Received: 2 March 2017 | Revised: 22 May 2017 | Accepted: 17 August 2017

DOI: 10.1002/gepi.22079

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

On the testing of Hardy-Weinberg proportions and equality of allele frequencies in males and females at biallelic genetic markers

Jan Graffelman[1,2] | Bruce S. Weir[2]

[1]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain

[2]Department of Biostatistics, University of Washington, Seattle, Washington, United States of America

Correspondence
Jan Graffelman, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Avinguda Diagonal 647, 08028 Barcelona, Spain.
Email: jan.graffelman@upc.edu

Funding information:
Grant sponsor: Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR); Grant number: 2014SGR551; Grant sponsor: Spanish Ministry of Economy and Competitiveness and European Regional Development Fund; Grant number: MTM2015-65016-C2-2-R (MINECO/FEDER); Grant sponsor: United States National Institutes of Health; Grant number: GM075091.

ABSTRACT

Standard statistical tests for equality of allele frequencies in males and females and tests for Hardy-Weinberg equilibrium are tightly linked by their assumptions. Tests for equality of allele frequencies assume Hardy-Weinberg equilibrium, whereas the usual chi-square or exact test for Hardy-Weinberg equilibrium assume equality of allele frequencies in the sexes. In this paper, we propose ways to break this interdependence in assumptions of the two tests by proposing an omnibus exact test that can test both hypotheses jointly, as well as a likelihood ratio approach that permits these phenomena to be tested both jointly and separately. The tests are illustrated with data from the 1000 Genomes project.

KEYWORDS

Akaike's information criterion, exact test, inbreeding coefficient, likelihood ratio test, ternary diagram

1 | INTRODUCTION

Quality control filtering of genetic data is a crucial procedure in modern genetic studies. Extensive procedures and protocols are used to filter genetic data prior to their use in association tests (Laurie et al., 2010). Such procedures include, but are not limited to, gender checks, assessment of relatedness between individuals, population substructure investigation, tests for Hardy-Weinberg equilibrium (Gomes et al., 1999; Hosking et al., 2004; Leal, 2005), and missing data analysis.

In this paper, we focus on two closely related aspects of the quality control of biallelic genetic markers, the equality of allele frequencies (EAF) in the sexes and Hardy-Weinberg proportions (HWP). Under normal conditions, we expect an autosomal genetic marker to have equal allele frequencies in males and females, and with genotype frequencies that agree with the Hardy-Weinberg law. EAF can be tested by a chi-square or Fisher's exact test on a two-way table where all alleles are cross-classified according to sex and type of allele (A or B). If we let n represent the sample size (number of individuals), then such testing assumes the $2n$ alleles to be independent, and therefore the EAF test relies on the assumption of HWP. It thus seems natural to test for HWP prior to testing for EAF. A genetic marker can be tested for HWP by means of a chi-square or an exact test, among others (Weir, 1996, Chapter 3). These tests assess to what extent observed genotypic proportions (f_{AA}, f_{AB}, f_{BB}) deviate from the theoretically expected proportions $(p^2, 2pq, q^2)$, p and q being the A and B allele frequency, respectively, with $p + q = 1$. It is thereby implicitly assumed that the allele frequencies p and q are the same in males and females. This assumption might be true or not, and it thus seems necessary to test for EAF prior to testing for HWP. We are thus caught in a vicious testing circle depicted in Figure 1.

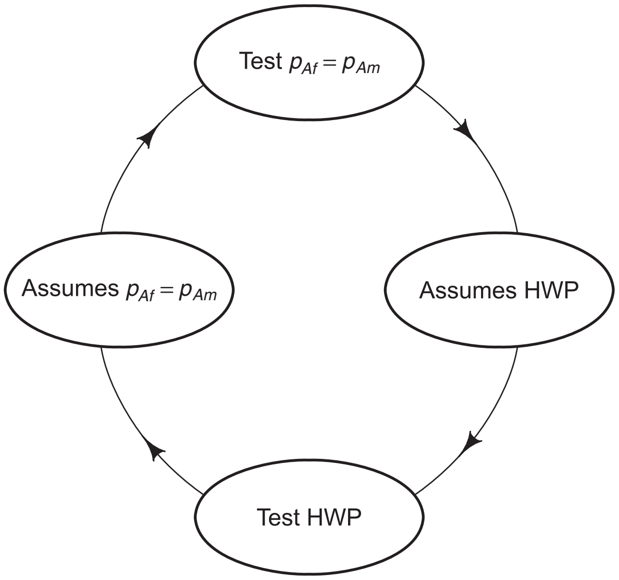

FIGURE 1 Vicious testing circle: mutual dependency of a test for EAF in males and females and a test for HWP
Notes: *A* allele frequencies in males and females are represented by p_{Am} and p_{Af}, respectively.

