ORIGINAL ARTICLE

Testing for Hardy–Weinberg equilibrium at biallelic genetic markers on the X chromosome

J Graffelman[1] and BS Weir[2]

Testing genetic markers for Hardy–Weinberg equilibrium (HWE) is an important tool for detecting genotyping errors in large-scale genotyping studies. For markers at the X chromosome, typically the χ^2 or exact test is applied to the females only, and the hemizygous males are considered to be uninformative. In this paper we show that the males are relevant, because a difference in allele frequency between males and females may indicate HWE not to hold. The testing of markers on the X chromosome has received little attention, and in this paper we lay down the foundation for testing biallelic X-chromosomal markers for HWE. We develop four frequentist statistical test procedures for X-linked markers that take both males and females into account: the χ^2 test, likelihood ratio test, exact test and permutation test. Exact tests that include males are shown to have a better Type I error rate. Empirical data from the GENEVA project on venous thromboembolism is used to illustrate the proposed tests. Results obtained with the new tests differ substantially from tests that are based on female genotype counts only. The new tests detect differences in allele frequencies and seem able to uncover additional genotyping error that would have gone unnoticed in HWE tests based on females only.
Heredity (2016) **116**, 558–568; doi:10.1038/hdy.2016.20; published online 13 April 2016

INTRODUCTION

The Hardy–Weinberg law is a fundamental principle in population genetics, and is of relevance in many related areas of the life sciences, such as epidemiology, bioinformatics and biostatistics. In its most elementary form, the law states that the genotype frequencies AA, AB and BB for an autosomal biallelic marker will occur in the relative proportions p^2, $2pq$ and q^2, where p is the allele frequency of A, and $q = 1 - p$ the allele frequency of B. In the absence of disturbing forces (migration, selection and so on) these proportions will be reached in one generation of random mating and afterwards genotype and allele frequencies will remain unaltered over the generations. In this state, the population is said to be in Hardy–Weinberg equilibrium (HWE).

It has become common practice in gene–disease association studies to test genetic markers for HWE. Significant deviations from HWE are often the consequence of genotyping error, and HWE tests are an efficient way of detecting (gross) genotyping error. There are many other possible causes for disequilibrium. Disequilibrium can also arise from population substructure or inbreeding (Laird and Lange, 2011). Several statistical tools can be used to test markers for equilibrium (Weir, 1996). Formal hypothesis testing is possible by means of the χ^2 test, an exact test, a likelihood ratio test, a permutation test or by using Bayesian procedures (Ayres and Balding, 1998; Lindley, 1988; Shoemaker *et al.*, 1998; Consonni *et al.*, 2010; Wakefield, 2010). Graphics like ternary diagrams (Graffelman and Morales-Camarena, 2008; Graffelman, 2015) and Q–Q plots of *P*-values are valuable tools if large numbers of markers are tested for HWE simultaneously.

It is well known that traits and markers on the X chromosome are different from autosomal markers with respect to HWE. As males are hemizygous and receive their X chromosome from their mother, male allele frequencies of X-chromosomal markers equal the female allele frequencies of the previous generation. If male and female allele frequencies initially differ, then it will take several generations before HWE is reached. The implications of X-chromosomal inheritance for HWE are well described in most text books on population genetics (Crow and Kimura, 1970; Hartl, 1980; Lange, 2002). Figure 1a summarizes the situation: if males are fixed for the A allele, and females are fixed for the B allele, then it takes 8 generations before the difference in allele frequency is below 0.01. Figure 1b shows how female genotype frequencies change over the generations, converging to the HWE proportions once the allele frequencies have stabilized.

Biallelic markers, which are the focus of this paper, are usually tested for HWE by means of well-known tests like the χ^2 test or the exact test. The relative merits of these test procedures have been the subject of several studies (Elston and Forthofer, 1977; Emigh, 1980; Wigginton *et al.*, 2005; Graffelman and Moreno, 2013). However, bibliographical research shows that there are many papers dedicated to statistical tests for HWE for autosomal markers, but that little or no attention has been devoted to sex-linked markers (Zheng *et al.*, 2007). The usual approach for testing X-chromosomal markers for HWE is to test it in females only, because the males are hemizygous. There is plenty of software available for analyzing genetic marker data for HWE. Available software often does treat X-linked markers. Plink (Purcell *et al.*, 2007) tests markers on the X chromosome by doing the standard tests (χ^2 or exact) using only the females in the database. The R package GWASTools (Gogarten *et al.*, 2012) explicitly states to exclude males for X-chromosomal markers. Other programs do not

[1]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain and [2]Department of Biostatistics, University of Washington, Seattle, WA, USA
Correspondence: Dr J Graffelman, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Avinguda Diagonal 647, 08028 Barcelona, Spain.
E-mail: jan.graffelman@upc.edu

a

Expected male and female A allele frequencies

b

Expected female genotype frequencies
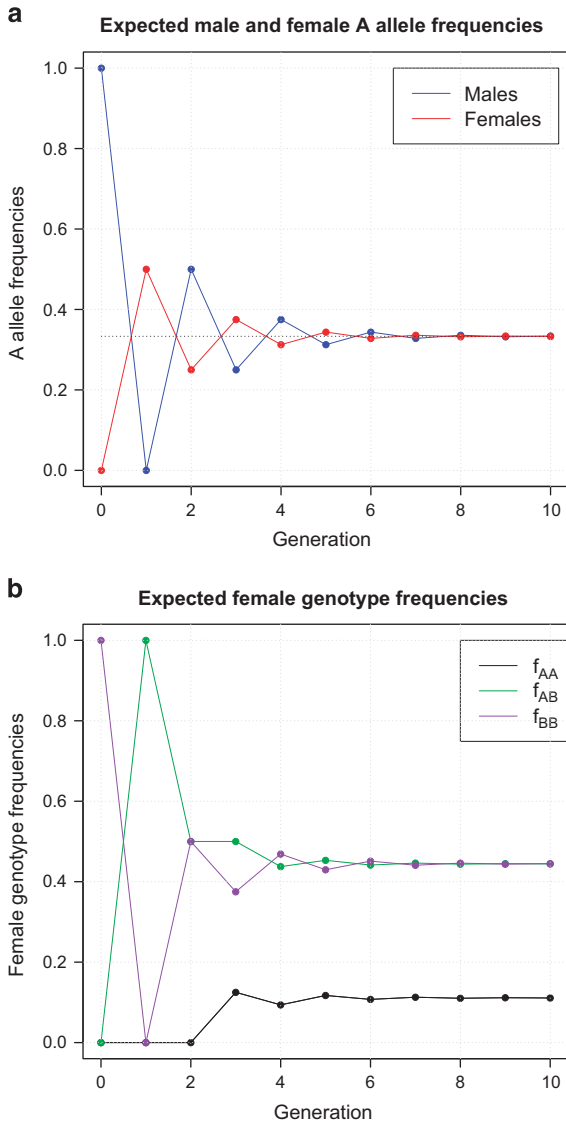
Figure 1 HWE for a biallelic marker on the X chromosome. (**a**) Evolution of male and female allele frequencies over time after an initial difference of 1. (**b**) Simultaneous evolution of the female genotype frequencies over time.

distinguish between autosomal and X-linked markers, leaving it to the user to decide what to do with the hemizygous males. Because of the use of standard file formats, males are often coded in the database as if they were diploid, that is, the hemizygous genotypes A and B appear as AA and BB in the data set. This has the danger that males actually enter the HWE test as if they would be diploid individuals if the sexes are not distinguished. In this case the user has to explicitly discard the males before testing for HWE. It is apparently assumed that the hemizygous males are uninformative with respect to HWE. In this paper, we argue that testing HWE for X-chromosomal markers by applying the standard χ^2 test or the exact test to females only is inadequate. Males should be taken into account when testing X-chromosomal marker data for HWE, and the purpose of this paper is to lay down the statistical foundation for testing biallelic X-chromosomal markers for HWE. We develop four frequentist procedures in full detail: the χ^2 test, the likelihood ratio test, the exact test and the permutation test. There has been an upsurge of Bayesian tests for HWE. This paper gives the first full frequentist treatment of HWE

tests for X-chromosomal markers, and Bayesian procedures as well as multiallelic markers are left for future work and considered beyond the scope of the current paper.

