A Likelihood Ratio-Based Mann-Whitney Approach Finds Novel Replicable Joint Gene Action for Type 2 Diabetes

Qing Lu,[1] Changshuai Wei,[1] Chengyin Ye,[1] Ming Li,[1] and Robert C. Elston[2]*

[1]*Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan*
[2]*Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio*

The potential importance of the joint action of genes, whether modeled with or without a statistical interaction term, has long been recognized. However, identifying such action has been a great challenge, especially when millions of genetic markers are involved. We propose a likelihood ratio-based Mann-Whitney test to search for joint gene action either among candidate genes or genome-wide. It extends the traditional univariate Mann-Whitney test to assess the joint association of genotypes at multiple loci with disease, allowing for high-order statistical interactions. Because only one overall significance test is conducted for the entire analysis, it avoids the issue of multiple testing. Moreover, the approach adopts a computationally efficient algorithm, making a genome-wide search feasible in a reasonable amount of time on a high performance personal computer. We evaluated the approach using both theoretical and real data. By applying the approach to 40 type 2 diabetes (T2D) susceptibility single-nucleotide polymorphisms (SNPs), we identified a four-locus model strongly associated with T2D in the Wellcome Trust (WT) study (permutation P-value < 0.001), and replicated the same finding in the Nurses' Health Study/Health Professionals Follow-Up Study (NHS/HPFS) (P-value $= 3.03 \times 10^{-11}$). We also conducted a genome-wide search on 385,598 SNPs in the WT study. The analysis took approximately 55 hr on a personal computer, identifying the same first two loci, but overall a different set of four SNPs, jointly associated with T2D (P-value $= 1.29 \times 10^{-5}$). The nominal significance of this same association reached 4.01×10^{-6} in the NHS/HPFS. *Genet. Epidemiol.* 36:583–593, 2012.

Key words: gene-gene interaction; genome-wide search; forward selection

*Correspondence to: Robert C. Elston, Department of Epidemiology and Biostatistics, Case Western Reserve University, Wolstein Research Building, 2103 Cornell Road, Cleveland, OH 44106. E-mail: robert.elston@cwru.edu
Received 12 January 2012; Revised 9 April 2012; Accepted 9 May 2012
Published online 3 July 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21651

INTRODUCTION

With high-throughput genotyping technology and ever-decreasing genotyping cost, the analysis of genome-wide association studies (GWAS) has become commonplace [Hardy and Singleton, 2009; Hirschhorn and Daly, 2005]. To date, hundreds of GWAS have been conducted and significant progress has been made in discovering previously unknown genetic variants predisposing to common complex diseases [Goldstein, 2009]. The findings from GWAS are expected to advance our knowledge of the genetic causes of diseases, and this may lead to more effective methods for prevention or treatment [Hirschhorn, 2009]. Despite these achievements, for most if not all of the common complex diseases, a large proportion of the genetic variants contributing to them still remains undiscovered [Manolio et al., 2009]. Gene-gene interactions may play important roles in the biological pathways causing common complex diseases. However, they have not been taken fully into account by many GWAS, which have used a single locus approach. The identification of such interactions should help elucidate how genetic variants interplay with each other within biological pathways to cause disease [Cordell 2009; Eichler et al., 2010; Maher, 2008].

Recognizing this, recent studies have been conducted to evaluate the joint action among existing genetic variants [Culverhouse et al., 2011; Kirchhoff et al., 2008; Liu et al., 2011], whether with or without considering a statistical interaction term. Most of these studies have focused on evaluating two-way joint action among a limited number of loci. In reality, we are interested in discovering the joint action of as many genes as possible [Wang et al., 2010]. By not limiting the search to two-way joint action, we could have a better chance of identifying important high-order interactions. Moreover, strong interactions may exist among risk loci with low to intermediate marginal effects [Cordell, 2009; Wan et al., 2010; Yung et al., 2011]. Evaluating joint gene action among only previously recognized genetic variants can limit the discovery process with respect to finding interactions among loci. Therefore, for the purpose of discovering novel findings, it would be better to look beyond existing knowledge and conduct a genome-wide study by searching for joint action among all available genetic variants [Cordell, 2009].

It could be statistically complicated and computationally time consuming to evaluate the joint gene action among 500K or 1 M (single-nucleotide polymorphisms [SNPs]). Numerous statistical approaches have been developed for a joint gene action analysis, such as MDR [Ritchie et al., 2001],

BEAM [Zhang and Liu, 2007], PLINK [Purcell et al., 2007], Random Jungle [Schwarz et al., 2010], BOOST [Wan et al., 2010], and Kernel Machine [Wu et al., 2010]. Only BOOST, PLINK, and Random Jungle have been shown to have the capacity to perform genome-wide analysis. BOOST evaluates all pair-wise gene-gene interactions across the genome via a regression-based approach, and is more computationally efficient than PLINK. However, similar to other pair-wise analysis tools such as PLINK, BOOST has reduced power when high-order interactions are present. Performing a large number of tests on the genome-wide scale raises a serious multiple testing issue, requiring a very stringent threshold (e.g., 4×10^{-13}) to reach a genome-wide significance level [Cordell, 2009]. Pair-wise analysis using a regression-based approach could also be subject to a multicollinearity problem [Yung et al., 2011]). Unlike pair-wise analysis approaches, data-mining approaches such as Random Jungle are capable of capturing high-order interactions without suffering from this multicollinearity issue. Random Jungle, based on random forests [Breiman, 1996], builds an ensemble of decision trees for the purpose of classification. It runs much faster than the original random forest packages, making it more appealing for genome-wide analysis. However, similar to them, it was designed for the purpose of classification and lacks an asymptotic test statistic to assess the significance level of any particular joint gene action.

Targeting these issues, we propose a likelihood ratio (LR) based Mann-Whitney (LRMW) test to find genome-wide joint gene action. It extends the traditional univariate Mann-Whitney test [Mann and Whitney, 1947; Wilcoxon, 1945] to deal with multiple genetic variables, which facilitates a general multilocus association test that allows for high-order interactions. It conducts one overall significance test for the entire genome-wide analysis, and thus has the advantage of not being subject to any multiple testing issue. Moreover, the approach adopts a computationally efficient algorithm that makes it feasible to perform on a genome-wide scale, without being subject to any multicollinearity issue. We performed a simulation study to validate the method and then applied it to candidate SNPs in the Wellcome Trust (WT) and Nurses' Health Study/Health Professionals Follow-Up (NHS/HPFS) type 2 diabetes (T2D) GWAS data, replicating the same four-locus model. Finally, we conducted a genome-wide search in the WT data, with approximately 386K SNPs [Wellcome Trust Case Control Consortium, 2007]. Using a Dell workstation equipped with two 2.5 GHz quad-core processors and a 4GB memory, the genome-wide analysis took approximately 55 hr to identify the most parsimonious model and estimate its nominal significance level. This finding was then replicated using data from the NHS/HPFS [Cornelis et al., 2009].

METHODS

Current GWAS for common complex diseases involve up to a million or more SNPs. Searching such high-dimensional data requires a powerful and computationally efficient statistical tool. We introduce here a LRMW method to test the joint action of multiple loci allowing for high-order interaction, incorporating a forward selection algorithm into the method to efficiently search the entire genome. Both a permutation and a kernel density method [Parzen, 1962; Rosenblatt, 1956] (the latter in much less time) can be applied to assess the empirical P-value of the identified model, adjusting for any inflated Type I error due to model selection.

