Multilocus Informativeness: A Novel Method for Fine-Mapping Disease-Causing Variants in Genome-Wide Association Studies

Daniel Ben-Isvy

Thesis Advisor: Sorin Istrail
Second Reader: Emilia Huerta-Sanchez

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science with Honors

Center for Computational Molecular Biology
Brown University

April 2021

Abstract

Genome-wide association studies (GWAS) are powerful tools for identifying associations between genetic variants and complex human diseases. However, attempts to pinpoint the exact genetic variants that contribute to disease phenotypes are often obstructed by blocks of high linkage disequilibrium (LD) spread throughout the genome. Consequently, effective fine-mapping methods are needed to successfully localize the sections of the genome containing disease-causing variants. To achieve this goal, we develop a novel method called multilocus informativeness that leverages both LD and association data to effectively fine-map disease-causing variants. We show that multilocus informativeness is a robust fine-mapping method that can successfully locate causative variants for several different disease phenotypes in the same population. Additionally, we demonstrate that multilocus informativeness can effectively localize disease association signals in a variety of simulated study populations, and we show that multilocus informativeness compares favorably to multiple existing mapping methods.

Contents

1 Introduction

Genome-wide association studies (GWAS) have become popular approaches for identifying the genetic determinants of complex human diseases. While GWAS methodologies can be used to uncover associations between disease phenotypes and any types of genetic variants, they are most commonly applied to single-nucleotide polymorphisms, or SNPs. One of the central challenges that confounds GWAS attempts to pinpoint disease-causing SNPs is linkage disequilibrium (LD), or the correlation between genotypes at different loci across the genome. Specifically, the human genome contains many LD blocks, or regions characterized by high LD between SNPs located within the region and low LD between SNPs located inside and outside the region[1]. Consequently, because the genotypes of SNPs within LD blocks are highly correlated, it can be difficult to determine exactly which SNP in an LD block contributes to the disease phenotype when a statistical association is detected.

Fine-mapping methods seek to resolve this problem by further localizing the association signal obtained from a GWAS, ideally pinpointing a specific variant suspected to cause disease. One approach to fine-mapping genetic variants is to apply the same single-SNP association tests often used to identify genome-wide associations on this smaller candidate region of the genome[2,3]. However, the applications of these statistical tests typically assume that the genotypes at all of the examined loci are independent, an assumption which does not hold in smaller candidate regions with high LD. Additionally, because these tests only consider the genetic variation at one SNP at a time, they lose statistical power that could be gained from simultaneously considering the data at other nearby loci.

As a result of these disadvantages, a variety of other fine-mapping methods have been developed, including methods built on LD-based heuristics, penalized regression, Bayesian statistics, and genome annotation[4]. However, each of these methods also has important limitations. Specifically, previously developed LD-based fine-mapping heuristics typically rely on arbitrary LD thresholds to identify potentially causative SNPs, and these approaches do not leverage association data from multiple SNPs to improve statistical power. Additionally, while other fine-mapping approaches often do simultaneously consider data from multiple loci, they do not explicitly model the significant amount of local genomic structure present in LD blocks. Consequently, we sought to develop a fine-mapping method that both explicitly models the genomic structure of LD blocks and simultaneously incorporates association data from several nearby SNPs to improve statistical power.

2 Mapping Disease-Causing Variants in GWAS

2.1 Existing Methods

While many methods for mapping disease-causing variants in GWAS exist, we focus here on the single-SNP methods that are commonly used to identify genotype-phenotype associations in genome-wide analyses. The computation of these test statistics is based on a contingency table of observations specifying the number of individuals in the study population with each possible genotype-phenotype combination (Table 1).

Genotype	**AA**	**Aa**	**aa**	**Total**
Controls	r_0	r_1	r_2	r
Cases	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

Table 1: A contingency table for the A/a locus.

2.1.1 Chi-Squared Test

The chi-squared test measures the likelihood that an observed deviation from the expected contingency table occurred due to chance. Under the null hypothesis of no association between the genotype and phenotype, the expected contingency table can be easily computed from a contingency table of observations (Table 2).

Genotype	**AA**	**Aa**	**aa**
Controls	n_0r/n	n_1r/n	n_2r/n
Cases	n_0s/n	n_1s/n	n_2s/n

Table 2: The expected contingency table for the observations shown in Table 1.

The observed and expected contingency tables can then be used to compute the chi-squared test statistic, which is χ^2-distributed with 2 degrees of freedom:

$$\chi_2^2 = \sum_{i=0}^{2} \frac{(r_i - n_ir/n)^2}{n_ir/n} + \frac{(s_i - n_is/n)^2}{n_is/n} \tag{1}$$

2.1.2 Cochran-Armitage Trend Test

The Cochran-Armitage trend test modifies the chi-squared test to investigate a suspected trend in the genotypes' effects[5,6]. The suspected trend is specified by a weights vector $x = (x_0, x_1, x_2)$ where x_i is the weight for the genotype with i copies of the a allele. When compared to the chi-squared test, the Cochran-Armitage trend test has more statistical power to detect deviations from the expected contingency table that follow the suspected trend but less statistical power to detect deviations from the expected contingency table that follow other trends.

Several commonly examined trends for how the a allele at the A/a locus may affect disease risk are shown in Table 3. Notably, the Cochran-Armitage trend test can detect both increases and decreases in disease risk associated with the a allele, indicating that the trends shown in Table 3 can be used even if it is unclear which allele may contribute to the disease phenotype. However, it is also important to note that for some weights vectors, the suspected trend for the a allele may be different than the suspected trend for the A allele. Specifically, while an additive trend for the a allele also investigates an additive trend for the A allele, a dominant (or recessive) trend for the a allele examines a recessive (or dominant) trend for the A allele. As a result, it may sometimes be necessary to test for multiple suspected trends if either allele could contribute to the disease phenotype.

Suspected Trend	$\mathbf{x_0}$	$\mathbf{x_1}$	$\mathbf{x_2}$
Additive	0	1	2
Dominant	0	1	1
Recessive	0	0	1

Table 3: Common suspected trends for how the a allele at the A/a locus may affect disease risk.

The Cochran-Armitage trend test statistic is χ^2-distributed with 1 degree of freedom and can be computed from a contingency table of observations as follows:

$$T^2 = \frac{\sum_{i=0}^{2} [x_i(sr_i - rs_i)]^2}{\frac{rs}{n}\left[\sum_{i=0}^{2} x_i^2 n_i(n-n_i) - 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} x_i x_j n_i n_j\right]} \tag{2}$$

2.2 Multilocus Informativeness

In this section, we develop a new measure called multilocus informativeness that adapts the directed informativeness measure of LD to the GWAS setting.

2.2.1 Directed Informativeness

Directed informativeness is a graph-theoretic LD measure that can be conservatively extended to multiple loci[7,8]. Consider a population of n haplotypes genotyped at m biallelic SNPs. At each SNP, let 0 denote the major, or more frequent, allele and 1 denote the minor, or less frequent, allele. Additionally, let $A_{i,j}$ denote the allele of haplotype i at SNP j.

We define the distinguishability graph, or D graph, for a SNP t to be a directed graph denoted D_t with vertex set V and edge set E_t defined as follows:

$$V = \{1, 2, ..., n\}$$
$$E_t = \{(i, j) : A_{i,t} = 0, A_{j,t} = 1\}$$

Notably, each vertex in D_t represents a haplotype in the population, and two haplotypes are connected by an edge if they have different alleles at SNP t. Edges in D_t are oriented from the haplotype containing the major allele at SNP t to the haplotype containing the minor allele at SNP t.

