Original Paper

Human Heredity

Hum Hered 2010;69:212–218
DOI: 10.1159/000289597

Received: July 23, 2009
Accepted after revision: January 4, 2010
Published online: March 5, 2010

Using Both Cases and Controls for Testing Hardy-Weinberg Proportions in a Genetic Association Study

Jian Wang Sanjay Shete

Department of Epidemiology, The University of Texas M.D. Anderson Cancer Center, Houston, Tex., USA

Key Words
Hardy-Weinberg proportion · Case-control association study

Abstract

Objectives: Assessment of the Hardy-Weinberg proportion (HWP) in controls has been widely used as a quality control measure in case-control association studies. However, when the disease being studied is common, controls might not represent the general population, which could result in inaccurate HWP test results. Such results could lead investigators to discard important single-nucleotide polymorphisms (SNPs) that could potentially be causal. In this paper, we showed the inappropriateness of the HWP test in controls and proposed a mixture HWP (mHWP) exact test using a mixture sample that mimics the general population. ***Methods:*** The mHWP exact test estimates HWP in a mixture sample that is a combination of both cases and controls proportional to the prevalence of disease. We implemented a re-sampling procedure to construct mixture samples and then obtained the empirical p value of HWP in the general population. Simulation studies were performed to investigate the performance of the proposed mHWP exact test. The method was also applied to a genetic association study of obesity. ***Results:*** The results showed that the mHWP exact test is more likely than either the traditional HWP method in controls or the likelihood-based approach to keep causal SNPs for further analysis when the disease is more common. ***Conclusion:*** The mHWP exact test using a mixture sample is a better HWP test for case-control genetic association studies than the traditional HWP in controls or the likelihood-based approach, and it will improve our ability to keep causal SNPs in the case-control genetic association studies.

Introduction

Assessment of the Hardy-Weinberg proportion (HWP) in control subjects has been widely used as a quality control measure for identifying questionable genotypes in case-control association studies [1–8]. Consider a simple situation with two alleles, A and a, at a single locus. If the allele frequency of A is p and the allele frequency of a is $(1 – p)$, then the expected genotype frequencies of AA, Aa, and aa are p^2, $2p\ (1 – p)$, and $(1 – p)^2$, respectively, assuming HWP in the population. In a case-control study, the deviation from HWP in controls, which is assessed by comparing the difference between observed genotype frequencies and the corresponding expected frequencies [7], is used to identify potential genotyping errors. Using only controls for HWP assessment is reasonable when assuming a rare disease in the study. However, when the disease of interest is common, controls might not represent the general population, as cases account for a relatively large portion of the general population. Therefore, it would be problematic to use only controls when evalu-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2010 S. Karger AG, Basel
0001–5652/10/0693–0212$26.00/0

Accessible online at:
www.karger.com/hhe

Dr. Sanjay Shete
Department of Epidemiology, Unit 1340
The University of Texas M.D. Anderson Cancer Center
1155 Pressler Street, CPB4.3628, Houston, TX 77030 (USA)
Tel. +1 713 745 2483, Fax +1 713 792 8261, E-Mail sshete@mdanderson.org

ating the expected genotype frequencies in the general population. In these situations, using the HWP test in controls might lead to discarding important single-nucleotide polymorphisms (SNPs) that could potentially be causal SNPs of the disease.

In this paper, we first show the inappropriateness of using the HWP test in only controls and then propose an improved HWP test, called the mixture HWP (mHWP) exact test, using a mixture sample that mimics the general population for the case-control genetic association study. The mHWP exact test estimates HWP in a mixture sample that is a combination of cases and controls. The cases and controls in the mixture sample were selected randomly, and the number of cases in the mixture sample was proportional to the prevalence of the disease of interest. We implemented a re-sampling procedure to obtain empirical p values for the mHWP test. For each step of re-sampling, we constructed a mixture sample and obtained an HWP p value. Non-parametric density estimation (kernel density estimation) was used to estimate the density function of the empirical p values based on mixture samples. The maximum likelihood estimator (MLE) of this empirical density was then evaluated as the p value of HWP in the general population. We compared our mHWP exact test with the traditional HWP exact test in controls and with the likelihood-based approach recently proposed by Li and Li [4]. A recent study of Yu et al. [9] proposed a likelihood ratio test for HWP that is similar to the likelihood-based approach proposed by Li and Li [4]. The results from our simulation studies showed that, when there is no genotyping error, the mHWP exact test is more likely to keep causal SNPs for further analysis in a case-control association study when the disease of interest is more common. We also applied the proposed mHWP exact test to the real case-control genetic association study of obesity.

