

# Testing Hardy–Weinberg disequilibrium using the generalized linear model

SHIZHONG XU

Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA

(Received 11 August 2012; revised 31 October 2012; accepted 12 November 2012; first published online 18 December 2012)

## Summary

Current methods for detecting Hardy–Weinberg disequilibrium (HWD) only deal with one locus at a time. We developed a method that can jointly detect HWD for multiple loci. The method was developed under the generalized linear model (GLM) using the probit link function. When applied to a single locus, the new method is more powerful than the exact test. When applied to two or more loci, the method can reduce false positives caused by linkage disequilibrium (LD). We applied the method to 24 single nucleotide polymorphism (SNP) markers of a single human gene and eliminated several false positive HWDs due to LD. We developed an R package ‘hwdglm’ for joint HWD detection, which can be downloaded from our personal website ([www.statgen.ucr.edu](http://www.statgen.ucr.edu)).

## 1. Introduction

Hardy–Weinberg equilibrium (HWE) is an important phenomenon in population genetics and it has been a subject of intense consideration and a powerful research tool in population genetics (Crow, 1988). Deviation from HWE is called Hardy–Weinberg disequilibrium (HWD), which may be generated by many different evolutionary forces, including non-random mating, selection, genetic drift and so on. Testing HWD can help detection of such evolutionary forces (Weir, 1996; Deng *et al.*, 2001). In addition, HWD may also be introduced by genotyping errors (Lee, 2003), which is a factor responsible for low power of association studies. A very important application of HWD is in case–control disease association studies (Chen & Chatterjee, 2007) or case-only association studies (Nielsen *et al.*, 1999; Lee, 2003; Weinberg & Morris, 2003). In case–control studies, genotype frequencies of the case population are compared with HWE predicted genotype frequencies from the control population. A significant deviation of a marker from HWD implies the association of the locus with the disease. In case-only studies, a significant HWD of a marker within the case population implies marker-disease association if the entire population is assumed to be in HWE. A more detailed

review of application of HWD to genetic association studies can be found in Salanti *et al.* (2005).

Many statistical methods have been developed to test HWE, including the  $\chi^2$  test (Li, 1955; Hernandez & Weir, 1989) and the exact test (Fisher, 1935; Haldane, 1954; Chapco, 1976; Guo & Thompson, 1992; Rohlf & Weir, 2008). Bayesian statistics have also been applied to HWD analysis by drawing posterior credibility of the estimated HWD parameter (Ayres & Balding, 1998; Shoemaker *et al.*, 1998). These methods have been substantially validated and applied to population genetics. They can be summarized in two general categories: (1) a parametric hypothesis test and (2) an exact test. In the parametric hypothesis test, there is a HWD parameter to estimate and a test is performed on this parameter. The most common parameters are the inbreeding coefficient ( $f$ ) for two alleles, the fixation index ( $f_{ij}$ ) for multiple alleles, the additive HWD parameter ( $D$ ) for two alleles and the corresponding multiple allelic HWD parameter ( $D_{ij}$ ). Please see Weir (1996) for details of these parameters. There are two versions of the exact test. The first version remains in its original form of the exact test (Fisher, 1935; Guo & Thompson, 1992), in which the proportion of the extreme samples was counted as the  $P$ -values. The second version of the exact test requires a parametric estimation of the HWD parameter and a test statistic. All possible samples from the hypergeometric distribution of

E-mail: shizhong.xu@ucr.edu

genotype counts under the null model are evaluated. The test statistics are then ranked to form a null distribution (Rousset & Raymond, 1995). Different test statistics may generate different ranks and thus have different  $P$ -values for the same data.

It appears that existing methods are already mature and do not have much room to improve. However, one aspect has been ignored in all HWD studies, which is a conditional test for one locus given the genotypes of other loci. If the multiple loci are in low linkage disequilibrium (LD), conditional tests would not offer any advantage over the individual locus tests (one locus at a time). However, if these loci are in high LD, conditional tests will lead to different results. A locus in HWD detected separately may not be truly in HWD, but caused by the HWD of another locus in high LD with the one of interest. This Type I error has not been recognized in the literature. None of the existing methods have addressed this problem. Some of the pooling methods deal with multiple loci (Hill *et al.*, 1995; Ayres & Balding, 1998). However, these methods were based on the assumption of same inbreeding coefficient for all loci and thus gained power when different loci are combined. There has been no report in the literature to jointly estimate HWD for multiple loci.

One approach to tackling this problem is to extend the contingency table to cover multiple loci. The dimension of the table would rise rapidly as the number of loci increases. This may explain why people choose to avoid HWD testing for multiple loci. It is feasible to test HWD for a pair of loci simultaneously, but the method may not be easily extended to multiple loci. Here, we proposed an entirely different approach to handle multiple locus HWD test, a generalized linear model (GLM) approach. The GLM is a well-known model for analysing discrete data (McCullagh & Nelder, 1989). Connection of the HWD problem to GLM is not obvious. If we treat the genotype or allele count as the response variable, there are no predictors. If we treat the genotype counts as predictors, there is no response variable. We previously developed a GLM for testing multiple locus segregation distortion (Zhan & Xu, 2011). In that model, we proposed a hidden linear predictor, called liability. The response variables are the genotype counts. In this study, we adopted that model to perform HWD test. The key difference between the GLM and the existing methods is that we can use a linear predictor to measure the strength of HWD. With the proposed linear predictor, we can handle multiple loci easily by combining all locus-specific HWD parameters into a single linear predictor. Parameter estimation and hypothesis test can be performed under the GLM framework.

We first introduced the GLM for a single locus. We then extended the model to conditional analysis for

two loci. Finally, we addressed conditional tests for multiple loci. Simulation experiments were performed to validate the GLM for the single locus and two loci analyses. A real single nucleotide polymorphism (SNP) dataset was used to demonstrate the multiple locus conditional tests.

## 2. Theory and method

### (i) Model for single locus

Let  $A_1$  and  $A_2$  be two alleles in a biallelic population and  $A_1A_1$ ,  $A_1A_2$  and  $A_2A_2$  be the three possible genotypes. Denote the observed counts of the three genotypes by  $n_{11}$ ,  $n_{12}$  and  $n_{22}$ . The estimated frequencies of the two alleles are  $p_1 = (2n_{11} + n_{12})/(2n)$  and  $p_2 = (2n_{22} + n_{12})/(2n)$ , respectively, where  $n = n_{11} + n_{12} + n_{22}$  is the sample size. Under HWE, the predicted genotypic frequencies are  $\varphi_{11} = p_1^2$ ,  $\varphi_{12} = 2p_1p_2$  and  $\varphi_{22} = p_2^2$ , respectively. Rather than using the classical heterozygosity reduction index  $f$  or the disequilibrium  $D$  as the parameter to measure the amount of departure from HWE, we proposed a new parameter for HWD. This parameter takes any real number without the awkward constraint as that in the  $D$  parameter. The new parameter can be derived based on the selection theory, in which the heterozygosity deficiency (or excess) is formulated as selection against (or in favour of) heterozygote. Let  $\psi_{11} = \psi_{22}$  be the fitness of the two homozygotes and  $\psi_{12}$  be the fitness of the heterozygote. The relative fitness of the heterozygote over the homozygotes is  $\psi_{12}/\psi_{11}$ . This relative fitness of a non-unity value represents the departure from HWE. This is the dominance model of fitness selection.

