Genome-wide Marginal Epistatic Association Mapping in Case-Control Studies

Lorin Crawford[1-3,†] and Xiang Zhou[4,5,†]

1 Department of Biostatistics, Brown University, Providence, RI, USA
2 Center for Statistical Sciences, Brown University, Providence, RI, USA
3 Center for Computational Molecular Biology, Brown University, Providence, RI, USA
4 Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
5 Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA

† **E-mail: lorin_crawford@brown.edu; xzhousph@umich.edu**

Abstract

Epistasis, commonly defined as the interaction between genetic loci, is an important contributor to the genetic architecture underlying many complex traits and common diseases. Most existing epistatic mapping methods in genome-wide association studies explicitly search over all pairwise or higher-order interactions. However, due to the potentially large search space and the resulting multiple testing burden, these conventional approaches often suffer from heavy computational cost and low statistical power. A recently proposed attractive alternative for mapping epistasis focuses instead on detecting marginal epistasis, which is defined as the combined pairwise interaction effects between a given variant and all other variants. By searching for marginal epistatic effects, one can identify genetic variants that are involved in epistasis without the need to identify the exact partners with which the variants interact — thus, potentially alleviating much of the statistical and computational burden associated with conventional epistatic mapping procedures. However, previous marginal epistatic mapping methods are based on quantitative trait models. As we will show here, these lack statistical power in case-control studies. Here, we develop a liability threshold mixed model that extends marginal epistatic mapping to case- control studies. Our method properly accounts for case-control ascertainment and the binary nature of case-control data. We refer to this method as the liability threshold marginal epistasis test (LT-MAPIT). With simulations, we illustrate the benefits of LT-MAPIT in terms of providing effective type I error control, and being more powerful than both existing marginal epistatic mapping methods and conventional explicit search-based approaches in case-control data. We finally apply LT-MAPIT to identify both marginal and pairwise epistasis in seven complex diseases from the Wellcome Trust Case Control Consortium (WTCCC) 1 study.

Introduction

Epistasis, commonly defined as the interaction between genetic loci, has long been thought to play a key role in defining the genetic architecture underlying many complex traits and common diseases [1, 2]. Indeed, previous studies have detected pervasive epistasis in many model organisms [3–26]. Substantial contributions of epistasis to phenotypic variance have been revealed for many complex traits [27, 28] and have been suggested to constitute the genetic basis of evolution [29, 30]. Furthermore, modeling epistasis in addition to additive/dominant effects has been shown to increase phenotypic prediction accuracy in model organisms [31] and facilitate genomic selection in animal breeding programs [32, 33]. Despite some controversies [34], recent genetic mapping studies have also identified candidates of epistatic interactions that significantly contribute to quantitative traits and diseases [35–39]. Importantly, epistasis has been proposed as a key contributor to missing heritability — the proportion of heritability not explained by the top associated variants in GWASs [40–43].

The recent development of genome-wide association studies (GWASs) [42] has provided an unique opportunity to search across the whole genome to detect genetic variants that are involved in epistasis. Many statistical works have been developed to facilitate the identification of epistasis in GWASs [44, 45]. Generally, these existing tools can be classified into two categories. The first category of methods is conventional and explicitly searches for pairwise or higher-order interactions when identifying epistatic effects. These approaches rely on distinct criteria for selecting a testing unit based on variants or genes [46]. In particular, they use different searching strategies including exhaustive search [47–49], probabilistic search [50], or prioritization based on a preselected candidate set [51]. Specific statistical paradigms have been implemented for explicit search-based approaches including various frequentist tests [52] and Bayesian inferences [53–55]. Unfortunately, because of the extremely large search space (e.g. $p(p-1)/2$ possible pairwise combinations for p-variants), these methods often suffer from heavy computational burden. Computationally, despite various efficient computational improvements [48, 56] and recently developed search algorithms [50], exploring over a large combinatorial domain remains a daunting task for large epistatic mapping studies. Statistically, because of a lack of *a priori* knowledge about epistatic loci, exploring all combinations of genetic variants could result in low statistical power — on the other hand, restricting to a subset of prioritized combinations based on prior knowledge or marginal additive effects could also miss important genetic interactions.

The second category of epistatic mapping methods, exemplified by the recently developed MArginal ePIstasis Test (MAPIT) [39], attempts to address the previously mentioned challenges by alternatively detecting *marginal* epistasis. Specifically, instead of directly identifying individual pairwise or higher-order interactions, these approaches focus on identifying variants that have a non-zero interaction effect with any other variants. For example, MAPIT tests each variant (in turn) for its *marginal epistatic effect* — the combined pairwise interaction effects between a given variant and all other variants. By testing marginal epistatic effects, one can identify candidate markers that are involved in epistasis without the need to identify the exact partners with which the variants interact — thus, potentially alleviating much of the statistical and computational burden associated with the first category of epistatic mapping methods. In addition, evidence of significant marginal epistasis detected by MAPIT can be used to further prioritize the search for meaningful pairwise interactions. Despite its potential advantages, however, MAPIT has previously been restricted to only quantitative traits. As we will show below, a naïve application of MAPIT to case-control studies yields surprisingly low power. Thus, an overall restriction to quantitative traits limits the potential impact of marginal epistatic mapping methods.

Here, we develop a new statistical method that allows us to detect marginal epistasis in case-control studies. Our method relies on a liability threshold mixed model [57–59] to properly account for case-control ascertainment and the binary nature of case-control data. By combining efficient variance component estimation and hypothesis testing procedures [39,60] together with reasonable model approximations, our proposed approach is scalable to moderately sized GWASs. We refer to this method as the liability threshold marginal epistasis test (LT-MAPIT). With simulations, we illustrate the benefits of LT-MAPIT in terms of providing effective type I error control, as well as being more powerful than both existing marginal epistatic models and conventional search-based approaches in case-control data. Lastly, we use LT-MAPIT to identify both marginal and pairwise epistasis in seven complex diseases from the Wellcome Trust Case Control Consortium (WTCCC) 1 study [61].

Material and Methods

Modeling Marginal Epistasis for Case-Control Data

We begin by briefly reviewing the notation and intuition behind the MArginal ePIstasis Test (MAPIT) for case-control studies. Our goal is two-fold. First, we aim to identify variants that interact with other variants without having to explicitly search over the entire model space of all possible interactions.

Second, we aim to properly account for case-control ascertainment and the binary nature of case-control data. To achieve these objectives, we consider a liability threshold mixed effects model. Here, we assume that the case-control status y_i for the i-th individual is determined by the their corresponding underlying liability value l_i. Namely,

$$y_i = \begin{cases} 1 & \text{if} \quad l_i \geq 0 \\ 0 & \text{if} \quad l_i < 0 \end{cases} \quad \text{for} \quad i = 1, \ldots, n \tag{1}$$

where $y_i = 1$ if the i-th individual is a case, and $y_i = 0$ when the i-th individual is a control. It is important to note that while the disease status $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ is binary, the underlying liability $\mathbf{l} = (l_1, \ldots, l_n)^\intercal$ is continuous. The liability threshold model allows us to adapt classical statistical methods, that were initially developed for quantitative phenotypes, and use them to analyze the unobserved continuous trait of liability [57–59]. In our case, we examine one variant at a time (indexed by r), and consider the following regression

$$\mathbf{l} = \mu\mathbf{1} + \mathbf{x}_r\beta_r + \mathbf{m}_r + \mathbf{g}_r + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{I}). \tag{2}$$

where 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 with effect sizes β_s, and effectively represents the polygenic background of all variants except for 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 $s \neq r$ with regression coefficient θ_s; $\boldsymbol{\varepsilon}$ is an n-vector of residual errors; τ^2 is the residual error variance; $\mathbf{I}$ denotes the identity matrix; and $\mathcal{N}(\bullet, \bullet)$ denotes a multivariate normal distribution. The term $\mathbf{g}_r$ is the main focus of the model and represents the marginal epistatic effect of r-th variant — formally defined as the summation of its epistatic interaction effects with all other variants [39]. Here, each genotypic vector is assumed to have been centered and scaled to have mean 0 and standard deviation 1, respectively. Also note that while we limit ourselves to the task of identifying second order (i.e. pairwise) epistatic relationships between genetic variants (i.e. $\mathbf{x}_r \circ \mathbf{x}_s$), extensions to higher-order epistatic effects are straightforward to implement [39, 60].

There are a few assumptions that need to be made in order to ensure model identifiability in Equation (2). First, because we consider application scenarios where $p > n$, we recall our previous approach and assume that individual effect sizes follow univariate normal distributions, or $\beta_s \sim \mathcal{N}(0, \omega^2/(p-1))$ and $\theta_s \sim \mathcal{N}(0, \sigma^2/(p-1))$ for $s \neq r$ [39]. With the assumption of normally distributed effect sizes, the model defined in Equation (2) is equivalent to a variance component model where $\mathbf{m}_r \sim \mathcal{N}(\mathbf{0}, \omega^2\mathbf{K}_r)$ 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; 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, $\mathbf{D}_r = \text{diag}(\mathbf{x}_r)$ denotes an $n \times n$ diagonal matrix with the genotype vector of interest 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. For the second assumption, the mean and variance of the liability $\mathbf{l}$ determine a known disease prevalence γ in the population — however, these two parameters are not identifiable from each other given only γ. As a result, we follow previous approaches by constraining the variance of $\mathbf{l}$ to be one (i.e. $\omega^2 + \sigma^2 + \tau^2 = 1$) [57].

We will formally refer to the liability threshold mixed model defined in Equations (1) and (2) as LT-MAPIT. This approach properly handles the binary nature of case-control data through the direct modeling of $\mathbf{y}$. Furthermore, it also appropriately accounts for case-control ascertainment through the direct modeling of liability $\mathbf{l}$. The difference between the proportion of cases observed in the population versus those sampled in the case-control study, is accounted for by the distributions of the latent liability values in the population versus in the study. Furthermore, the variance component σ^2 represents a

measure on the marginal epistatic effect for the r-th variant. Therefore, testing the null hypothesis $H_0 : \sigma^2 = 0$ will allow us to examine whether the r-th SNP interacts with any other SNPs to exhibit a significant pairwise epistatic interaction.