Methods

For our studies, we assumed a diallelic locus, with A as the risk allele and a as the normal allele. We denoted the three genotypes – aa, Aa, and AA – as a categorical random variable, $X = (0, 1, 2)$. This coding assumed an additive model, but different coding for representing a dominant or recessive model was also used in the simulations. We defined another categorical random variable, $Y = (0, 1)$, to indicate the case-control status, with 0 representing controls and 1 representing cases.

mHWP Exact Test

Given a dataset of observations of random variables X and Y corresponding to the genotypes of a SNP and the case-control status, respectively, and the known prevalence of the disease, we first constructed a mixture sample to represent the general population.

Consider a case-control study with n individuals, $n = n_0 + n_1$, where n_0 is the number of controls and n_1 is the number of cases. Denote $n_m < n$ as the sample size of the mixture sample. Let $\hat{f}$ be the estimated prevalence of the disease of interest in the general population. To represent the general population, the number of cases in the mixture sample should be $\lfloor n_m \times \hat{f} \rfloor$ and the number of controls should be $\lfloor n_m \times (1 - \hat{f}) \rfloor$ (online suppl. fig. 1, www.karger.com/doi/10.1159/000289597). It should be noted that $\lfloor n_m \times \hat{f} \rfloor \leq n_1$ and $\lfloor n_m \times (1 - \hat{f}) \rfloor \leq n_0$; therefore, $n_m \leq \min(\lfloor n_1/\hat{f} \rfloor, \lfloor n_0/(1 - \hat{f}) \rfloor)$. One could choose $n_m = \min(\lfloor n_1/\hat{f} \rfloor, \lfloor n_0/(1 - \hat{f}) \rfloor)$ to achieve the largest possible mixture sample size. We randomly sampled $\lfloor n_m \times \hat{f} \rfloor$ individuals from the cases and $\lfloor n_m \times (1 - \hat{f}) \rfloor$ individuals from the controls. This mixture sample should represent the general population. In the mixture sample, we calculated the counts of the three genotypes, aa, Aa, and AA. The exact HWP test was then applied to the mixture sample [7]. To allow for variability in the mixture sampling, we repeated the procedure to obtain the mixture sample L times and then obtained L HWP exact p values. The empirical distribution-based non-parametric density was constructed based on L mixture sample p values (see details of kernel density estimation in Appendix 1). Then the MLE of this empirical distribution was obtained as the final estimate of p value for HWP in the general population. We conducted simulations to decide the number of mixture samples L. We found that when $L \geq 500$, the empirical distributions and the corresponding MLEs approach stability. Therefore, we selected $L = 500$ in our study.

Simulation Studies

We performed simulation studies to compare three approaches for HWP testing for case-control association study: (i) the proposed mHWP exact test; (ii) the likelihood-based approach proposed by Li and Li [4], and (iii) the traditional HWP exact test using controls. We considered two independent SNPs at two different genetic loci: SNP_1 and SNP_2. In addition to the genetic risk factors, we also accounted for environmental risk or protective factors, such as sex, ethnicity, physical activity, and age, in the simulation models. The case-control status was simulated based on a logistic model. We defined all the odds ratios (ORs) and prevalences of the genetic and environmental factors for the purpose of the simulation studies, as listed in table 1. We assumed SNP_1 is a causal SNP of the disease (OR = 1.5) and SNP_2 is non-causal (OR = 1.0). We simulated genotypes of both SNPs under the assumption of HWP as in the general population.

