1-2017

Powerful Method for Including Genotype Uncertainty in Tests of Hardy-Weinberg Equilibrium

Andrew Beck
University of Michigan-Ann Arbor

Alexander Luedtke
University of California - Berkeley

Keli Liu
Harvard University

Nathan L. Tintle
Dordt College, nathan.tintle@dordt.edu

Powerful Method for Including Genotype Uncertainty in Tests of Hardy-Weinberg Equilibrium

Abstract
The use of posterior probabilities to summarize genotype uncertainty is pervasive across genotype, sequencing and imputation platforms. Prior work in many contexts has shown the utility of incorporating genotype uncertainty (posterior probabilities) in downstream statistical tests. Typical approaches to incorporating genotype uncertainty when testing Hardy-Weinberg equilibrium tend to lack calibration in the type I error rate, especially as genotype uncertainty increases. We propose a new approach in the spirit of genomic control that properly calibrates the type I error rate, while yielding improved power to detect deviations from Hardy-Weinberg Equilibrium. We demonstrate the improved performance of our method on both simulated and real genotypes.

Keywords
Hardy-Weinberg formula, probabilities, multiple imputation, statistical hypothesis testing

Disciplines
Genetics and Genomics

Comments
Presented at the 2017 Pacific Symposium on Biocomputing, held on the Big Island of Hawaii, January 3-7, 2017.

A POWERFUL METHOD FOR INCLUDING GENOTYPE UNCERTAINTY IN TESTS OF HARDY-WEINBERG EQUILIBRIUM

ANDREW BECK

Department of Biostatistics, University of Michigan
Ann Arbor, MI, 48109 USA
Email: beckandy@umich.edu

ALEXANDER LUEDTKE

Department of Biostatistics, University of California- Berkeley
Berkeley, CA, 94720 USA
Email: aluedtke@berkeley.edu

KELI LIU

Department of Statistics, Harvard University
Cambridge, MI, 02138 USA
Email: kliu@college.harvard.edu

NATHAN TINTLE

Department of Mathematics, Statistics, and Computer Science, Dordt College
Sioux Center, IA 51250, USA
Email: Nathan.Tintle@dordt.edu

The use of posterior probabilities to summarize genotype uncertainty is pervasive across genotype, sequencing and imputation platforms. Prior work in many contexts has shown the utility of incorporating genotype uncertainty (posterior probabilities) in downstream statistical tests. Typical approaches to incorporating genotype uncertainty when testing Hardy-Weinberg equilibrium tend to lack calibration in the type I error rate, especially as genotype uncertainty increases. We propose a new approach in the spirit of genomic control that properly calibrates the type I error rate, while yielding improved power to detect deviations from Hardy-Weinberg Equilibrium. We demonstrate the improved performance of our method on both simulated and real genotypes.

1. Introduction

With recent advances in high-throughput gene sequencing technologies, it is now possible to obtain measurements on millions of single nucleotide variants (SNVs) throughout the human genome. Large scale genetic data sets, whether from microarray, sequencing or imputation, contain genotype uncertainty which, if unaccounted for in downstream analyses, can significantly decrease power to detect disease-variant associations [1,2] if the uncertainty is not associated with the phenotype, or affect the corresponding type I error rate [3,4] if the uncertainty is associated with the phenotype. To minimize the impact of genotype uncertainty, a standard pre-processing step in most studies is to remove markers that are not in Hardy-Weinberg Equilibrium (HWE), since genotyping errors due to factors like DNA contamination and allelic dropout can cause deviation from HWE [5,6].

The standard approach to testing HWE uses a χ^2_{GOF} test whereby observed genotype frequencies at a variant site are used to obtain maximum likelihood estimates (MLEs) of the minor allele frequency (MAF; f) at the site. A one degree of freedom χ^2_{GOF} statistic is then computed to test the null hypothesis that the observed genotype frequencies follow HWE, namely $(1-f)^2, 2f(1-f)$ and f^2 for the major homozygote, heterozygote and minor homozygote, respectively. While this version of the test is the most straightforward and widely used, alternatives exist including methods for testing HWE in datasets with excess correlation between subject genotypes [7,8], missing genotypes [9] and those that account for covariates [10].