We now reparameterize the fitness using  $\psi_{11} = \psi_{22} = \Phi(0) = 1/2$  and  $\psi_{12} = \Phi(\theta)$ , where  $\Phi(\cdot)$  is the standardized cumulative normal distribution and  $\theta$  is the new HWD parameter. If  $\theta = 0$ , then  $\psi_{12} = \Phi(0) = 1/2$  and the relative fitness of the heterozygote is 1 (compared with that of the homozygotes) and the population is in HWE. If  $\theta < 0$ , selection is against the heterozygote while  $\theta > 0$  means selection against homozygotes. Note that HWD may be caused by many other factors, including inbreeding, genetic drift and population subdivision. The selection parameter  $\theta$  also indicates heterozygote deficiency if  $\theta < 0$  and heterozygote excess if  $\theta > 0$ . We can estimate and test  $\theta$  as the parameter of departure from HWE. The null hypothesis is  $H_0: \theta = 0$ . We can see that  $-\infty < \theta < +\infty$  and thus there is no constraint on this parameter. We could have set  $\psi_{11} = \psi_{22} = 1$ , but parameter  $\theta$  under this set up would not be zero under HWE. It is equally valid to define  $\psi_{12} = \Phi(-\theta)$  so that  $\theta > 0$  represents heterozygosity deficiency. More discussion on the relative fitness is given in the last section of the manuscript.

Using the Bayes theorem, we can find the probabilities of the three genotypes, as shown below:

$$\begin{aligned} \pi(11) &= \frac{\varphi_{11}\Phi(0)}{\varphi_{11}\Phi(0) + \varphi_{12}\Phi(\theta) + \varphi_{22}\Phi(0)}, \\ \pi(12) &= \frac{\varphi_{12}\Phi(\theta)}{\varphi_{11}\Phi(0) + \varphi_{12}\Phi(\theta) + \varphi_{22}\Phi(0)}, \\ \pi(22) &= \frac{\varphi_{22}\Phi(0)}{\varphi_{11}\Phi(0) + \varphi_{12}\Phi(\theta) + \varphi_{22}\Phi(0)}. \end{aligned} \tag{1}$$

These probabilities are considered the posterior probabilities of the three genotypes under the Bayesian framework. The HWE predicted probabilities serve as the prior probabilities. The fitness of each genotype then serves as the likelihood. One can easily see that, when  $\theta=0$ , the posterior probabilities converge to the HWE probabilities.

The log-likelihood function is constructed using the multinomial distribution of the data, as shown by

$$L(\theta) = n_{11} \ln[\pi(11)] + n_{12} \ln[\pi(12)] + n_{22} \ln[\pi(22)]. \tag{2}$$

The derivative of  $L(\theta)$  with respect to  $\theta$  is

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{n_{12}\phi(\theta)}{\Phi(\theta)} - \frac{2(n_{11} + n_{12} + n_{22})\varphi_{12}\phi(\theta)}{\varphi_{11} + 2\varphi_{12}\Phi(\theta) + \varphi_{22}}, \tag{3}$$

where  $\phi()$  is the standardized normal density. Setting the derivative to zero and solving for  $\theta$  leads to the maximum likelihood estimate (MLE) of the parameter,

$$\hat{\theta} = \Phi^{-1} \left[ 0.5 \times \frac{n_{12}(\varphi_{22} + \varphi_{11})}{\varphi_{12}(n_{11} + n_{22})} \right]. \tag{4}$$

The variance of the estimate is approximated by the inverse of the information,

$$\text{var}(\hat{\theta}) \approx I^{-1}(\hat{\theta}) = \frac{\Phi(\hat{\theta}) \left( \varphi_{11} + 2\varphi_{12}\Phi(\hat{\theta}) + \varphi_{22} \right)^2}{2n\varphi_{12}(\varphi_{11} + \varphi_{22})\phi^2(\hat{\theta})}. \tag{5}$$

Significance test for  $H_0:\theta=0$  may be performed in two ways, the likelihood ratio test (LRT) and the Wald test. The former is defined as

$$\text{LRT} = -2 \left[ L(0) - L(\hat{\theta}) \right]. \tag{6}$$

The Wald test is

$$\text{Wald} = \frac{\hat{\theta}^2}{\text{var}(\hat{\theta})}. \tag{7}$$

Under the null hypothesis, both statistics asymptotically follow the  $\chi^2_1$  distribution and thus the critical value  $\chi^2_{1,0.95} = 3.84$  can be used to declare significance at the  $\alpha=0.05$  Type I error rate. One caveat of the Wald test is that if the genotype counts extremely deviate from HWE, the second derivative of  $L(\theta)$  with respect to  $\theta$  may not be defined, leading to no estimate of the variance. This problem does not apply to the

LRT because it does not require the second derivative. We call this model the GLM because the HWD parameter  $\theta$  is a linear predictor and the link function is probit. The novelty of the model is that we used a linear predictor  $\theta$  to measure the strength of HWD.

This paragraph shows an example using the proposed GLM and compared it with the Fisher’s exact test. Let the counts of the three genotypes be  $n_{11} = 24$ ,  $n_{12} = 39$  and  $n_{22} = 37$  with a sample size  $n = 100$ . The genotypic frequencies predicted under HWE are  $\varphi_{11} = 0.1892$ ,  $\varphi_{12} = 0.4916$  and  $\varphi_{22} = 0.3192$ . The GLM estimated HWD parameter is  $\hat{\theta} = -0.4380$  with an estimation error  $\sqrt{\text{var}(\hat{\theta})} = \sqrt{0.03499} = 0.1870$ . The Wald and LRT statistics are  $\text{Wald} = 5.4844$  and  $\text{LRT} = 4.1467$ , respectively. The corresponding  $P$ -values for the two tests are 0.0192 and 0.0413, respectively. Using Fisher’s exact test, the  $P$ -value is 0.0423, close to the  $P$ -value of the LRT.

(ii) Model for two loci

The single locus model is just an alternative way to test HWD. The main goal of proposing this new test statistic is to facilitate a conditional test for each of two loci that are in LD. If two loci are in linkage equilibrium, there is no advantage for the conditional test. However, if the two loci are in high LD, the HWD of one locus may be caused by the HWD of the other locus. The conditional test may correct the confounding. It is easy to use an indicator variable for the genotype of each locus for each individual. Let  $A_1A_1$ ,  $A_1A_2$  and  $A_2A_2$  be the three ordered genotypes for the first locus. The genotype indicator variable for individual  $j$  is defined by a  $1 \times 3$  vector  $X_j$ . For example, if individual  $j$  is of type  $A_1A_1$ , then  $X_j = [1 \ 0 \ 0]$ . Let  $B_1B_1$ ,  $B_1B_2$  and  $B_2B_2$  be the three ordered genotypes for the second locus. The genotype indicator variable of individual  $j$  for the second locus is defined by a  $1 \times 3$  vector  $Y_j$ . For example, if individual  $j$  is of type  $B_1B_2$  for the second locus, then  $Y_j = [0 \ 1 \ 0]$ . We now use  $\theta_X$  and  $\theta_Y$ , respectively, to denote the HWD parameters for the two loci. Let us define  $\delta = [0 \ 1 \ 0]^T$  as a  $3 \times 1$  vector of constants. We can see that  $Y_j\delta = 1$  if  $j$  is heterozygote for the second locus and  $Y_j\delta = 0$  otherwise. Conditional on  $\theta_Y$  for the second locus, the posterior probabilities of the three genotypes of individual  $j$  for the first locus are

$$\begin{aligned} \pi_j(11) &= \frac{\varphi_{11}\Phi(Y_j\delta\theta_Y)}{\varphi_{11}\Phi(Y_j\delta\theta_Y) + \varphi_{12}\Phi(\theta_X + Y_j\delta\theta_Y) + \varphi_{22}\Phi(Y_j\delta\theta_Y)}, \\ \pi_j(12) &= \frac{\varphi_{12}\Phi(\theta_X + Y_j\delta\theta_Y)}{\varphi_{11}\Phi(Y_j\delta\theta_Y) + \varphi_{12}\Phi(\theta_X + Y_j\delta\theta_Y) + \varphi_{22}\Phi(Y_j\delta\theta_Y)}, \\ \pi_j(22) &= \frac{\varphi_{22}\Phi(0 + Y_j\delta\theta_Y)}{\varphi_{11}\Phi(Y_j\delta\theta_Y) + \varphi_{12}\Phi(\theta_X + Y_j\delta\theta_Y) + \varphi_{22}\Phi(Y_j\delta\theta_Y)}. \end{aligned} \tag{8}$$