It should be fairly obvious that ignoring males is inadequate for the following reasons. First of all, the sample size is reduced by ~50% by ignoring the males, and this evidently brings about a loss of power. Moreover, if the male and female allele frequencies differ then the marker cannot be in HWE. Testing females only ignores this possibility. Depending on the allele frequencies and genotype frequencies of males and females, four possible situations can arise, as depicted in the ternary diagrams in Figure 2. The ternary plot, also known in genetics as a de Finetti diagram (De Finetti, 1926), is a useful tool, as it simultaneously displays genotype frequencies, allele frequencies and the deviation from HWE in a single graph (Graffelman and Morales-Camarena, 2008). The Hardy–Weinberg (HW) law is represented by a parabola in the ternary diagram (Cannings and Edwards, 1968). The base of the triangle is a 0–1 axis for the allele frequency, and the male allele frequency can be conveniently represented by a single point on this axis.

Figure 2a shows a ternary diagram representing male and female genotype and allele frequencies for an X-linked marker in HWE. Males and females have equal allele frequencies, and female genotype frequencies occur in HW proportions. For this reason, the female composition is on the HWE parabola. Figure 2b shows a disequilibrium situation. The two sexes have the same allele frequencies, but disequilibrium arises from the fact that the female genotype frequencies do not correspond to HW proportions. Figure 2c shows a different condition that gives rise to disequilibrium: the females do occur in HW proportions, but male and female allele frequencies differ. Figure 2d shows the combination of both phenomena: disequilibrium arises from different allele frequencies and female genotype frequencies not corresponding to HW proportions. Thus, two conditions have to be met for an X-chromosomal marker to be in equilibrium: equal allele frequencies *and* HW proportions for females.

The structure of the remainder of this paper is as follows. In the next Section we develop statistical tests for markers at the X chromosome. The newly developed tests are illustrated with empirical data sets in the Example Section. Discussion and bibliography complete the paper.

STATISTICAL TESTS FOR HWE FOR MARKERS AT THE X CHROMOSOME

Several statistical tests are available for investigating genetic marker data for HWE. The classical χ^2 test for goodness of fit has been the most popular test for HWE for decades, although nowadays exact procedures are more and more often employed. A likelihood ratio test is also available. A description of the different tests is given by Weir (1996). In the following paragraphs, we develop the χ^2 test, the likelihood ratio test, the exact test and the permutation test for X-linked markers. We then compare the Type I error rate of the proposed tests.

The χ^2 test
The χ^2 test is the classical test for HWE. Let n_{AA}, n_{AB} and n_{BB} represent the observed genotype counts. Let $\hat{p}_A$ be the maximum likelihood estimator for the population allele frequency given by:

$$\hat{p}_A = \frac{(2n_{AA} + n_{AB})}{2n}, \tag{1}$$

and let $e_{AA} = n\hat{p}_A^2$, $e_{AB} = 2n\hat{p}_A(1 - \hat{p}_A)$ and $e_{BB} = n(1 - \hat{p}_A)^2$ be the

expected genotype counts under HWE. The χ^2 statistic can be computed as

$$X^2 = \frac{(n_{AA} - e_{AA})^2}{e_{AA}} + \frac{(n_{AB} - e_{AB})^2}{e_{AB}} + \frac{(n_{BB} - e_{BB})^2}{e_{BB}}. \quad (2)$$

Under the null hypothesis, this statistic has a χ_1^2 reference distribution. This χ^2 test is typically explained in textbooks on population genetics (Hartl, 1980; Hedrick, 2005). We now develop the χ^2 test for markers on the X chromosome that takes the males into account. The following notation will be used. We first specify the theoretical population parameters, and let p_A be the frequency of the A allele, and $\hat{\phi}$ the fraction of males in the population. The sample quantities are as follows: let m_A and m_B be the number of males carrying the A and B allele, respectively, and let f_{AA}, f_{AB} and f_{BB} be the number of females of each of the three possible genotypes. Let n_m be the number of males, n_f the number of females and $n = n_m + n_f$ the total sample size. The total number of alleles is given by $n_t = 2n_f + n_m$. The distribution of these genotype counts is multinomial. The probabilities of the five categories, the observed and expected counts under HWE are shown in Table 1.

The maximum likelihood estimators for $\hat{\phi}$ and p_A are given by

$$\hat{\phi} = \frac{n_m}{n}, \quad \hat{p}_A = \frac{n_A}{2n_f + n_m}. \quad (3)$$

Let e_i be the expected count of genotype i. Then, Pearson's χ^2 statistic for goodness of fit is given by

$$X^2 = \frac{(m_A - e_A)^2}{e_A} + \frac{(m_B - e_B)^2}{e_B} + \frac{(f_{AA} - e_{AA})^2}{e_{AA}}$$
$$+ \frac{(f_{AB} - e_{AB})^2}{e_{AB}} + \frac{(f_{BB} - e_{BB})^2}{e_{BB}}. \quad (4)$$

Under HWE, this statistic has a χ_2^2 distribution. Note that the proportion of males (ϕ) is assumed unknown, and estimated from the data. If the sample is known to come from a population of exactly equally frequent sexes, then ϕ may assumed to be known and set to $\frac{1}{2}$. In that case, the reference distribution of X^2 is the χ^2 distribution with 3 degrees of freedom. Note that this χ^2 test assumes both homogeneity of allele frequencies and HW proportions in females, and thus constitutes an omnibus test for HWE. Rejection can occur if female
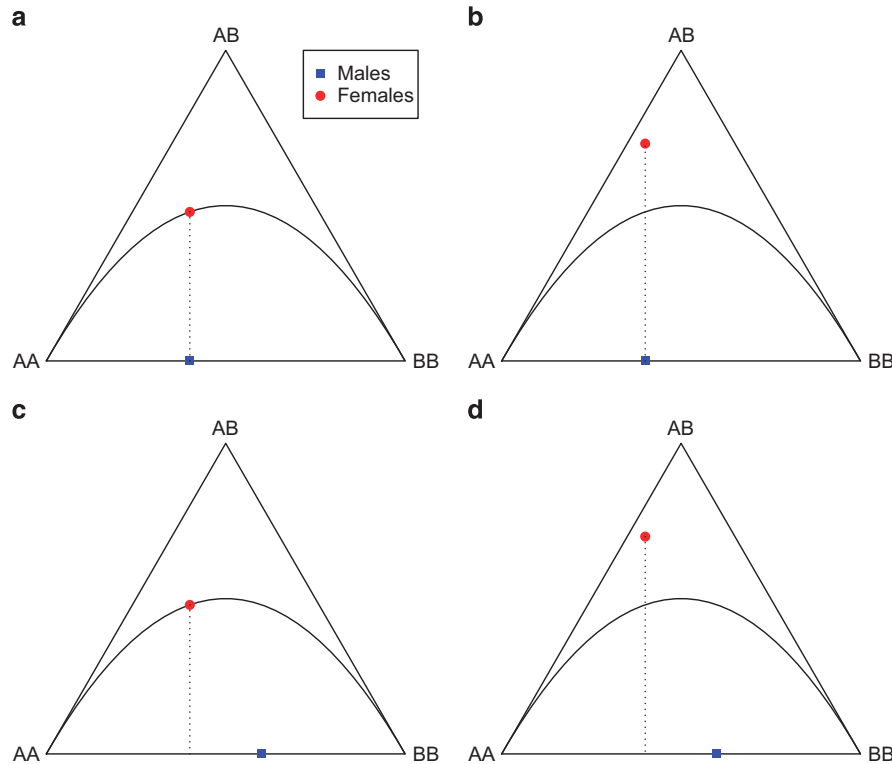
Figure 2 Hardy–Weinberg (dis)equilibrium for a biallelic marker on the X chromosome. (**a**) Equilibrium. (**b**) Disequilibrium due to deviating female genotype frequencies. (**c**) Disequilibrium due to nonhomogeneous allele frequencies. (**d**) Disequilibrium due to deviating female genotype frequencies and nonhomogeneous allele frequencies.