In this paper, we studied minor allele frequencies (MAFs) of 10, 30, and 50% (from rare to more common) for both SNPs. By defining different intercept coefficients of the logistic model, we considered different levels of prevalence, ranging from 19–36%, which can represent different common diseases. For example, the prevalence of current smoking was about 20% in the U.S. in 2004 [10], the prevalence of obesity among adults in the U.S. was about 32% in 2004 [11], and the prevalence of overweight in the U.S. was about 66% [11]. We did not study the scenarios of rare diseases because we, as well as Li and Li [4], have found that the traditional approach of testing HWP only in controls works well in this situation. Meanwhile, we also studied three different genetic models: dominant, additive, and recessive. The type I error probabilities reported in the results section were based on 10,000 replicates, which included 1,000 cases and 1,000 controls. For the mHWP exact test, the significance of each replicate was determined on the basis of 500 re-sampling-based mixtures. The sample size of the mixture samples n_m was set to be 1,000 in the simulation studies. We compared the type I error rates of the traditional HWP exact test in controls, the likelihood-based approach [4], and our mHWP exact test in the mixture sample. For the likelihood-based approach, we applied the 'fminsearchcon' function [12] in Matlab to implement the simplex algorithm when maximizing the likelihood, as suggested by [4].

In order to assess the ability of the three approaches to detect genotyping errors, we introduced genotyping errors into the simulation studies using the GLHO genotyping error model described by Gordon et al. [13–15] and the 'empirical' error model used in Fardo et al. [16]. The GLHO model introduces errors for each allele independently, with probabilities of ε_1 and ε_2, respectively, where ε_1 is the probability of allele A incorrectly coded as allele a, and ε_2 is the probability of allele a incorrectly coded as allele A. In our study, we assumed that $\varepsilon_1 = \varepsilon_2 = \varepsilon$. Therefore, the probabilities of a homozygous genotype being miscoded as the other homozygous genotype or a heterozygous genotype are ε^2 and $\varepsilon(1 - \varepsilon) + (1 - \varepsilon)\varepsilon = 2\,\varepsilon(1 - \varepsilon)$, respectively, while the probability of a heterozygous genotype being miscoded as a homozygous genotype is $\varepsilon(1 - \varepsilon)$. The expected genotyping error rate is $2\varepsilon - \varepsilon^2(1 + 2p - 2p^2)$. When using the GLHO model, we assumed that $\varepsilon = 0.5$ or 2.5%, so the expected genotyping error rate in our simulation was approximately 1 or 5%. On the other hand, the 'empirical' error model is based on a real genome-wide association data in which errors were estimated based on re-sequencing. According to this model, the genotyping error rate is very high (~12%) [16].

Relative Rejection Probability

In order to compare the different approaches for testing HWP in a case-control genetic association study, we measured the relative probability of one approach rejecting HWP in causal or non-causal SNPs, which were assumed in HWP, compared to the other approaches at a given significance level α and called this the relative rejection probability (RRP). Consider two methods for testing HWP in a case-control study, M_1 and M_2. The RRP of M_1 compared to M_2 at significance level α is defined as the following [17]:

RRP =
[P(reject HWP hypothesis using M_1 at $\alpha|$HWP) – P(reject HWP hypothesis using M_2 at $\alpha|$HWP)]/P(reject HWP hypothesis using M_2 at $\alpha|$HWP),

where $P(\cdot)$ is the probability. Note that if RRP is positive, using M_1 is more likely to result in rejection of SNPs than using M_2, when the SNPs are in HWP.

Results

Simulation Study Results

We first studied the models without any genotyping errors. We reported the observed type I error rates of the three approaches for testing HWP at the defined significance of 0.05 for all the scenarios based on 10,000 replicates. We reported the RRPs for the likelihood-based approach versus the mHWP exact test in the mixture sample. The results of SNP_1 (causal SNP) and SNP_2 (non-causal SNP) are reported in table 2 and 3, respectively. Both tables are organized into two panels with respect to type I error rates and RRPs.

