectively. Under Hardy-Weinberg assumptions, if the minor allele frequencies in two loci are 0.10, then we only expect to observe a genotype frequency for a double minor homozygote of 0.0001. We select the cutoff observations of 10 and 5 in rats and mice, respectively to ensure the combinatory genotype frequencies of the MLGs to be sufficiently above what would be expected if minor allele frequencies are 0.10 in both epistatic loci under Hardy-Weinberg assumptions.

Identifying QTL-Associated and non-QTL-Associated Epistatic Events and Epistatic Hubs

After our initial pruning steps, we determine if a locus was associated with a putative QTL for BMI ("BMI with tail" in rats and "BMI" in mice) from the original GWAS studies ([21–25]) to observe if most epistatic events occurs at or near loci with large main effects. To accomplish this, we use lists of the putative single locus QTL from the original GWAS studies and their respective genomic locations. We record if either locus 1 or locus 2 of a pair is within 10Mb upstream or downstream of a putative QTL to count the number of epistatic pairs with one putative GWAS QTL involved. We also note if both loci of a pair are associated with a GWAS QTL to record QTL to QTL epistatic interactions. If a locus is within 10Mb upstream or downstream of a GWAS QTL, its identification is replaced by the putative QTL's identification. We then count how many epistatic events involve each GWAS QTL. If non-QTL-associated loci are within 10Mb upstream or downstream of each other on the same chromosome, the average genomic location is calculated and all non-QTL epistatic loci's bp positions used to calculate the average are replaced by the average bp location. This allows us to identify non-QTL-associated epistatic loci. For our analyses, any locus with 10 or more

epistatic interactions are defined as an epistatic hub (GWAS QTLs included).

Under XOR, we also identify loci that are within 10Mb upstream or downstream of a GWAS QTL and replace their identifications with the respective GWAS QTL. Next, we determine which non-QTL-associated loci are within 10Mb upstream or downstream of the non-QTL loci determined under Cartesian encoding and replace their identification with the identification of the Cartesian locus to determine how many epistatic loci are detected under both encodings. However, there are loci that did not meet these criteria. For these, we apply the same procedure used to determine non-QTL-associated loci under Cartesian encoding of calculating the average genomic location among loci if on the same chromosome and within 10Mb upstream or downstream of each other. Thus, we identify loci specific to either encoding as well as overlapping loci and hubs shared between encodings.

Gene Set Enrichment and Kegg Pathway Analysis

To perform functional annotation for loci involved in epistatic events, we query protein-coding gene models, non-protein coding gene models, and pseudogene models from the Rat Genome Database (https://rgd.mcw.edu/) for rats and mice. We retrieve any model that is 1Mb upstream or downstream from each epistatic locus. We then convert the species-specific gene model symbols from the respective database to Entrez IDs. Any returned queries that do not have associated Entrez IDs are omitted. We use this list to identify enrichment of GO (gene ontology) terms (cellular component, molecular function, and biological process) using the BioConductor ([61]) package clusterProfiler ([62, 63]) in R with FDR correction (p-value cutoff= 0.01; q-value cutoff= 0.05). We also perform KEGG pathway enrichment analysis using clusterpPofiler (p-value cutoff= 0.05; q-value cutoff= 0.02). Dotplot figures of enriched ontological terms are made using the DOSE ([64]) package in R.

Identifying Three-Way Epistatic Interactions between Putative QTLs

To test the extension of our algorithm for identifying higher-order epistasis, we explore three-way epistatic interactions among the 18 putative QTL identified in the original rat GWAS ([21–23]) under both Cartesian and XOR interaction encodings. Triplets that have an associated experimental p-value < 0.05 are retained and we record occurrences of three-way epistasis for each GWAS QTL. In addition, the triplets between each encoding (Cartesian or XOR) are compared for overlap. Since all significant triplets in both experiments are unique in terms of genomic location and between putative QTL, no pruning or epistatic locus identifications are performed for this analysis.

Results

Most Epistatic Interactions and Hubs Occur at non-QTL-Associated Loci

For the Cartesian experiment in rats, our method detects 4,158 (86.8%) interchromosomal and 634 (13.2%) intrachromosomal significant pairs (4,792 total; File S1). After MLG pruning, this reduces to 3,109 (90.4%) interchromosomal and 329 (9.6%) intrachromosomal significant pairs (3,438 total; File S1). After redundancy pruning, this is further reduced to 175 interchromosomal pairs (96.2%) and seven intrachromosomal pairs (3.8%) (182 total) (Fig 1A; File S1). There are 182 Cartesian pairs after all pruning. Of these, there are 66 pairs (36.3%) containing one

QTL-associated locus and 9 (4.9%) QTL-QTL interactions (Fig 1C; File S1). However, most pairs are between two non-QTL-associated loci (107 pairs; 58.8%). There are a total of 91 unique Cartesian epistatic loci. Of these, 75 are non-QTL-associated (82.4%) (Fig 1E; File S1). Of the 18 putative GWAS QTL, chr7:8599340_A, chr18.32316331_A, and chr5:72916242_T are not involved in epistasis (File S1). It is important to note that chr18.32316331_A is physically close to another putative QTL, chr18.27348077_G (4,968,254 bp apart), and is not represented as loci we detect are physically closer to chr18.27348077_G.

Fig 1. Proportions of epistasis detected in rats. A and B: Interchromosomal (blue squares) vs intrachromosomal (light purple squares) pairs under Cartesian (A) and XOR (B) models. C and D: Number of epistatic pairs involving no (pink squares), one (green squares), or two (yellow squares) putative GWAS QTL under Cartesian (C) and XOR (D) models. E and F: Non-QTL-associated (gray squares) and QTL-associated (dark purple squares) epistatic loci under Cartesian (E) and XOR (F) models.

In mice, 81 epistatic pairs are detected under Cartesian encoding. However, 65 of these involve at least one locus on the X chromosome and are omitted as we are only investigating autosomal loci, leaving 16 pairs. None of these pairs are intrachromosomal (File S3). After MLG pruning eight pairs remain and after redundancy pruning only one pair remains between chr3.44666611, and chr12.7079769 (File S3). In rats, XOR encoding yields more significant pairs compared to Cartesian (31,182 vs. 4,792) (File S1). Of the 31,182 significant pairs, 27,774 are interchromosomal (89.1%) and 3,408 are intrachromosomal (10.9%; File S1). After MLG pruning, this is reduced to 11,016 (86.4%) interchromosomal pairs and 1733 (13.6%) intrachromosomal pairs (12,749 pairs total; File S1). After redundancy pruning, this is further reduced to 296 (94.9%) interchromosomal pairs and 16 (5.1%) intrachromosomal pairs (312 pairs total) (Fig 1B; File S1). Most pairs (181 (58.0%)) are between two non-QTL-associated loci while there are 125 (40.1%) pairs involving one QTL-associated locus and six (1.9%) QTL-QTL interactions (Fig 1D; File S1). Epistatic pairs involve 100 loci where 82 (82.0%) are non-QTL-associated (Fig 1F; File S1). All 18 putative GWAS QTL are represented under XOR encoding. In mice, under XOR, there are 468,596 significant pairs. After we map these loci to genomic locations and remove any pair involving an non-autosomal locus, 25,055 pairs remain with 23,688 (94.5%) interchromosomal and 1,367 (5.5%) intrachromosomal (File S3). After MLG pruning, this is reduced to 13,328 (95.0%) interchromosomal pairs and 707 (5.0%) intrachromosomal pairs (14,035 pairs total; File S3). Finally, after redundancy pruning, this is reduced to 341 (92.4%) interchromosomal pairs and 28 (7.6%) intrachromosomal pairs (369 pairs total) (Fig 2A; File S3). As in rats, and due to there only being 3 putative GWAS QTL for BMI in mice we were able to map to a genomic location, most pairs (351 (95.1%)) are between two non-QTL-associated loci while there are 18 (4.9%) pairs containing one QTL-associated locus (Fig 2B; File S1). There are no QTL-QTL interactions in mice. In mice, XOR yields 115 unique epistatic loci where 112 (97.4%) are non-QTL-associated.

It is important to note that a large amount of significant epistatic pairs involve non-autosomal or unmappable loci in mice. In future analyses, we aim to explore interactions that involve non-autosomal loci within the mouse cohort and in other species.

The rank order and number of interactions per GWAS QTL differ between experiments and across species. In rats, under Cartesian encoding, the largest QTL-associated hubs are chr1.281788173_G and chr5.107167969_G with 12 interactions each (Fig 3A; File S1). chr1.281788173_G is also the locus with the largest main effect signal in the rat GWAS (File S1). Under XOR encoding, there are four QTL-associated hubs identified (Fig 3B; File S1). As in Cartesian, chr1.281788173_G is the largest

Fig 2. Representations of epistasis in mice. A: Interchromosomal (blue squares) vs. intrachromosomal (light purple squares) pairs under XOR model. B: Number of pairs involving no (pink squares), one (green squares), or two (yellow squares) putative GWAS QTL under XOR model. C: Non-QTL-associated (gray squares) and QTL-associated (dark purple squares) epistatic loci under XOR model. D: Interaction plot of XOR epistasis across mouse autosomes. Autosome numbers are depicted as roman numerals and increase clockwise. Orange bars represent counts of epistatic instances per locus. E: Venn diagram of epistatic pairs under Cartesian (left circle) model, the intersect between models (center), and under XOR (right circle) model. F: Venn diagram of epistatic loci under Cartesian (left circle) model, the intersect between models (center), and under XOR (right circle) model. E and F: Color gradient illustrates low (white) to high (blue) occurrences.

QTL-associated hub under XOR with 34 interactions. This is followed by chr18:27348077_G with 25 interactions. XOR QTL-associated hubs are also identified at chr5:107167969_G, and chr8:103608382_G. Under Cartesian encoding, 10 of the putative GWAS QTLs are in QTL-QTL interactions (Fig 3C; File S1). Two putative GWAS QTLs are in three QTL-QTL interactions while four are in two QTL-QTL interactions. The remaining four are in one. The Cartesian QTL hubs, chr5.107167969_G and chr1.281788173_G are involved in QTL-QTL interactions, but there doesn't seem to be a clear relationship between number of epistatic interactions and number of QTL-QTL interactions. Under XOR encoding, seven putative GWAS QTL are in QTL-QTL interactions (Fig 3D; File S1). However, the rank orders and QTL representations are distinct between encodings. For example, under XOR, chr1.281788173_G has the largest occurrences of QTL-QTL interactions with five. Under XOR, there also is not a clear relationship between the number of QTL epistatic events a hub is involved in and the propensity of QTL-QTL interactions.