Liability Threshold Mixed Model Approximations

In order to estimate and test σ^2 in LT-MAPIT, we first write down the full likelihood of the liability threshold mixed model as a joint function of the variance components $(\omega^2, \sigma^2, \tau^2)$

$$\mathcal{L}(\omega^2, \sigma^2, \tau^2 \,|\, \mathbf{y}) = \int_{\mathbf{l}} \mathcal{P}(\mathbf{y} \,|\, \mathbf{l}) \mathcal{P}(\mathbf{l} \,|\, \omega^2, \sigma^2, \tau^2) \, \mathrm{d}\mathbf{l}, \tag{3}$$

where $\mathcal{P}(\mathbf{y} \,|\, \mathbf{l})$ is the conditional probability of the observed case-control status $\mathbf{y}$ given the latent liability $\mathbf{l}$ as defined in Equation (1), and $\mathcal{P}(\mathbf{l} \,|\, \omega^2, \sigma^2, \tau^2)$ denotes the probability of the latent liability given all other parameters as defined in Equation (2). The likelihood function in Equation (3), unfortunately, involves an n-dimensional integration over $\mathbf{l}$ that cannot be solved analytically. Consequently, we have to rely on approximation methods for inference. Here, we follow previous works [62–65] and approximate the likelihood function with a first order Taylor expansion equated at an estimate of the posterior mean for the liabilities, say $\widehat{\mathbf{l}} = \mathbb{E}[\mathbf{l} \,|\, \mathbf{y}]$. Note that the posterior distribution of the liabilities is given as $\mathcal{P}(\mathbf{l} \,|\, \mathbf{y}) \propto \mathcal{P}(\mathbf{y} \,|\, \mathbf{l}) P(\mathbf{l} \,|\, \omega^2, \sigma^2, \tau^2)$, which is proportional to the integrand in Equation (3). The resulting approximate likelihood is then written as

$$\mathcal{L}(\omega^2, \sigma^2, \tau^2 \,|\, \mathbf{y}) \approx \mathcal{P}(\mathbf{y} \,|\, \widehat{\mathbf{l}}) \mathcal{P}(\widehat{\mathbf{l}} \,|\, \omega^2, \sigma^2, \tau^2), \tag{4}$$

where the first term $\mathcal{P}(\mathbf{y} \,|\, \widehat{\mathbf{l}})$ can be ignored as the posterior mean $\widehat{\mathbf{l}}$ satisfies the constraint set by $\mathbf{y}$. The above approximated likelihood may be interpreted as the likelihood of a linear mixed model with the posterior mean estimate $\widehat{\mathbf{l}}$ as the response variable. Therefore, the liability threshold mixed model defined in Equations (1) and (2) can be effectively approximated by a linear mixed effects model on $\widehat{\mathbf{l}}$, where

$$\widehat{\mathbf{l}} = \mu \mathbf{1} + \mathbf{x}_r \beta_r + \mathbf{m}_r + \mathbf{g}_r + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{I}). \tag{5}$$

In practice, the posterior mean estimates of the liabilities can be obtained by using a Gibbs sampler, which is an iterative algorithm that generates posterior samples from conditional distributions [66–68]. Thus, given the current parameter estimates of $(\omega^2, \sigma^2, \tau^2)$, new values of $\widehat{\mathbf{l}}$ are generated by taking draws from the posterior distribution $\mathcal{P}(\mathbf{l} \,|\, \mathbf{y})$. These generated values of $\widehat{\mathbf{l}}$ can then be used further to refine the estimates of $(\omega^2, \sigma^2, \tau^2)$ through optimizing the approximated likelihood defined in Equation (5). Hence, parameter estimates of $(\omega^2, \sigma^2, \tau^2)$ can also be obtained through an iterative optimization procedure. As previously mentioned [63], optimizing a multivariate truncated normal distribution $\mathcal{P}(\mathbf{l} \,|\, \mathbf{y})$ to obtain $\widehat{\mathbf{l}}$ is, unfortunately, not straightforward. As a result, instead of a formal Gibbs sampler, we apply a second approximation based on the properties of GWAS data to obtain $\widehat{\mathbf{l}}$ [65]. Specifically, we assume that individuals are essentially unrelated in the sense that $\mathbf{K} \approx \mathbf{I}$. We also assume that both the additive and the marginal epistatic effects are small (which is typical for most complex traits [60]), such that that we can ignore the terms $\mathbf{x}_r \beta_r$, $\mathbf{m}_r$, and $\mathbf{g}_r$. With these additional assumptions, the multivariate normal distribution $\mathcal{P}(\mathbf{l} \,|\, \omega^2, \sigma^2, \tau^2)$ defined in Equation (2) is approximated by a joint product of independent normal distributions $l_i \sim \mathcal{N}(\mu, 1)$. We then apply Markov Chain Monte Carlo (MCMC) to obtain the posterior estimates of $\mathbf{l}$. Specifically, in each iteration t, posterior samples of $\mathbf{l}$ are generated from constrained conditional univariate normal distributions

$$l_i^{(t)} \,|\, \mathbf{y} \sim \begin{cases} \mathcal{N}(\mu, 1) & \text{with } l_i^{(t)} \geq 0 \quad \text{if} \quad y_i = 1 \\ \mathcal{N}(\mu, 1) & \text{with } l_i^{(t)} < 0 \quad \text{if} \quad y_i = 0 \end{cases} \quad \text{for} \quad i = 1, \ldots, n, \tag{6}$$

from which $\widehat{\mathbf{l}}$ is analytically computed by averaging over samples drawn during the MCMC iterations. The parameter $\mu = \Phi^{-1}(\gamma)$ represents the truncation (liability) threshold determined by the known disease prevalence γ in the population, with $\Phi^{-1}(\bullet)$ being the inverse cumulative probability density of the standard normal distribution. For all examples in the text, we follow previous approaches [59] and take $\widehat{\mathbf{l}}$ to be a single instance drawn from the posterior in Equation (6).

Parameter Estimation and Variance Component Test

With the approximations described in the previous section, we are now at a stage to perform joint estimation of the all variance component parameters in the LT-MAPIT model. To do so, we make use of the MQS approach for parameter estimation and hypothesis testing [60]. Briefly, MQS is based on the computationally efficient method of moments and produces estimates that are mathematically identical to the Haseman-Elston (HE) cross-product regression [69]. To estimate the variance components with MQS, we first multiply Equation (5) by a variant specific projection (hat) matrix $\mathbf{H}_r$ onto both the null space of the intercept and the corresponding genotypic vector $\mathbf{x}_r$. Namely, $\mathbf{H}_r = \mathbf{I} - \mathbf{b}_r(\mathbf{b}_r^\intercal \mathbf{b}_r)^{-1}\mathbf{b}_r^\intercal$ with $\mathbf{b}_r = [\mathbf{1}, \mathbf{x}_r]$. After this projection procedure, we obtain a simplified liability threshold mixed model

$$\mathbf{l}_r^* = \mathbf{m}_r^* + \mathbf{g}_r^* + \boldsymbol{\varepsilon}_r^*, \quad \mathbf{m}_r^* \sim \mathcal{N}(\mathbf{0}, \omega^2 \mathbf{K}_r^*), \quad \mathbf{g}_r^* \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{G}_r^*), \quad \boldsymbol{\varepsilon}_r^* \sim \mathcal{N}(\mathbf{0}, \tau^2 \mathbf{H}_r), \tag{7}$$

where $\mathbf{l}_r^* = \mathbf{H}_r\widehat{\mathbf{l}}$; $\mathbf{m}_r^* = \mathbf{H}_r\mathbf{m}$; $\mathbf{K}_r^* = \mathbf{H}_r\mathbf{K}_r\mathbf{H}_r$; $\mathbf{g}_r^* = \mathbf{H}_r\mathbf{g}_r$; $\mathbf{G}_r^* = \mathbf{H}_r\mathbf{G}_r\mathbf{H}_r$; and $\boldsymbol{\varepsilon}_r^* = \mathbf{H}_r\boldsymbol{\varepsilon}$, respectively. Then for each variant considered, the MQS estimate for the marginal epistatic effect is computed as

$$\widehat{\sigma}^2 = \mathbf{l}_r^{*\intercal} \mathbf{A}_r \mathbf{l}_r^* \tag{8}$$

where $\mathbf{A}_r = (\mathbf{S}_r^{-1})_{21}\mathbf{K}_r^* + (\mathbf{S}_r^{-1})_{22}\mathbf{G}_r^* + (\mathbf{S}_r^{-1})_{23}\mathbf{H}_r$ with elements $(\mathbf{S}_r)_{jk} = \mathrm{tr}(\boldsymbol{\Sigma}_{rj}\boldsymbol{\Sigma}_{rk})$ for the covariance matrices subscripted as $[\boldsymbol{\Sigma}_{r1}; \boldsymbol{\Sigma}_{r2}; \boldsymbol{\Sigma}_{r3}] = [\mathbf{K}_r^*; \mathbf{G}_r^*; \mathbf{H}_r]$. Here, $\mathrm{tr}(\bullet)$ is used to denote the matrix trace function. It has been shown that a given marginal epistatic variance component estimate $\widehat{\sigma}^2$ follows a mixture of chi-square distributions under the null hypothesis [39]. Namely, $\widehat{\sigma}^2 \sim \sum_{i=1}^n \lambda_i \chi_{1,i}^2$, where χ_1^2 are chi-square random variables with one degree of freedom and $(\lambda_1, \ldots, \lambda_n)$ are the eigenvalues of the matrix

$$\left(\widehat{\omega}_0^2 \mathbf{K}_r^* + \widehat{\tau}_0^2 \mathbf{H}_r\right)^{1/2} \mathbf{A}_r \left(\widehat{\omega}_0^2 \mathbf{K}_r^* + \widehat{\tau}_0^2 \mathbf{H}_r\right)^{1/2}$$

with $(\widehat{\omega}_0^2, \widehat{\tau}_0^2)$ being the MQS estimates of (ω^2, τ^2) under the null hypothesis. Several approximate and exact methods have been suggested to obtain p-values under the distribution of $\widehat{\sigma}^2$. One frequented choice is Davies exact method [70, 71], which we will use here.

Because of the approximations we use when specifying LT-MAPIT (in the presence of case-control ascertainment), the MQS based testing procedure can result in conservative p-values — even when the model fit is conducted on the liability scale [72]. To circumvent the conservative p-value issue, we propose a recalibration approach following the main idea of genomic control [73] by using the following readjusted test statistics. Specifically, we denote $\widetilde{\mathbf{p}} = (\widetilde{p}_1, \ldots, \widetilde{p}_p)$ as a vector of p-values computed by using Davies exact method. We consider the genomic control transformation

$$\mathbf{p} = 1 - P_1(Q_1(\widetilde{\mathbf{p}})/\lambda_{GC}), \tag{9}$$

where $Q_1(\bullet)$ denotes the quantile function of the standard chi-square distribution with one degree of freedom; $P_1(\bullet)$ denotes the cumulative density function of the standard chi-square distribution with one degree of freedom; and $\lambda_{GC} = \widehat{m}/\widetilde{m}$ is a genomic control factor with $\widehat{m}$ being the median of $Q_1(\widetilde{\mathbf{p}})$ and $\widetilde{m} \approx 0.455$ being the median of the standard chi-square distribution [74–76]. With this readjustment, the ranking of variants that are most likely to be involved in epistatic interactions remains the same. However, the readjusted p-values now become well-calibrated, which we demonstrate later via simulations. All MAPIT-derived results that we present throughout the rest of the paper will be based on the readjusted marginal epistatic statistics.

Software Availability

Code for implementing LT-MAPIT is freely available within the "MArginal ePIstasis Test" (MAPIT) software located at `https://github.com/lorinanthony/MAPIT`. All MAPIT functions use the CompQuadForm R package to compute p-values with the Davies method. Note that the Davies method can sometimes yield a p-value that exactly equals 0. This can occur when the true p-value is extremely small [77]. In this case, we report p-values as approximately 0. If this is of concern, practitioners can compute p-values for all MAPIT based functions using Kuonen's saddlepoint method [77,78] or Satterthwaite's approximation equation [79].