Recently, another alternative HWE testing approach was proposed, $\chi^2_{Posterior}$ [6], which extends the standard χ^2_{GOF} approach to allow for the incorporation of genotype uncertainty. The method has widespread application since for all common genotyping technologies (SNP microarray technology [11], imputation [12] and next-generation sequencing technology [13,14]), probabilistic genotypes are obtained as part of the standard genotype calling pipeline. Such probabilistic genotypes typically take the form of a vector of three posterior probabilities for each individual at each variant site, representing the posterior probability that the individual is actually each of the three possible genotypes. While standard analysis techniques typically "call" genotypes by summarizing the posterior probability by a single discrete genotype (e.g., mode posterior probability), researchers are increasingly using alternative approaches. For example, researchers may use of the entire vector of posterior probabilities or they may use the expected genotype (dosage) [15]. The simulation results of Zheng et al. [15], which were recently made rigorous [16], demonstrate substantial power loss from the use of the modal genotype in many realistic situations and approximately equivalent power from use of the dosage or the entire vector of posterior probabilities in case-control tests of genetic association. These results underscore the importance of considering HWE testing methods, which incorporate genotype uncertainty via the underlying posterior probabilities.

The traditional χ^2_{GOF} makes the key assumption that genotype counts are non-negative integers at each variant site, an assumption that is violated with the inclusion of probabilistic calls. A recently proposed alternative approach, $\chi^2_{Posterior}$, allows for the incorporation of probabilistic genotypes. However, $\chi^2_{Posterior}$ has been shown to be overly conservative (empirical type I error

rate less than nominal) as uncertainty at the variant site increases [6]. In this manuscript, we explore reasons for the conservative nature of $\chi^2_{Posterior}$ and propose an alternative approach to HWE testing which incorporates genotype uncertainty while maintaining the type I error rate at nominal levels. We then evaluate the type I error and power of the new approach across a variety of realistic HWE and non-HWE settings to identify powerful and robust HWE tests for probabilistic genotypes. Finally, we implement the new method on a real data set illustrating its improved ability to maintain the type I error rate, while improving power to detect variants not in HWE.

2. Methods

2.1. Notation

To facilitate the presentation and evaluation of existing and novel approaches to testing for HWE while incorporating genotype uncertainty, we start by defining some basic notation we will use throughout the manuscript. Genotypes for a given individual i can be represented as a vector of three posterior probabilities, $\alpha_i \triangleq (\alpha_{i0}, \alpha_{i1}, \alpha_{i2})$, where α_{ik} is the posterior probability that individual i has k minor alleles, $k = 0,1,2$ at a variant site of interest. The vector of posterior probabilities, α_i, suggests that the true minor allele count for individual i, denoted $x_i \in 0,1,2$, can be modeled as being a single random draw from a multinomial distribution with probabilities indicated by α_i. We assume that α_i is available for each individual.

2.2. Existing approaches to incorporating genotype uncertainty

The most straightforward and widely used approach to manage genotype uncertainty is to summarize the vector of posterior probabilities α_i with the modal genotype, namely, $m_i \triangleq \arg\max_k(\alpha_i)$ in place of the individuals true genotype. When the modal genotype is used as the true genotype, a standard χ^2 goodness of fit test can be used to test for HWE (χ^2_{Mode}). However, when using such a method we expect an increase in the type I error rate and/or decrease in power due to the introduction of genotype errors caused by ignoring the genotype uncertainty represented in the posterior probabilities vector [2,6]. For example, if $\alpha_{i0} = 0.95$ (the mode), we "call" the individual as having no rare alleles and, thus, there is a 5% chance we are incorrect.

A recently proposed test for HWE, $\chi^2_{Posterior}$, utilizes the entire vector of posterior probabilities [6]. This method starts by computing three, non-discrete, genotype counts based on α_i: $A_0^* = \sum_{i=1}^N \alpha_{i0}$, $A_1^* = \sum_{i=1}^N \alpha_{i1}$, and $A_2^* = \sum_{i=1}^N \alpha_{i2}$, where N is the total sample size and we use A^* to represent genotype counts computed by summing the posterior probabilities across the sample. This approach applies a standard χ^2 goodness of fit test as follows

$$\chi^2_{Posterior} = \chi^2_{GOF}(A^*) = N\left[\frac{\left|\frac{A_0^*}{N} - (1-\hat{f})^2\right| - c/N}{(1-\hat{f})^2} + \frac{\left|\frac{A_1^*}{N} - 2(1-\hat{f})\hat{f}\right| - c/N}{2(1-\hat{f})\hat{f}} + \frac{\left|\frac{A_2^*}{N} - (\hat{f})^2\right| - c/N}{(\hat{f})^2}\right] \quad (1)$$

where c is a continuity correction, e.g. 0.5 [17], and where the maximum likelihood estimate (MLE) of the minor allele frequency (MAF), $\hat{f}$, at the site is estimated as $\frac{A_1^* + 2A_2^*}{2N}$. The test uses as its null hypothesis that the variant site is in HWE, and as the alternative hypothesis that the variant

site is not in HWE. This approach uses a central χ^2 distribution with a single degree of freedom as the null distribution for $\chi^2_{Posterior}$.

