RESEARCH ARTICLE

Compositional epistasis detection using a few prototype disease models

Lu Cheng, Mu Zhu *

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada

* mu.zhu@uwaterloo.ca

Abstract

We study computational approaches for detecting SNP-SNP interactions that are characterized by a set of "two-locus, two-allele, two-phenotype and complete-penetrance" disease models. We argue that existing methods, which use data to determine a best-fitting disease model for each pair of SNPs prior to screening, may be too greedy. We present a less greedy strategy which, for each given pair of SNPs, limits the number of candidate disease models to a set of prototypes determined a priori.

Citation: Cheng L, Zhu M (2019) Compositional epistasis detection using a few prototype disease models. PLoS ONE 14(3): e0213236. https://doi.org/10.1371/journal.pone.0213236

Editor: Li Chen, Auburn University - Harrison School of Pharmacy, UNITED STATES

Received: August 2, 2017

Accepted: February 19, 2019

Published: March 27, 2019

Data Availability Statement: This study makes use of data generated by and obtained from the Wellcome Trust Case-Control Consortium (WTCCC). Researchers can go to the WTCCC web site (https://www.wtccc.org.uk/) and email the WTCCC data access committee (cdac@sanger.ac.uk) to apply for access to use their data.

Funding: This work was funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada (http://www.nserc-crsng.gc.ca/), grant no. RGPIN-2016-03876. Funding for the project was also provided by the Wellcome Trust under award 076113, 085475 and 090355. The

1 Introduction

For many years, scientists have tried to identify single-nucleotide polymorphisms (SNPs) that are associated with various diseases, but over the years it is becoming apparent that single genetic variations can explain only very little heritability. This has come to be known as the so-called "missing heritability problem" [1–3], and has prompted many scientists to conjecture that perhaps SNP-SNP interactions are more prevalent than we had previously thought [4].

In genetics, the term "epistasis" refers to the phenomenon that the effect of one gene (or SNP) is dependent on the presence of others. Different definitions of epistasis exist. For example, in biochemical genetics, the term "functional epistasis" is sometimes used to mean the molecular interactions that proteins (and/or other genetic elements) have with one another; whereas in population and/or quantitative genetics, the terms "statistical epistasis" and "compositional epistasis" are often used. The former is due to Fisher [5] and usually taken to mean deviation from additive genetic effects, while the latter emphasizes the notion of having a masking effect—as such, some [6–8] believe it to be closer to the original meaning of the word "epistasis" when Bateson [9] first coined it in 1909. As Phillips [7] wrote, "[c]ompositional epistasis measures the effects of allele substitution against a particular fixed genetic background, while statistical epistasis measures the average effect of allele substitution against the population average genetic background."

To characterize different compositional epistatic effects, we follow various researchers who have studied this problem and focus on a set of "two-locus, two-allele, two-phenotype, and complete-penetrance" (TTTC) disease models [10]. Table 1 shows a few examples. Often, these disease models can be interpreted as one SNP having a certain masking effect on the other. For instance, the recessive-recessive disease model [Table 1(c)] can be viewed as the

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Table 1. Examples of TTTC disease models. A "1" means the corresponding genotype combination, e.g., "aabb" in (c), would elevate the risk of disease, whereas a "0" means it would not.

(a)				(b)			
	AA	Aa	aa		AA	Aa	aa
BB	**1**	**1**	0	BB	0	**1**	0
Bb	**1**	**1**	0	Bb	**1**	0	**1**
bb	0	0	0	bb	0	**1**	0
(c)				(d)			
	AA	Aa	aa		AA	Aa	aa
BB	0	0	0	BB	0	**1**	0
Bb	0	0	0	Bb	**1**	**1**	**1**
bb	0	0	**1**	bb	0	**1**	0

https://doi.org/10.1371/journal.pone.0213236.t001

major allele "A" from one SNP having a masking effect on the causal genotype "bb" from the other SNP, or as the major allele "B" having a similar masking effect on the causal genotype "aa".

Clearly, these TTTC disease models can describe only two-way interactions between two SNPs, and the notion of epistasis itself certainly does not preclude higher-order interactions among more than two SNPs. At a *genome-wide* level, however, screening for higher-order interactions is still largely impractical. For example, even with 100,000 SNPs, there would be $\binom{100,000}{2} \approx 5.0 \times 10^9$ or about 5 *billion* SNP-pairs to screen already if we limited ourselves to 2-way interactions only, and $\binom{100,000}{3} \approx 1.7 \times 10^{14}$ SNP-triplets to screen if 3-way interactions were to be considered. Therefore, in this paper, we take a "narrow" point of view by restricting ourselves to consider only two-way interactions.

What's more, these TTTC models are practically useful, especially when the minor allele frequency (MAF) is low. A TTTC model has two degrees of freedom, corresponding to the two penetrance levels, denoted respectively by "1" and "0" in Table 1; whereas a "full model" will have a 9 degrees of freedom, one for each of the nine genotype combinations. When the MAF is low, there can be insufficient data for some of the rare genotype combinations, making it hard to obtain reliable parameter estimates. (In the extreme case, we may have no data in the sample for a particular genotype combination.) Under such circumstances, it is beneficial to reduce the number of parameters, or the degree of freedom. Using a TTTC model, one only has to estimate two parameters. By limiting the degree of freedom in this way, the power of the statistical test can be improved.

Thus, when we say "epistasis" in this paper, we are largely referring to these TTTC disease models only. Even so, there are still 2^9 possible TTTC disease models in theory [10] for *each pair* of SNPs, and it is generally not possible to screen them all. But a bad choice of the disease model can be detrimental, in that a pair of SNPs may appear highly associated with an outcome under one disease model and not associated under another. For example, studies on single-locus effects have generally confirmed that the power (of detecting an existing effect) is largest when the correct genetic model—e.g., recessive, dominant, additive, and so on—is specified [11–13], and there is all the reason to expect that the same conclusion will hold for detecting epistatic effects between two SNPs.

Among methods available for choosing a disease model for each pair of SNPs prior to screening, two popular ones are: the multi-factor dimensionality reduction (MDR) method by Ritchie *et al.* [14], and the method by Wan *et al.* [15], which we shall refer to throughout the paper simply as the "ratio split" (RS) method. Both of these methods rely on the case-control ratios of different genotype combinations (i.e., AABB, AABb, and so on) in order to decide on a particular disease model to use for a given pair of SNPs. Specifically, the MDR method determines a disease model by thresholding the case-control ratios; typically, genotype combinations with ratios ≥ 1 (on a balanced case-control sample) are regarded as high risk. The RS method, on the other hand, first sorts the case-control ratios in descending order and evaluates 8 different disease models by sequentially considering the top x genotype combinations as high risk, for $x = 1, 2, \ldots, 8$. Then, it chooses the one that best predicts the outcome (e.g., disease).

Both of these methods are essentially greedy and use the data twice: first, to determine the disease model for each pair of SNPs; then, to determine whether each pair of SNPs is associated with the outcome. As such, they can be *overly* adaptive to data, and have a tendency to produce many false positives. The cost of using the data twice is especially pronounced if the sample size is relatively small (which is almost always the case for genome-wide association studies), and/or if the data quality is not so good. Indeed, this kind of concern has been reported in the literature [16], especially in the context of genome-wide studies where extra out-of-sample validation, which can help mitigate such problems, is computationally prohibitive.

Instead of relying on the case-control ratios to determine what disease model to use for a pair of SNPs, our main idea is based on the observation that some of these TTTC disease models are more similar than others. In Table 1, for example, arguably the two models on the left [(a) and (c)] are quite different from, whereas the two on the right [(b) and (d)] are somewhat similar to, each other. While we shall be more specific later (Section 3) about how we propose to measure the similarity between two disease models, such an observation nonetheless means that we can first group all possible disease models into a few clusters, and then select a representative prototype from each cluster for screening purposes. The set of prototype models can be seen to constrain the search space somewhat, in the sense that only disease models in the prototype set are now "permitted". This allows our method to be less data-adaptive, while including a prototype from each cluster still ensures that we are not systematically missing important parts of the search space. In what follows, we will sometimes use the acronym "PTY" (for "prototype") to refer to our method, especially in tables and figures.

It is worth mentioning that a cluster analysis of all disease models is beneficial in its own right. For example, it may allow us to better understand and characterize different epistatic effects (more on this below in Section 3.2), for which there have been a few previous endeavours [10, 17–19].

1.1 Marginal versus sequential screening

Throughout the paper, we will use the following empirical protocol repeatedly to compare different methods. For any given pair of SNPs, e.g., (i, j), each method has its own way of determining a "best-fitting" disease model—call it $M_{i,j}$. A *nominal* measure of association between the (i, j)-pair and the outcome is then computed as the $\chi^2_{(1)}$-statistic for testing whether the risky/non-risky assignment by $M_{i,j}$ is statistically independent of the outcome (i.e., disease or no disease), which we simply denote as $\widehat{\chi}^2_{i,j}$. (We will explain in more detail later in Section 2.3 why we use the adjective "nominal" to describe these association measures.) The pair (i, j) can then be ranked according to $\widehat{\chi}^2_{i,j}$ or considered having been "selected" or "detected" by the method if $\widehat{\chi}^2_{i,j}$ exceed a certain significance threshold. We refer to this as the "marginal screening procedure". (One also can use only part of the data to determine $M_{i,j}$, and compute an *out-of-sample* measure of association by testing $M_{i,j}$ against the outcome on the remaining data. For example, MDR is usually applied in this manner when the number of candidate SNPs being studied is relatively small. To reduce variation caused by chance division of the data, however, such a process often needs to be repeated a few times and the resulting measures averaged, thus making it computationally prohibitive for genome-wide screening [20, 21]).

Alternatively, we also can combine the effects of multiple SNP-pairs *sequentially*. For example, after having selected the top pair—call it (i_1, j_1), we can *re-assess* each remaining pair (i, j) by testing whether the *combined* risky/non-risky assignment by

$$M_{i_1,j_1} \text{ or } M_{i,j} \tag{1}$$

is independent of the outcome. We use $\widehat{\chi}^2_{i,j|\mathcal{H}}$ to denote the corresponding test statistic, where $\mathcal{H}$ means the entire *history* of pairs already selected so far. (After the top pair has been selected, $\mathcal{H} = \{M_{i_1,j_1}\}$; after two pairs have been selected, $\mathcal{H} = \{M_{i_1,j_1}, M_{i_2,j_2}\}$; and so on.) The pair to be selected next is

$$\underset{M_{i,j}\notin\mathcal{H}}{\arg\max} \quad \widehat{\chi}^2_{i,j|\mathcal{H}}, \tag{2}$$

rather than

$$\underset{M_{i,j}\notin\mathcal{H}}{\arg\max} \quad \widehat{\chi}^2_{i,j}. \tag{3}$$

We refer to this as the “sequential screening procedure”.

2 Motivation

Before we describe our approach in more detail, we first provide some motivations by discussing some weaknesses of existing methods. We should emphasize that these are merely some *examples* of scenarios in which PTY can be seen to have certain advantages over MDR and RS. They are by no means the only—or even necessarily the main —such scenarios. The reason why they are being presented, rather than others, is because they are still relatively easy for us to describe with a reasonable amount of clarity, whether algebraically (Section 2.1), verbally (Section 2.2), or both (Section 2.3).

2.1 A pathological scenario

We begin by considering a pathological scenario. Suppose that two pairs of SNPs (e.g., {A/a, B/b}, {C/c, D/d}) are independent (Table 2). For $i = 1, 2, \ldots, 9$, let w_i be the probability of having the i-th genotype combination in the first pair, and likewise v_j for the second pair. For simplicity, suppose each genotype combination is either risky ($\in R$) or non-risky ($\in N$). For $k, \ell \in \{0, 1\}$, let $p_{k\ell}$ be the penetrance level for individuals having risky combinations from both pairs ($k = \ell = 1$), the first pair only ($k = 1, \ell = 0$), the second pair only ($k = 0, \ell = 1$), or neither pair ($k = \ell = 0$). Then, derivations contained in S1 Appendix show that, if

$$p_{11}\sum_{j\in R} v_j + p_{10}\sum_{j\in N} v_j = p_{01}\sum_{j\in R} v_j + p_{00}\sum_{j\in N} v_j, \tag{4}$$

Table 2. Analytic framework for Section 2.1. Two SNP-pairs (where each w_i, v_j denotes the probability of the respective genotype combination) and four penetrance levels ($p_{k\ell}$, $k, \ell \in \{0, 1\}$). Certain relationships among the four penetrance parameters, e.g., Eq (4), can make it impossible for us to determine an appropriate disease model for the underlying pair based on the case-control ratios.