Let  $\pi_j = [\pi_j(11) \ \pi_j(12) \ \pi_j(22)]^T$  and thus  $\ln(\pi_j)$  is a  $3 \times 1$  column vector as shown below:

$$\ln(\pi_j) = [\ln \pi_j(11) \ \ln \pi_j(12) \ \ln \pi_j(22)]^T. \tag{9}$$

We define the conditional log likelihood function for  $\theta_X$  given  $\theta_Y$  by

$$L(\theta_X|\theta_Y) = \sum_{j=1}^n X_j \ln(\pi_j). \tag{10}$$

Maximizing this likelihood function and finding the solution of  $\theta_X$  gives the conditional MLE of  $\theta_X$ . Unfortunately, an explicit solution for the MLE of  $\theta_X$  is hard to find. A numerical solution may have to be resorted to e.g. the Newton method, which has the following iteration form:

$$\theta_X^{(t+1)} = \theta_X^{(t)} - \frac{L'(\theta_X^{(t)}|\theta_Y)}{L''(\theta_X^{(t)}|\theta_Y)}, \tag{11}$$

where  $L'(\theta_X^{(t)})$  and  $L''(\theta_X^{(t)})$  are the first- and second-order derivatives of the log-likelihood function with respect to  $\theta_X$  evaluated at  $\theta_X = \theta_X^{(t)}$ . When the iteration process converges, we get the MLE of  $\theta_X$  conditional on  $\theta_Y$ , denoted by  $\hat{\theta}_X$ , whose variance can be approximated by

$$\text{var}(\hat{\theta}_X) \approx -\frac{1}{L''(\hat{\theta}_X|\theta_Y)}. \tag{12}$$

Once  $\theta_X$  is estimated, we construct the likelihood function for  $\theta_Y$  conditional on  $\theta_X$  using the same approach. Define the posterior probabilities of the three genotypes for the second locus by

$$\begin{aligned} \xi_j(11) &= \frac{\omega_{11}\Phi(X_j\delta\theta_X)}{\omega_{11}\Phi(X_j\delta\theta_X) + \omega_{12}\Phi(\theta_Y + X_j\delta\theta_X) + \omega_{22}\Phi(X_j\delta\theta_X)}, \\ \xi_j(12) &= \frac{\omega_{12}\Phi(\theta_Y + X_j\delta\theta_X)}{\omega_{11}\Phi(X_j\delta\theta_X) + \omega_{12}\Phi(\theta_Y + X_j\delta\theta_X) + \omega_{22}\Phi(X_j\delta\theta_X)}, \\ \xi_j(22) &= \frac{\omega_{22}\Phi(X_j\delta\theta_X)}{\omega_{11}\Phi(X_j\delta\theta_X) + \omega_{12}\Phi(\theta_Y + X_j\delta\theta_X) + \omega_{22}\Phi(X_j\delta\theta_X)}, \end{aligned} \tag{13}$$

where  $\omega_{11} = q_1^2$ ,  $\omega_{12} = 2q_1q_2$  and  $\omega_{22} = q_2^2$  are the HWE predicted genotype frequencies for the second locus. The conditional log likelihood function for  $\theta_Y$  is

$$L(\theta_Y|\theta_X) = \sum_{j=1}^n Y_j \ln(\xi_j), \tag{14}$$

where

$$\ln(\xi_j) = [\ln \xi_j(11) \ \ln \xi_j(12) \ \ln \xi_j(22)]^T. \tag{15}$$

Conditional on  $\theta_Y$  to estimate  $\theta_X$  and then conditional on  $\theta_X$  to estimate  $\theta_Y$  will complete just one cycle of iterations. The Newton iteration process continues until the sequence converges. After convergence, we obtain the MLE of both parameters, denoted by  $\hat{\theta}_X$  and  $\hat{\theta}_Y$  for the two loci.

Again, both the likelihood ratio and the Wald test statistics are used to perform the HWD tests. The LRT statistics are

$$L_{X|Y} = -2[L(0|\hat{\theta}_Y) - L(\hat{\theta}_X|\hat{\theta}_Y)] \tag{16}$$

and

$$L_{Y|X} = -2[L(0|\hat{\theta}_X) - L(\hat{\theta}_Y|\hat{\theta}_X)] \tag{17}$$

for the two loci. The corresponding Wald tests are

$$W_{X|Y} = \frac{\hat{\theta}_X^2}{\text{var}(\hat{\theta}_X)} \tag{18}$$

and

$$W_{Y|X} = \frac{\hat{\theta}_Y^2}{\text{var}(\hat{\theta}_Y)}. \tag{19}$$

We now show the result of an example using the GLM method. The joint genotype counts of the two loci ( $X$  and  $Y$ ) of the example are given below,

Locus $X$	Locus $Y$			$N_X$
	$B_1B_1$	$B_1B_2$	$B_2B_2$	
$A_1A_1$	27	6	1	34
$A_1A_2$	3	23	6	32
$A_2A_2$	1	10	23	34
$N_Y$	31	39	30	100

We performed a contingency table association test for the data. The  $\chi^2$  test statistic for the association is 78.52 with a  $P$ -value of less than 0.0001, showing strong LD between the two loci. The marginal tests of HWD for individual loci showed that  $\hat{\theta}_X = -0.7215 \pm 0.1640$  and  $\hat{\theta}_Y = -0.4684 \pm 0.1833$  for the two loci. The corresponding  $P$ -values are 0.00001 for locus  $X$  and 0.01064 for locus  $Y$ , meaning that both loci deviate from HWE. The conditional tests, however, showed that  $\hat{\theta}_X = -0.6294 \pm 0.1630$  and  $\hat{\theta}_Y = -0.2489 \pm 0.1850$  for the two loci. The corresponding  $P$ -values are 0.00011 for locus  $X$  and 0.1785 for locus  $Y$ . The conclusion is that the marginal HWD for locus  $Y$  is actually caused by the HWD of locus  $X$  due to LD.

We now show another example where the two loci are in linkage equilibrium. The frequency table is

Locus $X$	Locus $Y$			$N_X$
	$B_1B_1$	$B_1B_2$	$B_2B_2$	
$A_1A_1$	10	12	12	34
$A_1A_2$	10	12	10	32
$A_2A_2$	12	13	9	34
$N_Y$	32	37	31	100

The  $\chi^2$  test statistic for LD between  $X$  and  $Y$  is 0.6629 with a very large  $P$ -value of 0.9558, i.e. the LD is not significant. The marginal HWD analysis for individual loci produced  $\hat{\theta}_X = -0.7215 \pm 0.1640$  and  $\hat{\theta}_Y = -0.5426 \pm 0.1766$  for the two loci. The corresponding  $P$ -values are 0.0000109 for locus  $X$  and 0.00213 for locus  $Y$ , meaning that both loci deviate from HWE. The conditional analysis showed that  $\hat{\theta}_X = -0.6471 \pm 0.1544$  and  $\hat{\theta}_Y = -0.4632 \pm 0.1610$  for the two loci. The corresponding  $P$ -values are 0.0000278 for locus  $X$  and 0.00403 for locus  $Y$ . The marginal tests and conditional tests gave the same conclusion that both loci deviate from HWE.

