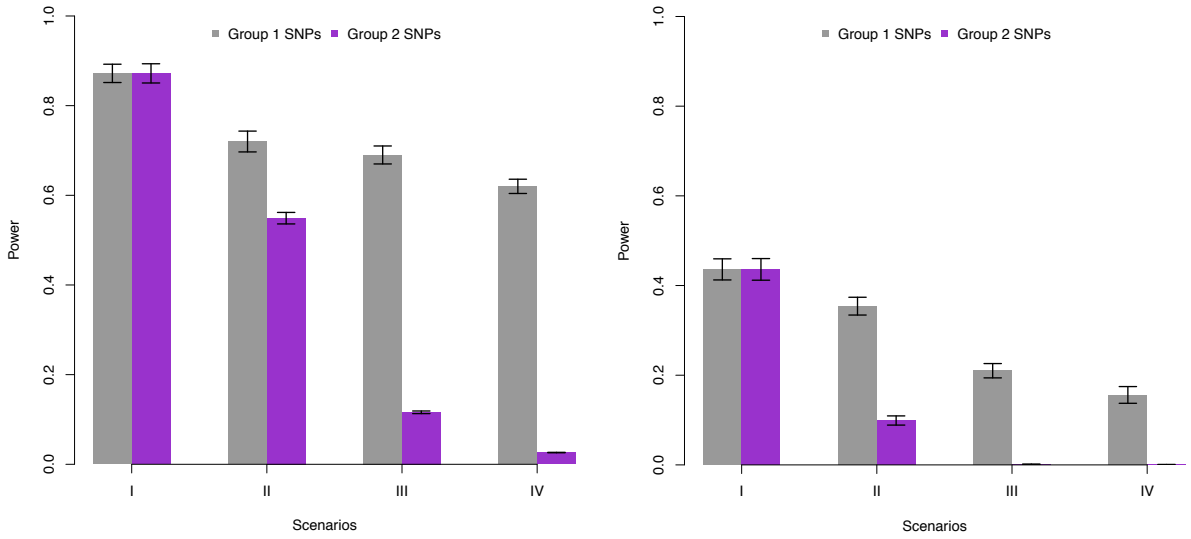
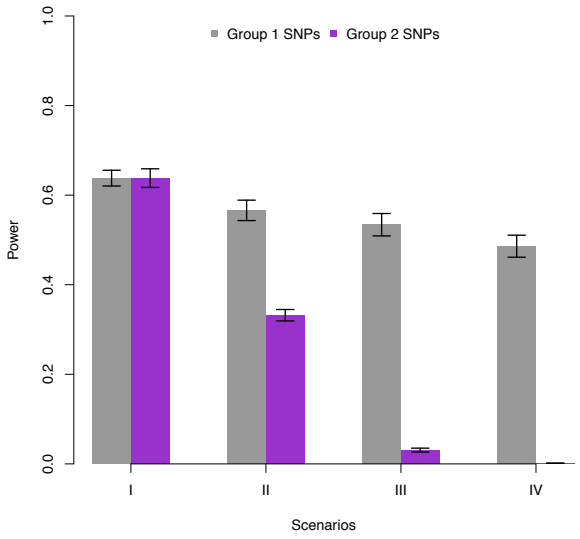
Supplementary Figures

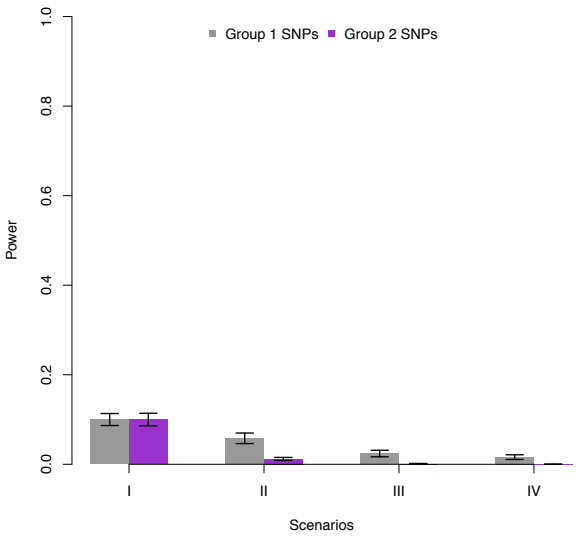


(a) $\gamma = 0.1\%$ and $\rho = 0.5$

(b) $\gamma = 0.1\%$ and $\rho = 0.8$

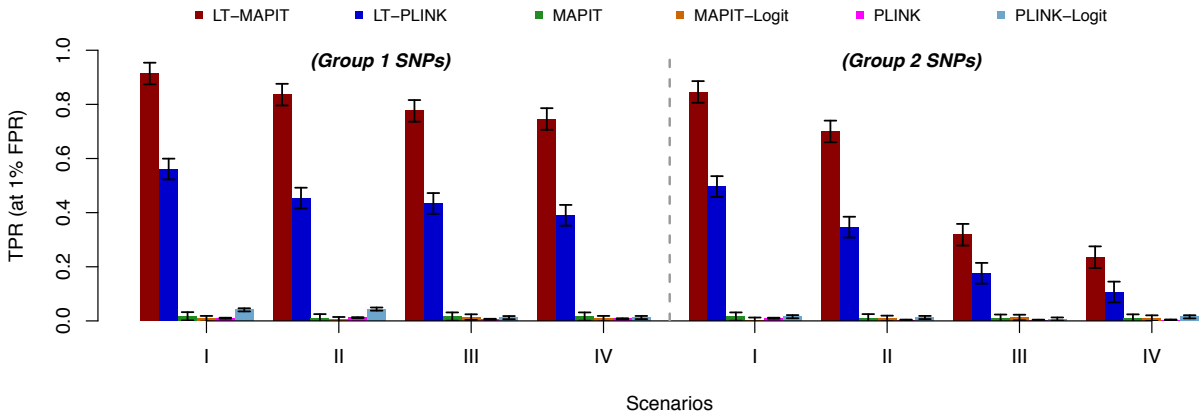


(c) $\gamma = 0.5\%$ and $\rho = 0.5$

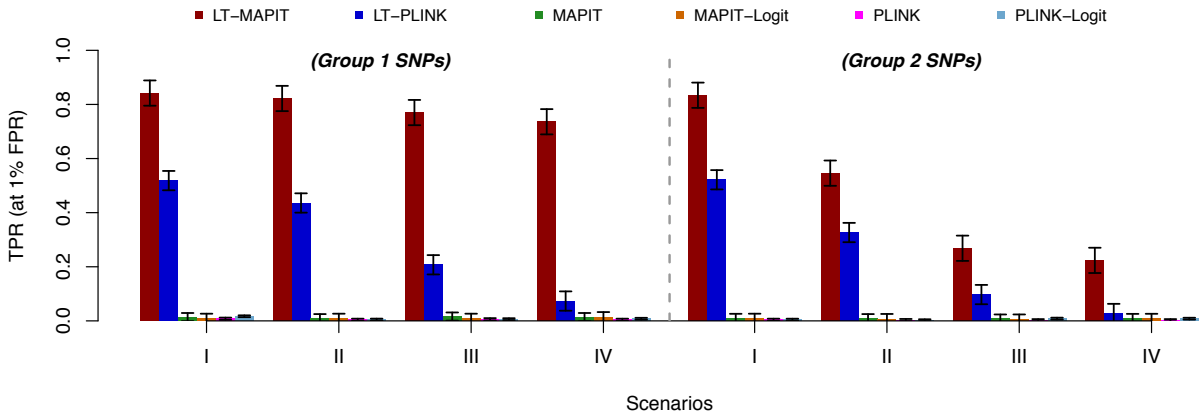


(d) $\gamma = 0.5\%$ and $\rho = 0.8$

Figure S1. Empirical power to detect simulated causal interacting makers at a genome-wide significance level. Group 1 and 2 causal markers are colored in grey and purple, respectively. Figures show the power of LT-MAPIT to identify SNPs in each causal group under the Bonferroni-corrected significance level $\alpha = 8.3 \times 10^{-6}$. Phenotypes were generated on the liability scale with disease prevalence in the population set to $\gamma = 0.1\%$ in Figures (a) and (b), and $\gamma = 0.5\%$ in Figures (c) and (d). Here, the parameters $\rho = 0.5$ in (a) and (c), and $\rho = 0.8$ in (b) and (d), are used to determine the proportion of phenotypic variance (on the liability scale) that is contributed by interaction effects. Results are based on 100 simulated replicates and the lines on each bar represent a 95% standard error interval due to resampling.

