Published in final edited form as:

Genet Epidemiol. 2011 November ; 35(7): 632–637. doi:10.1002/gepi.20612.

Approximate and Exact Tests of Hardy-Weinberg Equilibrium Using Uncertain Genotypes

Daniel Shriner*

Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, Maryland

Abstract

Testing for Hardy-Weinberg equilibrium (HWE) is commonly used as a quality control filter in genome-wide scans for markers with experimentally determined genotypes. In contrast, for markers with imputed genotypes, there are post-imputation metrics of quality that can be used as screens but there are no formal tests of deviation from HWE. Similarly, there are no formal tests of deviation from HWE for probabilistic genotypes that are generated by sequencing projects. Here, I describe generalizations of the approximate χ^2 and exact tests of HWE for use with uncertain genotypes. The tests fully account for the probabilities of all possible genotypes at a marker for each individual. By computer simulation, the approximate and exact tests are shown to maintain valid control of the type I error rate. Calculations of the loss of power as the uncertainty in genotypes increases are illustrated. The tests are compatible with chip-based genotypes for single-nucleotide polymorphisms and copy number polymorphisms, imputed genotypes, and probabilistic assignments of genotype from variable-coverage sequence data.

Keywords

quality control; measurement error; genotype probabilities

INTRODUCTION

Hardy-Weinberg equilibrium (HWE) is an application of the binomial theorem to population genetics [Hardy, 1908; Weinberg, 1908]. Removing markers that are not in HWE is a standard pre-processing quality control step in genome-wide scans [Gomes et al., 1999]. Many tests of HWE have been considered, including approximate χ^2 tests [Emigh, 1980], exact tests [Levene, 1949; Haldane, 1954; Chapco, 1976; Louis and Dempster, 1987; Wellek, 2004; Wigginton et al., 2005; Schaid et al., 2006], likelihood tests [Elston and Forthofer, 1977; Li and Li, 2008; Yu et al., 2009], Bayesian tests [Lindley, 1988; Montoya-Delgado et al., 2001; Pereira et al., 2006], and Monte Carlo-based methods [Guo and Thompson, 1992; Lazzeroni and Lange, 1997]. The data required to evaluate any of these tests are the observed genotype counts.

Two methods for high-throughput generation of genome-wide genotypes are microarray-based hybridization and sequencing. For both methods, genotype calling algorithms produce genotypes that are assigned probabilistically but that are typically provided to end-users as discrete calls [Korn et al., 2008; The 1000 Genomes Project Consortium, 2010]. Whereas the genotyping error rates for chip-based typing and high-coverage sequencing are generally low, the error rate for low-coverage sequencing, e.g., with read depths of $2\times-4\times$, is considerably higher [The 1000 Genomes Project Consortium, 2010]. To account for uncertainty, Bayesian genotype calling algorithms calculate posterior probabilities for all possible genotypes at a marker [Bentley et al., 2008; Browning and Yu, 2009; McKenna et al., 2010; Li et al., 2011; Le and Durbin, 2011]. When there are too few reads to call a genotype unambiguously, the genotype can be discarded and subsequently treated as missing or the genotype can be retained and uncertainty as reflected by the posterior probabilities can be taken into account in downstream analyses.

Imputation of missing genotypes has become a standard part of genome-wide association studies [Marchini and Howie, 2010]. There are two basic uses of imputation: to fill in missing genotypes among otherwise experimentally determined genotypes at a marker or to fill in all genotypes for a completely untyped marker. In either case, imputation yields the posterior probabilities for all possible genotypes, just as with Bayesian genotype calling algorithms. Although four post-imputation information measures have been developed to assess the quality of imputed genotypes for completely untyped SNPs [Marchini and Howie, 2010], a critical void exists in that there are no formal hypothesis tests for post-imputation quality control.

A key assumption of traditional tests of HWE is that the genotype counts are non-negative integers. However, due to uncertainty, genotypes are not discretely countable. By extending previous work on models of genotype error [Gordon et al., 2002; Leal, 2005; Cox and Kraft, 2006], I derive both approximate and exact tests of HWE that make full use of the distribution of genotype probabilities. I show that both approximate and exact tests are valid, in the sense that they maintain appropriate control of the type I error rate. Using the approximate χ^2 test, I illustrate the loss of power as the amount of uncertainty increases.