In this paper, we address ways to break the mutual dependency between the HWP and EAF test outlined above, but first motivate the relevance of the issue with an empirical example. SNP rs147120681 at chromosome 1 of the 104 individuals of the Japanese (JPT) sample of the 1000 Genomes project (The 1000 Genomes Project Consortium, 2015) has genotype counts of AA = 23, AB = 18, BB = 15 for males and AA = 7, AB = 32, BB = 9 for females, summing to AA=30, AB=50, BB=24 in total. Applying standard quality control, an exact test for HWP clearly finds no evidence for disequilibrium ($P = 0.6981$). When we test for EAF by a Fisher's exact test, we also obtain a nonsignificant result ($P = 0.2107$). Using these tests separately and observing *both* to be nonsignificant, we are led to believe that the marker is well-behaved, and that there are no reasons to suspect any genotyping error.

However, strictly speaking we do not know if equilibrium holds, or that we failed to reject it because the assumptions of the test were not met, and we neither know if the allele frequencies are really homogeneous, or we failed to reject the null because the HWP assumption was not met. Preferably, one would like to test these phenomena independently, or jointly in one step. We will reanalyze SNP rs147120681 in Section 5, once we have developed the statistical procedures that avoid the dependence in assumptions, to arrive at a different conclusion about this variant.

Two ways to break the mutual dependence between the HWP test and the EAF test are considered. One approach is to test HWP and EAF simultaneously in a single omnibus test. This approach has been used by Graffelman and Weir (2016) to test biallelic variants on the X chromosome for HWP. An omnibus test seems attractive, as it allows two aspects of quality control to be tested with a single statistical test. Alternatively, with a flexible likelihood ratio (LR) approach, disequilibrium and allele frequency differences can be modeled with multiple parameters, allowing both phenomena to be tested jointly or separately. In this paper, we develop an omnibus exact procedure to test HWP and EAF jointly and we also develop LR procedures for testing HWP and EAF both jointly and separately. Extensions for multiple alleles, and a Bayesian approach, are considered beyond the scope of the current paper and left for future work.

For biallelic markers, the Hardy-Weinberg law can be graphically represented by a parabola in a ternary diagram (Cannings & Edwards, 1968; Li, 1976; Graffelman & Morales-Camarena, 2008). If autosomal genotype frequencies of both sexes are distinguished, then several scenarios are possible, which are also conveniently represented in ternary diagrams, as is shown in Figure 2. Under normal conditions, we expect a marker to be in Hardy-Weinberg equilibrium with equal allele frequencies in males and females, as represented by Figure 2 A. If a marker is out of equilibrium, then in general we expect this to affect males and females in the same manner. This is represented in Figure 2 B, where males and females have the same allele frequencies and the same inbreeding coefficient. Alternatively, as represented in Figure 2 C, both sexes can have equal allele frequencies but different inbreeding coefficients (in magnitude and, possibly, in direction too). When the allele frequencies of the sexes differ: males and females can still be in HWP, as shown in Figure 2 D; can have similar inbreeding coefficients as in Figure 2 E; or can have different inbreeding coefficients as in Figure 2 F.

The structure of the remainder of this article is as follows. In Sections 2 and 3, we develop an omnibus exact test and LR tests, respectively. In Section 4, we study the Type I error rate and the power of the omnibus tests. Section 5 shows applications of exact and LR tests. Section 6 presents a discussion. Some mathematical derivations are given in an Appendix.