LRMW TEST

Suppose we are interested in evaluating the joint action of p disease-susceptibility loci, comprising M possible p-locus genotypes. If we knew the underlying mode of inheritance, we could cluster all the M p-locus genotypes into R risk groups ($R \leq M$). We define a risk group as a cluster of p-locus genotypes that all have the same risk predisposing to disease. For instance, if the ith locus follows a dominant model, we could cluster the p-locus genotypes containing either one or two copies of the risk alleles at locus i and the same genotypes at the remaining $p-1$ loci into one group. Similarly, we could group the p-locus genotypes according to a particular statistical interaction model. For example, if two of p loci follow what was first called the action of either of two dominant genes [Elston, 1981; Tiwari and Elston, 1998], but more recently the "threshold" model [Marchini et al., 2005], we would cluster into one group the p-locus genotypes having at least one of the disease-susceptibility alleles at each of the two loci, but the same genotype at the remaining $p-2$ loci, and into another group those containing no risk allele at the two loci, but the same genotype at the rest of the loci. By clustering the p-locus genotypes into different risk groups according to this underlying model and evaluating their difference in risk, we take the known mode of inheritance into account.

Given R genomically defined risk groups, we rank them according to their LR. The LR is defined as $LR(G) = \frac{P(G|D)}{P(G|\bar{D})}$), where G denotes a particular risk group, D represents cases, and $\bar{D}$ represents controls. The LRMW statistic can then be estimated as

$$U_{LRMW} = \sum_{i=1}^{R} \sum_{j=1}^{R} n_{G_i}^{D} n_{G_j}^{\bar{D}} \psi [LR(G_i), LR(G_j)], \qquad (1)$$

where $n_{G_i}^{D}$ ($n_{G_j}^{\bar{D}}$) denotes the number of persons carrying risk group G_i (G_j) in cases (controls), and ψ is a kernel function. When the kernel function has the following form,

$$\psi [LR(G_i), LR(G_j)] = \begin{cases} 1 & \text{if } LR(G_j) < LR(G_i) \\ 0.5 & \text{if } LR(G_j) = LR(G_i) \\ 0 & \text{if } LR(G_j) > LR(G_i) \end{cases}, \qquad (2)$$

U_{LRMW} is equivalent to a Mann-Whitney statistic comparing the difference in LR risk scores between cases and controls. This link provides us with a simple way to test for the joint action of p loci, because we can then apply the known properties of the univariate Mann-Whitney test [Mann and Whitney, 1947; Wilcoxon, 1945]. Note that a variety of link functions (e.g., a logit link) can be used to map the p-locus genotypes into a one-dimensional risk score. However, as we have shown elsewhere, among all the link functions, the LR has optimal properties [Lu and Elston, 2008].

Under the null hypothesis that there is no association between case-control status and the genotypes carried, U_{LRMW}

equals $\frac{N_D N_{\bar{D}}}{2}$, and so we can form the test statistic

$$Z = (U_{LRMW} - \frac{N_D N_{\bar{D}}}{2})/\sqrt{\text{Var}(U_{LRMW})}, \quad (3)$$

where N_D and $N_{\bar{D}}$ are, respectively, the total numbers of cases and controls. For a large sample size and under the null, Z follows a standard normal distribution. The variance of U_{LRMW} can be estimated using the result of Delong et al. [1988],

$$\text{Var}(U) = S_D + S_{\bar{D}}, \quad (4)$$

where $\begin{cases} S_D = \sum_{i=1}^{R} n_{G_i}^{D} \left\{ \sum_{j=1}^{R} n_{G_j}^{\bar{D}} \psi\left[LR(G_i), LR(G_j)\right] - \frac{U_{LRMW}}{N_D} \right\}^2 \\ S_{\bar{D}} = \sum_{j=1}^{R} n_{G_j}^{\bar{D}} \left\{ \sum_{i=1}^{R} n_{G_i}^{D} \psi\left[LR(G_i), LR(G_j)\right] - \frac{U_{LRMW}}{N_{\bar{D}}} \right\}^2 \end{cases}$,

where $n_{G_i}^{D}$, $n_{G_j}^{D}$, $n_{G_i}^{\bar{D}}$, and $n_{G_j}^{\bar{D}}$ are the numbers of individuals, respectively, in risk groups G_i and G_j among cases and controls.

By comparing all p-locus risk groups, the above test statistic assesses the overall significance of all p loci associated with the disease, taking into account their interactions. If the test is significant, we might be further interested in evaluating the difference between any two risk groups. The odds ratio comparing two risk groups is commonly used for such a purpose, and can be easily obtained using the LR values:

$$OR_{ij} = \frac{LR(G_i)}{LR(G_j)}, \quad (5)$$

where $LR(G_i)$ and $LR(G_j)$ are the LR values of risk groups G_i and G_j, respectively. A $(1-\alpha)$ confidence interval (CI) for the Odds Ratio (OR) can be estimated based on Woolf's method [Woolf, 1955],

$$OR_{ij} \cdot e^{\pm z_{1-\alpha/2} \cdot \sqrt{\frac{1}{n_{G_i}^{D}} + \frac{1}{n_{G_j}^{D}} + \frac{1}{n_{G_i}^{\bar{D}}} + \frac{1}{n_{G_j}^{\bar{D}}}}}, \quad (6)$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ fractile of the standard normal distribution.

MANN-WHITNEY-BASED FORWARD SELECTION ALGORITHM

In a genetic study of a common complex disease, the genetic model is usually not known and most of the genetic markers studied are probably not disease related, which makes it difficult to determine the underlying risk groups. Simply clustering all the individuals into just two risk groups could lead to low power of the approach, while treating each p-locus genotype as a separate risk group will lead to inflated Type 1 error [Lu et al., 2010]. Based on the idea of forward selection [Li et al., 2011; Ye et al., 2011], we introduce a computationally feasible and efficient forward selection algorithm to cluster individuals into an optimal number of risk groups, i.e., the number that best represents the underlying risks. We assume diallelic loci, though in principle (but not with the same ease of computation), the method can be extended to any number of alleles. In the first step of the algorithm, we select a single locus and divide all individuals into two risk groups according to their genotypes at that locus. For instance, if the ith locus, which we call locus A, has genotypes AA, Aa, and aa, we consider three possible models to divide the individuals into two groups: (i) individuals carrying the AA or Aa genotype vs. individuals carrying the aa genotype (i.e., $\{AA/Aa\}$, $\{aa\}$), (ii) individuals carrying the Aa or aa genotype vs. individuals carrying the AA genotype (i.e., $\{Aa/aa\}$, $\{AA\}$), and (iii) individuals carrying the AA or aa genotype vs. individuals carrying the Aa genotype (i.e., $\{AA/aa\}$, $\{Aa\}$). For each case, we estimate the U_{LRMW} statistic using Equation 1. The process is repeated for each locus in turn and the model with the highest U_{LRMW} statistic is chosen as the best model at step one. Once we have found the best two risk groups at step one, for step two, we further divide these two specific risk groups. Given locus j has three genotypes $\{BB, Bb, bb\}$, we divide the two risk groups already found into four risk groups by again considering three possibilities. Suppose, for example, model (i) ($\{AA/Aa\}$, $\{aa\}$) was chosen as the best model in step one; then the three new possibilities would be:

(1) $\{AA/Aa, BB/Bb\}$; $\{AA/Aa, bb\}$; $\{aa, BB/Bb\}$; $\{aa, bb\}$

(2) $\{AA/Aa, BB/bb\}$; $\{AA/Aa, Bb\}$; $\{aa, BB/bb\}$; $\{aa, Bb\}$, and

(3) $\{AA/Aa, Bb/bb\}$; $\{AA/Aa, BB\}$; $\{aa, Bb/bb\}$; $\{aa, BB\}$.