Pair 1				Pair 2				Penetrance		
	BB	**Bb**	**bb**		**DD**	**Dd**	**dd**	**Pair$_1$\Pair$_2$**	**R**	**N**
AA	w_1	w_2	w_3	CC	v_1	v_2	v_3	R	p_{11}	p_{10}
Aa	w_4	w_5	w_6	Cc	v_4	v_5	v_6	N	p_{01}	p_{00}
aa	w_7	w_8	w_9	cc	v_7	v_8	v_9			

https://doi.org/10.1371/journal.pone.0213236.t002

Table 3. Simulated examples for Section 2.1. Disease models for the two pairs of SNPs that contribute to the simulated outcome. The penetrance parameters, $(p_{00}, p_{01}, p_{10}, p_{11})$, are chosen so that the case-control ratio is the same for all genotype combinations $i = 1, 2, \ldots, 9$ in the first pair, {A/a, B/b}.

Ex. 1: Two SNP-pairs, identical disease models (MAF = 0.3).

	BB	Bb	bb		DD	Dd	dd
AA	0	0	1	CC	0	0	1
Aa	0	1	0	Cc	0	1	0
aa	1	0	0	cc	1	0	0

$(p_{10} = 0.1, p_{01} = 0.28, p_{00} = 0.01 \overset{\text{Eq (4)}}{\Rightarrow} p_{11} = 0.03.)$

Ex. 2: Two SNP-pairs, different disease models (MAF = 0.2).

	BB	Bb	bb		DD	Dd	dd
AA	0	0	0	CC	0	1	1
Aa	0	1	1	Cc	1	0	0
aa	0	1	1	cc	1	0	0

$(p_{10} = 0.09, p_{01} = 0.12, p_{00} = 0.001 \overset{\text{Eq (4)}}{\Rightarrow} p_{11} = 0.016.)$

https://doi.org/10.1371/journal.pone.0213236.t003

the case-control ratio will be the same for all genotype combinations $i = 1, 2, \ldots, 9$ in the first pair, regardless of whether $i \in R$ or $i \in N$. It is thus a pathological case, in which it would be impossible to rely on the case-control ratios to determine the disease model.

Since both MDR and RS rely on the case-control ratios to determine disease models, we can expect their powers (of detecting the relevant pair) to be greatly affected if Eq (4) holds, *even if only approximately*.

To offer a more concrete illustration, we simulated two examples (see Table 3). In the first one, the true disease models were the same for the two relevant SNP-pairs; in the second, they were different. The penetrance parameters p_{10}, p_{01} and p_{00} were predetermined, and we explored a few different values for the last penetrance parameter, p_{11}, around the value implied by Eq (4). Keep in mind, however, that unless p_{11} is equal to the value implied by Eq (4) *exactly*, there is still some weak signal left that is, in principle, detectable by considering the case-control ratios. The simulation was repeated for 100 times, with a total of 100 SNPs and a sample size of $n = 800$.

We then assessed the number of times each relevant pair was successfully detected by each method with the sequential screening procedure, out of 100 repetitions. A relevant pair was considered to have been successfully detected if it was among the top two pairs selected by the method. Here, the effect of the second pair ({C/c, D/d}) was stronger than the first ({A/a, B/b}) —e.g., $p_{01} > p_{10}$—and all three methods detected it perfectly (i.e., 100 times out of 100 replications), but as the parameter p_{11} dropped, both MDR and RS started to deteriorate in terms of their ability to detect the first pair ({A/a, B/b}), whereas our method, PTY, remained largely unaffected (Fig 1).

To better understand Eq (4), notice that it can be rearranged slightly as

$$p_{11} = p_{01} - \frac{\sum_{j \in R} v_j}{\sum_{j \in N} v_j} (p_{10} - p_{00}). \qquad (5)$$

Since we typically expect $p_{10} > p_{00}$, i.e., having a risky combination in the first pair will increase the probability of having the disease, Eq (5) implies that $p_{11} < p_{01}$, or that having risky combinations from *both* pairs will actually lead to a *lower* probability of having the disease than having risky combinations only from the second pair. This is analogous to the logical operator, "exclusive or" (XOR).

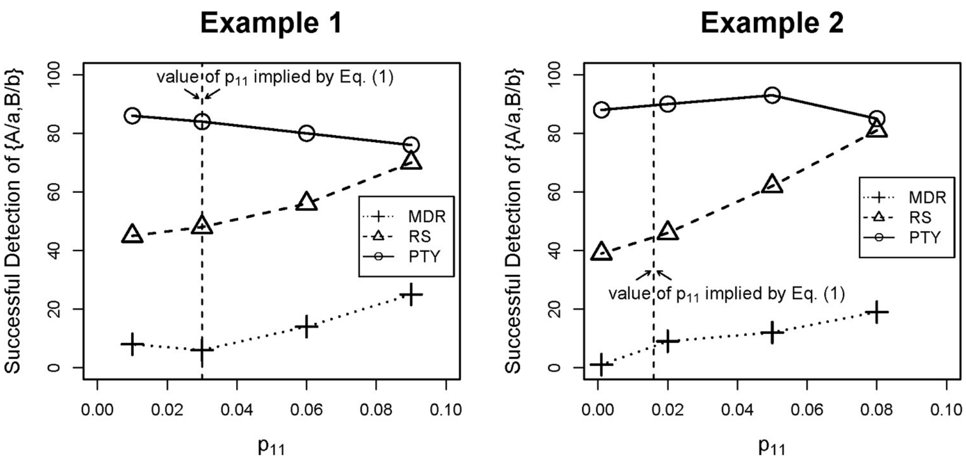


Fig 1. Simulated examples for Section 2.1. Number of times the first pair, {A/a, B/b}, was successfully detected (out of 100 repetitions) as the parameter p_{11} varied.

https://doi.org/10.1371/journal.pone.0213236.g001

While one certainly can argue that this may be a totally hypothetical scenario that is not likely to occur in nature, it is nonetheless a theoretical possibility against which our method, PTY, is robust.

Of course, the aforementioned XOR-type relationship means the two pairs, {A/a, B/b} and {C/c, D/d}, are interacting with each other, so there is actually a four-way interaction across the four SNPs involved. Such a high-order interaction still could be detectable by methods such as the MDR or the RS if four-way disease models were considered and screened but, as we stated earlier (Section 1), in this study we are taking a "narrow" point of view by restricting ourselves to consider only two-way interactions. Indeed, there is nothing "pathological" about having a high-order interaction; it is only "pathological" when one is restricted to consider only two-way interactions.

2.2 Detection of spurious effects

We have also observed that being overly adaptive to data can cause a method to be more easily tricked into detecting spurious epistatic effects, e.g., by SNPs with large individual effects. To demonstrate this, we simulated 100 SNPs on a case-control sample of size $n = 200$. Two pairs of SNPs—say, {A/a, B/b} and {C/c, D/d}—contributed to the simulated outcome independently, each according to an additive disease model (see Table 4). The SNPs A/a and C/c were simulated to have higher minor allele frequencies (MAFs) than B/c and D/d so that, according to the underlying additive disease model, they had larger marginal, individual effects than the other two.

We repeated the simulation for 100 times and counted those pairs most frequently ranked by each method—with the sequential screening procedure—to be among the top two (Table 5).

Table 4. Simulated examples for Section 2.2. Disease models for the two pairs of SNPs that contribute to the simulated outcome. Numeric values (e.g., 0.1, 0.2) are penetrance parameters for the corresponding genotype combinations. (MAF = 0.5 and 0.3, respectively for the two SNPs in each pair).

	BB	Bb	bb		DD	Dd	dd
AA	0	0	0.1	CC	0	0	0.1
Aa	0	0	0.1	Cc	0	0	0.1
aa	0.1	0.1	0.2	cc	0.1	0.1	0.2

https://doi.org/10.1371/journal.pone.0213236.t004

Table 5. Simulated examples for Section 2.2. Number of times different pairs of SNPs were among the top two pairs detected, out of 100 replications. The truly relevant pairs are emboldened.

MDR		RS		PTY	
{A/a, C/c}	75	{A/a, C/c}	74	**{A/a, B/b}**	43
{B/b, D/d}	32	{B/b, D/d}	51	**{C/c, D/d}**	42
{C/c, D/d}	13	{A/a, D/d}	12	{A/a, C/c}	32
{A/a, D/d}	13	{B/b, C/c}	9	{A/a, D/d}	23
{A/a, B/b}	10	**{C/c, D/d}**	8	{B/b, C/c}	9
{B/b, C/c}	10	**{A/a, B/b}**	7	{B/b, D/d}	6
Other Pairs	≤ 9	Other Pairs	≤ 6	Other Pairs	≤ 5

https://doi.org/10.1371/journal.pone.0213236.t005

Both MDR and RS were more likely to select a spurious pair, {A/a, C/c}, due to the large marginal effects of both of these SNPs. They were much less effective than our method, PTY, in identifying the truly relevant pairs.

2.3 Exaggeration of effects and false positives

Earlier in Section 1, we already stated that both MDR and RS tend to produce many false positives. To demonstrate this point more concretely, we conducted another experiment. We simulated 100 SNPs on a case-control sample of size $n = 400$, except that, this time, *none* of the SNPs was related to the simulated outcome. We then allowed all three methods, MDR, RS, and PTY, to assess the resulting $\binom{100}{2} = 4,950$ pairs of SNPs and examined the distributional properties of the resulting association measures (see Section 1.1) produced by each method for all pairs, $\{\widehat{\chi}^2_{i,j} : 1 \leq i, j \leq 100\}$.

Fig 2 shows various Q-Q plots of these association measures, produced by different methods under different MAF settings, against the theoretical quantiles of the $\chi^2_{(1)}$-distribution. We can see that all methods produced *inflated* association measures, leading to false discoveries. This is not a big surprise; after all, $M_{i,j}$ was not just any disease model but the one deemed "best-fitting" for the underlying pair (i, j). Though the meaning of "best-fitting" differed for the three methods, a post-hoc test of independence based on $M_{i,j}$ was clearly biased toward being significant. This is why we used the adjective "nominal" earlier in Section 1.1 to describe these association measures.

However, the main point here is that our method, PTY, suffered the *least* from this tendency to produce false positives. As the MAF increased, the tendency to produce false positives also became more pronounced for both MDR and RS, but not for PTY. To further make this point, we repeated the aforementioned "null simulation" 400 times. For each repetition, we computed the mean value of the (nominal) association measure,

$$\frac{1}{4950}\sum_{i,j}\widehat{\chi}^2_{i,j}, \qquad (6)$$

across all 4,950 SNP-pairs. The average of these mean values and its standard error over the 400 repetitions are shown in Table 6 for each method under different MAF settings. Clearly, this value is more inflated for MDR and especially for RS than it is for PTY.

