A fast and powerful W-test for pairwise epistasis testing

Maggie Haitian Wang[12†*], Rui Sun[12†], Junfeng Guo[3], Haoyi Weng[12], Jack Lee[1], Inchi Hu[4], Pak Chung Sham[5], and Benny Chung-Ying Zee[12*]

[1]Division of Biostatistics and Centre for Clinical Research and Biostatistics, JC School of Public Health and Primary Care, the Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[2]CUHK Shenzhen Research Institute, Shenzhen, China
[3] the Australian National University, Canberra, Australia
[4]ISOM Department and Biomedical Engineering Division, the Hong Kong University of Science and Technology, Kowloon, Hong Kong SAR
[5]Department of Psychiatry; Centre for Genomic Sciences, the University of Hong Kong, Pok Fu Lam, Hong Kong SAR, China

*To whom correspondence should be addressed. Tel: +852 2252 8725; Fax: +852 2646 7297; Email: maggiew@cuhk.edu.hk

† The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

ABSTRACT

Epistasis plays an essential role in the development of complex diseases. Interaction methods face common challenge of seeking a balance between persistent power, model complexity, computation efficiency, and validity of identified bio-markers. We introduce a novel W-test to identify pairwise epistasis effect, which measures the distributional difference between cases and controls through a combined log odds ratio. The test is model-free, fast, and inherits a Chi-squared distribution with data adaptive degrees of freedom. No permutation is needed to obtain the p-values. Simulation studies demonstrated that the W-test is more powerful in low frequency variants environment than alternative methods, which are the Chi-squared test and logistic regression. In two independent real bipolar disorder genome-wide associations (GWAS) datasets, the W-test identified significant interactions pairs that can be replicated, including *SLIT3-CENPN*, *SLIT3-TMEM132D*, *CNTNAP2-NDST4* and *CNTCAP2-RTN4R*. The genes in the pairs play central roles in neurotransmission and synapse formation. A majority of the identified loci are undiscoverable by main effect and are low frequency variants. The proposed method offers a powerful alternative tool for mapping the genetic puzzle underlying complex disorders.

INTRODUCTION

Genetic association studies have identified a repertoire of susceptible loci that are associated with common diseases. However, they have collectively explained only a small fraction of disease heritability (1-3). It was widely accepted that epistasis, or gene-gene interactions, plays an essential role in the development of complex disorders (4,5). Past epistasis studies mainly focused on the

genomic region in which the minor allele frequency (MAF) is greater than 5%, while rare variant analysis focused on main effect in the exome region where MAF is less than 1%. The low frequency variants, in which MAF is between 1% and 5%, remain largely understudied. Available tools to calculate genome-wide epistasis can be broadly grouped into three categories: the parametric methods represented by logistic regression, non-parametric methods represented by the classic Pearson's Chi-squared test, and the machine learning method by the multifactor dimensionality reduction (MDR) (6). Logistic regression assumes a linear relationship between phenotype and genotypic combinations. It has the unique property of providing an odds ratio interpretation, which allows it to give prospective inferences from retrospective case-control datasets (7). The Pearson's Chi-squared test is fast and non-parametric. However, it requires some minimum cell counts for the test statistic to follow a Chi-squared distribution. The MDR is a very powerful machine learning approach that first pools the genotype combinations into a low risk and a high risk group to achieve dimensionality reduction, evaluates the multi-locus model through cross-validations, and then estimates the model p-values through permutations. Despite the specialties of various methods, the detection of interaction effects faces the common challenges of bringing persistent power in intricate genetic architectures, varying sample sizes, computing efficiency, and reproducibility of results towards mapping clinically relevant biomarkers.

Bipolar disorder (BD) is a serious mental disorder that is characterized by episodes of mania and deep depression. Family studies suggest that the heritability of BD is 80–85% (8-10). Overwhelming evidence shows that genetics play a fundamental role in the onset of BD besides the influence of environmental factors. However, in past decades, genetic association studies had difficulty in identifying suspect genes relating to BD with large effect sizes, and the explained heritability is less than 5% (11,12). Interplay of multiple genetic markers is crucial for the etiology of bipolar disorder; therefore, we hypothesized that we would have interesting findings when epistasis effects were considered in BD datasets.

In this paper, we introduce a W-test for pairwise epistasis testing that has robust power covering the genome where MAF > 1%. The W-statistic tests the null hypothesis that the joint distribution of a set of single nucleotide polymorphisms (SNPs) is different in the cases from that in the control group. The distributional differences are measured by a combined log odds ratio from the contingency table, with two scalars estimated from the null hypothesis. The method is advantageous in several respects. First, it is model-free, such that it makes no assumption about genotypic effect model. Second, it is very fast; it only uses a subset of bootstrap samples to estimate two distribution parameters and calculate p-values, and genome-wide screening can be performed efficiently. Third, the W-test incorporates a statistical distribution that is data-adaptive, such that the association measurement is robust for various genetic scenarios. In principle, the W-test takes the form of Chi-squared distribution, and its degrees of freedom are estimated from the covariance structure of a contingency table formed by the interaction set. The data-dependent degrees of freedom allow the method to cope with low frequency genotypes, which, for classic tests, will result in low power from imperfect statistical distributions. The W-test showed robust power and reasonable type I error in various genetic environments; when the variants frequency is low, it outperforms all alternative methods.

The remainder of the article is organized as follows. In the next section, we describe the proposed method, including its formulation and distribution. We then will test the power and type I error of the proposed methods and alternative methods under different genetic models and genetic architectures, using simulated phenotype generated from real data. The method will then be applied to an American Caucasian's bipolar GWAS data and an independent European Caucasian's data. We identified a number of genes that are highly relevant to neuronal function and depressive disorders, which can be replicated by the two datasets. To our knowledge, this is also the first report of successful replication of the genes with significant epistasis effect in GWAS. The method proposed also has general application values for identifying disease-susceptible interactions in other types of data.