2.3. Direct likelihood approach

As shown via simulation in prior work [6], and confirmed in our simulations (see *Results*), the $\chi^2_{Posterior}$ test has an overly conservative type I error rate, which becomes more pronounced as genotype uncertainty increases. We now argue that the reason for this overly conservative type I error rate is due to a change in the covariance structure of the genotypes when using probabilistic genotypes (α_i). In particular, the $\chi^2_{Posterior}$ test assumes that each individual genotype occurs according to a multinomial distribution. However, this is no longer the case when observed genotype counts are obtained by summing over the posterior probability vectors [18]. Thus, the covariance structure assumed by the $\chi^2_{Posterior}$ test is not true in practice when using probabilistic genotypes. In situations where the alternative covariance structure due to probabilistic genotypes can be explicitly modeled or otherwise controlled for, likelihood based approaches to testing with uncertain genotypes are possible [15,18]. However, that is not the case for HWE testing, as we explain in the following paragraph.

In particular, in order to develop a likelihood ratio test you must have an explicit expression for the likelihood function of the population genotype frequencies, $G_0, G_1, and\ G_2$. Here the likelihood function can be written as $L(G_0, G_1, G_2; \alpha_1, \dots, \alpha_N) = P(\alpha_1, \dots, \alpha_N | G_0, G_1, G_2) = P(\alpha_1, \dots, \alpha_N | g_1, \dots, g_N, G_0, G_1, G_2) P(g_1, \dots, g_N | G_0, G_1, G_2)$, where g_i indicates the true genotype of individual *i*. Thus, you must have knowledge of the true uncertainty mechanism, $P(\alpha_1, \dots, \alpha_N | g_1, \dots, g_N, G_0, G_1, G_2)$ in order to develop a likelihood ratio test based on the posterior probabilities alone. Because explicit knowledge of the true uncertainty mechanism is unlikely, a likelihood approach to HWE testing using $\alpha_1, \dots, \alpha_N$ will not be possible without making unwarranted assumptions.

2.4. Alternative approach

Because of the overly conservative nature of existing approaches and the limitations we describe above when deriving an explicit likelihood approach, we present an alternative strategy: a post-hoc empirical correction in the spirit of genomic-control. Genomic control [19] is a widely-utilized post-hoc correction factor in genome-wide association studies. When systematic inflation of SNP-association statistics occurs in the data, which can occur due to population stratification or differential genotyping errors, dividing the distribution of observed chi-squared statistics by the median observed chi-squared statistic properly controls the empirical type I error rate. Essentially, this approach assumes that when testing thousands of variant sites for association with the phenotype, the vast majority of sites will not be associated with the phenotype. Thus, the observed distribution of test statistics, aside from the extreme upper-tail, can, in essence, be used as its own null distribution.

To extend the notion of genomic control to HWE testing, we argue that in most real testing situations, the majority of variant sites in a sample of many thousands of variants will be in HWE. Thus, we propose computing $\chi^2_{Posterior,j}$ from A^* as shown above for all variants of interest,

j=1,…,*m*, where *m* is large. Then the measure of inflation/deflation in the null distribution of test statistics is computed as $\hat{\lambda} = \frac{median(\chi^2_{GOF,1},\chi^2_{GOF,2},\ldots,\chi^2_{GOF,m})}{median(\chi^2_1)}$, where $median(\chi^2_1) = 0.455$ [19]. The genomic control-like test statistic for HWE is then computed as $\chi^2_{GC,j} = \frac{\chi^2_{GOF,j}}{\hat{\lambda}}$ for all j=1,…,*m*. We consider four different versions of χ^2_{GC}: $\chi^2_{GC,overall}$, $\chi^2_{GC,MAF}$, χ^2_{GC,r^2} and χ^2_{GC,MAF,r^2}, where $\hat{\lambda}$ is computed on different subsets of the data. Overall indicates that $\hat{\lambda}$ is computed across all *m* SNPs in the set. MAF indicates that $\hat{\lambda}$ is computed separately by MAF group (0.05-0.10, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5). r^2 indicates that $\hat{\lambda}$ is computed separately by r^2 group (0-0.5, 0.5-0.75, 0.75-0.85, 0.85-0.95 and 0.95-1), where r^2 is a measure of genotype uncertainty- see next section for details. And, MAF, r^2 computes $\hat{\lambda}$ in groups defined by both MAF and r^2 (25 separate groups).