METHODS

Consider a marker with minor allele A and major allele B . Let p_A be the frequency of allele A and let p_B be the frequency of allele B , with the constraint that $p_B = 1 - p_A$. Let n_{AA} , n_{AB} , and n_{BB} be the number of individuals with genotypes AA , AB , and BB , respectively. For a sample size N , $N = n_{AA} + n_{AB} + n_{BB}$. Let n_A be the number of A alleles and let n_B be the number of B alleles, subject to the constraint that $n_A + n_B = 2N$. Furthermore, $n_{AA} = (n_A - n_{AB})/2$ and $n_{BB} = (n_B - n_{AB})/2$. The estimate of the frequency of allele A is $\hat{p}_A = (n_{AA} + 0.5n_{AB})/N$.

Under the assumption of HWE, the expected probability of AA genotypes is $\hat{p}_A^2$, the expected frequency of AB heterozygotes is $2\hat{p}_A\hat{p}_B = 2\hat{p}_A(1 - \hat{p}_A)$, and the expected frequency of BB homozygotes is $\hat{p}_B^2 = (1 - \hat{p}_A)^2$. For the approximate χ^2 test,

$$\chi^2 = \frac{(|n_{AA} - \hat{p}_A^2 N| - c)^2}{\hat{p}_A^2 N} + \frac{(|n_{AB} - 2\hat{p}_A(1 - \hat{p}_A)N| - c)^2}{2\hat{p}_A(1 - \hat{p}_A)N} + \frac{(|n_{BB} - (1 - \hat{p}_A)^2 N| - c)^2}{(1 - \hat{p}_A)^2 N},$$

in which c is a continuity correction such as 0.5 [Yates, 1934] or 0.25 [Emigh, 1980]. This test has one degree of freedom. For the exact test,

$$P(n_{AB}|n_A, N) = \frac{2^{n_{AB}} N! n_A! n_B!}{n_{AA}! n_{AB}! n_{BB}! (2N)!}.$$

The P -value for the two-tailed test is determined by summing the table probabilities for all tables with probabilities smaller than the probability of the observed table.

The approximate and exact tests can be generalized to allow for nondiscrete genotype data. For the approximate test, let $p_{AA} = n_{AA}/N$, $p_{AB} = n_{AB}/N$, and $p_{BB} = n_{BB}/N$ be the estimates of the genotype probabilities. Then,

$$\chi^2 = N \left[\frac{\left(|\hat{p}_{AA} - \hat{p}_A^2| - \frac{c}{N} \right)^2}{\hat{p}_A^2} + \frac{\left(|\hat{p}_{AB} - 2\hat{p}_A(1 - \hat{p}_A)| - \frac{c}{N} \right)^2}{2\hat{p}_A(1 - \hat{p}_A)} + \frac{\left(|\hat{p}_{BB} - (1 - \hat{p}_A)^2| - \frac{c}{N} \right)^2}{(1 - \hat{p}_A)^2} \right].$$

The approximate tests are based on the differences of the observed and expected genotype probabilities, with no requirement for discrete genotype counts. For the exact test, by definition $x! = \Gamma(x+1) = \int_0^\infty t^x e^{-t} dt$. The argument of $\Gamma(x+1)$ need not be an integer, again relaxing the requirement for discrete genotype counts.

Let $\mathbf{G}$ be the $N \times 3$ genotype matrix, in which g_{i1} is the probability of genotype AA for the i th individual, g_{i2} is the probability of genotype AB for the i th individual, and g_{i3} is the probability of genotype BB for the i th individual, subject to the constraint that $g_{i1} + g_{i2} + g_{i3} = 1$. Then, $n_{AA} = \sum_{i=1}^N g_{i1}$, $n_{AB} = \sum_{i=1}^N g_{i2}$, and $n_{BB} = \sum_{i=1}^N g_{i3}$. The approximate test follows immediately. For the exact test, there is one additional issue. When n_A is odd, the possible values of n_{AB} are $n_A, n_A-2, \dots, 1$. Similarly, when n_A is even, the possible values of n_{AB} are $n_A, n_A-2, \dots, 0$. If $n_A = 2n_{AA} + n_{AB}$ is not an integer, the appropriate sequence is $n_A, n_A-2, \dots$, stopping at the last value > 0 . The exact test then follows.

For the exact test, calculations can be simplified by taking advantage of recurrence [Guo and Thompson, 1992; Wigginton et al., 2005]. Specifically,

$$P(n_{AB} - 2|n_A, N) = P(n_{AB}|n_A, N) \frac{n_{AB}(n_{AB} - 1)}{4(n_{AA} + 1)(n_{BB} + 1)}.$$

Using this relationship, the recurrence is initialized by calculating the table probability corresponding to the largest possible number of heterozygotes. All other table probabilities are then calculated recursively.