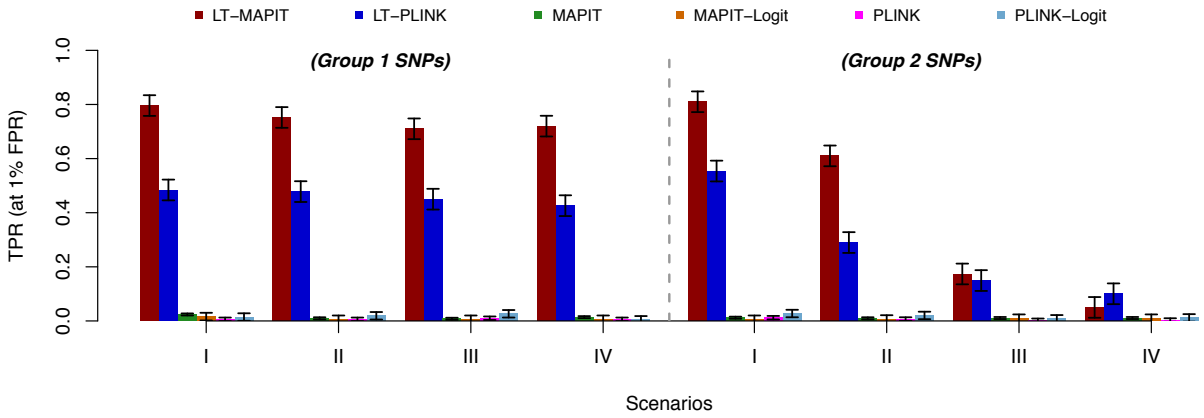


(a) $\gamma = 0.1\%$ and $\rho = 0.5$

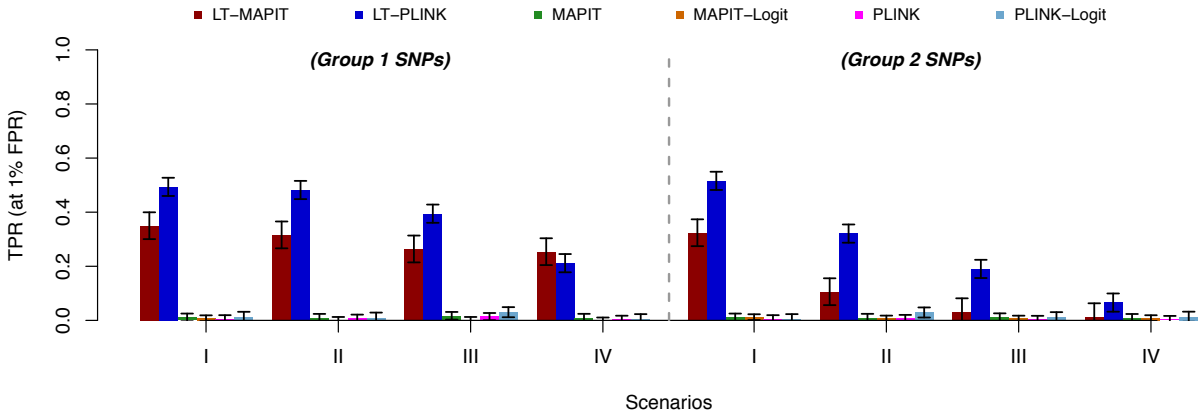


(b) $\gamma = 0.1\%$ and $\rho = 0.8$

Figure S2. Power comparison for detecting group 1 and group 2 causal SNPs. The association mapping ability of LT-MAPIT (red) is compared to five competing methods: (i) the original linear mixed model version of MAPIT (green), (ii) the binary mixed model MAPIT-Logit (orange) fit by using penalized quasi-likelihoods (PQL), (iii) the linear exhaustive search model PLINK (pink), (iv) the generalized linear model PLINK-Logit (light blue), and (v) the liability threshold model LT-PLINK (blue). Each approach is assessed under four simulation scenarios. Phenotypes were generated on the liability scale with disease prevalence in the population set to $\gamma = 0.1\%$. Here, the parameters $\rho = 0.5$ in Figure (a) and $\rho = 0.8$ in Figure (b) are used to determine the proportion of phenotypic variance (on the liability scale) that is contributed by interaction effects. The y-axis gives the rate at which true causal variants were identified (TPR) at a 1% false positive rate (FPR). Results are based on 100 simulated replicates and the lines on each bar represent a 95% standard error interval due to resampling.

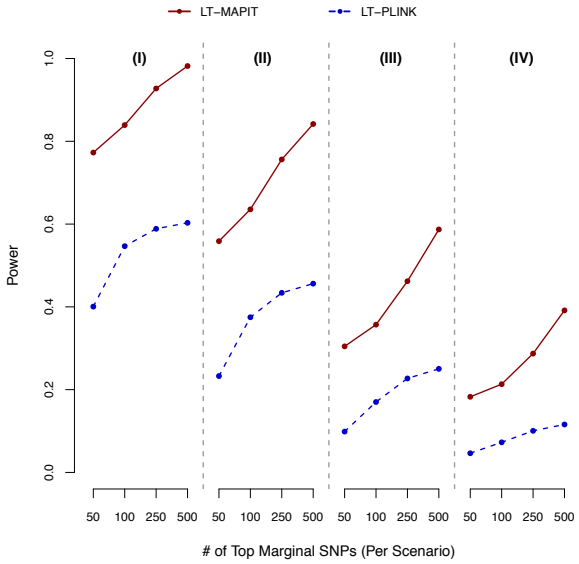


(a) $\gamma = 0.5\%$ and $\rho = 0.5$

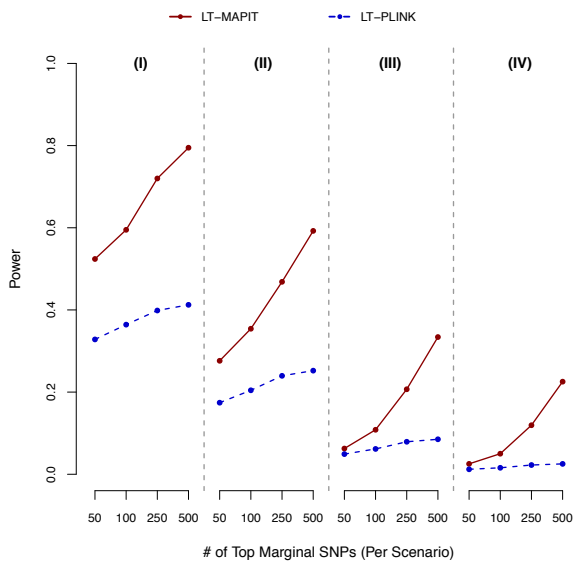


(b) $\gamma = 0.5\%$ and $\rho = 0.8$

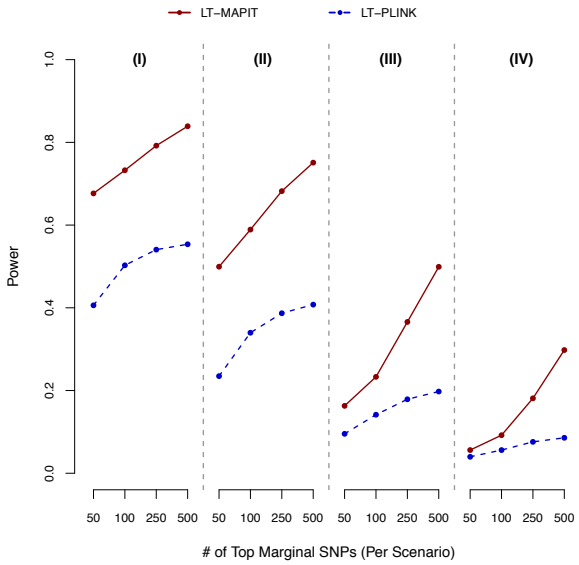
Figure S3. Power comparison for detecting group 1 and group 2 causal SNPs. The association mapping ability of LT-MAPIT (red) is compared to five competing methods: (i) the original linear mixed model version of MAPIT (green), (ii) the binary mixed model MAPIT-Logit (orange) fit by using penalized quasi-likelihoods (PQL), (iii) the linear exhaustive search model PLINK (pink), (iv) the generalized linear model PLINK-Logit (light blue), and (v) the liability threshold model LT-PLINK (blue). Each approach is assessed under four simulation scenarios. Phenotypes were generated on the liability scale with disease prevalence in the population set to $\gamma = 0.5\%$. Here, the parameters $\rho = 0.5$ in Figure (a) and $\rho = 0.8$ in Figure (b) are used to determine the proportion of phenotypic variance (on the liability scale) that is contributed by interaction effects. The y-axis gives the rate at which true causal variants were identified (TPR) at a 1% false positive rate (FPR). Results are based on 100 simulated replicates and the lines on each bar represent a 95% standard error interval due to resampling.