Table 1. Parameters for simulation studies

Factors	Coefficients of logistic model	Prevalence, %
Intercept	–3.4/–2.5/–1.9	
SNP_1	0.4055 (OR = 1.5)	10/30/50
SNP_2	0 (OR = 1)	10/30/50
Sex	0.6931 (OR = 2)	50 (male)
Ethnicity	0.4055 (OR = 1.5)	75 (Caucasian)
Physical activity	–0.4055 (OR = 0.67)	50 (yes)
Age		
0–30 years	0.4055 (OR for additive model = 1.5)	36
31–50 years		39

When the SNP was associated with the disease (SNP_1), the type I error rates for the traditional approach (using controls only) were inflated dramatically as MAFs and prevalence of disease increased, when the dominant or recessive model was assumed. For example, for the dominant model, when the MAF was 0.1 and the prevalence of disease was 19.56%, the type I error rate for the traditional approach was 0.051, and when the MAF was 0.5 and the prevalence of disease was 34.12%, the type I error rate increased to 0.152, which is about three times the nominal significance level (0.05). We observed a similar trend for the recessive model. The traditional approach could control the type I error rate when the additive model was assumed; however, in reality, one would not know the real underlying genetic model, so the traditional approach evaluating HWP using only controls would lead to inflated type I error in many situations. On the other hand, the likelihood-based approach and the mHWP exact test could control type I errors in all scenarios in the simulation studies. For example, when the dominant model was assumed, MAF was set as 0.3, and prevalence was 32.20%, the type I error rate of causal SNP_1 was 0.111 using the traditional approach in controls, but the type I error rates were 0.050 and 0.033, respectively, using the likelihood-based approach and mHWP exact test, which agree well with the nominal value of 0.05.

Table 2. Estimated type I error probability and relative rejection probability of causal SNP$_1$, at 0.05 significance level in simulation studies based on 10,000 replicates, each replicate with 1,000 cases and 1,000 controls

Model	MAF	Prev %	Type I errors			Relative rejection probability
			controls only	likelihood-based	mHWP	likelihood vs. mHWP
Dominant	0.1	19.56	0.051	0.057	0.039	0.470
		29.64	0.058	0.052	0.029	0.808
	0.3	21.58	0.077	0.053	0.044	0.203
		32.20	0.111	0.050	0.033	0.545
	0.5	23.09	0.095	0.053	0.047	0.146
		34.12	0.152	0.054	0.035	0.551
Additive	0.1	19.64	0.036	0.053	0.033	0.613
		29.73	0.041	0.052	0.026	1.031
	0.3	22.25	0.047	0.049	0.041	0.197
		32.99	0.049	0.050	0.031	0.604
	0.5	24.96	0.056	0.054	0.045	0.205
		36.31	0.056	0.049	0.032	0.553
Recessive	0.1	18.43	0.041	0.052	0.037	0.403
		28.20	0.041	0.051	0.030	0.678
	0.3	18.93	0.080	0.052	0.046	0.144
		28.84	0.115	0.051	0.037	0.393
	0.5	19.94	0.095	0.048	0.045	0.060
		31.12	0.150	0.049	0.038	0.284

MAF = Minor allele frequency; Prev = prevalence of disease; mHWP = mixture Hardy-Weinberg proportion.

Table 3. Estimated type I error probability and relative rejection probability of non-causal SNP$_2$, at 0.05 significance level in simulation studies based on 10,000 replicates, each replicate with 1,000 cases and 1,000 controls

MAF	Prev %	Type I errors			Relative rejection probability
		controls only	likelihood-based	mHWP	likelihood vs. mHWP
0.1	19.21	0.041	0.052	0.040	0.296
	29.19	0.041	0.053	0.033	0.619
0.3	20.92	0.045	0.050	0.045	0.114
	31.34	0.046	0.051	0.035	0.449
0.5	22.66	0.048	0.051	0.045	0.127
	33.85	0.049	0.052	0.035	0.501

MAF = Minor allele frequency; Prev = prevalence of disease; mHWP = mixture Hardy-Weinberg proportion.