Competing Methods

In the simulations and real data applications presented in this paper, we compare the LT-MAPIT model to five different epistatic mapping methods. The first is the original MAPIT method [39], which treats the case-control labels as a quantitative trait by fitting a linear mixed model directly onto the binary response variable (following the argument that the first order Taylor series expansion is an approximation to a generalized linear model [80–82]). The second method we consider reframes the MAPIT framework within the context of a generalized linear mixed model (GLMM) with a canonical logit link function (MAPIT-Logit). Briefly, the MAPIT-Logit takes an alternative route for GLMM inference and identifies marginal epistatic effects using a penalized quasi-likelihood (PQL) approach (specific details in Supporting Information) [83–85]. The last three methods we use for model comparisons are variations of the PLINK epistatic mapping software (version 1.9) [86]. Here, we use the `--epistasis` argument to fit an exhaustive search (generalized) linear model, which iteratively tests all possible pairwise interactions to identify the exact pairs of variants involved in significant epistasis [48, 49]. Specifically, we consider the following specifications for PLINK: (a) the standard linear regression model where the binary class labels are again treated as quantitative traits, (b) the standard logistic regression model for case-control phenotypes (PLINK-Logit), and (c) a liability threshold exhaustive search model (LT-PLINK). In the last case, we simply fit the linear regression version of PLINK with the same estimated liability values that are given to LT-MAPIT. The PLINK software is publicly available at `https://www.cog-genomics.org/plink2/epistasis`.

Simulated Data Generation

To validate LT-MAPIT in terms of its ability to preserve type I error control, and to assess its power compared to the aforementioned competing methods, we carried out a series of simulation studies under a wide range of configurations. Each of these experiments make use of synthetic genotypes that are independently generated to have $p = 6000$ SNPs with allele frequencies randomly sampled from a uniform distribution over values ranging from $[0.05, 0.5]$. We do not simulate these variants in linkage disequilibrium, partly because of the heavy computational burden involved in generating correlated genotypes under the liability threshold model setting [58], and partly because it has been shown that the distribution of genetic similarity matrices is not affected much by its presence [87]. These genotypes are simulated for a population of size $n = 2.5 \times 10^6$. Phenotypes are then generated on the continuous latent liability scale. Disease status for each individual is determined by comparing each simulated l_i with respect to a threshold of liability that is predetermined by the population prevalence parameter γ. In this work, we consider three levels of prevalence $\gamma = 0.1\%$, 0.25%, and 0.5%. For each simulated run, an equal number of (randomly selected) cases and controls contribute to the case-control sample. All simulations are replicated 100 times and used a total of $n = 5000$ individuals (2500 cases and 2500 controls).

Preprocessing of WTCCC GWAS Data

The Wellcome Trust Case Control Consortium (WTCCC) 1 study GWAS data [61] (`http://www.wtccc.org.uk/`) consists of about 14,000 cases of seven common diseases, including 1,868 cases of bipolar disorder (BD), 1,926 cases of coronary artery disease (CAD), 1,748 cases of Crohn's disease (CD), 1,952 cases of hypertension (HT), 1,860 cases rheumatoid arthritis (RA), 1,963 cases of type 1 diabetes (T1D) and 1,924 cases of type 2 diabetes (T2D), as well as 2,938 shared controls. We selected an initial set of 458,868 total shared SNPs following the quality control procedures of previous studies [28, 82, 88]. Then for each trait, we retained final data sets with about 363,000 SNPs that had a minor allele frequency (MAF) above 0.05. The disease prevalence γ for each trait was determined by referencing past work [89–95] (see Table S1 in Supporting Information).

In the WTCCC GWAS data set, we perform three sets of analyses. The first involves using LT-MAPIT with the usual genome-wide genetic relatedness matrix, where for each SNP r in turn, the matrix $\mathbf{K}_{GW}^{(r)}$ is computed using all other SNPs $s \neq r$ in the study. The second analysis involves using LT-MAPIT with a *cis*-genetic relatedness matrix, where alternatively for each SNP r, the matrix $\mathbf{K}_{cis}^{(r)}$ is computed using only variants within a 1 Mb window of SNP r on the same chromosome. The last analysis is centered around fitting LT-MAPIT with a genetic relatedness matrix $\mathbf{K}_{trans}^{(r)}$, which is computed using only the corresponding *trans*-SNPs located outside of the previously defined 1 Mb *cis*-window for a given SNP r. In each of the analyses, SNPs are mapped to the closest neighboring gene(s) using the dbSNP database provided by the National Center for Biotechnology Information (`http://www.ncbi.nlm.nih.gov/SNP/`).

Results

Null Simulations and Type I Error Control

We make use of the following simulation scheme in order to investigate whether LT-MAPIT preserves the desired type I error rate and produces well-calibrated p-values under the null hypothesis. In these numerical experiments, we begin by simulating $p = 6000$ SNPs for a population of $n = 2.5 \times 10^6$ individuals, and generate their liability phenotypes using a linear regression model based on various genetic architectures. Specifically, with these simulated genotypes, we randomly select 1000 causal additive SNPs and draw their corresponding effect sizes from a standard normal distribution. Next, we create residual errors also from a standard normal distribution. Both the genetic effects of the causal SNPs and the effects of the residual errors are scaled to ensure a narrow-sense heritability of 60% on the liability scale. Afterwards, based on predetermined population prevalence levels $\gamma = \{0.1\%, 0.25\%, 0.5\%\}$, we select the γ proportion of individuals who have the largest liability values to be cases and the remaining to be controls. We then randomly select 2500 cases and 2500 controls as our case-control data to perform analysis. For each γ, we performed 500 simulation replicates. Note that, due to the desired experimental property of case-control ascertainment (and the consequent necessary large population size), we are unable to use real genotypes and hence make use of synthetic data.

In these simulations, the idea of the null model holds because there are no interaction effects present during the generation of phenotypes, and LT-MAPIT solely searches for significant marginal epistatic effects that are a summation of pairwise interactions. All evaluations of test calibration and type 1 error control are strictly based on the liability threshold mixed model fitting algorithm presented in the Equations (5)-(8); and the p-values used for model assessment correspond to the readjustment procedure proposed in Equation (9). Figure 1 shows the quantile-quantile (QQ) plots based on the distribution under the null hypothesis. Similarly, Table 1 shows the empirical type I error rates estimated for LT-MAPIT at significance levels $\alpha = 0.05$, 0.01, 0.001, and 0.0001, respectively. Despite being slightly liberal in the assignment of p-values at smaller prevalences (i.e. $\gamma = 0.1\%$ case) and at larger critical values (i.e. $\alpha = 0.05$ or 0.01 cases), LT-MAPIT is able to appropriately control the type I error rate for

smaller critical values across various prevalences.

Power Assessment and Method Comparisons

The performance of LT-MAPIT approach is evaluated through comparative studies with existing works (see Material and Methods for details). Our goal here is two-fold. First, we want to detect individual variants that are involved in epistasis. Second, we want to explicitly identify pairwise (or higher-order) epistatic interactions. Among many recently proposed epistatic mapping procedures, we focus on comparing LT-MAPIT against two marginal epistasis detection methods (MAPIT and MAPIT-Logit [39]), and three exhaustive search models (PLINK, PLINK-Logit, and LT-PLINK [86]).

We utilize the following simulation design to assess the utility of LT-MAPIT for case-control studies. Here, we again use simulated genotypes to generate continuous phenotypes (on the liability scale) that this time mirror genetic architectures affected by a combination of additive and pairwise epistatic effects. Specifically, we randomly choose 1,000 causal SNPs to directly affect the liability and which classify into three groups: (1) a small set of interaction SNPs, (2) a larger set of interaction SNPs, and (3) a large set of additive SNPs. Here, SNPs interact between sets, so that SNPs in the first group interact with SNPs in the second group, but do not interact with variants in their own group (the same rule applies to the second group). All causal SNPs in both the first and second groups have additive effects and are involved in pairwise interactions, while causal SNPs in the third set only have additive effects. Both the additive and epistatic interaction effect sizes are drawn from standard normal distributions. We scale these genetic effects so that collectively they explain a fixed proportion of the total variance on the liability scale, which again is set to be 60%. In addition, this scaling ensures that the additive effects make up a proportion ρ of the total genetic variance, while the pairwise interactions make up the remaining proportion $(1-\rho)$. Altogether, we consider simulations that depend on the three aforementioned parameters:

- γ: the assumed disease prevalence in the population, as previously defined;
- $(1-\rho)$: the proportion of genetic variance (on the liability scale) that is contributed by the interaction effects between the first and second groups of causal SNPs;
- $p_1/p_2/p_3$: the number of causal SNPs in each of the three groups, respectively.

We set $\rho = \{0.5, 0.8\}$, and define four scenarios corresponding to $p_1/p_2/p_3 = 10/10/980$ (scenario I), $10/20/970$ (scenario II), $10/50/940$ (scenario III), and $10/100/890$ (scenario IV). In the case where $\rho = 0.5$, the additive and epistatic effects are assumed to equally contribute to the liability phenotypes. In the alternative case, where $\rho = 0.8$, the liabilities are dominated by additive effects. We analyze 100 different simulated datasets for each value of ρ, across each of the four scenarios. Unless stated otherwise, all of the results described in the main text are based on the cases in which case-control data are simulated with disease prevalence $\gamma = 0.25\%$. The results for $\gamma = 0.1\%$ and 0.5% can be found in Supporting Information (see Figures S1-S4).

Figure 2 shows the empirical power of LT-MAPIT to detect both group 1 and 2 causal variants, respectively, under a Bonferroni corrected genome-wide significance threshold. Overall, we see that LT-MAPIT's ability to detect epistatic markers mostly depends on the variance explained by the epistatic SNPs. For example, in Figure 2(a), each causal variant in group 1 is expected to explain $(1-\rho)/p_1 = 5\%$ of the genetic variance on the liability scale — while in Figure 2(b), these same variants are expected to only explain 2%. In these situations, due to these large effects, LT-MAPIT's power is greater for detecting group 1 variants in all four scenarios. It is important to note that, unlike in the case for purely quantitative traits (without ascertainment) [39], the liability threshold specification of MAPIT demonstrates an empirical power that also depends on the total interaction effect size. More specifically, LT-MAPIT's ability to detect group 1 causal epistatic SNPs is slightly affected by the number of group 2 causal SNPs, which also determines the pairwise interaction effect size in simulations. Therefore, LT-MAPIT displays great power when there are only a small number of interactions, each with a large effect

size (e.g. scenarios I and II); however, its power marginally decreases in the more polygenic setting where there are a large number of interactions, each with a small effect size (e.g. scenarios III and IV). We hypothesize that this behavior occurs for case-control data because the posterior sampling of the latent liability phenotypes is directly impacted by the magnitude of the pairwise interaction effect size.