MATERIAL AND METHODS

The W-test formulation

The basic hypothesis of the W-test is that the statistical distributions of a set of disease-associated markers are different in the case group from that in the control group. Under a co-dominant model, the genotype data X can be coded by minor allele count to take values (0, 1, 2). The phenotype Y is binary for the case and control dataset. To test the association of a pair of SNPs (X_1, X_2), a 2 by 9 contingency table can be formed. Let k denote the number of columns of the table. The cell distribution of (X_1, X_2) in the case and control group can be written as:

$$\hat{p}_{1i} = \Pr(X \mid Y = 1) = \frac{n_{1i}}{N_1}, \quad \hat{p}_{0i} = \Pr(X \mid Y = 0) = \frac{n_{0i}}{N_0}, \quad i = 1,...,k$$

where n_{1i} is the number of case subjects in the i^{th} cell, N_1 is the total number of cases, n_{0i} is the number of control subjects in the i^{th} cell, and N_0 is the total number of controls. For pair-wise interactions, k = 9. The method can also accommodate main effect testing. When a single SNP is considered, k=3. For both case and control samples, we have:

$$\sum_{i=1}^{k} \hat{p}_{1i} = 1, \text{ and } \sum_{i=1}^{k} \hat{p}_{0i} = 1,$$

To measure the discordance between the two distributions, we first use the following measure to combine the normalized log odds ratios of the cell probability distributions:

$$X^2 = \sum_{i=1}^{k} \left[\log \frac{\hat{p}_{1i}/(1-\hat{p}_{1i})}{\hat{p}_{0i}/(1-\hat{p}_{0i})} \Big/ SE_i \right]^2 \qquad \text{Equation 1}$$

Where,

$$SE_i = \sqrt{\frac{1}{n_{0i}} + \frac{1}{n_{1i}} + \frac{1}{N_0 - n_{0i}} + \frac{1}{N_1 - n_{1i}}} \ .$$

Diagram 1 shows the decomposition of X^2. Since the contingency table's margins are fixed, the cell probabilities are not entirely independent. If they are independent, the X^2 would follow a k degrees of freedom Chi-squared distribution. The mutual dependency among the cells decreases as k becomes

large. The distribution of the X^2 can be estimated by matching its first two moments to the moments of the following variable R of a known Chi-squared distribution with f degrees of freedom (13):

$$R = c\chi_f^2$$

The first two moments of X^2 are:

$$E(X^2) = k$$

$$\sigma^2(X^2) = \sum_i \sum_j \mathrm{cov}(x_i^2, x_j^2) = \sum_{i=1}^{k} Var(x_i^2) + 2\sum_{i<j} \sum \mathrm{cov}(x_i^2, x_j^2)$$

$$= 2k + 2\sum_{i<j} \sum \mathrm{cov}(x_i^2, x_j^2).$$

x_i and x_j are the components in the summation sign in Equation 1, which are the single cell's normalized log of odds ratio. And the first two moments of R are:

$$\begin{cases} E(X^2) = cf \\ \sigma^2(X^2) = 2c^2 f \end{cases}$$

The c and f can be obtained accordingly:

$$c = \frac{\sigma^2(X^2)}{2E(X^2)} = \frac{2k + 2\sum_{i<j} \sum \mathrm{cov}(x_i^2, x_j^2)}{2k}$$

$$f = \frac{2[E(X^2)]^2}{\sigma^2(X^2)} = \frac{2k^2}{2k + 2\sum_{i<j} \sum \mathrm{cov}(x_i^2, x_j^2)}$$

Let h=1/c, we define the W-test, with a scalar h before the X^2, as:

$$W = h\sum_{i=1}^{k} \left[\log \frac{\hat{p}_{1i}/(1-\hat{p}_{1i})}{\hat{p}_{0i}/(1-\hat{p}_{0i})} \Big/ SE_i \right]^2 \sim \chi_f^2 \qquad \text{Equation 2}$$

Thus the W-test follows a Chi-squared distribution with f degrees of freedom. The approximation is shown to give accurate probability by numerical studies (14). Theoretical justification for the validity of the approximation is given by Chuang and Shih (2012) (15). In real data analysis, it might be difficult to obtain the covariance matrix for X^2. Large sample theory can be used to estimate the covariance from smaller bootstrap samples under the null hypothesis. Each bootstrap sample consists of the real genotype data and permutated Y. Suppose total number of subjects is N, and total number of pairs is P. Converging estimates for h and f can be achieved by setting bootstrap times B > 200, subjects number N_B= min (1000, N), and number of pairs P_B= min (1000, P) (Supplementary Information S1). Empirical studies give $h \approx (k-1)/k$ and $f \approx k-1$. A table of estimated h and f in real data is provided in the Supplementary Information S2. Frequently, the degrees of freedom f are non-integer. Then the Chi-squared distribution is in fact a Gamma (f /2, 2) distribution. The covariance of X^2 is dataset dependent, so for every set of new genotypes, h and f need to be estimated. When there is an empty cell, a continuity correction is applied by adding 0.5 to all cells.

The W-test is a combined log of odds ratio test based on maximum likelihood probabilities conditioned on disease status. It is equivalent to testing the following null hypothesis:

H_0: P(X_1, X_2 | Y = 1) = P(X_1, X_2 | Y = 0), such that OR_i = 1, for i = 1,…, k

The test is model-free, and does not assume the form of interactions. Because of its odds ratio form, it is suitable to be applied to retrospective case-control datasets, which is how data are collected in most of the genome studies. If a W-value is large and the null hypothesis is rejected, we would conclude that the joint probability distribution has a significant difference between cases and controls, which indicates the interactive set (X_1, X_2) has association effect. The classic odds ratio test is a special case of W-test for a single marker with two levels. The W-test can be extended to higher order interactions mapping. In this paper, we mainly focus on pair-wise interactions tests in simulation studies. Simulation study of main effect is provided in Supplementary Materials S3.

Application on simulated datasets

Simulated datasets are composed of genotypes from real GWAS datasets and simulated phenotypes. Different genetic architectures considered include minor allele frequency (MAF) in the common range (MAF>5%) and in the low frequency range (1%< MAF< 5%); linkage disequilibrium (LD) in the high range (r^2>80%), medium range (20%<r^2<80%), and low range (r^2<20%) (16,17). In each of the six genetic architecture combinations, 50 SNPs and 1,000 subjects are randomly drawn from real GWAS datasets. The original phenotype label is removed, and binary response variables are simulated using two types of model, specified as follows.