To model uncertainty in genotypes, consider a model in which errors are introduced randomly and independently for alleles. Let the error model consist of ε_1 , the probability of observing B when A is the true allele, and ε_2 , the probability of observing A when B is the true allele [Gordon et al., 2002]. The error-contaminated probability of the AA genotype is $(1 - \varepsilon_1)^2 p_A^2 + \varepsilon_2(1 - \varepsilon_1)2p_A(1 - p_A) + \varepsilon_2^2(1 - p_A)^2$, the error-contaminated probability of the AB genotype is $2\varepsilon_1(1 - \varepsilon_1)p_A^2 + [\varepsilon_1\varepsilon_2 + (1 - \varepsilon_1)(1 - \varepsilon_2)]2p_A(1 - p_A) + 2\varepsilon_2(1 - \varepsilon_2)(1 - p_A)^2$, and the error-contaminated probability of the BB genotype is $\varepsilon_1^2 p_A^2 + \varepsilon_1(1 - \varepsilon_2)2p_A(1 - p_A) + (1 - \varepsilon_2)^2(1 - p_A)^2$.

All calculations were performed in R, version 2.12.0 [R Development Core Team, 2010]. R functions are provided as Supplementary Material.

RESULTS

Table I illustrates the calculations using the classical tests of HWE with discrete genotype data. As is well known, a small number of distinct configurations exist. This granularity leads to the exact test being conservative. For small numbers of heterozygotes, the approximate tests overestimate significance relative to the exact test. These features also apply to the tests with nondiscrete genotype data (Table II).

The validity of the tests was assessed by simulating under the null hypothesis of HWE. For a given minor allele frequency and sample size, all tests showed an empirical type I error rate below the significance level, indicating that all tests were valid but conservative (Fig. 1). All tests became more conservative as the genotype uncertainty increased (Fig. 1).

To illustrate the loss of power for the approximate χ^2 test as genotype uncertainty increases, consider the most extreme configuration wherein all individuals are heterozygotes. In this configuration, $p_{AA} = 0$, $p_{AB} = 1$, $p_{BB} = 0$, and $p_A = 0.5$. Then, the test statistic is

$$N \left[\frac{(0 - 0.25)^2}{0.25} + \frac{(1 - 0.5)^2}{0.5} + \frac{(0 - 0.25)^2}{0.25} \right] = N[1].$$

Now incorporate genotype uncertainty such that $p_{AA} = 0.025$, $p_{AB} = 0.95$, and $p_{BB} = 0.025$. The estimate p_A remains 0.5. The test statistic is

$$N \left[\frac{(0.025 - 0.25)^2}{0.25} + \frac{(0.95 - 0.5)^2}{0.5} + \frac{(0.025 - 0.25)^2}{0.25} \right] = N[0.81].$$

To maintain power, the sample size has to increase by $1/0.81 \approx 1.23$ fold, or 23%. Similarly, if $p_{AA} = 0.05$, $p_{AB} = 0.90$, and $p_{BB} = 0.05$, the sample size has to increase by 56% to maintain power. Identical results hold for the other most extreme configuration comprising no heterozygotes.

Using the allelic error model, as ε_1 and ε_2 both approach 0.5, p_{AA} approaches 0.25, p_{AB} approaches 0.5, and p_{BB} approaches 0.25, regardless of the true underlying genotype

probabilities. Consequently, the tests lose power as ε_1 and ε_2 both approach 0.5 and the tests have no power if ε_1 and ε_2 both equal 0.5, regardless of either the sample size or the true underlying genotype and allele probabilities.

The uncertainty-aware tests require more memory in order to store the genotype probabilities. However, after calculating n_{AA} , n_{AB} , and n_{BB} , the algorithmic costs for performing the uncertainty-aware approximate and exact tests are identical to the algorithmic costs of the conventional approximate and exact tests, respectively.

DISCUSSION

The Hardy-Weinberg law provides a theoretical starting point for much of population genetics. Testing HWE for genotyped markers is a standard quality control step in genome-wide association studies. Imputation has also become a standard part of genome-wide association studies. However, quality control measures for untyped markers are not as well developed as they are for genotyped markers. This study describes approximate and exact tests of HWE using uncertain genotypes. These tests may be applied to genotyped single-nucleotide polymorphisms in the presence of genotyping error, genotyped copy-number polymorphisms with probabilistic assignments of copy-number states, and untyped markers with probabilistic assignment of genotype via imputation. These tests may also be applied to probabilistic genotype calls produced by sequencing data, which is especially important for low-coverage data.

Both the approximate and exact tests are valid but conservative. There are two reasons for this behavior. One, both approximate and exact tests are increasingly conservative as the minor allele frequency decreases. Two, the exact test is conservative because the number of genotypic configurations conditioned on marginal allele frequencies and the sample size is discrete, even if the genotype data are not discrete.