3 Method

We now describe our approach in more detail. First, we derive a metric to measure the similarity (or equivalently, difference) between two disease models. Then, we cluster all disease

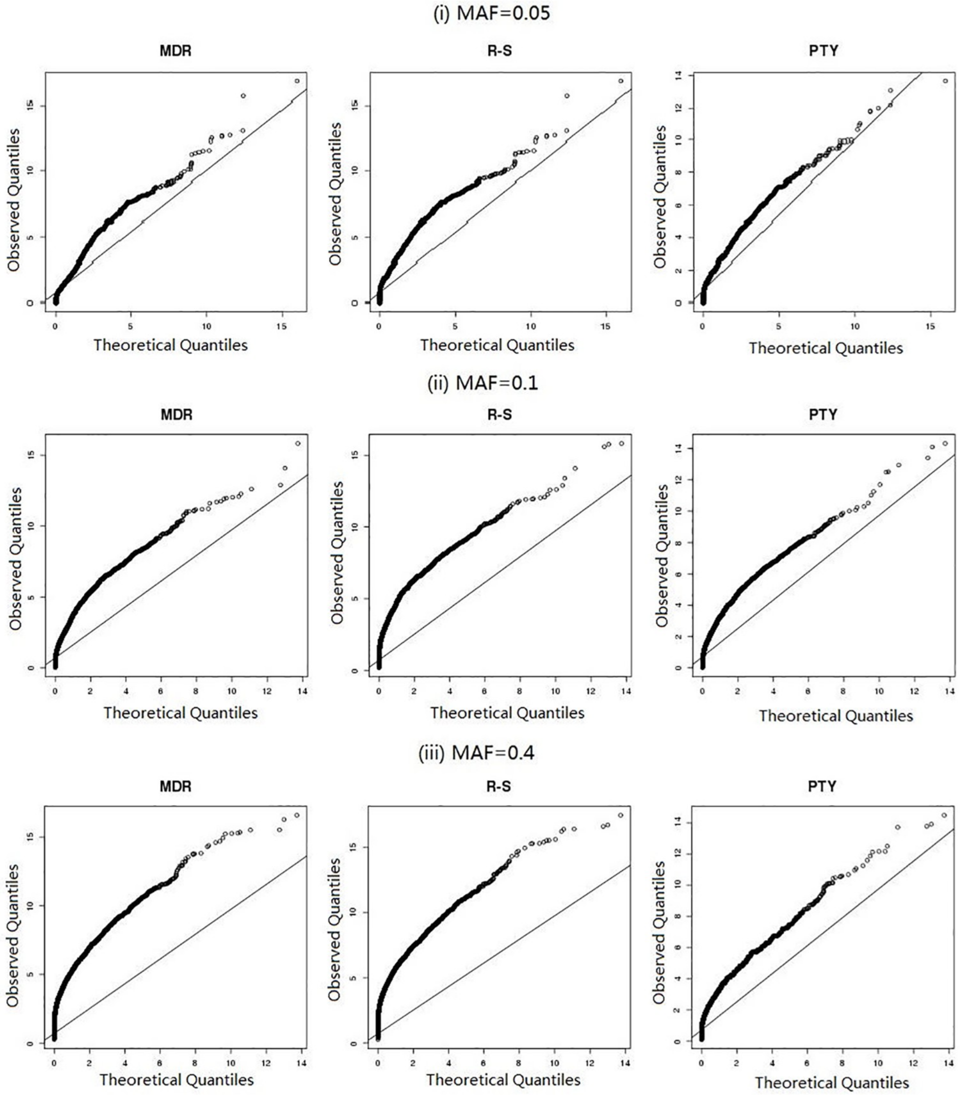


Fig 2. Results from simulation study (Section 2.3). Q-Q plots of nominal association measures $\{\widehat{\chi}^2_{i,j} : 1 \leq i, j \leq 100\}$ against their theoretical quantiles.

https://doi.org/10.1371/journal.pone.0213236.g002

models into a few groups and select a prototype model from each group. Finally, we screen each pair of SNPs against the set of prototype models. The set of prototype models is decided *a priori*, without considering the disease status of individuals in the data set. This is what makes our approach less greedy, and less data-adaptive, than existing methods such as MDR and RS.

Table 6. Results from simulation study (Section 2.3). Average values of the nominal association measures $\{\widehat{\chi}^2_{i,j} : 1 \leq i, j \leq 100\}$ across all 4,950 SNP-pairs, together with their standard errors, over 400 repetitions.

MAF	MDR	RS	PTY
0.05	1.837 (0.229)	2.115 (0.233)	1.792 (0.214)
0.10	2.457 (0.239)	2.975 (0.230)	2.202 (0.200)
0.40	4.748 (0.289)	5.101 (0.303)	3.033 (0.221)

https://doi.org/10.1371/journal.pone.0213236.t006

3.1 Similarity measure

Earlier in Section 1, we already alluded to the intuition that some disease models appear to be more similar than others (Table 1). Such intuition can be formalized in many different ways; for instance, some researchers have used a geometric approach to categorize them [17]. In this paper, we take a more pragmatic approach.

We measure the similarity of two disease models—say, M and M'—according to how much they agree in terms of their assignment of individuals into high- and low-risk groups. For $k, \ell = \{0, 1\}$, suppose $n_{k\ell}$ is the number of individuals classified to be high-risk by both models ($k = \ell = 1$), by M only ($k = 0, \ell = 1$), by M' only ($k = 1, \ell = 0$), or by neither model ($k = \ell = 0$), out of a *hypothetical* group of $n_{\cdot\cdot}$ individuals (Table 7). We then use the so-called Φ-coefficient [22], defined as

$$\Phi = \frac{(n_{11})(n_{00}) - (n_{10})(n_{01})}{\sqrt{(n_{1\cdot})(n_{0\cdot})(n_{\cdot 1})(n_{\cdot 0})}} \tag{7}$$

to measure the concordance between M and M'. A high (low) value of Φ means the two models classify many (few) individuals to be in the same high- or low-risk group.

For $i = 1, 2, \ldots, 9$, let G_i denote a genotype combination formed by a pair of SNPs; and let $\mathbb{P}(D|G_i)$ denote the penetrance (or probability of trait/disease) of the particular combination G_i. Suppose that M is the true disease model with only two unique penetrance levels,

$$\mathbb{P}(D|G_i) = \begin{cases} P_1, & M(G_i) = 1, \\ P_0, & M(G_i) = 0; \end{cases} \tag{8}$$

whereas M' is a different disease model used in place of the true model M. Then, derivations contained in S2 Appendix show that the Φ-coefficient between M and M' can be expressed as

$$\Phi(M', M) = \frac{(W_{11})(W_{00}) - (W_{10})(W_{01})}{\sqrt{\left(\frac{U}{V}W_{11} + W_{01}\right)\left(W_{10} + \frac{V}{U}W_{00}\right)(W_{1\cdot})(W_{0\cdot})}}, \tag{9}$$

where

$$W_{k\ell} = \sum_{\substack{M(G_i) = k \\ M'(G_i) = \ell}} \mathbb{P}(G_i) \quad \text{for} \quad k, \ell \in \{0, 1\}; \tag{10}$$

$$U = rP_1[1 - \mathbb{P}(D)] + (1 - P_1)\mathbb{P}(D); \tag{11}$$

$$V = rP_0[1 - \mathbb{P}(D)] + (1 - P_0)\mathbb{P}(D); \tag{12}$$

r is the case-control ratio of the sample; and $\mathbb{P}(D)$ is the prevalence of the trait/disease.

Table 7. Assignment of individuals into high- and low-risk groups by two disease models, M and M'.

$M' \backslash M$	High Risk	Low Risk	Total
High Risk	n_{11}	n_{10}	$n_{1\cdot}$
Low Risk	n_{01}	n_{00}	$n_{0\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	$n_{\cdot\cdot}$

https://doi.org/10.1371/journal.pone.0213236.t007

Eq (9) allows us to use

$$d(M', M) = 1 - \Phi(M', M) \tag{13}$$

as a distance metric for two disease models. It is important to note here that our distance metric $d(M', M)$ is not symmetric but directional. In particular, M is assumed to be the true model, and M' is the prototype used in its place. We shall say more about the similarity measure, $\Phi(M', M)$, later in Section 6; here, we first give some details about how values of various parameters can be obtained in order to compute the expression on the right-hand side of Eq (9).

$W_{k\ell}$: Assuming Hardy-Weinberg equilibrium, the MAFs of the two SNPs can be estimated from the control sample, and used to determine $\mathbb{P}(G_i)$ for each genotype combination G_i and hence $W_{k\ell}$ as well for $k, \ell \in \{0, 1\}$.

r: For any given data set, the case-control ratio r is known, e.g., $r = 1$ for a balanced case-control data set.

$\mathbb{P}(D)$: The prevalence, $\mathbb{P}(D)$, of a particular trait/disease can often be obtained from external sources, e.g., published studies and/or expert opinions. (More on this below in Sections 3.2 and 6).

P_1, P_0: To determine the value of these parameters, we make a convenient assumption that the underlying pair of SNPs is the actual pair associated with the outcome. Then, the prevalence is simply

$$\mathbb{P}(D) = \sum_{i=1}^{9} \mathbb{P}(D|G_i)\mathbb{P}(G_i) \tag{14}$$

and the heritability (the amount of genetic contribution to overall phenotype variation [23]) is given by

$$h^2 = \frac{1}{[\mathbb{P}(D)][1 - \mathbb{P}(D)]} \sum_{i=1}^{9} [\mathbb{P}(D|G_i) - \mathbb{P}(D)]^2 \mathbb{P}(G_i). \tag{15}$$

Since we have assumed in Eq (8) that M has only two unique penetrance levels, i.e., each $\mathbb{P}(D|G_i)$ is either P_1 and P_0, they can now be uniquely determined from the two Eqs (14) and (15), provided that information is available about the heritability parameter, h^2. This can often be obtained from external sources as well, much like the prevalence parameter. (More on this below in Sections 3.2 and 6).

3.2 Clustering

There are altogether $2^9 - 2 = 510$ non-trivial TTTC disease models—the trivial ones are those such that $M(G_i) = 1$ or $M(G_i) = 0$ for all G_i. For clustering purposes, we need not consider disease models that are symmetric with respect to (i) the exchange of locus, i.e., swapping the two SNPs, or (ii) the exchange of disease status, i.e., flipping the binary values of each $M(G_i)$ from a zero to a one, and vice versa. The set of models that remain, which we denote as $\mathcal{M}$, is listed in S3 Appendix.

Fig 3 shows the 2-dimensional coordinates of all models $\in \mathcal{M}$ as estimated by the multidimensional scaling (MDS) technique from their pairwise distances, assuming that the MAFs of both SNPs are equal to 0.1, 0.2, 0.3, and 0.4, respectively, while fixing the prevalence and heritability parameters at $\mathbb{P}(D) = h^2 = 0.02$. While we used the directional distance metric for prototype identification (see Table 8 below), we used a symmetrized distance metric, $d_s(M_i, M_j) \equiv [d(M_i, M_j) + d(M_j, M_i)]/2$, for performing MDS so that the resulting

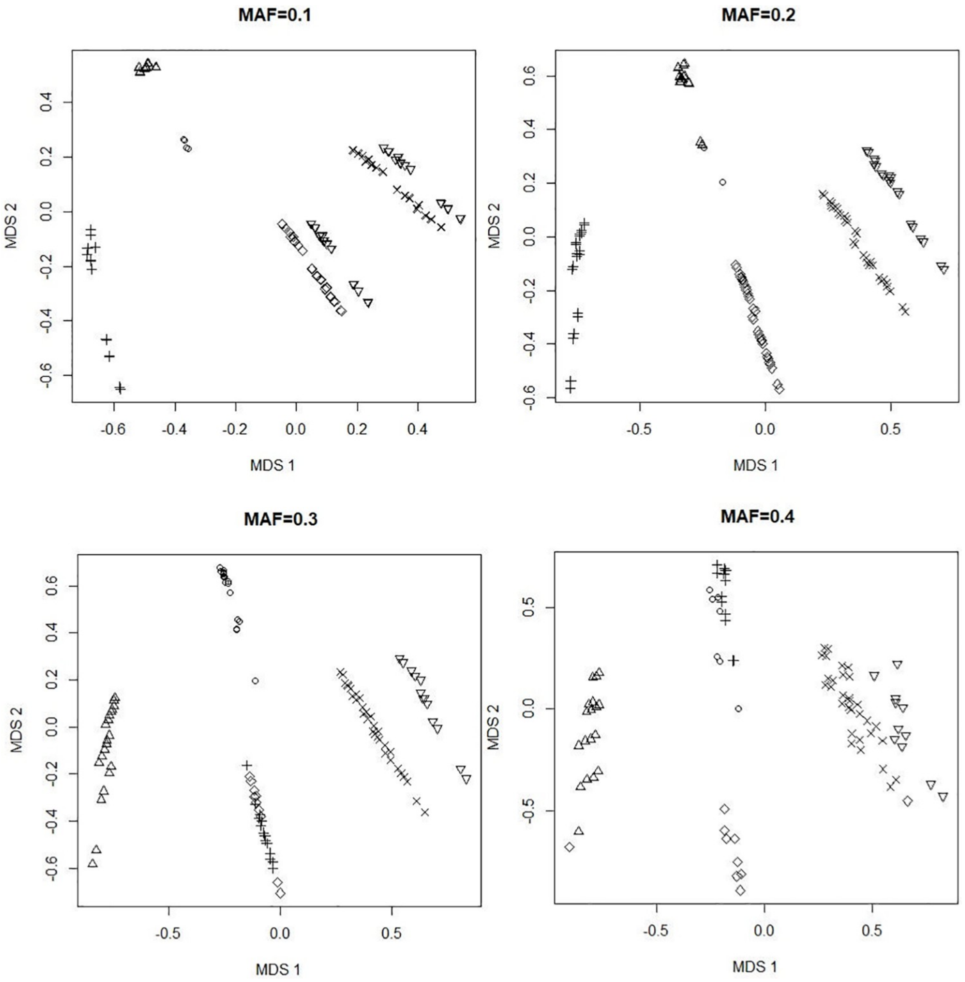


Fig 3. A two-dimensional map of disease models in $\mathcal{M}$. The coordinates are estimated by applying the multi-dimensional scaling (MDS) technique to the symmetrized pairwise distances, $d_s(M_i, M_j) \equiv [d(M_i, M_j) + d(M_j, M_i)]/2$, for all $i \neq j$. Models clustered into the same group are depicted by the same symbol (e.g., '+', '○', '×'). These two-dimensional coordinates explain about 50-70% of the variation in $d_s(\cdot, \cdot)$, so there is some loss of information—in particular, some disease models may be closer to (or farther apart from) each other than how they appear in this map.

https://doi.org/10.1371/journal.pone.0213236.g003

2-dimensional coordinate-map (Fig 3) is more meaningful. It is clear from Fig 3 that these disease models form several clusters.