Next, we define the directed informativeness of a SNP t with respect to a SNP u to be:

$$DI(t, u) = \frac{\sum_{(i,j) \in E_u} \delta_{i,j}(t)}{|E_u|} \tag{3}$$

where:

$$\delta_{i,j}(t) = \begin{cases} 1 & \text{if } (i, j) \in E_t \\ -1 & \text{if } (j, i) \in E_t \\ 0 & \text{otherwise} \end{cases}$$

Directed informativeness ranges from -1 to 1 with values further from 0 signaling stronger LD. The measure is computed by counting the fraction of the edges in the D graph for SNP u that are also in the D graph for SNP t, with edges oriented in opposite directions in the two graphs contributing -1 instead of 1 to the numerator. Consequently, positive directed informativeness values indicate an association between the major alleles at both loci, whereas negative directed informativeness values indicate an association between the major allele at one locus and the minor allele at the other locus.

2.2.2 Adaptation to GWAS

Now, consider a population of n diploid individuals genotyped at m biallelic SNPs. Let $G_{i,j}$ be the number of copies of the minor allele that individual i has at SNP j, and let P_i be the phenotype of individual i where $P_i = 0$ for individuals in the control group and $P_i = 1$ for individuals in the case group.

We define the GWAS distinguishability graph for a SNP t to be a directed graph denoted GD_t with vertex set V and edge set E'_t defined as follows:

$$V = \{1, 2, ..., n\}$$
$$E'_t = \{(i, j) : G_{i,t} < G_{j,t}, P_i \neq P_j\}$$

Notably, each vertex in GD_t represents a diploid individual in the population, and two individuals are connected by an edge if they have different phenotypes and different genotypes at SNP t. This construction reflects the fact that we are primarily interested in distinguishing case individuals from control individuals in GWAS, so we only draw edges between individuals with different phenotypes. Additionally, edges in GD_t are oriented from the individual with fewer copies of the minor allele at SNP t to the individual with more copies of the minor allele at SNP t.

Next, we define the single-locus informativeness of a SNP t with respect to the phenotype to be:

$$I_1(t) = \frac{\sum_{(i,j)\in E'_t} \delta'(i,j)}{|E'_t|} \tag{4}$$

where:

$$\delta'(i,j) = \begin{cases} 1 & \text{if } P_i = 0 \text{ and } P_j = 1 \\ -1 & \text{if } P_i = 1 \text{ and } P_j = 0 \end{cases}$$

Single-locus informativeness ranges from -1 to 1 with values further from 0 signaling a more significant association between the SNP t and the phenotype. The measure is computed by counting the fraction of the edges in the D graph for SNP t that point in the same phenotypic direction, with edges oriented from controls to cases contributing 1 to the numerator and edges oriented from cases to controls contributing -1 to the numerator. Consequently, positive single-locus informativeness values associate the minor allele with the case phenotype, and negative single-locus informativeness values associate the major allele with the case phenotype.

Furthermore, we define the GWAS directed informativeness of a SNP t with respect to a SNP u equivalently to how directed informativeness was previously defined for LD, with the exception that the definition of the D graph has now changed:

$$GDI(t,u) = \frac{\sum_{(i,j)\in E'_u} \delta_{i,j}(t)}{|E'_u|} \tag{5}$$

GWAS directed informativeness has the same general properties as directed informativeness for LD, as edges in the D graph are still oriented from the vertex with fewer copies of the minor allele to the vertex with more copies of the minor allele.

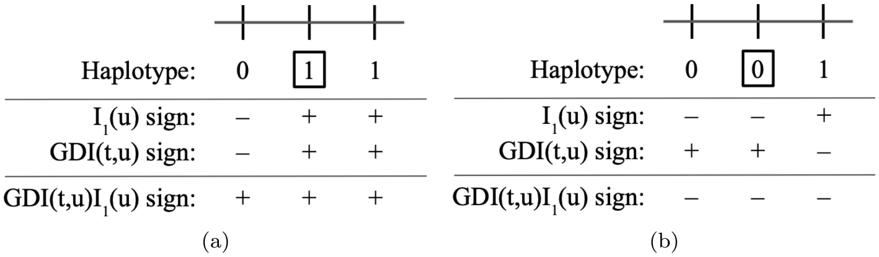
Finally, we define the multilocus informativeness of a SNP t with respect to the phenotype to be:

$$I_w(t) = \frac{\sum_{u\in W_t} GDI(t,u) I_1(u)}{w} \tag{6}$$

where w is an odd-numbered window size specifying how many SNPs are contained within the set W_t, and W_t is the set containing the SNP t as well as the first $(w-1)/2$ SNPs to the left and right of SNP t. Therefore, the multilocus informativeness for a SNP t is computed as a weighted average of the single-locus informativeness values of all SNPs u within $(w-1)/2$ of SNP t, where each SNP u is weighted according to the strength of the LD between SNPs t and u. This formulation seeks to use the fact that SNPs in high LD with a disease-causing SNP will also exhibit associations with the disease phenotype to more effectively identify the disease-causing variant from a GWAS.

More specifically, the expectation that disease-causing SNPs will often exist within regions of high LD in GWAS study populations is supported by the typical sampling methodology of GWAS. In particular, because GWAS typically include a larger fraction of case individuals than the general population, we expect that disease-causing SNPs, along with tracts of local SNPs in high LD with those SNPs, will often appear more frequently in GWAS study populations than the general population. This observation indicates that levels of genomic structure may frequently be elevated around disease-causing SNPs in GWAS, suggesting that regions with higher LD may be more likely to contain causative SNPs. As a result, multilocus informativeness may be able to more effectively localize disease-causing SNPs in GWAS by integrating this expectation of local genomic structure with the association data at several nearby SNPs.

Moreover, while multilocus informativeness incorporates information from several nearby SNPs, it also possesses many of the same general properties that characterize single-locus informativeness. Namely, multilocus informativeness ranges from -1 to 1 with values further from 0 signaling a more significant association between the SNP t and the phenotype. Additionally, positive multilocus informativeness values associate the minor allele with the case phenotype, and negative multilocus informativeness values associate the major allele with the case phenotype. Figure 1 demonstrates how the association and LD data at SNPs surrounding a disease locus contribute to the computation of multilocus informativeness to produce these trends.



Haplotype:	0	$\boxed{1}$	1
I_1(u) sign:	–	+	+
GDI(t,u) sign:	–	+	+
GDI(t,u)I_1(u) sign:	+	+	+

(a)

Haplotype:	0	$\boxed{0}$	1
I_1(u) sign:	–	–	+
GDI(t,u) sign:	+	+	–
GDI(t,u)I_1(u) sign:	–	–	–

(b)

Figure 1: Computation of multilocus informativeness for disease-causing SNPs in strong LD with nearby SNPs. Haplotypes characteristic of case individuals are shown, with the boxed SNP t as the disease-causing SNP. The signs of single-locus informativeness and GWAS directed informativeness are depicted for each nearby SNP u as they would be computed in a case/control cohort with strong LD. (a) Association of a minor allele with the disease phenotype produces a positive value of multilocus informativeness. (b) Association of a major allele with the disease phenotype produces a negative value of multilocus informativeness.