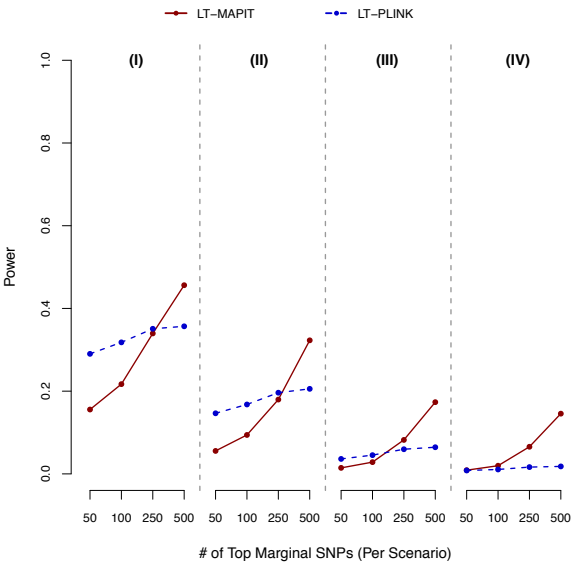
(a) $\gamma = 0.1\%$ and $\rho = 0.5$



(b) $\gamma = 0.1\%$ and $\rho = 0.8$

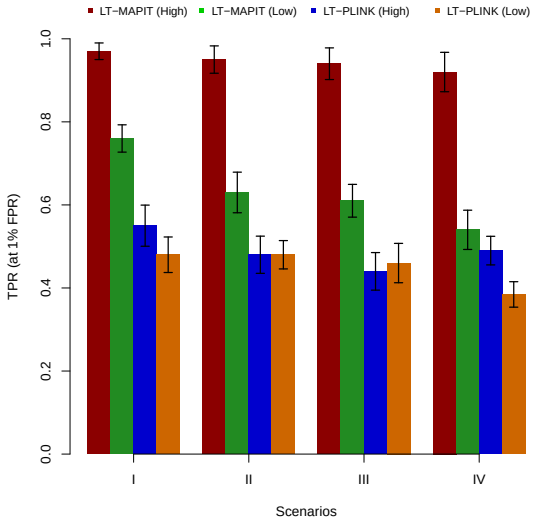


(c) $\gamma = 0.5\%$ and $\rho = 0.5$

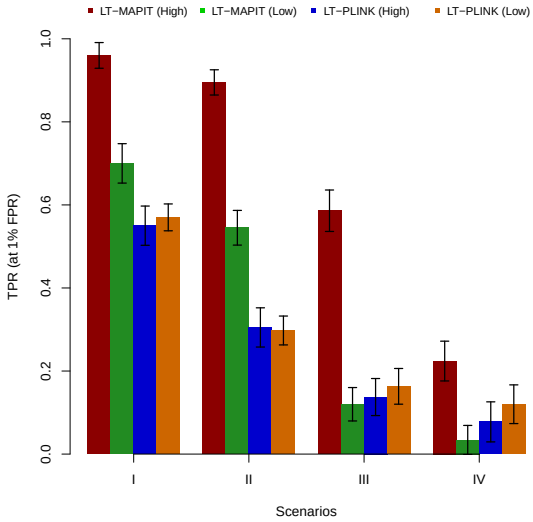


(d) $\gamma = 0.5\%$ and $\rho = 0.8$

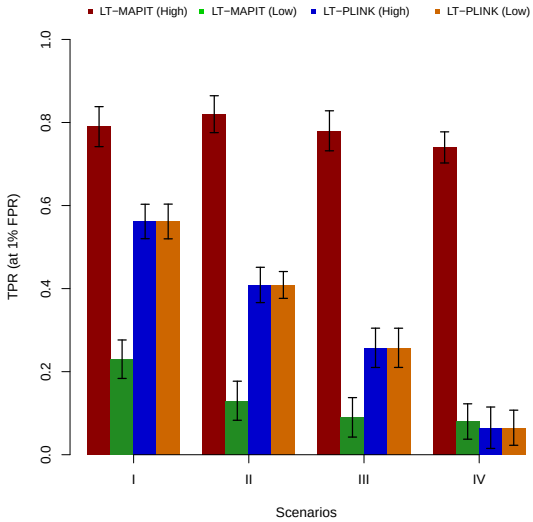
Figure S4. Power comparison with exhaustive search procedures to detect epistatic pairs. LT-MAPIT (red) is assessed as an initial step in a pairwise epistatic filtration process compared against the more conventional exhaustive search LT-PLINK (blue), which serves as a baseline. While using LT-MAPIT, the search for epistatic pairs occurs between the top ranked $\{50, 100, 250, 500\}$ significant marginally associated SNPs. Both methods are evaluated under four simulation scenarios. Phenotypes were generated on the liability scale with disease prevalence in the population set to $\gamma = 0.1\%$ in Figures (a) and (b), and $\gamma = 0.5\%$ in Figures (c) and (d). Here, the parameters $\rho = 0.5$ in (a) and (c), and $\rho = 0.8$ in (b) and (d), are used to determine the proportion of phenotypic variance (on the liability scale) that is contributed by interaction effects. The y-axis gives the rate at which true causal epistatic pairs were identified. Results are based on an average performance across 100 simulated replicates.



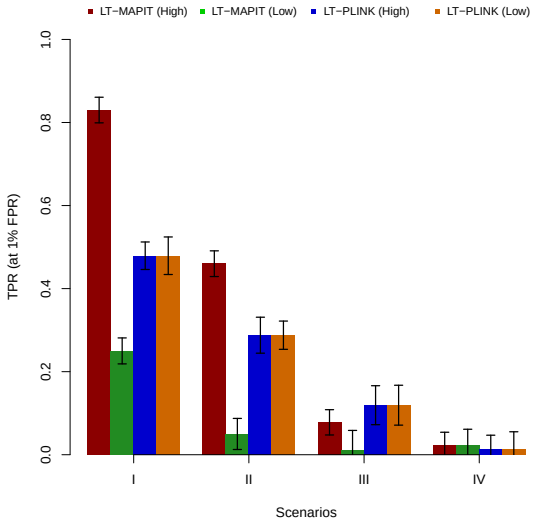
(a) Group 1 SNPs ($\rho = 0.5$)



(b) Group 2 SNPs ($\rho = 0.5$)

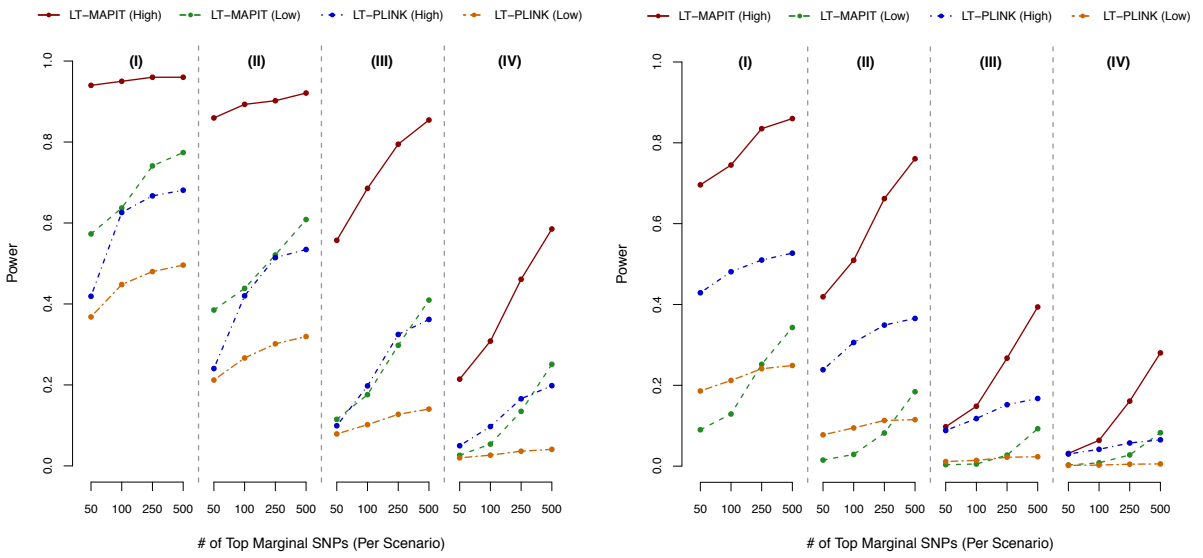


(c) Group 1 SNPs ($\rho = 0.8$)



