Theor Appl Genet (2007) 115:393–398
DOI 10.1007/s00122-007-0573-6

ORIGINAL PAPER

An incomplete enumeration algorithm for an exact test of Hardy–Weinberg proportions with multiple alleles

H. P. Maurer · A. E. Melchinger · M. Frisch

Received: 17 October 2006 / Accepted: 27 April 2007 / Published online: 3 July 2007

Abstract Testing of Hardy–Weinberg proportions (HWP) with asymptotic goodness-of-fit tests is problematic when the contingency table of observed genotype counts has sparse cells or the sample size is low, and exact procedures are to be preferred. Exact *p*-values can be (1) calculated via computational demanding enumeration methods or (2) approximated via simulation methods. Our objective was to develop a new algorithm for exact tests of HWP with multiple alleles on the basis of conditional probabilities of genotype arrays, which is faster than existing algorithms. We derived an algorithm for calculating the exact permutation significance value without enumerating all genotype arrays having the same allele counts as the observed one. The algorithm can be used for testing HWP by (1) summation of the conditional probabilities of occurrence of genotype arrays with smaller probability than the observed one, and (2) comparison of the sum with a nominal Type I error rate α. Application to published experimental data from seven maize populations showed that the exact test is computationally feasible and reduces the number of enumerated genotype count matrices about 30% compared with previously published algorithms.

Communicated by D. A. Hoisington.

H. P. Maurer · A. E. Melchinger (✉) · M. Frisch
Institute of Plant Breeding, Seed Science,
and Population Genetics, University of Hohenheim,
70593 Stuttgart, Germany
e-mail: melchinger@uni-hohenheim.de

Introduction

The Hardy–Weinberg law (Hardy 1908; Weinberg 1908) states that in a large random mating population of diploid individuals in the absence of mutation, selection, and drift (1) allele and genotype frequencies remain constant from generation to generation and (2) for an autosomal locus with m alleles A_1, A_2, ..., A_m the expected genotype frequencies are

$$\begin{aligned} P_{ii} &= p_i^2 \quad \text{for homozygotes } A_iA_i \text{ and} \\ P_{ij} &= 2p_ip_j \quad \text{for heterozygotes } A_iA_j \end{aligned} \tag{1}$$

where p_i is the allele frequency of A_i. Testing H_0: "The genotype frequencies in a population follow the distribution described by Eq. 1" are commonly referred to as tests for Hardy–Weinberg proportions (HWP, Weir 1996). The assumption of HWP is the basis of many concepts in population genetics and quantitative genetics (Crow 1988). For example, HWP tests can be applied (1) to gather information on the mating system and genetic structure of wild and breeding populations (Semerikov et al. 2002; Reif et al. 2004), (2) to detect population admixture (Deng et al. 2001), and (3) to detect marker phenotype associations (Nielsen et al. 1999). Therefore, tests of HWP are of crucial importance in plant, animal and human genetics as well as evolutionary studies.

In principle two approaches are possible to test for HWP: (1) Asymptotic goodness-of-fit tests, such as χ^2 or likelihood ratio tests, for which the distribution of the test statistic under the null hypothesis H_0 is approximately known for large samples. (2) Exact tests based on the probability of occurrence of genotype arrays (Weir 1996, chap. 3).

Asymptotic goodness-of-fit tests are computationally simple and fast, yet it is known that they can lead to false rejection or non-rejection of the null-hypothesis of HWP, especially with small sample sizes and/or sparse cells (Elston and Forthofer 1977). Several corrections for small sample sizes have been proposed, but in general these do not substantially improve the performance of the tests (Emigh 1980; Hernández and Weir 1989). With the advent of highly polymorphic DNA markers such as SSRs, an additional problem has emerged. It is not unusual that more than two alleles per marker locus are observed in population samples, most of which occur with rather low frequencies. Consequently, for many cells of the contingency table of observed genotype counts, only a few or even no individuals are observed. Applying goodness-of-fit tests to such data is problematic, because small expected cell counts can inflate the test statistic (Hernández and Weir 1989). In such cases, alleles with low frequencies are often pooled to meet the prerequisites of the test, yet this is expected to result in a loss of power.