Table 1 Observed and expected genotype counts for an X-chromosomal marker under Hardy–Weinberg equilibrium

Genotype	Males		Females		
	A	B	AA	AB	BB
Probability	ϕp_A	$\phi(1 - p_A)$	$(1 - \phi)p_A^2$	$2(1 - \phi)p_A(1 - p_A)$	$(1 - \phi)(1 - p_A)^2$
Observed	m_A	m_B	f_{AA}	f_{AB}	f_{BB}
Expected	$n\hat{\phi}\hat{p}_A$	$n\hat{\phi}(1 - \hat{p}_A)$	$n(1 - \hat{\phi})\hat{p}_A^2$	$2n(1 - \hat{\phi})\hat{p}_A(1 - \hat{p}_A)$	$n(1 - \hat{\phi})(1 - \hat{p}_A)^2$

genotype proportions deviate from HW proportions, if female and male allele frequencies differ, or if both these phenomena occur simultaneously.

Likelihood ratio test

A likelihood ratio test for autosomal biallelic markers has been described by Weir (1996). It is based on a multinomial likelihood, the latter evaluated under the null hypothesis of HWE and under the alternative. It is straightforward to extend the likelihood ratio test for markers on the X chromosome. For autosomal makers, the logarithm of the likelihood ratio is given by

$$\ln\left(\frac{L_0}{L_1}\right) = -2n\ln(2) - n\ln(n) + n_{AB}\ln(2) + n_A\ln(n_A)$$
$$+ n_B\ln(n_B) - n_{AA}\ln(n_{AA}) - n_{AB}\ln(n_{AB})$$
$$- n_{BB}\ln(n_{BB}), \qquad (5)$$

where n_{AA}, n_{AB} and n_{BB} contain the sums of male and female genotype frequencies. The statistic $G^2 = -2\ln(L_0/L_1)$ has, asymptotically, a χ_1^2 distribution. For a marker on the X chromosome, using a multinomial distribution with 5 categories, the likelihood ratio statistic becomes

$$\ln\left(\frac{L_0}{L_1}\right) = -n_t\ln(n_t) + n_m\ln(n_m) + n_f\ln(n_f) + f_{AB}\ln(2)$$
$$+ n_A\ln(n_A) + n_B\ln(n_B) - f_{AA}\ln(f_{AA}) - f_{AB}\ln(f_{AB})$$
$$- f_{BB}\ln(f_{BB}) - m_A\ln(m_A) - m_B\ln(m_B).$$

Asymptotically, $G^2 = -2\ln(L_0/L_1)$ has a χ_2^2 distribution. For large samples the likelihood ratio test is equivalent to a χ^2 test for HWE. A different likelihood ratio test assuming a fixed number of males and females has recently been proposed by You *et al.* (2015).

Exact test

Exact test procedures for HWE are based on the conditional distribution of the number of heterozygotes (N_{AB}) given the minor allele count (N_A). This distribution was derived by Levene (1949) and Haldane (1954) and is given by:

$$P(N_{AB}|N_A) = \frac{n_A!n_B!n!2^{n_{AB}}}{n_{AA}!n_{AB}!n_{BB}!(2n)!}. \qquad (6)$$

The standard way to compute the *P*-value of an exact test is to sum probabilities according to Equation (6) for all samples that are as likely or less likely than the observed sample. We propose the following exact test for X-chromosomal biallelic markers. We derive, under the assumption of HWE, the joint distribution of the number of A males (M_A) and the number of female heterozygotes (F_{AB}), conditional on sample size, number of A alleles (n_A) and the observed male and female frequencies. The corresponding probability density is given by:

$$P(M_A = m_a \cap F_{AB} = f_{AB}|n, n_A, n_m)$$
$$= \frac{n_A!n_B!n_m!n_f!}{m_A!m_B!f_{AA}!f_{AB}!f_{BB}!n_t!}2^{f_{AB}}. \qquad (7)$$

This density is seen to be a straightforward generalization of the autosomal density in Equation (6), with the sample size factorial $n!$ split in separate factorials for males and females, additional factorials for the male genotype counts in the denominator and $2n$ replaced by n_t. In fact, the autosomal exact test (Equation 6) is a special case of the X-chromosomal exact test, because for a sample without males ($m_A = m_B = 0$) Equation (7) reduces to Equation (6). The *P*-value of the exact test for HWE of an X-linked marker is then obtained by summing Equation (7) over all possible outcomes that are as likely or

less likely than the observed sample. We consider a numerical example to illustrate the computations. A sample of 20 individuals (10 males and 10 females) with a total of 6 A alleles has been observed, with genotype counts $m_A = 3$, $m_B = 7$, $f_{AA} = 0$, $f_{AB} = 3$ and $f_{BB} = 7$. Table 2 lists all 16 possible samples together with their probability according to Equation (7).

The number of possible outcomes for the distribution in Equation (7) is given by

$$\sum_{m_A = \max(n_m - n_B, 0)}^{\min(n_A, n_m)} (\lfloor\min(n_A - m_A, n_B - (n_m - m_A))/2\rfloor + 1),$$

where n_A represents the overall minor allele count and $\lfloor\cdot\rfloor$ the flooring operation. The observed sample (12) has a probability of 0.1940. All samples except sample 10 have a smaller probability than that for sample 12. The probabilities in Table 2 are the probabilities under the assumption of HW proportions and equal allele frequencies. The *P*-value of the test is thus equal to $1 - 0.2546 = 0.7454$. For autosomal markers, the mid *P*-value has been proposed as a more appropriate *P*-value for an exact test for HWE (Graffelman and Moreno, 2013). The mid *P*-value could also be used in the present exact test for markers on the X chromosome, and corresponds to half the probability of the observed sample plus the probabilities of all samples that are more extreme. For the example at hand, the mid *P*-value is 0.6484.

The joint density for the example in Table 2 is shown in Figure 3. The standard exact *P*-value corresponds to the height of the bar representing the observed sample at ($m_A = 3$, $f_{AB} = 3$), plus the height of all bars lower than the observed sample.

Permutation test

HWE refers to the statistical independence of alleles within individuals at a marker. For autosomal markers this independence can be assessed by a permutation test, where all 2n alleles of all individuals are written out as a single sequence (for example, BABBABABAABBAA....). This sequence is then permuted many times, and for each permuted sequence alleles are grouped in pairs that are taken as individuals. For each permuted sequence a test statistic (the pseudo-statistic) for disequilibrium is computed. The test statistic for the original observed sample is compared against the distribution of the pseudo-statistic, where the latter has been generated under the null hypothesis.