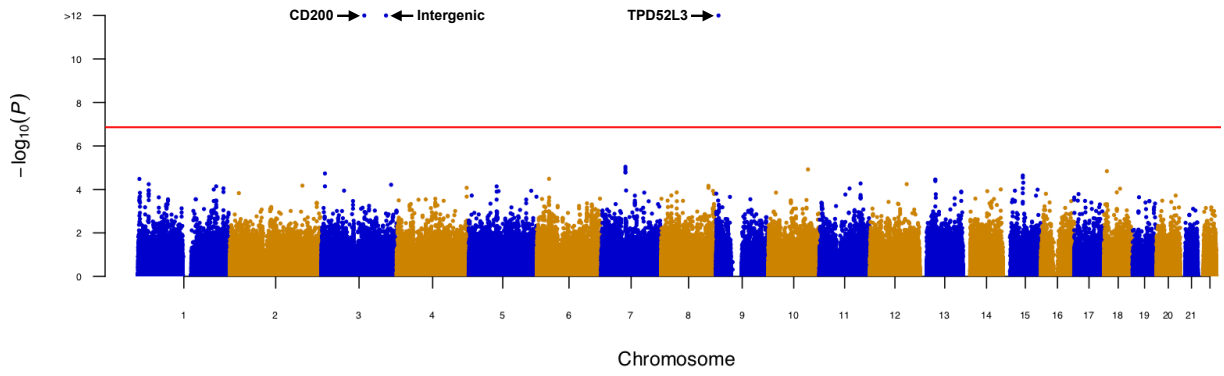
(d) Group 2 SNPs ($\rho = 0.8$)

Figure S5. Power comparison for detecting group 1 and group 2 causal SNPs when the assumed disease prevalence (γ) is misspecified. In simulations, the true disease prevalence in the population is set to $\gamma = 0.25\%$. Compared are two sets of liability threshold (LT) regression models that are fit by incorrectly assuming $\widetilde{\gamma} = 0.1\%$ (low) and $\widetilde{\gamma} = 0.5\%$ (high), respectively. Specifically, LT-MAPIT (red and green) is compared to LT-PLINK (blue and orange) across all four simulation scenarios. Assessments for the group 1 SNPs are given in Figures (a) and (c), while evaluations for the group 2 SNPs are in Figures (b) and (d). The parameters $\rho = 0.5$ in (a) and (c), and $\rho = 0.8$ in (b) and (d), are used to determine the proportion of phenotypic variance (on the liability scale) that is contributed by interaction effects. The y-axis gives the rate at which true causal variants were identified (TPR) at a 1% false positive rate (FPR). Results are based on 100 replicates. The lines represent 95% confidence interval due to simulated replicates.

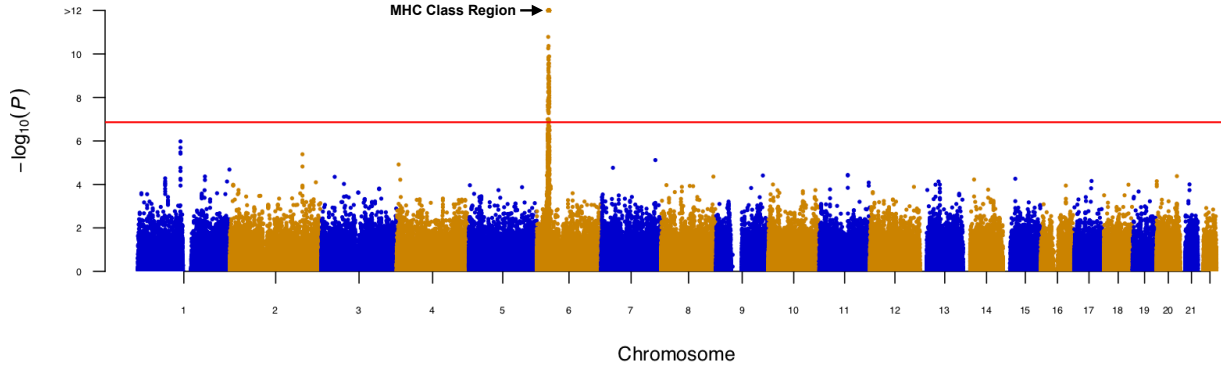


(a) $\rho = 0.5$ (b) $\rho = 0.8$

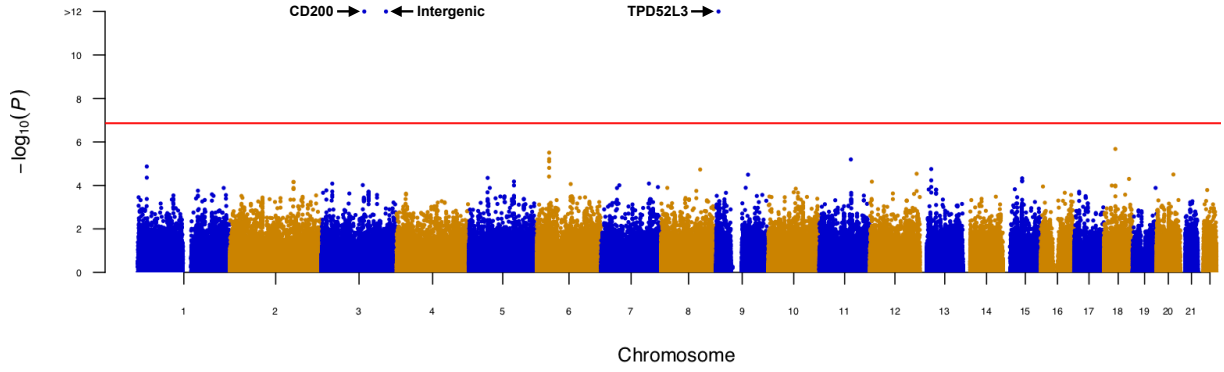
Figure S6. Power comparison with exhaustive search procedures to detect epistatic pairs when the assumed disease prevalence (γ) is misspecified. In simulations, the true disease prevalence in the population is set to $\gamma = 0.1\%$. Compared are two sets of liability threshold (LT) regression models that are fit by incorrectly assuming $\widetilde{\gamma} = 0.1\%$ (low) and $\widetilde{\gamma} = 0.5\%$ (high), respectively. Specifically, LT-MAPIT (red and green) is assessed as an initial step in a pairwise epistatic filtration process compared against the more conventional exhaustive search LT-PLINK (blue and orange), which serves as a baseline. While using LT-MAPIT, the search for epistatic pairs occurs between the top ranked $\{50, 100, 250, 500\}$ significant marginally associated SNPs. Both methods are evaluated under four simulation scenarios. The parameters $\rho = 0.5$ in Figure (a) and $\rho = 0.8$ in Figure (b) are used to determine the proportion of phenotypic variance (on the liability scale) that is contributed by interaction effects. The y-axis gives the rate at which true causal epistatic pairs were identified. Results are based on an average performance across 100 simulated replicates.