Exact tests are computationally demanding, but they are preferred over asymptotic goodness-of-fit tests when the sample size is small and/or table cells are sparse, because they do not require large sample assumptions. When the computational demand exceeds the available capacities, the distribution of the test statistic under the null hypothesis can be approximated with Monte Carlo methods (Guo and Thompson 1992; Huber et al. 2006).

Wellek (2004) constructed a computationally efficient exact HWP test for loci with two alleles. For loci with $m = 2, 3, 4$ alleles Louis and Dempster (1987) suggested to determine the distribution of the test statistic under H_0 for an exact HWP test with a complete enumeration of all possible contingency tables with the given marginals. They noted that their method could be extended to arbitrary numbers of alleles. However, it has the shortcoming that it needs to be elaborated and also programmed separately for each possible allele number m. Aoki (2003) adopted the network algorithm of Mehta and Patel (1983) for tests in $r \times c$ contingency tables to reduce the computational effort of the complete enumeration algorithm. Pagano and Taylor Halvorsen (1981) introduced an incomplete enumeration technique for finding exact significance levels in $r \times c$ contingency tables, which omits enumeration of contingency tables with a larger conditional probability of occurrence than the observed one.

Our objectives were to (1) develop an incomplete enumeration algorithm for an exact HWP test with multiple alleles by adopting the concept of Pagano and Taylor Halvorsen (1981), and (2) illustrate its computational advantages in comparison to complete enumeration and the network algorithm of Aoki (2003) with an example of experimental data from seven maize populations.

Exact HWP test for *m* alleles

We follow the notation of Guo and Thompson (1992) and consider an autosomal locus with m alleles A_1, A_2, ..., A_m. The observed genotype counts in a sample of n individuals can be presented as the array

$$\begin{array}{c|c|c|c|c|}
 & A_1 & A_2 & \cdots & A_m \\
\hline
A_1 & f_{1,1} & f_{1,2} & \cdots & f_{1,m} \\
\hline
A_2 & & f_{2,2} & \cdots & f_{2,m} \\
\hline
\vdots & & & \cdots & \cdots \\
\hline
A_m & & & & f_{m,m} \\
\hline
\end{array}$$

where $f_{i,\,j}$ ($1 \le i \le j \le m$) is the count of genotype $A_i\,A_j$. The upper diagonal matrix of genotype counts is denoted by $\mathbf{F} = \{f_{i,\,j}\}_{1 \le i \le j \le m}$. The allele counts in the sample are denoted by the vector $\mathbf{f} = \{f_i\}_{1 \le i \le m}$ with $f_i = \sum_{j=1}^{i} f_{j,i} + \sum_{j=i}^{m} f_{i,j}$.

Under HWP and conditional on the allele counts $\mathbf{f}$, the probability of occurrence of the genotype count matrix $\mathbf{F}$ is (Levene 1949)

$$P(\mathbf{F}|\mathbf{f}) = \frac{n!\prod_{i=1}^{m} f_i!}{(2n)!} \times \frac{2^{\sum_{i<j} f_{i,j}}}{\prod_{i \le j} f_{i,j}!}.$$

For testing the null hypothesis H_0: “Genotypes in the population occur in HWP”, we need to determine the probability

$$p = \sum_{\mathbf{G} \in \mathcal{G}_{f,F}} P(\mathbf{G}|\mathbf{f}), \tag{2}$$

where the set $\mathcal{G}_{f,F}$ contains all genotype count matrices $\mathbf{G}$ which have the same allele counts $\mathbf{f}$ as the observed genotype count matrix $\mathbf{F}$, but have a smaller or equal conditional probability of occurrence