We estimate the U_{LRMW} statistic for each of these three situations (or analogous situations had we chosen model (ii) or (iii) at step one), and repeat the process for all the $p-1$ loci other than locus A. Similarly, we choose the model with the highest U_{LRMW} statistic as the best model at step two. Since the step-two model is obtained by further dividing the risk groups found in the step-one model, it is easy to show that the U_{LRMW} statistic obtained at step two is greater than the one at step one. We continue dividing the individuals in further steps and, in principle, we could continue until a full model is reached (i.e., the samples cannot be further split). As we divide the samples, the value of U_{LRMW} increases, as well as the model complexity as measured by the number of risk groups. However, after a certain number of steps, the number of risk groups starts to over-represent the underlying true number of risk groups and the model tends to over-fit the data. To identify the point at which this occurs, and hence the best model with an optimal number of risk groups, K-fold cross-validation is used. The data are first randomly divided into K subsets. Then the selection algorithm is performed on $K-1$ of the subsets, used as the training dataset, to obtain a series of models for each step. Each of these models is then applied on the remaining subset, used as a validation dataset, to calculate U_{LRMW}. This is done by assigning to the validation dataset the LR values of the risk groups estimated in the training dataset. We repeat this process K times, with each of the K subsets used exactly once as the validation dataset, then average the K U_{LRMW} statistics, one from each validation dataset. The final model with an optimal number of risk groups is chosen to be the one that leads to the highest average U_{LRMW} statistic over the validation datasets. Note that we stop as soon as the highest statistic averaged over the validation datasets no longer increases.

PERMUTATION AND KERNEL DENSITY ESTIMATION PROCEDURES

Given the R risk groups determined by the final model, we can use Equation (3) to carry out an LRMW test, assessing the joint action of the multiple risk loci on the disease. However, selecting the model and performing the asymptotic test on the same dataset could lead to an inflated Type 1 error. For an association study with a limited number of loci, we can control the Type 1 error by using a permutation test. We generate a certain number of permutation replicates (e.g., 1,000), in which the disease status of each individual has been randomly permuted. On each of the permutation replicates, we apply the forward selection algorithm to choose the best model and calculate Z. By repeating this for each permutation replicate, the empirical null distribution of Z is generated and the empirical P-value for the association test is obtained by comparing the final observed statistic Z^O to this null distribution, so that we have P-value $= \mathrm{P}\,(Z \leq Z^O)$.

Permutation is a computationally intensive procedure, especially when it involves complicated modeling such as the forward selection procedure just described. One thousand permutation replicates are commonly required to assess the test's significance, and even more permutation replicates would be required if the association is highly significant (e.g., for a significance level 10^{-8}). This could be time consuming and even infeasible for a genome-wide study of joint gene action. To reduce the computation time and make it feasible for a genome-wide search, we use a kernel density estimation (KDE) procedure, which is a nonparametric way to estimate the probability density function of a distribution [Parzen, 1962; Rosenblatt, 1956]. We draw a certain number of permutation replicates and for each we estimate Z. By treating these values of Z as a random sample drawn from the null distribution of Z, we can use the KDE procedure to estimate the null distribution:

$$\hat{f}(Z) = \frac{1}{N\lambda}\sum_{i=1}^{N} K_\lambda(Z, Z_i), \tag{7}$$

where Z_i is Z calculated from the ith permutation replicate, K_λ is a kernel function—the most commonly used form is the Gaussian Kernel, $K_\lambda(Z, Z_i) = \phi(|Z - Z_i|/\lambda)$, where ϕ is the standard normal density function, N is the number of permutation replicates, and λ is the bandwidth. For our genome-wide study, we applied the forward selection algorithm to 100 permutation replicate samples to estimate the distribution of Z, using a Gaussian kernel with the standard deviation as bandwidth.

RESULTS

PAIR-WISE JOINT ACTION ANALYSIS VS. MULTILOCUS JOINT ACTION ANALYSIS

Given the underlying disease model, we calculated the U_{LRMW} statistic and the theoretical power of the test at the significance level 0.05. In the pair-wise analysis, power was calculated for the most "important" pair of loci (i.e., the pair associated with the highest test statistic), while in the multilocus analysis, power was obtained based on all disease loci.

Four sets of hypothesized disease models were considered, varying from simple models including only one two-way statistical interaction, to more complicated models containing a higher-order statistical interaction. Details of the settings considered are described in Table I (Settings I). The first set of models comprised only one two-way statistical interaction. The interaction in the model was assumed to follow a multiplicative model (i.e., the odds increases multiplicatively within and among loci), an interaction-multiplicative model (i.e., the odds increases multiplicatively with the number of disease-susceptibility alleles given both loci have at least one disease-susceptibility allele), or a threshold model (i.e., dominant disease-susceptibility alleles at both loci). As expected, for all three interaction models, we observed that power increases when the risk allele frequencies of the interacting loci increase. Note that, when there is only one two-way interaction in the disease model, the theoretical power of the pair-wise analysis is equivalent to that of the multilocus analysis (Figure 1).

We then considered a scenario where multiple gene interactions exist. For simplicity, we included just two two-way interactions in the second set of models and assumed the two interactions were the same. Similarly, we included in a third set of models a two-way interaction and two independent risk loci to evaluate the test performance when both interaction and independent risk loci are present. In a fourth set of models, we studied the new approach in the presence of higher order interactions by introducing a three-way interaction. In these three further sets of models, we also observed that power increased when the risk allele frequency increases. More importantly, significant power increase was obtained from the multilocus analysis as compared to the pair-wise analysis (Figure 1).

PERMUTATION VS. KDE

A small simulation experiment was conducted to compare the performance of the permutation and KDE procedures for the LRMW approach under various scenarios. The disease model was simulated to contain one two-way interaction and two independently acting risk loci. The interaction model was assumed to be a locus-multiplicative model, multiplicative-interaction model, or a threshold model, and the effect size was varied in the simulations. The details of the simulation settings are described in Table I (Settings II). One thousand permutation replicates were used to estimate the significance level by the permutation test. For the KDE procedure, we fitted the probability density function of the null distribution on 100 permutation replicates to estimate the null distribution. Type I error from both procedures was well controlled at the 0.05 significance level (Table II). For all three interaction models, we observed a similar performance of the two procedures, indicating that KDE based on 100 permutation replicates well approximates the 1,000 replicate permutations. We also evaluated the permutation and KDE procedures in a disease scenario where more than one interaction or a high-order interaction is present (Table 1). The results were consistent with the findings from disease models with only one interaction, confirming that the two procedures of assessing the null distribution give similar results (Table II).