Model 1. A linear regression model with interaction effect. A linear model can be prescribed by the following logistic regression (18):

$$LOGIT[P(Y=1)] = \begin{cases} \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 & p = 0.3 \\ \beta_4 + \beta_5 X_3 + \beta_6 X_4 + \beta_7 X_3 X_4 & p = 0.3 \\ \beta_8 & p = 0.4 \end{cases}$$

The logit has 30% probability to be equal to the first equation, 30% probability to take the second equation, and 40% chance of equaling β_8, which is a random real number. The coefficients are chosen such that the case and control ratio is balanced. This model contains both the main and the interaction effect terms, and the coefficients of the cross-product β_3 and β_7 terms are tested for interaction effects. The genotype takes values of 0, 1, or 2, under a co-dominant genetic model assumption.

Model 2. A non-linear interaction model. The Y has a non-linear association to (X_1, X_2) and (X_3, X_4) (19,20):

$$Y = \begin{cases} X_1 + X_2 \,(\mathrm{mod}\, 2) & p = 0.3 \\ X_3 + X_4 \,(\mathrm{mod}\, 2) & p = 0.3 \\ 0,1 & p = 0.4 \end{cases},$$

where the Ys are (X_1 + X_2) mod 2 or (X_3 + X_4) mod 2 with 30% probability, and are randomly assigned to be 0 or 1 with 40% probability. Only pure interaction effect is present in this model. The null hypothesis is that none of the X predictor is associated with Y.

Power and type I error rate calculation using simulated datasets

For power calculation, simulated dataset using Model 1 or Model 2 for each genetic architecture block is generated 1,000 times. Pairwise interactions are calculated exhaustively by the W-test and alternative methods. For 50 SNPs, 1,225 pairs are evaluated. An interaction set is significant if its p-value is smaller than 4.1×10^{-5}, which is the Bonferroni corrected p-value at a family-wise error rate (FWER) of 0.05. Methods that result significant p-values of all interaction effect variables are said to have successfully identified the causal markers. Power is the averaged true positive proportion in 1,000 simulated datasets. For type I error estimation, Y is permutated 10^6 times, and type I error rate is the average false positive proportion in one million permutated datasets.

Application to real GWAS datasets

Datasets

The W-test is then applied to two real bipolar GWAS datasets. The first dataset is from the Wellcome Trust Case Control Consortium (WTCCC), comprising 2,000 bipolar cases and 3,000 controls of European Caucasians (21). The dataset includes 500,568 SNPs genotyped with GeneChip 500K Mapping Array Set (Affymetrix Chip). The second dataset is from the Genetic Association Information Network (GAIN) project (22). The GAIN data is composed of an American ancestry population with a bipolar phenotype, which contains 653 cases and 1,767 controls. The data includes 906,600 SNPs genotyped with Affymetrix 6.0 platform. Quality control is performed: SNPs with high missing rates (> 5%), MAF < 1%, and which depart from the Hardy-Weinberg equilibrium are excluded (23). After the quality control, the WTCCC dataset includes 414,682 SNPs and the GAIN data contains 729,304 SNPs.

Pair-wise interaction search

A three-step procedure is used to search for pair-wise interaction, described as follows.

Step 1. Main effect search The main effects are evaluated exhaustively on the whole genome by the W-test. The SNPs with p-value that is less than 0.01 are passed to the next step SNP-SNP interaction test. The p-value < 0.01 compared to the genome-wide significance 10^{-8} is trivial (24). Thus this filtering can include the weak effect markers that are potentially influential in an interaction setting, while downsizing the candidate sets.

Step 2. Two-way interaction search The epistasis test is performed on the candidate markers. A pair of SNPs will be selected if its p-value passes the Bonferroni corrected alpha at FWER 5%. The p-value of the pair should be more significant than component SNPs' p-value.

Step 3. Map SNP-SNP to gene-gene interaction pair The SNP-pairs are matched to corresponding genes using web-based databases (25,26); interaction networks are created by linking

significant pairs (27,28). Steps 1-3 are performed on the WTCCC and the GAIN data sets respectively, and interactions that are replicated by the two data sets are reported.

RESULTS

Simulation study statistical power and type I error rate

Linear interaction model with main effect

The W-test is compared to the logistic regression and Chi-squared test under different genetic architectures (Table 1). Under the common variant (MAF > 5%) and high LD environment, the power for logistic regression is 83.3%, Chi-squared test is 74.5% and W-test is 86.7%. However, the nominal type I error of the MDR is greater than 1; while W-test's nominal type I error is 5.5% (Table 1). When MAF is low (1% <MAF< 5%), the W-test has the highest power for all LD scenarios. Specifically, in the mid LD range, the power of W is 79.5%, compared to the Chi-square's 65.2% and logistic regression's 62.5%. The nominal type I error of the W-test is 5.1%. Interestingly, the model-free W-test outperforms the logistic regression even when the underlying model is linear. In general, the power of all methods improves when the variables are in high LD, and drop as their mutual correlation diminish (Figure 1a). This is likely due to the presence of main effect terms in the linear model, such that a causal marker can pair with another one due to high LD, which makes it easier to be identified.

Non-linear interaction model without main effect

In the common MAF environment, the average power of W-test's is 84.7%, higher than Chi-square's 68.3%, and the nominal type I error for the W-test is around 5% (Table 1). In the low frequency variant environment (1%<MAF<5%), the average power of W-test is 87.6%, 63.8% greater than the Chi-square's. The average nominal type I error of W-test is 5.3%. Specifically, when LD is medium, the power of W is 83.3%, compared to the Chi-squared's 43.9% and the logistic regression's 31.8%. The nominal type I errors for the methods are 5.1%, 0.2% and 5.3%, respectively. The results show that the W-test has robust power and reasonable type I errors in both the common and low frequency variant environment and various LD scenarios. When necessary, the type I error of the W-test can be refined using permutation method for selected markers. Now we briefly describe how LD pattern affects performance for the non-linear two-locus model. Model 2 does not contain any main effect terms, so a high LD environment will not form many strong signal-noise combinations as by Model 1. When the LD is low, the signal-signal pairs could be easier to be distinguished against a low noise background, thus all methods showed higher power in this scenario.