$$\mathcal{G}_{f,F} = \left\{\mathbf{G}|\mathbf{G} \in \mathcal{G}_f, P(\mathbf{G}|\mathbf{f}) \le P(\mathbf{F}|\mathbf{f})\right\}, \tag{3}$$

where $\mathcal{G}_f$ is the set of all genotype count matrices $\mathbf{G} = \{g_{i,\,j}\}_{1 \le i \le j \le m}$ having the same allele counts $\mathbf{f}$ as the observed genotype count matrix $\mathbf{F}$:

$$\mathcal{G}_f = \left\{\mathbf{G}|\forall i = 1, \ldots, m : f_i = \sum_{j=1}^{i} g_{j,i} + \sum_{j=i}^{m} g_{i,j}\right\}$$

If p is smaller than a given Type I error rate α, then H_0 is rejected.

Incomplete enumeration algorithm

Notation and outline of the algorithm

Let the function $\xi : \mathcal{G}_f \to \mathbb{N}$ uniquely assign to each genotype matrix **G** a positive integer, which is denoted in a number system with radix $r = \max\,(f_i + 1 \mid 1 \le i \le m)$:

$$\xi(\mathbf{G}) = (g_{1,1}g_{1,2}\cdots g_{1,m}g_{2,2}\cdots g_{2,m}\cdots g_{m,m})_r$$

The digits of the $\xi(\mathbf{G})$ are the elements $\{g_{i,j}\}_{i\le j\le m}$ of the upper diagonal of the genotype count matrix $\mathbf{G} \in \mathcal{G}_f$. We denote with $n_f = \| \Phi_f \|$ the number of elements in the image set

$$\Phi_f = \{\xi(\mathbf{G}_1), \ldots, \xi(\mathbf{G}_k), \xi(\mathbf{G}_{k+1}), \ldots, \xi(\mathbf{G}_{n_f})\} = \xi(\mathcal{G}_f)$$

where

$$\xi(\mathbf{G}_k) < \xi(\mathbf{G}_l) \quad \text{for } 1 \le k < l \le n_f.$$

For convenience in notation we define for each $i = 1, ..., m$

$$\mathbf{s}_i = \{g_{i,j}\}_{i \le j \le m}.$$

We further define an ordering on the set

$$I = \{(i,j) | i = 1 \ldots m, i \le j \le m\}$$

such that

$$(i,j) \succ (l,k) \quad \text{iff } i > l \text{ or } (i = l \text{ and } j > k).$$

For calculating p according to Eq. 2, the probability $P(\mathbf{G} \mid \mathbf{f})$ is summed over all elements $\mathbf{G} \in \mathcal{G}_{f,F}$. In consequence, we need a method for enumerating every genotype count matrix $\mathbf{G} \in \mathcal{G}_{f,F}$ without enumerating all genotype count matrices $\mathbf{G} \in \mathcal{G}_f$. We start by describing a general algorithm for complete enumeration of all genotype count matrices $\mathbf{G} \in \mathcal{G}_f$ consisting of two steps: (1) Construction of the smallest element $\xi(\mathbf{G}_1) \in \Phi_f$ and (2) construction of all elements of Φ_f recursively by finding $\xi(\mathbf{G}_{k+1})$ on the basis of $\xi(\mathbf{G}_k)$ until $\xi(\mathbf{G}_{n_f})$ is reached. Subsequently, we describe the rules to determine the $\mathbf{G} \notin \mathcal{G}_{f,F}$ which are not enumerated by the incomplete enumeration algorithm.