One can easily expect from Eq (9) that our distance metric will be affected by the MAFs of the underlying SNPs, but Fig 3 shows that the resulting clusters do not change significantly. Therefore, it is not necessary to repeat the prototype selection step for every individual SNP-pair. Instead, we simply discretized the MAF-scale into 6 bins, {0.05, 0.1, 0.2, 0.3, 0.4, 0.45}, and created 36 different *sets* of prototypes for all 6 × 6 pairwise combinations. For example, when screening a SNP-pair (i, j) with $(\text{MAF}_i, \text{MAF}_j) = (0.068, 0.182)$, we would use the set of prototypes for $(\text{MAF}_i, \text{MAF}_j) = (0.05, 0.2)$, and so on.

We also examined similar plots (not shown) produced with different values of $\mathbb{P}(D)$ and h^2. While these parameters also affected the distance metric, they did not produce any substantial changes to the clustering. Intuitively, this is because there has to be a fairly drastic warping of the relative distances between objects in order to alter their grouping; we shall come back to this point again later in Section 6. Hence, for this paper we simply used $\mathbb{P}(D) = h^2 = 0.02$.

Table 8. The global *K*-means algorithm for identifying prototype disease models.

1. Let $\mathcal{M}$ be the set of all disease models and $\mathcal{M}^*$, the prototype set (initially empty).
2. Evaluate each $M_i \in \mathcal{M} \setminus \mathcal{M}^*$ as a potential new prototype, as follows:
   a. For each $M_k \in \mathcal{M} \setminus \{\mathcal{M}^* \cup M_i\}$, calculate the distances $d(M_i, M_k)$, and $d(M_j^*, M_k)$ for all $M_j^* \in \mathcal{M}^*$ if $\mathcal{M}^*$ is not empty.
   b. Assign M_k eitder to an existing cluster—e.g., C_j^*, with center M_j^*— or to a potentially new cluster—say C_i, with center M_i—depending on which of $d(M_i, M_k)$ and $d(M_j^*, M_k)$ is the shortest.
   c. After all $M_k \in \mathcal{M} \setminus \{\mathcal{M}^* \cup M_i\}$ are assigned, calculate the total within-cluster distances,

$$D(M_i) \equiv \sum_{M_k \in C_i} d(M_i, M_k) + \sum_{M_j^* \in \mathcal{M}^*} \sum_{M_k \in C_j^*} d(M_j^*, M_k),$$

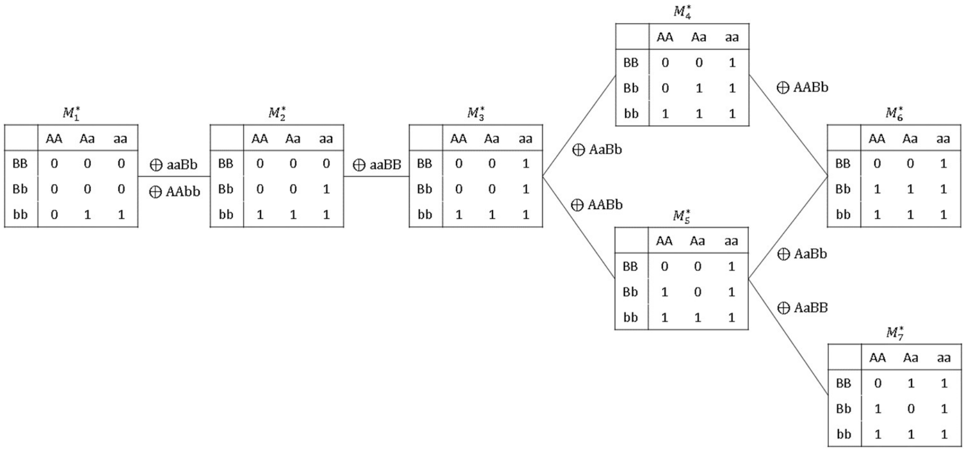
   as a result of using M_i as an additional cluster center.
3. Identify a new prototype model as the one that minimizes the total within-cluster distances, i.e.,

$$M^* \quad = \quad \underset{M_i \in \mathcal{M} \setminus \mathcal{M}^*}{\arg\min} \quad D(M_i),$$

   and insert it into the set $\mathcal{M}^* \leftarrow \mathcal{M}^* \cup M^*$.
4. Repeat steps 2-3 until a certain number of prototypes are identified.

https://doi.org/10.1371/journal.pone.0213236.t008

In principle, we could use any distance-based clustering algorithm. In our implementation, we used the "global *K*-means" algorithm [24]. The steps of our algorithm are given in Table 8. Based on Fig 3, we selected 7 prototypes for each MAF-combination. As an illustration, the prototypes for SNP-pairs with (MAF_i, MAF_j) = (0.2, 0.2) are displayed in Fig 4 with manual annotations to reveal their relationships with one another. One may interpret this figure to mean that, for a pair of SNPs, both of which have MAF around 0.2, these are the primary epistatic effects to consider, and their structural relationships; any other will likely be very similar to one of these—in terms of how they would classify individuals into high- versus low-risk groups, that is. This is also a unique piece of insight from our overall methodology that is not otherwise available from MDR or RS.



M_1^*

	AA	Aa	aa
BB	0	0	0
Bb	0	0	0
bb	0	1	1

M_2^*

	AA	Aa	aa
BB	0	0	0
Bb	0	0	1
bb	1	1	1

M_3^*

	AA	Aa	aa
BB	0	0	1
Bb	0	0	1
bb	1	1	1

M_4^*

	AA	Aa	aa
BB	0	0	1
Bb	0	1	1
bb	1	1	1

M_5^*

	AA	Aa	aa
BB	0	0	1
Bb	1	0	1
bb	1	1	1

M_6^*

	AA	Aa	aa
BB	0	0	1
Bb	1	1	1
bb	1	1	1

M_7^*

	AA	Aa	aa
BB	0	1	1
Bb	1	0	1
bb	1	1	1

Fig 4. The set of prototype disease models selected by the global *K*-means algorithm ($K = 7$) for SNP-pairs (*i*, *j*) with (MAF_i, MAF_j) ≈ (0.2, 0.2). The structural relationships between the seven prototypes were manually annotated; the clustering algorithm itself was not capable of making this type of discoveries.

https://doi.org/10.1371/journal.pone.0213236.g004

Finally, it is worth emphasizing that the reduction of disease models to the set $\mathcal{M}$, due to various symmetry considerations we mentioned in the opening paragraph of this subsection, is *only* applicable to the clustering and prototype selection stage. When screening each candidate SNP-pair, prototype disease models that are asymmetric with respect to the exchange of locus, such as M_1^* in Fig 4, are always tested both for {A/a,B/b} and for {B/b,A/a}, and so on.

4 Simulation study

To motivate our approach, we already presented a few simulated examples in Section 2, where we concentrated on evidence that our approach appears to overcome various weaknesses of existing approaches. In this section, we assess our approach more generally with a number of simulated examples that are commonly examined in the literature.

4.1 Set-up

In each simulation, we generated 100 SNPs, but only the first two determined the simulated outcome according to a particular disease model (more details below in Section 4.2). To evaluate the performance of a method, we used a metric known as the F-measure, defined as

$$\text{F-measure} \equiv \frac{2 \times (\text{precision}) \times (\text{recall})}{(\text{precision}) + (\text{recall})}, \tag{16}$$

where

$$\text{precision} = \begin{cases} \dfrac{1}{\#\ (\text{pairs detected})}, & \text{if the true pair was detected,} \\ 0, & \text{otherwise;} \end{cases} \tag{17}$$

and

$$\text{recall} = \begin{cases} 1, & \text{if the true pair was detected,} \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Each simulation was repeated for 400 times, and the average F-measure and its standard error were recorded (Table 11). To avoid excessive computation, we used the marginal screening procedure for all methods; see Section 1.1.

The F-measure is a widely used criterion in the field of information retrieval; it is a single numeric metric that balances the trade-off between true positives and false positives. We adopted the F-measure, instead of other metrics such as the "balanced accuracy", because the underlying problem really is more of an "information retrieval" problem than a "classification" problem, not only because there are far more true negatives than true positives, but also because detecting the positives—here, the relevant SNP-pair—is a much more important objective than correctly calling out the negatives. Imagine the experience of conducting a Google search. For each given query, most of the web pages on the Internet are irrelevant. Therefore, from a customer's perspective, the most important measure must concentrate on the set of detected web pages retrieved by the search engine, for example, the top twenty. How many of these are relevant (true positives), and how many are irrelevant (false positives)? By and large, the customer does *not* care how many truly irrelevant web pages have been correctly left out of the search result—that is, the customer does not care about the true negatives. Moreover, because the set of truly irrelevant pages is so large, the true negative rate will also be difficult to distinguish meaningfully for most "reasonable" search engines; any "reasonable" search engine will have a true negative rate of >99%. Therefore, measures like the “balanced accuracy” actually places an undue amount of emphasis on this rather inconsequential side of the performance. This is also why the information retrieval community tend to largely favor metrics such as the F-measure to those more commonly used by the classification community, such as “balanced accuracy”. The situation of detecting relevant SNP-pairs is very much akin to performing a Google search in that (i) most pairs are not signals; (ii) we care *not very much* about getting the true negatives right; (iii) instead, we care *a lot* about how many of the detected pairs are true positives or false positives.

4.2 Disease models

Our primary focus was to evaluate the ability of different methods to detect different epistatic effects as represented by different disease models.

First, we included six disease models with main effects (Table 9). They were among the most commonly used examples in various studies [25–30]. Here in Table 9, these models are parameterized in terms of odds, $\mathbb{P}(D|G_i)/[1-\mathbb{P}(D|G_i)]$, rather than penetrance, $\mathbb{P}(D|G_i)$. The parameters α and θ were determined by simultaneously solving Eqs (14) and (15), given the prevalence $\mathbb{P}(D)$ and heritability h^2 of the outcome, as well as the MAF of each SNP. We simply fixed $\mathbb{P}(D) = 0.02$, but repeated each of these simulations with MAF = 0.1 and 0.4 for all SNPs. Assuming Hardy-Weinberg equilibrium, the MAF determined $\mathbb{P}(G_i)$ for each genotype combination G_i, leaving α and θ to be the only unknowns in Eqs (14) and (15) so that they could be uniquely determined.

Next, we included four disease models without main effects (Table 10), taken from an earlier study conducted by Ritchie *et al.* [31], in which these disease models were created to have purely epistatic effects in the sense that no marginal effect existed for either SNP involved.

The disease models, T, DD, MOD and XOR, all have two penetrance levels (Table 9), and so do our prototype disease models (see Fig 4). However, in designing our simulations we took care to ensure that, while some of these models (e.g., XOR) were relatively close to a prototype,

Table 9. Simulated examples for Section 4. Disease models with main effects. The parameters α and θ were uniquely determined given prevalence $\mathbb{P}(D)$, heritability h^2, and MAF. We fixed $\mathbb{P}(D) = 0.02$, and repeated each simulation with MAF = 0.1 and 0.4 for all SNPs.

(a) Threhold (T) h^2 = 0.02				(b) Dominant-Dominant (DD) h^2 = 0.02			
	BB	Bb	bb		BB	Bb	bb
AA	α	α	α	AA	α	α	α
Aa	α	α	$\alpha(1+\theta)$	Aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$	aa	α	$\alpha(1+\theta)$	$\alpha(1+\theta)$
(c) Modifying Effect (MOD) h^2 = 0.02				(d) Exclusive Or (XOR) h^2 = 0.02			
	BB	Bb	bb		BB	Bb	bb
AA	α	α	α	AA	α	α	$\alpha(1+\theta)$
Aa	α	α	$\alpha(1+\theta)$	Aa	α	α	$\alpha(1+\theta)$
aa	$\alpha(1+\theta)$	$\alpha(1+\theta)$	$\alpha(1+\theta)$	aa	$\alpha(1+\theta)$	$\alpha(1+\theta)$	α
(e) Multiplicative (ME) h^2 = 0.015				(f) Threshold Multiplicative (MET) h^2 = 0.015			
	BB	Bb	bb		BB	Bb	bb
AA	α	$\alpha\ (1+\theta)$	$\alpha(1+\theta)^2$	AA	α	α	α
Aa	$\alpha\ (1+\theta)$	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$	Aa	α	$\alpha\ (1+\theta)$	$\alpha(1+\theta)^2$
aa	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^3$	$\alpha(1+\theta)^4$	aa	α	$\alpha(1+\theta)^2$	$\alpha(1+\theta)^4$

https://doi.org/10.1371/journal.pone.0213236.t009

Table 10. Simulated examples for Section 4. Disease models without main effects, taken from [31], where they were specifically constructed in such a way that there is no individual association between either SNP and the disease.