Fig 3. Bar graphs of occurrences of QTL-associated epistasis in rats. A and B: Bar graphs of instances of epistatic events for GWAS putative QTL under Cartesian (A, orange bars) and XOR (B, blue bars) models. C and D: Bar graphs of instances of QTL to QTL epistatic events for GWAS putative QTL under Cartesian (C, orange bars) and XOR (D, blue bars) models.

In mice under Cartesian encoding, QTL and QTL-QTL interactions do not occur as only one significant pair, involving two non-QTL-associated loci, is detected. Under XOR encoding, all three putative QTL are involved in a least one epistatic interaction, however none are hubs (Fig 4A; File S3). The largest QTL epistatic loci is chr2.68831331_G with eight interactions. No QTL-QTL interactions occur in mice under XOR encoding.

Fig 4. Bar graphs of epistatic interactions and dot plots of GO term gene set enrichment in mice under XOR encoding. A: Bar graph of instances of QTL-associated epistatic events for GWAS putative QTL under XOR model. B: Bar graph of instances of epistasis for all epistatic hubs (QTL-associated and non-QTL-associated) under XOR model. C: dot plots for enriched cellular component GO terms. D through F: dot plots for enriched GO terms for biological processes (D), cellular components (E), and molecular functions (F) for XOR-specific epistatic loci. C through E: size of dot corresponds to the number of genes associated with that enrichment. Color gradient illustrates level of significance with higher p-values in blue and lower p-values in red.

In rats, there are four Cartesian non-QTL-associated epistatic hubs under Cartesian and 18 under XOR (Fig 5A,C; File S1). The largest non-QTL-associated epistatic hubs are chr16.54146824 under Cartesian and chr9.60993465_G under XOR. As we observe with GWAS QTL epistasis, the rank order and hub sizes of non-QTL-associated hubs differ between experiments (Fig 5A,C; File S1). When including putative GWAS QTLs as hubs, there are a total of six epistatic hubs identified under Cartesian encoding (Fig 5B; File S1) and 22 under XOR encoding (Fig 5D; File S1). GWAS QTLs account for 33% (2/6) and 18.2% (4/22) of Cartesian and XOR hubs, respectively. Despite the rank order of hubs being largely distinct between experiments, the largest hub under both Cartesian and XOR encodings is the putative GWAS QTL with the largest signal in the rat GWAS([21–23]), chr1.281788173_G. In mice, under XOR encoding, all 25 hubs are non-QTL-associated. The largest is chr12.42624574 with 19 interactions (Fig 4B, File S3).

Fig 5. Bar graphs of epistatic events in non-QTL-associated hubs and all hubs in rats. A and B: Bar graphs of instances of epistatic events for non-QTL-associated hubs (A) and all hubs (B) under Cartesian model (orange bars). C and D: Bar graphs of instances of epistatic events for non-QTL-associated hubs (C) and all hubs (D) under XOR model (blue) bars). B and D: Loci on y axes ending with "_" and allele designation are GWAS putative QTL.

In rats, there is a significant correlation between the significance of SNPs in the GWAS considering its $p$value (i.e. -log10p-value) and the number of epistatic interactions under XOR encoding when considering all epistatic SNPs ($p = 0.0112$, r = 0.253; File S2) and only QTL-associated SNPs ($p = 0.00249$, r = 0.667; File S2). This relationship does not occur under Cartesian encoding. We did not perform this analysis in mice as we did not have access to the GWAS statistics for all SNPs.

Cartesian and XOR Share Common Epistatic Loci while Epistatic Landscapes are Distinct

In rats, the interaction landscapes of Cartesian and XOR two-way epistasis are mostly distinct (Fig 6A,B; File S1). Out of the 182 Cartesian and 312 XOR pairs, only 16 pairs (3%) occur under both encodings (Fig 6C; File S1). However, of the 91 and 100 epistatic loci that occur under Cartesian and XOR encodings, respectively, 75 (65%) are shared Fig 6D; File S1). Thus, although most epistatic loci are shared between Cartesian and XOR, distinct significant two-way epistatic interactions occur under each encoding.

Fig 6. Interaction plots and epistatic model comparisons in rats. A and B: Interaction plots of Cartesian (A) XOR (B) epistasis across rat autosomes. Autosome numbers are depicted as roman numerals and increase clockwise. Orange bars represent counts of epistatic instances per locus. C: Venn diagram of epistatic pairs under Cartesian (left circle) model, the intersect between models (center) and under XOR (right circle) model. D: Venn diagram of epistatic loci under Cartesian (left circle) model, the intersect between models (center) and under XOR (right circle) model. C and D: Color gradient illustrates low (white) to high (blue) occurrences.

In mice, only one epistatic pair is detected and is distinct to Cartesian encoding (i.e., not detected using an XOR encoding) (Fig 2E; File S3). Only two epistatic loci are detected under Cartesian encoding. These two loci are also detected under XOR encoding (Fig 2F; File S3).

Enriched Terms and Pathways Detected for XOR-Specific Epistatic Loci

In rats, the 16 Cartesian-unique and 25 XOR-unique epistatic loci are analyzed using gene set enrichment. Additionally, the 75 loci that are shared between the encodings and the 18 putative QTLs are also analyzed. In Cartesian encoding, the 16 Cartesian-unique loci led to no significant enrichment of cellular component, biological process, or molecular function. However, one KEGG pathway was enriched - Ribosome (File S1). Enrichment of the 75 shared loci reveal biological processes, cellular components, and molecular functions associated largely with immunity (Fig 7A-C; File S2). Under XOR encoding, five enriched biological processes are identified, all of which are metabolic in nature (Fig 7D; File S1; File S2). A total of 10 molecular functions are enriched with the vast majority being oxidoreductase activities and carboxylic acid binding (Fig 7E; File S1; File S2). No cellular components are significantly enriched. A total of 17 KEGG pathways are enriched. Notable metabolism-associated KEGG pathways are nitrogen metabolism, gastric acid secretion, and glucagon signaling pathway (File S1). In mice, the two shared loci between encodings reveal significant enrichments of metabolic terms (Fig 4C; File S3). The enrichments from the 113 XOR-specific loci are mostly involved in immunity (Fig 4D,E; File S3). There are a total of eight biological functions, five cellular components, 15 molecular functions, and two KEGG pathways significantly enriched from XOR-specific loci. However, "detection of chemical stimulus involved in sensory perception of smell", "fatty acid transmembrane transporter activity", and "hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds, in linear amides" are enriched metabolically-relevant processes and functions. Additionally, both KEGG pathways are metabolic in nature (File S3).

Fig 7. Dotplots of gene set enrichment in rats. A through C: dot plots for enriched GO terms for biological processes (A), cellular components (B), and molecular functions (C) for shared genes between Cartesian and XOR models. D and E: dot plots for enriched GO terms for biological processes (D) and molecular functions (E) for XOR-specific epistatic loci. A through E: size of dot corresponds to the number of genes associated with that enrichment. Color gradient illustrates level of significance with higher p-values in blue and lower p-values in red.

In rats, GO term enrichments for the 75 shared epistatic loci between encodings yielded 75 significantly enriched biological processes, three cellular components, five molecular functions, and 46 KEGG pathways. The vast majority of these terms and pathways are involved in immunity (File S1). In mice, The two shared epistatic loci between Cartesian and XOR encodings significantly enriched 10 cellular components and four KEGG pathways, all of which are involved in metabolism (File S3). This is likely due to the gene *ApoB* being located in close proximity to the chr12.7079769 epistatic locus. Interestingly, SNPs near this gene were not implicated in the original GWAS study ([24, 25]; File S3). in rats, The 18 putative GWAS QTL yield 14 significantly enriched biological processes and one molecular function (File S1). Outside of "cellular response to alcohol", all other enriched biological processes are related to immunity. The lone enriched molecular function is "type I interferon receptor binding". The 18 putative GWAS QTL yield KEGG pathway enrichment mostly associated with immunity. However, metabolic-related pathways like "alcoholic liver disease" and "lipid and atherosclerosis" are also enriched (File S1). In mice, the three putative GWAS QTL are enriched for one cellular component ("inner dynein arm") and two metabolic KEGG pathways: "Glycoaminoglycan biosynthesis" and "Cholesterol metabolism" (File S3).

Three-Way Epistatic Landscapes for QTL are Distinct Between Encodings

XOR encoding of significant three-way epistatic interactions between putative GWAS QTL results in more and larger epistatic hubs compared to Cartesian encoding (seven GWAS QTL hubs in Cartesian vs. all 18 GWAS QTLs in XOR) (Fig 8A,B; File S1). Additionally, as we also observe in the two-way experiments, the epistatic landscapes and rank order of epistatic loci are distinct between the two encodings (Fig 8A-E; File S1). Furthermore, out of the 51 significant epistatic triplets in Cartesian and 90 in XOR, only 4 triplets are shared between them despite sharing all epistatic loci (the 18 putative GWAS QTLs) (Fig 8E; File S1). A similar result occurred in the two-way experiment (Fig 6D; File S1). There is no significant correlation between the absolute value of the GWAS betas or the GWAS -log$P$ and the number of three-way interactions for the 18 putative GWAS QTL under either encoding (File S2).

Fig 8. Representations of three-way epistasis in rats. A and B: Bar graphs of occurrences of three-way epistatic events in GWAS QTLs under Cartesian (A, orange bars) and XOR (B, blue bars) models. C and D: Interaction plots of Cartesian (C) XOR (D) three-way epistasis of GWAS QTLs across rat autosomes. Autosome numbers are depicted as roman numerals and increase clockwise. Orange bars represent counts of epistatic instances per locus. E: Venn diagram of epistatic triplets for GWAS QTLs under Cartesian (left circle) model, the intersect between models (center) and under XOR (right circle) model. E: Color gradient illustrates low (white) to high (blue) occurrences.