Part 1: Construction of $\xi(\mathbf{G}_1)$

The smallest number $\xi(\mathbf{G}_1)$ can be constructed by first determining $\mathbf{s}_1$ and then determining in sequential manner $\mathbf{s}_i$ proceeding from $i = 2$ to $i = m$. The elements of a sequence $\mathbf{s}_i$ ($i = 1 \ldots m$) are determined by starting with $g_{i,m}$ and then determining in sequential manner $g_{i,j}$ by proceeding from $j = m–1$ to $j = i$. Consider a sequence $\mathbf{s}_i$ ($i = 1 \ldots m$) and assume that all sequences $\mathbf{s}_k$ ($k \in \mathcal{K}$, with $\mathcal{K} = \{k \in \mathbb{N} | 1 \le k < i\}$) were already determined in previous steps and, hence, fixed. (Note that for $i = 1$ we have $\mathcal{K} = \emptyset$, because $\mathbf{s}_1$ is the first sequence that is determined.) We define

$$C_{i,j} = \begin{cases} \sum\limits_{1 \le k < i} g_{k,j} & \text{for } i > 1 \\ 0 & \text{for } i = 1 \end{cases} \tag{4}$$

Likewise, for fixed elements $g_{i,l}$ ($l = j + 1, ..., m$) of $\mathbf{s}_i$ we define

$$R_{i,j} = \begin{cases} \sum\limits_{j < l \le m} g_{i,l} & \text{for } j < m \\ 0 & \text{for } j = m. \end{cases} \tag{5}$$

Using these definitions we have for a fixed i and $1 \le i \le j \le m$

$$f_i = C_{i,i} + g_{i,i} + \sum_{i \le l \le j} g_{i,l} + R_{i,j} \tag{6}$$

and for a fixed j and $1 \le i < j \le m$

$$f_j = C_{i,j} + g_{i,j} + \sum_{i < k \le j} g_{k,j} + \sum_{j \le l \le m} g_{j,l}. \tag{7}$$

In consequence, by defining

$$g_{i,j} = \begin{cases} \min\left(f_i - C_{i,i} - R_{i,j}, f_j - C_{i,j}\right) & \text{for } i < j \\ (f_i - C_{i,i} - R_{i,i})/2 & \text{for } i = j \end{cases} \tag{8}$$

we choose the maximum element $g_{i,j}$, which fulfills the conditions (6) and (7). The resulting number is the smallest element of Φ_f, because the sequences $\mathbf{s}_i$ are constructed from right to left.

Part 2: Construction of $\xi(\mathbf{G}_{k+1})$ given $\xi(\mathbf{G}_k)$

Let $g_{i,j}$ and $g^*_{i,j}$ denote the digits of $\xi(\mathbf{G}_k)$ and $\xi(\mathbf{G}_{k+1}) \in \Phi_f$, respectively. First, the indices $(i', j') \in I$ are determined according to the following rules. We define for each (i, j)

$$q_i^{(i,j)} = R_{i,j} \quad \text{for } i \le j$$

and

$$q_j^{(i,j)} = f_j - C_{i,j} - g_{i,j} \quad \text{for } i < j.$$

We then start with

$$(i,j) = (m-1, m-1) \tag{9}$$

and check whether

$q_i^{(i,j)} > 0$ and $q_j^{(i,j)} > 0 \quad$ for $i < j$

or

$q_i^{(i,j)} > 1 \quad$ for $i = j$

holds true. If so, then (i', j') = (i, j), if not the next smallest (i, j) is checked. This procedure is continued until (i', j') is found or until $i = j = 1$ and $q_i^{(i,j)} = 0$. In the latter case $\xi(\mathbf{G}_k) = \xi(\mathbf{G}_{n_f})$ is the largest number and the enumeration stops. Note, that we start with (i, j) = (m–1, m–1) because $R_{i,m} = 0$ (Eq. 5) and in consequence $q_i^{(i,m)} = 0$.