Power and type I error as sample size changes

We reduce the sample size from 1,000 to 300 subjects, and compare the different methods' power and type I error performances under the non-linear model and low frequency variants scenario. When the sample size decreases, the W-test still demonstrates the highest and most robust power (Figure 2, Table 2). At N=800, the performance of W is 82.2%, which drops 2% compared to the power at N=1,000; while the Chi-square's power goes down to 37.8%, dropped 18% compared to N=1,000. The logistic regression's power dropped from 29.1% to 17.7% (Table 2). A sample size ranging from

300 to 500 is common for small scale biomedical studies. At N=400, the W-test's power is 28.8%, while the alternative methods' power falls below 5%. Furthermore, the type I errors of the W are very stable, averaging 4.6×10^{-5} with standard deviation of 4.5×10^{-6} (Table 2), while the alternative methods display stringent type I errors that could have affected their power. When N is smaller than 700, the Chi-squared test has type I errors below 1.0×10^{-5}, which are conservative compared to the multiple testing error rate at 4.1×10^{-5}.

Computing time

On a laptop computer with 2.4 GHz CPU and 8GB memory, for the W-test, the time elapsed for computing one simulation study of 1,225 SNP-pairs and 1,000 subjects is 7.4 seconds(s); the Chi-squared took 7.7s and the logistic regression took 45. For real data analysis, the W-test takes 3.4 hours for genome-wide main effect evaluation, and takes about the same time for the stage-wise interaction effects.

Real datasets applications

SNP-SNP interaction has identified a number of replicated genes from the two independent GWAS datasets (Table 3). The Q-Q plot of pair-wise interactions shows no inflation of spurious association (Figure 3). Interestingly, a majority of these replicated genes are marginally insignificant, which means that they are undiscoverable through main effect screening. Furthermore, these markers show a highly relevant biological function to autism spectrum disorder. The replicated and significant gene-gene interactions can be summarized in two networks (Figure 4 and Supplementary Information S6). The first network (Figure 4a) consists of 8 genes, in which only one gene, *RTN4R*, has significant main effect. It encodes a Nogo receptor that mediates axonal growth inhibition and may play a role in regulating axonal regeneration and plasticity in the central nervous system. Studies reported that the deletion of the gene would cause microstructural anomalies in brain white matters (29); human and mouse genetic studies suggested the gene to be a candidate marker for schizophrenia (30). The gene *SLIT3* is coupled with *CENPN* and *TMEM132D*, forming two significant interaction pairs. Experiments showed that the gene *SLIT3* (5q35) decreases neurogenesis and may play a role in regulating neuron-vessel interactions (31). The gene *DPP10* is marginally insignificant but is replicated by significant interactions in both datasets. *DPP10* facilitates neuronal excitability and its aberrant distribution is associated with Alzheimer's disease as revealed by immunohistochemistry (32). Other genes in the network are also highly related to autism and neurodegenerative disorders. For example, *NRXN3* encodes a protein that functions as a synaptic adhesion protein; *TMEM132D* is a transmembrane protein expressed in white matter in the spinal cord and optic nerve (33); *PTPRT* is a receptor-type protein tyrosine phosphatase for signal transduction and neurite extension, which promotes synapse formation and is reported to be highly expressed in the central nervous system (34). Chromosome position, MAF and p-value of the replicated genes are reported in Table 3.

In the second network (Figure 4b), the well-known autism spectrum disorder-associated loci *PARK2* (rs2849605, 6q5.2) has been identified. The pair *ELMO1-A2BP1* has significant epistasis

effect (p-value =3.9E-18), while the component SNPs are non-significant with p-values of 3.0×10^{-6} and 4.0×10^{-3}, respectively (Supplementary Information S6). Both SNPs are replicated in the GAIN dataset through significant interactions with *RTN4R*. *ELMO1* encodes a protein that interacts with *DOCK180* to promote phagocytosis and cell migration. The *A2BP1*, other name *RBFOX1*, is an RNA-binding protein that regulates alternative splicing in neurons and plays a key role in the development of human neurons reported in RNA-sequencing, cytogenetic, and molecular characterization studies (35,36). The gene *CNTNPA2* has weak main effects in both datasets, yet the gene is replicated through its significant interactions with *NDST4* and *RTN4R*. The product of this gene functions in the vertebrate nervous system as cell adhesion molecules and receptor.

DISCUSSION

We propose the W-test as a general measure for epistasis testing. It is fast, model-free, and powerful. We have demonstrated that the W-test has robust power for linear and non-linear genetic models over a range of genetic environments. The method is especially advantageous for low frequency variants and has persistent power when the sample size is small. The advantages of the W-test are explained through the following characteristics.

The proposed method aims to test the distributional differences between cases and controls, using the sum of squared log odds ratio over the complete cell distribution in a contingency table. The cell distribution that is formed by a pair of markers has the overall probability to be one, in the control group and the case group, respectively. This constraint keeps the cell proportions to reflect distributional differences, which are tested cell by cell using the odds ratio. The W-test is different from the Chi-squared test in three aspects: first, the W tests the case-control distributional differences, while the Chi-squared tests the observed distribution against the joint distribution under an independence assumption. Second, the W-test does not depend on the total sample size, but is a function of the cell proportions; the Pearson Chi-squared test is a function of both the cell proportions and total sample size, such that it can measure the association significance but not the association magnitude (37,38). Third, the Pearson's Chi-squared test has a more stringent requirement on cell sample size. It is well known that the minimum expected cell counts should be no less than 5 for good approximation to a Chi-squared distribution. For the W-test, at extreme cell count, the distribution approximation is corrected through the implementation of *h* and *f* that are estimated from the sample covariance.