(iii) *Model for multiple loci*

Extension to multiple loci is straightforward except that the notation becomes a little bit more complicated. We will not provide the detail of the multiple locus model, but only point out the key step of the extension. Recall that the conditional posterior probabilities for the three genotypes of the first locus given the HWD parameter of the second locus requires an additional term in the linear predictor,  $Y_j \delta \theta_j$ , as shown in eqn (8). For multiple loci, this additional term is replaced by a term called offset, which includes effects of all other loci except the current one of interest. Suppose that we have  $m$  loci for the joint analysis. When the  $k$ th locus is considered given the effects of all other loci, the relative fitness for this locus are individual specific,  $\psi_{11} = \psi_{22} = \Phi(o_j)$  and  $\psi_{12} = \Phi(\theta_k + o_j)$ , where

$$o_j = \sum_{k' \neq k}^m X_{jk'} \delta \theta_{k'} \tag{20}$$

is the offset and  $X_{jk'}$  is the genotype indicator vector ( $1 \times 3$ ) for individual  $j$  at locus  $k'$ .

Theoretically, one can perform such a conditional test for as many loci as desired. There is a computational issue when the number of loci is too large. Another issue is the model identifiability problem. Since the  $\theta$  parameter is a linear predictor (a first moment parameter), we can easily assign each  $\theta$  a normal prior distribution, which makes the model a generalized linear mixed model (GLMM). The GLMM may be considered as a penalized GLM. If the variance of the normal prior is further assigned a hyper prior, the model will then become the Bayesian hierarchical GLM. If the number of loci is extremely large, we must take the penalized model using the GLMM approach, which deserves further investigation.

**3. Simulation studies**

All analyses were conducted using an R package ‘hwdglm’ developed in this study. The package can be

downloaded from our personal website: www.stat-gene.ucr.edu. There are three functions within the package: (1) hwd1.glm() for single locus HWD detection; (2) hwd2.glm() for joint detection of HWD for two loci; and (3) hwdm.glm() for joint detection of HWD for  $m$  loci ( $m > 1$ ).

(i) *Model for single locus*

This simulation experiment aimed to validate the new model and compare the new model with the exact test. The following factors were considered in the simulation: (i) minor allele frequency ( $p_1$ ), (ii) sample size ( $n$ ) and (iii) strength of HWD. We used the traditional inbreeding coefficient  $f$  as the degree of HWD to generate the data, although we never tried to estimate  $f$ . For the Fisher’s exact test, we presented the estimated  $D$  and the  $P$ -value. The HWE.test () function of the R package ‘genetics’ was used for the exact test. This method is labelled ‘EXACT’. For the new method, we used the hwd1.glm() function of the ‘hwdglm’ package to perform the analysis. We reported the estimated  $\theta$  and its estimation error. In addition, we also reported the  $P$ -values drawn from the LRT and the Wald test. The new method is labelled ‘GLM’.

The data were generated under the following multinomial distribution:

$$\Pr(n_{11}, n_{12}, n_{22}) = \frac{n!}{n_{11}!n_{12}!n_{22}!} (p_1^2 + p_1 p_2 f)^{n_{11}} \times (2p_1 p_2 - 2p_1 p_2 f)^{n_{12}} (p_2^2 + p_1 p_2 f)^{n_{22}}, \tag{21}$$

where  $p_1$  is the minor allele frequency and  $p_2 = 1 - p_1$  is the major allele frequency. This model was used to generate the data and is not the model for estimating the HWD parameter.

The minor allele frequency was examined at five levels: 0.1, 0.2, 0.3, 0.4 and 0.5. The sample size was set at the following levels: 50, 100, 200, 300 and 500. The  $f$  value (representing the degree of HWD) was examined at the following levels: 0.0, 0.1, 0.2, 0.3, 0.4 and 0.5. The situation of  $f=0$  represents the null model, which was simulated to examine the Type I errors of various methods compared. The simulation experiments with  $f>0$  were used to examine the statistical powers. The total number of combinations of the three factors is  $5 \times 5 \times 6 = 150$ . Combining with the two methods, the simulation would generate too many data points for presentation. Therefore, we used a subset of the 150 cases to perform the simulations and drew general conclusions about the new method. We did not expect any better performance of the new method than the exact method. We would be satisfied as long as the performance of the new method is not too much worse than the existing method. That would

Table 1. Empirical Type I errors for the new method compared with the exact test obtained from a simulation experiment with 1000 replicates

Allele frequency	Sample size	Method		
		GLM (Wald)	GLM (LRT)	EXACT
$p_1$	$n$			
0.1	50	0.002	0.001	0.018
	100	0.001	0.001	0.019
	200	0.003	0.001	0.030
	300	0.001	0.001	0.038
	500	0.000	0.000	0.038
0.2	50	0.010	0.002	0.017
	100	0.012	0.006	0.036
	200	0.009	0.007	0.040
	300	0.004	0.001	0.032
	500	0.005	0.004	0.039
0.3	50	0.018	0.015	0.039
	100	0.022	0.012	0.031
	200	0.015	0.023	0.041
	300	0.015	0.019	0.041
	500	0.021	0.022	0.041
0.4	50	0.028	0.034	0.032
	100	0.036	0.028	0.040
	200	0.037	0.045	0.047
	300	0.040	0.049	0.049
	500	0.036	0.040	0.040
0.5	50	0.038	0.053	0.034
	100	0.044	0.039	0.055
	200	0.033	0.039	0.044
	300	0.047	0.055	0.057
	500	0.039	0.051	0.047

Table 2. Empirical critical values from  $\chi^2_1$  distribution to achieve 0.05 Type I error rate for the new method compared with the exact method obtained from a simulation study with 1000 replicates

Allele frequency	Sample size	Method		
		GLM (Wald)	GLM (LRT)	EXACT
$p_1$	$n$			
0.1	50	1.0090	0.8406	2.3404
	100	0.8535	0.8556	2.9449
	200	0.6936	0.7052	2.7447
	300	0.8448	0.8354	3.1618
	500	0.9022	0.8883	3.3160
0.2	50	1.9415	1.5638	3.0088
	100	1.5650	1.5660	3.0255
	200	1.7218	1.7745	3.3791
	300	1.6743	1.7102	3.3955
	500	1.8567	1.8739	3.8098
0.3	50	2.7500	2.7019	3.1727
	100	2.0475	2.2982	2.9886
	200	2.5307	2.7923	3.6152
	300	2.6805	2.9328	3.8471
	500	2.6354	2.8227	3.6957
0.4	50	2.6522	3.7519	3.4703
	100	3.5930	3.2863	3.6168
	200	3.1476	3.2818	3.4325
	300	3.4330	3.6250	3.7051
	500	3.4289	3.6253	3.8049
0.5	50	3.2831	3.9282	3.3527
	100	3.2342	3.5309	3.9437
	200	3.2684	3.5344	3.3926
	300	3.9445	4.0755	3.8675
	500	3.6144	3.5499	3.5174

make the new method a reasonable substitution for the exact method so that extension to two or more loci would be well justified.

Under each scenario, the simulation was replicated 1000 times. For each replicate, a *P*-value less than a nominal 0.05 indicated a detection of HWD. Under the null model, the proportion of replicates in which HWD was detected represents the empirical Type I error. If this Type I error is around 0.05, we say that the Type I error is under our control. Otherwise, the test is either too conservative or too liberal. Under the alternative models ( $f > 0$ ), the proportion of replicates with significant HWD detection becomes the statistical power.

Table 1 shows the empirical Type I error when  $\chi^2_{0.95,1} = 3.84$  was used as the critical value for the test. Both the Wald test and the LRT under the GLM are over conservative when the minor allele frequency is small. However, when the minor allele frequency reaches 0.4, the empirical Type I error is close to the theoretical nominal value of 0.05. The Fisher's exact test is slightly conservative when the minor allele frequency is small. In conclusion, the GLM is more conservative than the exact test. In practice, if one uses  $\chi^2_{0.95,1} = 3.84$  as the critical value for the test, the

result will be over-conservative, which implies a low power. Wigginton *et al.* (2005) stated that even in a sample size of 1000, the actual type I error rates for both goodness-of-fit tests and exact tests may be much different from the nominal values. Our conclusion is consistent with Wigginton *et al.* (2005). Hernandez & Weir (1989) also observed the same conservativeness.