If (i', j') is found, the digits of $\xi(\mathbf{G}_{k+1})$ are determined with a four step procedure: First

$g^*_{i,j} = g_{i,j}$ for all $(i,j) \prec (i',j')$;

second

$g^*_{i',j'} = g_{i',j'} + 1$;

third, the digits $g^*_{i',j}(j' < j \le m)$ are determined in descending order of (i', j), starting with $j = m$ and proceeding until $j = j' + 1$ reached, as

$$g^*_{i',j} = \min\left(\left(f_{i'} - C_{i',i'} - R^*_{i',j} - g^*_{i',i'} - \sum_{i' \le k \le j'} g^*_{i',k}\right), \left(f_j - C_{i',j}\right)\right)$$

where $R^*_{i,j}$ is defined according to Eq. 5 but using $g^*_{i,j}$ instead of $g_{i,j}$. By analogy we define $C^*_{i,j}$ according to Eq. 4 but using $g^*_{i,j}$ instead of $g_{i,j}$. Fourth, the elements of $\mathbf{s}^*_{i'+1}, \ldots, \mathbf{s}^*_m$ are determined in recursive manner analogously to the digits of $\xi(\mathbf{G}_1)$ but using $C^*_{i,j}$ and $R^*_{i,j}$ instead of $C_{i,j}$ and $R_{i,j}$ in Eq. 8.

Applying ξ^{-1} to all elements of Φ_f results in an enumeration of all elements $\mathbf{G} \in \mathcal{G}_f$, which is required for determining $\mathcal{G}_{f,F}$ in order to calculate p (Eq. 2) from the probabilities $P(\mathbf{G} \mid \mathbf{f})$.

Part 3: Incomplete enumeration

We now extend the complete enumeration algorithm such that only the elements $\mathbf{G} \in \mathcal{G}_{f,F}$ are enumerated. The procedure is analogous to a technique proposed by Pagano and Taylor Halvorsen (1981) for computing the exact significance levels of $r \times c$ contingency tables.

For loci with two alleles the application of the previously described complete enumeration algorithm results in an enumeration of all genotype count matrices $\mathbf{G}_i \in \mathcal{G}_f$ with $i = 1, \ldots, n_f$ having the same allele counts $\mathbf{f}$ as the observed genotype count matrix $\mathbf{F}$. The respective probabilities of occurrence $P(\mathbf{G}_i|\mathbf{f})$ with $i = 1, \ldots, n_f$ form a unimodal sequence. Thus, there exists a t with $1 \le t \le n_f$ such that $P(\mathbf{G}_1|\mathbf{f}) \le \cdots \le P(\mathbf{G}_t|\mathbf{f})$ and $P(\mathbf{G}_t|\mathbf{f}) \ge \cdots \ge P(\mathbf{G}_{n_f}|\mathbf{f})$ holds true. The idea behind the incomplete enumeration algorithm can be summarized as follows: First, enumerate all genotype count matrices in ascending order starting from $\mathbf{G}_1, \ldots, \mathbf{G}_l$ until a genotype count matrix $\mathbf{G}_l \notin \mathcal{G}_{f,F}$ with $P(\mathbf{G}_l|\mathbf{f}) > P(\mathbf{F}|\mathbf{f})$ is found. Second, enumerate all genotype count matrices in descending order starting from $\mathbf{G}_{n_f}, \ldots, \mathbf{G}_r$ until a genotype count matrix $\mathbf{G}_r \notin \mathcal{G}_{f,F}$ with $P(\mathbf{G}_r|\mathbf{f}) > P(\mathbf{F}|\mathbf{f})$ is found. Thus, only genotype count matrices $\mathbf{G}_1, \ldots, \mathbf{G}_l, \quad \mathbf{G}_r, \ldots, \mathbf{G}_{n_f}$ with $1 \le l \le r \le n_f$ are enumerated and the enumeration of genotype count matrices between $\mathbf{G}_{l+1}, \ldots, \mathbf{G}_{r-1}$ is omitted.

For loci with more than two alleles, the sequence of conditional probabilities $P(\mathbf{G}_i|\mathbf{f})$ with $i = 1, \ldots, n_f$ is multimodal and can be dissected into unimodal sequences. Two genotype count matrices $\mathbf{G}$ and $\mathbf{G}^*$ are elements of the same unimodal sequence if their digits

$g_{i,j} = g^*_{i,j}$ for all $(i,j) \prec (m-1, m-1)$.