(a) DMN 1 MAF = 0.25, h^2 = 0.016				(b) DMN 2 MAF = 0.25, h^2 = 0.04			
	BB	Bb	bb		BB	Bb	bb
AA	0.08	0.07	0.05	AA	0	0.1	0.09
Aa	0.1	0	0.1	Aa	0.04	0.01	0.08
aa	0.03	0.1	0.04	aa	0.07	0.09	0.03
(c) DMN 3 MAF = 0.1, h^2 = 0.002				(d) DMN 4 MAF = 0.1, h^2 = 0.015			
	BB	Bb	bb		BB	Bb	bb
AA	0.07	0.05	0.02	AA	0.09	0.001	0.02
Aa	0.05	0.09	0.01	Aa	0.08	0.07	0.005
aa	0.02	0.01	0.03	aa	0.003	0.007	0.02

https://doi.org/10.1371/journal.pone.0213236.t010

others (e.g., DD) were relatively far from all prototypes, as measured by our metric Φ. The disease models, ME, MET, and DMN 1-4, on the other hand, all have more than two penetrance levels (Tables 9 and 10). They were chosen so that a wider variety of epistatic effects could be studied.

4.3 Thresholds

The nominal association measures produced by different methods for each pair of SNPs (see Section 1.1) were thresholded by their corresponding (nominal) p-values,

$$\widehat{p}_{i,j} \equiv \mathbb{P}(\chi^2_{(1)} > \widehat{\chi}^2_{i,j}), \tag{19}$$

and a pair was considered "detected" if $\widehat{p}_{i,j} < \alpha$, where α was a significance threshold. For convenience, we applied simple Bonferroni corrections to determine the threshold α. As there were a total of $\binom{100}{2} = 4,950$ pairs of SNPs, it was natural to first consider a threshold of

$$\alpha^{\text{easy}} = 0.05 \div 4,950 \approx 10^{-5}. \tag{20}$$

To account for the fact that these nominal association measures were biased (see Section 2.3), however, we also considered applying a more stringent threshold. But since there was not a clear way to pick such a threshold that easily could be considered "fair" for all methods, as each method considers a different number of (almost certainly) correlated disease models for a pair of SNPs, we simply settled on a convenient choice of

$$\alpha^{\text{hard}} = 0.05 \div 4,950 \div 8 \approx 1.26 \times 10^{-6}, \tag{21}$$

based on the fact that RS would always consider 8 different disease models. Correcting significance thresholds for simultaneous tests of correlated hypotheses is an intricate inferential problem, for which there is no good solution yet. It is not clear whether α^{hard} is really the "correct" threshold for RS but, as a *rough* guideline, one may think that this choice would favor RS slightly. Our empirical results below do support this interpretation to some extent.

4.4 Results

Results are given in Table 11. We used a relatively large sample size of *n* = 600 when the MAF was relatively low (e.g., 0.1, 0.25), and a relatively small sample size of *n* = 300 when it was high (e.g., 0.4). This is because, when the MAF was relatively high (low), the underlying signals

Table 11. Results from simulation study (Section 4). Average F-measures (and their standard errors) over 400 replications. A star (*) in front of the number indicates the best performer for that simulation.

			$\alpha^{\text{easy}} = 1.00 \times 10^{-5}$						$\alpha^{\text{hard}} = 1.26 \times 10^{-6}$					
n	**MAF**	**Model**	**MDR**		**RS**		**PTY**		**MDR**		**RS**		**PTY**	
600	0.1	T	0.012	(0.005)	0.080	(0.013)	*0.083	(0.015)	0.000	(0.000)	*0.046	(0.012)	0.030	(0.010)
		MOD	0.062	(0.009)	0.067	(0.007)	*0.075	(0.010)	0.063	(0.011)	*0.109	(0.013)	0.090	(0.013)
		DD	0.183	(0.016)	0.172	(0.015)	*0.278	(0.019)	0.341	(0.023)	0.372	(0.021)	*0.449	(0.024)
		XOR	0.199	(0.014)	0.198	(0.013)	*0.328	(0.020)	0.283	(0.022)	0.449	(0.022)	*0.531	(0.024)
		ME	0.011	(0.000)	0.011	(0.000)	*0.012	(0.000)	0.013	(0.000)	0.013	(0.001)	*0.015	(0.001)
		MET	0.234	(0.019)	0.245	(0.016)	*0.294	(0.021)	0.335	(0.025)	*0.414	(0.023)	0.350	(0.024)
300	0.4	T	0.167	(0.011)	0.152	(0.010)	*0.223	(0.014)	*0.357	(0.017)	0.346	(0.017)	0.317	(0.018)
		MOD	0.076	(0.007)	0.064	(0.006)	*0.110	(0.010)	0.163	(0.013)	0.154	(0.013)	*0.202	(0.015)
		DD	0.015	(0.001)	0.014	(0.001)	*0.135	(0.009)	0.035	(0.005)	0.032	(0.005)	*0.276	(0.014)
		XOR	0.195	(0.012)	0.164	(0.011)	*0.306	(0.016)	0.441	(0.018)	0.409	(0.018)	*0.561	(0.019)
		ME	0.022	(0.002)	0.020	(0.002)	*0.033	(0.002)	0.043	(0.004)	0.039	(0.003)	*0.086	(0.006)
		MET	0.073	(0.006)	0.074	(0.007)	*0.076	(0.008)	0.132	(0.011)	*0.141	(0.011)	0.103	(0.010)
600	0.25	DMN 1	0.729	(0.015)	0.686	(0.016)	*0.935	(0.009)	0.951	(0.008)	0.939	(0.009)	*0.992	(0.003)
	0.25	DMN 2	0.743	(0.015)	0.705	(0.016)	*0.938	(0.010)	0.959	(0.007)	0.944	(0.008)	*0.972	(0.009)
	0.1	DMN 3	0.700	(0.018)	0.675	(0.018)	*0.832	(0.019)	0.822	(0.021)	*0.852	(0.019)	0.722	(0.026)
	0.1	DMN 4	0.752	(0.015)	0.720	(0.015)	*0.897	(0.014)	0.912	(0.013)	*0.921	(0.012)	0.831	(0.021)

https://doi.org/10.1371/journal.pone.0213236.t011

became stronger (weaker) and easier (harder) to detect, and all the methods would perform quite well (badly) if given a sample that was "too large (small)", making it difficult to tell them apart. For our simulated cases with 100 SNPs, we found that all methods essentially became indistinguishable when the sample size reached as low as $n = 100$ or as high as $n = 1000$.

As we explained previously, the threshold, α^{easy}, only includes a simple Bonferroni correction—here, for multiple testing of 4,950 pairs—and does not account for the fact that a method, whether MDR, RS, or PTY, has usually tested a few disease models already before testing the significance of the SNP-pair against the outcome. Strictly speaking, therefore, the Bonferroni correction alone is not enough, and often leads to inflated false positive rates. Among the three methods, PTY is the least prone to false positives, which explains why its performance is the best under α^{easy}. Generally speaking, our results confirm that there is some practical value to consider a less greedy and less data-adaptive procedure such as ours for epistasis detection.

4.5 Comments

Throughout our simulation study, we have assessed the performance of each screening method by its ability to detect the underlying SNP-pair, but not by whether the true disease model is correctly identified as well. The detection of the relevant SNP-pair is certainly the more fundamental task. Once the relevant SNP-pairs are identified, further studies can be conducted to determine the real underlying mechanism. Such an approach is certainly not unusual in the context of genome-wide association (GWA) studies. For most GWA studies in the literature, single SNPs are often tested and reported using disease models—e.g., additive, dominant, and so on—that are not necessarily the correct ones. Ascertaining the true disease model is almost never the goal of the initial GWA study; detecting the affected SNPs is.

In fact, this is also the very reason why our method works, because one need *not* always use exactly the true disease model in order to detect a pair of affected SNPs. While using a "very wrong" disease model can negatively affect the chances of detecting an affected SNP-pair, one has a good chance of making the detection as long as the disease models used for screening is "close enough" to the true one. Due to the way our prototype models are selected—i.e., as representative models from each cluster, there is a very good chance that at least one of our models is "close enough" to the true one.

5 Analysis of bipolar disorder data

In this section, we report our analysis of the phase I bipolar disorder data from the Wellcome Trust Case Control Consortium (WTCCC) [32]. Because our method is aimed at screening SNP-pairs for different epistatic effects (rather than individual SNPs for main effects), we focus on the *complementary value* that our method offers—in particular, its ability to find relevant SNPs that other methods may still miss.

The WTCCC project involves genotyping of 500K SNPs on humans of British ancestry. Bipolar disorder is one of seven diseases being studied by the WTCCC, and the shared control samples consist of 1, 500 individuals from the 1958 British Birth Cohort and another 1, 500 individuals selected from blood donors recruited as part of their project.

Identical-twin studies have shown that bipolar disorder has a strong genetic component [33]. Current findings from genome-wise association studies (GWAS) demonstrate that bipolar disorder shares many genetic overlaps with schizophrenia and other major depressive disorders, and that it is also characteristic of being polygenic, i.e., many variants that coalesce into functional pathways contribute to the disorder with small effects. The current understanding of its neurobiology is that changes in inflammatory cytokines, corticosteroids, neurotrophins, mitochondrial energy generation, oxidative stress, and neurogenesis are all involved in a comprehensive way to explain its various clinical features [34].

5.1 Pre-processing

We began by applying the same data quality control procedures as described in [32]—excluding SNPs with $> 5\%$ missing observations ($> 1\%$ for SNPs with MAF < 0.05), Hardy-Weinberg exact p-value $< 5.7 \times 10^{-7}$, p-value $< 5.7 \times 10^{-7}$ for either a one- or two degree-of-freedom test of association between the two control groups, and genome-wide heterozygosity $< 23\%$ or $> 30\%$; as well as samples with $> 3\%$ missing across all SNPs. In addition, we also filtered out SNPs with MAF $< 1\%$, p-value $< 10^{-7}$ for a univariate test of association, and p-value $< 10^{-5}$ for a test of Hardy-Weinberg equilibrium. The remaining data contained 1, 868 cases (individuals with bipolar disorder), 2, 938 controls, and 405, 524 SNPs. Eliminating "easily detectable" SNPs with "obvious" main effects is not uncommon for studies that focus on the detection of SNP-SNP interactions—for example, the paper by Wan *et al.* [15] that proposed the RS method also did this.

5.2 Mapping SNPs to genes

We used the marginal screening procedure (see Section 1.1) to screen and rank all pairs of SNPs. Here, we focus on the 100 unique SNPs appearing in the top 85 pairs (nominal p-value $< 10^{-11}$). We used the "Ensembl gene annotation system" [35] as well as SNPnexus [36] to map these SNPs to the genes in which they most likely reside. Altogether, we identified 75 genes in this manner.