2.5. Simulation

We simulated genotype data in order to explore the performance of our proposed new approach under a wide variety of situations. We simulated approximately 850,000 SNPs where HWE was maintained (HWE SNPs). To ensure that the characteristics of this simulation reflected both a realistic allele frequency distribution as well as genotype uncertainty, we randomly sampled (*f, r*^{2}) pairs with replacement from a large dataset of genotypes from the FUSION study [20] that were imputed using MaCH [12]. For each (*f, r*^{2}) pair, we then simulated the 'real' genotypes of 10,000 individuals according to the specified allele frequency, *f*, assuming the population was in Hardy-Weinberg Equilibrium (HWE) $((1-f)^2, 2f(1-f), f^2)$. To model genotype uncertainty at the appropriate level, *r*^{2}, we drew from one of the following Dirichlet distributions conditional on the true genotype [16].

If $g_i = 2$ then $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2}) \sim Dirichlet(aq^2, 2aq(1-q), a(1-q)^2)$
If $g_i = 1$ then $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2}) \sim Dirichlet(aq(1-q), a(1-q)^2 + aq^2, aq(1-q))$
If $g_i = 0$ then $\alpha_i = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2}) \sim Dirichlet(a(1-q)^2, 2aq(1-q), aq^2)$
for $a > 0$ and $0 < q < 1$, where *a* and *q* are chosen to yield a desired *r*^{2} value. This model is chosen to simulate symmetric noise in posterior probabilities while maintaining HWE. Further details are available in *Appendix #1* and elsewhere [16]. In short, parameter *q* is the "average" amount of error. For example, if *q*=0.05 (5% noise/error level in posterior probabilities), then for the major homozygote, $g_i = 2$, $E(\alpha_i) = (\alpha_{i0}, \alpha_{i1}, \alpha_{i2})$ = (0.9025, 0.0.95, 0.0025) and, likewise, if there is no noise/error (*q*=0), then $\alpha_i = (0,0,1)$. Parameter *a* is the variation in the error from person to person. For example, as *a* increases, then $Var(\alpha_i)$ also increases, and so for very small values of a (e.g., *a*=0.01), there is virtually no variation in the values of α_i from person to person.

We also simulated three sets, each with approximately 75,000 SNPs, that were not in HWE (non-HWE SNPs). To do this we randomly sampled two SNPs (*i* and *j*) that were in HWE from the set of 850,000 SNPs described above, keeping track of the difference in the allele frequencies of the two SNPs, $d_{i,j}=f_i-f_j$. We then randomly sampled *n*(1-*k*) individuals from SNP *i* and *nk* individuals from SNP *j*, combining the individuals into a single sample of *n* individuals. We used values of *k*=0.1, 0.3 and 0.5, and continued to use a total sample size of 10,000. Thus, the resulting sample is not in HWE because the observed genotype frequencies were generated from two subpopulations with different allele frequencies.

The three resulting sets of 75,000 simulated SNP genotypes were analyzed using (a) a standard HWE test on the simulated 'real' genotypes (χ^2_{True}), (b) chi-squared on the modal genotype χ^2_{Mode}, (c) the approach utilizing posterior probabilities ($\chi^2_{Posterior}$) and (d) four different GC-like approaches (χ^2_{GC}; see previous section for details). For the purposes of the GC-like approach we combined random subsets of 25,000 non-HWE SNPs with the 850,000 HWE SNPs and applied the adjustment, keeping the total proportion of non-HWE SNPs in the set below 3%.

Type I error rates were computed on the 850,000 HWE SNPs as the proportion of SNPs that were detected to be 'not in HWE' at a particular significance level and for a particular combination of MAF and r^2 levels. Power was computed as the fraction of non-HWE SNPs with a p-value less than the significance level in 300 separate groups created by values of *k (*0.1, 0.2, 0.5*)*, difference in MAF between the two SNPs being mixed together (0.1, 0.1-0.2, 0.2-0.3 or >0.3), observed MAF of the combined variant (0.05-0.10, 0.10-0.20, 0.20-0.30, 0.30-0.40 and 0.40-0.50) and observed r^2 of the combined variant (0-0.50, 0.50-0.75, 0.75-0.85, 0.85-0.95 and 0.95-1.0). We examined significance levels of 0.01, $1x10^{-3}$, and $1x10^{-5}$. We computed power and type I error rates across a variety of subsets of the variants including minor allele frequency, genotype uncertainty (r^2), and deviation from HWE.