Although both the likelihood-based approach and the mHWP exact approach can control type I error rates, the mHWP exact test proposed in this paper is more likely to keep causal SNPs than the likelihood-based approach, according to the RRP results shown in table 2. Considering the dominant model, with MAF being 0.3 and prevalence being 21.58%, the type I error rates for the likelihood-based and mHWP approaches were 0.053 and 0.044, respectively, at significance 0.05, and the RRP of the likelihood-based approach compared to the mHWP exact test for SNP$_1$ was 0.203, which means that the likelihood-based approach is 20.3% more likely to reject the causal SNP$_1$ than the mHWP exact test. Because the RRPs of the likelihood-based approach versus the mHWP exact test in all the scenarios reported in table 2 are positive, we can conclude that the likelihood-based approach is more likely to reject causal SNPs, or in other words, the mHWP exact test is more likely to keep the causal SNPs for analysis. For a fixed MAF, as prevalence of the disease increases, we observed that the mHWP exact test is even more likely to keep causal SNPs.

When the SNP is non-causal (SNP$_2$, table 3), all three approaches control type I error well. The mHWP exact test has a type I error rate very similar to that of the traditional approach, while the likelihood-based approach has a slightly inflated type I error rate. For example, when MAF was set as 0.1 and prevalence was 19.21%, the observed type I error rates of non-causal SNP$_2$ at 0.05 significance were 0.041, 0.052, and 0.040 for the traditional, likelihood-based, and mHWP exact test approaches, respectively. The corresponding RRP for this scenario, comparing the likelihood-based approach to the exact mHWP approach, was 29.6%. As all the RRPs were positive, the exact mHWP approach was still better than the likelihood-based approach for keeping non-causal SNPs.

When genotyping errors were introduced in the simulation, we found that all three approaches worked in a very similar way. When the GLHO error model was used with low genotyping error rates (1 and 5%), all three approaches had low power (5 ~ 10% power at the 5% level of significance) to detect the errors. This is because when the genotyping error rates are small, the observed genotype counts will not be significantly different from the expected genotype counts under HWP and, therefore, any test that attempts to detect such errors based on HWP testing will have very little power [18]. On the oth-

Table 4. Genotype counts of SNP rs9939609 in normal weight (BMI <25) and obese (BMI ≥30) individuals and p values for three approaches of HWP test of case-control genetic association studies in UK T2D GCC cases

Cohort	Genotype counts						Prev	p values		
	normal weight			obese				controls only	likelihood-based	mHWP
	TT	AT	AA	TT	AT	AA				
UK T2D GCC cases	113	174	37	524	818	321	0.52	0.018	0.038	0.054

Prev = Prevalence of disease; mHWP = mixture Hardy-Weinberg proportion.

er hand, when the genotyping error rates are higher (as with the 'empirical' error model), the genotyping error can generate extreme deviation from HWP, and therefore, we found that all three approaches have almost 100% power to detect the genotyping errors. These results are consistent with previous studies of the relationship between genotyping errors and the HWP test [18–21].

In the simulation studies, we assumed that the prevalence of the disease was known. However, in reality, prevalence is not known and is estimated from data. Here, we assessed the sensitivity of the mHWP exact test to the estimated prevalence of disease, $\hat{f}$. We considered the models with MAF = 0.1, and the real prevalence f values used to simulate the data were 29.64, 29.73, and 28.20% for the dominant, additive, and recessive genetic models, respectively. We evaluated the mHWP exact test p values using a range of prevalence centered on real prevalence [f – 2%, f + 2%]. All the results were very similar to those obtained with the use of the real prevalence. For example, consider the causal SNP$_1$. When the dominant model was assumed, the type I error rates of the mHWP exact test were 0.029 using the real prevalence of 29.64%, 0.030 using the estimated prevalence of 27.64%, and 0.026 using the estimated prevalence of 31.64%.