The odds ratio has a unique property for drawing prospective inference from a retrospective data set. The GWAS case-and-control dataset has a retrospective nature, so the W-test with an odds ratio interpretation is especially advantageous. The logistic regression also has an odds ratio interpretation. However, it assumes that the *logit* has a linear relationship with the interaction term x_1x_2. Under a co-dominant genetic model, the heterozygous *Aa* genotype may correspond to an over-expression of the phenotype, while the homozygous *aa* and *AA* genotype may associate with suppressed phenotypes. This non-linear relationship will be missed by logistic regression unless indicator variables for genotypes are specified. On the other hand, the W-test takes a sum of squared form, such that the

genotypic combinations can have opposite effect directions; and no genetic model is assumed. In the simulation studies, all the non-parametric tests performed better than the logistic regression when the underlying model is non-linear.

The W-test inherits a statistical distribution that is adaptive to the data. A direct benefit is having a built-in distribution that saves the computational cost for permutations to calculate p-values, although a small proportion of time is needed for estimating the *h* and *f* from bootstrapped samples. The original purpose of employing the parameters is to handle correlation among the odds ratios in W, and the implementation has several bonuses. First, deviation from an ideal distribution caused by sparse data can be corrected by *h* and *f*. The accurately approximated distribution leads the W-test to have persistent power at small sample size. The second bonus is that the adjustment has the spirit and effect of performing the genomic control (39). The covariance used to calculate the parameters is estimated from bootstrapped subjects and permuted *Y*, under the hypothesis of no disease association and sample independence, which is similar to the genomic control procedure. Consequently, the *h* and *f* absorb the extra variance caused by population stratification. The effect of this correction is evident from the Q-Q plots of W-test on the real GWAS dataset (Figure 3), which showed perfect null distributions. This property can make the W-test robust against false positives arising from population structure, and can assist the replication of genetic markers using other independent datasets.

Therefore, the proposed method showed better power in the low frequency variables environment. In fact, in the two real datasets applications, 76.4% of the identified significant main effect SNPs have MAF less than 5%, which explains the large number of SNPs passing the genome-wide significant level compared to previous GWAS studies of Bipolar Disorder (21,22,40). In the two significant interaction networks, 10 out of the 11 replicated genes are identified by SNPs with MAF in the 1% to 5% range (Table 3). Another interesting observation from the real data analysis is that we have identified a number of genes that are highly relevant to neuron disorder but have not been found by previous GWAS studies, even when the SNPs are common variants. The standing example is SNP A-84299018 of gene *RTN4R* in the GAIN dataset. The SNP has a large MAF 12.2%, and p-value is 5.9E-18. The *RTN4R* acts as a hub gene connecting to all important genes in the GAIN dataset (Figure 4b). Furthermore, there are converging evidences of this Nogo receptor gene's association with schizophrenia through brain cell, animal and candidate gene studies (41-43). Nevertheless, so far we have not seen a GWAS study to report its significant association with autism related diseases. Similar to the *RTN4R*, *SLIT3* (SNP_A-2229791, MAF=0.22, p-value= 3.7E-03) has been identified from copy-number variations analysis by whole genome sequencing (44) and mRNA expression studies (31). *TMEM132D* (SNP_A-8630842, MAF=0.14, p-value = 5.9E-03) is found to associate with anxiety co-morbidity in depression and panic disorder by brain mRNA analysis (45); the *NRXN3* (rs17108944, MAF = 0.03, P-value = 5.1E-3) has been reported to have a strong association with schizophrenia via studies of gene expression (46), DNA-pooling (47), and candidate genes .

To conclude, we proposed the W-test as a model-free and dataset adaptive method for detecting epistasis in genotype dataset. It is fast, robust, and possesses statistical distributions. Real data

analysis replicated important genes with epistasis effect, which were undiscoverable through main effect evaluations. These results showed that the W-test is a very powerful and practical tool for detecting functional variants thereby helping to solve the genetic puzzle underlying complex diseases.

AVAILABILITY

The W-test software is available at: http://www2.ccrb.cuhk.edu.hk/wtest/download.html

ACKNOWLEDGEMENT

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. We thank Tony Liu and Sammy Tang from Li Ka Shing Institute of Health Science CUHK, and Morris Law from HPCCC, Hong Kong Baptist University for providing technical support of computer cluster. We thank Xin Lai and William K.K. Wu for providing valuable comments on the manuscript. We thank both reviewers for their constructive comments.

FUNDING

This work has been supported the Chinese University of Hong Kong Direct Grant [4054169] to MHW; Research Grant Council – General Research Fund [476013] to MHW; and National Science Foundation of China [81473035, 31401124] to MHW.

Conflict of Interest: none declared.

REFERENCES

1. Cantor, R.M., Lange, K. and Sinsheimer, J.S. (2010) Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. *Am J Hum Genet*, **86**, 6-22.
2. Kraft, P. and Hunter, D.J. (2009) Genetic Risk Prediction - Are We There Yet? *New Engl J Med*, **360**, 1701-1703.
3. Witte, J.S., Visscher, P.M. and Wray, N.R. (2014) The contribution of genetic variants to disease depends on the ruler. *Nat Rev Genet*, **15**, 765-776.
4. Phillips, P.C. (2008) Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet*, **9**, 855-867.
5. Cordell, H.J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*, **10**, 392-404.
6. Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. and Moore, J.H. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet*, **69**, 138-147.
7. Agresti, A. (1996) *An introduction to categorical data analysis*. Wiley, New York.
8. Smoller, J.W. and Finn, C.T. (2003) Family, twin, and adoption studies of bipolar disorder. *Am J Med Genet C*, **123C**, 48-58.
9. McGuffin, P., Rijsdijk, F., Andrew, M., Sham, P., Katz, R. and Cardno, A. (2003) The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch Gen Psychiat*, **60**, 497-502.
10. Cardno, A.G., Marshall, E.J., Coid, B., Macdonald, A.M., Ribchester, T.R., Davies, N.J., Venturi, P., Jones, L.A., Lewis, S.W., Sham, P.C. *et al.* (1999) Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. *Arch Gen Psychiat*, **56**, 162-168.
11. Craddock, N. and Sklar, P. (2013) Bipolar Disorder 1 Genetics of bipolar disorder. *Lancet*, **381**, 1654-1662.