Table 2 All possible samples for a set of 20 individuals (10 males and 10 females) with a total of 6 A alleles

	m_A	m_B	f_{AA}	f_{AB}	f_{BB}	Prob
1	0	10	3	0	7	0.0002
2	0	10	2	2	6	0.0085
3	0	10	1	4	5	0.0340
4	0	10	0	6	4	0.0226
5	1	9	2	1	7	0.0121
6	1	9	1	3	6	0.1132
7	1	9	0	5	5	0.1358
8	2	8	2	0	8	0.0034
9	2	8	1	2	7	0.1091
10	2	8	0	4	6	0.2546
11	3	7	1	1	8	0.0364
12	3	7	0	3	7	0.1940
13	4	6	1	0	9	0.0035
14	4	6	0	2	8	0.0637
15	5	5	0	1	9	0.0085
16	6	4	0	0	10	0.0004
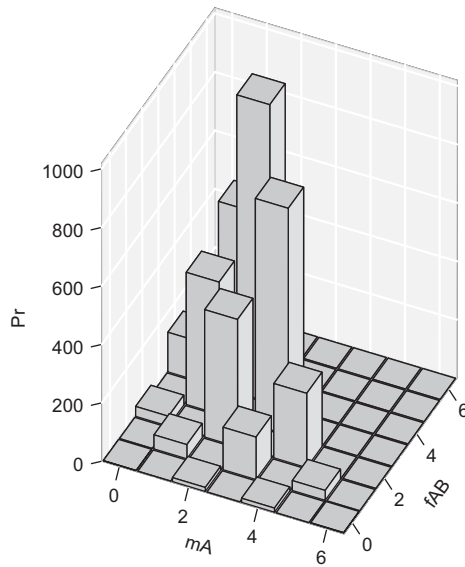
Figure 3 Joint distribution of m_A and f_{AB} for given sample size, minor allele count and number of males and females. Example for $n=20$, $n_m=10$, $n_f=10$ and $n_A=6$.

The P-value of the test is calculated as the fraction of permuted samples for which the pseudo-statistic is equal to or exceeds the test statistic. Such a test is computer intensive but has the advantage that it does not rely on asymptotic assumptions. The permutation test is easily adapted for X-chromosomal markers as follows. Again all alleles are written out as a single sequence of $n_t = 2n_f + n_m$ alleles, and this sequence is shuffled many times. The first n_m elements from this sequence are taken to be males, and the male allele counts are determined. The remaining $2n_f$ elements of the sequence are grouped in pairs as females and the female genotype counts are determined. The test statistic for disequilibrium (for example, the χ^2 statistic in Equation (4)) is calculated for each shuffled sequence, and the P-value of the test is calculated as before. This permutation test conditions on the observed allele frequency and also on the observed gender ratio.

Type I error rate

In this section we evaluate the proposed statistical tests in terms of their Type I error rate. We consider the effect of the sex ratio and the minor allele frequency (MAF) on the Type I error rate. Type I error rates were calculated by exhaustive enumeration (Graffelman and Moreno, 2013). This avoids error in the obtained rates because of simulation or asymptotic approximations.

Figure 4 shows the Type I error rate for the most common tests for HWE, the exact test and the χ^2 test. We use a sample of 100 individuals and report the Type I error as a function of the minor allele count and the sex ratio. We note that the maximum minor allele count is given by $\frac{1}{2}(2n_f + n_m)$ for each graph, but that the limits of all graphs have been kept fixed to facilitate comparison. Figure 4 shows that the standard exact test is conservative for low MAF and never exceeds the nominal rate ($\alpha = 0.05$). The standard exact test underrates the nominal rejection level for low MAF. The χ^2 test can largely exceed the nominal rate, especially for low MAF markers. These findings are essentially consistent with those reported for autosomal markers (Wigginton et al., 2005; Graffelman and Moreno, 2013). When there are no males in the sample, the Type I error rate is exactly the same as the one obtained for an autosomal marker (see Figure 2; Graffelman

and Moreno, 2013). When there are no females in the sample Hardy-Weinberg proportions (HWP) can never be rejected, and therefore the Type I error rate is 0. Extremely biased sex ratios affect the χ^2 test, giving it an increased Type I error rate. The χ^2 statistic is not defined for samples without males or females, because expected counts for males or females are zero in those cases. The best Type I error rate profiles are observed when the number of males is approximately equal to the number of females. With regard to the exact test, the rejection level of the mid P-value is most close to the nominal rejection level, for all allele frequencies and sex ratios.

We also compare the Type I error rate of the exact test for the all-individual test and the females-only test, as a function of sex ratio and minor allele count, in Figure 5. This figure shows the Type I error rates to be identical (as expected) when all sampled individuals are females (upper left panel). For samples that do contain males, the Type I error rate profile of a test ignoring males is typically below the profile of an all-individual test, and farther below the nominal rejection level. The convergence to the nominal rate is faster for the all-individual exact test. We note that the maximal minor allele count on the horizontal axis is $\frac{1}{2}n_t$ for the all-individual test, but smaller ($\frac{1}{2}n_f$) for the females-only test because of the exclusion of males.

EXAMPLES: THE GENEVA VENOUS THROMBOSIS PROJECT

We present applications of the tests developed in this paper to data from a genome-wide association study on venous thrombosis that formed part of the GENEVA project (www.geneva.org). The original genome-wide data set contains 561 490 single-nucleotide polymorphisms (SNPs) typed for 2600 study subjects from the Mayo Clinic. Details of the study have been posted on the database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap) and are available under accession number phs000289.v2.p1. We use a subset of X-chromosomal SNPs of this project, to which we apply the tests proposed in this paper. Only control subjects were considered, and for pairs with a first- or second-degree family relationship, one individual was removed in order to create a subset of independent individuals. Statistical tests for HWE assume a random sample of independent individuals from an infinite universe. The removal of related individuals is an effort to try to satisfy that assumption as best as possible. Individuals potentially stemming from a different human population were identified as such by principal component analysis and removed. In order to obtain a set of approximately independent markers, SNPs were linkage disequilibrium pruned with Plink (Purcell et al., 2007). More details on the filtering of the database can be found in its associated technical report (GENEVA, 2010). After applying these filters, a database of 1256 control subjects genotyped for 4158 X-chromosomal SNPs was used for the statistical analysis described in this paper.

In order to emphasize the difference between the tests proposed in this paper and the habitual approach of testing HWE on females only, we report χ^2 and exact tests with and without males for four SNPs in Table 3 as an example. The reported χ^2 tests used 1 and 2 degrees of freedom for a females-only and an all-individual test, respectively.