Fig 5 shows the number of SNPs appearing in the top 85 pairs identified by PTY, MDR and RS, respectively. While 15 SNPs were identified by all three methods, 42 were identified by our method alone and they were mapped to 18 unique genes. Five of them—specifically, UNC13A [37], RGS6 [38], DPP10 [39], FGF14 [40] and TLE4 [40]—had been associated with bipolar disorder or related suicide attempts. Moreover, the SNP that was mapped to FGF14 had a p-

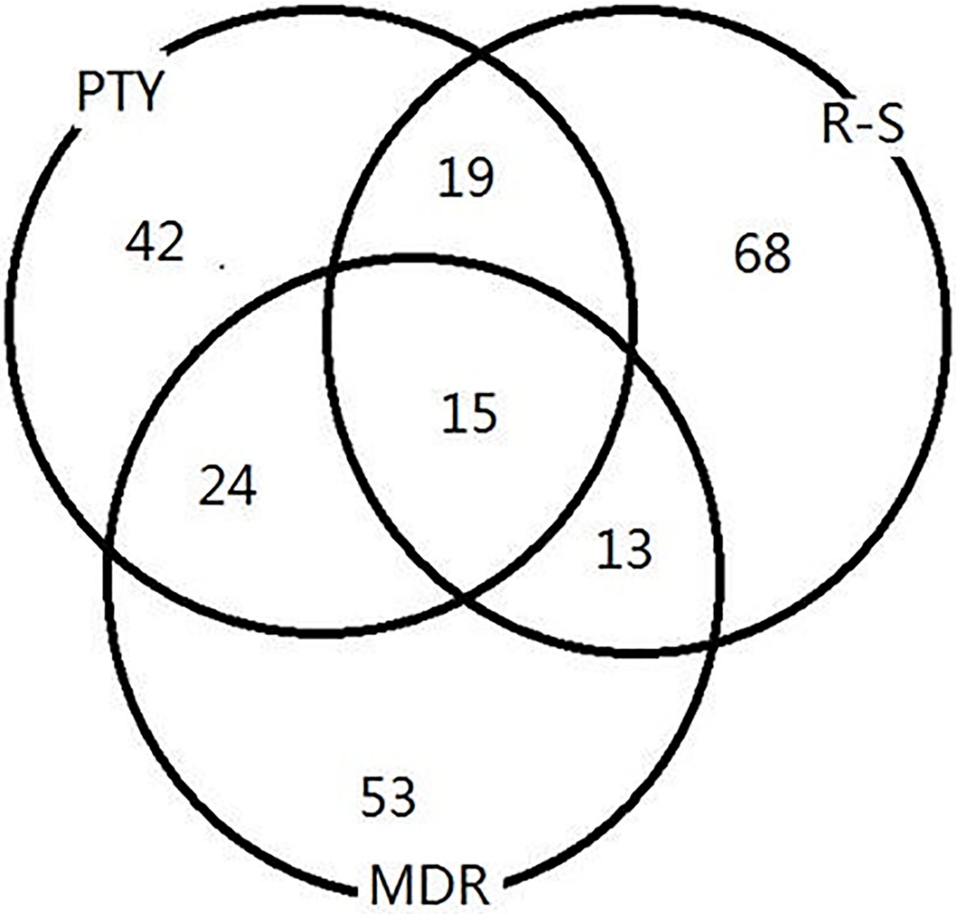


Fig 5. Analysis of bipolar disorder data. Venn diagram of unique SNPs appearing in the top 85 pairs detected by PTY, RS, and MDR, respectively. SNPs detected multiple times (e.g., occurring in multiple pairs) were counted only once.

https://doi.org/10.1371/journal.pone.0213236.g005

value of 0.03 on a *univariate* test of association, indicating that it would have had no chance of being detected in a genome-wide screening of individual SNPs. Here, it was detected as a result of pairwise screening that focused on epistatic effects.

Fig 6 shows the largest interaction network based on the 75 genes we identified. The hub gene, AQP1, encodes a small integral membrane protein that functions as a water channel protein and is potentially involved in a human neurological disorder called "central pontine myelinolysis" [41]. The specific SNP that was mapped to this gene (rs4299909) had a p-value of 0.0002 based on a univariate test of association; hence, it would have had no chance of being detected by marginal screening of individual SNPs, either. Here again, it was detected as a result of pairwise screening that focused on epistatic effects.

Among other genes in this network, ST6GALNAC5 is known to catalyse the transfer of sialic acid to cell surface proteins, and sialic acid has been suggested as an essential nutrient for brain development and cognition [42]. RGS6 regulates G protein signaling and may modulate neuronal activities; in previous studies, SNPs in this gene have been reported to be associated with schizophrenia [43]. MAN2A1 encodes a glycosyl hydrolase (a common enzyme) and catalyses the final hydrolytic step in the N-glycan maturation pathway; many SNPs in this gene have been reported to be associated with various phenotypes and diseases, including

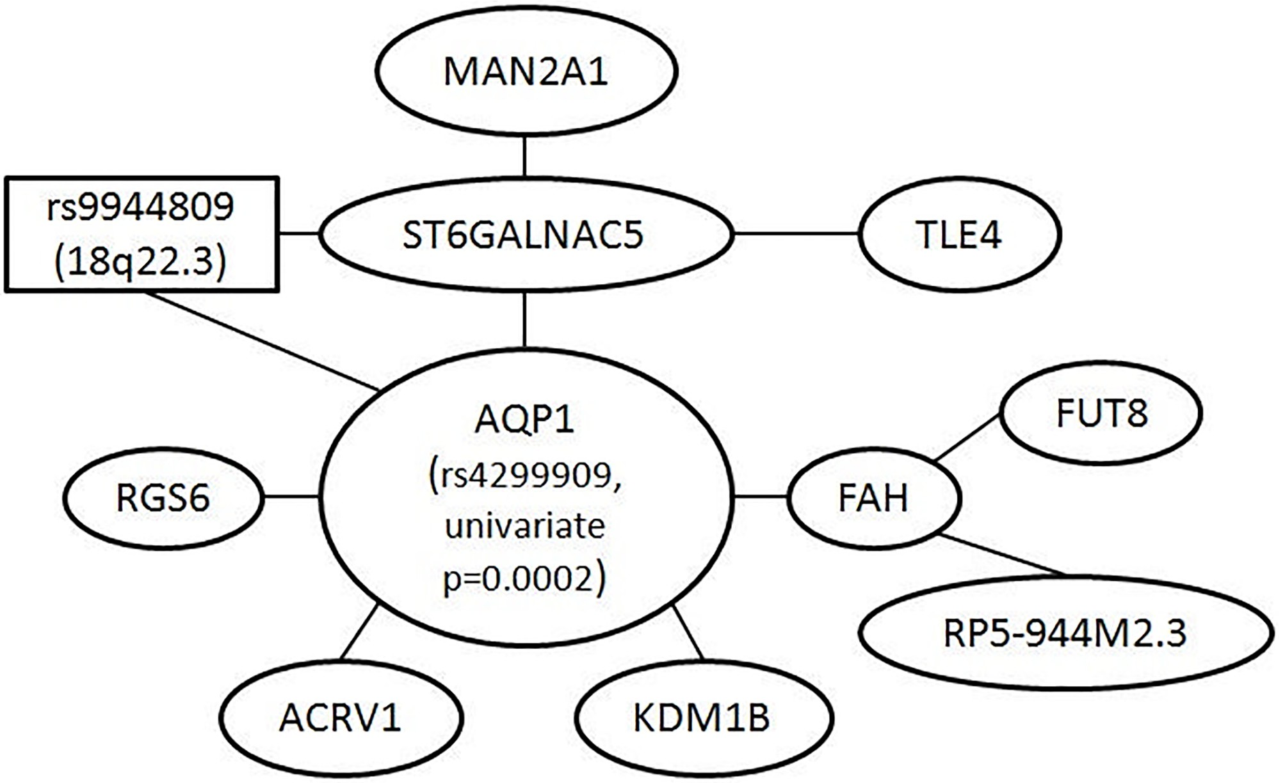


Fig 6. Analysis of bipolar disorder data. Largest interaction network formed by genes mapped from SNPs appearing in the top 85 pairs. Each node is either a gene (oval), or a SNP (rectangle) itself if it cannot be mapped to any gene. The size of the node is irrelevant—it is determined by the amount of text inside rather than anything scientific. A link between two nodes means the SNPs underlying the nodes are from the same pair detected, so, for example, a link between AQP1 and FAH means that a pair of SNPs—one of which was mapped to AQP1 and another, to FAH—was among the top 85 pairs detected. The resulting network contains many disjoint components. The one presented here is the biggest component.

https://doi.org/10.1371/journal.pone.0213236.g006

Alzheimer's disease [44, 45]. TLE4 inhibits the transcriptional activation mediated by PAX5, and by CTNNB1 and TCF family members in Wnt signaling, which has been suggested to be potentially involved in the pathophysiology of bipolar disorder [46]. FAH encodes the last enzyme in the tyrosine metabolism pathway; the amino acid, tyrosine, is a precursor to neurotransmitters and increases plasma neurotransmitter levels—particularly dopamine and norepinephrine, both important neurotransmitters in the brain [47]. FUT8 encodes an enzyme belonging to the family of fucosyltransferases; a variant in this gene has been reported to influence glutamate concentrations in brains of patients with multiple sclerosis [48]—glutamate is a neurotransmitter accounting in total for well over 90% of the synaptic connections in the human brain.

Out of the 75 genes we identified, the following have also been reported by various independent studies to be associated with bipolar disorder, or suicides related to bipolar disorder: ANK3 [49], CNTNAP2 [50], PTPRN2 [51], DSCAM [37], PSD3 [37], RAPGEF4 [52], CPN1 [53], EPHB2 [40], CAP2 [40], NAV2 [40], and ABCB1 [40].

5.3 Gene set enrichment analysis

To further validate our findings, we also performed gene set enrichment analysis (GSEA) [54] on the aforementioned set of 75 genes. GSEA identifies classes of genes (e.g., those involved in specific pathways) that are over-represented in a given gene set (e.g., the ones we discovered)

Table 12. Analysis of bipolar disorder data. GSEA results from KEGG. O = number of genes in the discovered set; C = total number of genes in the given pathway.

Line	Name	O	C	p-value	
				Nominal	Adjusted
1	metabolic pathways	13	1130	$\ll 0.01$	$\ll 0.01$
2	thyroid cancer	2	29	< 0.01	0.01
3	tyrosine metabolism	2	41	< 0.01	0.01
4	N-glycan biosynthesis	2	49	< 0.01	0.01
5	arginine and proline metabolism	2	54	< 0.01	0.01
6	melanoma	2	71	0.01	0.02
7	ErbB signaling pathway	2	87	0.01	0.02
8	hepatitis C	2	134	0.02	0.03
9	lysosome	2	121	0.02	0.03
10	axon guidance	2	129	0.02	0.03
11	pathways in cancer	3	326	0.02	0.03
12	cell adhesion molecules	2	133	0.02	0.03
13	endocytosis	2	201	0.05	0.05
14	regulation of actin cytoskeleton	2	213	0.05	0.05

https://doi.org/10.1371/journal.pone.0213236.t012

and may have an association with disease phenotypes, by comparing the candidate set against background databases. Gene Ontology [55] is one such database, which annotates and classifies genes in terms of their associated biological processes, cellular components and molecular functions. Other popular databases include KEGG [56] and Pathway Commons [57]. To compare a candidate gene set to various background databases and determine whether certain gene groups (e.g., those occurring in known pathways) appear statistically more or less often than expected, we used a tool called WebGestalt [58].

Table 12 lists the statistically enriched pathways from KEGG (multiple-testing adjusted p-value ≤ 0.05). Many of them have been associated with bipolar disorder or related diseases. For instance, the neurotransmitter dopamine, which is believed to have connections to bipolar disorder, is part of the tyrosine metabolism pathway (line 3). The N-Glycan biosynthesis pathway (line 4) has been reported to be significantly enriched by a study of bipolar disorder in Canadian and UK populations [59]. Both arginine and proline (line 5) have been related to schizophrenia [60]. The ErbB signaling pathway (line 7) regulates a diverse range of physiological responses, such as cell proliferation, migration, differentiation, apoptosis and motility; and insufficient ErbB signaling has been associated with the development of neuro-degenerative diseases in humans [61]. The regulations of the lysosome pathway (line 9) and of the actin cytoskeleton pathway (line 14) were found in a transcriptome sequencing and GWA study to be statistically enriched in genes associated with schizophrenia [62].

For comparison, the corresponding results for MDR and RS are provided in S4 Appendix, while enriched pathways from Gene Ontology and Pathway Commons (for PTY identified genes only) are provided in S5 Appendix.

6 Discussion

This paper is concerned with screening pairs of SNPs, rather than just individual SNPs, for their association with various phenotypes. The complication is that there are many mechanisms—corresponding to different epistatic effects and described by different disease models —for a pair of SNPs to be associated with the outcome.

At the highest level, our main point is that we would be better off using less greedy approaches to determine the "best" disease model for each pair of SNPs. While there are certainly many different ways to achieve this goal, some of which are currently under our active investigation, in this paper we have concentrated on the simple idea of first clustering the disease models and then limiting the candidates to a set of prototypes selected from each respective cluster.

Earlier in Section 1, we stated that screening for higher-order interactions at a genome-wide level is still largely impractical at the present time, but when the time does become ripe for doing so, we think our idea of using prototype disease models will become even more attractive because, as higher-order interactions are considered, there will be combinatorial growth in the number of disease models and a heightened tendency for greedy approaches to produce false positives.