Discussion

Epistasis Occurs Widely at Non-QTL-Associated Locations

Our findings point the potential ubiquity of epistasis in living systems as we have detected numerous epistatic pairs and loci under two distinct interaction encodings with many of them unique to either Cartesian or XOR. Furthermore, most of interactions detected occur at loci not associated with a GWAS QTL (more than 10Mb upstream or downstream) in both species. However, in rats, the largest epistatic hub under both Cartesian and XOR encoding is chr1.281788173_G, which was also the putative QTL with the largest signal found in the rat GWAS study. In mice, under XOR encoding, all epistatic hubs were non-QTL-associated. One explanation for this is that there are only three putative QTL for BMI from the mouse GWAS study that were mapped to a genomic location. Two studies investigating genome-wide non-additive effects in yeast found results to the contrary where a strong positive correlation between the main effect size of a locus and the number of interactions it was involved in is observed [7, 8]. Although we observed strong correlations in the two-way experiment between the GWAS p-value of a locus and number of interactions under an XOR encoding in rats, it is important to consider that most hubs, both under Cartesian and XOR encodings and in both species, are not located near a putative QTL. There are several possible explanations as to why most instances of detected epistasis occur at non-QTL-associated locations. The first is that there are more possible non-QTL-associated genomic locations that could serve as epistatic loci or hubs. The rat GWAS study identified 18 putative QTL. Therefore, only 360Mb of the *R. norvegicus* genome would be considered QTL-associated under our pruning and categorization strategy. The R. norvegicus genome assembly used in this study is approximately 2.6 Gb in size [65], meaning that there are more possibly

non-QTL-associated regions we could have detected in this study compared to the 18 QTL-associated regions. The mouse GWAS identified four QTL for BMI, three of which mapped to a genomic location, representing only 30Mb of the mouse genome. This same reasoning as to why there are more non-QTL-associated epistatic loci can be applied to why we observe so many more interchromosomal epistatic pairs compared to intrachromosomal pairs. Simply, there are more possible pairwise combinations that can occur across chromosomes than can occur within the same chromosome. The second explanation is that main effects derived from significant GWAS summary statistics are not adequate predictors of epistasis. 82.4% and 82.0% of unique epistatic loci are non-QTL-associated under Cartesian and XOR encodings, respectively. Furthermore, the dataset of SNPs in mice were not selected based upon GWAS summary statistics. Yet, we still detected significant levels of statistical epistasis. Thus, the mouse dataset is a more pertinent example of why searching for epistasis solely based upon univariate effects is not optimal. Even though, to test our methodology, we pruned the original rat dataset using GWAS p-values, we suggest alternative pruning strategies, perhaps based on expert knowledge or biological relevance, to test for epistasis. A third possible explanation is our pruning strategy for epistatic pairs itself. Our initial step is to remove redundant pairs within 10Mb upstream or downstream each other and only the most significant pair (lowest p-value) is retained. In rats, there may be more redundant pairs involving loci in LD with putative QTL compared to non-QTL-associated pairs due our dataset being derived from the 10,000 most significant loci in terms of GWAS p-values. Perhaps with a different pruning strategy, QTL-associated loci would have possibly been highlighted more. However, our pruning strategy may highlight other loci, with non-significant main effects, as centers for epistasis. Taken together, our results highlight the importance of loci with non-significant main effects. It is plausible that combinatory mutations in multiple loci, interacting in networks or pathways, may be required to explain much of the variation observed in phenotypes as canalized systems are likely resistant to alterations to one or few loci ([16–19]. It is possible that we have detected loci underlying BMI in *R. norvegicus* and *M. musculus* that would otherwise go undetected if only main effects were considered. Additionally, the sheer number of loci involved in significant two-way interactions compared to the number of loci with significant main effects detected in the original GWAS point to the potential importance of epistasis in complex biological traits like BMI in two closely-related species.

XOR Model Uncovers Numerous Epistatic Events not Detected under Cartesian Encoding

Although the number of epistatic loci detected in both Cartesian and XOR encodings is comparable and generally overlaps in rats, there are more significant pairs detected under XOR encoding compared to Cartesian. These pairings are mostly distinct with only 16 pairs shared between the encodings. In mice, there is only one significant Cartesian pair detected. This highlights that using a Cartesian encoding, at least in mice, may not uncover any substantial level of epistasis as if we did not apply an XOR encoding, nearly all of the detected epistasis in mice would have been missed. Furthermore, in rats, XOR hubs are larger and more numerous than Cartesian hubs. We also observe a similar results in our 3-way experiment in rats with XOR hubs being larger and triplets being mostly unique between encodings. Taken together, these results provide evidence that interaction encoding is an important consideration when investigating epistatic events in biological systems. The high amount of overlap between epistatic loci is partly attributable to us re-using Cartesian hub identifications for XOR epistasis and our conservative pruning and identification strategy. Despite this, the ways in which these loci interact are mostly distinct in rats. This is explained by the many

possible ways in which epistasis can theoretically occur which Li and Reich illustrate in their enumeration of possible full penetrance models [42]. For two-way interactions, Cartesian encodings model epistasis multiplicatively where the slope of one MLG is zero while another is double that of a third intermediate MLG (File S2). As we have shown, these interactions are indeed possible and detectable. However, genetic systems can be complex, leading to an array of possible interaction mechanisms, like in XOR (File S2). Li and Reich only highlight full penetrance models in their work which serve as adequate models for detecting non-linear and/or non-additive interactions, as we have illustrated. However, we can imagine infinite penetrance functions when incomplete penetrance is considered [66]. Therefore, a Cartesian interaction term likely neglects many possible epistatic relationships as it describes one possible relationship among loci. Our results with the mouse data even suggest that the Cartesian encoding may not be suitable to detect epistasis in that system. This criticism can also be made for XOR, of course. However, our results illustrate that XOR captures more epistatic events than Cartesian in two systems. This begs the question, what about the Cartesian interaction term, or any one interaction term, makes it suitable to represent all forms of epistasis in biological systems? These results illustrate that multiple models of epistasis should be considered when investigating interactions among loci. Despite XOR being an extreme model of epistasis, we identify epistatic interactions in two natural systems, perhaps for the first time, that are detected under a full penetrance XOR model. Thus, our statistical evidence illustrates that biological interactions following an XOR-like model can evolve and be maintained in natural systems and also describe epistatic events and loci not detectable via Cartesian interaction encoding. In the future, we aim to include other penetrance functions to determine the degree of overlap between interaction encodings using our methodology. It is likely that the XOR model, as with the Cartesian model, can only detect a portion of all epistatic events occurring across a genome.

Genes near XOR-Unique Loci are Associated with Metabolism in rats

In rats, functional Annotation of gene models 1Mb upstream and downstream of the 16 loci unique to the Cartesian encoding are not enriched for any GO terms and only one KEGG pathway (Ribosome). However, gene models near the 25 loci unique to XOR yield enrichment results primarily associated with metabolism and catalysis. Of these enrichments, a KEGG pathway of primary interest is glucagon signaling pathway as glucagon-like-protein-1 receptor agonists (GLP-1 RAs) are emerging as important medications to lower plasma glucose and induce weight loss in Murine models and humans [67, 68]. SNPs near the genes that enriched this pathway were not implicated in the associated GWAS and could present possible targets for further study with additional experimentation. When we investigate enrichment in the 75 epistatic loci shared between Cartesian and XOR encoding and the 18 putative QTL from the GWAS, immunity-related processes make up most enriched GO terms and pathways. In mice, XOR-specific GO term enrichments are mostly centered around immunity, with some notable metabolic exceptions. Due to there being no Cartesian-specific epistatic loci in mice, we can not investigate Cartesian-specific enrichment. However, one of the two loci shared between encodings, chr12.7079769, is in close proximity to the *ApoB* gene. *ApoB* is an apolipoprotein, a structural component involved with the formation and synthesis of low density lipoproteins [69]. It has been shown that heterozygous mice for a knockout in *ApoB* were protected against diet-induced hypercholesterolemia after being fed a diet rich in fat and cholesterol [69]. Further, in a more recent study, disruption of *ApoB* led to high incidences of non-alcoholic fatty liver disease [70]. Thus, variants in this gene are likely linked to changes in cholesterol/lipid metabolism and BMI. In rats,

use of the XOR encoding uncovers an enrichment signal of metabolism that would have been missed if only using Cartesian interaction encoding. However, in mice, strong metabolic signal is captured using Cartesian encoding. However, numerous examples of epistasis and notable metabolic-related enrichments are also detected with XOR. It is important to note that because so many XOR-specific epistatic loci are discovered in mice, enrichment becomes more challenging as many gene models are considered, potentially weakening enrichment signals. It is extremely difficult to extrapolate if a similar enrichment of immunity in shared genes, as in rats, would have occurred if more epistasis was detected using Cartesian encoding in mice. More experiments across breeding designs, phenotypes, and species are needed to better understand if certain interaction encodings as associated with specific biological enrichments.

Although links between immunity and obesity-related phenotypes have been well-documented [71–76], immunity-related enrichments and associations are commonly reported across diverse taxa and phenotypes [74, 77–80]. This is largely due to the inherent complexity of gene networks underlying immune systems across animal taxa [81], not excluding invertebrates [77, 79, 80]. Additionally, and perhaps even more importantly, genes related to immunity are likely core to general stress responses in all forms of life [72, 74, 78]. Since most phenotypes of interest to biologists and clinicians center around extreme perturbations from homeostasis (stressful conditions), it perhaps is not surprising to see the over-representation of immunity-related GO terms in functional annotations across biology and medicine. Our results may suggest that, in rats, shared epistatic loci between encodings capture the dynamic and feedback-regulated nature of immunity observed in biological systems. Yet, in a closely-related species (*M. musculus*), XOR-specific epistatic loci uncover a mostly immunity-related signal, again with notable metabolic exceptions. An important consideration is that the mice and rats used in our associated experiments, despite being closely-related biological systems, are reared in different environments, are exposed to unique stressors, eat different foods, and come from different pedigrees. One plausible explanation for the difference in enrichment profiles is the role of gut microbiota in the link between immunity and obesity. There is strong evidence in the literature of the link between the gut microbiome and immunity in affecting obesity-related phenotypes [82–85]. Furthermore, there have been notable differences identified between rats and mice concerning the role of the microbiome in immunity and obesity [86, 87]. It may be possible that our differing enrichment profiles may be highlighting species-level differences between these systems in how obesity is related to immunity. Alternatively, we could also be capturing differences in methodologies and/or environments. Additional experiments are required to elucidate any potential differences in the development of obesity between these species.