The first marker in Table 3 is seen to be almost in perfect HWE when only females are tested. Males have, however, a different allele frequency, and the test including the males is highly significant. This shows that the inclusion of the males can drastically change the statistical inference on HWE. The second marker is monomorphic in females. The inclusion of the males increases the sample size and males with B alleles do exist, showing the marker is in fact polymorphic. Inclusion of the males brings the marker close to statistically significant. For the third marker, HWE is rejected when we
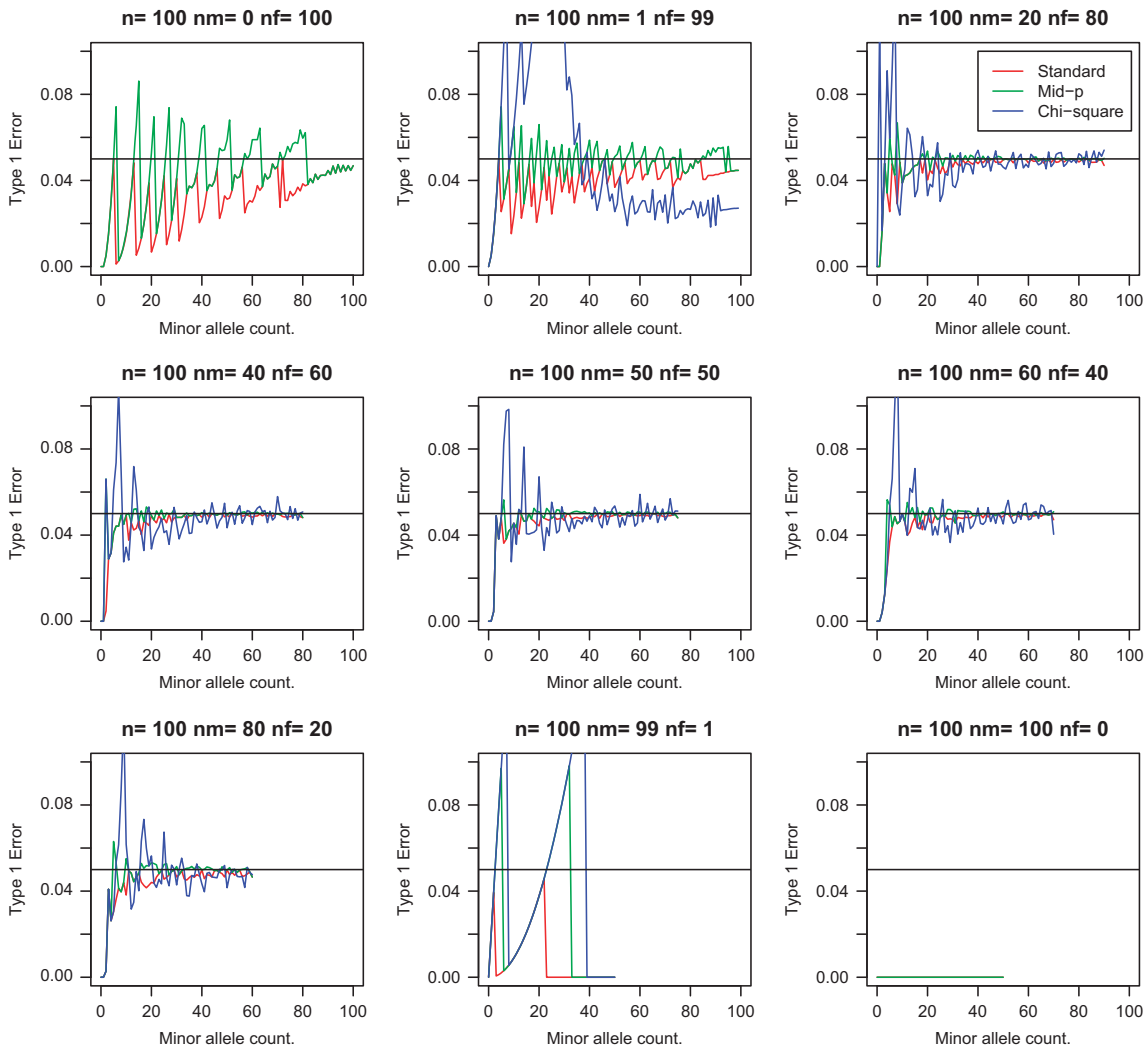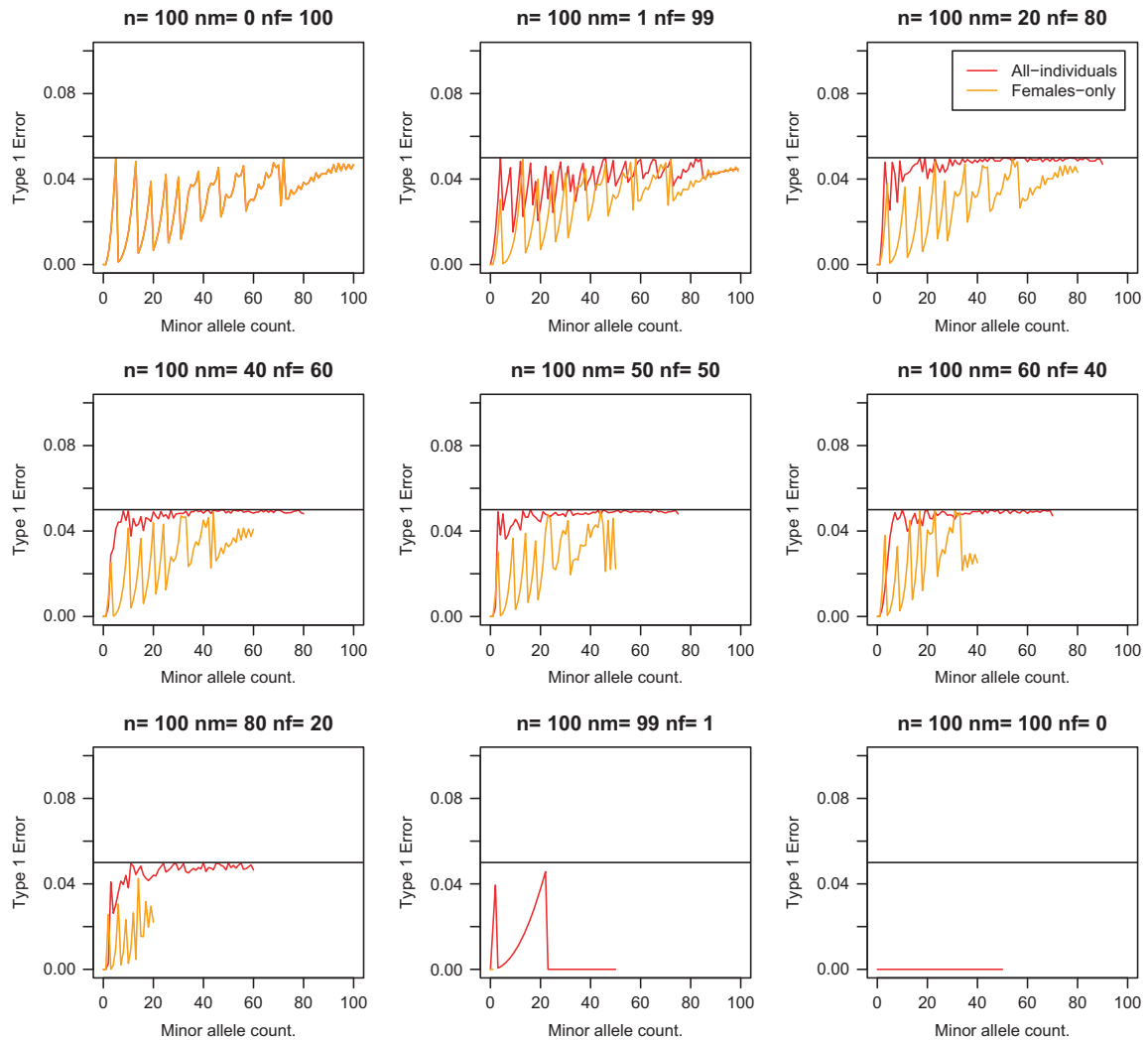
Figure 4 Type I error rate of X-chromosomal tests for HWP as a function of sex ratio and MAF. The Type I error rate of an all-individual test for HWP is plotted for the exact test with the standard *P*-value (red), the exact test with the mid *P*-value (green) and the χ^2 test without continuity correction (blue). *n*, sample size (100); nf, number of females; nm, number of males.

look at females only. Males have virtually the same allele frequency. Inclusion of the males lessens the evidence for disequilibrium to some extent, bringing the *P*-value above the 5% threshold. For the fourth marker, males and females have almost the same allele frequency, and the female genotypes are close to HW proportions. In this case, the two tests agree. For all four SNPs, χ^2 and exact *P*-values do closely agree. We applied the new tests to all 4158 selected SNPs of the database and compare the *P*-values of tests with all individuals with those obtained by discarding males. Plots of the χ^2 and exact *P*-values are shown in Figure 6 in their original scale (Figures 6a and c for χ^2 and exact respectively) and after a $-\log_{10}$ transformation (Figures 6b and d). The latter transformation is usually applied to focus on the lower tail of the distribution.