Results of Real Disease Application

We also applied our approach to the real case-control genetic association study of adult obesity. We used the case-control data from an association study of a common variant in the FTO gene that is associated with obesity [22]. In that study, the investigators found that the SNP rs9939609 predisposes individuals to diabetes through an effect on body mass index (BMI). Although BMI is a continuous measure, standard cut-off points can be applied to define the cases and controls. Therefore, for the case-control study of obesity, individuals with a BMI ≥30 were classified as cases (obesity) and individuals with a BMI <25 were classified as controls (normal weight). The original study involved association studies in 13 cohorts, but for the purpose of our study, we selected one United Kingdom (UK) cohort: UK Type 2 Diabetes Genetics Consortium Collection Cases (UK T2D GCC Cases).

The genotype counts of SNP rs9939609 in normal weight (controls) and obese (cases) individuals and all the resulting HWP test p values are listed in table 4. The prevalence of obesity among type 2 diabetes patients in UK was estimated as 52% [23]. When using the HWP exact test with controls only, the p value was 0.018; when using the likelihood-based approach, the p value was 0.038; when using our approach, the p value was 0.054. Therefore, at a 5% level of significance both the traditional and likelihood-based approaches will remove this SNP from further association analyses. On the other hand, our approach will retain this SNP in analysis. Furthermore, our mHWP approach allows for using multiple categories to make mixture sample, i.e. normal weight (BMI <25), overweight (25≤ BMI<30), and obese (BMI ≥30). Using prevalence of 14, 34, and 52% for normal weight, overweight, and obese individuals, respectively, our mixture approach resulted in a p value of 0.123. Importantly, the investigators kept this SNP in the analyses of BMI because originally this study was for the investigation of type 2 diabetes, and in the type 2 diabetes GCC controls the HWP test gave a p value of 0.83 for this SNP.

Discussion

Traditional quality control methods test HWP using only controls and remove the SNPs that deviate from HWP as genotyping errors. However, when the disease is common, controls might not be representative of the general population, and the traditional approach may lead to the removal of causal SNPs from further analysis. In this paper, we have shown that this is indeed the case. When the prevalence of the disease is large (ranging from 19 to 36% in our simulation studies), the type I error probability of the traditional approach was inflated for the disease-associated SNPs when either the dominant or recessive model was assumed. This range of prevalence is realistic for common diseases, such as smoking, obesity, and hypertension. Therefore, we developed an mHWP exact test based on a mixture sample that can represent the general population. In the mixture sample, a certain proportion was randomly sampled from cases proportional to the prevalence of the disease, and the rest was sampled from controls. A re-sampling procedure was applied to obtain the empirical p values, and the MLE of the empirical distribution of re-sampled p values was the HWP p value in the general population. The mHWP approach was compared to the traditional approach and the likelihood-based approach.

On the basis of the results of our simulation studies, the mHWP exact test can effectively control the type I error probability in all scenarios examined, including models with causal or non-causal SNPs, different genetic models, and different MAFs and prevalence. Furthermore, on average, the mHWP exact test proposed in this paper is more likely than the likelihood-based approach to keep causal SNPs in the analysis when the disease is common. In genome-wide association studies, using the improved mHWP exact test in the discovery stage will increase the chance that causal SNPs will be carried over for replication. However, to achieve a more stringent significance level (e.g. 10^{-5} used in GWAS), more mixture samples will be needed to obtain robust MLE of the empirical mHWP p value. We also considered smaller numbers of cases and controls (500 cases and 500 controls), and the results had a similar pattern. Therefore, we conclude that the mHWP exact test is better for testing HWP for case-control genetic association study than either the traditional HWP method using only controls or the likelihood-based approach. Furthermore, when data is available, the mHWP exact approach allows for using multiple categories to build the mixture samples, which could better represent the general population.

In addition to simulation studies, we also applied the mHWP exact approach to the real case-control genetic association study of adult obesity. The comparison of the p values from three approaches show that the mHWP approach has higher likelihood to keep the SNP rs9939609 for further analysis.