To control the Type I error at the claimed 0.05 level, one would need an empirical critical value less than  $\chi^2_{0.95,1} = 3.84$ . Hernandez & Weir (1989) suggested to use  $\chi^2_{0.95,0.5} = 2 \times 4$  as the critical value. We decided to use the  $\chi^2_1$  distribution to draw the critical value. We performed additional simulations to draw the empirical critical values for the test statistics (Wald and LRT). The results are listed in Table 2. The critical values are clearly less than  $\chi^2_{0.95,1} = 3.84$  when the minor allele frequency is small. For the Fisher's exact test, we should also correct the Type I error at low minor allele frequencies.

We now use these empirical critical values (Table 2) to examine the powers under various scenarios. We examined the power using  $f = 0.2$  as an example to show the powers of the new method. The empirical

Table 3. Empirical statistical power of HWD detection for the new method compared with the exact method ( $f=0.2$ )

Allele frequency	Sample size	Method		
		GLM (Wald)	GLM (LRT)	EXACT
$p_1$	$n$			
0.1	50	0.348	0.348	0.327
	100	0.493	0.479	0.481
	200	0.738	0.726	0.740
	300	0.872	0.861	0.863
	500	0.963	0.959	0.966
0.2	50	0.365	0.339	0.313
	100	0.580	0.533	0.543
	200	0.823	0.784	0.781
	300	0.932	0.915	0.914
	500	0.995	0.993	0.990
0.3	50	0.335	0.256	0.246
	100	0.631	0.563	0.550
	200	0.855	0.780	0.790
	300	0.941	0.914	0.913
	500	0.992	0.989	0.990
0.4	50	0.453	0.279	0.286
	100	0.589	0.542	0.520
	200	0.866	0.811	0.806
	300	0.938	0.915	0.917
	500	0.992	0.989	0.989
0.5	50	0.430	0.253	0.295
	100	0.626	0.521	0.473
	200	0.885	0.824	0.826
	300	0.945	0.918	0.920
	500	0.997	0.997	0.997

powers are given in Table 3. First, the LRT test of the GLM is similar to the exact test in all situations. When the minor allele frequency is low, the power of the Wald test is also close to that of the exact test. However, as the minor allele frequency reaches 0.4 or above, the Wald test is more powerful than both the LRT test and the exact test, especially when the sample size is small. This observation is unexpected. We proposed this new GLM method as a suitable substitute for the exact test and never expected it to outperform the exact test. Similar trends were also observed for other  $f$  values (see supplemental Tables S1–S4 for the power comparisons under other  $f$  values).

#### (ii) Model for two loci

Although the two loci model does not depend on whether LD is gametic or zygotic, for simplicity of simulation, we only considered gametic LD between two loci. Let the first locus ( $X$ ) be the primary locus with HWD and the second locus ( $Y$ ) be the secondary locus in HWE. The two loci have LD measured by  $r$ , the correlation between the indicator variables of allele  $A_1$  of the first locus and allele  $B_1$  of the second

locus. The genotype transition matrix from the first locus to the second locus is

$$T_{XY} = \frac{1}{4} \begin{bmatrix} (1+r)^2 & 2(1-r^2) & (1-r)^2 \\ 1-r^2 & (1+r)^2 + (1-r)^2 & 1-r^2 \\ (1-r)^2 & 2(1-r^2) & (1+r)^2 \end{bmatrix}. \quad (22)$$

For example, the conditional probability of  $B_1B_1$  given  $A_1A_2$  takes the element of the second row and the first column of the above transition matrix,

$$\Pr(B_1B_1|A_1A_2) = T_{XY}(2, 1) = \frac{1}{4}(1-r^2). \quad (23)$$

Given  $f$  for the primary locus and  $r$  for the LD, we simulated genotypes of the two loci one individual at a time. First, we calculated  $\pi = [\pi(11) \ \pi(12) \ \pi(22)]$  using equation (1) and simulated  $X_j$  from the multivariate Bernoulli distribution with a probability vector  $\pi$ . We then simulated  $Y_j$  from the multivariate Bernoulli distribution with a probability vector taking one of the three rows of matrix  $T_{XY}$ , depending on the simulated genotype for the first locus. When  $f > 0$ , both loci would show HWD, but only the first locus is of true HWD and the appeared HWD of the second locus is purely caused by LD. We used  $f=0.2$  as an example to demonstrate the empirical powers of the conditional and marginal analyses. The marginal analysis was conducted using the single locus model, ignoring the LD. The simulation was replicated 1000 times under each scenario. For the first locus, the detection of HWD is called the power. For the second locus, the detection of HWD is called the Type I error. The results are given in Table 4. The powers for the first locus of the conditional and marginal analyses are very similar. The Type I error for the second locus, however, is different when the LD level is high. Compared with the marginal analysis, the conditional analysis has lowered the Type I error for the second locus, but cannot fully control the Type I error when LD is too high. The power and Type I error analysis under other  $f$  values showed a similar trend (see supplemental Tables S5–S8). The conclusion from the two loci analysis was that the conditional method can control the Type I error for the second locus at some degree.

#### 4. Real data analysis

The purpose of the real data analysis is to demonstrate the difference between the marginal and conditional analyses. The data were downloaded from the website of a book entitled Applied Statistical Genetics with R (Foulkes, 2009: [http://people.umass.edu/foulkes/asg/data/FMS\\_data.txt](http://people.umass.edu/foulkes/asg/data/FMS_data.txt)).

The dataset (FAMuSS) contains 1397 college students genotyped for 225 SNPs across multiple genes. We analysed a gene named ‘akt1’, which included 24

Table 4. Empirical statistical power (locus *X*) and Type I error (locus *Y*) of HWD detection for the new method ( $f=0.2$ )

LD ( <i>r</i> )	Sample size ( <i>n</i> )	Conditional		Marginal	
		<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>
0.1	50	0.3630	0.0380	0.3620	0.0420
	100	0.6160	0.0330	0.6180	0.0360
	200	0.8500	0.0420	0.8510	0.0360
	300	0.9580	0.0460	0.9600	0.0440
	500	0.9950	0.0510	0.9960	0.0500
0.2	50	0.3950	0.0430	0.3770	0.0450
	100	0.5890	0.0390	0.5900	0.0460
	200	0.8310	0.0350	0.8350	0.0350
	300	0.9470	0.0370	0.9520	0.0410
	500	0.9950	0.0490	0.9950	0.0590
0.3	50	0.3940	0.0440	0.3860	0.0550
	100	0.5860	0.0440	0.5780	0.0600
	200	0.8550	0.0630	0.8590	0.0760
	300	0.9360	0.0580	0.9450	0.0780
	500	0.9950	0.0460	0.9940	0.0720
0.4	50	0.3920	0.0600	0.3640	0.0730
	100	0.5970	0.0390	0.5920	0.0590
	200	0.8520	0.0590	0.8580	0.1030
	300	0.9420	0.0600	0.9440	0.1040
	500	0.9950	0.0580	0.9950	0.1280
0.5	50	0.3930	0.0630	0.3650	0.0990
	100	0.6040	0.0560	0.5960	0.1080
	200	0.8340	0.0590	0.8350	0.1230
	300	0.9520	0.0680	0.9580	0.1580
	500	0.9960	0.0920	0.9980	0.2550

SNPs. We used both the marginal and the conditional methods to analyse the HWD for the 24 SNPs.

First, we expected the 24 SNPs to have high pairwise LD because they all came from the same gene. We used the LD(snp<sub>1</sub>,snp<sub>2</sub>) function of the R package 'genetics' to test all pairwise LD. Among the 24(24-1)/2=276 pairs of SNPs, a total of 243 pairs have *P*-values less than 0.05. The average *P*-value for all the 276 pairs is 0.04. The LD is very strong among the 24 SNPs. This means that the HWD tests for the conditional and marginal analyses should be very different.