Figure 3 depicts the direct comparison between the MAPIT-based approaches and the PLINK-based exhaustive searches to accurately identify marginal epistatic markers in each of the two causal groups. Here, we assess power by comparing each model's true positive rate (TPR) at a 1% false positive rate (FPR). We consider the marginal epistatic detection in the PLINK exhaustive searches by first ordering the resulting p-values for all possible interactions, and then computing the power for correctly identifying unique causal SNPs. For example, if the top p-values from an exhaustive search include interactions SNP1-SNP2, SNP2-SNP3, SNP4-SNP5, and only SNP2 is the true causal epistatic variant, then the top three pairs marginally identify only one true variant and four false variants. Under this evaluation, the methods that do not account for ascertainment (MAPIT, MAPIT-Logit, PLINK, and PLINK-Logit) consistently perform the worst. LT-MAPIT constantly performs the best when the proportions of phenotypic variance explained by additive and epistatic effects are equal (i.e. $\rho = 0.5$). However, when additivity dominates this variance, LT-MAPIT's advantage over LT-PLINK depends on the disease prevalence. When γ is small, LT-MAPIT maintains comparable power; while, LT-PLINK gains an advantage as γ increases.

In addition to detecting marginal epistatic effects, we also assessed each method's ability to explicitly identify significant interactions. Specifically, we investigate the use of LT-MAPIT as the initial step in a two-step prioritization based association mapping procedure [39]. Under this type of approach [45,49,51], we first apply LT-MAPIT as a single-SNP test to identify associated genetic variants with non-zero marginal epistatic effects, and then focus on the significantly identified variants to test all pairwise interactions between them. Figure 4 compares the utility of this prioritization procedure when selecting the top $v = \{50, 100, 250, 500\}$ SNPs as opposed to ranking significant interactions estimated via the liability threshold exhaustive search LT-PLINK. Altogether, just searching over the top 250 prioritized SNPs using LT-MAPIT provides more power in finding true pairwise epistatic interactions across all scenarios. This holds true even for most of the cases in which the additive effects contribute to a majority of the variance on the liability scale.

Lastly, we check method sensitivity to correct prevalence specification. To evaluate this, we run a final simulation study where we set the true disease prevalence in the population to be $\gamma = 0.25\%$, but then generate latent liabilities for LT-MAPIT and LT-PLINK using misspecified thresholds based on prevalences $\widetilde{\gamma} = 0.1\%$ (low) and $\widetilde{\gamma} = 0.5\%$ (high). In Supporting Information (see Figures S5-S6), we show that LT-MAPIT is generally more robust to a liberal misspecification of the ascertainment threshold (i.e. $\widetilde{\gamma} = 0.5\%$), as opposed to a conservative one (i.e. $\widetilde{\gamma} = 0.1\%$). This is particularly apparent in the $\rho = 0.8$ simulation cases where additive effects dominant the phenotypic variance. While the performance of LT-PLINK is seemingly unaffected by the misspecification of the liability threshold when it comes to identifying marginal epistatic variants, it also follows the same liberal versus conservative behavior as LT-MAPIT when detecting explicit epistatic pairs (see Figure S6). Nonetheless, LT-PLINK still underperforms overall when compared to LT-MAPIT in most of the simulation scenarios we consider.

Practical Application to WTCCC Data

In addition to our simulation study, we further assess LT-MAPIT's ability to detect marginal and pairwise epistasis in the seven complex diseases provided by the Wellcome Trust Case Control Consortium (WTCCC) 1 study [61]. Briefly, these diseases 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) — where the disease prevalence γ for each trait has been previously documented [89–95]. For an in-depth analysis on the summed epistatic effects involving any given SNP, we implement LT-MAPIT using three distinct genetic relatedness matrices: (i) $\mathbf{K}_{GW}$ which tests against all SNPs genome-wide, (ii) $\mathbf{K}_{cis}$ which tests against all *cis*-SNPs within a given 1 Mb genomic window,

and (iii) $\mathbf{K}_{trans}$ which tests against all other SNPs located outside the *cis*-window (details in Materials and Methods). Intuitively, using $\mathbf{K}_{cis}$ will be more powerful than using either $\mathbf{K}_{trans}$ or $\mathbf{K}_{GW}$ if epistatic interactions are more likely to happen between *cis*-SNP pairs, rather than between SNPs that are located genome-wide — however, less powerful otherwise.

For each trait, we provide summary tables which list the marginal epistatic p-values for all SNPs when the LT-MAPIT model is fit using the different genetic relatedness matrices (see Table S2-S8 in Supporting Information). Corresponding QQ-plots of the $-\log_{10}(P)$ transformed p-values visually display this information in Figure 5. To contrast LT-MAPIT's marginal association findings, we again directly compare results from the liability threshold exhaustive search model LT-PLINK. The first half of Table S9 details comparisons between these two approaches and their effectiveness in finding marginally epistatic variants across all seven traits. We are reminded that LT-MAPIT searches over a reduced space, and thus only requires p tests for a data set with p genetic markers. Therefore, a significant marginal association for a particular SNP identified by LT-MAPIT is determined by using a (traditional) trait specific Bonferroni-corrected significance p-value threshold $P = 0.05/p_d$, where p_d is the number of SNPs for disease trait d. For LT-PLINK, a single SNP was deemed marginally significant if it belonged to an epistatic pair with a p-value below the genome-wide threshold $P = 0.05/C(p_d, 2)$ — where $C(\bullet, 2)$ denotes the binomial coefficient enumerating all possible SNP pairs. Note that this is equivalent to implementing a Bonferroni-corrected threshold for exhaustive search methods [49].

In Table 2, we list all genes found by LT-MAPIT with at least two SNPs having marginal epistatic p-values that satisfy the Bonferroni-corrected significance cutoffs. Overall, LT-MAPIT was able to identify noteworthy SNPs found within the coding regions of genes that are known to have relevant associations. For example, in BD and RA, the genome-wide and *trans*-based analyses detected variants with the lowest marginal epistatic p-values near the gene *CD200*. In the context of BD, *CD200* is known to be involved in microglial activation and plays a pivotal role in pathways connected to psychiatric disorders [96, 97]. Alternatively, when studying RA, irregular expression of *CD200* has been shown to contribute to abnormal Th17 cell differentiation [98], regulate myeloid cell function [99], and has been proposed as a potential therapeutic target [100, 101]. For HT, the *cis*-window version of LT-MAPIT identified significant variants near the gene *PSAT1*, whose low expression levels serve as a biomarker of decreased serine biosynthesis pathway activity and increased risk for pulmonary hypertension [102, 103]. Similarly, *IFNB1* has been validated as a gene that affects immunity in T1D [104–106] and is a key member of the inflammatory signature for progressive diabetic nephropathy in T2D [106, 107].

As expected from previous statistical epistatic methods that have heavily studied this data [49, 55, 56, 108, 109], LT-MAPIT most noticeably identified the major histocompatibility complex (MHC) region on chromosome 6 in RA and T1D as having many genetic variants that are most likely involved in prominent pairwise interactions. This is clearly visible in Figures S7 and S8 which display manhattan plots of our marginal epistatic genome-wide scans for these two traits. In these figures, spikes across chromosomes suggest loci where members involved in epistatic interactions can be found. Moreover, when strictly considering *cis*-relationships between SNPs via the relatedness matrix $\mathbf{K}_{cis}$, LT-MAPIT was able to detect more marginal signal than LT-PLINK in both RA and T1D (again see Table S9). These results are unsurprising since it is well known that the MHC region holds significant clinical relevance in complex traits and diseases with respect to infection, inflammation, autoimmunity, and transplant medicine [110, 111]. Since this region has also been consistently implicated by other epistatic analyses, we focus our search for significant pairwise interactions to these two phenotypic traits (see the second half of Table S9 for results pertaining to the other five traits).

Similar to what was done in the simulation studies, we utilize the proposed LT-MAPIT prioritization approach to detect explicit epistatic pairs in RA and T1D. To evaluate the power of this method, we first rank SNPs according to their marginal epistatic p-values and then search for significantly associated epistatic pairs between the top v SNPs. Interactions between two variants were then called significant if they had a joint p-value below the Bonferroni-corrected threshold $\alpha = 0.05/C(v, 2)$. Here, we choose

two sets of prioritization thresholds: one low-dimensional threshold $v = \{10, 50, 100, 200\}$ (see Figure S9(a)) and one high-dimensional threshold $v = \{500, 1000, 2000, 5000\}$ (see Figure S9(b)). In each case, we compared results with significant pairs detected by exhaustive search via LT-PLINK, as a baseline. Naturally, this comparison serves as a way for us to continue our assessment of LT-MAPIT as a viable prioritization approach within the context of real case-control data. Altogether, the exhaustive search method identified 75 significant epistatic pairs in RA and 2753 epistatic pairs in T1D. As demonstrated in the numerical experiments, using LT-MAPIT (particularly with $\mathbf{K}_{cis}$) efficiently improves upon this power.

Discussion

We have presented LT-MAPIT for identifying variants that are involved in epistasis in ascertained case-control GWASs. For each variant in turn, LT-MAPIT tests for marginal epistatic effects — the combined epistatic effect between the examined variant and all other variants — to identify variants that exhibit non-zero interactions with another variant without the need to select the specific marker combinations that drive the epistatic association. LT-MAPIT extends the previous MAPIT method for quantitative traits [39] to case-control studies and represents an attractive alternative to conventional epistatic mapping methods [47–49, 56]. With both simulations and real data applications, we have illustrated the benefits of LT-MAPIT.

We have primarily focused on using a liability threshold mixed model to account for case-control ascertainment and the binary nature of case-control data. An alternative approach commonly used for modeling case-control studies and performing variance component test is the logistic mixed model. Both the liability threshold mixed model and the logistic mixed model are prospective models that treat the binary label of case-control status as the outcome variable. The liability threshold mixed model directly accounts for the retrospective sampling in case-control studies by explicitly modeling case-control ascertainment. The prospective logistic model without random effects, on the other hand, is also well known to yield valid likelihood based inference of the odds ratio which is equivalent to a corresponding retrospective logistic model under case-control ascertainment [112]. Unfortunately, the equivalence of the likelihood based inference between a prospective logistic model and a retrospective logistic model does not directly imply an equivalence of likelihood based inference between a prospective logistic mixed model and a retrospective logistic mixed model. Indeed, it is not trivial to derive a corresponding retrospective model for a prospective logistic mixed model, or the other way around, due to case-control ascertainment [113,114]. Consequently, it is also not straightforward to treat the multiple genetic variance components in the prospective logistic mixed model as population level parameters — which we can do with a liability threshold mixed model. Because of the lack of a corresponding retrospective model, it has been shown that the prospective logistic mixed model may not be robust to various modeling misspecifications [115]. Indeed, here we also found that the logistic mixed model does not provide sufficient power in the presence of retrospective sampling and case-control ascertainment. Nevertheless, the inference of logistic mixed model has been well established in the statistics literature [116–119], and the model has been recently applied to perform association tests in case-control studies [120] as well as in RNA sequencing and bisulfite sequencing studies [83–85]. Extending and adapting the logistic mixed model for mapping marginal epistasis would be an interesting avenue to explore in the future.