Incomplete enumeration is done by finding for each unimodal sequence of conditional probabilities the first genotype count matrix $\mathbf{G} \notin \mathcal{G}_{f,F}$ with digits $g_{i,j}$. Subsequently, the digits $g^*_{i,j}$ of the next enumerated genotype count matrix $\mathbf{G}^* \in \mathcal{G}_f$ are determined with a four step procedure: First, set (i', j') = (m–1, m–1),

$g^*_{i,j} = g_{i,j}$ for all $(i,j) \prec (i',j')$;

second

$$g^*_{i',j'} = g_{i',j'} + \left\lfloor \frac{g_{i',j'+1}}{2} \right\rfloor; \tag{10}$$

third, the digit $g^*_{i',j'+1}$ is determined as

$$g^*_{i',j'+1} = g_{i',j'+1} \mod 2; \tag{11}$$

fourth, the digit $g^*_{i'+1,j'+1}$ is determined as

$$g^*_{i'+1,j'+1} = g_{i'+1,j'+1} + \left\lfloor \frac{g_{i',j'+1}}{2} \right\rfloor. \tag{12}$$

If the genotype count matrix $\mathbf{G}^* \in \mathcal{G}_{f,F}$, then the next smaller genotype count matrix is determined analogously, but using $g^*_{i',j'} = g_{i',j'} - 1$ in Eq. 10, $g^*_{i',j'+1} = g_{i',j'+1} + 2$ in Eq. 11, and $g^*_{i'+1,j'+1} = g_{i'+1,j'+1} - 1$ in Eq. 12. Otherwise the algorithm is continued by using (i, j) = (m–2, m–1) in Eq. 9.

Part 4: Hybrid algorithm

The incomplete enumeration algorithm for testing Hardy-Weinberg proportions as described in the previous section

Table 1 Mean number of enumerated genotype count matrices required to calculate the exact test of HWP in the sample dataset for different number of alleles and four algorithms

No. of alleles m	No. of markers with m alleles	Mean no. of enumerated genotype count matrices			
		CE	IE	NE	IN
1	14	–	–	–	–
2	96	12	11	12	11
3	170	297	224	201	147
4	122	24 208	18 557	11 368	7 918
5	82	3 194 751	2 346 638	766 303	596 553
6	51	68 808 888	47 689 855	31 202 750	21 441 493
7	27	5 787 883 573	3 605 185 045	1 033 897 968	760 752 194
> 7	33	–	–	–	–

CE complete enumeration, *IE* incomplete enumeration, *NE* network algorithm, *IN* incomplete network algorithm

and the network algorithm proposed by Aoki (2003) can be combined into a single hybrid algorithm. In a first step, a network representation of all elements of the set $\mathcal{G}_f$ is constructed, in which each element of $\mathcal{G}_f$ corresponds to a distinct path from the initial node to the terminal node through a network consisting of nodes and arcs. In a second step, all paths through the network are enumerated in the order of the described complete enumeration algorithm. However, using the Theorems 1 and 2 of Aoki (2003), some of the paths can be trimmed and their complete enumeration is omitted. If no trimming occurred for a single path through the network, then the incomplete enumeration technique as described in the previous section is applied.

Discussion

Computing time is often a limiting factor in carrying out exact tests. We illustrate the required number of enumerated genotype count matrices to calculate the exact test of HWP which directly affects the required computing time for the exact test of HWP with an example from a genetic diversity study in maize (Reif et al. 2003). Its objective was to investigate the relationship of genetic distance based on molecular markers with heterosis in population crosses in order to establish heterotic pools for hybrid breeding. To estimate genetic distances between seven tropical maize populations, 48 individuals were sampled from each population and investigated with 85 simple sequence repeat (SSR) markers. For a correct quantitative genetic interpretation of the heterosis observed in population crosses in a previous study (Vasal et al. 1992), it is important to know, whether the base populations are in HWP or whether they deviate from the HWP owing to inbreeding, selection or other reasons.