12. So, H.C., Gui, A.H.S., Cherny, S.S. and Sham, P.C. (2011) Evaluating the Heritability Explained by Known Susceptibility Variants: A Survey of Ten Complex Diseases. *Genet Epidemiol*, **35**, 310-317.
13. Brown, M.B. (1975) Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics*, **31**, 987-992.
14. Hou, C.D. (2005) A simple approximation for the distribution of the weighted combination of non-independent or independent probabilities. *Stat Probabil Lett*, **73**, 179-187.
15. Chuang, L.L. and Shih, Y.S. (2012) Approximated distributions of the weighted sum of correlated chi-squared random variables. *J Stat Plan Infer*, **142**, 457-472.
16. Erdmann, J., Grosshennig, A., Braund, P.S., Konig, I.R., Hengstenberg, C., Hall, A.S., Linsel-Nitschke, P., Kathiresan, S., Wright, B., Tregouet, D.A. *et al.* (2009) New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet*, **41**, 280-282.
17. Corvin, A., Craddock, N. and Sullivan, P.F. (2010) Genome-wide association studies: a primer. *Psychol Med*, **40**, 1063-1077.
18. He, J., Wang, K., Edmondson, A.C., Rader, D.J., Li, C. and Li, M.Y. (2011) Gene-based interaction analysis by incorporating external linkage disequilibrium information. *Eur J Hum Genet*, **19**, 164-172.
19. Chernoff, H., Lo, S.H. and Zheng, T.A. (2009) Discovering Influential Variables: A Method of Partitions. *Ann Appl Stat*, **3**, 1335-1369.
20. Wang, H.T., Lo, S.H., Zheng, T. and Hu, I.C. (2012) Interaction-based feature selection and classification for high-dimensional biological data. *Bioinformatics*, **28**, 2834-2842.
21. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., Samani, N.J. *et al.* (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661-678.
22. McInnis, M.G., Dick, D.M., Willour, V.L., Avramopoulos, D., MacKinnon, D.F., Simpson, S.G., Potash, J.B., Edenberg, H.J., Bowman, E.S., McMahon, F.J. *et al.* (2003) Genome-wide scan and conditional analysis in bipolar disorder: Evidence for genomic interaction in the National Institute of Mental Health Genetics Initiative bipolar pedigrees. *Biol Psychiat*, **54**, 1265-1273.
23. Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J. *et al.* (2010) Quality Control and Quality Assurance in Genotypic Data for Genome-Wide Association Studies. *Genet Epidemiol*, **34**, 591-602.
24. Liu, Y., Xu, H.M., Chen, S.C., Chen, X.F., Zhang, Z.G., Zhu, Z.H., Qin, X.Y., Hu, L.D., Zhu, J., Zhao, G.P. *et al.* (2011) Genome-Wide Interaction-Based Association Analysis Identified Multiple New Susceptibility Loci for Common Diseases. *Plos Genet*, **7**.
25. Wang, J.B., Ronaghi, M., Chong, S.S. and Lee, C.G.L. (2011) pfSNP: An Integrated Potentially Functional SNP Resource That Facilitates Hypotheses Generation Through Knowledge Syntheses. *Human Mutation*, **32**, 19-24.
26. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, **35**, D61-65.
27. Lo, S.H., Chernoff, H., Cong, L., Ding, Y.J. and Zheng, T. (2008) Discovering interactions among BRCA1 and other candidate genes associated with sporadic breast cancer. *Proc Natl Acad Sci U S A*, **105**, 12387-12392.
28. Hu, T., Sinnott-Armstrong, N.A., Kiralis, J.W., Andrew, A.S., Karagas, M.R. and Moore, J.H. (2011) Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *Bmc Bioinformatics*, **12**.
29. Perlstein, M.D., Chohan, M.R., Coman, I.L., Antshel, K.M., Fremont, W.P., Gnirke, M.H., Kikinis, Z., Middleton, F.A., Radoeva, P.D., Shenton, M.E. *et al.* (2014) White matter abnormalities in 22q11.2 deletion syndrome: preliminary associations with the Nogo-66 receptor gene and symptoms of psychosis. *Schizophrenia research*, **152**, 117-123.
30. Hsu, R., Woodroffe, A., Lai, W.S., Cook, M.N., Mukai, J., Dunning, J.P., Swanson, D.J., Roos, J.L., Abecasis, G.R., Karayiorgou, M. *et al.* (2007) Nogo Receptor 1 (RTN4R) as a Candidate Gene for Schizophrenia: Analysis Using Human and Mouse Genetic Approaches. *PLoS ONE*, **2**.
31. Greaves, E., Collins, F., Esnal-Zufiaurre, A., Giakoumelou, S., Horne, A.W. and Saunders, P.T.K. (2014) Estrogen Receptor (ER) Agonists Differentially Regulate Neuroangiogenesis in Peritoneal Endometriosis via the Repellent Factor SLIT3. *Endocrinology*, **155**, 4015-4026.
32. Chen, T., Gai, W.P. and Abbott, C.A. (2014) Dipeptidyl Peptidase 10 (DPP10(789)): A Voltage Gated Potassium Channel Associated Protein Is Abnormally Expressed in Alzheimer's and Other Neurodegenerative Diseases. *Biomed Res Int*.
33. Nomoto, H., Yonezawa, T., Itoh, K., Ono, K., Yamamoto, K., Oohashi, T., Shiraga, F., Ohtsuki, H. and Ninomiya, Y. (2003) Molecular cloning of a novel transmembrane protein MOLT expressed by mature oligodendrocytes. *J Biochem*, **134**, 231-238.
34. Lim, S.H., Kwon, S.K., Lee, M.K., Moon, J., Jeong, D.G., Park, E., Kim, S.J., Park, B.C., Lee, S.C., Ryu, S.E. *et al.* (2009) Synapse formation regulated by protein tyrosine phosphatase receptor T through interaction with cell adhesion molecules and Fyn. *Embo J*, **28**, 3564-3578.
35. Fogel, B.L., Wexler, E., Wahnich, A., Friedrich, T., Vijayendran, C., Gao, F.Y., Parikshak, N., Konopka, G. and Geschwind, D.H. (2012) RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. *Hum Mol Genet*, **21**, 4171-4186.
36. Martin, C.L., Duvall, J.A., Ilkin, Y., Simon, J.S., Arreaza, M.G., WilkeS, K., Alvarez-Retuerto, A., Whichello, A., Powell, C.M., Rao, K. *et al.* (2007) Cytogenetic and molecular characterization of A2BP1/FOX1 as a candidate gene for autism. *Am J Med Genet B*, **144B**, 869-876.
37. Fleiss, J.L. (1981) *Statistical methods for rates and proportions*. 2d ed. Wiley, New York.
38. Fisher, R.A. (1954) *Statistical methods for research workers*. 12th ed. Oliver and Boyd, Edinburgh,.
39. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997-1004.
40. Smith, E.N., Koller, D.L., Panganiban, C., Szelinger, S., Zhang, P., Badner, J.A., Barrett, T.B., Berrettini, W.H., Bloss, C.S., Byerley, W. *et al.* (2011) Genome-wide association of bipolar disorder suggests an enrichment of replicable associations in regions near genes. *Plos Genet*, **7**, e1002134.
41. Budel, S., Padukkavidana, T., Liu, B.P., Feng, Z., Hu, F.H., Johnson, S., Lauren, J., Park, J.H., McGee, A.W., Liao, J. *et al.* (2008) Genetic Variants of Nogo-66 Receptor with Possible Association to Schizophrenia Block Myelin Inhibition of Axon Growth. *J Neurosci*, **28**, 13161-13172.