There are several potential extensions of LT-MAPIT. For example, while we have focused on detecting variants that are involved in pairwise interactions with other variants, LT-MAPIT can be easily extended to detect variants that are involved in higher-order interactions. In particular, we can introduce extra random effects terms to represent the combined higher-order interaction effects between the examined variant and all other variants. Under the normality assumption for the interaction effect sizes, the introduced random effects terms would all follow multivariate normal distributions with the covariance matrices determined as a function of the Hadamard product of the additive genetic relatedness matrix,

where the power of the Hadamard product depends on the particular higher-order term one is interested in modeling [33,121,122]. Therefore, in the presence of higher-order interactions, we can apply a multiple variance component model with additional random effects to map these epistatic variants. As another example, we have assumed in LT-MAPIT that the interaction effect between the examined variant and every other variant follows a normal distribution. This normality assumption has been widely applied in many areas of genetics [71, 82, 123] and is known to produce practically unbiased estimates of the total effects, regardless of whether or not the underlying effect sizes actually follow a normal distribution [82,123]. However, the main idea in LT-MAPIT of mapping marginal epistatic effects is not restricted to the normality assumption for the interaction effect sizes. Indeed, we can incorporate sparsity-inducing priors for effect sizes if the proportion of interaction pairs is known to be small *a priori*, or we can use hybrid effect size priors to accommodate a larger variety of distributions [82, 88]. Different interaction effect size assumptions can be advantageous under different genetic architectures and will likely improve the power of LT-MAPIT further.

LT-MAPIT is not without its limitations. Like other marginal epistatic mapping methods, LT-MAPIT is unable to directly identify detailed interaction pairs despite being able to identify SNPs that are involved in epistasis. However, being able to identify SNPs involved in epistasis allows us to come up with an initial (likely) set of variants that are worth further exploration, and thus represents an important first step towards identifying and understanding detailed epistatic associations. Indeed, we have illustrated a two-step *ad hoc* epistatic association mapping procedure, where we first identify individual SNP associations with LT-MAPIT and then we focus on the most significant associations from the first step to further test all of the pairwise interactions among them in order to identify specific epistatic interaction pairs. Unlike the previous prioritization strategies commonly used in this space, which prioritize SNPs based on additive effects, our two-step procedure is unique in the sense that the SNP set identified in our first step contains SNPs that already display strong epistatic effects with other variants. Therefore, our two-step procedure outperforms alternative prioritization strategies in simulations and real data applications. Nonetheless, we caution that the two-step procedure is still *ad hoc* in nature and could miss important epistatic associations. Exploring statistical approaches that can unify the two steps in a joint fashion would be an interesting area for future research. Besides this main limitation, we also note that LT-MAPIT can be computationally expensive. LT-MAPIT requires fitting a different variance component model for every SNP in turn, and fitting variance component models is known to be computationally challenging [81,124]. In this study, we rely on various approximations to the liability threshold model and the recently developed efficient MQS method to enable efficient variance component test. Compared with the standard REML method, MQS is computationally efficient, allows for accurate p-value computation based on the Davies method, and is statistically produces more accurate estimates than the REML when the variance component is small [60] — a property that is particularly relevant here considering the marginal epistatic effect size is often small in real data. With these techniques, LT-MAPIT is applicable to moderately sized genetic mapping studies with thousands of samples and millions of variants. Still, new algorithms are likely needed to scale LT-MAPIT up to datasets that are orders of magnitude larger.

Acknowledgements

LC would like to acknowledge the support of Institutional Development Award Number P20GM109035 from the National Institute of General Medical Sciences of the National Institutes of Health, which funds COBRE Center for Computational Biology of Human Disease. XZ would like to acknowledge the support of National Institutes of Health (NIH) Grant R01HG009124 and the support of National Science Foundation (NSF) Grant DMS1712933. This study also makes use of data generated by the Wellcome Trust Case Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the data is available from `www.wtccc.org.uk`. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113 and 085475. Any opinions, findings, and conclusions

or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders or supporters.

Figures and Tables

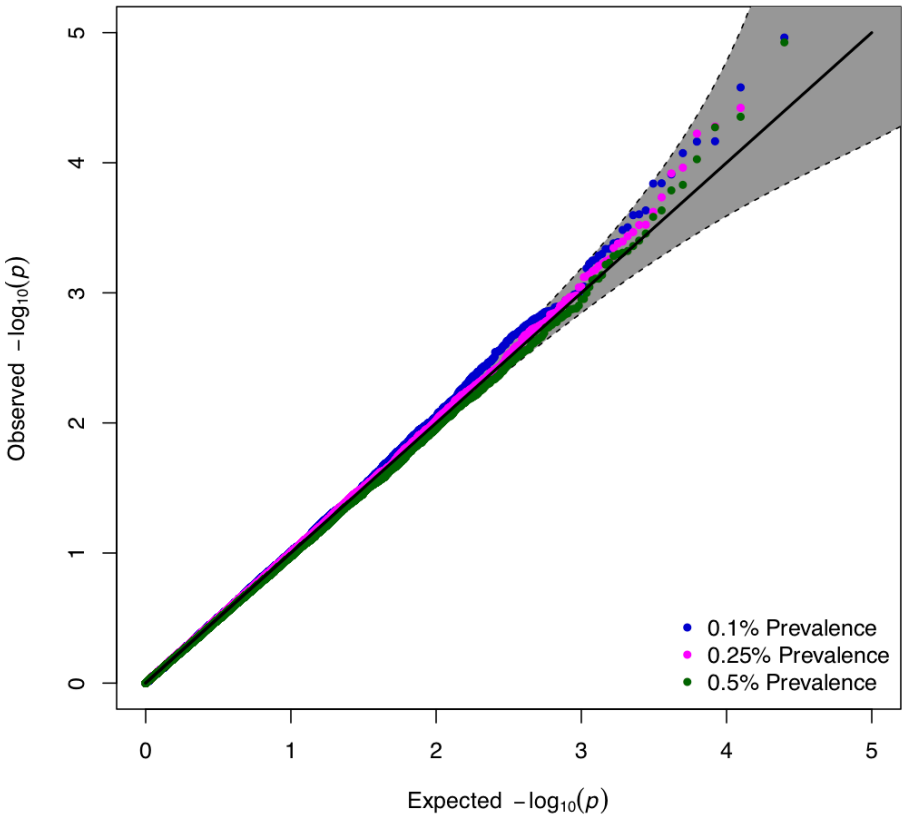
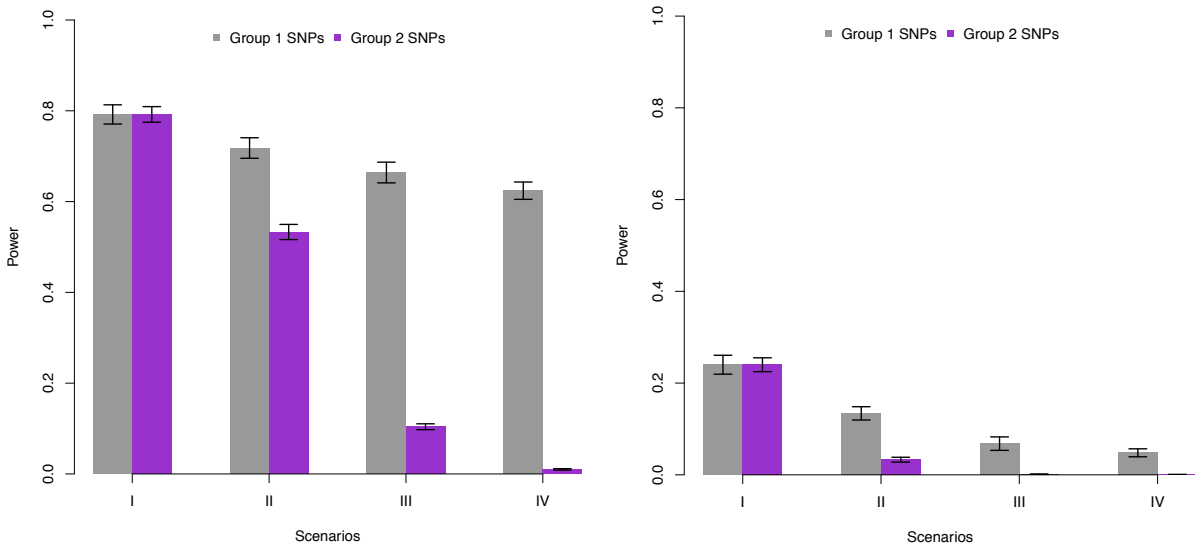


Figure 1. Calibration of p-values produced by LT-MAPIT via QQ-plots. The QQ-plots applying MAPIT to 500 simulated null datasets assuming disease prevalence in the population set to $\gamma = 0.1\%$ (blue), 0.25% (pink), and 0.5% (green). The 95% confidence interval for the null hypothesis of no association is shown in grey. Sample size was 5000 (2500 cases, 2500 controls) for all cases.

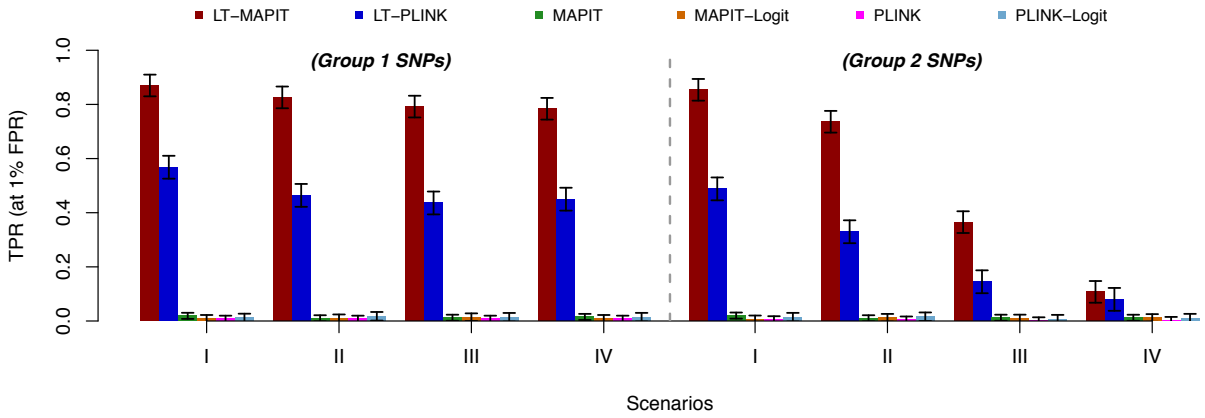
Table 1. Empirical type I error estimates of LT-MAPIT. Each entry represents type I error rate estimates as the proportion of significant p-values under the null hypothesis. Phenotypes were generated on the liability scale with disease prevalence in the population set to $\gamma = 0.1\%$, 0.25%, and 0.5%. Empirical size for the analyses used significance thresholds of $\alpha = 0.05$, 0.01, and 0.001. Sample size was 5000 (2500 cases, 2500 controls) for all cases. Results are based on 500 simulated data sets. Values in the parentheses are the standard deviations across replicates.