Prototype disease models can be selected in many different ways, although we do not expect that using different sets of prototypes will make a substantial difference. The specific proposal we have outlined in this paper is based on using a particular metric, $\Phi(M', M)$, to quantify the similarity of disease models. We now say more about the intuitive appeal of this metric, as promised earlier in Section 3.1.

Let $r_0 = \mathbb{P}(D)/[1 - \mathbb{P}(D)]$ denote the population-wide case-control ratio. Then, the ratio U/V appearing in the denominator of Eq (9) is simply

$$\frac{U}{V} = \frac{(P_1)r + (1 - P_1)r_0}{(P_0)r + (1 - P_0)r_0} = \frac{r_0 + (r - r_0)P_1}{r_0 + (r - r_0)P_0}. \tag{22}$$

This makes it clear that, if $r = r_0$, then $U/V = 1$. In this case, it is easy to see that the denominator of the Φ-coefficient can be interpreted as $\sqrt{\mathbb{V}\text{ar}(M')\mathbb{V}\text{ar}(M)}$. This is because M can be viewed as a Bernoulli random variable mapping various genotype combinations to either 0 or 1, with $\mathbb{P}(M = 1) = W_{1\cdot}$ and $\mathbb{P}(M = 0) = W_{0\cdot}$, so $\mathbb{V}\text{ar}(M) = W_{1\cdot}W_{0\cdot}$. Likewise,

$$\begin{aligned} \mathbb{V}\text{ar}(M') &= W_{\cdot 1}W_{\cdot 0} \\ &= (W_{11} + W_{01})(W_{10} + W_{00}) \\ &= \underbrace{W_{11}W_{10}}_{M=1} + \underbrace{W_{01}W_{10}}_{M \neq M'} + \underbrace{W_{11}W_{00}}_{M=M'} + \underbrace{W_{01}W_{00}}_{M=0}. \end{aligned} \tag{23}$$

We can decompose $\mathbb{V}\text{ar}(M')$ into four terms, as shown above in Eq (23), where each successive term can be seen to measure the variability in M' when $M = 1$, when M and M' completely disagree, when they completely agree, and when $M = 0$, respectively.

However, for a case-control sample, it is often the case that $r \gg r_0$, in which case Eq (22) implies that $U/V \approx P_1/P_0 > 1$. We can now see that, in this case, Eq (9) implicitly tells us to calculate $\mathbb{V}\text{ar}(M')$, the variance of the potential prototype model M' used to approximate/represent M, differently:

$$\mathbb{V}\text{ar}(M') = \frac{U}{V}W_{11}W_{10} + W_{01}W_{10} + W_{11}W_{00} + \frac{V}{U}W_{01}W_{00}. \tag{24}$$

In particular, among genotypes considered to be risky by M (the set for which $M = 1$), the variability in M' should be up-weighted, which reduces their similarity; whereas, among those considered to be non-risky by M (the set for which $M = 0$), the variability in M' should be down-weighted, which increases their similarity. In other words, when considering M' as a potential prototype for representing M, the metric $\Phi(M', M)$ "thinks" it is more important for M' to agree with M on their assignments of the risky genotypes than for them to agree on the non-risky ones. This is intuitively appealing; a concrete numeric example is given in S6 Appendix.

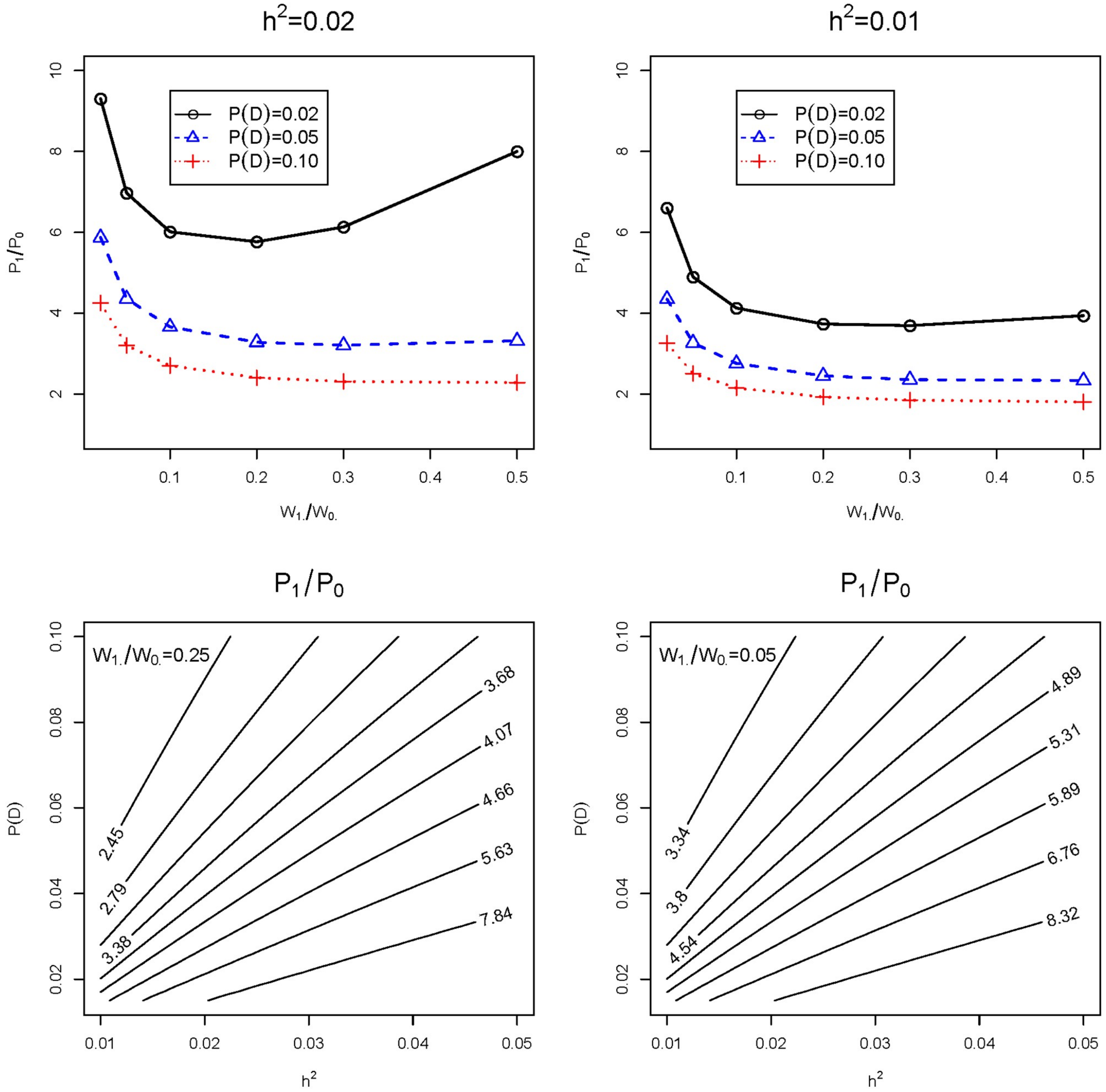


Fig 7. Different views of the odds, P_1/P_0, as a function of the ratio $W_{1\cdot}/W_{0\cdot}$, prevalence $\mathbb{P}(D)$, and heritability h^2, where P_1, P_0 are solutions to Eqs (14) and (15). While the parameters $\mathbb{P}(D)$ and h^2 do affect the odds P_1/P_0 and hence the metric $\Phi(M', M)$, their impact is similar at different values of $W_{1\cdot}/W_{0\cdot}$ and hence similar for different M.

https://doi.org/10.1371/journal.pone.0213236.g007

The approximation that $U/V \approx P_1/P_0$ also allows us to see more clearly how the parameters, $\mathbb{P}(D)$ and h^2, affect the metric Φ. The solution to Eqs (14) and (15) is:

$$P_1 = \mathbb{P}(D) + \sqrt{\frac{W_{0\cdot}}{W_{1\cdot}}\mathbb{P}(D)[1-\mathbb{P}(D)]h^2}, \quad P_0 = \mathbb{P}(D) - \sqrt{\frac{W_{1\cdot}}{W_{0\cdot}}\mathbb{P}(D)[1-\mathbb{P}(D)]h^2}. \qquad (25)$$

Fig 7 contains various views of the odds, P_1/P_0, as a function of the ratio $W_{1\cdot}/W_{0\cdot}$, prevalence $\mathbb{P}(D)$, and heritability h^2. For any given disease model M with a specific ratio $W_{1\cdot}/W_{0\cdot}$, the odds P_1/P_0 is certainly affected by the choices of $\mathbb{P}(D)$ and h^2; but these parameters also affect the odds of other disease models with different $W_{1\cdot}/W_{0\cdot}$-ratios in a similar manner. For example, for fixed h^2, a large (and potentially wrong) choice of $\mathbb{P}(D)$ lowers the odds—whereas, for fixed $\mathbb{P}(D)$, a large (and potentially wrong) choice of h^2 elevates it—for *all* disease models. As a result, even though the distances do change between different disease models and their candidate prototypes, the *relative* distances are not drastically warped. That's why we were able to observe that the resulting prototypes were fairly robust to different choices of $\mathbb{P}(D)$ and h^2.

We also note that the sequential screening procedure we outlined in Section 1.1 is not yet widely considered. This is understandable due to the extra computational burden—every time a pair of SNPs is added, all remaining pairs must be re-assessed. Even with today's technology, such a procedure is still largely infeasible on the genome-wide scale. In fact, we also avoided it in our real-data analysis (Section 5) and simulation study (Section 4) for the same practical reasons, but it may deserve some attention and more systematic treatment in the future.

Finally, we note an important limitation of our current study is that we did *not* consider linkage disequilibrium (LD). It is well-known that, in GWA studies, screening algorithms can declare a specific SNP to be significant only because it is in linkage disequilibrium with another truly-associated SNP. In other words, the presence of LD can lead to false discoveries. Undoubtedly, such concerns also apply to our study here. In fact, properly accounting for LD is much more challenging and complex when we screen for epistatic effects than it is when we screen for main effects only, because the underlying question now becomes whether a *pair* of SNPs is in LD with another *pair*, rather than merely whether an individual SNP is in LD with another SNP. Thus, although studies are available for considering LD in genetic simulations [4], the complexity involved is by no means trivial on a pairwise level. We anticipate that full considerations of LD in the context of epistatis will require a considerable amount of effort in the next few years.

Supporting information

S1 Appendix. Derivation of Eq (4).
(PDF)

S2 Appendix. Derivation of Eq (9).
(PDF)

S3 Appendix. All disease models in $\mathcal{M}$.
(PDF)

S4 Appendix. Analysis of bipolar disorder data. GSEA results from KEGG for genes identified by MDR and RS.
(PDF)

S5 Appendix. Analysis of bipolar disorder data. GSEA results from Gene Ontology and Pathway Commons for genes identified by PTY.
(PDF)

S6 Appendix. A numeric example illustrating the similarity measure $\Phi(M', M)$.
(PDF)

Acknowledgments

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

Author Contributions

Conceptualization: Lu Cheng, Mu Zhu.

Data curation: Lu Cheng.

Formal analysis: Lu Cheng.

Funding acquisition: Mu Zhu.

Investigation: Lu Cheng, Mu Zhu.

Methodology: Lu Cheng, Mu Zhu.

Project administration: Mu Zhu.

Software: Lu Cheng.

Supervision: Mu Zhu.

Writing – original draft: Lu Cheng.

Writing – review & editing: Mu Zhu.