2 | OMNIBUS EXACT TEST

In this section we develop an exact test that jointly tests HWP and EAF for an autosomal marker. We will use the following notation for developing our test procedures. Let $P_{AAm}, P_{ABm}, P_{BBm}, P_{AAf}, P_{ABf}$, and P_{BBf} be the male and female genotype frequencies in the population with $P_{AAm} + P_{ABm} + P_{BBm} = P_{AAf} + P_{ABf} + P_{BBf} = 1$. Let M_{AA}, M_{AB}, M_{BB}, F_{AA}, F_{AB} , and F_{BB} represent random variables for the male (M) and female (F) genotype counts, respectively, that take on observed values m_{AA}, m_{AB}, m_{BB}, f_{AA}, f_{AB} , and f_{BB} in the sample. If no distinction is made between the sexes, as in the classical autosomal case, the notation N_{AA}, N_{AB} , and N_{BB} with observed values n_{AA}, n_{AB} , and n_{BB} will be used. Let n_m be the num-

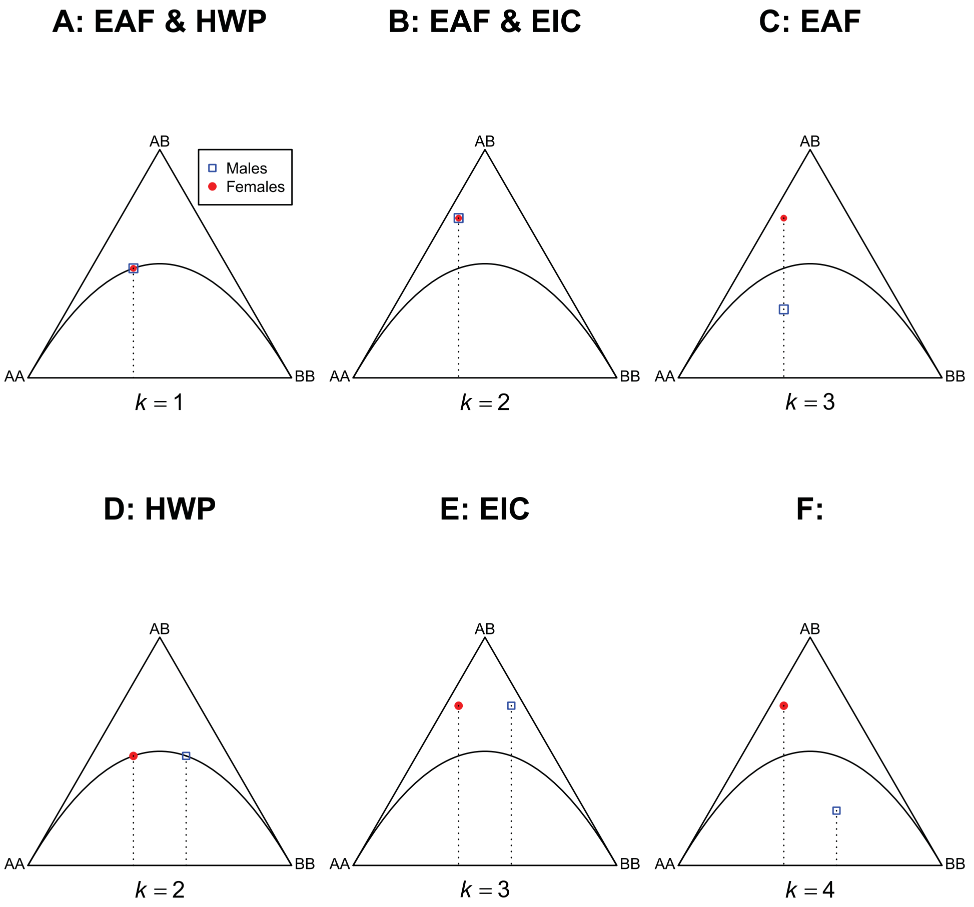

FIGURE 2 Ternary diagrams for male and female genotype frequencies

Notes: (A) HWP and EAF. (B) Equality of inbreeding coefficients, EAF, and both sexes out of HWP. (C) Unequal inbreeding coefficients, both sexes out of equilibrium but with equal allele frequencies. (D) Both sexes in HWP but with different allele frequencies. (E) Each sex out of equilibrium with identical inbreeding coefficients and different allele frequencies. (F) Both sexes out of equilibrium, with different inbreeding coefficients and different allele frequencies. The number of free parameters k is given below the basis of each scenario.