These plots show a large degree of scatter and a relatively poor agreement between the *P*-values of the all-individual and females-only tests. It is possible that a marker has a *P*-value close to 1 in a test based on females only, whereas it is significant in a test that uses males and females. When we focus on the lower tail of the distribution by applying a $-\log_{10}$ transformation, a set of markers for which inference changes qualitatively (from significant to nonsignificant or the reverse) is uncovered. This set comprises ~5% of the total number of SNPs

analysed. SNPs in Figure 6 are colour coded according to their significance level in Fisher's exact test for the equality of male and female allele frequencies. This shows that SNPs that become significant upon changing from a females-only test to an all-individual test typically have significant differences in male and female allele frequencies, whereas markers that become nonsignificant typically do not. Of the SNPs, 4.3% show a significant difference in allele frequency between males and females. If there are significant differences in allele frequencies between the sexes, then the all-individual test tends to have a smaller *P*-value, whereas if there are no such differences, it tends to have a larger *P*-value. The χ^2 and exact *P*-values correlate well (plots not shown), and particularly so if low MAF markers are excluded. The large difference between an all-individual and a females-only test cannot be explained by the presence of low MAF markers. Figure 6 largely remains the same when low MAF markers are removed (result not shown). We study the distribution of the *P*-values by making Q–Q plots for the *P*-values of both tests, as is shown in Figure 7.

The χ^2 tests (Figures 7a and b) show outliers and a lower tail of the *P*-value distribution that deviates from the uniform distribution. This is because of the presence of highly significant low MAF markers.

Figure 5 Type I error rate of tests for HWP as a function of sex ratio and MAF. The Type I error rate of an all-individual test for HWP on the X chromosome is plotted for the exact test with the standard *P*-value (red), and for the exact test that excluded the males (orange). *n*, sample size (100); nf, number of females; nm, number of males.

Table 3 Genotype counts (m_A, m_B, f_{AA}, f_{AB}, f_{BB}), allele frequencies of males and females (p_{Am}, p_{Af}), χ^2 *P*-values, standard exact *P*-values and exact mid *P*-values for the all-individual test and the female-only test for four single-nucleotide polymorphisms (SNPs) of the venous thrombosis database

SNP	Genotypes					Allele frequencies		χ^2 P-values		Exact P-values		Exact mid P-values	
	m_A	m_B	f_{AA}	f_{AB}	f_{BB}	p_{Am}	p_{Af}	χ^2_2 (all)	χ^2_1 (female)	Exact (all)	Exact (female)	Exact (all)	Exact (female)
rs6646338	399	205	230	314	107	0.661	0.594	0.022	0.992	0.021	1.000	0.021	0.968
rs12010339	603	2	651	0	0	0.997	1.000	0.116	1.000	0.101	1.000	0.051	0.500
rs5935567	372	233	231	337	83	0.615	0.614	0.064	0.019	0.067	0.021	0.067	0.019
rs5968922	392	212	275	296	80	0.649	0.650	0.999	0.980	1.000	1.000	0.999	0.966

The χ^2 test for autosomal markers has been reported to be often too liberal for low MAF markers (Graffelman and Moreno, 2013) and this also happens for the X-chromosomal χ^2 test proposed here. The exact *P*-values show better agreement with a uniform distribution (Figures 7c and d). We focus on the most significant markers in the all-individual and the females-only tests.

We found the all-individual exact test to detect more significant SNPs. The all-individual test detected 188, 40 and 4 significant SNPs at

the 5%, 1% and 0.1% level, respectively. For the females-only test, there were 154, 22 and 2 significant SNPs at these levels. The most significant markers detected in the all-individual exact test are basically SNPs that have different allele frequencies for males and females. In females-only tests, significant disequilibrium was mostly observed because of a lack of heterozygotes (60% of the 154 SNPs).

Cluster plots of four markers are shown in Figure 8. Figures 8a–c are among the most significant in an all-individual exact test for HWE.
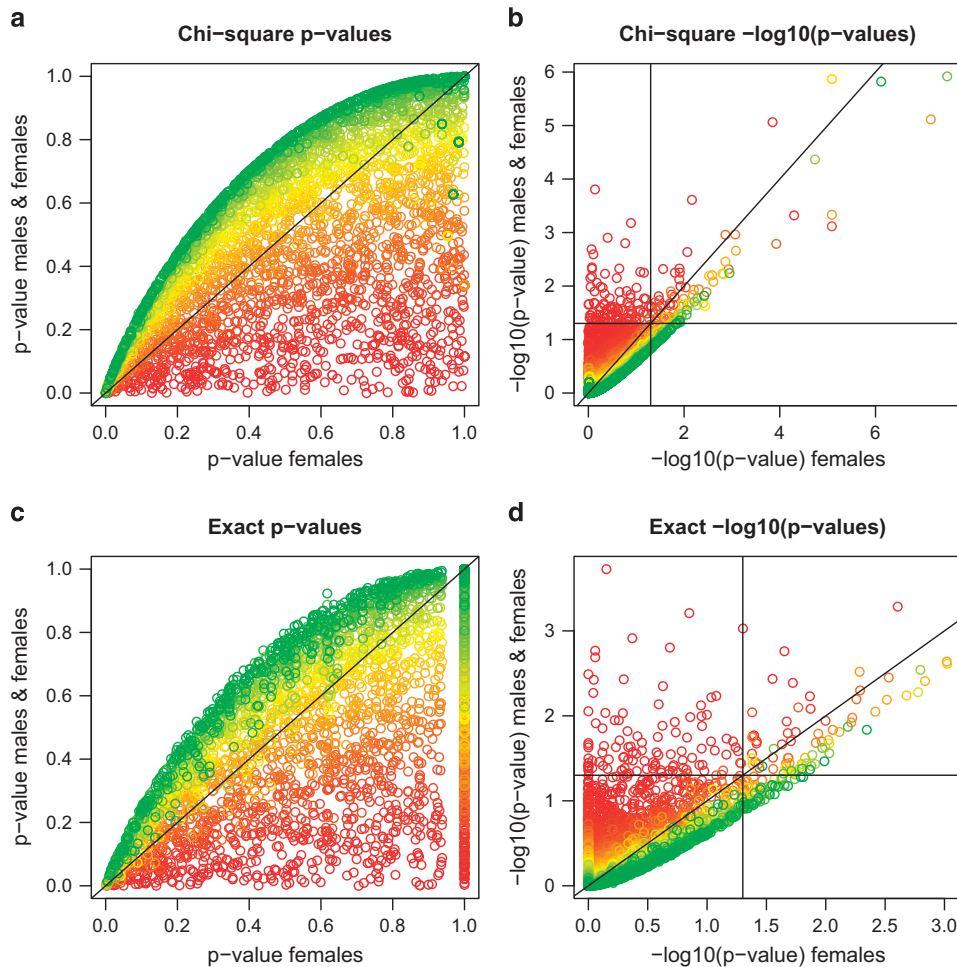
Figure 6 Scatter plots of *P*-values in original and -log10 scale for χ^2 tests (**a**, **b**) and exact tests (**c**, **d**) for HWE using females only and using both males and females for 4158 SNPs at the X chromosome of the venous thrombosis database. The horizontal and vertical black lines in (**b**) and (**d**) correspond to a significance level of 5%. Points colored according to their significance level in Fisher's test for equality of allele frequencies (range 0–1 from red to green).

Figure 8a shows males that have allele intensities close to zero, and suggests the possible existence of null alleles. Figure 8b is representative for many significant SNPs and shows no apparent sign of genotyping errors. Figure 8c suggests that there is copy number variation, and Figure 8d shows null alleles and missing values. In general, X-chromosomal cluster plots often show poor separation of homozygous females and the corresponding hemizygous males.