In rats, the XOR model identifies epistatic loci enriched for more biologically relevant functions and processes. It is perhaps surprising that an extreme model of epistasis such as XOR yields enrichment more closely associated with metabolism and BMI. Potentially, in its purest, full penetrance form, XOR logic is not biologically plausible due to genetic constraints. However, it is easy to imagine a scenario where in the presence of one activator (A), encoded by gene one, while in the presence of another activator (B), encoded by gene two, results in the transcription of gene three. However, in the presence of both activators, gene three is not expressed. This could be due to activators A and B binding when in the presence of one another, inhibiting their DNA binding motifs, or due both activators collectively blocking other transcription enzymes [44]. Although not perfect, this is just one example of a possible scenario where XOR logic could exist in biological systems. Mechanical examples of XOR interactions in transcriptional regulation, as the one postulated above, are somewhat easy to imagine. However, it is more difficult to imagine a real-life example of XOR in

terms of genetic interactions between two loci. The XOR model assumes a phenotypic score in one extreme when either locus is in a heterozygous state but the other extreme if both loci are heterozygous (File S2). In terms of BMI, where higher phenotypic scores are deleterious in a static environment where food is abundant, this would translate to heterozygote disadvantages in locus one and locus two when the other locus is in a homozygous state. However, if both loci are in a heterozygous state, then the MLG is associated with lower BMI, leading to a heterozygote advantage. Heterosis is commonly described in natural systems, primarily in crop plants [88] but also in humans as observed in the genetic mechanisms underlying sickle-cell anemia and malaria resistance [89]. Yet, the variable phenotypic diversity of sickle cell morphology and/or malaria resistance violates Mendelian expectations [11]. It is possible that one explanation for these abnormalities is epistatic interactions between loci, including those described by the XOR model, which may assist in explaining extreme deviations from heterosis in malaria and other phenotypes.

Conclusions and Future Work

Regardless of the underlying mechanisms or real-world examples, our results show that in our rat sample, some epistatic loci are located near genes that are enriched for metabolic functions and interact in a manner more accurately described, or are only detectable, by the XOR model. In our mouse sample, we illustrate that many epistatic events underlying BMI are only detectable using an XOR model. It is important to keep in mind that we are applying a full penetrance model to a continuous phenotype (BMI). The XOR model assumes a binary, discrete outcome of presence or absence of a certain phenotypic state (commonly a disease phenotype). Despite this, we still detect many pairwise and three-way interactions between SNPs with the XOR model, validating it as a proper model to detect epistasis regardless of phenotype. Our two-way and three-way results in rats suggest that epistatic loci interact in different ways depending on the interaction encoding used. This is evidence that suggests distinct interaction mechanisms exist for different biological networks involving shared loci. Taken together, we provide evidence that epistatic interactions in biological systems are far more complex and ubiquitous than previously thought. In the future, we will apply our methodology to diverse taxa and phenotypes to investigate the complexity of epistasis and warrant the development of validation studies to describe biological interactions. We hope to further elucidate complex traits, including many human diseases, by uncovering genetic relationships that have thus far been elusive to standard analyses. Extending the algorithms given for logistic regression for case/control studies and generalized linear models would be very useful and important especially for handling population structure directly. Li and Reich presented many different models for epistasis and corresponding penetrance functions [42]. A natural extension to this work would be to use more penetrance functions for encoding the interactions. This can already be done with the algorithm given with the only change being how the interaction terms are encoded according the penetrance function. This work for this software extension is already under development.

Acknowledgements

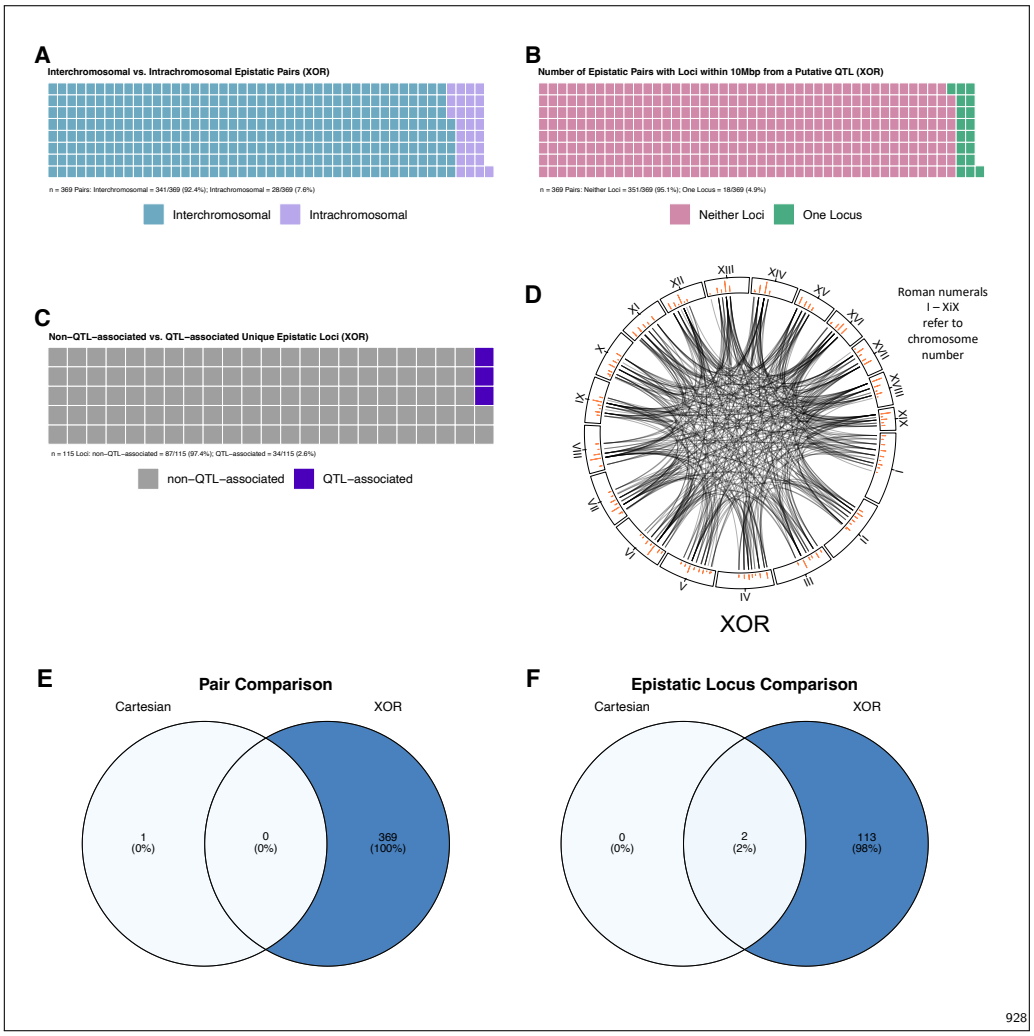
We would like to thank Drs. Takeshi Matsui (Stanford), Sasha Levy (Stanford), and Ian Ehrenreich (University of Southern California, Dornsife) for providing R scripts and data to generate interaction plots in Figures 2, 6, and 8. We would also like to thank Dr. Karl Broman (University of Wisconsin, Madison) for his suggestions in handling the

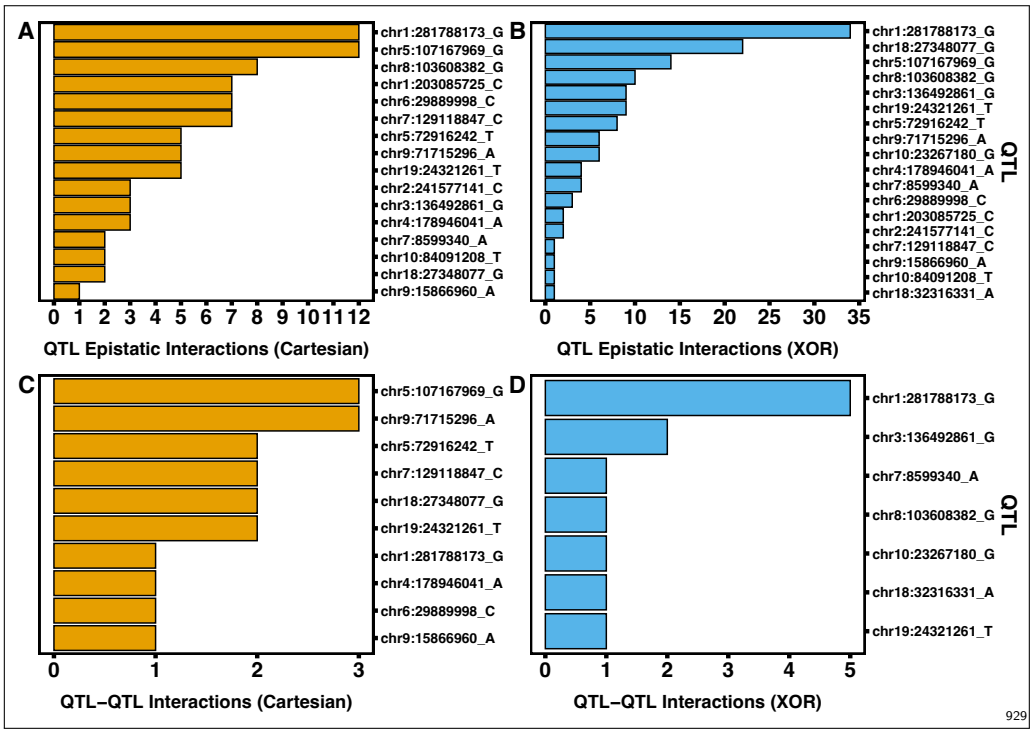
relatedness structure in the rat data set.