(a) Results with $\mathbf{K}_{GW}$

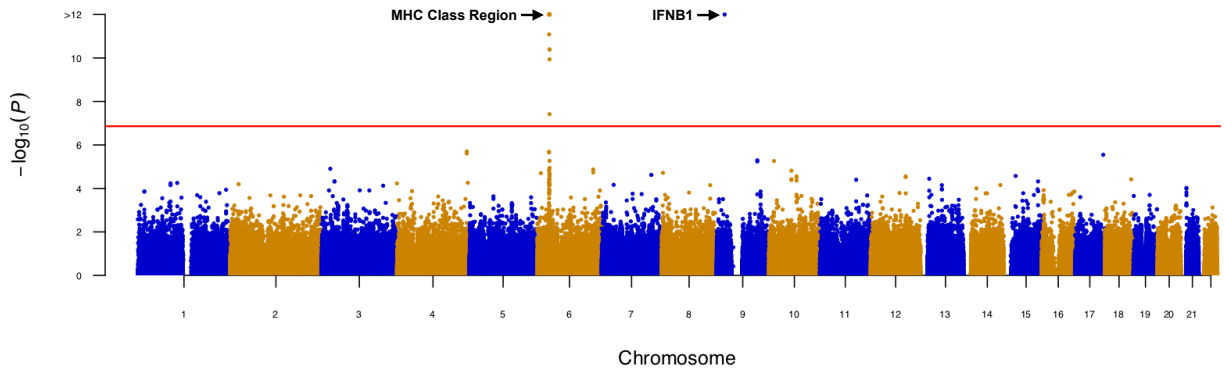


(b) Results with $\mathbf{K}_{cis}$

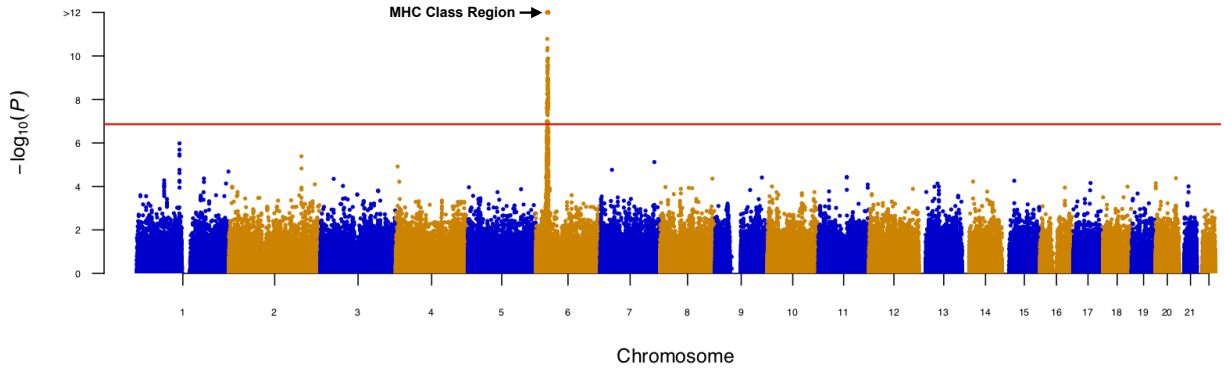


(c) Results with $\mathbf{K}_{trans}$

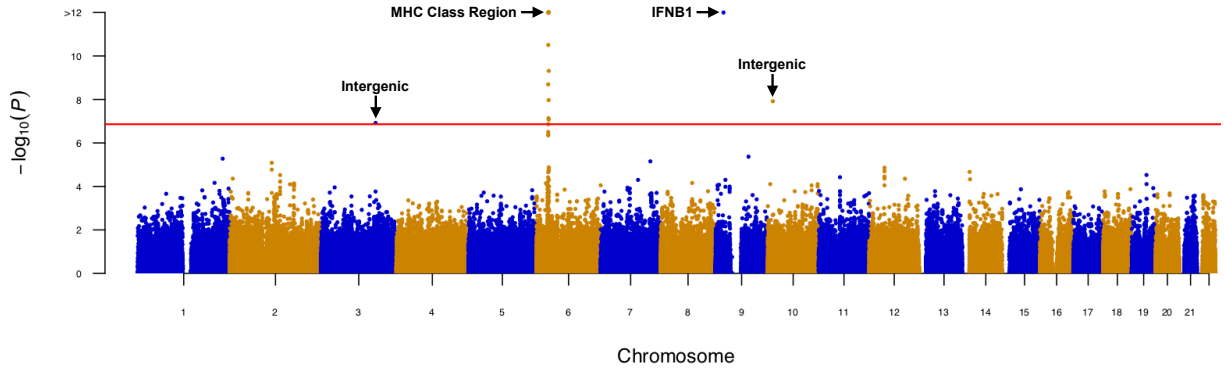
Figure S7. Genome-wide scans for marginal epistatic effects in rheumatoid arthritis (RA) from the WTCCC data set. Depicted are the $-\log_{10}(P)$ transformed LT-MAPIT p-values of quality-control-positive SNPs plotted against their genomic positions. Note that LT-MAPIT was implemented with three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ in Figure (a), a *cis*-based matrix $\mathbf{K}_{cis}$ in Figure (b), and a *trans*-based matrix $\mathbf{K}_{trans}$ in Figure (c), respectively. The (red) horizontal line indicates a Bonferroni-corrected significance threshold $\alpha = 1.37 \times 10^{-7}$. Note that all panels are truncated at $-\log_{10}(P) = 12$ for consistency and presentation, although for some SNPs have stronger marginally epistatic associations with p-values $P \approx 0$ (indicated by > 12). Gene annotations are derived from the dbSNP database provided by the National Center for Biotechnology Information.



(a) Results with $\mathbf{K}_{GW}$

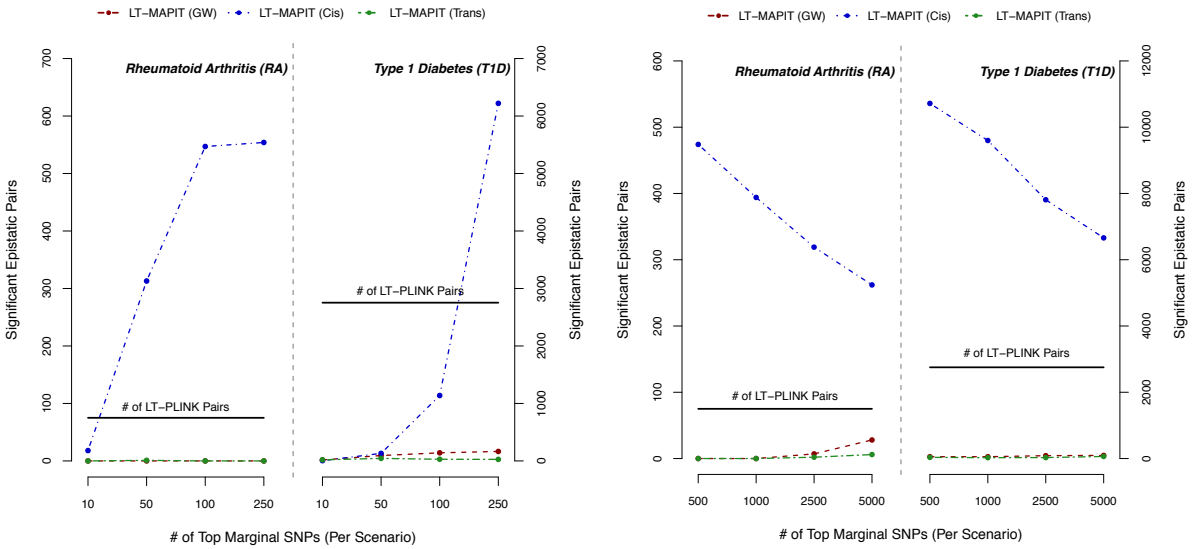


(b) Results with $\mathbf{K}_{cis}$



(c) Results with $\mathbf{K}_{trans}$

Figure S8. Genome-wide scans for marginal epistatic effects in type 1 diabetes (T1D) from the WTCCC data set. Depicted are the $-\log_{10}(P)$ transformed LT-MAPIT p-values of quality-control-positive SNPs plotted against their genomic positions. Note that LT-MAPIT was implemented with three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ in Figure (a), a *cis*-based matrix $\mathbf{K}_{cis}$ in Figure (b), and a *trans*-based matrix $\mathbf{K}_{trans}$ in Figure (c), respectively. The (red) horizontal line indicates a Bonferroni-corrected significance threshold $\alpha = 1.37 \times 10^{-7}$. Note that all panels are truncated at $-\log_{10}(P) = 12$ for consistency and presentation, although for some SNPs have stronger marginally epistatic associations with p-values $P \approx 0$ (indicated by > 12). Gene annotations are derived from the dbSNP database provided by the National Center for Biotechnology Information.



(a) Low-Dimensional Filter

(b) High-Dimensional Filter

Figure S9. Power comparison with exhaustive search procedures to detect epistatic pairs for rheumatoid arthritis (RA) and type 1 diabetes (T1D) from the WTCCC data set. Figures depict the number of significant pairwise interactions (y-axis) identified by LT-MAPIT when using three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ (red), a *cis*-based matrix $\mathbf{K}_{cis}$ (blue), and a *trans*-based matrix $\mathbf{K}_{trans}$ (green), respectively. Figure (a) is based on searching between the top $v = \{10, 50, 100, 200\}$ marginally associated variants, while results in Figure (b) considers the top $v = \{500, 1000, 2000, 5000\}$ variants (labeled on the x-axis). We use the number of significant pairs identified by fully exhaustive search model LT-PLINK as a baseline comparison (solid black line). An interaction found by using the LT-MAPIT filtration was deemed signifiant if it had a joint p-value below the Bonferroni-corrected threshold $\alpha = 0.05/C(v, 2)$ — where $C(\bullet, 2)$ denotes the binomial coefficient enumerating all possible SNP pairs. In the case of LT-PLINK, we consider two variants to be a significantly associated epistatic pair if they have a joint p-value below the threshold $\alpha = 7.6 \times 10^{-13}$, which corresponds to the Bonferroni-correction used after examining all possible genome-wide SNP pairs across the data set. Overall, at this level, LT-PLINK 75 detected significant pairs in RA, and 2753 significant pairs in T1D.