ber of males and n_f the number of females, such that $n_m = m_{AA} + m_{AB} + m_{BB}$ and $n_f = f_{AA} + f_{AB} + f_{BB}$ and the total sample size is $n = n_m + n_f$. Let F_A, and F_B, be the number of A and B alleles in females, and M_A and M_B the number of these alleles in males. The total A and B allele counts are $N_A = M_A + F_A$ and $N_B = M_B + F_B$, respectively, with sample values n_A, m_A, f_A, n_B, m_B , and f_B. Finally, let ρ_m and ρ_f be the inbreeding coefficients of males and females, respectively, given by:

$$\rho_m = \frac{P_{AAm} - p_{Am}^2}{p_{Am}(1 - p_{Am})}, \qquad \rho_f = \frac{P_{AAf} - p_{Af}^2}{p_{Af}(1 - p_{Af})}.$$

We base our inference for HWP *and* EAF on the joint distribution of the number of male and female heterozygotes. Under the assumptions of HWP and EAF, this joint distribution is given by:

$$P\left(M_{AB}, F_{AB} \mid n, n_A, n_m\right) = \frac{n_A! n_B! n_m! n_f!}{m_{AA}! m_{AB}! m_{BB}! f_{AA}! f_{AB}! f_{BB}! (2n)!} 2^{m_{AB}+f_{AB}}. \quad (1)$$

This joint density resembles the density used in the omnibus exact test for markers on the X chromosome recently proposed by Graffelman and Weir (2016). A derivation of this joint density is given in the Appendix, where its relationship with the classical autosomal and the X chromosomal test is shown as well. Rejection of the null may be caused by genotype frequencies in the population deviating from HWP, by unequal allele frequencies, or by both these factors simultaneously, or can be a chance effect in the sample. We consider a toy example sample of six males and seven females with genotype counts ($m_{AA} = 1, m_{AB} = 2, m_{BB} = 3, f_{AA} = 0, f_{AB} = 2, f_{BB} = 5$) to illustrate the calculations. Table 1 shows all possible samples for the given minor allele (A) count of six, together with their probabilities according to Equation (1).

The observed sample (row 27 of Table 1) has probability 0.0876. The sum of all probabilities of all samples having a probability smaller or equal to 0.0876 is 0.5500. At a usual significance level of $\alpha = 0.05$, the composite null hypothesis of HWP *and* EAF is not rejected. Recently, the use of the mid P-value has been recommended for exact tests

TABLE 1 All possible samples (30) for a set of 13 individuals (six males and seven females) with a total of six A alleles, and their probabilities

Sample	m_{AA}	m_{AB}	m_{BB}	f_{AA}	f_{AB}	f_{BB}	$P(M_{AB}, F_{AB})$	Cum.
1	3	0	3	0	0	7	0.0001	0.0001
2	0	0	6	3	0	4	0.0001	0.0002
3	0	6	0	0	0	7	0.0003	0.0005
4	2	0	4	1	0	6	0.0005	0.0010
5	1	0	5	2	0	5	0.0006	0.0015
6	2	2	2	0	0	7	0.0016	0.0031
7	0	0	6	0	6	1	0.0019	0.0050
8	1	4	1	0	0	7	0.0021	0.0071
9	0	0	6	2	2	3	0.0036	0.0108
10	2	0	4	0	2	5	0.0055	0.0162
11	0	2	4	2	0	5	0.0055	0.0217
12	0	4	2	1	0	6	0.0073	0.0290
13	0	0	6	1	4	2	0.0073	0.0363
14	2	1	3	0	1	6	0.0073	0.0436
15	1	2	3	1	0	6	0.0073	0.0509
16	1	0	5	1	2	4	0.0109	0.0618
17	0	1	5	2	1	4	0.0109	0.0728
18	0	5	1	0	1	6	0.0117	0.0845
19	1	0	5	0	4	3	0.0146	0.0991
20	1	1	4	1	1	5	0.0219	0.1210
21	1	3	2	0	1	6	0.0292	0.1501
22	0	1	5	0	5	2	0.0350	0.1852
23	0	3	3	1	1	5	0.0584	0.2435
24	0	1	5	1	3	3	0.0584	0.3019
25	1	1	4	0	3	4	0.0730	0.3749
26	0	4	2	0	2	5	0.0876	0.4625
27	1	2	3	0	2	5	0.0876	0.5500
28	0	2	4	1	2	4	0.1095	0.6595
29	0	2	4	0	4	3	0.1459	0.8054
30	0	3	3	0	3	4	0.1946	1.0000

The last column (Cum.) gives the cumulative probabilities. The observed sample is marked in red.

for HWP (Graffelman & Moreno, 2013). The mid *P*-value, calculated as half the probability of the observed sample plus the sum of the probabilities of more extreme samples, for this example is 0.5062 and points to the same conclusion. Note that samples 26 and 27 have the same probability and that 26 is therefore included in the sum that constitutes the *P*-value.