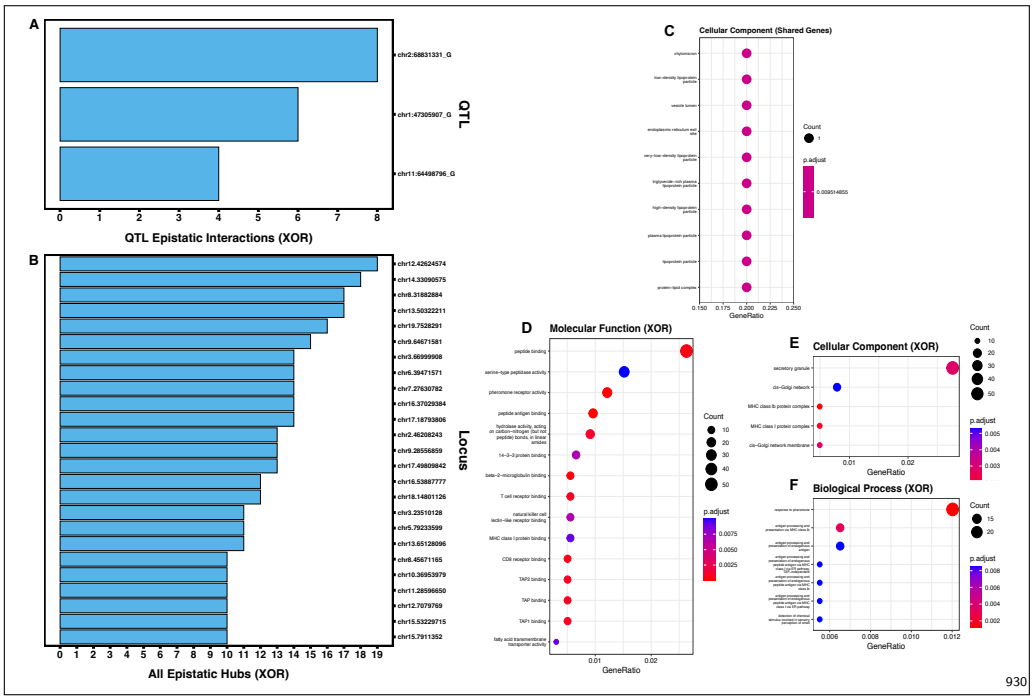
Data Availability

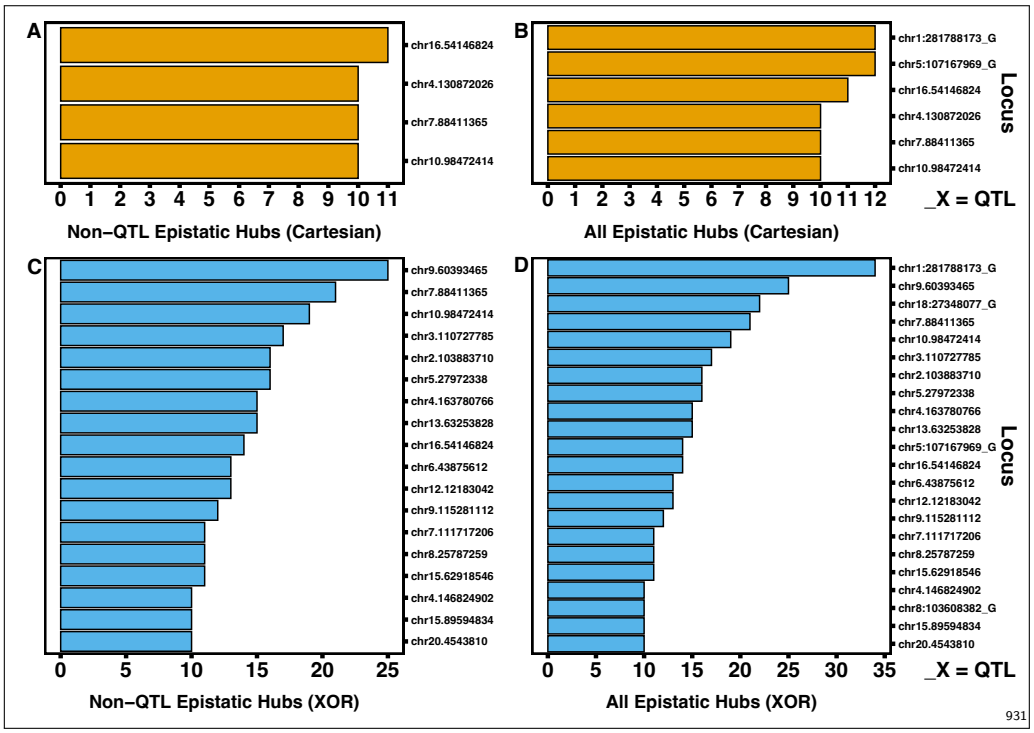
Rat phenotype data and GWAS summary statistics are available at https://library.ucsd.edu/dc/object/bb83725195. Rat genotype data are available at: https://library.ucsd.edu/dc/object/bb15123938. Mouse genotype and phenotype data are available in the "BGLR" R package.

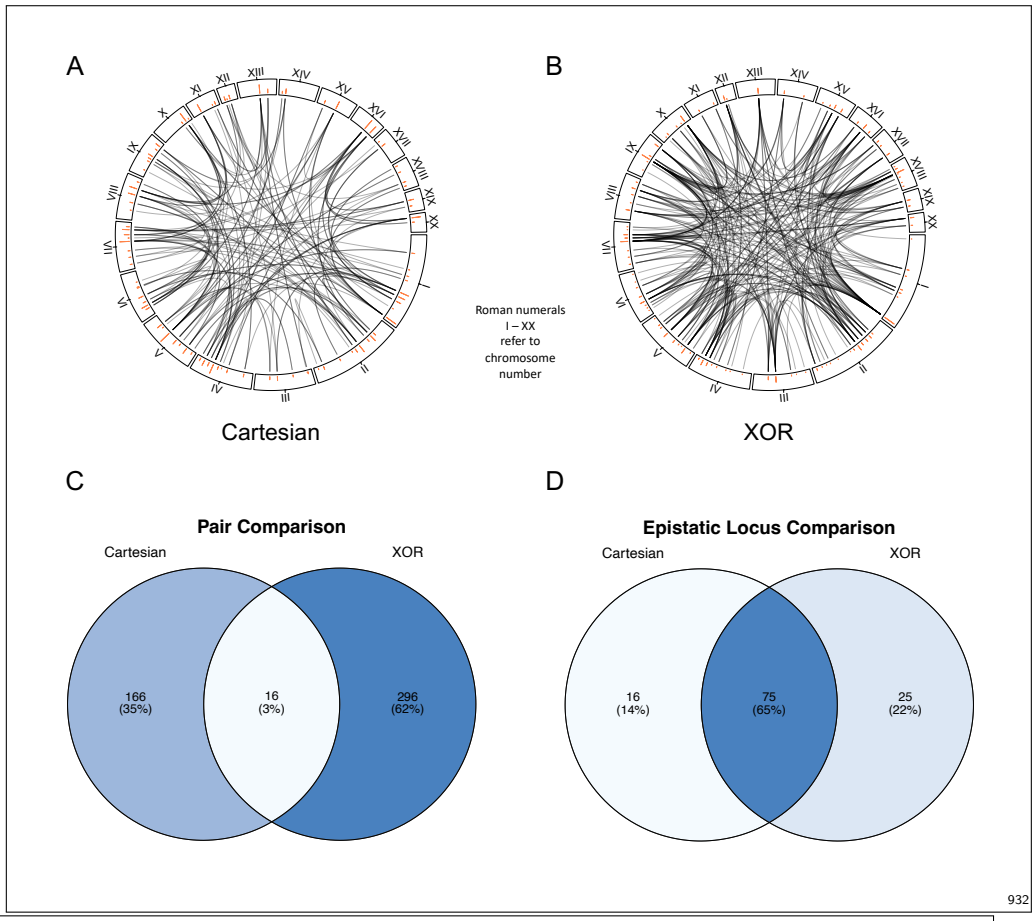
1 Figures


A
Interchromosomal vs. Intrachromosomal Epistatic Pairs (XOR)
n = 369 Pairs: Interchromosomal = 341/369 (92.4%); Intrachromosomal = 28/369 (7.6%)
Interchromosomal
Intrachromosomal
B
Number of Epistatic Pairs with Loci within 10Mbp from a Putative QTL (XOR)
n = 369 Pairs: Neither Loci = 351/369 (95.1%); One Locus = 18/369 (4.9%)
Neither Loci
One Locus
C
Non−QTL−associated vs. QTL−associated Unique Epistatic Loci (XOR)
n = 115 Loci: non−QTL−associated = 87/115 (97.4%); QTL−associated = 34/115 (2.6%)
non−QTL−associated
QTL−associated
D
Roman numerals
I – XiX
refer to
chromosome
number
XOR
E
Pair Comparison
Cartesian
XOR
1
(0%)
0
(0%)
369
(100%)
F
Epistatic Locus Comparison
Cartesian
XOR
0
(0%)
2
(2%)
113
(98%)
928