TABLE I. Summary of the simulation settings for four risk loci: *A*, *B*, *C*, and *D*

		Gene-gene interaction			Each single locus		
		Interaction	RR[a]	MAF[b]	Risk locus	RR[c]	MAF
Settings I	Two-way interaction only[d]	$A\times B$	1.5	[0.05, 0.5][f]			
	Two two-way interactions[d]	$A\times B$; $C\times D$	1.5; 1.5	[0.05, 0.5]			
	Two-way interaction[c] + two loci [e]	$A\times B$	1.5	[0.05, 0.5]	*C*; *D*	1.4; 1.3	0.2; 0.3
	Three-way interaction[d]	$A\times B\times C$	1.5	[0.05, 0.5]			
Settings II	Two-way interaction[c] + two loci[e]	$A\times B$	1; 1.2; 1.5; 2	0.25	*C*; *D*	1.4; 1.3	0.2; 0.3
	Two two-way interactions[d]	$A\times B$; $C\times D$	1.5	0.25			
	Three-way interaction [c] + one locus [e]	$A\times B\times C$	1.5	0.25	*D*	1.3	0.3

[a]Relative risk of a high-risk group(s) (i.e., a multilocus genotype combination) to a low-risk group(s).
[b]Minor allele frequency.
[c]Relative risk of a high-risk genotype(s) to a low-risk genotype(s).
[d]The interaction was assumed to follow a locus-multiplicative model, a multiplicative-interaction model, or a threshold model.
[e]Risk loci *C* and *D* follow recessive and additive modes of inheritance, respectively.
[f]Minor allele frequency simulated in the models ranged from 0.05 to 0.5.

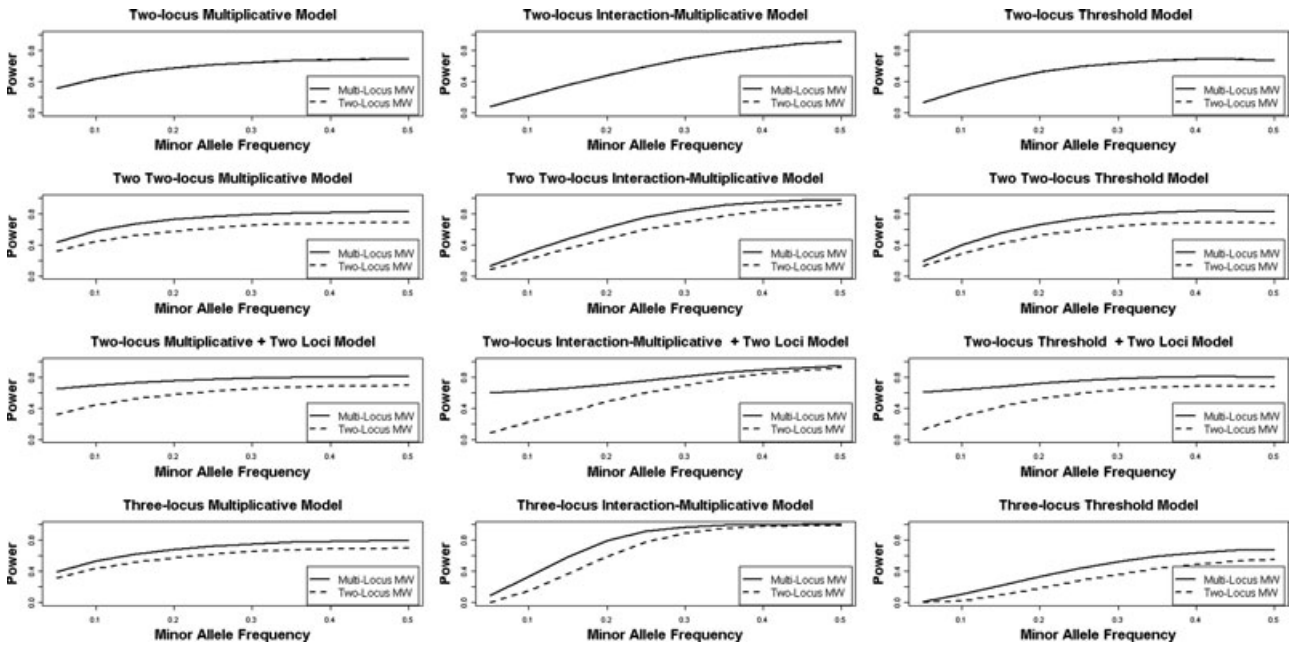

Fig. 1. Theoretical power of a two-locus LRMW test vs. a multilocus LRMW test.

APPLICATION TO T2D

T2D is a chronic disease that is believed to be caused by the interplay of multiple genetic and environmental risk factors [Frayling, 2007]. *PPARG* [Altshuler et al., 2000], *KCNJ11* [Gloyn et al., 2003], and *TCF7L2* [Grant et al., 2006; Zeggini et al., 2007] were the first identified and well-replicated T2D genes. Current extensive genetic research, in particular the completion of multiple large-scale GWAS [Saxena et al., 2007; Scott et al., 2007; Sladek et al., 2007; Steinthorsdottir et al., 2007; Voight et al., 2010; Zeggini et al., 2007, 2008], has uncovered multiple novel sites associated with T2D. We used the data from two large-scale GWAS, the WT GWAS [Wellcome Trust Case Control Consortium, 2007] and the NHS/HPFS GWAS [Cornelis et al., 2009], to conduct joint analyses. We first applied our approach to 40 previously identified T2D SNPs, to assess their joint association with T2D. Then, to facilitate the discovery of novel interactions, we extended our analysis to nearly 386K SNPs to conduct a genome-wide search for joint gene action.

WT GWAS AND NHS/HPFS GWAS

The WT GWAS included 1,924 T2D cases and 2,938 controls that passed the WT quality control criteria [Wellcome Trust Case Control Consortium 2007]. All individuals were genotyped by using the Affymetrix 500K chip. Among the 40 T2D SNPs, 22 can be directly found in the WT T2D GWAS dataset, and the remaining 18 were imputed using the model-based imputation software IMPUTE [Marchini et al., 2007]. For the genome-wide analysis, we used the 385,598 SNPs that met WT quality control criteria and had a missing rate less than 0.01.

TABLE II. Type I error and power comparison of the permutation procedure with the KDE procedure

	Type I error		Power									
	RR = 1		RR = 1.2		RR = 1.5		RR = 2		Two two-way interactions		Three-way interactions	
Model	PERM[a]	KDE[b]	PERM	KDE	PERM	KDE	PERM	KDE	PERM	KDE	PERM	KDE
Locus-multiplicative	0.045	0.047	0.337	0.349	0.682	0.696	0.996	0.996	0.694	0.714	0.773	0.790
Interaction-multiplicative	0.040	0.032	0.225	0.157	0.649	0.644	0.970	0.970	0.806	0.798	0.804	0.809
Threshold	0.047	0.050	0.308	0.327	0.693	0.690	0.984	0.980	0.847	0.843	0.806	0.808

[a]Calculated based on 1,000 permutation replicates.
[b]Calculated using 100 permutation replicates and a Gaussian kernel with one SD bandwidth.

The NHS/HPFS GWAS dataset, genotyped using the Affymetrix 6.0 platform, served as a replication dataset and comprised 2,725 cases and 3,120 controls after excluding low-quality samples using the NHS/HPFS quality control filter. We directly selected or imputed SNPs that were significant in the initial joint gene action analysis to seek confirmation of their association with T2D.