References

1. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–753. https://doi.org/10.1038/nature08494 PMID: 19812666
2. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics. 2010; 11(6):446–450. https://doi.org/10.1038/nrg2809 PMID: 20479774
3. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. The American Journal of Human Genetics. 2011; 88(3):294–305. https://doi.org/10.1016/j.ajhg.2011.02.002 PMID: 21376301
4. Li P, Guo M, Wang C, Liu X, Zou Q. An overview of SNP interactions in genome-wide association studies. Briefings in Functional Genomics. 2015; 14:143–155. https://doi.org/10.1093/bfgp/elu036 PMID: 25241224
5. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh. 1919; 52(2):399–433. https://doi.org/10.1017/S0080456800012163
6. Phillips PC. The language of gene interaction. Genetics. 1998; 149(3):1167–1171. PMID: 9649511
7. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nature Reviews Genetics. 2008; 9(11):855–867. https://doi.org/10.1038/nrg2452 PMID: 18852697
8. Suzuki E, VanderWeele TJ. Compositional epistasis: An epidemiologic perspective. In Moore JH, Williams SM (eds), Epistasis: Methods and Protocols, Springer New York. 2015;197–216.
9. Bateson W, Waunders E, Punnett RC. Experimental studies in the physiology of heredity. Molecular and General Genetics MGG. 1909; 2(1):17–19. https://doi.org/10.1007/BF01975751
10. Li W, Reich J. A complete enumeration and classification of two-locus disease models. Human Heredity. 2000; 50(6):334–349. https://doi.org/10.1159/000022939 PMID: 10899752
11. Lewis CM. Genetic association studies: design, analysis and interpretation. Briefings in Bioinformatics. 2002; 3(2):146–153. https://doi.org/10.1093/bib/3.2.146 PMID: 12139434
12. Minelli C, Thompson JR, Abrams KR, Thakkinstian A, Attia J. The choice of a genetic model in the meta-analysis of molecular association studies. International Journal of Epidemiology. 2005; 34 (6):1319–1328. https://doi.org/10.1093/ije/dyi169 PMID: 16115824
13. Lettre G, Lange C, Hirschhorn JN. Genetic model testing and statistical power in population-based association studies of quantitative traits. Genetic Epidemiology. 2007; 31(4):358–362. https://doi.org/10.1002/gepi.20217 PMID: 17352422

14. 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. https://doi.org/10.1086/321276 PMID: 11404819

15. Wan X, Yang C, Yang Q, Zhao H, Yu W. The complete compositional epistasis detection in genome-wide association studies. BMC Genetics. 2013; 14(1):7. https://doi.org/10.1186/1471-2156-14-7 PMID: 23421496

16. Oki NO, Motsinger-Reif AA. Multifactor dimensionality reduction as a filter-based approach for genome wide association studies. Frontiers in Genetics. 2011; 2:80. https://doi.org/10.3389/fgene.2011.00080 PMID: 22303374

17. Hallgrímsdóttir IB, Yuster DS. A complete classification of epistatic two-locus models. BMC Genetics. 2008; 9(1):17. https://doi.org/10.1186/1471-2156-9-17 PMID: 18284682

18. Gao H, Granka JM, Feldman MW. On the classification of epistatic interactions. Genetics. 2010; 184 (3):827–837. https://doi.org/10.1534/genetics.109.111120 PMID: 20026678

19. Urbanowicz RJ, Granizo-Mackenzie AL, Kiralis J, Moore JH. A classification and characterization of two-locus, pure, strict, epistatic models for simulation and detection. BioData Mining. 2014; 7(1):8. https://doi.org/10.1186/1756-0381-7-8 PMID: 25057293

20. Oki NO, Motsinger-Reif AA. Multifactor dimensionality reduction as a filter-based approach for genome wide association studies. Frontiers in Genetics. 2011; 2:80. https://doi.org/10.3389/fgene.2011.00080 PMID: 22303374

21. Mei H, Ma D, Ashley-Koch A, Martin ER. Extension of multifactor dimensionality reduction for identifying multilocus effects in the GAW14 simulated data. BMC Genetics. 2005; 6(Suppl 1):S145. https://doi.org/10.1186/1471-2156-6-S1-S145 PMID: 16451605

22. Yule GU. On the methods of measuring association between two attributes. Journal of the Royal Statistical Society. 1912; 75(6):579–652. https://doi.org/10.2307/2340126

23. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. Nature Reviews Genetics. 2013; 14(2):139–149. https://doi.org/10.1038/nrg3377 PMID: 23329114

24. Likas A, Vlassis N, Verbeek JJ. The global K-means clustering algorithm. Pattern Recognition. 2003; 36 (2):451–461. https://doi.org/10.1016/S0031-3203(02)00060-2

25. Dong C, Chu X, Wang Y, Wang Y, Jin L, Shi T, et al. Exploration of gene–gene interaction effects using entropy-based methods. European Journal of Human Genetics. 2008; 16(2):229–235. https://doi.org/10.1038/sj.ejhg.5201921 PMID: 17971837

26. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nature Genetics. 2007; 39(9):1167–1173. https://doi.org/10.1038/ng2110 PMID: 17721534

27. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, 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. https://doi.org/10.1016/j.ajhg.2010.07.021 PMID: 20817139

28. Emily M. IndOR: A new statistical procedure to test for SNP–SNP epistasis in genome-wide association studies. Statistics in Medicine. 2012; 31(21):2359–2373. https://doi.org/10.1002/sim.5364 PMID: 22711278

29. Park MY, Hastie T. Penalized logistic regression for detecting gene interactions. Biostatistics. 2008; 9 (1):30–50. https://doi.org/10.1093/biostatistics/kxm010 PMID: 17429103

30. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics. 2005; 37(4):413–417. https://doi.org/10.1038/ng1537 PMID: 15793588

31. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genetic Epidemiology. 2003; 24(2):150–157. https://doi.org/10.1002/gepi.10218 PMID: 12548676

32. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447 (7145):661–678. https://doi.org/10.1038/nature05911

33. Craddock N, Sklar P. Genetics of bipolar disorder. The Lancet. 2013; 381(9878):1654–1662. https://doi.org/10.1016/S0140-6736(13)60855-7

34. Berk M, Kapczinski F, Andreazza A, Dean O, Giorlando F, Maes M, et al. Pathways underlying neuroprogression in bipolar disorder: focus on inflammation, oxidative stress and neurotrophic factors. Neuroscience & Biobehavioral Reviews. 2011; 35(3):804–817. https://doi.org/10.1016/j.neubiorev.2010.10.001

35. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, et al. The Ensembl gene annotation system. Database. 2016; 2016. https://doi.org/10.1093/database/baw093 PMID: 27337980

36. Ullah AZD, Lemoine NR, Chelala C. SNPnexus: A web server for functional annotation of novel and publicly known genetic variants (2012 update). Nucleic Acids Research. 2012; 40(Web Server issue): W65–W70. https://doi.org/10.1093/nar/gks364

37. Wang KS, Liu XF, Aragam N. A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder. Schizophrenia Research. 2010; 124(1):192–199. https://doi.org/10.1016/j.schres.2010.09.002 PMID: 20889312

38. Yosifova A, Mushiroda T, Kubo M, Takahashi A, Kamatani Y, Kamatani N, et al. Genome-wide association study on bipolar disorder in the Bulgarian population. Genes, Brain and Behavior. 2011; 10(7):789–797. https://doi.org/10.1111/j.1601-183X.2011.00721.x

39. Perlis RH, Huang J, Purcell S, Fava M, Rush AJ, Sullivan PF, et al. Genome-wide association study of suicide attempts in mood disorder patients. American Journal of Psychiatry. 2010. https://doi.org/10.1176/appi.ajp.2010.10040541

40. Willour VL, Seifuddin F, Mahon PB, Jancic D, Pirooznia M, Steele J, et al. A genome-wide association study of attempted suicide. Molecular Psychiatry. 2012; 17(4):433–444. https://doi.org/10.1038/mp.2011.4 PMID: 21423239

41. Popescu BFG, Bunyan RF, Guo Y, Parisi JE, Lennon VA, Lucchinetti CF. Evidence of aquaporin involvement in human central pontine myelinolysis. Acta Neuropathologica Communications. 2013; 1(1):40. https://doi.org/10.1186/2051-5960-1-40 PMID: 24252214

42. Wang B. Molecular Mechanism Underlying Sialic Acid as an Essential Nutrient for Brain Development and Cognition. Advances in Nutrition. 2012; 3(3):465S–472S. https://doi.org/10.3945/an.112.001875 PMID: 22585926

43. Ripke S, Neale BM, Corvin A, Walters JT, Farh KH, Holmans PA, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014; 511(7510):421. https://doi.org/10.1038/nature13595

44. Kathiresan S, Manning AK, Demissie S, D'agostino RB, Surti A, Guiducci C, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC Medical Genetics. 2007; 8(1):S17. https://doi.org/10.1186/1471-2350-8-S1-S17 PMID: 17903299

45. Sherva R, Tripodis Y, Bennett DA, Chibnik LB, Crane PK, De Jager PL, et al. Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. Alzheimer's & Dementia. 2014; 10(1):45–52. https://doi.org/10.1016/j.jalz.2013.01.008

46. Hu LW, Kawamoto EM, Brietzke E, Scavone C, Lafer B. The role of Wnt signaling and its interaction with diverse mechanisms of cellular apoptosis in the pathophysiology of bipolar disorder. Progress in Neuro-Psychopharmacology and Biological Psychiatry. 2011; 35(1):11–17. https://doi.org/10.1016/j.pnpbp.2010.08.031 PMID: 20828594

47. Rasmussen DD, Ishizuka B, Quigley ME, Yen SS. Effects of tyrosine and tryptophan ingestion on plasma catecholamine and 3,4-dihydroxyphenylacetic acid concentrations. Journal of Clinical Endocrinology and Metabolism. 1983; 57(4):760–763. https://doi.org/10.1210/jcem-57-4-760 PMID: 6885965

48. Baranzini SE, Srinivasan R, Khankhanian P, Okuda DT, Nelson SJ, Matthews PM, et al. Genetic variation influences glutamate concentrations in brains of patients with multiple sclerosis. Brain. 2010; 133(9):2603–2611. https://doi.org/10.1093/brain/awq192 PMID: 20802204

49. Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, Jones L, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nature Genetics. 2008; 40(9):1056–1058. https://doi.org/10.1038/ng.209 PMID: 18711365

50. O'Dushlaine C, Kenny E, Heron E, Donohoe G, Gill M, Morris D, et al. Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. Molecular Psychiatry. 2011; 16(3):286–292. https://doi.org/10.1038/mp.2010.7 PMID: 20157312

51. Curtis D, Vine AE, McQuillin A, Bass NJ, Pereira A, Kandaswamy R, et al. Case-case genome wide association analysis reveals markers differentially associated with schizophrenia and bipolar disorder and implicates calcium channel genes. Psychiatric Genetics. 2011; 21(1):1. https://doi.org/10.1097/YPG.0b013e3283413382 PMID: 21057379

52. Smith EN, Bloss CS, Badner JA, Barrett T, Belmonte PL, Berrettini W, et al. Genome-wide association study of bipolar disorder in European American and African American individuals. Molecular Psychiatry. 2009; 14(8):755–763. https://doi.org/10.1038/mp.2009.43 PMID: 19488044

53. Sklar P, Smoller J, Fan J, Ferreira M, Perlis R, Chambert K, et al. Whole-genome association study of bipolar disorder. Molecular Psychiatry. 2008; 13(6):558–569. https://doi.org/10.1038/sj.mp.4002151 PMID: 18317468

54. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102

55. Consortium GO, et al. Gene ontology consortium: going forward. Nucleic Acids Research. 2015; 43 (D1):D1049–D1056. https://doi.org/10.1093/nar/gku1179

56. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Research. 2016; 44(D1):D457–D462. https://doi.org/10.1093/nar/gkv1070 PMID: 26476454

57. Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur Ö, Anwar N, et al. Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research. 2011; 39(suppl 1):D685–D690. https://doi.org/10.1093/nar/gkq1039 PMID: 21071392

58. Wang J, Duncan D, Shi Z, Zhang B. WEB-based gene set analysis toolkit (WebGestalt): update 2013. Nucleic Acids Research. 2013; 41(W1):W77–W83. https://doi.org/10.1093/nar/gkt439 PMID: 23703215

59. Xu W, Cohen-Woods S, Chen Q, Noor A, Knight J, Hosang G, et al. Genome-wide association study of bipolar disorder in Canadian and UK populations corroborates disease loci including SYNE1 and CSMD1. BMC Medical Genetics. 2014; 15(1):1. https://doi.org/10.1186/1471-2350-15-2

60. He Y, Yu Z, Giegling I, Xie L, Hartmann A, Prehn C, et al. Schizophrenia shows a unique metabolomics signature in plasma. Translational Psychiatry. 2012; 2(8):e149. https://doi.org/10.1038/tp.2012.76 PMID: 22892715

61. Roy K, Murtie JC, El-Khodor BF, Edgar N, Sardi SP, Hooks BM, et al. Loss of erbB signaling in oligodendrocytes alters myelin and dopaminergic function, a potential mechanism for neuropsychiatric disorders. Proceedings of the National Academy of Sciences. 2007; 104(19):8131–8136. https://doi.org/10.1073/pnas.0702157104

62. Zhao Z, Xu J, Chen J, Kim S, Reimers M, Bacanu SA, et al. Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder. Molecular Psychiatry. 2015; 20(5):563–572. https://doi.org/10.1038/mp.2014.82 PMID: 25113377