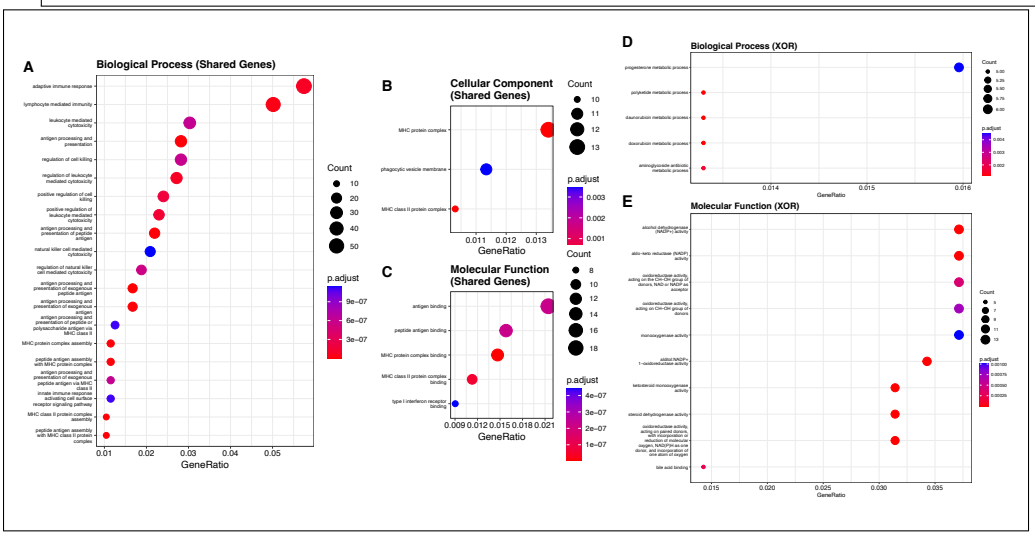

A
chr1:281788173_G
chr5:107167969_G
chr8:103608382_G
chr1:203085725_C
chr6:29889998_C
chr7:129118847_C
chr5:72916242_T
chr9:71715296_A
chr19:24321261_T
chr2:241577141_C
chr3:136492861_G
chr4:178946041_A
chr7:8599340_A
chr10:84091208_T
chr18:27348077_G
chr9:15866960_A
0 1 2 3 4 5 6 7 8 9 10 11 12
QTL Epistatic Interactions (Cartesian)
B
chr1:281788173_G
chr18:27348077_G
chr5:107167969_G
chr8:103608382_G
chr3:136492861_G
chr19:24321261_T
chr5:72916242_T
chr9:71715296_A
chr10:23267180_G
chr4:178946041_A
chr7:8599340_A
chr6:29889998_C
chr1:203085725_C
chr2:241577141_C
chr7:129118847_C
chr9:15866960_A
chr10:84091208_T
chr18:32316331_A
0 5 10 15 20 25 30 35
QTL Epistatic Interactions (XOR)
QTL
C
chr5:107167969_G
chr9:71715296_A
chr5:72916242_T
chr7:129118847_C
chr18:27348077_G
chr19:24321261_T
chr1:281788173_G
chr4:178946041_A
chr6:29889998_C
chr9:15866960_A
0 1 2 3
QTL−QTL Interactions (Cartesian)
D
chr1:281788173_G
chr3:136492861_G
chr7:8599340_A
chr8:103608382_G
chr10:23267180_G
chr18:32316331_A
chr19:24321261_T
0 1 2 3 4 5
QTL−QTL Interactions (XOR)
QTL
929


A
chr2:68831331_G
chr1:47305907_G
chr11:64498796_G
QTL
0 1 2 3 4 5 6 7 8
QTL Epistatic Interactions (XOR)
B
chr12.42624574
chr14.33090575
chr8.31882884
chr13.50322211
chr19.7528291
chr9.64671581
chr3.66999908
chr6.39471571
chr7.27630782
chr16.37029384
chr17.18793806
chr2.46208243
chr9.28556859
chr17.49809842
chr16.53887777
chr18.14801126
chr3.23510128
chr5.79233599
chr13.65128096
chr8.45671165
chr10.36953979
chr11.28596650
chr12.7079769
chr15.53229715
chr15.7911352
Locus
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
All Epistatic Hubs (XOR)
C
Cellular Component (Shared Genes)
chylomicron
low−density lipoprotein particle
vesicle lumen
endoplasmic reticulum exit site
very low−density lipoprotein particle
triglyceride−rich plasma lipoprotein particle
high−density lipoprotein particle
plasma lipoprotein particle
lipoprotein particle
protein−lipid complex
0.150 0.175 0.200 0.225 0.250
GeneRatio
Count
1
p.adjust
0.009514855
D
Molecular Function (XOR)
peptide binding
serine−type peptidase activity
pheromone receptor activity
peptide antigen binding
hydrolase activity, acting on carbon−nitrogen (but not peptide) bonds, in linear amides
14−3−3 protein binding
beta−2−microglobulin binding
T cell receptor binding
natural killer cell lectin−like receptor binding
MHC class I protein binding
CD8 receptor binding
TAP2 binding
TAP binding
TAP1 binding
fatty acid transmembrane transporter activity
0.01 0.02
GeneRatio
Count
10 20 30 40 50
p.adjust
0.0075 0.0050 0.0025
E
Cellular Component (XOR)
secretory granule
cis−Golgi network
MHC class Ib protein complex
MHC class I protein complex
cis−Golgi network membrane
0.01 0.02
GeneRatio
Count
10 20 30 40 50
p.adjust
0.005 0.004 0.003
F
Biological Process (XOR)
response to pheromone
antigen processing and presentation via MHC class Ib
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP−independent
antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP−dependent
antigen processing and presentation of endogenous peptide antigen via MHC class I via TAP−pathway
detection of chemical stimulus involved in sensory perception of smell
0.006 0.008 0.010 0.012
GeneRatio
Count
15 20
p.adjust
0.008 0.006 0.004 0.002
930

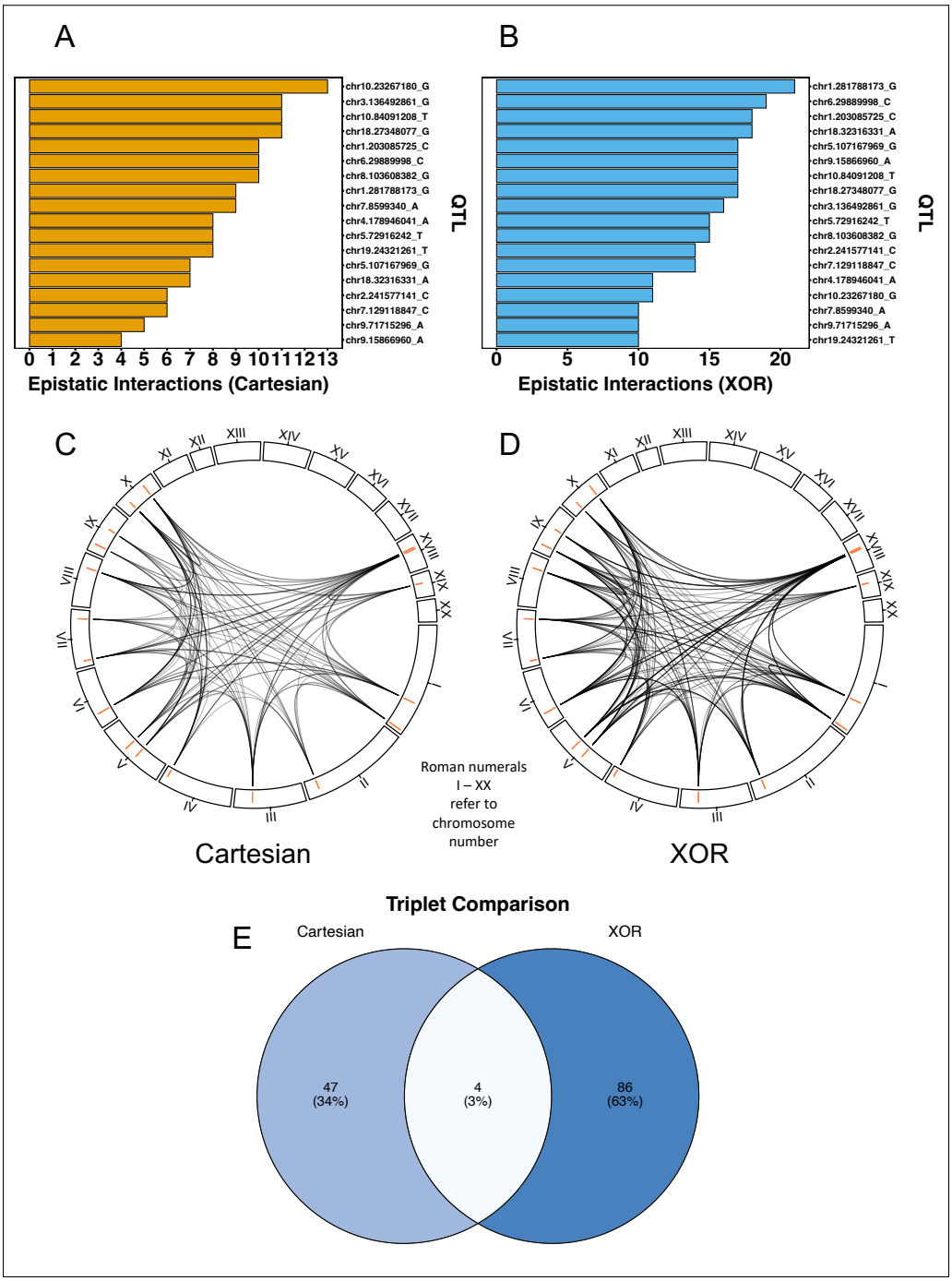

A
chr16.54146824
chr4.130872026
chr7.88411365
chr10.98472414
0 1 2 3 4 5 6 7 8 9 10 11
Non−QTL Epistatic Hubs (Cartesian)
B
chr1:281788173_G
chr5:107167969_G
chr16.54146824
chr4.130872026
chr7.88411365
chr10.98472414
Locus
0 1 2 3 4 5 6 7 8 9 10 11 12
_X = QTL
All Epistatic Hubs (Cartesian)
C
chr9.60393465
chr7.88411365
chr10.98472414
chr3.110727785
chr2.103883710
chr5.27972338
chr4.163780766
chr13.63253828
chr16.54146824
chr6.43875612
chr12.12183042
chr9.115281112
chr7.111717206
chr8.25787259
chr15.62918546
chr4.146824902
chr15.89594834
chr20.4543810
0 5 10 15 20 25
Non−QTL Epistatic Hubs (XOR)
D
chr1:281788173_G
chr9.60393465
chr18:27348077_G
chr7.88411365
chr10.98472414
chr3.110727785
chr2.103883710
chr5.27972338
chr4.163780766
chr13.63253828
chr5:107167969_G
chr16.54146824
chr6.43875612
chr12.12183042
chr9.115281112
chr7.111717206
chr8.25787259
chr15.62918546
chr4.146824902
chr8:103608382_G
chr15.89594834
chr20.4543810
Locus
0 5 10 15 20 25 30 35
_X = QTL
All Epistatic Hubs (XOR)
931