Supplementary Tables

Table S1. Disease prevalences (γ) for each trait in the WTCCC data set. The seven diseases analyzed include: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). References denote the past work from which these values were taken.

	BD	CAD	CD	HT	RA	T1D	T2D
Disease Prevalence (γ)	0.5%	5.6%	0.1%	26.4%	1%	0.5%	3%
Reference(s)	[1]	[2]	[3]	[4]	[5]	[6]	[7]

Table S2. The marginal epistatic p-values for all SNPs as computed by LT-MAPIT for bipolar disease (BD) in the WTCCC data set. LT-MAPIT was fit while using three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ (GW) on page 1, a *cis*-based matrix $\mathbf{K}_{cis}$ (Cis) on page 2, and a *trans*-based matrix $\mathbf{K}_{trans}$ (Trans) on page 3, respectively. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S3. The marginal epistatic p-values for all SNPs as computed by LT-MAPIT for coronary artery disease (CAD) in the WTCCC data set. LT-MAPIT was fit while using three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ (GW) on page 1, a *cis*-based matrix $\mathbf{K}_{cis}$ (Cis) on page 2, and a *trans*-based matrix $\mathbf{K}_{trans}$ (Trans) on page 3, respectively. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S4. The marginal epistatic p-values for all SNPs as computed by LT-MAPIT for Crohn's disease (CD) in the WTCCC data set. LT-MAPIT was fit while using three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ (GW) on page 1, a *cis*-based matrix $\mathbf{K}_{cis}$ (Cis) on page 2, and a *trans*-based matrix $\mathbf{K}_{trans}$ (Trans) on page 3, respectively. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S5. The marginal epistatic p-values for all SNPs as computed by LT-MAPIT for hypertension (HT) in the WTCCC data set. LT-MAPIT was fit while using three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ (GW) on page 1, a *cis*-based matrix $\mathbf{K}_{cis}$ (Cis) on page 2, and a *trans*-based matrix $\mathbf{K}_{trans}$ (Trans) on page 3, respectively. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S6. The marginal epistatic p-values for all SNPs as computed by LT-MAPIT for rheumatoid arthritis (RA) in the WTCCC data set. LT-MAPIT was fit while using three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ (GW) on page 1, a *cis*-based matrix $\mathbf{K}_{cis}$ (Cis) on page 2, and a *trans*-based matrix $\mathbf{K}_{trans}$ (Trans) on page 3, respectively. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S7. The marginal epistatic p-values for all SNPs as computed by LT-MAPIT for type 1 diabetes (T1D) in the WTCCC data set. LT-MAPIT was fit while using three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ (GW) on page 1, a *cis*-based matrix $\mathbf{K}_{cis}$ (Cis) on page 2, and a *trans*-based matrix $\mathbf{K}_{trans}$ (Trans) on page 3, respectively. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S8. The marginal epistatic p-values for all SNPs as computed by LT-MAPIT for type 2 diabetes (T2D) in the WTCCC data set. LT-MAPIT was fit while using three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ (GW) on page 1, a *cis*-based matrix $\mathbf{K}_{cis}$ (Cis) on page 2, and a *trans*-based matrix $\mathbf{K}_{trans}$ (Trans) on page 3, respectively. Also listed are the chromosome location and physical position (bp) for each SNP. (XLSX)

Table S9. Summary table of the number of epistatic associations identified in the analysis of the WTCCC data set. The seven diseases analyzed include: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D). Compared are the number of SNPs with significant marginal epistatic associations, as well as the number of significant epistatic pairs, found by LT-MAPIT and LT-PLINK. LT-MAPIT was fit while using three different genetic relatedness matrices: the genome-wide matrix $\mathbf{K}_{GW}$ (GW), a *cis*-based matrix $\mathbf{K}_{cis}$ (Cis), and a *trans*-based matrix $\mathbf{K}_{trans}$ (Trans), respectively. We use the fully exhaustive search model LT-PLINK as a baseline comparison. Interactions were found with LT-MAPIT by first implementing an initial filter to find significant marginally epistatic SNPs, and then searching among these select variants to find significant pairs. Alternatively, marginal associations were found via LT-PLINK if a given SNP is a member of a significant epistatic pair.

Association Type	Analysis	BD	CAD	CD	HT	RA	T1D	T2D
Marginal Effects	GW	3	3	2	0	3	19	3
	Cis	0	0	0	2	60	397	0
	Trans	3	3	2	0	3	23	2
	LT-PLINK	3	3	0	3	49	366	5
Pairwise Interactions	GW	0	0	0	0	0	39	0
	Cis	0	0	0	0	443	10392	0
	Trans	0	0	0	0	0	44	0
	LT-PLINK	2	2	0	2	75	2753	3

Supplementary Text: MAPIT Binary Mixed Model Extension

We describe the logistic modeling approach for the MArginal ePIstasis Test (MAPIT-Logit) in detail here. Begin by considering the following generalized linear mixed model (GLMM)

$$y_i \sim \mathcal{B}(\pi_i), \quad i = 1, \ldots, n$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ is an n-vector of Bernoulli distributed (binary) traits for n individuals, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^\intercal$ is an unknown parameter that represents the underlying probability that the i-th individual is a case. Under this modeling approach, we use a logit link function $\psi(\bullet)$ and model the data as a linear combination of the following parameters

$$\psi(\boldsymbol{\pi}) = \log\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) = \boldsymbol{\eta}, \quad \boldsymbol{\eta} = \mu\mathbf{1} + \mathbf{x}_r\beta_r + \mathbf{m}_r + \mathbf{g}_r \tag{S1}$$

where (as in the main text) μ is an intercept term multiplied with an n-dimensional vector of ones $\mathbf{1}$; $\mathbf{x}_r$ is an n-dimensional genotype vector for the r-th variant that is the focus of the model; β_r is the corresponding additive effect size; $\mathbf{m}_r = \sum_{s\neq r}\mathbf{x}_s\beta_s$ is the combined additive effects from all other $s \neq r$ variants, and effectively represents the additive effect of the r-th variant under the polygenic background of all other variants; $\mathbf{m}_r \sim \mathcal{N}(\mathbf{0}, \omega^2\mathbf{K}_r)$ with $\mathbf{K}_r = \mathbf{X}_{-r}\mathbf{X}_{-r}^\intercal/(p-1)$ being the genetic relatedness matrix computed using genotypes from all variants other than the r-th; $\mathbf{g}_r = \sum_{s\neq r}(\mathbf{x}_r \circ \mathbf{x}_s)\theta_s$ is the summation of all pairwise interaction effects $\mathbf{x}_r \circ \mathbf{x}_s$ between the r-th variant and all other variants $s \neq r$; and $\mathbf{g}_r \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{G}_r)$ with $\mathbf{G}_r = \mathbf{D}_r\mathbf{K}_r\mathbf{D}_r$ representing a relatedness matrix computed based on pairwise interaction terms between the r-th variant and all other variants. Here, we denote $\mathbf{D}_r = \mathrm{diag}(\mathbf{x}_r)$ to be an $n \times n$ diagonal matrix with the genotype vector $\mathbf{x}_r$ as its diagonal elements. It is important to note that both $\mathbf{K}_r$ and $\mathbf{G}_r$ change with every new marker r that is considered.

PQL Inference Algorithm

We fit the above MAPIT-logistic model using the penalized quasi-likelihood (PQL) approach [8–10]. Briefly, PQL employs an iterative numerical optimization procedure where, in each iteration, we introduce a set of pseudo-data $\widetilde{\mathbf{y}}$ to replace the originally observed binary data $\mathbf{y}$. The generation of this pseudo-data $\widetilde{\mathbf{y}}$ is based on the second order Taylor expansion of the conditional distribution $\mathcal{P}(y_i \,|\, \mathbf{m}_r, \mathbf{g}_r)$ where we make use of the first and second order moments $\mathbb{E}(\mathbf{y} \,|\, \mathbf{m}_r, \mathbf{g}_r)$ and $\mathbb{V}(\mathbf{y} \,|\, \mathbf{m}_r, \mathbf{g}_r)$, respectively. With the realization of the pseudo-data $\widetilde{\mathbf{y}}$, the complex likelihood function for the original data $\mathbf{y}$ is replaced by a much simpler likelihood function, thereby alleviating much of the computational burden that comes with fitting GLMMs. As a result, we can perform marginal epistatic inference and update parameters using a standard average information (AI) algorithm for linear mixed models [11–13].