Finally, we would like to highlight the importance of choosing an appropriate window size w when computing multilocus informativeness. The window size w should be chosen so that each set W_t is expected to contain all of the SNPs in high LD with the candidate SNP t and few SNPs that are not in high LD with the candidate SNP t. If the window size w is chosen to be too small, multilocus informativeness will not adequately incorporate local genomic structure and association data from nearby SNPs into the calculation, which will diminish the potential benefits of using this measure over other single-SNP statistical tests. Additionally, if the window size w is chosen to be too large, the multilocus informativeness calculation may become less accurate due to the noisy contributions of the additional low LD SNPs. However, because low LD SNPs do not contribute as much to the calculation of multilocus informativeness, having a window size that is slightly too large will likely not significantly affect results.

3 Simulating GWAS Data

In order to test the performance of the various methods for fine-mapping disease-causing variants in GWAS, we develop a pipeline for simulating a realistic GWAS study population. This pipeline is based on methodologies that have been used in previous fine-mapping studies[9], and it includes three main steps. First, a forward evolutionary simulation is used to generate a population with realistic patterns of genetic variation. Second, a model of disease incidence is specified. And third, a case/control sample is drawn from the general population. In this section, we will examine each of these steps in greater detail.

3.1 Forward Simulations

To perform forward simulations, we use the evolutionary simulation framework SLiM[10]. We model the candidate region of the genome selected for fine-mapping as a 1 Mb region of one chromosome. Then, we simulate a population of 10,000 diploid individuals forwards through time for 20,000 generations under specific models of mutation and recombination to generate a population with realistic patterns of genetic variation in the region. In this study, we use a uniform mutation model in which novel neutral mutations are randomly introduced into the population at a rate of m mutations/base pair/individual/generation across the entire candidate region.

Additionally, we examine two models of recombination in this study. First, we investigate a uniform recombination model in which recombinations randomly occur in the population at a rate of r_U breakpoints/base pair/individual/generation across the entire candidate region. And second, we consider a recombination hotspot model that more explicitly models the elevated recombination rates typically found on the boundaries of LD blocks[1]. In this model, 1% of the total length of the candidate region consists of 5 recombination hotspots evenly spaced across the region, and 60% of all recombination events occur within the hotspots. Furthermore, recombinations randomly occur inside hotspots at a higher rate of r_H breakpoints/base pair/individual/generation, and recombinations randomly occur outside hotspots at a lower rate of r_L breakpoints/base pair/individual/generation.

Moreover, to relate the parameters of the two recombination models, we can compute the values of r_H and r_L that are consistent with an overall average recombination rate of r_U by solving the following linear system of equations:

$$\begin{cases} 0.99r_L + 0.01r_H = r_U \\ 1.5(0.99r_L) - 0.01r_H = 0 \end{cases} \tag{7}$$

In this system, the first equation specifies that the average recombination rate across the entire candidate region is r_U, and the second equation specifies that 60% of all recombination events occur within recombination hotspots. Finally, Table 4 depicts the default parameter values that are used for the mutation and recombination models in this study unless alternative parameter values are specified.

Parameter	**Default Value**
m	1.1×10^{-8}
r_U	2.2×10^{-8}
r_H	1.32×10^{-6}
r_L	8.89×10^{-9}

Table 4: Default parameter values for mutation and recombination models in this study.

3.2 Disease Models

Suppose that allele a at locus A/a contributes to the disease phenotype. To quantify how allele a affects phenotypes in the population, we define several parameters that allow us to specify a disease model for this situation. First, we define the penetrance of a particular genotype XX, denoted $P(XX)$, to be the probability that an individual has the disease phenotype given that they have genotype XX. Second, we define the genotype relative risk of a particular genotype XX, denoted $GRR(XX)$, to be how much more likely an individual with genotype XX is to have the disease phenotype than an individual with genotype AA. More precisely, genotype relative risk is computed as a ratio of penetrances:

$$GRR(XX) = \frac{P(XX)}{P(AA)} \tag{8}$$

From this definition, it is clear that we will always have $GRR(AA) = 1$. However, $GRR(Aa)$ and $GRR(aa)$ can vary considerably based on the disease model used. Table 5 describes four commonly used types of disease models along with the values of $GRR(Aa)$ and $GRR(aa)$ that are consistent with these models. For the complex disease phenotypes analyzed in GWAS, an additive disease model is most frequently assumed, although other types of genotype-phenotype relationships are also possible[2].

Next, we define the disease prevalence, denoted p, to be the fraction of individuals in the entire population who have the disease phenotype. And lastly, we define the disease allele frequency, denoted q, to be the fraction of individuals in the entire population who have the disease allele a. Now, we can completely specify a model for how the genotype at locus A/a affects disease phenotypes in the population by providing the parameters $GRR(Aa)$, $GRR(aa)$, p, and q. Notably, if these four parameters are given, then we can compute the penetrance for each possible genotype at locus A/a under the assumption that this locus is

Model Type	**Description**	**Genotype Relative Risks**
Additive	Each copy of the disease allele increases disease risk by the same additive amount	$GRR(aa) = 2 \times GRR(Aa) - 1$
Multiplicative	Each copy of the disease allele increases disease risk by the same multiplicative amount	$GRR(aa) = [GRR(Aa)]^2$
Dominant	All individuals with at least one copy of the disease allele have the same disease risk	$GRR(aa) = GRR(Aa)$
Recessive	Individuals need two copies of the disease allele to have an increased disease risk	$GRR(Aa) = 1$

Table 5: Commonly used types of disease models and their genotype relative risks.

in Hardy-Weinberg equilibrium (HWE) by solving the following linear system of equations:

$$\begin{cases} p = (1-q)^2 \times P(AA) + 2q(1-q) \times P(Aa) + q^2 \times P(aa) \\ 0 = GRR(Aa) \times P(AA) - P(Aa) \\ 0 = GRR(aa) \times P(AA) - P(aa) \end{cases} \tag{9}$$

In this system, the first equation expresses the disease prevalence as a weighted average of penetrances where the penetrance for each genotype is weighted by the frequency of that genotype in the population under HWE. Additionally, the second and third equations are derived from the definition of genotype relative risk (Equation 8).

Therefore, if the locus A/a is in HWE, the four parameters $GRR(Aa)$, $GRR(aa)$, p, and q are sufficient to fully specify the disease model. In this study, we assume that all disease loci are in HWE, which is consistent with the forward simulation methodology used here. Table 6 depicts the default parameter values and computed penetrances that are used for the disease model in this study unless alternative parameter values are specified.

Parameter	**Default Value**
Model type	Additive
$GRR(Aa)$	1.3
$GRR(aa)$	1.6
p	0.1
q	0.2
$P(AA)$	0.089
$P(Aa)$	0.116
$P(aa)$	0.143

Table 6: Default parameter values for the disease model in this study.

Finally, we would like to note that if the candidate locus A/a is not in HWE, the methodology from Equation 9 can still be used to compute the disease penetrance for each possible genotype at locus A/a. However, now that the population is not in HWE, the genotype frequencies in the population cannot be simply computed from the disease allele frequency q. Consequently, in this case, the genotype frequencies must be separately specified in the disease model.

3.3 Case/Control Sampling

After conducting a forward simulation and specifying a disease model, a case/ control cohort can be sampled from the simulated population using Algorithm 1. The algorithm takes as input a simulated population S, the parameters of a disease model, the number of control individuals to sample n_C, the number of case individuals to sample n_D, and a small number ϵ specifying the amount of allowed variation in the disease allele frequency. Table 7 depicts the default parameter values that are used for this algorithm in this study unless alternative parameter values are specified.

The basic steps of the algorithm are as follows. First, select a disease locus by randomly choosing a locus with allele frequency between $q - \epsilon$ and $q + \epsilon$ in the simulated population. Next, while the case and control groups are not both full, randomly choose two haplotypes from the population to form an individual. Then, randomly draw the phenotype of this individual based on the penetrance for their specific genotype, which gives the probability that the individual has the disease phenotype given their genotype. Finally, add the individual to the appropriate study group (case or control) based on their phenotype if that group is not yet full.