Prevalence (γ)	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.0001$
0.1%	0.0572 (0.0102)	0.0136 (0.0033)	0.0010 (0.0012)	0.0001 (0.0001)
0.25%	0.0484 (0.0025)	0.0114 (0.0018)	0.0008 (0.0013)	0.0002 (0.0005)
0.5%	0.0454 (0.0050)	0.0106 (0.0029)	0.0008 (0.0013)	0.0002 (0.0005)

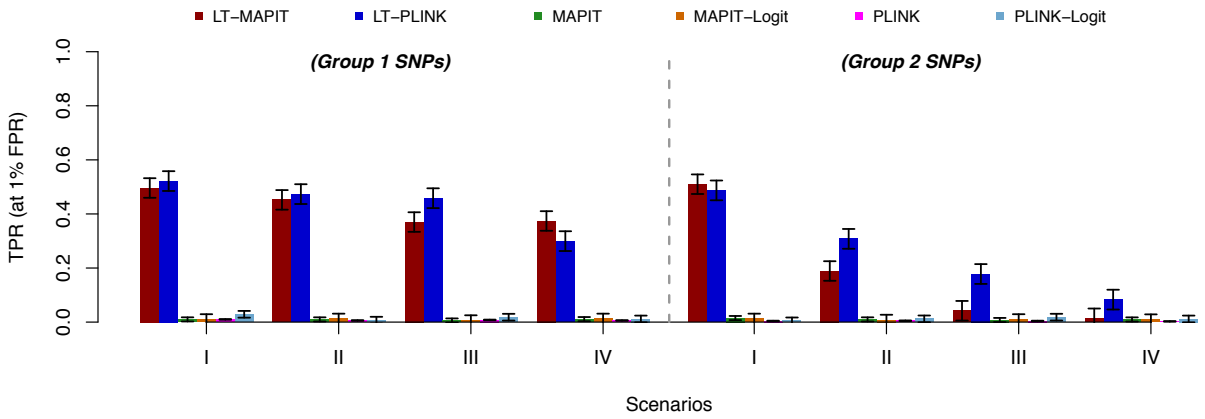


(a) $\gamma = 0.25\%$ and $\rho = 0.5$ (b) $\gamma = 0.25\%$ and $\rho = 0.8$

Figure 2. 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 (a) and (b) 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 $\gamma = 0.25\%$. 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. Results are based on 100 simulated replicates and the lines on each bar represent a 95% standard error interval due to resampling.

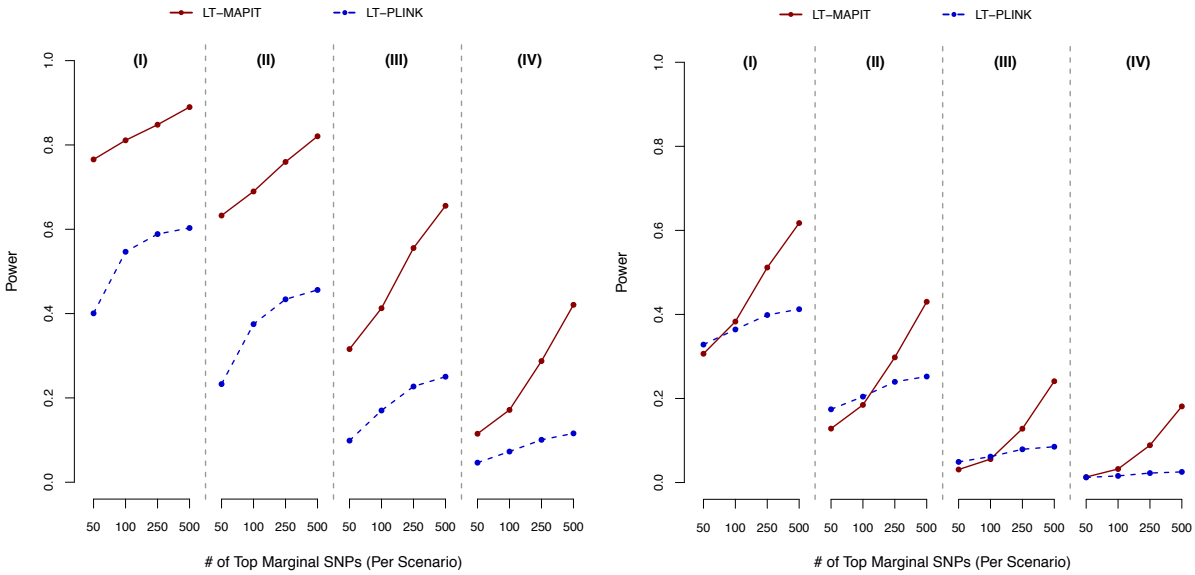


(a) $\gamma = 0.25\%$ and $\rho = 0.5$



(b) $\gamma = 0.25\%$ and $\rho = 0.8$

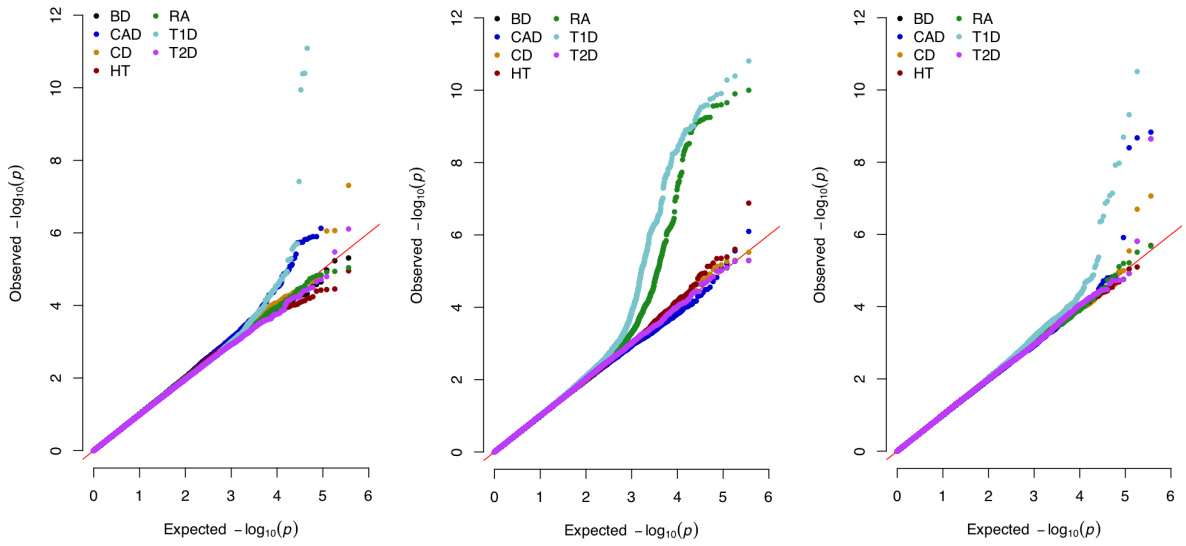
Figure 3. Power comparison for detecting group 1 and group 2 causal SNPs. The association mapping ability of LT-MAPIT (red) is compared to five competing methods: (i) the original linear mixed model version of MAPIT (green), (ii) the binary mixed model MAPIT-Logit (orange) fit by using penalized quasi-likelihoods (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 $\gamma = 0.25\%$. 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) $\gamma = 0.25\%$ and $\rho = 0.5$

(b) $\gamma = 0.25\%$ and $\rho = 0.8$

Figure 4. Power comparison with exhaustive search procedures to detect epistatic pairs. LT-MAPIT (red) is assessed as an initial step in a pairwise epistatic prioritization 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 $\gamma = 0.25\%$.



(a) Results with $\mathbf{K}_{GW}$ (b) Results with $\mathbf{K}_{cis}$ (c) Results with $\mathbf{K}_{trans}$

Figure 5. QQ-plots of LT-MAPIT p-values 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). Each figure corresponds to using LT-MAPIT with a different genetic relatedness matrix. Namely, Figure (a) corresponds to fitting the liability threshold mixed model with a genetic relatedness matrix $\mathbf{K}_{GW}$, which is computed using all SNPs in the study. Figure (b) corresponds to using $\mathbf{K}_{cis}$, which is computed using only variants within a 1 Mb window (on the same chromosome) from the SNP of interest. Figure (c) are results from using $\mathbf{K}_{trans}$, which is computed using only the corresponding *trans*-SNPs located outside of the previously defined 1 Mb *cis*-window.

Table 2. Gene coding regions with at least two SNPs having marginal epistatic p-values that satisfy Bonferroni-corrected thresholds 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). Listed for all regions are the SNPs with the lowest p-values. The reference column gives literature sources that have previously suggested some level of association between a given gene and disease. Columns 4-6 correspond to using LT-MAPIT with different genetic relatedness matrices. Namely, these include: 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. §: Multiple SNPs in the MHC gene class regions are significant, so a single SNP with the lowest marginal epistatic p-value is reported as the most extreme. There are three subregions that divide the MHC region: the class I region (29.8Mb-31.6Mb), the class III region (31.6Mb-32.3Mb), and the class II region (32.3Mb-33.4Mb).

Disease	Chromosome	Nearest Gene(s)	P-Value (GW)	P-Value (Cis)	P-Value (Trans)	Reference(s)
BD	3	CD200	≈ 0 (rs6438075)	—	≈ 0 (rs6438075)	[96, 97]
CAD	1	LMX1A	1.62×10^{-14} (rs16841238)	—	1.47×10^{-9} (rs16841238)	—
HT	9	PSAT1	—	1.32×10^{-7} (rs944513)	—	[102, 103]
RA	3	CD200	≈ 0 (rs6438075)	—	≈ 0 (rs6438075)	[98–101]
RA	6	MHC Class I§	—	8.93×10^{-8} (rs9266775)	—	[125, 126]
RA	6	MHC Class III§	—	1.00×10^{-10} (rs9268480)	—	[125, 126]
RA	6	MHC Class II§	—	5.85×10^{-10} (rs2857154)	—	[125, 126]
RA	9	TPD52L3	≈ 0 (rs2290913)	—	≈ 0 (rs2290913)	—
T1D	6	MHC Class I§	—	5.47×10^{-11} (rs261948)	2.03×10^{-9} (rs2523485)	[127, 128]
T1D	6	MHC Class III§	4.87×10^{-23} (rs3132959)	3.73×10^{-10} (rs3131294)	2.82×10^{-22} (rs2736177)	[127, 128]
T1D	6	MHC Class II§	1.82×10^{-17} (rs9272723)	1.29×10^{-10} (rs2858308)	4.85×10^{-10} (rs2647015)	[127, 128]
T1D	9	IFNB1	≈ 0 (rs6475502)	—	≈ 0 (rs6475502)	[104–106]
T2D	1	LMX1A	≈ 0 (rs16841238)	—	—	[129, 130]
T2D	9	IFNB1	≈ 0 (rs6475502)	—	—	[106, 107]

References

1. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008;9(11):855–867. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2689140/.

2. Mackay TFC. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat Rev Genet. 2014;15(1):22–33. Available from: http://dx.doi.org/10.1038/nrg3627.

3. Peripato A, De Brito R, Matioli S, Pletscher L, Vaughn T, Cheverud J. Epistasis affecting litter size in mice. Journal of Evolutionary Biology. 2004;17(3):593–602.

4. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, et al. Global Mapping of the Yeast Genetic Interaction Network. Science. 2004;303(5659):808. Available from: http://science.sciencemag.org/content/303/5659/808.abstract.

5. Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature. 2005;436(7051):701–703. Available from: http://dx.doi.org/10.1038/nature03865.

6. Deutschbauer AM, Davis RW. Quantitative trait loci mapped to single-nucleotide resolution in yeast. Nat Genet. 2005;37(12):1333–1340. Available from: http://dx.doi.org/10.1038/ng1674.

7. Kroymann J, Mitchell-Olds T. Epistasis and balanced polymorphism influencing complex trait variation. Nature. 2005;435(7038):95–98.

8. Collins SR, Schuldiner M, Krogan NJ, Weissman JS. A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biology. 2006;7(7):R63. Available from: http://dx.doi.org/10.1186/gb-2006-7-7-r63.

9. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. Nat Genet. 2006;38(8):896–903. Available from: http://dx.doi.org/10.1038/ng1844.

10. Onge RPS, Mani R, Oh J, Proctor M, Fung E, Davis RW, et al. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. Nat Genet. 2007;39(2):199–206. Available from: http://dx.doi.org/10.1038/ng1948.

11. Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet. 2007;3(9):e162.

12. Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ. Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. The Plant Cell. 2008;20(5):1199–1216.

13. Shao H, Burrage LC, Sinasac DS, Hill AE, Ernest SR, O'Brien W, et al. Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. Proceedings of the National Academy of Sciences. 2008;105(50):19910–19914. Available from: http://www.pnas.org/content/105/50/19910.abstract.

14. Flint J, Mackay TFC. Genetic architecture of quantitative traits in mice, flies, and humans. Genome Research. 2009;19(5):723–733. Available from: http://genome.cshlp.org/content/19/5/723.abstract.

15. Gerke J, Lorenz K, Cohen B. Genetic Interactions Between Transcription Factors Cause Natural Variation in Yeast. Science. 2009;323(5913):498. Available from: http://science.sciencemag.org/content/323/5913/498.abstract.

16. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. The Genetic Landscape of a Cell. Science. 2010;327(5964):425. Available from: http://science.sciencemag.org/content/327/5964/425.abstract.

17. He X, Qian W, Wang Z, Li Y, Zhang J. Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. Nat Genet. 2010;42(3):272–276. Available from: http://dx.doi.org/10.1038/ng.524.

18. Horn T, Sandmann T, Fischer B, Axelsson E, Huber W, Boutros M. Mapping of signaling networks through synthetic genetic interaction analysis by RNAi. Nat Meth. 2011;8(4):341–346. Available from: http://dx.doi.org/10.1038/nmeth.1581.

19. Jarvis JP, Cheverud JM. Mapping the Epistatic Network Underlying Murine Reproductive Fatpad Variation. Genetics. 2011;187(2):597. Available from: http://www.genetics.org/content/187/2/597.abstract.

20. Leamy L, Gordon R, Pomp D. Sex-, Diet-, and Cancer-Dependent Epistatic Effects on Complex Traits in Mice. Frontiers in Genetics. 2011;2:71. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2011.00071.

21. Pettersson M, Besnier F, Siegel PB, Carlborg Ö. Replication and explorations of high-order epistasis using a large advanced intercross line pedigree. PLoS Genet. 2011;7(7):e1002180.

22. Szappanos B, Kovacs K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, et al. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet. 2011;43(7):656–662. Available from: http://dx.doi.org/10.1038/ng.846.

23. Gaertner BE, Parmenter MD, Rockman MV, Kruglyak L, Phillips PC. More Than the Sum of Its Parts: A Complex Epistatic Network Underlies Natural Variation in Thermal Preference Behavior in *Caenorhabditis elegans*. Genetics. 2012;192(4):1533. Available from: http://www.genetics.org/content/192/4/1533.abstract.

24. Bloom JS, Ehrenreich IM, Loo WT, Lite TLV, Kruglyak L. Finding the sources of missing heritability in a yeast cross. Nature. 2013;494(7436):234–237. Available from: http://dx.doi.org/10.1038/nature11867.

25. Chari S, Dworkin I. The Conditional Nature of Genetic Interactions: The Consequences of Wild-Type Backgrounds on Mutational Interactions in a Genome-Wide Modifier Screen. PLoS Genet. 2013;9(8):e1003661. Available from: http://dx.doi.org/10.1371%2Fjournal.pgen.1003661.

26. Monnahan PJ, Kelly JK. Epistasis Is a Major Determinant of the Additive Genetic Variance in *Mimulus guttatus*. PLoS Genet. 2015;11(5):e1005201. Available from: http://dx.doi.org/10.1371%2Fjournal.pgen.1005201.

27. Huang W, Mackay TFC. The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. PLOS Genetics. 2016;12(11):e1006421. Available from: http://dx.doi.org/10.1371%2Fjournal.pgen.1006421.

28. Crawford L, Wood KC, Zhou X, Mukherjee S. Bayesian Approximate Kernel Regression With Variable Selection. Journal of the American Statistical Association. 2018;p. 1–12. Available from: https://doi.org/10.1080/01621459.2017.1361830.

29. Carlborg O, Jacobsson L, Ahgren P, Siegel P, Andersson L. Epistasis and the release of genetic variation during long-term selection. Nat Genet. 2006;38(4):418–420. Available from: http://dx.doi.org/10.1038/ng1761.

30. Martin G, Elena SF, Lenormand T. Distributions of epistasis in microbes fit predictions from a fitness landscape model. Nat Genet. 2007;39:555–560.

31. Forsberg SKG, Bloom JS, Sadhu MJ, Kruglyak L, Carlborg Ö. Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. Nature Genetics. 2017;49:497–503. Available from: http://dx.doi.org/10.1038/ng.3800.

32. Muñoz PR, Resende MFR, Gezan SA, Resende MDV, de los Campos G, Kirst M, et al. Unraveling Additive from Non-Additive Effects Using Genomic Relationship Matrices. Genetics. 2014;198(4):1759–1768. Available from: http://www.genetics.org/content/early/2014/10/15/genetics.114.171322.abstract.

33. Jiang Y, Reif JC. Modeling Epistasis in Genomic Selection. Genetics. 2015;201(2):759–768. Available from: http://www.genetics.org/content/201/2/759.abstract.

34. Wood AR, Tuke MA, Nalls MA, Hernandez DG, Bandinelli S, Singleton AB, et al. Another explanation for apparent epistasis. Nature. 2014;514(7520):E3–E5.

35. The Wellcome Trust Case Control Consortium 2 G. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. Nat Genet. 2010;42(11):985–990. Available from: http://dx.doi.org/10.1038/ng.694.

36. Evans DM, Spencer CCA, Pointon JJ, Su Z, Harvey D, Kochan G, et al. Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nat Genet. 2011;43(8):761–767. Available from: http://dx.doi.org/10.1038/ng.873.

37. Brown AA, Buil A, Viñuela A, Lappalainen T, Zheng HF, Richards JB, et al. Genetic interactions affecting human gene expression identified by variance association mapping. eLife. 2014;3:e01381. Available from: https://dx.doi.org/10.7554/eLife.01381.

38. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, McRae AF, et al. Detection and replication of epistasis influencing transcription in humans. Nature. 2014;508(7495):249–253. Available from: http://dx.doi.org/10.1038/nature13005.

39. Crawford L, Zeng P, Mukherjee S, Zhou X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet. 2017;13(7):e1006869. Available from: https://doi.org/10.1371/journal.pgen.1006869.

40. 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. Nat Rev Genet. 2010;11(6):446–450. Available from: http://dx.doi.org/10.1038/nrg2809.

41. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences. 2012;109(4):1193–1198. Available from: http://www.pnas.org/content/109/4/1193.abstract.

42. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery. The American Journal of Human Genetics. 2012;90(1):7–24. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3257326/.

43. Hemani G, Knott S, Haley C. An Evolutionary Perspective on Epistasis and the Missing Heritability. PLoS Genet. 2013;9(2):e1003295. Available from: http://dx.doi.org/10.1371%2Fjournal.pgen.1003295.

44. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009;10(6):392–404. Available from: http://dx.doi.org/10.1038/nrg2579.

45. Wei WH, Hemani G, Haley CS. Detecting epistasis in human complex traits. Nat Rev Genet. 2014;15(11):722–733. Available from: http://dx.doi.org/10.1038/nrg3747.

46. Ma L, Clark AG, Keinan A. Gene-Based Testing of Interactions in Association Studies of Quantitative Traits. PLoS Genet. 2013;9(2):e1003321. Available from: http://dx.doi.org/10.1371%2Fjournal.pgen.1003321.

47. Zhang X, Huang S, Zou F, Wang W. TEAM: efficient two-locus epistasis tests in human genome-wide association study. Bioinformatics. 2010;26(12):i217–i227. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2881371/.

48. Hemani G, Theocharidis A, Wei W, Haley C. EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. Bioinformatics. 2011;27(11):1462–1465.

49. Lippert C, Listgarten J, Davidson RI, Baxter J, Poon H, Kadie CM, et al. An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. Sci Rep. 2013;3:1099 EP. Available from: http://dx.doi.org/10.1038/srep01099.

50. Prabhu S, Pe'er I. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. Genome Research. 2012;22(11):2230–2240. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3483552/.

51. Lewinger JP, Morrison JL, Thomas DC, Murcray CE, Conti DV, Li D, et al. Efficient two-step testing of gene-gene interactions in genome-wide association studies. Genet Epidemiol. 2013;37(5):440–451.

52. Ueki M, Cordell HJ. Improved Statistics for Genome-Wide Interaction Analysis. PLoS Genet. 2012;8(4):e1002625. Available from: http://dx.doi.org/10.1371%2Fjournal.pgen.1002625.

53. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. Nat Genet. 2007;39(9):1167–1173. Available from: http://dx.doi.org/10.1038/ng2110.

54. Tang W, Wu X, Jiang R, Li Y. Epistatic Module Detection for Case-Control Studies: A Bayesian Model with a Gibbs Sampling Strategy. PLoS Genet. 2009;5(5):e1000464. Available from: http://dx.doi.org/10.1371%2Fjournal.pgen.1000464.

55. Zhang Y, Zhang J, Liu JS. Block-based bayesian epistasis association mapping with application to WTCCC type 1 diabetes data. The Annals of Applied Statistics. 2011;5(3):2052.

56. 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. Am J Hum Genet. 2010;87(3):325–340.

57. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011;88(3):294–305. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059431/.

58. Golan D, Lander ES, Rosset S. Measuring missing heritability: Inferring the contribution of common variants. Proc Natl Acad Sci USA. 2014;111(49):E5272–E5281. Available from: http://www.pnas.org/content/111/49/E5272.abstract.

59. Weissbrod O, Lippert C, Geiger D, Heckerman D. Accurate liability estimation improves power in ascertained case-control studies. Nat Meth. 2015;12:332–334. Available from: http://dx.doi.org/10.1038/nmeth.3285.

60. Zhou X. A unified framework for variance component estimation with summary statistics in genome-wide association studies. Ann Appl Stat. 2017;11(4):2027–2051. Available from: https://projecteuclid.org:443/euclid.aoas/1514430276.

61. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–678. Available from: http://dx.doi.org/10.1038/nature05911.

62. Tempelman R, Gianola D. Marginal maximum likelihood estimation of variance components in Poisson mixed models using Laplacian integration. Genet Sel Evol. 1993;25(4):305–319. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2710342/.

63. Engel B, Buist W, Visscher A. Inference for threshold models with variance components from the generalized linear mixed model perspective. Genet Sel Evol. 1995;27(1):15–32. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2708264/.

64. Williams CK, Barber D. Bayesian classification with Gaussian processes. IEEE Trans Pattern Anal Mach Intell. 1998;20(12):1342–1351.

65. Kuss M, Rasmussen CE. Assessing approximate inference for binary Gaussian process classification. J Mach Learn Res. 2005;6(10):1679–1704.

66. Zaitlen N, Lindström S, Pasaniuc B, Cornelis M, Genovese G, Pollack S, et al. Informed conditioning on clinical covariates increases power in case-control association studies. PLoS Genet. 2012;8(11):e1003032. Available from: https://doi.org/10.1371/journal.pgen.1003032.

67. Hayeck TJ, Zaitlen NA, Loh PR, Vilhjalmsson B, Pollack S, Gusev A, et al. Mixed model with correction for case-control ascertainment increases association power. Am J Hum Genet. 2015;96(5):720–730. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4570278/.

68. Hayeck TJ, Loh PR, Pollack S, Gusev A, Patterson N, Zaitlen NA, et al. Mixed Model Association with Family-Biased Case-Control Ascertainment. American Journal of Human Genetics. 2017;100(1):31–39. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5223022/.

69. Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet. 1972;2(1):3–19. Available from: http://dx.doi.org/10.1007/BF01066731.

70. Davies RB. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. J R Stat Soc Ser C Appl Stat. 1980;29(3):323–333. Available from: http://www.jstor.org/stable/2346911.

71. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82–93. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3135811/.

72. Schweiger R, Weissbrod O, Rahmani E, Müller-Nurasyid M, Kunze S, Gieger C, et al. RL-SKAT: An exact and efficient score test for heritability and set tests. Genetics. 2017;207(4):1275–1283. Available from: http://www.genetics.org/content/207/4/1275.abstract.

73. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997–1004.

74. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 2004;55(4):997–1004. Available from: https://doi.org/10.1111/j.0006-341X.1999.00997.x.

75. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006;7:781–791. Available from: http://dx.doi.org/10.1038/nrg1916.

76. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet. 2011;19(7):807–812. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137506/.

77. Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol. 2013;37(2):196–204. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3642218/.

78. Kuonen D. Saddlepoint approximations for distributions of quadratic forms in normal variables. Biometrika. 1999;86(4):929–935. Available from: http://www.jstor.org/stable/2673596.

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

80. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–354. Available from: http://dx.doi.org/10.1038/ng.548.

81. Zhou X, Stephens M. Genome-wide efficient mixed model analysis for association studies. Nat Genet. 2012;44(7):821–824. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3386377/.

82. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.

83. 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.

84. 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.

85. 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.

86. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–575. Available from: //www.sciencedirect.com/science/article/pii/S0002929707613524.

87. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190. Available from: https://doi.org/10.1371/journal.pgen.0020190.

88. Zeng P, Zhou X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nature Communications. 2017;8(1):456. Available from: https://doi.org/10.1038/s41467-017-00470-2.

89. 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.

90. 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.

91. 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.

92. 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.

93. 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/.

94. 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.

95. 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.

96. Arloth J, Bogdan R, Weber P, Frishman G, Menke A, Wagner KV, et al. Genetic differences in the immediate transcriptome response to stress predict risk-related brain function and psychiatric disorders. Neuron. 2015;86(5):1189–1202.

97. Narayanan B, Soh P, Calhoun VD, Ruaño G, Kocherla M, Windemuth A, et al. Multivariate genetic determinants of EEG oscillations in schizophrenia and psychotic bipolar disorder from the BSNIP study. Transl Psychiatry. 2015;5(6):e588–. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4490286/.

98. Ren Y, Yang B, Yin Y, Leng X, Jiang Y, Zhang L, et al. Aberrant CD200/CD200R1 expression and its potential role in Th17 cell differentiation, chemotaxis and osteoclastogenesis in rheumatoid arthritis. Rheumatology. 2015;54(4):712–721. Available from: http://dx.doi.org/10.1093/rheumatology/keu362.

99. Jenmalm MC, Cherwinski H, Bowman EP, Phillips JH, Sedgwick JD. Regulation of Myeloid Cell Function through the CD200 Receptor. The Journal of Immunology. 2006;176(1):191. Available from: http://www.jimmunol.org/content/176/1/191.abstract.

100. Gorczynski RM, Chen Z, Yu K, Hu J. CD200 Immunoadhesin Suppresses Collagen-Induced Arthritis in Mice. Clinical Immunology. 2001;101(3):328–334. Available from: http://www.sciencedirect.com/science/article/pii/S1521661601951174.

101. Chakera A, Bennett SC, Morteau O, Bowness P, Luqmani RA, Cornall RJ. The phenotype of circulating follicular-helper T cells in patients with rheumatoid arthritis defines CD200 as a potential therapeutic target. Clin Dev Immunol. 2012;2012:948218.

102. Friese RS, Mahboubi P, Mahapatra NR, Mahata SK, Schork NJ, Schmid-Schönbein GW, et al. Common Genetic Mechanisms of Blood Pressure Elevation in Two Independent Rodent Models of Human Essential Hypertension. Am J Hypertens. 2005;18(5):633–652. Available from: http://dx.doi.org/10.1016/j.amjhyper.2004.11.037.

103. Guo D, Gu J, Jiang H, Ahmed A, Zhang Z, Gu Y. Inhibition of pyruvate kinase M2 by reactive oxygen species contributes to the development of pulmonary arterial hypertension. J Mol Cell Cardiol. 2016;91:179–187.

104. Alba A, Puertas MC, Carrillo J, Planas R, Ampudia R, Pastor X, et al. IFN beta accelerates autoimmune type 1 diabetes in nonobese diabetic mice and breaks the tolerance to beta cells in nondiabetes-prone mice. J Immunol. 2004;173(11):6667–6675.

105. Dou Y, Yim HC, Kirkwood CD, Williams BR, Sadler AJ. The innate immune receptor MDA5 limits rotavirus infection but promotes cell death and pancreatic inflammation. EMBO J. 2017;36(18):2742–2757.

106. Oshima M, Knoch KP, Diedisheim M, Petzold A, Cattan P, Bugliani M, et al. Virus-like infection induces human cell dedifferentiation. JCI Insight. 2018;3(3):e97732. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5821176/.

107. Schmid H, Boucherot A, Yasuda Y, Henger A, Brunner B, Eichinger F, et al. Modular activation of nuclear factor-kappaB transcriptional programs in human diabetic nephropathy. Diabetes. 2006;55(11):2993–3003.

108. 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. Available from: https://doi.org/10.1186/1471-2156-14-7.

109. Ning C, Wang D, Kang H, Mrode R, Zhou L, Xu S, et al. A rapid epistatic mixed-model association analysis by linear retransformations of genomic estimated values. Bioinformatics. 2018;34(11):1817–1825. Available from: http://dx.doi.org/10.1093/bioinformatics/bty017.

110. Fernando MMA, Stevens CR, Walsh EC, De Jager PL, Goyette P, Plenge RM, et al. Defining the Role of the MHC in Autoimmunity: A Review and Pooled Analysis. PLoS Genetics. 2008;4(4):e1000024. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2291482/.

111. Mosaad YM. Clinical Role of Human Leukocyte Antigen in Health and Disease. Scandinavian Journal of Immunology. 2015;82(4):283–306. Available from: http://dx.doi.org/10.1111/sji.12329.

112. Prentice RL, Pyke R. Logistic Disease Incidence Models and Case-Control Studies. Biometrika. 1979;66(3):403–411. Available from: http://www.jstor.org/stable/2335158.

113. Mukherjee B, Sinha S, Ghosh M. Bayesian analysis of case-control studies. Handbook of Statistics. 2005;25:793–819.

114. Staicu AM. On the equivalence of prospective and retrospective likelihood methods in case-control studies. Biometrika. 2010;97(4):990–996.

115. Zhong S, Jiang D, McPeek MS. CERAMIC: Case-Control Association Testing in Samples with Related Individuals, Based on Retrospective Mixed Model Analysis with Adjustment for Covariates. PLOS Genetics. 2016;12(10):e1006329–. Available from: https://doi.org/10.1371/journal.pgen.1006329.

116. 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.

117. Breslow NE, Lin X. Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika. 1995;82(1):81–91. Available from: http://dx.doi.org/10.1093/biomet/82.1.81.

118. Lin X, Breslow NE. Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion. Journal of the American Statistical Association. 1996;91(435):1007–1016. Available from: http://www.jstor.org/stable/2291720.

119. Lin X. Variance component testing in generalised linear models with random effects. Biometrika. 1997;84(2):309–326. Available from: http://biomet.oxfordjournals.org/content/84/2/309.abstract.

120. 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. Available from: http://dx.doi.org/10.1016/j.ajhg.2016.02.012.

121. Rönnegård L, Pong-Wong R, Carlborg Ö. Defining the Assumptions Underlying Modeling of Epistatic QTL Using Variance Component Methods. Journal of Heredity. 2008;99(4):421–425. Available from: http://jhered.oxfordjournals.org/content/99/4/421.abstract.

122. Young AI, Durbin R. Estimation of Epistatic Variance Components and Heritability in Founder Populations and Crosses. Genetics. 2014;198(4):1405–1416. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4256760/.

123. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565–569. Available from: http://dx.doi.org/10.1038/ng.608.

124. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Meth. 2014;11(4):407–409. Available from: http://dx.doi.org/10.1038/nmeth.2848.

125. Wei WH, Bowes J, Plant D, Viatte S, Yarwood A, Massey J, et al. Major histocompatibility complex harbors widespread genotypic variability of non-additive risk of rheumatoid arthritis including epistasis. Scientific Reports. 2016;6:25014 EP –. Available from: http://dx.doi.org/10.1038/srep25014.

126. Niel C, Sinoquet C, Dina C, Rocheleau G. SMMB: A stochastic Markov blanket framework strategy for epistasis detection in GWAS. Bioinformatics. 2018;p. bty154–bty154. Available from: http://dx.doi.org/10.1093/bioinformatics/bty154.

127. Nejentsev S, Howson JMM, Walker NM, Szeszko J, Field SF, Stevens HE, et al. Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. Nature. 2007;450(7171):887–892. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2703779/.

128. Todd JA. Etiology of type 1 diabetes. Immunity. 2010;32(4):457–467.

129. Thameem F, DiStefano J, Wang J, S German M, Bogardus C, Prochazka M. Cloning, expression and genomic structure of human LMX1A, and variant screening in Pima Indians. Gene. 2002;290:217–25.

130. Wongseree W, Assawamakin A, Piroonratana T, Sinsomros S, Limwongse C, Chaiyaratana N. Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses. BMC Bioinformatics. 2009;10:294.