JOINT GENE ACTION AMONG SNPS KNOWN TO BE ASSOCIATED WITH T2D

We applied the LRMW approach on 40 T2D susceptibility SNPs to evaluate their joint gene action. The new method identified a four-locus model in which the combined effect of rs4506565 (*TCF7L2*), rs8050136 (*FTO*), rs7961581 (*TSPAN8/LGR5*), and rs7756992 (*CDKAL1*) on T2D reached a nominal significance level of 8.91×10^{-28} (Table III). A permutation test was used to estimate the adjusted significance level. On the basis of 1,000 permutation replicates, the *P*-value for the joint model was less than 0.001, indicating a significant association of the four loci with T2D. We validated this association finding by applying the selected four-locus model to the NHS/PHFS dataset (i.e., by ranking the risk groups in the NHS/PHFS dataset based on their *LR* values in the WT dataset). The association of the same four-locus model reached the nominal significance level 3.03×10^{-11}, confirming its association with T2D.

Further analysis of the identified four-locus model suggested that rs8050136 (*FTO*) and rs7961581 (*TSPAN8/LGR5*) are likely independently associated with T2D (i.e., they follow a multiplicative model), while rs4506565 (*TCF7L2*) and rs7756992 (*CDKAL1*) jointly influence T2D through a protective double recessive allele model. In Figure 2, we plot the joint action model between rs4506565 (*TCF7L2*) and rs7756992 (*CDKAL1*) for each genotype combination of rs8050136 (*FTO*) and rs7961581 (*TSPAN8/LGR5*). The joint action models are consistent across the WT and NHS/PHFS: the TT-AA risk group in the two-locus model is associated with lower risk, while the other three risk groups are associated with higher risk of T2D. Figure 2 also shows that the joint action models are similar across all four strata. Analysis using Woolf's test [Woolf, 1955] confirmed homogeneity of the joint action model across all four strata in both the WT study(*P*-value = 0.16) and the NHS/PHFS (*P*-value = 0.97) and suggested that rs4506565 (*TCF7L2*) and rs7756992 (*CDKAL1*) are associated with T2D independently of rs8050136 (*FTO*) and rs7961581 (*TSPAN8/LGR5*).

Additional pair-wise joint gene action analysis using logistic regression also found an interaction-only effect between loci rs4506565 (*TCF7L2*) and rs7756992 (*CDKAL1*) (*P*-value = 0.0089) in the WT study, replicated in the NHS/PHFS (*P*-value = 0.0052). To follow up with this potential interaction, we calculated the odds ratios and corresponding CIs of the four risk groups identified (Table IV). Based on the WT data, we found the risk group TT-AA is associated with a significantly lower risk of T2D than the other three risk groups, while there is no significant difference among the other three risk groups. This suggests that the joint action model follows a double recessive model for low risk, which is same as the union of two dominantly acting genes for high risk. Further validation of the model in the NHS/HPFS found the same double recessive model. We found that the odds ratios associated with the risk groups in the NHS/HPFS were slightly lower, but not significantly different from, those estimated in the WT study.

GENOME-WIDE JOINT GENE ACTION

The above approach takes advantage of previous association findings and limits the search to a handful of known risk SNPs. Thus, it is ideal for detecting the joint action among SNPs with relatively strong single-locus effects. However, considering only currently known risk SNPs limits the discovery of novel findings. In particular, it may lack power to discover loci with low-to-medium marginal effects that could be acting jointly with other loci to cause disease. In order to explore novel joint gene actions, we conducted a genome-wide search by simultaneously analyzing all available SNPs in the WT GWAS dataset. This analysis took about 55 hr on a Dell workstation and identified a new four-locus model associated with T2D with a nominal *P*-value of 8.02×10^{-32}(Table V). The first two SNPs selected into the four-locus model were rs4506565 (*TCF7L2*) and rs7193144 (*FTO*), consistent with our previous finding among the known risk loci. In addition, the genome-wide analysis identified two new SNPs, rs8092098 and rs12508397. rs8092098 lies in the 18q21-18q23 region, which is potentially associated with fasting plasma glucose [Li et al., 2004] and diabetes-associated nephropathy [McDonough et al., 2009]; rs12508397 is located at the 3′ end of the transcribed region of *ANK2*, which encodes the adapter protein Ankyrin-B, related to the type 4 QT syndrome [Mohler et al., 2007]. Moreover, a recent study found that variants in *ANK2* both cause a loss of function

TABLE III. Stepwise result for joint gene action analysis among the known T2D loci using the LRMW approach

Steps	Selected SNPs	Wellcome Trust study		NHS/PHFS *P*-value[a]
		Nominal *P*-value[a]	Permutation *P*-value[b]	
1	rs4506565	4.77×10^{-11}		1.18×10^{-7}
2	rs4506565, rs8050136	6.18×10^{-17}		1.08×10^{-8}
3	rs4506565, rs8050136, rs7961581	5.62×10^{-22}		3.58×10^{-8}
4 [c]	rs4506565, rs8050136, rs7961581, rs7756992	8.91×10^{-28}	> 0.001	3.03×10^{-11}

[a]Calculated by using Equations 3 and 4.
[b]Estimated based on 1,000 permutation replicates.
[c]The most parsimonious model identified by the LRMW approach.

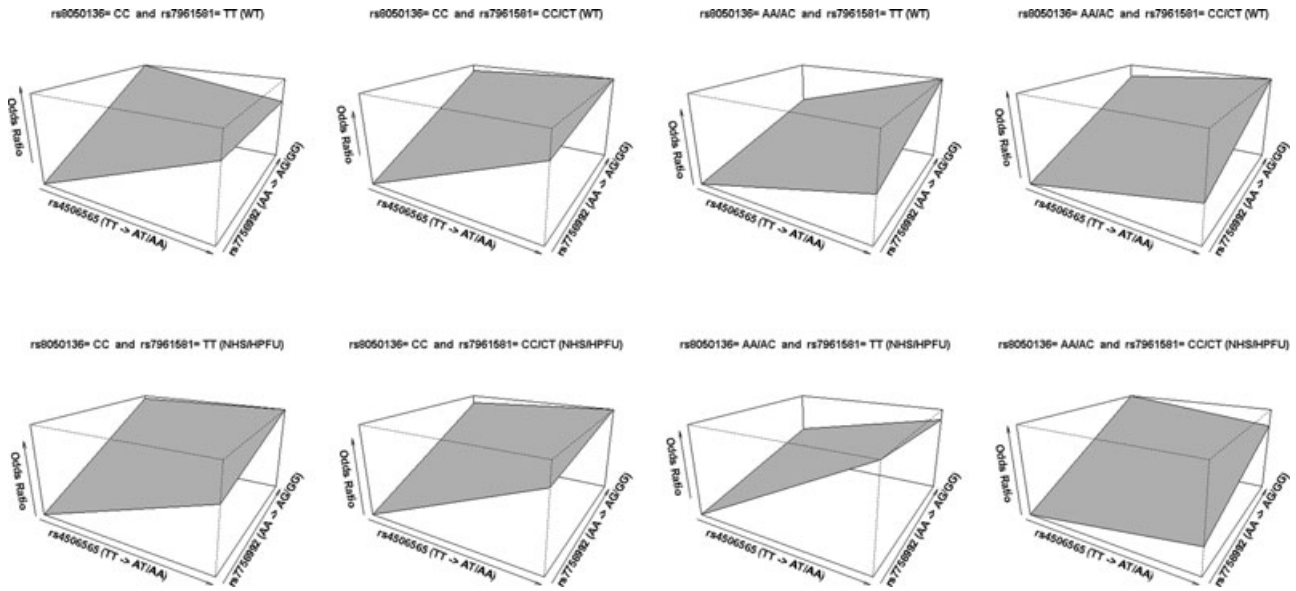

Fig. 2. The joint action between loci rs4506565 (*TCF7L2*) and rs7756992 (*CDKAL1*) at each combination of rs8050136 (*FTO*) and rs7961581 (*TSPAN8*/*LGR5*). The clustered genotypes (e.g., CC/CT) were determined by the LRMW approach. Top row: WT; bottom row: NHS/HPFS.