We now describe the detailed estimation and inference procedure for MAPIT-Logit. Here, the case-control labels $\mathbf{y}$ are assumed to be independently distributed conditional on the unobserved random effects $\mathbf{u}_r = \mathbf{m}_r + \mathbf{g}_r$ and fixed effects $\mu\mathbf{1} + \mathbf{x}_r\beta_r$. Each of these independent conditional distributions have the following mean and variance, respectively,

$$\mathbb{E}(\mathbf{y} \,|\, \mu, \beta_r, \mathbf{u}_r) = \boldsymbol{\vartheta}_r = \psi^{-1}(\mu\mathbf{1} + \mathbf{x}_r\beta_r + \mathbf{u}_r) \tag{S2}$$

$$\mathbb{V}(\mathbf{y} \,|\, \mu, \beta_r, \mathbf{u}_r) = v(\boldsymbol{\vartheta}_r), \tag{S3}$$

where, in addition to previous notation, $v(t) = nt(1-t)$ is the corresponding Bernoulli variance function. The joint quasi-likelihood function for the r-th variant is then given as

$$\mathcal{Q}(\mu, \beta_r, \omega^2, \sigma^2) = \log \int \left[\prod_{i=1}^{n} \mathcal{Q}(\mu, \beta_r, \mathbf{u}_r)\right] \mathcal{P}(\mathbf{u}_r \,|\, \omega^2, \sigma^2)\, \mathrm{d}\mathbf{u}_r,$$

which notedly also serves as an approximation to the joint conditional likelihood. Next, we may follow previous works [10] and use a Laplace approximation to further approximate the above joint function

$$\widetilde{\mathcal{Q}}(\mu, \beta_r, \omega^2, \sigma^2) = \frac{1}{2}\log|\mathbf{V}_r\mathbf{W}_r - \mathbf{I}| + \sum_{i=1}^{n} \mathcal{Q}(\mu, \beta_r, \widetilde{\mathbf{u}}_r) - \frac{1}{2}\widetilde{\mathbf{u}}_r^{\intercal}\mathbf{V}_r^{-1}\widetilde{\mathbf{u}}_r, \tag{S4}$$

where $\mathbf{V}_r = \omega^2\mathbf{K}_r + \sigma^2\mathbf{G}_r$, $\widetilde{\mathbf{u}}_r = \arg\max_{\mathbf{u}_r}(\sum_{i=1}^{n} \mathcal{Q}(\mu, \beta_r \,|\, \mathbf{u}_r) - \frac{1}{2}\mathbf{u}_r^{\intercal}\mathbf{V}_r^{-1}\mathbf{u}_r$, $\mathbf{W}_r = \text{diag}(1/\psi'(\boldsymbol{\vartheta}_r))$ is a diagonal weight matrix with corresponding elements $\boldsymbol{\vartheta}_r$, and $|\bullet|$ is used to denote the matrix determinant function.

We treat the approximated quasi-likelihood function $\widetilde{\mathcal{Q}}$ in Equation (S4) as the target function from which we obtain estimates for $(\mu, \beta_r, \mathbf{u}_r)$ and (ω^2, σ^2) via an iterative procedure. To begin, we first obtain estimates for $(\mu, \beta_r, \mathbf{u}_r)$ conditional on the current estimates of (ω^2, σ^2). Here, we assume that the iterative weights vary slowly with respect to the conditional mean. Namely,

$$\frac{\partial \mathbf{W}_r}{\partial \boldsymbol{\vartheta}_r} \approx \mathbf{0}.$$

We then obtain the first order derivatives with respect to either the fixed effects (μ, β_r) or the random effects $\mathbf{u}_r$, and set the corresponding two first order derivatives to zero

$$\widetilde{\mathbf{x}}_r^{\intercal}\mathbf{W}_r\boldsymbol{\Delta}_r(\mathbf{y} - \boldsymbol{\vartheta}_r) = \mathbf{0}, \tag{S5}$$

$$\boldsymbol{\Delta}_r(\mathbf{y} - \boldsymbol{\vartheta}_r) - \mathbf{V}_r^{-1}\mathbf{u}_r = \mathbf{0}, \tag{S6}$$

where $\widetilde{\mathbf{x}}_r = [\mathbf{1}, \mathbf{x}_r]$ and $\boldsymbol{\Delta} = \text{diag}(\psi'(\boldsymbol{\vartheta}_r))$. We may now formally define the pseudo-data as

$$\widetilde{\mathbf{y}}_r = \boldsymbol{\eta}_r + \psi'(\boldsymbol{\vartheta}_r)(\mathbf{y} - \boldsymbol{\vartheta}_r). \tag{S7}$$

With this formulation of pseudo-data, Equation (S5) simplifies to

$$\mathbf{u}_r = \mathbf{V}_r\mathbf{H}_r^{-1}\left(\widetilde{\mathbf{y}}_r - \widetilde{\mathbf{x}}_r\mathbf{b}_r\right), \tag{S8}$$

where $\mathbf{H}_r = \mathbf{W}_r^{-1} + \mathbf{V}_r$ and $\mathbf{b}_r = (\mu, \beta_r)^{\intercal}$. By substituting Equation (S8) into Equation (S6), we can obtain the following parameter estimates

$$\widehat{\mathbf{b}}_r = (\widetilde{\mathbf{x}}_r^{\intercal}\mathbf{H}_r^{-1}\widetilde{\mathbf{x}}_r)^{-1}\widetilde{\mathbf{x}}_r^{\intercal}\mathbf{H}_r^{-1}\widetilde{\mathbf{y}}_r \tag{S9}$$

$$\widehat{\mathbf{u}}_r = \mathbf{V}_r\mathbf{H}_r^{-1}\left(\widetilde{\mathbf{y}}_r - \widetilde{\mathbf{x}}_r\widehat{\mathbf{b}}_r\right) \tag{S10}$$

Note that both of these equations depend on estimates of the all variance components — we will use $\boldsymbol{\delta} = (\omega^2, \sigma^2)$ to represent their joint shorthand notation. As a result, we want to obtain the estimates for these parameters, conditional on the current values of $(\mu, \beta_r, \mathbf{u}_r)$. To do so, we first integrate out the fix effects from Equation (S4) to obtain the following restricted likelihood

$$\widetilde{\mathcal{Q}}_R(\boldsymbol{\delta}) = c_r + \frac{1}{2}\log|\mathbf{H}_r| - \frac{1}{2}\log|\widetilde{\mathbf{x}}_r^{\intercal}\mathbf{H}_r^{-1}\widetilde{\mathbf{x}}_r| - \frac{1}{2}\widetilde{\mathbf{y}}_r^{\intercal}\mathbf{P}_r\widetilde{\mathbf{y}}_r. \tag{S11}$$

where $\mathbf{P}_r = \mathbf{H}_r^{-1} - \mathbf{H}_r^{-1}\widetilde{\mathbf{x}}_r^{\intercal}(\widetilde{\mathbf{x}}_r^{\intercal}\mathbf{H}_r^{-1}\widetilde{\mathbf{x}}_r)^{-1}\widetilde{\mathbf{x}}_r\mathbf{H}_r^{-1}$ is a variant specific projection matrix, and $c_r = [(c + 1)\log(2\pi) - \log|\mathbf{W}_r|]/2$ is a variant specific constant term. The following AI algorithm is used to obtain the desired variance component estimates. In particular, we obtain the first derivatives as

$$\frac{\partial \widetilde{\mathcal{Q}}_R(\boldsymbol{\delta})}{\partial \omega^2} = \frac{1}{2}[\widetilde{\mathbf{y}}_r^{\intercal}\mathbf{P}_r\mathbf{K}_r\mathbf{P}_r\widetilde{\mathbf{y}}_r - \text{tr}(\mathbf{P}_r\mathbf{K}_r)],$$

$$\frac{\partial \widetilde{\mathcal{Q}}_R(\boldsymbol{\delta})}{\partial \sigma^2} = \frac{1}{2}[\widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r - \text{tr}(\mathbf{P}_r \mathbf{G}_r)],$$

where tr(•) denotes the matrix function. The second derivatives are then computed as the following. For the additive variance component of the polygenic background ω^2,

$$\frac{\partial^2 \widetilde{\mathcal{Q}}_R(\boldsymbol{\delta})}{\partial \omega^2 \partial \omega^2} = \frac{1}{2}\text{tr}(\mathbf{K}_r \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r) - \widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r,$$
$$\frac{\partial^2 \widetilde{\mathcal{Q}}_R(\boldsymbol{\delta})}{\partial \omega^2 \partial \sigma^2} = \frac{1}{2}\text{tr}(\mathbf{K}_r \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r) - \widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r.$$