2.6. Real data analysis - FUSION

As a proof of concept, we ran χ^2_{Mode}, $\chi^2_{Posterior}$ and χ^2_{GC,MAF,r^2} on 29,361 SNPs imputed with MaCH from chromosome 21 of the FUSION study (n=2456) [20]. We also created 2,377 new variants based on the 29,361 imputed variants, which were out of Hardy-Weinberg equilibrium. These 2,377 new variants were created by first randomly selecting two variants with differences in minor allele frequency of between 0.1 and 0.2 and r-squared values between 0.75 and 0.85. A new variant is created by randomly selecting 10% of the genotypes from one of the variants and 90% from the other. All three Hardy-Weinberg equilibrium tests (χ^2_{Mode}, $\chi^2_{Posterior}$ and χ^2_{GC,MAF,r^2}) were also applied to the 2,377 new non-HWE variants as well. We used a significance level of $1x10^{-5}$ on the 29,361 real and 2,377 new FUSION variants.

Table 1. Overall type I error rates

Method	Significance level		
	0.01	**0.001**	**$1x10^{-5}$**
$\chi^2_{Posterior}$	0.0067	0.00057	$3.5x10^{-6}$
χ^2_{Mode}	0.0134	0.00166	$2.6x10^{-5}$
$\chi^2_{GC,overall}$	0.0112	0.00127	$2.2x10^{-5}$
$\chi^2_{GC,MAF}$	0.0112	0.00128	$2.3x10^{-5}$
χ^2_{GC,r^2}	0.0104	0.0011	$1.3x10^{-5}$
χ^2_{GC,MAF,r^2}	0.0101	0.00105	$1.2x10^{-5}$
χ^2_{True}	0.0099	0.00097	$8.1x10^{-6}$

3. Results

3.1. Type I error simulation

Table 1 gives the overall type I error rates at three different significance levels for each of the six methods applied to posterior probabilities on SNPs in HWE, along with the significance level when using the true genotypes. As expected, use of the true genotypes yields type I error rates at the significance level. Overall, $\chi^2_{Posterior}$ yielded the most conservative type I error rates, while χ^2_{Mode} yielded anti-conservative type I error rates. The χ^2_{GC} corrected approaches tended to yield approximately correct type I error rates, with the version which adjusts statistics both within *MAF* and r^2 (χ^2_{GC,MAF,r^2}) bins providing the best Type I error control. A logistic regression model predicting the type I error rate $\chi^2_{Posterior}$ test across all

850,000 SNPs indicates that both MAF and r^2, as well as an interaction term between MAF and r^2, are significant predictors of the type I error rate, which further supports the necessity to use both bins for both MAF and r^2 when correcting statistics as is done by χ^2_{GC,MAF,r^2}.

The patterns observed in Table 1 remain true across all MAF and r^2 subgroups as shown in Supplemental Table 1. In particular we also see that $\chi^2_{Posterior}$ is the most conservative for less well imputed SNPs, though even well imputed SNPs are treated anti-conservatively by $\chi^2_{Posterior}$ ($8.5\text{x}10^{-3}$ for $r^2>0.95$). In contrast, χ^2_{Mode} is the most anti-conservative for less well imputed SNPs, with some inflation of the type I error rate for moderately well imputed SNPs (e.g., $0.85<r^2<0.95$). χ^2_{Mode} only controls the type I error rate for extremely well imputed SNPs ($r^2>0.95$). χ^2_{GC,MAF,r^2} controls the Type I error rate across MAF and r^2 strata. While Supplemental Table 1 only shows results for a significance level of 0.01, patterns remain the same across other more stringent significance levels (e.g., 0.001, $1\text{x}10^{-5}$, detailed results not shown). Figure 1 illustrates the anti-conservative performance of χ^2_{Mode}, the conservative performance of $\chi^2_{Posterior}$ and good control of the type I error rate by χ^2_{GC,MAF,r^2}

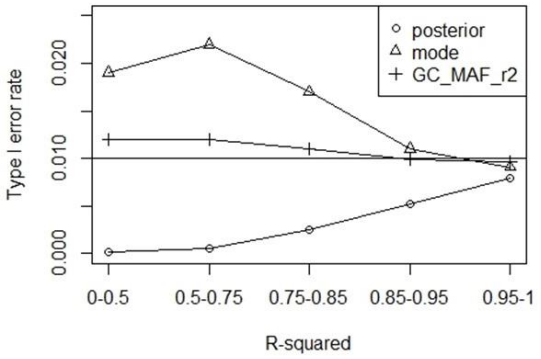

Figure 1. Type I error rate for three different HWE testing methods across different uncertainty levels. Type I error rate is shown across different r^2 settings for three different HWE testing approaches at the 1% significance level. SNPs in the low minor allele frequency range are depicted (MAF between 0.05 and 0.1)