Algorithm 1 Case/Control Sampling Algorithm

```
$l \leftarrow$ random locus with allele frequency between $q - \epsilon$ and $q + \epsilon$ in $S$
while the case and control groups are not both full do
    $(h1, h2) \leftarrow$ 2 random haplotypes from $S$
    $g \leftarrow$ genotype of $(h1, h2)$ at locus $l$
    $P_g \leftarrow$ phenotype of $(h1, h2)$, randomly drawn from Bernoulli$(P(g))$
    if group $P_g$ is not full then add $(h1, h2)$ to group $P_g$
end while
return $l$, the case group, and the control group
```

Parameter	Default Value
n_C	500
n_D	500
ϵ	0.005

Table 7: Default parameter values for the case/control sampling algorithm in this study.

4 Performance of Mapping Methods

In this section, we compare the performance of the various mapping methods described here on an assortment of simulated datasets.

4.1 Metrics of Performance

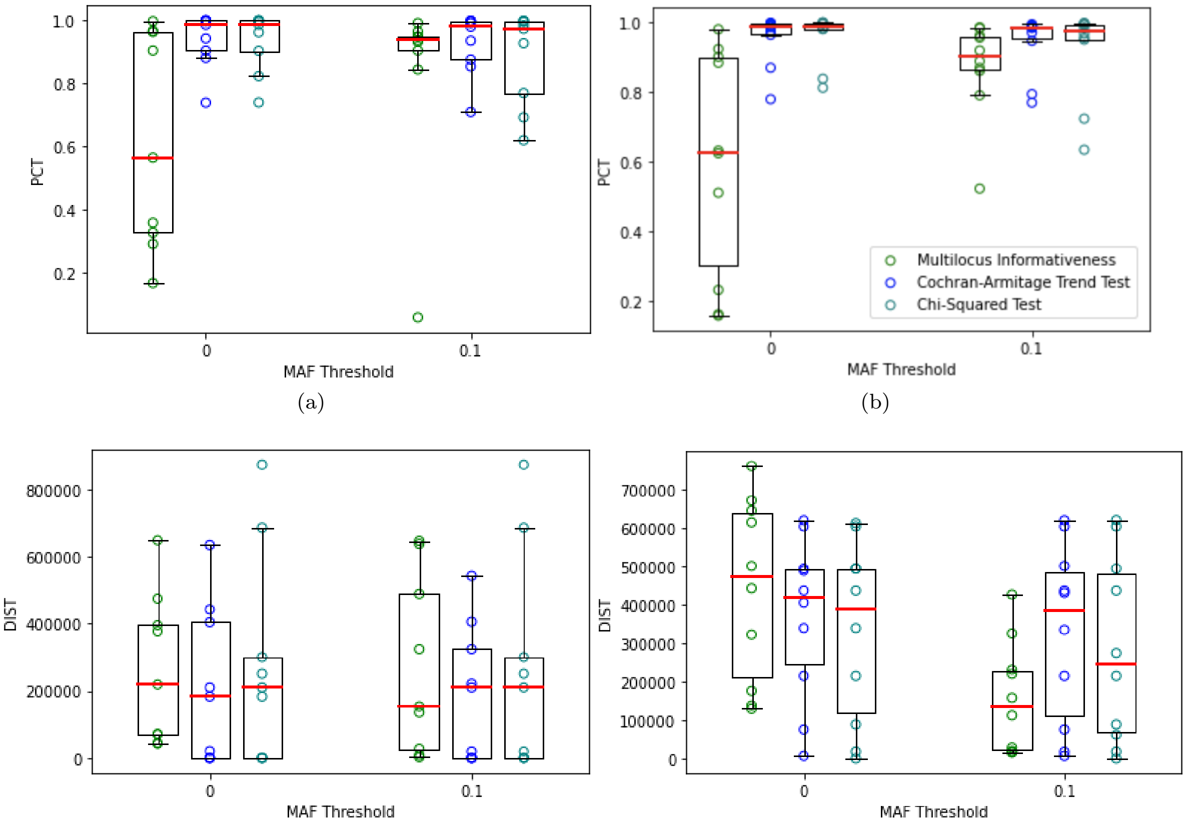
To compare the performance of the various mapping methods, we define two metrics that can be used to summarize how effectively a particular method localizes the disease-causing variant in the population. First, we examine the percentile of the causative SNP score in comparison to all SNP scores (PCT), which captures how highly a particular mapping method prioritizes the disease-causing SNP. And second, we investigate the distance from the causative SNP to the highest-scoring SNP (DIST), which measures whether a particular mapping method is able to effectively localize the disease-causing SNP to a small section of the entire candidate region. This second metric can be particularly useful for situations in which strong LD between the disease-causing SNP and other nearby SNPs makes it difficult for a particular mapping method to pinpoint the exact disease locus but the method can still effectively identify a small area of the candidate region that contains that locus. Additionally, because multilocus informativeness is a signed quantity, we use the magnitude of multilocus informativeness values to compute these metrics.

4.2 MAF Thresholds

In both LD-based algorithms and GWAS pipelines, SNPs with low minor allele frequency (MAF) are often filtered out of the set of examined SNPs. In LD-based algorithms, applying a MAF threshold can be useful to avoid the extreme values of LD that often result from the high variance in LD values at rare SNPs[11]. Additionally, in GWAS pipelines, removing low MAF variants can be useful to avoid unnecessary computations when the sample size is too small to have enough statistical power to effectively detect associations between these variants and the disease phenotype[12].

As a result of these practices, we examine whether removing low MAF variants from the simulated datasets improves the performance of any of the mapping methods (Figure 2). Under both the uniform recombination model and the recombination hotspot model, multilocus informativeness performs significantly better when a MAF threshold of 0.1 is applied compared to when no MAF threshold is applied. Specifically, the MAF threshold substantially increases the percentile of the causative SNP under both recombination models, and it substantially decreases the distance from the causative SNP to the highest-scoring SNP under the recombination hotspot model.

The improved performance of multilocus informativeness when a MAF threshold is applied is likely due to the ability of rare variants to significantly affect the computation of this measure. Specifically, because most individuals in a



(a)

(b)

(c)

(d)

Figure 2: Effect of MAF thresholds on the performance of mapping methods. Ten replicate populations are simulated under the uniform recombination model and the recombination hotspot model. Colored circles show performance on individual populations, red lines show median values of the examined metric, boxes extend to the upper and lower quartiles of the data, and whiskers extend up to 1.5 times the interquartile range away from the median. An additive trend is used for the Cochran-Armitage trend test, and window sizes of $w = 101$ and $w = 27$ are used for multilocus informativeness computations at MAF thresholds of 0 and 0.1, respectively. The cluster of boxes for each MAF threshold contains data for multilocus informativeness (left, green), the Cochran-Armitage trend test (middle, blue), and the chi-squared test (right, teal). (a) PCT under the uniform recombination model. (b) PCT under the recombination hotspot model. (c) DIST under the uniform recombination model. (d) DIST under the recombination hotspot model.