We implemented four different algorithms to calculate the exact p-values for testing HWP. CE: a complete enumeration algorithm, IE: the proposed incomplete enumeration algorithm, NE: the network algorithm according to Aoki (2003), and IN: a hybrid algorithm combining incomplete enumeration as described in the previous section with Theorems 1 and 2 of Aoki (2003).

For loci with less than six alleles, which comprised 87.6% of all 7 × 85 = 595 tests, the average computing time was about nine seconds per test, employing the complete enumeration on a personal computer (AMD Opteron 248 processor, program written in C). The maximum computing time required for a locus with six alleles was 18 min. For 60 loci with more than six alleles, carrying out the exact test with complete enumeration was not possible within 1 h.

Both, incomplete enumeration and network algorithm reduced the number of enumerated genotype count matrices considerably compared with complete enumeration (Table 1). For three or more alleles, the hybrid algorithm reduced the number of enumerated matrices up to a factor of 7.5 compared to complete enumeration and about 25–30% compared to the network algorithm.

The network algorithm performs best if the p-values are large (Aoki 2003), whereas our algorithm performs better with small p-values, because only genotype count matrices with a smaller conditional probability of occurrence than the observed one are enumerated. Therefore, both approaches are complementing each other, which explains the superiority of the hybrid algorithm. Summarizing, with the proposed hybrid algorithm the bounds of computational feasibility of the exact test of Hardy–Weinberg were significantly extended. We conclude that with today's computing resources, exact HWP tests are practicable either with large population sizes and biallelic markers or with medium population sizes and numbers of alleles. It must be noted, however, that for loci with more than two alleles the required computing time increases exponentially with the number of individuals and alleles investigated.

Pearson's χ^2 test is often used to test HWP. However, it is well known that the distribution of the test statistic is only poorly approximated by the χ^2 distribution if contingency tables are sparsely occupied (Agresti 1996; Hernández and Weir 1989; Wigginton et al. 2005). In our dataset, the distribution of allele frequencies was skewed, on average

4.02 alleles were observed per locus and the two most frequent alleles had a total frequency of 0.85. This resulted in a sparsely occupied contingency table of observed genotype counts. As a consequence, the χ^2 test is expected to result in a high rate of incorrectly rejecting or accepting the null hypothesis of HWP.

Monte Carlo approximations of the probability p can be obtained by evaluating only a random sample of all possible genotype count matrices $G \in \mathcal{G}_f$ (Guo and Thompson 1992; Huber et al. 2006). The estimated p values of a Monte Carlo test with 17,000 repetitions (as suggested by Guo and Thompson 1992) were in good accordance with those of the exact test (results not shown). Thus, Monte Carlo tests provide a valuable tool for approximating the probability p of the exact test. However, approximating p seems only appropriate when there are substantial problems in calculating the exact value of p, e.g., in the case of large population sizes and multiple alleles.

Extension of our method to one-sided tests of HWP is straightforward by defining $\mathcal{G}_{f,F}$ such that it contains only genotype count matrices, for which the count of homozygotes is greater (right-sided test) or smaller (left-sided test) than the number of homozygotes expected under HWP. The proposed algorithm can also be extended to exact tests for linkage disequilibrium, because these are also based on conditional probabilities of contingency tables of genotype counts.

Our results illustrate that the proposed incomplete enumeration algorithm for testing deviations from HWP significantly extends the range of computational feasible problems. Owing to the superior performance of the exact test, the presented approach can help experimenters to analyze datasets and extract the maximum possible information, even when only sparsely occupied contingency tables are available and, therefore, the large sample assumptions underlying the χ^2 goodness-of-fit tests do not apply, as frequently occurs in molecular marker studies with plants.

The routines developed for performing exact tests of HWP with the described enumeration algorithm are available in software Plabsoft (Maurer et al. 2004).