The relationship between genotyping error and the HWP test has been studied in the literature [16, 18–21]. These studies suggest that the traditional HWP test in controls has very low power for detecting genotyping errors, especially when the genotyping error rate is low and the MAF is not rare. However, the HWP test is always considered an essential procedure in genetic association study and has been widely used as a quality control tool in genetic case-control studies [1, 3, 24–26]. In this paper, our main purpose was developing an improved HWP test that is more likely to keep causal SNPs in the analyses. Like the traditional HWP test and the likelihood-based test, our test is not very sensitive for detecting genotyping errors when error rates are low. Furthermore, Fardo et al. [16] recently showed that genotyping errors will not increase the false-positive rate for detecting associated variants. Therefore, one may also consider a strategy of keeping all SNPs for the association study, performing the HWP test using our proposed approach only among significant SNPs.

The true prevalence of a disease in a population is not known with certainty, and Li and Li have shown, by using a sensitivity analysis, that the misspecification of disease prevalence would not inflate the type I error rate for their likelihood-based approach when the genetic effect size is moderate [4]. We also evaluated the sensitivity of the mHWP exact test to the estimated prevalence and found that the prevalence misspecification would not inflate the type I error rate of our approach either.

In conclusion, here in this paper, we proposed an improved HWP exact test (mHWP exact test) using a mixture sample, which is a better HWP test for case-control genetic association studies than the traditional HWP only in controls or the likelihood-based approach. This approach will improve our ability to keep causal SNPs in the case-control genetic association studies.

Appendix 1

Kernel Density Estimation

We estimated the density function of the empirical distribution of the p values from re-sampling-based mixtures with kernel density estimators. Kernel density estimator is the most popular and most widely used nonparametric approach for estimating the unknown probability density function of a random variable [27, 28]. It has been shown that the kernel density estimator is able to make efficient use of the data [27, 28]. The consistency of the kernel density estimator has been well studied [28–31]. Given a random sample $X_1, X_2, \ldots X_N$ from some density g, the kernel density estimate of g is defined by

$$\hat{g}_h(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x-X_i}{h}\right),$$

where $K(\cdot)$ is a kernel function and $h > 0$ is a smoothing parameter called the bandwidth. In order to estimate g, the choice of $K(\cdot)$ is not particularly crucial [27] and is usually taken to be a symmetric unimodal density centered at zero, such as the standard normal density. However, the bandwidth parameter h is extremely important to the performance of the estimator [32]. Without loss of generality, we implemented kernel density estimation using the function ‘ksdensity’ in Matlab [33] with default settings, which are the standard normal kernel and optimal bandwidth suggested by Bowman and Azzalini [34]. The optimal bandwidth is given as

$$h = \left(\frac{4}{3N}\right)^{1/5} \hat{\sigma}.$$

$\hat{\sigma}$ is the median absolute deviation estimator defined as $\hat{\sigma}$ = median $\{|X_i - \hat{u}|\}/0.6745$, where $\hat{u}$ denotes the median of the sample. $\hat{\sigma}$ defined here is a robust estimator compared to the usual sample standard deviation, which is not preferable as it is influenced by long-tailed distributions and possible outliers [35]. The MLE of this empirical distribution of p values based on re-sampled mixture samples was then obtained as the final p value of HWP in the general population.

Acknowledgement

Grant support: 1R01CA131324.

References

1 Gomes I, Collins A, Lonjou C, Thomas NS, Wilkinson J, Watson M, Morton N: Hardy-Weinberg quality control. Ann Hum Genet 1999;63:535–538.

2 Graffelman J, Camarena JM: Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Hum Hered 2008;65:77–84.

3 Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF: Detection of genotyping errors by Hardy-Weinberg equilibrium testing. Eur J Hum Genet 2004;12:395–399.

4 Li M, Li C: Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. Genet Epidemiol 2008;32: 589–599.

5 Schaid DJ, Batzler AJ, Jenkins GD, Hildebrandt MA: Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata. Am J Hum Genet 2006; 79:1071–1080.