The estimated HWD parameters ( $\theta$ ) for the 24 SNPs are plotted in Fig. 1 for both the marginal analysis (left panel) and the conditional analysis (right panel). For the marginal analysis, seven SNPs are significant ( $P < 0.05$ ). For the conditional analysis, six SNPs are significant ( $P < 0.05$ ). Among the significant SNPs detected here, only four were detected by both methods. Except for two SNPs, the estimated  $\theta$  are all negative for the marginal analysis. This means that almost all loci show heterozygote deficiency for the marginal analysis. The situation is quite different for the conditional analysis. Among the 24 SNPs, about half have positive estimates of  $\theta$ . An interesting

observation comes from akt1\_t10726c\_t12868c (the 5th SNP from the top). The marginal test shows that this SNP significantly deviates from HWE, but the conditional analysis shows that it does not. The appeared HWD in the marginal analysis may be caused by LD between this SNP and the next one (akt1\_t10598a\_t12740a, the 6th SNP from the top). The LD analysis (see Table S9) showed that the *D'* estimate between the two SNPs is 0.9785 with a correlation coefficient of 0.6169. The  $\chi^2$  test statistic is 589. This analysis clearly demonstrates the advantage of the conditional analysis over the marginal analysis. The two different estimates of  $\theta$  also share some level of similarity, with a Pearson correlation of 0.66. The Wald test statistics for the 24 SNPs obtained from both analyses are plotted in Fig. 2, again showing differences and similarities between the two analyses.

## 5. Discussion

We developed a GLM approach to testing HWD jointly for multiple loci. However, the method also works well for a single locus. Simulation studies showed that the method is over-conservative for low minor allele frequencies if the nominal 0.05 Type I error obtained from the  $\chi^2$  distribution is used as the criterion of test. The Fisher's exact test is also conservative when the minor allele frequency is low. This is consistent with other reports (Hernandez & Weir, 1989; Wigginton *et al.*, 2005). When both methods were compared under the same experimentally controlled Type I error rate of 0.05, the GLM method is more powerful than the exact test, which is a surprise. We never expected the new method to outperform the exact method for single locus analysis. Therefore, the new method can be a good (or even a better) substitute for the exact method for single locus analysis. For multiple locus joint analysis, the new method is the only one available in the literature.

The new parameter  $\theta$  is an alternative measurement of the heterozygote deficiency or excess. There is no explicit relationship between the new parameter and the *D* parameter of Weir (1996) or the Wright's fixation index *f*. However, the sign of  $\theta$  should be opposite to the sign of *D* or *f*. A negative  $\theta$  means heterozygote deficiency, which is represented by a positive *D* or *f*.

LD between two loci is contributed by digenic LD and higher-order trigenic and quadrigenic LD. Our GLM does not depend on the type of LD. In the simulation study, we considered only the digenic LD for the reason of simplicity. Higher-order LD may exist in real populations, but they are not expected to be strong in general (Jiang *et al.*, 2012). Therefore, the conclusion obtained from the simulation of gametic LD should be sufficiently general.

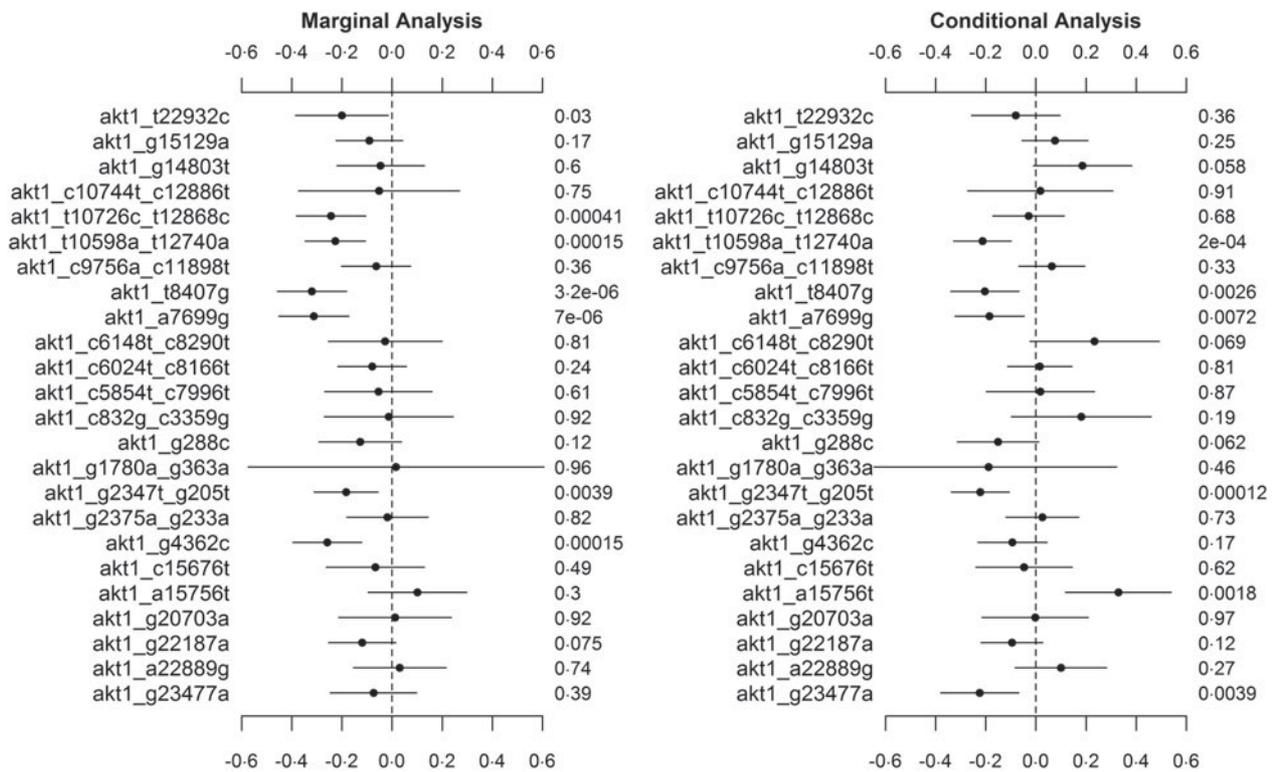


Fig. 1. Comparison of the estimated HWD parameters ( $\theta$ ) for the marginal (left panel) and conditional (right panel) analyses. The SNP names are shown in the left margin of each panel and the  $P$ -values are shown in the right margin of each panel. The  $x$ -axis represents the estimated  $\theta$  (dot) and the  $\theta \pm 2$  SE (bar).

Understanding the fact that the new method is conservative with a true Type I error lower than the nominal Type I error, we should correct this and use an empirical  $P$ -value criterion to test HWD in real data analysis. Two approaches may be taken to draw the empirical  $P$ -values for the Wald-test statistic, the exact test and the permutation test. The hypergeometric distribution of the genotype count under the biallelic system facilitates both tests. For the exact test, all possible outcomes from the hypergeometric distribution are evaluated and the empirical distribution of the Wald test statistic is drawn, from which the empirical  $P$ -value is calculated as the proportion of the more extreme samples than the actual observed Wald test statistic (Hernandez & Weir, 1989; Rousset & Raymond, 1995; Weir, 1996). The exact test applies to small samples, say  $<200$ . For the permutation test, one simply draws 1000 or more random samples from the hypergeometric distribution. Empirical  $P$ -value is then drawn from the proportion of extreme samples (Guo & Thompson, 1992). The permutation test applies to large samples.