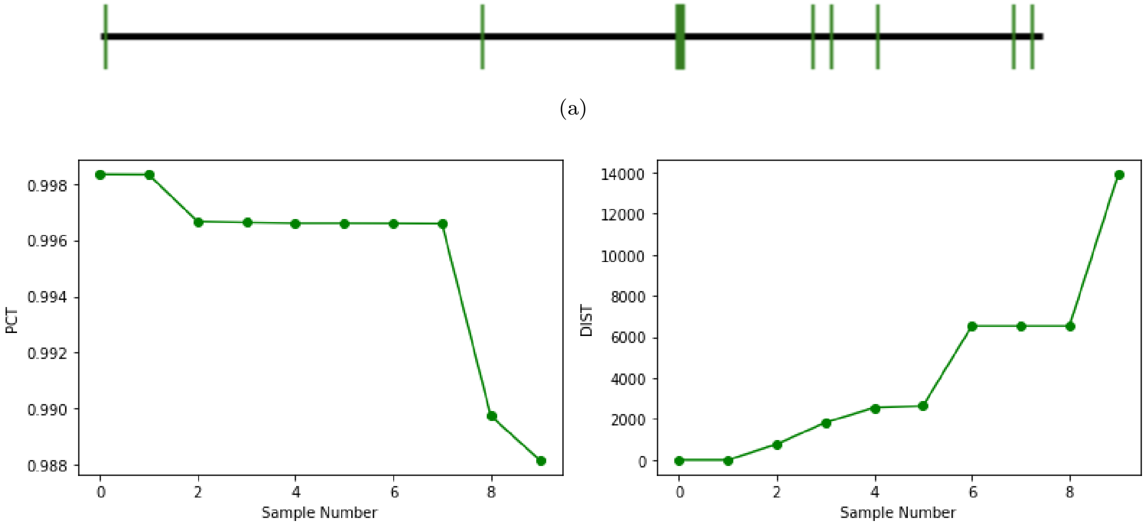
case/control cohort will have the same genotype at a low MAF SNP, this locus will only be able to distinguish a small number of cases and controls from each other. Consequently, the GWAS directed informativeness of other SNPs with respect to the low MAF SNP will often be high because the GWAS distinguishability graph for the low MAF SNP will only contain a small number of edges that the GWAS distinguishability graphs for other SNPs need to cover. Therefore, the low MAF SNP will often be weighted heavily in the computation of multilocus informativeness for other nearby SNPs. This heavy weighting can become problematic in insufficiently large case/control cohorts because rare SNPs that do not truly contribute to the disease phenotype can have a large variance in how often they appear in each study group. For these reasons, it is important to choose an appropriate MAF threshold based on the study size for multilocus informativeness to perform well.

While MAF thresholds significantly affect the performance of multilocus informativeness, they have little effect on the performance of the Cochran-Armitage trend test and the chi-squared test. This result is expected because the scores computed by single-SNP association tests for each SNP do not depend on which other SNPs are considered in the analysis. Consequently, the performance of these mapping methods should generally be consistent across MAF thresholds. In all subsequent analyses in this study, we apply a MAF threshold of 0.1 to the data to ensure that the performance of multilocus informativeness is not significantly worsened by the presence of rare variants in the population. Additionally, we use a window size of $w = 27$ for all subsequent multilocus informativeness calculations, as this window size is also used for the 0.1 MAF threshold here.

4.3 Robustness of Multilocus Informativeness

To verify that multilocus informativeness can robustly identify the genomic structure surrounding disease-causing SNPs, we draw ten different case/control samples from the same simulated population. Each of these samples uses the same disease model parameters but selects a different random disease locus based on those parameters. We set $GRR(Aa) = 3$ and $GRR(aa) = 5$ to ensure that the effect of the disease allele is always large enough to be detected, and we investigate whether multilocus informativeness can consistently locate the disease locus regardless of its location in the candidate region (Figure 3).

Multilocus informativeness locates the disease locus very accurately in all ten samples. Specifically, the causative SNP is placed above the 0.996 percentile in eight of the ten samples, and it is placed above the 0.988 percentile in all ten samples. Furthermore, even when multilocus informativeness places the causative SNP at a relatively lower percentile, the distance from the causative SNP to the highest-scoring SNP is still very small, indicating that the measure is still successfully localizing the disease association signal. Therefore, these results demonstrate that multilocus informativeness is an effective and robust method for fine-mapping disease-causing variants using local genomic structure and association data.



(a)

(b)

(c)

Figure 3: Robustness of multilocus informativeness under the recombination hotspot model. (a) Locations of the randomly selected disease loci across the candidate region. The bolded locus is randomly selected three times, and all other loci are randomly selected once. (b) Percentiles of the causative SNPs in decreasing order. (c) Distances from the causative SNPs to the highest-scoring SNPs in increasing order.

4.4 Comparison of Mapping Methods

4.4.1 Overall Performance

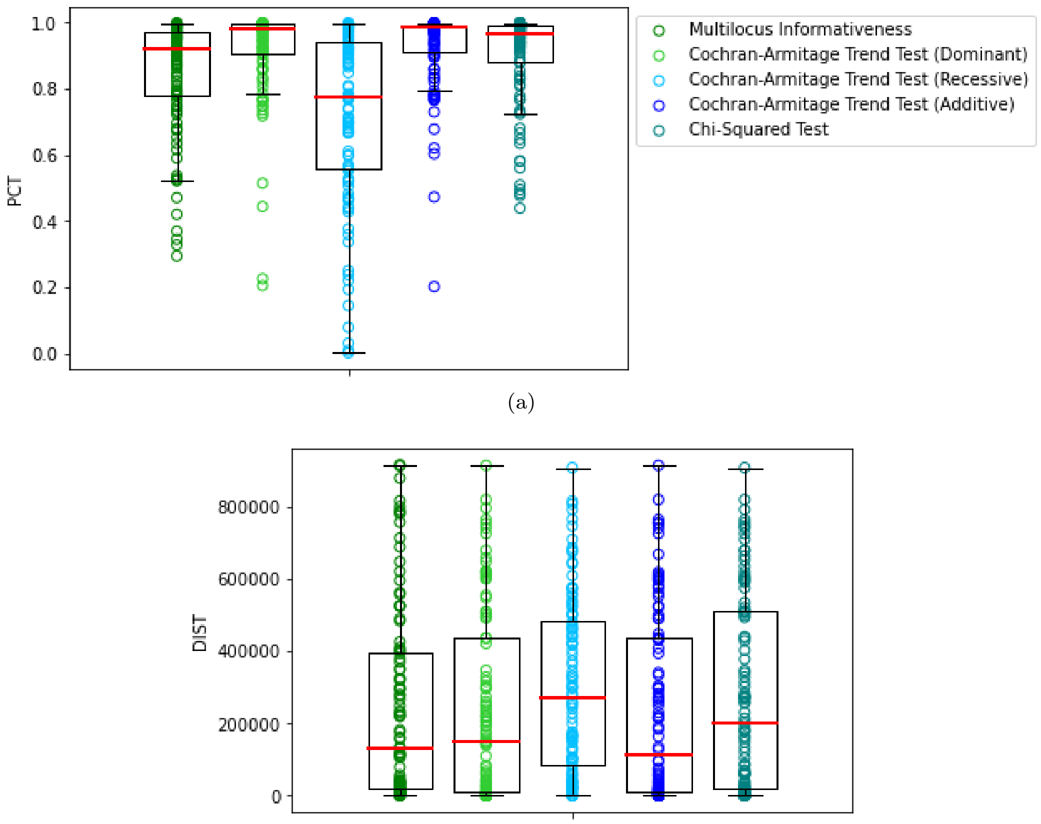
To compare the overall performance of multilocus informativeness to other mapping methods, we simulate 100 replicate populations under the recombination hotspot model (Figure 4). When the percentile of the causative SNP is examined, multilocus informativeness performs the fourth best of the five mapping methods tested, as it only outperforms the Cochran-Armitage trend test with a recessive trend. However, when the distance from the causative SNP to the highest-scoring SNP is investigated, multilocus informativeness compares much more favorably to the other mapping methods. Specifically, multilocus informativeness has the lowest upper quartile and the second lowest median of all the methods tested, and only the Cochran-Armitage trend test with an additive trend has a lower median. These results demonstrate that while multilocus informativeness may not place the causative SNP in the highest percentiles as frequently as some other mapping methods, it still often localizes the disease association signal to the correct area of the candidate region.