Power simulation

To understand the power of the different approaches for HWE testing, we considered 300 combinations of average minor allele frequency across SNPs *i* and *j*, observed r^2, difference in minor allele frequency and *k* (proportion of individuals from SNP *i*; where *1-k* is the proportion of individuals are from SNP *j*) across 225,000 SNPs which are a mixture of two different allele frequencies. One-hundred twenty-two of the settings yielded 100% power when using all methods, and another 40 combinations yielded no SNPs, and so these 162 settings are eliminated from further consideration. Due to the fact that χ^2_{Mode} has an inflated Type I error rate, we do not consider it in the following comparative analysis of the power of the different methods. Across these 162 settings the median number of SNPs per group was 490 (Min=5; Q1=182; Q3=1708; Max=4353), with only three settings having less than 20 SNPs.

Across the 138 remaining combinations of settings, χ^2_{GC,MAF,r^2} had higher power than $\chi^2_{Posterior}$ 122 times, by an average of 0.038 (SD=0.039). Across the 16 times that $\chi^2_{Posterior}$ yielded higher power than χ^2_{GC,MAF,r^2}, the average power gain was only 0.0029 (SD=0.0024). Table 2 illustrates a subset of 138 simulation settings, illustrating that χ^2_{GC,MAF,r^2} consistently yields higher power than $\chi^2_{Posterior}$ for all but the most certain SNPs, when performance is comparable. Largest gains in power were for the least certain

Table 2. Power[1] by MAF and r^2

MAF	r^2	Number of variants	$\chi^2_{Posterior}$	χ^2_{GC,MAF,r^2}	χ^2_{True}
0.05-0.1	**0-0.50**	122	0.91	0.98	1
	0.5-0.75	265	0.82	0.94	0.98
	0.75-0.85	166	0.79	0.84	0.98
	0.85-0.95	480	0.86	0.87	0.99
	0.95-1.0	695	0.85	0.85	0.99
0.1-0.2	**0-0.50**	123	0.67	0.84	0.86
	0.5-0.75	382	0.65	0.78	0.85
	0.75-0.85	441	0.66	0.76	0.82
	0.85-0.95	1411	0.62	0.65	0.82
	0.95-1.0	2561	0.62	0.61	0.81
0.2-0.3	**0-0.50**	152	0.53	0.66	0.7
	0.5-0.75	365	0.52	0.58	0.7
	0.75-0.85	489	0.56	0.67	0.74
	0.85-0.95	2029	0.53	0.57	0.72
	0.95-1.0	4217	0.52	0.51	0.7
0.3-0.4	**0-0.50**	81	0.43	0.51	0.57
	0.5-0.75	209	0.39	0.46	0.52
	0.75-0.85	277	0.4	0.52	0.58
	0.85-0.95	1324	0.36	0.41	0.54
	0.95-1.0	3321	0.38	0.38	0.53
MAF>0.4	**0-0.50**	25	0.32	0.32	0.44
	0.5-0.75	87	0.29	0.36	0.55
	0.75-0.85	160	0.33	0.43	0.48
	0.85-0.95	629	0.37	0.41	0.51
	0.95-1.0	1649	0.38	0.38	0.52

1. At the 1% significance level and when the observed SNP is a mix of two subgroups of individuals with a difference of between 0.10 and 0.20 in minor allele frequency between the two subgroups, and 10% of the individual are from one subgroup and 90% from the other (k=0.1).

SNPs, with overall higher power for all methods with lower MAF. Figure 2 illustrates this relative gain in power. Supplementary Table 1 gives the full power results for all 300 settings.

Real data example

When applying the three HWE testing methods to the 29,361 imputed FUSION SNPs, 237 variants were determined to be out of HWE by χ^2_{Mode}, none by $\chi^2_{Posterior}$ and two by χ^2_{GC,MAF,r^2} at a significance level of 1x10^{-5}. While true HWE status for these variants is unknown, these results suggest an inflated type I error rate for the χ^2_{Mode} test. When we applied the $\chi^2_{Posterior}$ and χ^2_{GC,MAF,r^2} tests to the 2,377 non-HWE variants, the power was always higher for the χ^2_{GC,MAF,r^2} test (see Table 3).