A
B
Roman numerals
I – XX
refer to
chromosome
number
Cartesian
XOR
C
D
Pair Comparison
Cartesian
XOR
166
(35%)
16
(3%)
296
(62%)
Epistatic Locus Comparison
Cartesian
XOR
16
(14%)
75
(65%)
25
(22%)

932


A
Biological Process (Shared Genes)
GeneRatio
Count
p.adjust
B
Cellular Component
(Shared Genes)
MHC protein complex
phagocytic vesicle membrane
MHC class II protein complex
GeneRatio
C
Molecular Function
(Shared Genes)
antigen binding
peptide antigen binding
MHC protein complex binding
MHC class II protein complex binding
type I interferon receptor binding
GeneRatio
D
Biological Process (XOR)
progesterone metabolic process
polyketide metabolic process
daunorubicin metabolic process
doxorubicin metabolic process
aminoglycoside antibiotic metabolic process
GeneRatio
E
Molecular Function (XOR)
alcohol dehydrogenase (NADP+) activity
alditol NADP+ 1-oxidoreductase activity
ketosteroid monooxygenase activity
steroid dehydrogenase activity
bile acid binding
GeneRatio

933



References

1. Leamy LJ, Routman EJ, Cheverud JM. An Epistatic Genetic Basis for Fluctuating Asymmetry of Mandible Size in Mice. Evolution. 2002;56(3):642–653. doi:10.1111/j.0014-3820.2002.tb01373.x.

2. Nelson MR, Kardia SLR, Ferrell RE, Sing CF. A Combinatorial Partitioning Method to Identify Multilocus Genotypic Partitions That Predict Quantitative

Trait Variation. Genome Research. 2001;11(3):458–470. doi:10.1101/gr.172901.

3. Zee RYL, Hoh J, Cheng S, Reynolds R, Grow MA, Silbergleit A, et al. Multi-locus interactions predict risk for post-PTCA restenosis: an approach to the genetic analysis of common complex disease. The Pharmacogenomics Journal. 2002;2(3):197–201. doi:10.1038/sj.tpj.6500101.

4. Rauscher R, Bampi GB, Guevara-Ferrer M, Santos LA, Joshi D, Mark D, et al. Positive epistasis between disease-causing missense mutations and silent polymorphism with effect on mRNA translation velocity. Proceedings of the National Academy of Sciences. 2021;118(4):e2010612118. doi:10.1073/pnas.2010612118.

5. Rohlfs EM, Shaheen NJ, Silverman LM. Is the Hemochromatosis Gene a Modifier Locus for Cystic Fibrosis? Genetic Testing. 1998;2(1):85–88. doi:10.1089/gte.1998.2.85.

6. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. The American Journal of Human Genetics. 2001;69(1):138–147. doi:10.1086/321276.

7. Hallin J, Märtens K, Young AI, Zackrisson M, Salinas F, Parts L, et al. Powerful decomposition of complex traits in a diploid model. Nature Communications. 2016;7(1):13311. doi:10.1038/ncomms13311.

8. Matsui T, Mullis MN, Roy KR, Hale JJ, Schell R, Levy SF, et al. The interplay of additivity, dominance, and epistasis on fitness in a diploid yeast cross. Nature Communications. 2022;13(1):1463. doi:10.1038/s41467-022-29111-z.

9. Greene CS, Penrod NM, Williams SM, Moore JH. Failure to Replicate a Genetic Association May Provide Important Clues About Genetic Architecture. PLOS ONE. 2009;4(6):e5639. doi:10.1371/journal.pone.0005639.

10. Gibson G. Epistasis and Pleiotropy as Natural Properties of Transcriptional Regulation. Theoretical Population Biology. 1996;49(1):58–89. doi:10.1006/tpbi.1996.0003.

11. Templeton AR. Epistasis and Complex Traits. In: Wolf J, Brodie III B, Wade M, editors. Epistasis and the Evolutionary Process. New York: Oxford University Press; 2000.

12. Gallie DR. Protein-protein interactions required during translation. Plant Molecular Biology. 2002;50(6):949–970. doi:10.1023/A:1021220910664.

13. Moore JH. The Ubiquitous Nature of Epistasis in Determining Susceptibility to Common Human Diseases. Human Heredity. 2003;56(1-3):73–82. doi:10.1159/000073735.

14. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. BioEssays. 2005;27(6):637–646. doi:10.1002/bies.20236.

15. Moore JH, Williams SM. Epistasis and Its Implications for Personal Genetics. The American Journal of Human Genetics. 2009;85(3):309–320. doi:10.1016/j.ajhg.2009.08.006.

16. Waddington CH. Canalization of Development and the Inheritance of Acquired Characters. Nature. 1942;150(3811):563–565. doi:10.1038/150563a0.

17. Rice SH. The Evolution of Canalization and the Breaking of Von Baer's Laws: Modeling the Evolution of Development with Epistasis. Evolution. 1998;52(3):647–656. doi:10.1111/j.1558-5646.1998.tb03690.x.

18. Gibson G, Wagner G. Canalization in evolutionary genetics: a stabilizing theory? BioEssays. 2000;22(4):372–380. doi:10.1002/(SICI)1521-1878(200004)22:4¡372::AID-BIES7¿3.0.CO;2-J.

19. Waddington CH. The Strategy of the Genes. London: Routledge; 2014. Available from: `https://www.taylorfrancis.com/books/mono/10.4324/9781315765471/strategy-genes-waddington`.

20. Madar VS, Batista SL. Solving The Ordinary Least Squares in Closed Form, Without Inversion or Normalization; 2023.

21. Chitre AS, Polesskaya O, Holl K, Gao J, Cheng R, Bimschleger H, et al. Genome-Wide Association Study in 3,173 Outbred Rats Identifies Multiple Loci for Body Weight, Adiposity, and Fasting Glucose. Obesity. 2020;28(10):1964–1973. doi:10.1002/oby.22927.

22. Chitre AS, Polesskaya O, Holl K, Gao J, Cheng R, Bimschleger H, et al.. Genome-Wide Association Study in 3,173 Outbred Rats for Body Weight, Adiposity, and Fasting Glucose; 2022. Available from: `https://cgord.org/dataset/2`.

23. Wright SN, Leger BS, Rosenthal SB, Liu S, Jia T, Chitre AS, et al. Genome-wide association studies of human and rat BMI converge on synapse, epigenome, and hormone signaling networks. Cell Reports In Press. 2023;.

24. Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO, et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nature Genetics. 2006;38(8):879–887. doi:10.1038/ng1840.

25. Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JNP, Mott R, et al. Genetic and environmental effects on complex traits in mice. Genetics. 2006;174(2):959–984. doi:10.1534/genetics.106.060004.

26. Lewontin RC. Annotation: the analysis of variance and the analysis of causes. American Journal of Human Genetics. 1974;26(3):400–411.

27. Wahlsten D. Insensitivity of the analysis of variance to heredity-environment interaction. Behavioral and Brain Sciences. 1990;13(1):109–120. doi:10.1017/S0140525X00077797.

28. Lewontin RC. Commentary: Statistical analysis or biological analysis as tools for understanding biological causes. International Journal of Epidemiology. 2006;35(3):536–537. doi:10.1093/ije/dyl070.

29. Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. Optimizationof neural network architecture using genetic programming improvesdetection and modeling of gene-gene interactions in studies of humandiseases. BMC Bioinformatics. 2003;4(1):28. doi:10.1186/1471-2105-4-28.

30. Ritchie MD, Motsinger AA, Bush WS, Coffey CS, Moore JH. Genetic programming neural networks: A powerful bioinformatics tool for human genetics. Applied Soft Computing. 2007;7(1):471–479. doi:10.1016/j.asoc.2006.01.013.

31. Abegaz F, Van Lishout F, Mahachie John JM, Chiachoompu K, Bhardwaj A, Gusareva ES, et al. Epistasis Detection in Genome-Wide Screening for Complex Human Diseases in Structured Populations. Systems Medicine. 2019;2(1):19–27. doi:10.1089/sysm.2019.0003.

32. Crawford L, Zeng P, Mukherjee S, Zhou X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genetics. 2017;13(7). doi:10.1371/journal.pgen.1006869.

33. Brandon Ogbunugafor C, Scarpino SV. In: Battiston F, Petri G, editors. Higher-Order Interactions in Biology: The Curious Case of Epistasis. Cham: Springer International Publishing; 2022. p. 417–433. Available from: `https://doi.org/10.1007/978-3-030-91374-8_18`.

34. Niel C, Sinoquet C, Dina C, Rocheleau G. A survey about methods dedicated to epistasis detection. Frontiers in Genetics. 2015;6. doi:10.3389/fgene.2015.00285.

35. Russ D, Williams JA, Cardoso VR, Bravo-Merodio L, Pendleton SC, Aziz F, et al. Evaluating the detection ability of a range of epistasis detection methods on simulated data for pure and impure epistatic models. PLoS One. 2022;17(2). doi:10.1371/journal.pone.0263390.

36. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics. 2007;81(3):559–575.

37. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, et al. BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. The American Journal of Human Genetics. 2010;87(3):325–340. doi:https://doi.org/10.1016/j.ajhg.2010.07.021.

38. Zhang X, Huang S, Zou F, Wang W. TEAM: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics. 2010;26(12):i217–i227. doi:10.1093/bioinformatics/btq186.

39. Bayat A, Hosking B, Jain Y, Hosking C, Kodikara M, Reti D, et al. Fast and accurate exhaustive higher-order epistasis search with BitEpi. Scientific Reports. 2021;11. doi:10.1038/s41598-021-94959-y.

40. Schüpbach T, Xenarios I, Bergmann S, Kapur K. FastEpistasis: a high performance computing solution for quantitative trait epistasis. Bioinformatics. 2010;26(11):1468–1469. doi:10.1093/bioinformatics/btq147.