Additionally, while the Cochran-Armitage trend test with an additive trend performs the best overall, as would be expected when an additive disease model is used, the Cochran-Armitage trend test with a recessive trend performs the worst overall. This discrepancy demonstrates the importance of choosing an accurate trend when using the Cochran-Armitage trend test, as inaccurate trend choices can significantly diminish statistical power. In this case, the superior performance of the additive and dominant trends compared to the recessive trend can be explained by analyzing the average contingency table for a causative SNP among these replicate study populations (Table 8).

Genotype	**AA**	**Aa**	**aa**
Controls	323.38 ± 11.99	157.68 ± 11.45	18.94 ± 4.43
Cases	285.57 ± 12.09	186.36 ± 11.36	28.07 ± 5.21

Table 8: The average contingency table for the causative SNPs in the study populations analyzed in Figure 4. The disease allele is denoted a, and the mean and standard deviation is shown for each cell in the contingency table.

Based on the simulation parameters used here, many of the excess disease alleles present in the case group appear in heterozygous individuals. Consequently, because the weights vector for a recessive trend does not differentiate between the AA and Aa genotypes, this trend test discards a lot of important data about how the disease allele affects phenotypes, resulting in diminished statistical power. In contrast, because the weights vector for both the additive and dominant trends do differentiate between the AA and Aa genotypes, these trend tests are able to identify the disease locus much more effectively. In all subsequent analyses in this study, we will use an additive trend for the Cochran-Armitage trend test, which should optimize the performance of this method under the additive disease models simulated here.



(a)

(b)

Figure 4: Performance of various mapping methods. Colored circles show performance on individual populations, red lines show median values of the examined metric, boxes extend to the upper and lower quartiles of the data, and whiskers extend up to 1.5 times the interquartile range away from the median. From left to right, the mapping methods examined are multilocus informativeness (green), the Cochran-Armitage trend test with dominant trend (light green), the Cochran-Armitage trend test with recessive trend (light blue), the Cochran-Armitage trend test with additive trend (blue), and the chi-squared test (teal). (a) Percentiles of the causative SNPs. (b) Distances from the causative SNPs to the highest-scoring SNPs.

4.4.2 Correlations in Performance

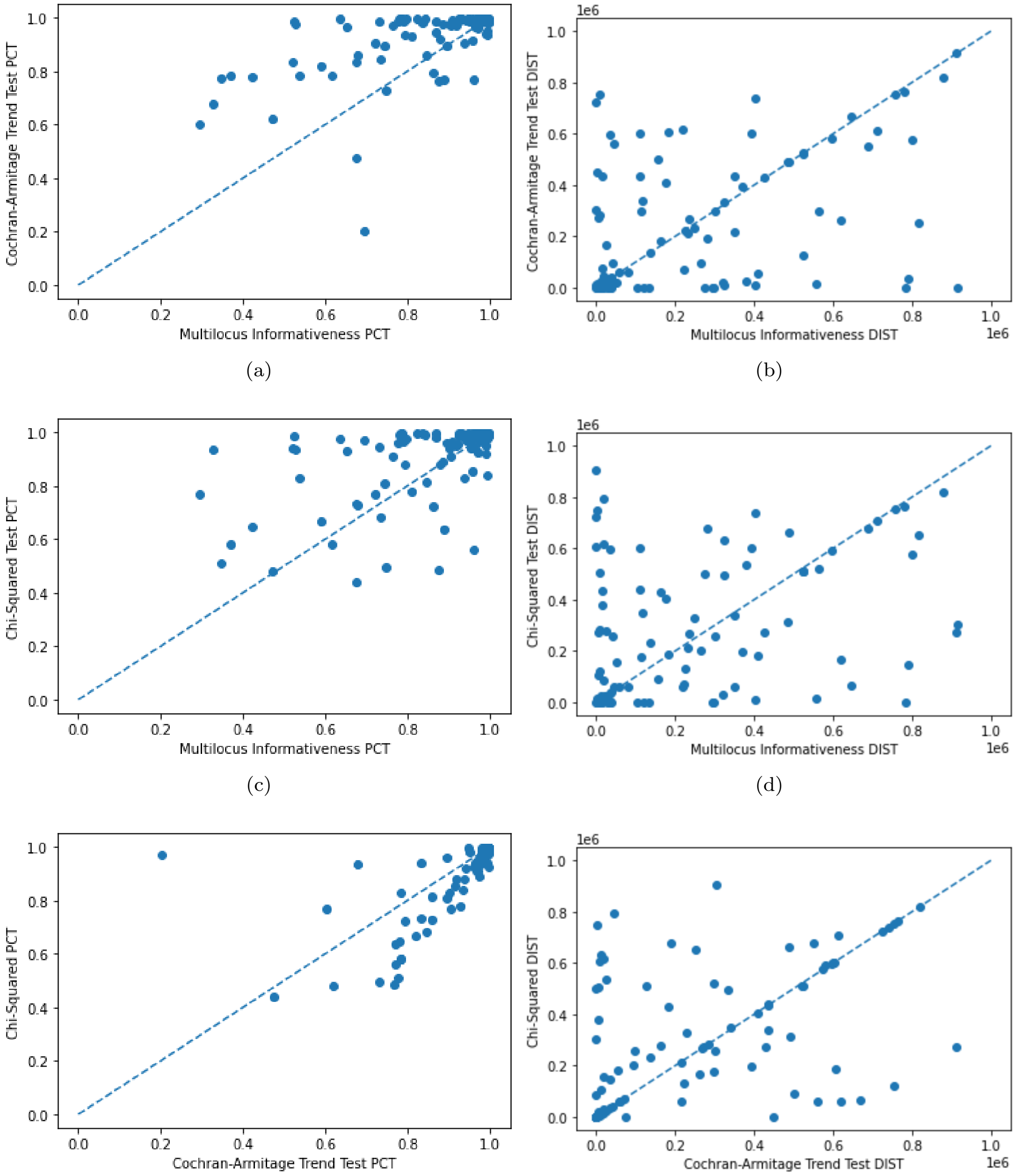
To determine whether different mapping methods are more effective in different study populations, we generate correlation plots that compare the performance of the mapping methods on individual case/control cohorts (Figure 5). When the percentile of the causative SNP is examined, the Cochran-Armitage trend test regularly outperforms both multilocus informativeness and the chi-squared test. Notably, the superior performance of the Cochran-Armitage trend test over the chi-squared test is consistent with the observation that the Cochran-Armitage trend test has higher statistical power than the chi-squared test when the investigated trend is accurate, as is the case here. Additionally, the chi-squared test also outperforms multilocus informativeness more often than not, although there are still several examples where multilocus informativeness is more effective.

Furthermore, when the distance from the causative SNP to the highest-scoring SNP is examined, there are fewer clear trends in which mapping method is most successful. Specifically, there are many study populations in which each mapping method outperforms each other mapping method, suggesting that each mapping method may perform best on study populations with different characteristics. Future work should investigate this hypothesis further and attempt to identify specific attributes of study populations that may correlate with the superior performance of a particular mapping method.