42. Karlsson, T.E., Karlen, A., Olson, L. and Josephson, A. (2013) Neuronal Overexpression of Nogo Receptor 1 in APPswe/PSEN1(Delta E9) Mice Impairs Spatial Cognition Tasks without Influencing Plaque Formation. *J Alzheimers Dis*, **33**, 145-155.
43. Sinibaldi, L., De Luca, A., Bellacchio, E., Conti, E., Pasini, A., Paloscia, C., Spalletta, G., Caltagirone, C., Pizzuti, A. and Dallapiccola, B. (2004) Mutations of the Nogo-66 receptor (RTN4R) gene in schizophrenia. *Hum Mutat*, **24**, 534-535.
44. Glessner, J.T., Wang, K., Sleiman, P.M.A., Zhang, H.T., Kim, C.E., Flory, J.H., Bradfield, J.P., Imielinski, M., Frackelton, E.C., Qiu, H.J. *et al.* (2010) Duplication of the SLIT3 Locus on 5q35.1 Predisposes to Major Depressive Disorder. *Plos One*, **5**.
45. Haaker, J., Lonsdorf, T.B., Raczka, K.A., Mechias, M.L., Gartmann, N. and Kalisch, R. (2014) Higher anxiety and larger amygdala volumes in carriers of a TMEM132D risk variant for panic disorder. *Transl Psychiat*, **4**.
46. Wolock, S.L., Yates, A., Petrill, S.A., Bohland, J.W., Blair, C., Li, N., Machiraju, R., Huang, K. and Bartlett, C.W. (2013) Gene X smoking interactions on human brain gene expression: finding common mechanisms in adolescents and adults. *J Child Psychol Psyc*, **54**, 1109-1119.
47. Li, W.Q., Zhang, Y.X., Gu, R.J., Zhang, P., Liang, F., Gu, J.P., Zhang, X.M., Zhang, H.Y. and Zhang, H.X. (2013) DNA Pooling Base Genome-Wide Association Study Identifies Variants at NRXN3 Associated with Delayed Encephalopathy after Acute Carbon Monoxide Poisoning. *Plos One*, **8**.

TABLES

Table 1. Power and type I error rates of alternative methods on pairwise epistasis effect

		MAF > 5%			1%< MAF <5%		
		LD			LD		
Model	Methods	Low	Medium	High	Low	Medium	High
Power (linear model)	Logistic	68.5%	76.9%	83.3%	47.1%	62.5%	71.1%
	Chi-squared	60.0%	67.2%	74.5%	42.2%	65.2%	74.0%
	W	71.1%	81.0%	86.7%	49.8%	79.5%	83.8%
Power (nonlinear model)	Logistic	5.9%	1.7%	0.6%	61.7%	31.8%	43.7%
	Chi-squared	72.6%	69.4%	62.8%	67.4%	43.9%	49.1%
	W	88.0%	86.6%	79.4%	95.6%	83.3%	83.9%
Type I Error Rate*	Logistic	3.2E-05	4.8E-05	3.2E-05	3.7E-05	4.3E-05	4.6E-05
	Chi-squared	2.3E-05	1.4E-05	2.5E-05	3.0E-06	2.0E-06	2.0E-06
	W	4.4E-05	4.9E-05	4.5E-05	3.3E-05	4.2E-05	5.5E-05

*Nominal type I error rate = type I error rate × 1,225 pairs

Table 2. Power and type I error rates of alternative methods at different sample sizes.

The simulation study is performed using a non-linear genetic model, 1%< MAF < 5%, and medium LD genetic architectures. As the sample size decreases, the W-test showed persistent better power and reasonable type I error rates.