The approximate and exact tests for HWE derived in this study all make full use of the distribution of genotype probabilities, similar to tests of association developed for imputed genotypes [Marchini et al., 2007]. Alternatively, one could use the genotype with the highest probability (i.e., the genotype corresponding to the posterior mode), which would yield discrete, non-negative counts. Two problems with this approach are that it is not apparent how to break ties and that the type I error rate increases as uncertainty increases. Another approach would be to use the allelic dosage (i.e., the posterior expectation). This approach also suffers from an inflated type I error rate as uncertainty increases.

In summary, there are no formal tests for quality control for uncertain genotypes. To address this problem, I derived approximate and exact tests of HWE that are based on genotype probabilities rather than discrete genotype counts. By computer simulation, the tests are shown to be valid across the range of minor allele frequencies with varying amounts of genotype uncertainty. The power to detect deviations from HWE decreases as genotype uncertainty increases. Equivalently, the sample size needs to be increased in order to maintain power in the presence of genotype uncertainty. The tests are compatible with chip-based genotypes for single-nucleotide polymorphisms and copy number polymorphisms,

imputed genotypes, and probabilistic assignments of genotype from variable-coverage sequence data. To avoid violating the simplifying assumptions of the approximate χ^2 tests [Graffelman, 2010], the exact test is preferred.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The contents of this publication are solely the responsibility of the author and do not necessarily represent the official view of the National Institutes of Health. This research was supported by the Intramural Research Program of the Center for Research on Genomics and Global Health (CRGGH). The CRGGH is supported by the National Human Genome Research Institute, the National Institute of Diabetes and Digestive and Kidney Diseases, the Center for Information Technology, and the Office of the Director at the National Institutes of Health (Z01HG200362).

REFERENCES

- Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Cheetham RK, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IMJ, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DMD, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Catenazzi MCE, Chang S, Cooley RN, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Furey WS, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Jones TAH, Kang G-D, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ng BL, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Pinkard DC, Pliskin DP, Podhasky J, Quijano VJ, Racz C, Rae VH, Rawlings SR, Rodriguez AC, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Sohna JE, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, vandeVondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008; 456:53–59. [PubMed: 18987734]
- Browning BL, Yu Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet*. 2009; 85:847–861. [PubMed: 19931040]
- Chapco W. An exact test of the Hardy-Weinberg law. *Biometrics*. 1976; 32:183–189. [PubMed: 1276367]
- Cox DG, Kraft P. Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Hum Hered*. 2006; 61:10–14. [PubMed: 16514241]
- Elston RC, Forthofer R. Testing for Hardy-Weinberg equilibrium in small samples. *Biometrics*. 1977; 33:536–542.
- Emigh TH. A comparison of tests for Hardy-Weinberg Equilibrium. *Biometrics*. 1980; 36:627–642.

- 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. [PubMed: 11246455]
- 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. [PubMed: 12446984]
- Graffelman J. The number of markers in the HapMap project: some notes on chi-square and exact tests for Hardy-Weinberg equilibrium. *Am J Hum Genet.* 2010; 86:813–818. author reply 818–819. [PubMed: 20466092]
- Guo SW, Thompson EA. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics.* 1992; 48:361–372. [PubMed: 1637966]
- Haldane JBS. An exact test for randomness of mating. *J Genet.* 1954; 52:631–635.
- Hardy GH. Mendelian proportions in a mixed population. *Science.* 1908; 28:49–50. [PubMed: 17779291]
- Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler KD. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet.* 2008; 40:1253–1260. [PubMed: 18776909]
- Lazzeroni LC, Lange K. Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. *Ann Stat.* 1997; 25:138–168.
- Le SQ, Durbin R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* 2011; 21:952–960. [PubMed: 20980557]
- Leal SM. Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. *Genet Epidemiol.* 2005; 29:204–214. [PubMed: 16080207]
- Levene H. On a matching problem arising in genetics. *Ann Math Stat.* 1949; 20:91–94.
- Li M, Li C. Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. *Genet Epidemiol.* 2008; 32:589–599. [PubMed: 18449919]
- Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* 2011; 21:940–951. [PubMed: 21460063]
- Lindley DV. Statistical inference concerning Hardy-Weinberg equilibrium. *Bayesian Stat.* 1988; 3:307–326.
- Louis EJ, Dempster ER. An exact test for Hardy-Weinberg and multiple alleles. *Biometrics.* 1987; 43:805–811. [PubMed: 3427165]
- Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet.* 2010; 11:499–511. [PubMed: 20517342]
- Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.* 2007; 39:906–913. [PubMed: 17572673]
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297–1303. [PubMed: 20644199]
- Montoya-Delgado LE, Irony TZ, Pereira CA de B, Whittle MR. An unconditional exact test for the Hardy-Weinberg equilibrium law: sample-space ordering using the Bayes factor. *Genetics.* 2001; 158:875–883. [PubMed: 11404348]
- Pereira CA de B, Nakano F, Stern JM, Whittle MR. Genuine Bayesian multiallelic significance test for the Hardy-Weinberg equilibrium law. *Genet Mol Res.* 2006; 5:619–631. [PubMed: 17183473]
- R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.
- Schaid DJ, Batzler AJ, Jenkins GD, Hildebrandt MAT. Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata. *Am J Hum Genet.* 2006; 79:1071–1080. [PubMed: 17186465]
- The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature.* 2010; 467:1061–1073. [PubMed: 20981092]