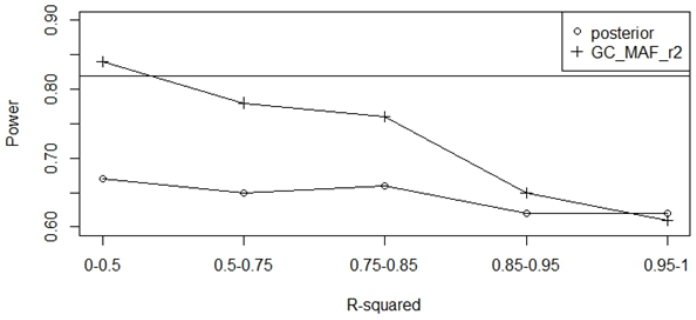

Figure 2 Power for two different approaches to HWE testing across different uncertainty levels
Power is illustrated across different r^2 settings for two different HWE testing approaches at the 1% significance level, with a horizontal line at the power of a test using the real genotypes. Power for SNPs with MAF between 0.1 and 0.2 are depicted, when the observed SNP is a mix of two subgroups of individuals where the difference in MAF between the two subgroups is between 0.1 and 0.2, and the 10% of the individuals are from one subgroup and 90% from the other.

Table 3. Power to detect pseudo variants not in Hardy-Weinberg Equilibrium from the FUSION study

Observed MAF	Number of variants	$\chi^2_{Posterior}$	χ^2_{GC,MAF,r^2}
0.05-0.10	375	5.3%	8.0%
0.10-0.20	731	28.6%	29.7%
0.20-0.3	412	28.9%	30.8%
0.3-0.4	385	26.8%	32.2%
0.4-0.5	374	2.9%	5.9%
Overall	2277	20.3%	22.8%

4. Discussion

We have proposed a new way to incorporate posterior probabilities in tests of HWE that provides a well-calibrated and more powerful way to incorporate genotype uncertainty. While it is common to use the modal posterior genotype, this approach inflates the type I error rate by failing to incorporate genotype uncertainty---treating uncertain genotypes as if they are error-free. Furthermore, another recent approach which explicitly incorporates posterior probabilities yields an overly conservative test (deflated type I error rate), due to an overestimation of the covariance of the posterior probability genotypes. Our approach applies a post-hoc correction to adjust the test statistic, yielding a calibrated type I error rate and improved power.

The proposed approach is approximately the same as other approaches when genotype uncertainty is low, but shows increasing benefit as genotype uncertainty increases. This result is in line with the fact that the genotype covariance estimates are increasingly biased when using $\chi^2_{Posterior}$ as genotype uncertainty increases. While it is common practice to simply drop markers with very high genotype uncertainty from analyses we've demonstrated that this may not be

necessary when using our approach. Furthermore, even if practitioners wish to drop markers with high genotype uncertainty (e.g., r^2<0.5), we've demonstrated that our approach to HWE testing still outperforms other HWE testing procedures for markers with modest genotype uncertainty (0.5<r^2<0.95). Importantly, recent work has shown that simply screening for HWE using r^2 is not sufficient, and that HWE testing is still necessary [21].

While the proposed approach performs well relative to the existing approaches by applying a post-hoc correction, a more explicit approach may also be possible. Preliminary exploration of such methods by our group has taken two separate paths to date. First, we considered multiple imputation by creating many, equally likely, versions of each individual's genotype according to the vector of calibrated posterior genotype probabilities and then computing the standard chi-squared GOF test on each multiply-imputed dataset. Methods for computing significance from a set of multiply-imputed datasets are standard [22–24], but may not be well-calibrated [25]. A lack of calibration was our experience for this application (detailed results not shown). A second approach is a Bayesian approach using the posterior probabilities for each individual's genotype explicitly. Evaluation of this method across a wide-range of simulation settings showed performance comparable to the $\chi^2_{Posterior}$ method and, thus, not as good as χ^2_{GC,MAF,r^2} in many cases (detailed results not shown).

We now make some important notes and comments on limitations of the χ^2_{GC,MAF,r^2}approach. While not considered here, the authors of the $\chi^2_{Posterior}$ approach also considered an exact test for small sample sizes. Future work is needed to evaluate the performance of the post-hoc correction strategy for small sample size situations (e.g., rare variants), though, in principle, there is no reason to believe that an approach in this same spirit is likely to perform well. A key assumption of χ^2_{GC,MAF,r^2} is that a relative small proportion of all markers overall will not be in HWE. In rare cases where a very large proportion of markers are out of HWE, the χ^2_{GC,MAF,r^2} approach may, in fact, be overly conservative by applying a correction factor based on markers not in HWE. However, these cases should be rare as a substantial portion of the markers in the correction set would need to be out of HWE in order to impact the median observed statistic and, hence, the lambda, in a practically significant way. However, since χ^2_{GC,MAF,r^2} computes a separate adjustment for many different MAF, r^2 'bins,' an aggregation of markers not in HWE in any bin could impact results. Finally, the size and quantity of MAF, r^2 bins selected in this study showed good performance, but may need adjustment in practice based on the MAF distribution, r^2 (or other uncertainty metric) distribution and number of variants. Care should be taken to ensure all bins have sufficient markers (generally recommended to be at least 100, but less may be fine) and examination of $\hat{\lambda}$ values within each bin is recommended. Future work may wish to explore the potential for a robust, continuous correction strategy.