DISCUSSION

There are many research papers on HWE testing for autosomal markers. The issue of testing markers at the X chromosome for HWE has apparently received little attention before, and the standard approach is to ignore males and test HWE with the autosomal procedures for the females only. In this paper we have presented four frequentist methods for testing markers at the X chromosome for HWE that take both males and females into account.

We note that the methodology outlined in this paper applies to all diploid species with a genetic sex determination system with a heterogametic and a homogametic sex. The GENEVA data on venous thrombosis analysed above stem from human genetics, but the presented methods are equally relevant for most other mammals and also for taxa with a different genetic sex determination system such as the ZW system in birds and Lepidoptera.

Some parts at the tips of the X chromosome, the pseudo-autosomal regions (Graves *et al.*, 1998), behave as an autosome. For markers in this region, the tests developed in this paper, of course, do not apply and the classical autosomal tests should be used. Though the pseudo-autosomal regions are small, positional information of the markers is thus needed in order to decide which type of HWE test (autosomal or X chromosomal) should be applied.

The proposed χ^2 test allows the user to fix the male to female ratio to 1:1 (or any other value) or to estimate it from the data, at the loss of one degree of freedom. We recommend to estimate ϕ always from the data. Fixing ϕ to 0.5 assumes that the data come from a statistical universe with a 1:1 sex ratio. This assumption is usually unwarranted, as the population sex ratio is typically unknown, and varies with age. Moreover, if males or females are oversampled (as often happens) then it surely makes more sense to estimate the male fraction ϕ from the data. Estimating ϕ from the data controls for possible unbalancedness of the sex ratio, and is the best default for the test.

An heuristic alternative procedure to test X-linked markers for HWE consists of testing first equality of allele frequencies (by χ^2 or Fisher's exact test on the two-way table of alleles by gender), and reject HWE if significant differences are found. For those markers that have no significant difference in allele frequencies, females can be tested for HW proportions in a second step, and HWE is rejected in second
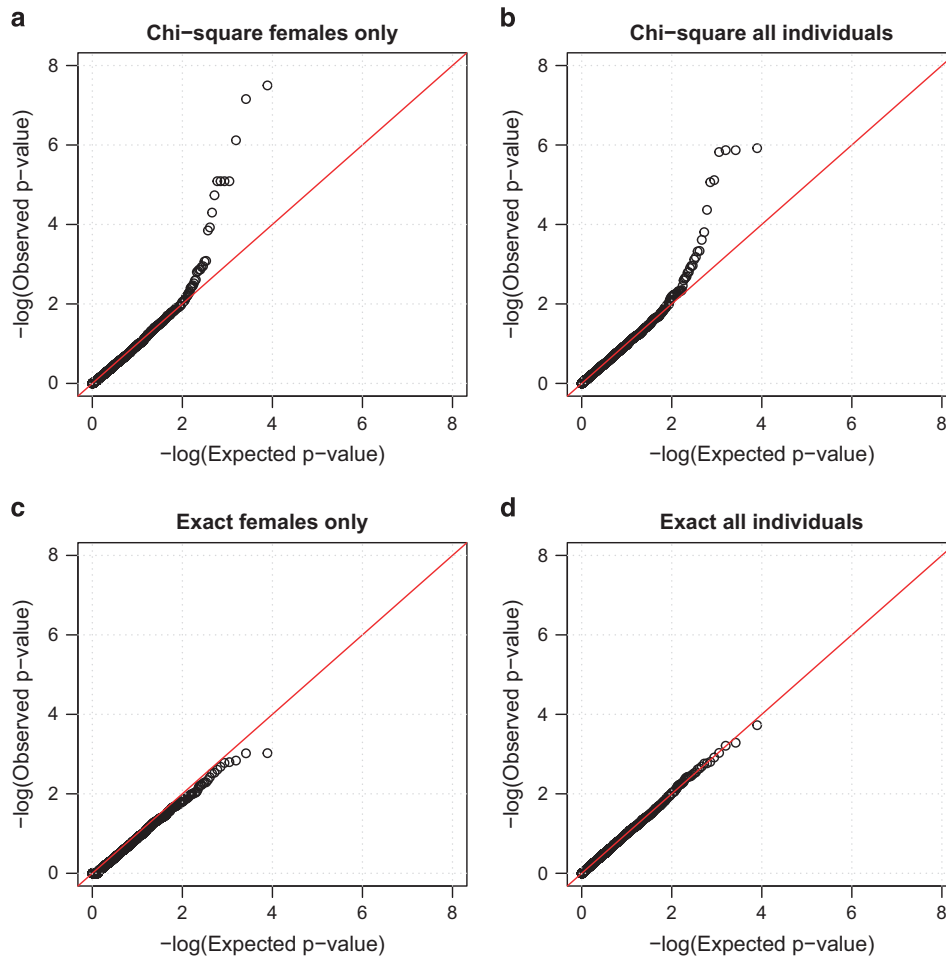
Figure 7 Q–Q plots of $-\log_{10}$ transformed P-values of χ^2 and exact tests for HWE for 4158 SNPs of the venous thrombosis database. (**a**, **c**) Females only and (**b**, **d**) all individuals.

instance if significant deviations are found. This approach unnecessarily increases the number of statistical tests performed and arbitrarily tests equality of allele frequencies before HW proportions. Moreover, a test for equal allele frequencies among the sexes typically assumes HWE from the onset, which seems circular. The tests proposed in this paper are omnibus tests in the sense that equality of allele frequencies and HW proportions for females are tested simultaneously in a single test.

The analysis of the venous thrombosis database shows that both HWE tests can detect markers with possible null alleles (see Figure 8). X-chromosomal null alleles behave like X-linked recessive alleles: they can be detected in hemizygous males for having a zero allele intensity but can easily go unnoticed in heterozygote females who are carriers but have a non-zero allele intensity. If males are ignored, it will become harder to detect X-chromosomal null alleles. The all-individual exact test finds more significant HWE test results for this database. This suggest that the latter has better power than a test based on the females alone. This can in fact be inferred from Figure 5 that shows the all-individual test to have a better Type I error rate, and thus better power.

In the database studied in this paper, the percentage of SNPs significant in an exact test for equality of allele frequencies (4.3%) is approximately what can be expected by chance alone at a 5% significance level (and, in fact, the distribution of the corresponding exact P-values is approximately uniform if the usual spike at P-value 1