Acknowledgments The financial support from a fellowship by KWS SAAT AG, Einbeck, Germany, to Hans Peter Maurer is gratefully acknowledged. The authors thank Professor Dr. Uwe Jensen and Professor Dr. Hans-Peter Piepho for their valuable suggestions on the manuscript.

References

Agresti A (1996) An introduction to categorical data analysis. Wiley, New York

Aoki S (2003) Network algorithm for the exact test of Hardy–Weinberg proportion for multiple alleles. Biometric J 4:471–490

Crow JF (1988) Eighty years ago: the beginnings of population genetics. Genetics 119:473–476

Deng HW, Chen WM, Recker RR (2001) Population admixture: Detection by Hardy–Weinberg test and its quantitative effects on linkage–disequilibrium methods for localizing genes underlying complex traits. Genetics 157:885–897

Elston RC, Forthofer R (1977) Testing for Hardy–Weinberg equilibrium in small samples. Biometrics 33:536–542

Emigh TH (1980) A comparison of tests for Hardy–Weinberg equilibrium. Biometrics 36:627–642

Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48:361–372

Hardy GH (1908) Mendelian proportions in a mixed population. Science 28:49–50

Hernández JL, Weir BS (1989) A disequilibrium coefficient approach to Hardy–Weinberg testing. Biometrics 45:53–70

Huber M, Chen Y, Dinwoodie I, Dobra A, Nicholas M (2006) Monte Carlo algorithms for Hardy–Weinberg proportions. Biometrics 62:49–53

Levene H (1949) On a matching problem arising in genetics. Ann Math Stat 20:91–94

Louis EJ, Dempster ER (1987) An exact test for Hardy–Weinberg and multiple alleles. Biometrics 43:805–811

Maurer HP, Melchinger AE, Frisch M (2004) Plabsoft: Software for simulation and data analysis in plant breeding. In: Vollmann J, Grausgruber H, Ruckenbauer P (eds) Genetic variation for plant breeding. (Proceedings of the 17th EUCARPIA General Congress, 8–11 September 2004, Tulln, Austria) BOKU, Vienna, pp 359–362

Mehta CR, Patel NR (1983) A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. J Am Stat Assoc 78:427–434

Nielsen DM, Ehm G, Weir BS (1999) Detecting marker-disease association by testing for Hardy–Weinberg Disequilibrium at a marker locus. Am J Hum Genet 63:1531–1540

Pagano M, Taylor Halvorsen K (1981) An algorithm for finding the exact significance levels for $r \times c$ contingency tables. J Am Stat Assoc 76:931–934

Reif JC, Melchinger AE, Xia XC, Warburton ML, Hoisington DA, Vasal SK, Srinivasan G, Bohn M, Frisch M (2003) Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci 43:1275–1282

Reif JC, Xia XC, Melchinger AE, Warburton ML, Hoisington DA, Beck D, Bohn M, Frisch M (2004) Genetic diversity determined within and among CIMMYT maize populations of tropical, subtropical, and temperate germplasm by SSR markers. Crop Sci 44:326–334

Semerikov V, Belyaev A, Lascoux M (2002) The origin of Russian cultivars of red clover (*Trifolium pratense* L.) and their genetic relationships to wild populations in the Urals. Theor Appl Genet 106:127–132

Vasal SK, Srinivasan DL, Beck DL, Crossa J, Pandey S, de Leon C (1992) Heterosis and combining ability of CIMMYT's tropical late white maize germplasm. Maydica 37:217–223

Weinberg W (1908) Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Württembergischen Vereins für vaterländische Naturkunde 64:369–382

Weir B (1996) Genetic data analysis II, 2nd edn. Sinauer Associates, Sunderland

Wellek S (2004) Tests for establishing compatibility of an observed genotype distribution with Hardy–Weinberg equilibrium in the case of a biallelic locus. Biometrics 60:694–703

Wigginton JE, Cutler DJ, Abecasis GR (2005) A note on exact tests of Hardy–Weinberg equilibrium. Am J Hum Genet 76:887–893