Supplemental Files

All supplemental and appendix files are available online at the following URL: http://homepages.dordt.edu/ntintle/hwe.zip

Acknowledgments

This work was funded by the National Human Genome Research Institute (R15HG006915). We acknowledge the use of the Hope College parallel computing cluster for assistance in data analysis. We are grateful for the generosity of the FUSION investigators for access to imputed genotypes on which we based our simulations and demonstrated a proof-of-concept of our proposed approach. The FUSION study is funded by the National Institute of Diabetes and Digestive and Kidney Diseases (U01DK062370).

References

1. Powers S, Gopalakrishnan S, Tintle N. Assessing the impact of non-differential genotyping errors on rare variant tests of association. Hum Hered. 2011;72: 153–60.

2. Gordon D, Finch SJ. Factors affecting statistical power in the detection of genetic association. J Clin Investig. 2005;115.

3. Mayer-Jochimsen M, Fast S, Tintle NL. Assessing the impact of differential genotyping errors on rare variant tests of association. PLoS One. 2013;8: e56626.

4. Moskvina V, Craddock N, Holmans P, Owen MJ, O'Donovan MC. Effects of differential genotyping error rate on the type I error probability of case-control studies. Hum Hered. 2006;61: 55–64.

5. Wang J, Shete S. Testing Hardy-Weinberg proportions in a frequency-matched case-control genetic association study. PLoS One. 2011;6: e27642.

6. Shriner D. Approximate and exact tests of Hardy-Weinberg equilibrium using uncertain genotypes. Genet Epidemiol. 2011;35: 632–7.

7. Li Y. A comparison of tests for Hardy-Weinberg Equilibrium in national genetic household surveys. BMC Genet. 2013;14: 14.

8. She D, Zhang H, Li Z. Testing Hardy-Weinberg equilibrium using family data from complex surveys. Ann Hum Genet. 2009;73: 449–55.

9. Graffelman J, Nelson S, Gogarten SM, Weir BS. Exact Inference for Hardy-Weinberg Proportions with Missing Genotypes: Single and Multiple Imputation. G3 Genes|Genomes|Genetics. 2015;5: g3.115.022111.

10. Schaid DJ, Sinnwell JP, Jenkins GD. Regression Modeling of Allele Frequencies and Testing Hardy Weinberg Equilibrium. Hum Hered. 2013;74: 71–82.

11. Hong H, Su Z, Ge W, Shi L, Perkins R, Fang H, et al. Evaluating variations of genotype calling: a potential source of spurious associations in genome-wide association studies. J Genet. 2010;89: 55–64.

12. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: Using sequence and genotype data to esimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34: 816–834.

13. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18: 1851–1858.

14. Nielson R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet. 2011;12: 443–451.

15. Zheng J, Li Y, Abecasis GR, Scheet P. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genet Epidemiol. 2011;35: 102–10.

16. Liu K, Luedtke A, Tintle NL. optimal methods for using posterior probabilities in association testing. Hum Hered. 2013;75: 2–11.

17. Yates F. Contingency table involving small numbers and the X2 test. Suppl to J Roayl Stat Soc. 1934;1: 217–235.

18. Tintle N, Gordon D, McMahon F, Finch SJ. Using Duplicate Genotyped Data in Genetic Analyses : Testing Association and Estimating Error Rates. Stat Appl Genet Mol Biol. 2007;6.

19. Devlin B, Roeder K. Genomic Control for Association Studies. Biometrics. 1999;55: 997–1004.

20. Scott L, Mohlke K, Bonnycastle L, Willer C, Li Y, Duren W, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science (80-). 2007;316: 1341–5.

21. Shriner D. Impact of Hardy-Weinberg disequilibrium on post-imputation quality control. Hum Genet. 2013;132: 1073–5.

22. Li K-H, Meng X-L, Raghunathan TE, Rubin DB. Signifiacnce levels from repeated p-values with multiply imputed data. Stat Sin. 1991;1: 65–92.

23. Meng X-L, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. Biometrika. 1992;79: 103–111.

24. Harel O, Zhou X-H. Multiple imputation: review of theory, implementation and software. Stat Med. 2007;26: 3057–3077.

25. Licht C. New methods for generating significance levels from multiply-imputed data Ph.D. Dissertation. University of Bamberg, Germany. 2010.