- Weinberg W. Über den Nachweis der Vererbung beim Menschen. *Jahresh Ver Vaterl Naturkd Württemb.* 1908; 64:369–382.
- Wellek S. Tests for establishing compatibility of an observed genotype distribution with Hardy-Weinberg equilibrium in the case of a biallelic locus. *Biometrics.* 2004; 60:694–703. [PubMed: 15339292]
- Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet.* 2005; 76:887–893. [PubMed: 15789306]
- Yates F. Contingency tables involving small numbers and the χ^2 test. *J R Stat Soc Suppl.* 1934; 1:217–235.
- 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. [PubMed: 19025784]

Fig. 1.

Type I error rates at a significance level of 0.05. Black represents the uncorrected χ^2 test, red represents the χ^2 test with the Yates continuity correction, and blue represents the exact test. (A and B) The error rates ε_1 and ε_2 were both 0. (C and D) The error rates ε_1 and ε_2 were both 0.01. (E and F) The error rates ε_1 and ε_2 were both 0.05. (G and H) The error rates ε_1 and ε_2 were both 0.10. (A, C, E, and G) The sample size was 100. (B, D, F, and H) The sample size was 1,000. Type I error rates were estimated from 10,000 independent replicates

with random sampling of the multinomial distribution given the sample size and the minor allele frequency, assuming Hardy-Weinberg equilibrium.

TABLE I

Sample configurations and probabilities for discrete genotype data

n_{AB}	$P(n_{AB} n_A=20, N=100)^a$	Exact P -value	Approximate P -value	Yates P -value
0	1.07×10^{-14}	1.07×10^{-14}	1.52×10^{-23}	2.20×10^{-21}
2	1.93×10^{-11}	1.93×10^{-11}	6.17×10^{-19}	5.12×10^{-17}
4	5.16×10^{-9}	5.18×10^{-9}	7.38×10^{-15}	3.52×10^{-13}
6	4.84×10^{-7}	4.89×10^{-7}	2.62×10^{-11}	7.18×10^{-10}
8	2.11×10^{-5}	2.15×10^{-5}	2.77×10^{-8}	4.39×10^{-7}
10	0.000483	0.000504	8.81×10^{-6}	8.14×10^{-5}
12	0.00622	0.00672	0.000858	0.00466
14	0.0459	0.0526	0.0263	0.0854
16	0.191	0.243	0.267	0.539
18	0.409	1.000	1.000	0.605
20	0.348	0.591	0.267	0.539

^aNote that the table probabilities sum to 1.

TABLE II

Sample configurations and probabilities for nondiscrete genotype data

n_{AB}	$P(n_{AB} n_A=20.5, N=100)^a$	Exact P -value	Approximate P -value	Yates P -value
0.5	5.52×10^{-14}	5.52×10^{-14}	2.29×10^{-22}	2.57×10^{-20}
2.5	5.27×10^{-11}	5.27×10^{-11}	5.56×10^{-18}	3.69×10^{-16}
4.5	1.07×10^{-8}	1.07×10^{-8}	4.22×10^{-14}	1.65×10^{-12}
6.5	8.35×10^{-7}	8.45×10^{-7}	9.99×10^{-11}	2.31×10^{-9}
8.5	3.17×10^{-5}	3.26×10^{-5}	7.44×10^{-8}	1.02×10^{-6}
10.5	0.000652	6.85×10^{-4}	1.76×10^{-5}	0.000144
12.5	0.00767	0.00835	0.00135	0.00664
14.5	0.0523	0.0607	0.0341	0.103
16.5	0.203	0.263	0.302	0.583
18.5	0.408	1.000	0.956	0.652
20.5	0.329	0.592	0.253	0.511

^aNote that the table probabilities sum to 1.