TABLE IV. Joint action between rs4506565 (*TCF7L2*) and rs7756992 (*CDKAL1*)

Wellcome Trust study			rs4506565		NHS/HPFS		rs4506565	
			TT	AT/AA			TT	AT/AA
Crude OR[a]	rs7756992	AA	Reference	1.75[c] [1.47, 2.08][d]	rs7756992	AA	Reference	1.52 [1.31, 1.75]
		AG/GG	1.53 [1.26, 1.86]	1.93 [1.61, 2.30]		AG/GG	1.37 [1.17, 1.60]	1.55 [1.33, 1.79]
MH OR[b]	rs7756992	AA	Reference	1.76 [1.48, 2.10]	rs7756992	AA	Reference	1.53 [1.32, 1.77]
		AG/GG	1.52 [1.25, 1.85]	1.93 [1.61, 2.31]		AG/GG	1.37 [1.17, 1.60]	1.56 [1.34, 1.81]

[a]Crude odds ratio estimated without adjusting for rs8050136 (*FTO*) and rs7961581 (*TSPAN8/LGR5*).
[b]The Mantel-Haenszel odds ratio estimated by adjusting for rs8050136 (*FTO*) and rs7961581 (*TSPAN8/LGR5*).
[c]Odds ratio.
[d]95% confidence interval.

in pancreatic islets and are associated with diabetes [Healy et al., 2010]. We simulated 100 permutation replicates and used the kernel density approach, to adjust for possible bias due to model selection in the same dataset. The adjusted *P*-value of the four-locus model was 1.29×10^{-5}. We further evaluated the selected four-locus model in the NHS/PHFS dataset and found the association reached a significance level of 4.01×10^{-6}.

Analysis showed that the joint action between rs4506565 (*TCF7L2*) and rs7193144 (*FTO*) was also associated with T2D in the NHS/PHFS (Step 2 in Table V). The joint action model was consistent across both studies and likely follows a multiplicative model (Figure 3). We also identified an interaction between rs8092098 and rs12508397 (*P*-value = 0.0001) in the WT study. However, this interaction did not reach the 0.05 significance level in the NHS/PHFS

TABLE V. Stepwise result for the genome-wide joint gene action analysis using the LRMW approach

Steps	Selected SNPs	Wellcome Trust study		NHS/PHFS *P*-value
		Nominal *P*-value	Adjusted *P*-value[a]	
1	rs4506565	4.77×10^{-11}		1.18×10^{-7}
2	rs4506565, rs7193144	1.29×10^{-16}		9.85×10^{-9}
3	rs4506565, rs7193144, rs8092098	1.59×10^{-23}		3.04×10^{-6}
4 [b]	rs4506565, rs7193144, rs8092098, rs12508397	8.02×10^{-32}	1.29×10^{-5}	4.01×10^{-6}

[a]Estimated based on the kernel density estimation procedure with 100 permutation replicates.
[b]The most parsimonious model identified by the LRMW approach.

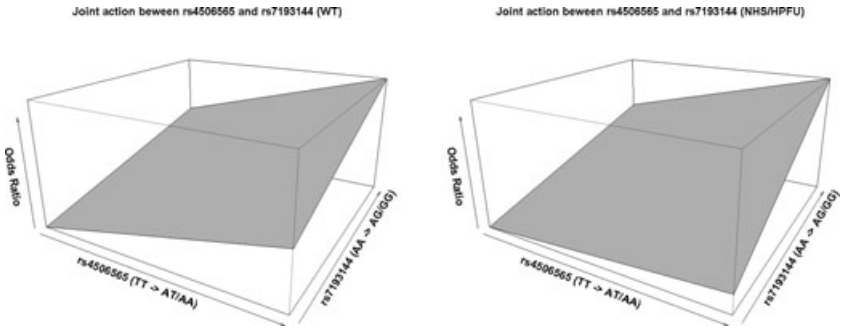

Fig. 3. The joint action between loci rs4506565 (*TCF7L2*) and rs7193144 (*FTO*). The clustered genotypes (e.g., CC/CT) were determined by the LRMW approach.

(*P*-value = 0.097). This may explain why the significance level of the joint association test decreased on including rs7193144 and rs8092098 in the NHS/PHFS (Table V). The lack of association of these two SNPs in the NHS/PHFS is likely due to chance. Further studies will be needed to evaluate the role of the two new putative loci in their association with T2D.

DISCUSSION

Following a genome-wide single SNP screening, a genome-wide study of joint gene action would be the natural next step to pursue [Cordell, 2009]. A genome-wide search is capable of considering all genetic variants simultaneously for identifying novel joint gene action. The findings from such a study are anticipated to lead to not only new genetic variants that may account for additional "missing heritability," but also a better understanding of the genetic architecture of common complex diseases [Cordell, 2009; Eichler et al., 2010; Maher, 2008]. To facilitate this kind of research, we propose here an LRMW approach, using a computationally efficient algorithm that allows for exploring joint gene actions on a genome-wide scale. As the number of loci increases, the computation time and memory needed for forward selection increase only linearly, while those for an exhaustive search (e.g., MDR) increase exponentially. This materially decreases the computational burden and makes it feasible for a genome-wide search. The method is nonparametric and thus does not suffer from the issue of an increasing number of parameters when modeling interactions statistically, especially high-order interactions, on a large number of SNPs. It is model-free in that it does not require specification of a particular mode of inheritance, and thus provides robust performance for various underlying disease models. The approach searches over all available SNPs to find a parsimonious model and conducts only one overall significance test on the selected model, and thus does not suffer from any multiple testing issue.

The LRMW test introduced here can be looked upon as an extension to multiple genetic loci of the traditional Mann-Whitney test [Mann and Whitney, 1947; Wilcoxon, 1945], taking into account their possible interactions. It is equivalent to a Mann-Whitney test comparing the LR scores between cases and controls, where the LR is defined as the ratio of the risk group's frequencies in cases vs. controls and has a range from 0 to $+\infty$. This link allows us easy access to the established properties of the Mann-Whitney test and provides a convenient way to test association. Instead of using Equation 4, we could also use the variance formula of the Mann-Whitney test with correction for ties [Lehmann and D'Abrera, 1975]. Other parameters of interest (e.g., odds ratios) can be easily obtained. The method is related to several *U*-statistic-based approaches [Schaid et al., 2005; Wei et al., 2008], sharing similar advantages—such as fewer statistical assumptions and more powerful and robust performance. Schaid et al. [2005] used a weighted sum kernel, averaging the difference in genotype scores between cases and controls, while Wei et al. [2008] used a symmetric kernel measuring a weighted Hamming distance between two individuals. Both kernels assume additivity across multiple SNPs, and can be used for a candidate gene association or pathway analysis. We use a kernel comparing the difference in LR values of risk groups between cases and controls. Since the LR is based on the joint probabilities of multiple SNPs, it does not assume independence of SNPs and can be used for assessing the joint action among multiple SNPs.