4.5 Effects of Altering Simulation Parameters

Finally, to identify any major trends in how the simulation parameters affect the performance of the mapping methods, we examine the effectiveness of each method when a variety of parameters are individually altered to a range of different values. Figure 6 depicts the effects of modifying the parameters of the forward simulation and the sampling algorithm, and Figure 7 depicts the effects of modifying the parameters of the disease model. For each set of parameters tested, ten replicate populations were generated for examination.

We observe that higher recombination rates generally improve the percentile of the causative SNP, a trend that likely reflects the difficulty in pinpointing the exact causative SNP in regions with low recombination rates and high LD. However, we also observe that the distance from the causative SNP to the highest-scoring SNP is generally smaller for lower recombination rates under the recombination hotspot model, which may indicate that recombination hotspots produce clearer LD blocks when the recombination rate outside of the hotspots is smaller. Additionally, we notice that the percentile of the causative SNP generally increases as the number of cases and controls increases, as would be expected. Interestingly, increasing the number of cases and controls in the study population only seems to decrease the distance from the causative SNP to the highest-scoring SNP when multilocus informativeness is used. This observation suggests that of the mapping methods examined here, multilocus informativeness may be the most effective at using additional cases and controls to further



(a) (b)

(c) (d)

(e) (f)

Figure 5: Correlations in performance of mapping methods on the study populations analyzed in Figure 4. (a)-(b) PCT and DIST of multilocus informativeness and the Cochran-Armitage trend test. (c)-(d) PCT and DIST of multilocus informativeness and the chi-squared test. (e)-(f) PCT and DIST of the Cochran-Armitage trend test and the chi-squared test.

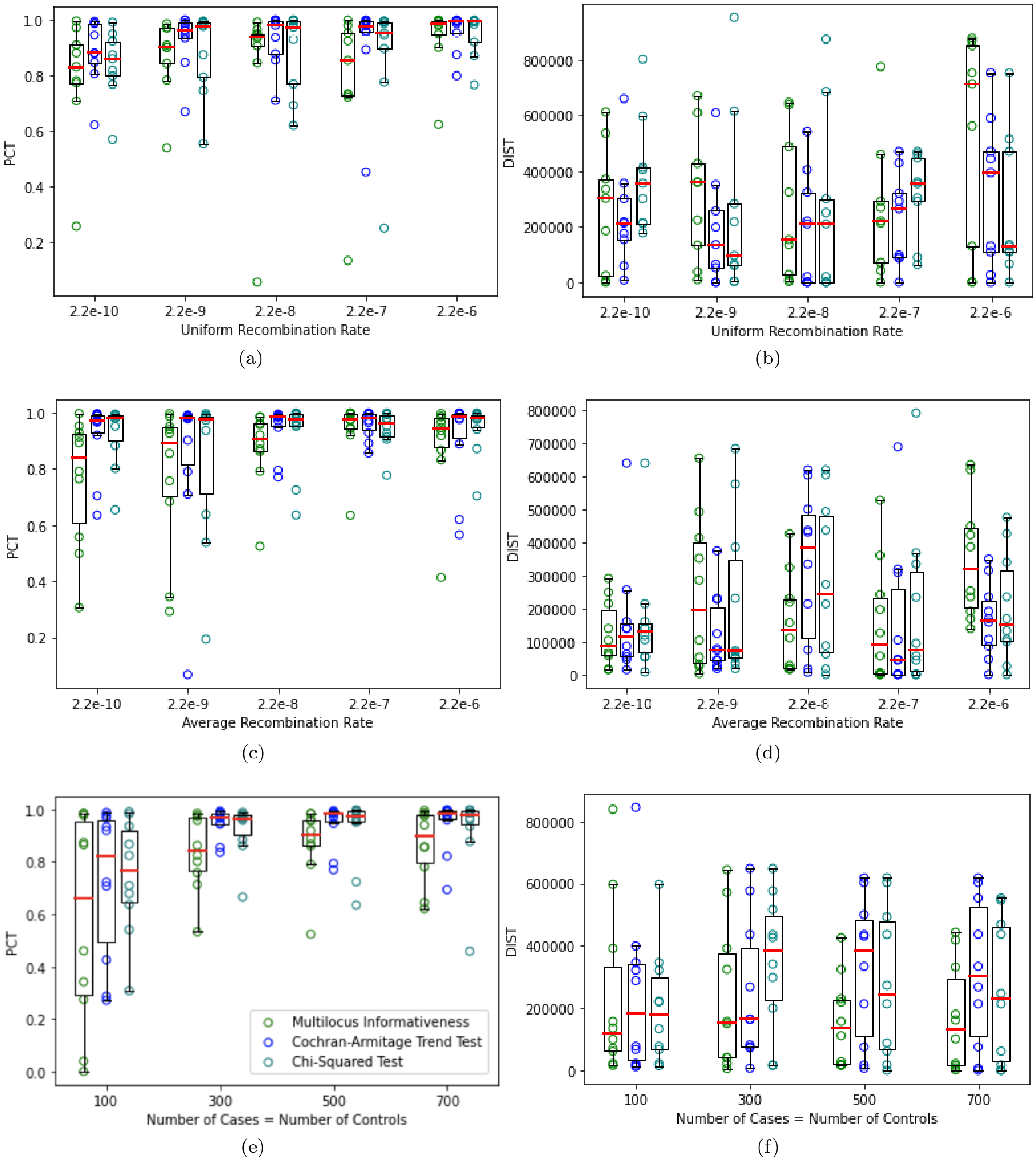
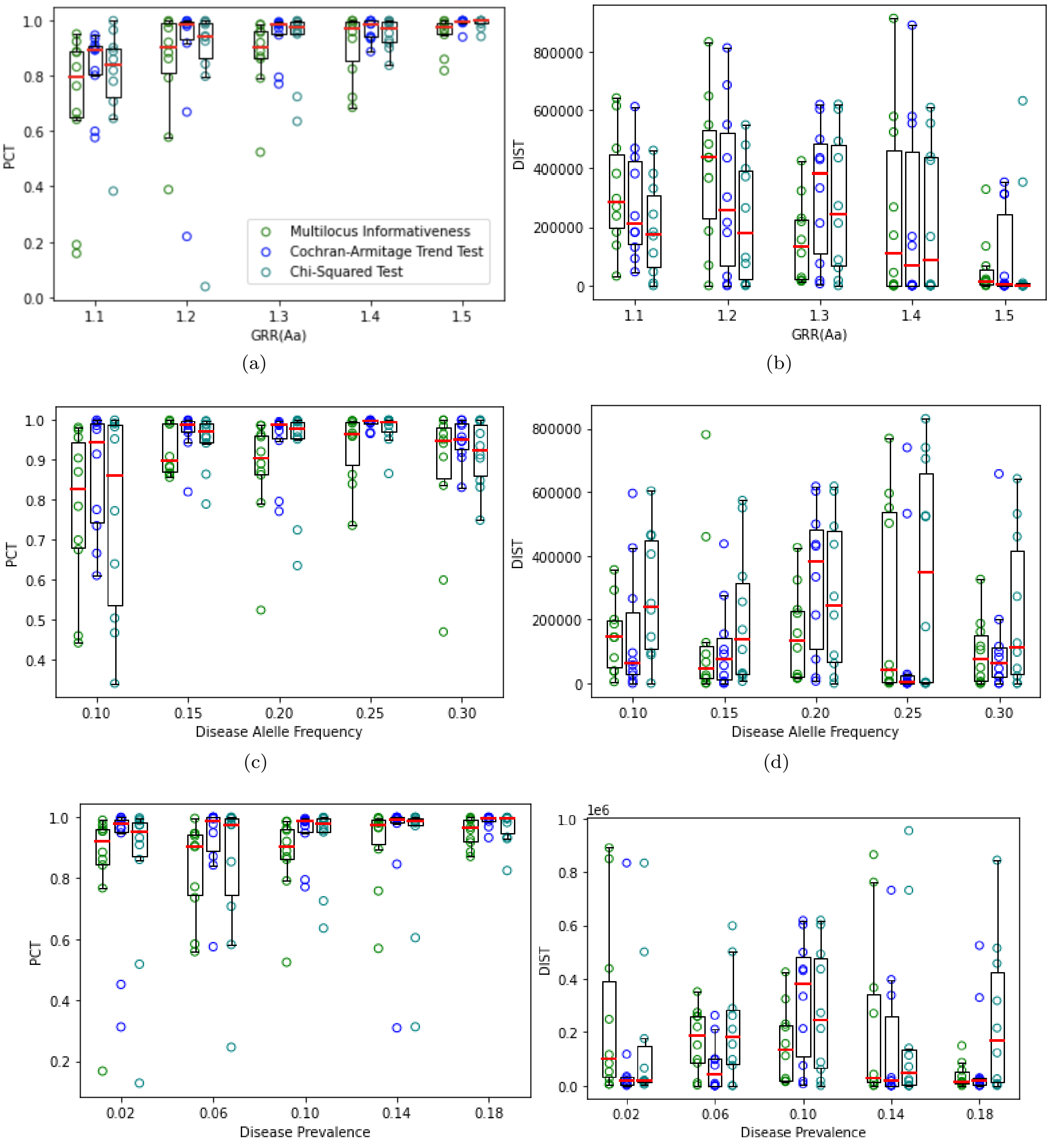