And for the marginal epistatic variance component σ^2 (i.e. the main parameter of interest),

$$\frac{\partial^2 \widetilde{\mathcal{Q}}_R(\boldsymbol{\delta})}{\partial \sigma^2 \partial \omega^2} = \frac{1}{2}\text{tr}(\mathbf{G}_r \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r) - \widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r,$$
$$\frac{\partial^2 \widetilde{\mathcal{Q}}_R(\boldsymbol{\delta})}{\partial \sigma^2 \partial \sigma^2} = \frac{1}{2}\text{tr}(\mathbf{G}_r \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r) - \widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r.$$

Note that, by definition, the second derivatives constitute the observed information matrix. Because the elements in the expected information matrix only include the trace terms of the above four equations, we can obtain the following variant specific average information (AI) matrix

$$\mathbf{A}_r = \begin{bmatrix} \widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r & \widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r \\ \widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \mathbf{K}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r & \widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r \end{bmatrix} \tag{S12}$$

With the first order derivatives and the AI matrix, we can perform a Newton-Raphson update with the AI algorithm to jointly obtain estimates $\widehat{\boldsymbol{\delta}} = (\widehat{\omega}^2, \widehat{\sigma}^2)$. In particular, this logistic version of MAPIT implements a PQL algorithm that consists of steps detailed in previous works [8–10]. Namely, we first initialize parameters $\mu^{(0)}$, $\beta_r^{(0)}$, $\boldsymbol{\delta}^{(0)} = (\omega^{2(0)}, \sigma^{2(0)})$. Next, we obtain the pseudo-data $\widetilde{\mathbf{y}}_r$ as detailed in Equation (S7). Finally, we set $t = 1$ and iterate between the following:

1. Update $\boldsymbol{\delta}^{(t)} = \boldsymbol{\delta}^{(t-1)} + \mathbf{A}_r^{-1}\left(\partial \widetilde{\mathcal{Q}}_R(\boldsymbol{\delta})/\partial \boldsymbol{\delta}\right)$.
2. Update $\left\{\mu^{(t)}, \beta_r^{(t)}, \mathbf{u}_r^{(t)}\right\}$ using $\boldsymbol{\delta}^{(t)}$ and $\widetilde{\mathbf{y}}_r^{(t-1)}$ as in Equations (S9) and (S10).
3. Update $\widetilde{\mathbf{y}}_r^{(t)}$ using $\left\{\mu^{(t)}, \beta_r^{(t)}, \mathbf{u}_r^{(t)}\right\}$ as in Equation (S7).

At the end of each iterate, reset $t = t + 1$, and repeat these steps until convergence. In practice we determine convergence to be the instance where the difference between two consecutive parameter estimates is smaller than $\epsilon = 1 \times 10^{-5}$.

Hypothesis Testing Procedure

Recall from the main text that our goal is to identify variants that interact with other variants, and to avoid explicitly searching for pairwise interactions. To do so, we want to test the appropriate null hypothesis $\text{H}_0 : \sigma_r^2 = 0$ for each variant r in the data. The following detailed hypothesis testing procedure stems from the fact that, under this null, the marginal epistatic variance component estimates follow a mixture of chi-square distributions $\widehat{\sigma}_r^2 \sim \sum_{j=1}^q \lambda_j \chi_{1,j}^2$. Here, $\chi_{1,j}^2$ denote chi-square random variables with one degree of freedom, and $(\lambda_1, \ldots, \lambda_q)$ represent the ordered (and quickly decaying) non-zero eigenvalues of the corresponding matrix $\mathbf{P}_r \mathbf{G}_r/2$. Unfortunately, the calculation of these eigenvalues is

computationally intensive and exact probabilities associated with a mixture of chi-square distributions is difficult to compute analytically [14]. As a result, we make use of the Satterthwaite method to approximate the mixture of chi-square distributions with a scaled chi-square distribution [15]. Namely,

$$\sum_{i=1}^{q} \lambda_j \chi^2_{1,j} \approx \kappa \chi^2_\nu$$

where the scale parameter κ and the degrees of freedom ν are empirically computed from

$$\kappa = I_r / e_r, \quad \nu = e_r^2 / 2 I_r,$$

where $e_r = \mathrm{tr}(\mathbf{P}_r \mathbf{G}_r)$ and $I_r = \mathrm{tr}(\mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \mathbf{G}_r)/2 - \mathrm{tr}(\mathbf{P}_r \mathbf{G}_r \mathbf{P}_r)^2 / \mathrm{tr}(\mathbf{P}_r \mathbf{P}_r)$. Altogether, the final bias-corrected test statistic for the Satterthwaite method is given by

$$\mathrm{S}(\widehat{\sigma}^2) = \frac{1}{2\kappa} (\widetilde{\mathbf{y}}_r^\intercal \mathbf{P}_r \mathbf{G}_r \mathbf{P}_r \widetilde{\mathbf{y}}_r), \tag{S13}$$

which, under the null hypothesis, follows a distribution that may be approximated by χ^2_ν.

Software Implementation

For the simulations detailed in the main text, we fit the MAPIT-Logit regression model using PQLseq [8–10], which is an R package for implementing mixed model association for count data via penalized quasi-likelihoods [16]. PQLseq is publicly available at `http://www.xzlab.org/software.html`.

References

1. Lichtenstein P, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: A population-based study. Lancet. 2009;373(9659):234–239. Available from: http://www.sciencedirect.com/science/article/pii/S0140673609600726.

2. Marenberg ME, Risch N, Berkman LF, Floderus B, de Faire U. Genetic susceptibility to death from coronary heart disease in a study of twins. N Engl J Med. 1994;330(15):1041–1046. Available from: https://doi.org/10.1056/NEJM199404143301503.

3. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 2015;11(4):e1004969. Available from: https://doi.org/10.1371/journal.pgen.1004969.

4. Kearney PM, Whelton M, Reynolds K, Muntner P, Whelton PK, He J. Global burden of hypertension: Analysis of worldwide data. Lancet. 2005;365(9455):217–223. Available from: http://www.sciencedirect.com/science/article/pii/S0140673605177411.

5. Symmons D, Turner G, Webb R, Asten P, Barrett E, Lunt M, et al. The prevalence of rheumatoid arthritis in the United Kingdom: New estimates for a new century. Rheumatology. 2002;41(7):793–800. Available from: http://dx.doi.org/10.1093/rheumatology/41.7.793.

6. Hyttinen V, Kaprio J, Kinnunen L, Koskenvuo M, Tuomilehto J. Genetic liability of type 1 diabetes and the onset age among 22,650 young Finnish twin pairs. Diabetes. 2003;52(4):1052–1055. Available from: http://diabetes.diabetesjournals.org/content/52/4/1052.abstract.

7. Das SK, Elbein SC. The genetic basis of type 2 diabetes. Cellscience. 2006;2(4):100–131. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1526773/.

8. Lea AJ, Tung J, Zhou X. A Flexible, Efficient Binomial Mixed Model for Identifying Differential DNA Methylation in Bisulfite Sequencing Data. PLoS Genet. 2015;11(11):e1005650. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005650.

9. Sun S, Hood M, Scott L, Peng Q, Mukherjee S, Tung J, et al. Differential expression analysis for RNAseq using Poisson mixed models. Nucleic Acids Res. 2017;45(11):e106. Available from: https://academic.oup.com/nar/article/45/11/e106/3093260.

10. Sun S, Zhu J, Mozaffari S, Ober C, Chen M, Zhou X. Heritability Estimation and Differential Analysis with Generalized Linear Mixed Models in Genomic Sequencing Studies. bioRxiv. 2018;Available from: http://biorxiv.org/content/early/2018/06/29/359265.abstract.

11. Gilmour AR, Thompson R, Cullis BR. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics. 1995;p. 1440–1450.

12. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.

13. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am J Hum Genet. 2016;98(4):653–666.

14. Zhang D, Lin X. Hypothesis testing in semiparametric additive mixed models. Biostatistics. 2003;4(1):57–74.

15. Satterthwaite FE. An approximate distribution of estimates of variance components. Biometrics. 1946;2(6):110–114. Available from: http://www.jstor.org/stable/3002019.

16. Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association. 1993;88(421):9–25. Available from: http://www.jstor.org/stable/2290687.