The method can be used to search for joint gene action both of candidate genes and on a genome-wide scale. Through a candidate gene approach, we identified four loci significantly associated with T2D in the WT study and confirmed the association in the NHS/HPFS. Assuming homogeneity of the other three genotypes, the disease odds ratio for the double recessive AATT to other three genotype is 0.57 (95% CI is [0.49, 0.66]) and 0.67 (95% CI is [0.59, 0.76]), respectively, in the WT study and the NHS/HPFS. Pooling the results of the two studies, the odds ratio for the double recessive AATT is 0.63 (95% CI is [0.57, 0.70]).

A candidate gene approach takes advantage of previous findings, and is powerful for capturing joint action among loci with strong marginal effects. The limitation of such an approach is that it could exclude potential statistical interactions between loci of small and medium-marginal effect size. A genome-wide analysis evaluates joint gene action over the entire genome and has more potential to identify novel interactions. However, it is more subject to false-positive results because it searches over a larger model space. In our genome-wide T2D analysis, we successfully identified two existing T2D loci without relying on any prior information. Results from the exploratory analysis (Figure 3) suggest the joint action between *TCF7L2* and *FTO* likely follows a multiplicative model. However, the two newly identified putative loci were found to be significant at the 0.05 level in the WT study but not in the NHS/HPFS study, so it may be a chance finding.

Accumulated evidence suggests that genes alone may play a limited role in T2D, thus a comprehensive investigation of both the genome and environment would be a natural next step. Note that the proposed approach can also be used for this purpose by simply coding environmental factors as categorical variables and evaluating all the combinations of genetic and environmental variables.

The proposed LRMW approach is computationally comparable to existing genome-wide approaches such as BOOST. The computational times for the genome-wide analysis of the WT T2D dataset using LRMW and BOOST are 55 hr and 60 hr [Wan et al., 2010], respectively. Compared with our T2D GWAS analysis, the pair-wise interaction analysis of the WT T2D GWAS dataset using BOOST identified 18 pair-wise statistically significant interactions. However, as reported by Wan et al., none of these interactions met the distance criterion (i.e., the distance between two interacting SNPs should be at least 1 MB), which suggests the significant interactions found are possibly just the result of linkage disequilibrium [Wan et al., 2010; Yung et al., 2011]. As demonstrated here, pair-wise analysis could be subject to low power if the underlying disease model is more complex than a simple two-locus model. As tree-based search algorithms, Random Jungle and the LRMW approach developed here both address the issue of high-order interactions and multicollinearity. However, tree-based algorithms require at least one of the selected loci to have a reasonably strong marginal effect. In contrast to Random Jungle, the LRMW approach provides a single test statistic and two measures of association—an odds ratio and the area under the receiver operating characteristic curve [DeLong et al., 1988; Lu and Elston, 2008], i.e., $U_{LRMW}/N_D N_{\bar{D}}$. This facilitates testing and quantifying the joint gene action, which is useful to evaluate both the likelihood of finding replication in a new study and possible clinical utility. Moreover, joint gene action models identified from the LRMW approach are easy to interpret, while models from Random Jungle can be hard to explain, being built on multiple trees that are each associated with a different set of SNPs. Since our focus here is genome-wide joint action analysis, we did not evaluate in detail approaches (e.g., MDR) that cannot be directly applied to a genome-wide search. Nevertheless, we carried out a small-scale simulation and found LRMW attained slightly more power than MDR. Using the simulation settings in Table II (Settings II), we found LRMW attained power of 0.649, 0.806, and 0.804, while MDR obtained power of 0.635, 0.779, and 0.792, respectively, under a two-way interaction model, a two two-way interactions model, and a three-way interaction model.

Finally, it must be stressed that the magnitude of any interaction found by statistical analysis depends on both the scale on which risk is measured and on how the marginal effects are defined [Wang et al., 2010]. Typically, logistic regression estimates marginal effects by using the numbers of observations in the cells of a multiway table as weights, leading to interaction terms that are not orthogonal to the main effects. Furthermore, data from a case-control study alone do not address the cause of any joint gene action. We tend to believe that any effect that cannot be explained by marginal effects is probably due to linkage disequilibrium if it occurs within a locus (i.e., the location of a gene), but to some kind of synergistic action between two genes if sites in different loci are involved. The truth is that in case-control data alone there is complete confounding between biological interaction and disequilibrium—linkage disequilibrium if the loci involved are linked, gametic phase disequilibrium if they are unlinked [Wang et al., 2010]. In the absence of experimental proof, inferring biological (i.e., physiological) interaction requires further analysis and a plausible causal mechanism. Nevertheless, even if no causal mechanism is involved, genome-wide genotypic associations can help predict risk to individuals in a population from which the cases and controls are a representative sample.

ACKNOWLEDGMENTS

This study makes use of data generated by the Wellcome Trust Case Control Consortium. Funding for the original WTCCC project was provided by the Wellcome Trust under award 076113. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk/info/participants.shtml. This work was supported by the National Institute of Dental and Craniofacial Research under Award Number R03DE022379, by the National Research Foundation of Korea Grant NRF-2011-220-C00004, funded by the Korean Government, by Cancer Center Support Grant Number P30CAD43703 from the National Cancer Institute, and Grant Numbers 1U01HG006382 and U01HG006382 from the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES. 2000. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat Genet 1:76–80.