	Sample size	300	400	500	600	700	800	900	1000
Power	Logistic	2.5%	4.0%	8.6%	12.4%	14.0%	17.7%	21.5%	29.1%
	Chi-squared	1.7%	2.2%	5.7%	18.9%	25.0%	37.8%	42.1%	44.7%
	W	16.0%	28.8%	38.5%	67.8%	72.8%	82.2%	83.2%	83.8%
Type I Error Rate	Logistic	4.1E-05	4.9E-05	3.9E-05	5.0E-05	4.4E-05	4.3E-05	4.6E-05	4.7E-05
	Chi-squared	2.0E-06	2.0E-06	1.0E-06	0	3.0E-06	4.0E-06	0	2.0E-06
	W	5.5E-05	4.9E-05	4.6E-05	4.6E-05	4.1E-05	4.4E-05	4.3E-05	4.2E-05

Table 3. Replicated bipolar disorder susceptible genes from two datasets

SNP	Gene	Position	MAF*	P-value of pair*
rs6741692	*DPP10*	2q14	0.303	5.8E-38
rs2407594	*CSMD1*	8p23	0.029	9.8E-36
rs1864952	*SLIT3*	5q35	0.046	1.9E-35
rs2849605	*PARK2*	6q5.2	0.021	3.3E-29
rs3867492	*TMEM132D*	12q24.33	0.030	1.0E-27
rs11222695	*HNT*	11q25	0.012	2.7E-25
rs1494451	*CNTNAP2*	7q35	0.025	1.3E-21
rs2785061	*ACCN1*	17q12	0.028	9.8E-19
rs17135053	*A2BP1*	16p13.3	0.025	3.9E-18
rs17170832	*ELMO1*	7p14.1	0.017	3.9E-18
rs9559408	*MYO16*	13q33.3	0.035	4.8E-17

***** MAF and p-value are presented using the WTCCC data. Detailed pair information can be found in Supplementary Information S6.

FIGURES LEGENDS

Diagram 1. Decomposition of the W-test.

The W-test measures the distributional differences between cases and controls using a combined log odds ratio. The dependency among the cells is handled by the data-dependent scalars h and f, estimated from the null hypothesis. The overall test statistic follows a Chi-squared distribution with f degrees of freedom.

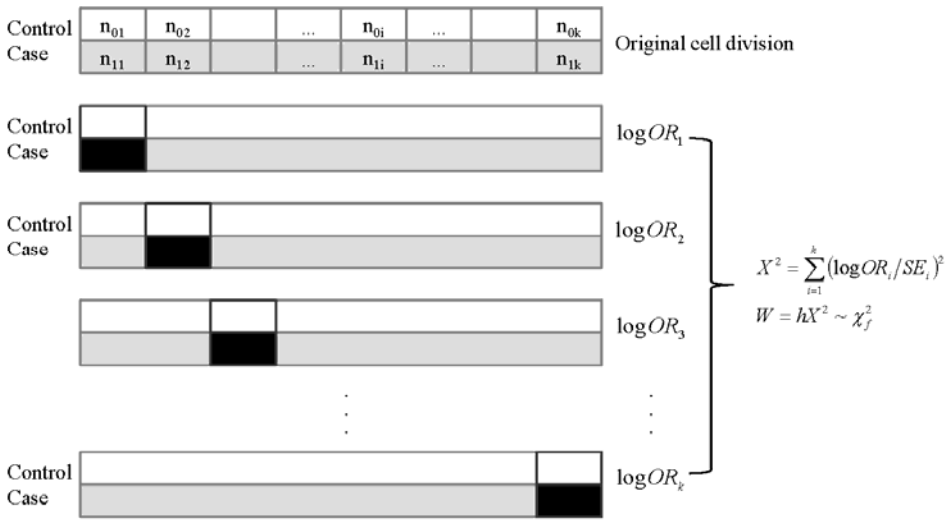

Figure 1. Power of alternative methods in low frequency variant environment

In the low frequency variant environment (1%<MAF<5%), the W-test outperforms alternative methods for both linear and non-linear models.

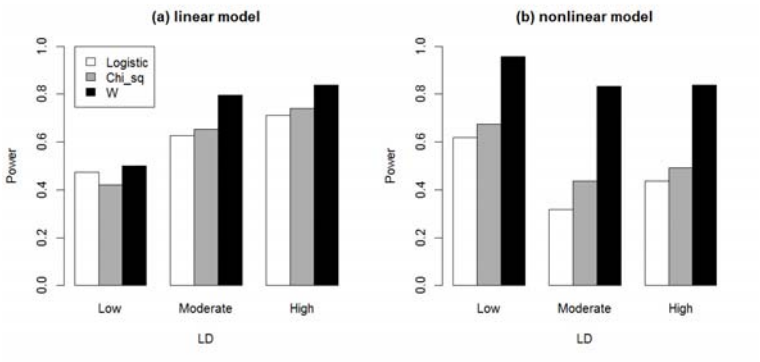

Figure 2. Power comparison of alternative methods at different sample sizes.

As the sample size reduces, the W-test shows a robust power compared to alternative methods. The power is calculated under the genetic environment of 1% <MAF< 5% and LD 20% < r^2< 80%, using a non-linear model.

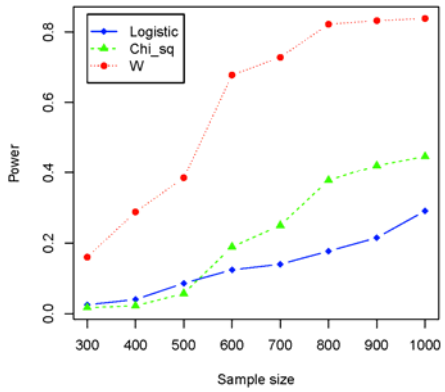

Figure 3. Q-Q Plot of W-test on real genome-wide data.

The W-test is computed on real genome-wide data with permuted phenotype for SNP-SNP interactions. No inflation of spurious association is observed.

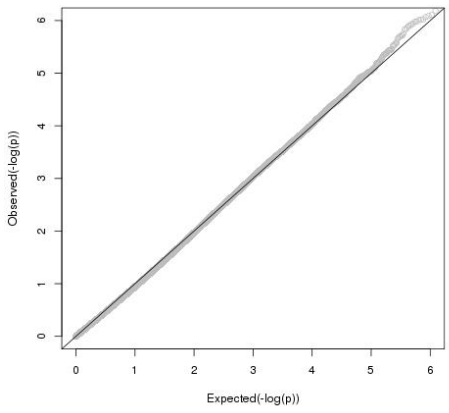

Figure 4. Gene-gene Interaction Networks.

The solid lines represent significant epistasis effect. Blue color indicates pairs found in the GAIN dataset and red color indicates that they are identified in the WTCCC dataset. Purple circles represent genes replicated by the two independent data; all of which play important roles in brain and neuronal function.