6 Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE: A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci USA 2005;102:11835–11839.

7 Weir BS: Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sunderland, Mass, Sinauer Associates, 1996.

8 Wittke-Thompson JK, Pluzhnikov A, Cox NJ: Rational inferences about departures from Hardy-Weinberg equilibrium. Am J Hum Genet 2005;76:967–986.

9 Yu C, Zhang S, Zhou C, Sile S: A likelihood ratio test of population Hardy-Weinberg equilibrium for case-control studies. Genet Epidemiol 2009;33:275–280.

10 CDC: Cigarette smoking among adults – United States, 2004. MMWR 2005;54:1121–1124.

11 Ogden CL, Carroll MD, Curtin LR, McDowell MA, Tabak CJ, Flegal KM: Prevalence of overweight and obesity in the United States, 1999–2004. JAMA 2006;295:1549–1555.

12 D’Errico J: Fminsearchcon. MATLAB Central File Exchange (http://www.mathworks.com/matlabcentral/fileexchange/?term=fminsearchcon).

13 Gordon D, Heath SC, Liu X, Ott J: A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. Am J Hum Genet 2001;69:371–380.

14 Gordon D, Ott J: Assessment and management of single nucleotide polymorphism genotype errors in genetic association analysis. Pac Symp Biocomput 2001;18–29.

15 Gordon D, Finch SJ, Nothnagel M, Ott J: Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. Hum Hered 2002;54:22–33.

16 Fardo DW, Becker KD, Bertram L, Tanzi RE, Lange C: Recovering unused information in genome-wide association studies: the benefit of analyzing SNPs out of Hardy-Weinberg equilibrium. Eur J Hum Genet DOI:10.1038/ejhg.2009.85.

17 Thomas GB, Finney RL: Calculus and Analytic Geometry, ed 9. Addison Wesley, 1995.

18 Cox DG, Kraft P: Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. Hum Hered 2006; 61:10–14.

19 Leal SM: Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. Genet Epidemiol 2005;29:204–214.

20 Zou GY, Donner A: The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note. Ann Hum Genet 2006;70:923–933.

21 Teo YY, Fry AE, Clark TG, Tai ES, Seielstad M: On the usage of HWE for identifying genotyping errors. Ann Hum Genet 2007;71: 701–703.

22 Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI: A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 5–11–2007;316:889–894.

23 Daousi C, Casson IF, Gill GV, MacFarlane IA, Wilding JP, Pinkney JH: Prevalence of obesity in type 2 diabetes in secondary care: association with cardiovascular risk factors. Postgrad Med J 2006;82:280–284.

24 Weiss ST, Silverman EK, Palmer LJ: Case-control association studies in pharmacogenetics. Pharmacogenomics J 2001;1:157–158.

25 Xu J, Turner A, Little J, Bleecker ER, Meyers DA: Positive results in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? Hum Genet 2002;111:573–574.

26 Wang J, Shete S: A test for genetic association that incorporates information about deviation from Hardy-Weinberg proportions in cases. Am J Hum Genet 2008;83:53–63.

27 Wand MP, Jones MC: Kernel Smoothing. London, Chapman and Hall, 1995.

28 Simonoff JS: Smoothing Methods in Statistics. New York, Springer-Verlag, 1996.

29 Devroye L: A Course in Density Estimation. Boston, Birkauser, 1987.

30 Izenman AJ: Recent developments in nonparametric density-estimation. J Am Stat Assoc 1991;86:205–224.

31 Silverman BW: Density Estimation for Statistics and Data Analysis. New York, Chapman and Hall, 1986.

32 Turlach BA. Bandwidth selection in kernel density estimation: a review. 1993. Univ. Catholique de Louvain. Ref Type: Report

33 Matlab. 2002. Cambridge, MA, Mathworks. Ref Type: Computer Program

34 Bowman AW, Azzalini A: Applied Smoothing Techniques for Data Analysis. New York, Oxford University Press, 1997.

35 Hogg RV: Statistical robustness – one view of its use in applications today. Am Stat 1979; 33:108–115.