Figure 6: Effects of altering parameters of the forward simulation and the sampling algorithm. Colored circles show performance on individual populations, red lines show median values of the examined metric, boxes extend to the upper and lower quartiles of the data, and whiskers extend up to 1.5 times the interquartile range away from the median. Each cluster of boxes contains data for multilocus informativeness (left, green), the Cochran-Armitage trend test (middle, blue), and the chi-squared test (right, teal). (a)-(b) PCT and DIST under the uniform recombination model with various uniform recombination rates r_U. (c)-(d) PCT and DIST under the recombination hotspot model with various average recombination rates r_U. (e)-(f) PCT and DIST under the recombination hotspot model with various numbers of cases and controls $n_D = n_C$.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 7: Effects of altering parameters of the disease model under the recombination hotspot model. Colored circles show performance on individual populations, red lines show median values of the examined metric, boxes extend to the upper and lower quartiles of the data, and whiskers extend up to 1.5 times the interquartile range away from the median. Each cluster of boxes contains data for multilocus informativeness (left, green), the Cochran-Armitage trend test (middle, blue), and the chi-squared test (right, teal). (a)-(b) PCT and DIST with various genotype relative risks $GRR(Aa)$ and an additive disease model. (c)-(d) PCT and DIST with various disease allele frequencies q. (e)-(f) PCT and DIST with various disease prevalence values p.

localize the region containing the causative SNP.

Furthermore, as the genotype relative risk at the disease locus increases, the percentile of the causative SNP generally increases and the distance from the causative SNP to the highest-scoring SNP generally decreases. These trends are expected given that disease loci with higher genotype relative risks have larger effects on the disease phenotype and are therefore easier to detect. Additionally, we notice that the percentile of the causative SNP is highest for intermediate disease allele frequencies. We believe that the difficulty in identifying rare causative SNPs likely results from these SNPs not appearing often enough in the study populations to have very strong disease association signals. Similarly, we believe that the difficulty in identifying very common causative SNPs likely results from these SNPs appearing too frequently in the control groups to have very strong disease association signals. Finally, we do not observe any significant trends in how disease prevalence affects the performance of the mapping methods.

5 Conclusion

In this study, we have found that multilocus informativeness is an effective method for fine-mapping disease-causing variants in GWAS using both LD and association data. We have determined that applying an appropriate MAF threshold is essential to ensure that multilocus informativeness performs well, and we have demonstrated that this method is sufficiently robust to detect multiple different disease associations in the same population. Moreover, while multilocus informativeness sometimes does not place the causative SNP in as high a percentile as other mapping methods, multilocus informativeness can often localize the disease association signal to a small area of the candidate region more effectively than alternative approaches.

Finally, we would like to highlight several important limitations of this study. First, we only compare the performance of multilocus informativeness to existing single-SNP tests of association. Future work should also investigate the performance of alternative fine-mapping methods such as those based on penalized regression, Bayesian statistics, and genome annotation. Second, we only analyze relatively small study populations, as most analyses presented here examine case/control cohorts with 500 individuals in each study group. Future studies should explore the performance of these mapping methods in much larger study populations, as modern GWAS can include tens of thousands of individuals. And third, we only inspect the performance of multilocus informativeness on simulated datasets. Future work should analyze the performance of multilocus informativeness on real GWAS datasets, as real data often contain more complex patterns and interactions that can make identifying disease-causing SNPs more difficult.

In conclusion, despite the limitations highlighted here, the favorable results obtained in this study demonstrate that multilocus informativeness is a promising new method for effectively fine-mapping disease-causing variants in GWAS.

6 References

[1] Slatkin, Montgomery (2008). "Linkage disequilibrium — understanding the evolutionary past and mapping the medical future", *Nature Reviews Genetics*, 9(6):477-485.

[2] Balding, David J. (2006). "A tutorial on statistical methods for population association studies", *Nature Reviews Genetics*, 7(10):781-791.

[3] Clarke, Geraldine M., et al. (2011). "Basic statistical analysis in genetic case-control studies", *Nature Protocols*, 6(2):121-133.

[4] Schaid, David J., Wenan Chen, and Nicholas B. Larson (2018). "From genome-wide associations to candidate causal variants by statistical fine-mapping", *Nature Reviews Genetics*, 19(8):491-504.

[5] Cochran, William G. (1954). "Some Methods for Strengthening the Common χ^2 Tests", *Biometrics*, 10(4):417-451.

[6] Armitage, P. (1955). "Tests for Linear Trends in Proportions and Frequencies", *Biometrics*, 11(3):375-386.

[7] Halldórsson, Bjarni V., et al. (2004). "Optimal Haplotype Block-Free Selection of Tagging SNPs for Genome-Wide Association Studies", *Genome Research*, 14(8):1633-1640.

[8] Tarpine, Ryan, Fumei Lam, and Sorin Istrail (2011). "Conservative Extensions of Linkage Disequilibrium Measures from Pairwise to Multi-Loci and Algorithms for Optimal Tagging SNP Selection", *Research in Computational Molecular Biology*, 468-482.

[9] Minichiello, Mark J. and Richard Durbin (2006). "Mapping Trait Loci by Use of Inferred Ancestral Recombination Graphs", *American Journal of Human Genetics*, 79(5):910-922.

[10] Haller, Benjamin C. and Philipp W. Messer (2019). "SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model", *Molecular Biology and Evolution*, 36(3):632-637.

[11] Carlson, Christopher S., et al. (2004). "Selecting a Maximally Informative Set of Single-Nucleotide Polymorphisms for Association Analyses Using Linkage Disequilibrium", *American Journal of Human Genetics*, 74(1):106-120.

[12] Marees, Andrew T., et al. (2018). "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis", *International Journal of Methods in Psychiatric Research*, 27(2):e1608.