Breiman L. 1996. Bagging predictors. Mach Learn 2:123–140.
Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 6:392–404.
Cornelis MC, Qi L, Zhang C, Kraft P, Manson J, Cai T, Hunter DJ, Hu FB. 2009. Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. Ann Intern Med 8:541–550.
Culverhouse RC, Saccone NL, Stitzel JA, Wang JC, Steinbach JH, Goate AM, Schwantes-An TH, Grucza RA, Stevens VL, Bierut LJ. 2011. Uncovering hidden variance: pair-wise SNP analysis accounts for additional variance in nicotine dependence. Hum Genet 2:177–188.
DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 3:837–845.
Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 6:446–450.
Elston RC. 1981. Segregation analysis. Adv Hum Genet 11:63–120.
Frayling TM. 2007. Genome-wide association studies provide new insights into type 2 diabetes aetiology. Nat Rev Genet 9:657–662.
Gloyn AL, Weedon MN, Owen KR, Turner MJ, Knight BA, Hitman G, Walker M, Levy JC, Sampson M, Halford S, McCarthy MI, Hattersley AT, Frayling TM. 2003. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. Diabetes 2:568–572.
Goldstein DB. 2009. Common genetic variation and human traits. N Engl J Med 17:1696–1698.
Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A, Styrkarsdottir U, Magnusson KP, Walters GB, Palsdottir E, Jonsdottir T, Gudmundsdottir T, Gylfason A, Saemundsdottir J, Wilensky RL, Reilly MP, Rader DJ, Bagger Y, Christiansen C, Gudnason V, Sigurdsson G, Thorsteinsdottir U, Gulcher JR, Kong A, Stefansson K. 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 3:320–323.
Hardy J, Singleton A. 2009. Genomewide association studies and human disease. N Engl J Med 17:1759–1768.
Healy JA, Nilsson KR, Hohmeier HE, Berglund J, Davis J, Hoffman J, Kohler M, Li LS, Berggren PO, Newgard CB, Bennett V. 2010. Cholinergic augmentation of insulin release requires ankyrin-B. Sci Signal 113:ra19.
Hirschhorn JN. 2009. Genomewide association studies—illuminating biologic pathways. N Engl J Med 17:1699–1701.
Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2:95–108.
Kirchhoff K, Machicao F, Haupt A, Schafer SA, Tschritter O, Staiger H, Stefan N, Haring HU, Fritsche A. 2008. Polymorphisms in the TCF7L2, CDKAL1 and SLC30A8 genes are associated with impaired proinsulin conversion. Diabetologia 4:597–601.
Lehmann EL, D'Abrera HJM. 1975. Nonparametrics Statistical Methods Based on Ranks. San Francisco: Holden-Day.
Li M, Ye C, Fu W, Elston RC, Lu Q. 2011. Detecting genetic interactions for quantitative traits with U-statistics. Genet Epidemiol 6:457–468.
Li WD, Dong C, Li D, Garrigan C, Price RA. 2004. A quantitative trait locus influencing fasting plasma glucose in chromosome region 18q22-23. Diabetes 9:2487–2491.
Liu C, Ackerman HH, Carulli JP. 2011. A genome-wide screen of gene-gene interactions for rheumatoid arthritis susceptibility. Hum Genet 5:473–485.
Lu Q, Elston RC. 2008. Using the optimal receiver operating characteristic curve to design a predictive genetic test, exemplified with type 2 diabetes. Am J Hum Genet 3:641–651.
Lu Q, Obuchowski N, Won S, Zhu X, Elston RC. 2010. Using the optimal robust receiver operating characteristic (ROC) curve for predictive genetic tests. Biometrics 2:586–593.
Maher B. 2008. Personal genomes: the case of the missing heritability. Nature 7218:18–21.
Mann HB, Whitney DR. 1947. On a test of whether one of 2 random variables is stochastically larger than the other. Ann Math Stat 1:50–60.
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 7265:747–753.
Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 4:413–417.
Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 7:906–913.
McDonough CW, Bostrom MA, Lu L, Hicks PJ, Langefeld CD, Divers J, Mychaleckyj JC, Freedman BI, Bowden DW. 2009. Genetic analysis of diabetic nephropathy on chromosome 18 in African Americans: linkage analysis and dense SNP mapping. Hum Genet 6:805–817.
Mohler PJ, Le SS, Denjoy I, Lowe JS, Guicheney P, Caron L, Driskell IM, Schott JJ, Norris K, Leenhardt A, Kim RB, Escande D, Roden DM. 2007. Defining the cellular phenotype of "ankyrin-B syndrome" variants: human ANK2 variants associated with clinical phenotypes display a spectrum of activities in cardiomyocytes. Circulation 4:432–441.
Parzen E. 1962. Estimation of a probability density-function and mode. Ann Math Stat 3:1065–1076.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 3:559–575.
Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 1:138–147.
Rosenblatt M. 1956. Remarks on some nonparametric estimates of a density-function. Ann Math Stat 3:832–837.
Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson BK, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 5829:1331–1336.
Schaid DJ, McDonnell SK, Hebbring SJ, Cunningham JM, Thibodeau SN. 2005. Nonparametric tests of association of multiple genes with human disease. Am J Hum Genet 5:780–793.
Schwarz DF, Konig IR, Ziegler A. 2010. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics 14:1752–1758.
Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 5829:1341–1345.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 7130:881–885.

Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S, Baker A, Snorradottir S, Bjarnason H, Ng MC, Hansen T, Bagger Y, Wilensky RL, Reilly MP, Adeyemo A, Chen Y, Zhou J, Gudnason V, Chen G, Huang H, Lashley K, Doumatey A, So WY, Ma RC, Andersen G, Borch-Johnsen K, Jorgensen T, van Vliet-Ostaptchouk JV, Hofker MH, Wijmenga C, Christiansen C, Rader DJ, Rotimi C, Gurney M, Chan JC, Pedersen O, Sigurdsson G, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K. 2007. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet 6:770–775.

Tiwari HK, Elston RC. 1998. Restrictions on components of variance for epistatic models. Theor Popul Biol 2:161–174.

Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarroll SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segre AV, van HM, Navarro P, Ardlie K, Balkau B, Benediktsson R, Bennett AJ, Blagieva R, Boerwinkle E, Bonnycastle LL, Bengtsson BK, Bravenboer B, Bumpstead S, Burtt NP, Charpentier G, Chines PS, Cornelis M, Couper DJ, Crawford G, Doney AS, Elliott KS, Elliott AL, Erdos MR, Fox CS, Franklin CS, Ganser M, Gieger C, Grarup N, Green T, Griffin S, Groves CJ, Guiducci C, Hadjadj S, Hassanali N, Herder C, Isomaa B, Jackson AU, Johnson PR, Jorgensen T, Kao WH, Klopp N, Kong A, Kraft P, Kuusisto J, Lauritzen T, Li M, Lieverse A, Lindgren CM, Lyssenko V, Marre M, Meitinger T, Midthjell K, Morken MA, Narisu N, Nilsson P, Owen KR, Payne F, Perry JR, Petersen AK, Platou C, Proenca C, Prokopenko I, Rathmann W, Rayner NW, Robertson NR, Rocheleau G, Roden M, Sampson MJ, Saxena R, Shields BM, Shrader P, Sigurdsson G, Sparso T, Strassburger K, Stringham HM, Sun Q, Swift AJ, Thorand B, Tichet J, Tuomi T, van Dam RM, van Haeften TW, van HT, van Vliet-Ostaptchouk JV, Walters GB, Weedon MN, Wijmenga C, Witteman J, Bergman RN, Cauchi S, Collins FS, Gloyn AL, Gyllensten U, Hansen T, Hide WA, Hitman GA, Hofman A, Hunter DJ, Hveem K, Laakso M, Mohlke KL, Morris AD, Palmer CN, Pramstaller PP, Rudan I, Sijbrands E, Stein LD, Tuomilehto J, Uitterlinden A, Walker M, Wareham NJ, Watanabe RM, Abecasis GR, Boehm BO, Campbell H, Daly MJ, Hattersley AT, Hu FB, Meigs JB, Pankow JS, Pedersen O, Wichmann HE, Barroso I, Florez JC, Frayling TM, Groop L, Sladek R, Thorsteinsdottir U, Wilson JF, Illig T, Froguel P, van Duijn CM, Stefansson K, Altshuler D, Boehnke M, McCarthy MI. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 7:579–589.

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL, Yu W. 2010. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. Am J Hum Genet 3:325–340.

Wang X, Elston RC, Zhu X. 2010. The meaning of interaction. Hum Hered 4:269–277.

Wei Z, Li M, Rebbeck T, Li H. 2008. U-statistics-based tests for multiple genes in genetic association studies. Ann Hum Genet 6:821–833.

Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 7145:661–678.

Wilcoxon F. 1945. Individual comparisons by ranking methods. Biometr Bull 6:80–83.

Woolf B. 1955. On estimating the relation between blood group and disease. Ann Hum Genet 4:251–253.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 6:929–942.

Ye C, Cui Y, Wei C, Elston RC, Zhu J, Lu Q. 2011. A non-parametric method for building predictive genetic tests on high-dimensional data. Hum Hered 3:161–170.

Yung LS, Yang C, Wan X, Yu W. 2011. GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. Bioinformatics 9:1309–1310.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jorgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjogren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 5:638–645.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 5829:1336–1341.

Zhang Y, Liu JS. 2007. Bayesian inference of epistatic interactions in case-control studies. Nat Genet 9:1167–1173.