41. Zhu S, Fang G. MatrixEpistasis: ultrafast, exhaustive epistasis scan for quantitative traits with covariate adjustment. Bioinformatics. 2018;34(14):2341–2348. doi:10.1093/bioinformatics/bty094.

42. Li W, Reich J. A Complete Enumeration and Classification of Two-Locus Disease Models. Human Heredity. 2000;50(6):334–349. doi:10.1159/000022939.

43. Buchler NE, Gerland U, Hwa T. On schemes of combinatorial transcription logic. Proceedings of the National Academy of Sciences. 2003;100(9):5136–5141. doi:10.1073/pnas.0930314100.

44. Tagkopoulos I, Liu YC, Tavazoie S. Predictive Behavior Within Microbial Genetic Networks. Science. 2008;320(5881):1313–1317. doi:10.1126/science.1154456.

45. Yule GU. On the Theory of Correlation for any Number of Variables Treated by a New System of Notation. Proc R Soc Lond A. 1907;79(529):182–193. doi:10.1098/rspa.1907.0028.

46. Morrison DF. Multivariate Statistical Methods. 4th ed. New York: McGraw-Hill; 2004.

47. Laurie C, Wang S, Carlini-Garcia LA, Zeng ZB. Mapping epistatic quantitative trait loci. BMC Genetics. 2014;15(1):112. doi:10.1186/s12863-014-0112-9.

48. Ang RML, Chen SAA, Kern AF, Xie Y, Fraser HB. Widespread epistasis among beneficial genetic variants revealed by high-throughput genome editing. Cell Genomics. 2023;(2666-979X):100260–100260. doi:https://doi.org/10.1016/j.xgen.2023.100260.

49. Hansen C, Spuhler K. Development of the National Institutes of Health Genetically Heterogeneous Rat Stock. Alcohol: Clinical and Experimental Research. 1984;8(5):477–479. doi:10.1111/j.1530-0277.1984.tb05706.x.

50. Pérez P, de los Campos G. Genome-Wide Regression and Prediction with the BGLR Statistical Package. Genetics. 2014;198(2):483–495. doi:10.1534/genetics.114.164442.

51. Team RC. R: A language and environment for statistical computing; 2022. Available from: `https://www.R-project.org/`.

52. Yang J, Lee S, Goddard M, Visscher P. GCTA: a tool for Genome-wide Complex Trait Analysis. American journal of human genetics. 2011;88:76–82. doi:10.1016/j.ajhg.2010.11.011.

53. Sul JH, Martin LS, Eskin E. Population structure in genetic studies: Confounding factors and mixed models. PLoS Genetics. 2018;14. doi:10.1371/journal.pgen.1007309.

54. Joo JWJ, Kang EY, Org E, Furlotte N, Parks B, Hormozdiari F, et al.; Los Angeles California. Bioinformatics Interdepartmental Ph.D. Program. Efficient and Accurate Multiple-Phenotype Regression Method for High Dimensional Data Considering Population Structure. Genetics. 2016;204(4):1379–1390.

55. Kang HM, Ye C, Eskin E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics. 2008;180:1909–25. doi:10.1534/genetics.108.094201.

56. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010;42:348–54. doi:10.1038/ng.548.

57. Covarrubias-Pazaran G. Genome-Assisted Prediction of Quantitative Traits Using the R Package sommer. PLOS ONE. 2016;11(6):1–15. doi:10.1371/journal.pone.0156744.

58. Higham NJ. Functions of Matrices: Theory and Computation. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics; 2008.

59. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272. doi:10.1038/s41592-019-0686-2.

60. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference; 2010.

61. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods. 2015;12(2):115–121. doi:10.1038/nmeth.3252.

62. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation. 2021;2(3):100141. doi:10.1016/j.xinn.2021.100141.

63. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A Journal of Integrative Biology. 2012;16(5):284–287. doi:10.1089/omi.2011.0118.

64. Yu G, Wang LG, Yan GR, He QY. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics. 2015;31(4):608–609. doi:10.1093/bioinformatics/btu684.

65. Smith JR, Hayman GT, Wang SJ, Laulederkind SJF, Hoffman MJ, Kaldunski ML, et al. The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. Nucleic Acids Research. 2020;48(D1):D731–D742. doi:10.1093/nar/gkz1041.

66. Urbanowicz RJ, Kiralis J, Sinnott-Armstrong NA, Heberling T, Fisher JM, Moore JH. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData Mining. 2012;5(1):16. doi:10.1186/1756-0381-5-16.

67. Nauck MA, Quast DR, Wefers J, Meier JJ. GLP-1 receptor agonists in the treatment of type 2 diabetes – state-of-the-art. Molecular Metabolism. 2021;46:101102. doi:10.1016/j.molmet.2020.101102.

68. Bendotti G, Montefusco L, Lunati ME, Usuelli V, Pastore I, Lazzaroni E, et al. The anti-inflammatory and immunological properties of GLP-1 Receptor Agonists. Pharmacological Research. 2022;182:106320. doi:10.1016/j.phrs.2022.106320.

69. Farese RV, Ruland SL, Flynn LM, Stokowski RP, Young SG. Knockout of the mouse apolipoprotein B gene results in embryonic lethality in homozygotes and protection against diet-induced hypercholesterolemia in heterozygotes. Proceedings of the National Academy of Sciences of the United States of America. 1995;92(5):1774–1778.

70. Li BT, Sun M, Li YF, Wang JQ, Zhou ZM, Song BL, et al. Disruption of the ERLIN–TM6SF2–APOB complex destabilizes APOB and contributes to non-alcoholic fatty liver disease. PLOS Genetics. 2020;16(8):e1008955. doi:10.1371/journal.pgen.1008955.

71. Andersen CJ, Murphy KE, Fernandez ML. Impact of Obesity and Metabolic Syndrome on Immunity. Advances in Nutrition. 2016;7(1):66–75. doi:10.3945/an.115.010207.

72. Chrousos GP. The stress response and immune function: clinical implications. The 1999 Novera H. Spector Lecture. Annals of the New York Academy of Sciences. 2000;917:38–67. doi:10.1111/j.1749-6632.2000.tb05371.x.

73. Schäffler A, Buechler C. CTRP family: linking immunity to metabolism. Trends in Endocrinology & Metabolism. 2012;23(4):194–204. doi:10.1016/j.tem.2011.12.003.

74. Saijo Y, Loo EPi. Plant immunity in signal integration between biotic and abiotic stress responses. New Phytologist. 2020;225(1):87–104. doi:10.1111/nph.15989.

75. Winn NC, Cottam MA, Wasserman DH, Hasty AH. Exercise and Adipose Tissue Immunity: Outrunning Inflammation. Obesity. 2021;29(5):790–801. doi:10.1002/oby.23147.

76. Pivonello C, Negri M, Patalano R, Amatrudo F, Montò T, Liccardi A, et al. The role of melatonin in the molecular mechanisms underlying metaflammation and infections in obesity: A narrative review. Obesity Reviews. 2022;23(3):e13390. doi:10.1111/obr.13390.

77. Loker ES, Adema CM, Zhang SM, Kepler TB. Invertebrate immune systems – not homogeneous, not simple, not well understood. Immunological reviews. 2004;198:10–24.

78. Sinclair BJ, Ferguson LV, Salehipour-shirazi G, MacMillan HA. Cross-tolerance and Cross-talk in the Cold: Relating Low Temperatures to Desiccation and Immune Stress in Insects. Integrative and Comparative Biology. 2013;53(4):545–556. doi:10.1093/icb/ict004.

79. Sun Y, Zhang X, Wang Y, Day R, Yang H, Zhang Z. Immunity-related genes and signaling pathways under hypoxic stresses in Haliotis diversicolor: a transcriptome analysis. Scientific Reports. 2019;9(1):19741. doi:10.1038/s41598-019-56150-2.

80. Freda PJ, Toxopeus J, Dowle EJ, Ali ZM, Heter N, Collier RL, et al. Transcriptomic and functional genetic evidence for distinct ecophysiological responses across complex life cycle stages. Journal of Experimental Biology. 2022;225(11):jeb244063. doi:10.1242/jeb.244063.

81. Subramanian N, Torabi-Parizi P, Gottschalk RA, Germain RN, Dutta B. Network representations of immune system complexity. Wiley Interdisciplinary Reviews Systems Biology and Medicine. 2015;7(1):13–38. doi:10.1002/wsbm.1288.

82. Bengmark S. Gut microbiota, immune development and function. Pharmacological Research. 2013;69(1):87–113. doi:10.1016/j.phrs.2012.09.002.

83. Manco M, Putignani L, Bottazzo GF. Gut Microbiota, Lipopolysaccharides, and Innate Immunity in the Pathogenesis of Obesity and Cardiovascular Risk. Endocrine Reviews. 2010;31(6):817–844. doi:10.1210/er.2009-0030.

84. Boulangé CL, Neves AL, Chilloux J, Nicholson JK, Dumas ME. Impact of the gut microbiota on inflammation, obesity, and metabolic disease. Genome Medicine. 2016;8(1):42. doi:10.1186/s13073-016-0303-2.

85. Sanz Y, Santacruz A, Gauffin P. Gut microbiota in obesity and metabolic disorders. Proceedings of the Nutrition Society. 2010;69(3):434–441. doi:10.1017/S0029665110001813.

86. Butler MJ, Perrini AA, Eckel LA. The Role of the Gut Microbiome, Immunity, and Neuroinflammation in the Pathophysiology of Eating Disorders. Nutrients. 2021;13(2):500. doi:10.3390/nu13020500.

87. Harris K, Kassis A, Major G, Chou CJ. Is the Gut Microbiota a New Factor Contributing to Obesity and Its Metabolic Disorders? Journal of Obesity. 2012;2012:e879151. doi:10.1155/2012/879151.

88. Fu D, Xiao M, Hayward A, Fu Y, Liu G, Jiang G, et al. Utilization of crop heterosis: a review. Euphytica. 2014;197(2):161–173. doi:10.1007/s10681-014-1103-7.

89. López C, Saravia C, Gomez A, Hoebeke J, Patarroyo MA. Mechanisms of genetically-based resistance to malaria. Gene. 2010;467(1):1–12. doi:10.1016/j.gene.2010.07.008.