Further investigation is required regarding the criterion of significance declaration for the multiple locus joint analysis. One can arbitrarily lower the  $P$ -value criterion to increase the stringency of the test. The change of stringency depends on the number of

SNPs included in the model (Lessios, 1992). The multiple test issue is a common problem in all genetic analyses, including the HWD test. Any approaches taken by the genome-wide association study (GWAS) for Type I error control can be adopted here for the HWD analysis. A typical approach, for example, is to control the false discovery rate (FDR).

HWD may be caused by any evolutionary forces. The one investigated in this study represents only one evolutionary force, non-random mating (including inbreeding and assortative mating), which only changes the heterozygosity to homozygote ratio (compared with the HWE predicted ratio) and does not distinguish between the two homozygotes. This can be reflected by our assignment of  $\psi_{11} = \psi_{22} = \Phi(0) = 1/2$ . We did not replace  $\Phi(0)$  by  $1/2$  in the formulation because that would make extension difficult. With  $\Phi(0)$  appearing in the model, we can extend the model easily to take into account other evolutionary forces, e.g. selection. In fact, none of the existing methods for HWD have taken into account selection. The extension is straightforward by adding another effect,  $\vartheta$ , that represents the fitness difference between the two homozygotes. The relative fitness of the three genotypes are then expressed by  $\psi_{11} = \Phi(\vartheta)$ ,  $\psi_{12} = \Phi(\theta)$  and  $\psi_{22} = \Phi(-\vartheta)$ , where  $\vartheta$  acts like the ‘additive’ effect and  $\theta$  the ‘dominance’ effect. The

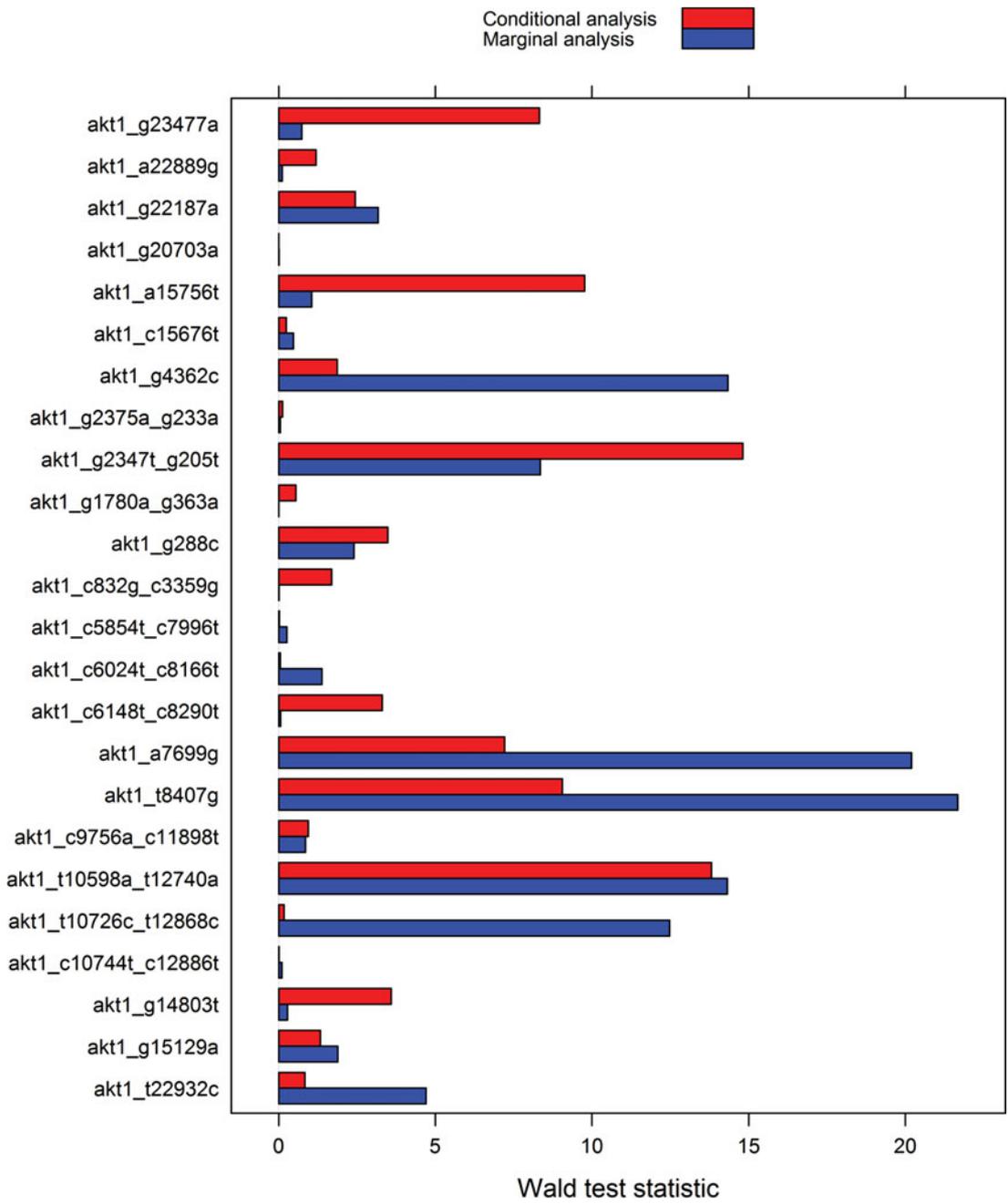


Fig. 2. Comparison of the Wald-test statistics for the 24 SNPs in gene akt1 obtained from the marginal (blue) and conditional (pink) analyses.

GLM then involves two HWD parameters. The algorithm developed in this study applies to two and multiple parameters with only slight modification. The two-parameter model for HWD test requires a  $\chi^2$  test with two degrees of freedom (asymptotically). Such a test is sufficiently general for HWD caused by any evolutionary forces. This topic deserves a further investigation.

Our initial study showed that choosing the relative fitness of  $\psi_{11} = \exp(0)$ ,  $\psi_{12} = \exp(\theta)$  and  $\psi_{22} = \exp(0)$  worked equally well (data not shown). Under this reparameterization, the link function is log. In the

single locus situation, the explicit MLE under the log link is

$$\hat{\theta} = \ln \left[ \frac{n_{12}(\varphi_{11} + \varphi_{22})}{\varphi_{12}(n_{11} + n_{22})} \right]. \tag{24}$$

The variance of  $\hat{\theta}$  is

$$\text{var}(\hat{\theta}) \approx - \left[ \frac{\partial^2 L(\hat{\theta})}{\partial \hat{\theta}^2} \right]^{-1} = \frac{(\varphi_{11} + \varphi_{12} \exp(\hat{\theta}) + \varphi_{22})^2}{n \varphi_{12} (\varphi_{11} + \varphi_{22}) \exp(\hat{\theta})}. \tag{25}$$

We now show an example of the log link analysis. Let the data be  $n_{11}=24$ ,  $n_{12}=39$  and  $n_{22}=37$ . The estimated  $\theta$  and the estimation error are

$$\begin{aligned}\hat{\theta} \pm \sqrt{\text{var}(\hat{\theta})} &= 0.413509 \pm \sqrt{0.0420345} \\ &= 0.413509 \pm 0.2050231\end{aligned}$$

and the Wald test statistic is  $\text{Wald}=4.0678$ . The results are similar to the probit link analysis. There are two advantages using the log link function. First, the computational behaviour of the log link function is better because the occurrence of floating errors is extremely low compared with the probit link function where  $\theta$  in  $\Phi(\theta)$  has to be constrained between  $-5$  and  $5$  to avoid numerical overfloating. Secondly,  $\psi_{12}=\exp(\theta)$  can be interpreted as the fitness odds ratio of the heterozygote to the homozygotes. An odds ratio of less than unity indicates heterozygosity deficit and an odds ratio of greater than unity indicates heterozygosity excess. We examined both the log and probit link functions, but decided to pursue the probit link because the log link did not work well for two and more loci. The reason is unclear and deserves further in-depth investigation.