is ignored). Though mere chance would be sufficient to explain the observed significant allele frequency differences, we think that the latter may at least in part be due to genotyping error, in particular if there are null alleles or if there is copy number variation. We refer to the cluster plots of some SNPs in the previous section to illustrate this issue. Marker rs3747393 (shown in Figure 8c) is nonsignificant in an exact test for HWP in females (with exact P-value 0.426 in a females-only test). However, Figure 8c shows many males with high allele intensities that are comparable to those of homozygote females, suggesting that males with two copies of the allele may exist. This, together with the inclined nature of the homozygote female clouds, suggests that copy number variation does exist for this SNP. In this case, this would have gone unnoticed in a test for HWP in females alone. Marker rs3747394 (shown in Figure 8d) is also nonsignificant in the HWP test in females ($P=0.873$), though the cluster plot shows missing values, evidence of null alleles and males with non-zero allele intensities for the allele they are inferred not to carry. We think Figure 8d represents signals of genotyping error undetected by a test for HWP in females alone. Difference in allele frequencies are significant for this case ($P=0.032$), and the proposed all-individual test is close to significant ($P=0.091$). We inspected all cluster plots with significant differences in allele frequencies and most of them look unproblematic like Figure 8b. This suggests that significant differences in allele frequencies in our database mostly represent chance effects.
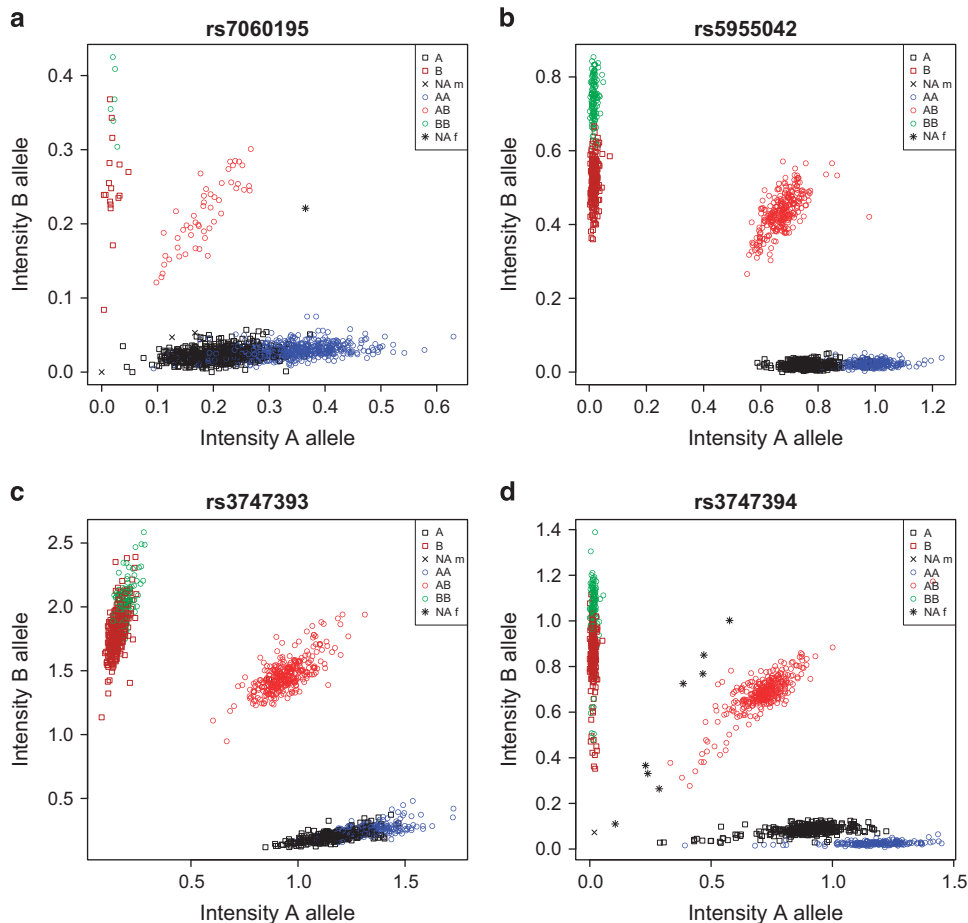
Figure 8 Cluster plots of allele intensities of four SNPs of the venous thrombosis database. (**a** and **b**) are significant in both the female-only (P=0.0025, P=0.0010) and all-individual test (P=0.0005, P=0.0023). (**c**) is non-significant in the female-only test (P=0.4261) but highly significant in the all-individual test (P=0.0012). (**d**) is non-significant in the female-only test (P=0.8732) and close to significant in the all-individual test (P=0.0914).

In many genetic studies, genome-wide association studies in particular, testing genetic markers for HWE is performed for reasons of quality control. This is because disequilibrium has often been found to be associated with genotyping error. We do however not recommend the blind elimination of all markers with significant HWD, precisely because HWD can also be a sign of disease association. Recent papers (Waples, 2015; Graffelman et al., 2015) discuss several factors that should be considered before discarding a marker: the number of missing values, the degree and nature of disequilibrium (lack or excess of heterozygotes), the MAF and the quality of the cluster plot. As shown in this paper, the corresponding HWE tests should take into account whether the marker is X chromosomal or autosomal.

SOFTWARE

All four X-chromosomal tests developed in this paper will be available in the R environment (R Core Team, 2014) if version 1.5.6 of the HardyWeinberg package (Graffelman, 2015) is installed.

DATA ARCHIVING

Genotype data are available at dbGaP (http://www.ncbi.nlm.nih.gov/gap accession number phs000289.v2.p1.)

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Ayres KL, Balding DJ (1998). Measuring departures from Hardy-Weinberg: a Markov chain monte carlo method for estimating the inbreeding coefficient. *Heredity* **80**: 769–777.

Cannings C, Edwards AWF (1968). Natural selection and the de Finetti diagram. *Ann Hum Genet* **31**: 421–428.

Consonni G, Moreno E, Venturini S (2010). Testing Hardy-Weinberg equilibrium: an objective bayesian analysis. *Stat Med* **30**: 62–74.

Crow JF, Kimura M (1970). *An Introduction to Population Genetics Theory*. Harper & Row Publishers.

De Finetti B (1926). Considerazioni matematiche sull'eredità mendeliana. *Metron* **6**: 3–41.

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.

GENEVA (2010). Geneva venous thromboembolism project quality control report. Technical report, Department of Biostatistics, University of Washington.

Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I et al. (2012). Gwastools: an r/bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* **28**: 3329–3331.

Graffelman J, Morales-Camarena J (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Hum Hered* **65**: 77–84.

568

Graffelman J, Moreno V (2013). The mid *p*-value in exact tests for Hardy-Weinberg equilibrium. *Stat Appl Genet Mol Biol* **12**: 433–448.

Graffelman J, Nelson SC, Gogarten SM, Weir BS (2015). Exact inference for Hardy-Weinberg proportions with missing genotypes: single and multiple imputation. *G3 (Genes, Genomes, Genetics)* **5**: 2365–2373.

Graffelman J (2015). Exploring diallelic genetic markers: the HardyWeinberg package. *J Stat Softw* **64**: 1–23.

Graves JAM, Wakefield MJ, Toder R (1998). The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Hum Mol Genet* **7**: 1991–1996.

Haldane JBS (1954). An exact test for randomness of mating. *J Genet* **52**: 631–635.

Hartl DL (1980). *Principles of Population Genetics*. Sinauer Associates.

Hedrick PW (2005). *Genetics of Populations*, 3rd edn. Jones and Bartlett Publishers.

Laird NM, Lange C (2011). *The Fundamentals of Modern Statistical Genetics*. Springer: New York, NY, USA.

Lange K (2002). *Mathematical and Statistical Methods for Genetic Analysis,* 2nd edn. Springer: New York, NY, USA.

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

Lindley DV (1988). Statistical inference concerning Hardy-Weinberg equilibrium. In: Bernardo JM, DeGroot MH, Lindley DV, Smith AFM (eds). *Bayesian Statistics, 3*. Oxford University Press. pp 307–326.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D *et al.* (2007). Plink: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* **81**: 559–575.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.

Shoemaker J, Painter I, Weir BS (1998). A bayesian characterization of Hardy-Weinberg disequilibrium. *Genetics* **149**: 2079–2088.

Wakefield J (2010). Bayesian methods for examining Hardy-Weinberg equilibrium. *Biometrics* **66**: 257–265.

Waples RS (2015). Testing for Hardy-Weinberg proportions: have we lost the plot? *J Hered* **106**: 1–19.

Weir BS (1996). *Genetic Data Analysis II*. Sinauer Associates: Massachusetts.

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

You X-P, Zou Q-L, Li J-L, Zhou J-Y (2015). Likelihood ratio test for excess homozygosity at marker loci on X chromosome. *PLoS One* **10**: e0145032.

Zheng G, Joo J, Zhang C, Geller NL (2007). Testing association for markers on the X chromosome. *Genet Epidemiol* **31**: 834–843.