Literature search did not show any evidence of using GLM for HWD analysis. The only closed method is the log linear approach of Lindley (1988), who defined two linear parameters,

$$\begin{aligned}\alpha &= \frac{1}{2} \ln(4P_{12}^2/P_{11}P_{22}), \\ \beta &= \frac{1}{2} \ln(P_{11}/P_{22}),\end{aligned}\quad (26)$$

where  $P_{11}$ ,  $P_{12}$  and  $P_{22}$  represent the observed frequencies of the three genotypes, and  $\alpha$  and  $\beta$  are analogous to the ‘dominance’ and ‘additive’ effects discussed earlier.

A typical GLM problem has closed forms of the first and second derivatives of the likelihood function with respect to the parameters for multiple independent variables. Our problem is atypical because it does not enjoy that property when two or more loci are considered. Fortunately, numerical differentiation is routinely conducted nowadays. With the high computer power, analytic and numerical differentiations do not seem to matter too much. In particular, finding the MLE of parameters is iterative anyway. Therefore, our atypical GLM problem is not much inferior to a typical GLM problem.

The number of SNPs handled by the joint analysis can be limited, say 100 at most. The method in its current form cannot be applied to genome-wide HWD in a simultaneous manner, although the single locus model is very practical. To perform a joint analysis for several thousand SNPs simultaneously, a penalty is required. Since the HWD parameter,  $\theta_k$  for the  $k$ th locus, has been formulated as a first moment

parameter, the  $L_1$  penalty (Lasso) or the  $L_2$  penalty (Ridge) or both (Elastic net) can be used. This opens a new avenue for the application of the penalized GLM.

Finally, if a large population is stratified and consists of many local populations of different demographic regions (population subdivision), the HWD of the whole population is also expected to be high. The  $\theta$  parameter estimated from the whole population would be denoted by  $\theta_{IT}$ , corresponding to  $F_{IT}$  (Holsinger & Weir, 2009). If the HWE predicted genotypic frequencies are calculated from each local population and these population-specific genotype frequencies are used as the prior probabilities for the corresponding local populations, the estimated  $\theta$  parameter (assuming all local populations have the same  $\theta$ ) would be denoted by  $\theta_{IS}$ , analogous to  $F_{IS}$ . The corresponding  $\theta$  parameter representing population differentiation may be derived from  $\theta_{ST} = \theta_{IT} - \theta_{IS}$ , corresponding to the population differentiation index  $F_{ST}$  (Holsinger & Weir, 2009). We may then use

$$H_{ST} = \frac{\theta_{ST}^2}{\theta_{ST}^2 + \theta_{IS}^2} = \frac{(\theta_{IT} - \theta_{IS})^2}{(\theta_{IT} - \theta_{IS})^2 + \theta_{ST}^2} \quad (27)$$

to measure population differentiation. This new parameter takes a domain between 0 and 1 with zero meaning no differentiation and one being maximum differentiation. This extension will shed new light on future population differentiation studies.

## 6. Supplemental Material

There are eight supplemental tables (Tables S1–S8). The first four tables (Tables S1–S4) show results of additional simulations for power analysis under the single locus model. The last four tables (Tables S5–S8) give the results of additional simulations for power (primary locus) and Type I error (secondary locus) under the two loci model. These supplemental documents are available under the Paper Information link at the Genetics Research website (<http://www.journals.cambridge.org/GrH>). The R package ‘hwdglm’ is available from the author’s personal website: <http://www.statgen.ucr.edu>.

We thank two anonymous reviewers for their informative comments on the first version of the manuscript. The project was supported by the USDA National Institute of Food and Agriculture Grant 2007-02784.

## References

- Ayres, K. L. & Balding, D. J. (1998). Measuring departures from Hardy–Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. *Heredity* **80**, 769–777.
- Chapco, W. (1976). An exact test of the Hardy–Weinberg law. *Biometrics* **32**, 183–189.

- Chen, J. & Chatterjee, N. (2007). Exploiting Hardy–Weinberg equilibrium for efficient screening of single SNP associations from case-control studies. *Human Heredity* **63**, 196–204.
- Crow, J. F. (1988). Eighty years ago: the beginnings of population genetics. *Genetics* **119**, 473–476.
- Deng, H.-W., Chen, W.-M. & Recker, R. R. (2001). Population admixture: detection by Hardy–Weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. *Genetics* **157**, 885–897.
- Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society* **98**, 39–54.
- Foulkes, A. S. (2009). *Applied Statistical Genetics with R – For Population-based Association Studies*. Springer, New York.
- Guo, S. W. & Thompson, E. A. (1992). Performing the exact test of Hardy–Weinberg proportion for multiple alleles. *Biometrics* **48**, 361–372.
- Haldane, J. B. S. (1954). An exact test for random mating. *Journal of Genetics* **52**, 631–635.
- Hernandez, J. L. & Weir, B. S. (1989). A disequilibrium coefficient approach to Hardy–Weinberg testing. *Biometrics* **45**, 53–70.
- Hill, W. G., Babiker, H. A., Ranford-Cartwright, L. C. & Walliker, D. (1995). Estimation of inbreeding coefficients from genotypic data on multiple alleles, and application to clonality in malaria parasites. *Genetical Research* **65**, 53–61.
- Holsinger, K. E. & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting FST. *EEB Articles Paper* 22.
- Jiang, Q., Wang, Z., Moore, S. S. & Yang, Y.-C. (2012). Genome-wide analysis of zygotic linkage disequilibrium and its components in crossbred cattle. *BMC Genetics* **13**, 1–13.
- Lee, W.-C. (2003). Searching for disease-susceptibility loci by testing for Hardy–Weinberg disequilibrium in a gene bank of affected individuals. *American Journal of Epidemiology* **158**, 397–400.
- Lessios, H. A. (1992). Testing electrophoretic data for agreement with Hardy–Weinberg expectations. *Marine Biology* **112**, 517–523.
- Li, C. C. (1955). *Population Genetics*. University of Chicago Press, Chicago.
- Lindley, D. (1988). Statistical inference concerning Hardy–Weinberg equilibrium. *Bayesian Statistics* **3**, 307–326.
- McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.
- Nielsen, D. M., Ehm, M. G. & Weir, B. S. (1999). Detecting marker-disease association by testing for Hardy–Weinberg disequilibrium at a marker locus. *American Journal of Human Genetics* **63**, 1531–1540.
- Rohlf, R. V. & Weir, B. S. (2008). Distributions of Hardy–Weinberg equilibrium test statistics. *Genetics* **180**, 1609–1616.
- Rousset, F. & Raymond, M. (1995). Testing heterozygote excess and deficiency. *Genetics* **140**, 1413–1419.
- Salanti, G., Amountza, G., Ntzani, E. E. & Ioannidis, J. P. A. (2005). Hardy–Weinberg equilibrium in genetic association studies: an empirical evaluation of reporting, deviations, and power. *European Journal of Human Genetics: EJHG* **13**, 840–848.
- Shoemaker, J., Painter, I. & Weir, B. S. (1998). A Bayesian characterization of Hardy–Weinberg disequilibrium. *Genetics* **149**, 2079–2088.
- Weinberg, C. R. & Morris, R. W. (2003). Invited commentary: testing for Hardy–Weinberg disequilibrium using a genome single-nucleotide polymorphism scan based on cases only. *American Journal of Epidemiology* **158**, 401–403.
- Weir, B. S. (1996). *Genetic Data Analysis II – Methods for Discrete Population Genetic Data*. Sinauer Associates Inc., Sunderland, MA.
- Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. (2005). A note on exact tests of Hardy–Weinberg equilibrium. *American Journal of Human Genetics* **76**, 887–893.
- Zhan, H. & Xu, S. (2011). Generalized linear mixed model for segregation distortion analysis. *BMC Genetics* **12**, 97.