3 | LIKELIHOOD RATIO TESTS

In this section, we develop LR tests for HWP and EAF. Similar work has been done by Zheng, Joo, Zhang, and Geller (2007) and You, Zou, Li, and Zhou (2015) for the X chromosome. To the best of our knowledge, a likelihood framework for jointly addressing HWP and EAF on the autosomes has hitherto not been developed. The LR approach is flexible, because it allows us to test HWP and EAF jointly, but also separately and it can avoid the dependence outlined in Figure 1. The probabilistic model used to describe the data is again the multinomial distribution, but with different allele frequencies for males and females and different inbreeding coefficients for males and females. The full model for the data is, conditioning on the observed number of males and females, obtained by multiplying the multinomial likelihoods of males and females:

$$L(\theta) = \binom{n_m}{m_{AA}, m_{AB}, m_{BB}} P_{AAm}{}^{m_{AA}} P_{ABm}{}^{m_{AB}} P_{BBm}{}^{m_{BB}} \times \binom{n_f}{f_{AA}, f_{AB}, f_{BB}} P_{AAf}{}^{f_{AA}} P_{ABf}{}^{f_{AB}} P_{BBf}{}^{f_{BB}} \quad (2)$$

with

$$\begin{aligned} P_{AAm} &= p_{Am}^2 + p_{Am}(1-p_{Am})\rho_m, \\ P_{ABm} &= 2p_{Am}(1-p_{Am})(1-\rho_m), \\ P_{BBm} &= (1-p_{Am})^2 + p_{Am}(1-p_{Am})\rho_m, \\ P_{AAf} &= p_{Af}^2 + p_{Af}(1-p_{Af})\rho_f, \\ P_{ABf} &= 2p_{Af}(1-p_{Af})(1-\rho_f), \\ P_{BBf} &= (1-p_{Af})^2 + p_{Af}(1-p_{Af})\rho_f, \end{aligned} \quad (3)$$

where $\theta = (p_{Am}, p_{Af}, \rho_m, \rho_f)$ is the parameter vector. Closed form expressions for the maximum likelihood estimators exist, and are given by:

$$\hat{p}_{Am} = \frac{2m_{AA} + m_{AB}}{2n_m}, \quad \hat{\rho}_m = \frac{4m_{AA}m_{BB} - m_{AB}^2}{n_{Am}n_{Bm}}, \quad (4)$$

$$\hat{p}_{Af} = \frac{2f_{AA} + f_{AB}}{2n_f}, \quad \hat{\rho}_f = \frac{4f_{AA}f_{BB} - f_{AB}^2}{f_A f_B}.$$

These expressions are the same as the well-known autosomal estimators, but then applied to the genotype counts of each gender separately. Several hypothesis tests of interest can now be developed and are detailed in the following sections. For each hypothesis, we initially use the unrestricted full four parameter model as the alternative.

Scenario A: EAF and HWP

If no disturbing factors (selection, migration, etc.) are operating, one expects EAF and HWP, which can be phrased as the null hypothesis $H_0 : p_{Af} = p_{Am} \cap \rho_m = \rho_f = 0$. Under the null, the sexes are not distinguished, which we parametrize as $p_A = p_{Af} = p_{Am}$ and $\rho = \rho_f = \rho_m = 0$. The ML estimator of p_A is the usual autosomal allele count estimator given by $\hat{p}_A = (2n_{AA} + n_{AB})/(2n)$. We can test this hypothesis by using the generalized LR statistic, $\Lambda_A = L(\hat{\theta}_0)/L(\hat{\theta}_1)$, where $\hat{\theta}_0 = (\hat{p}_A, \hat{p}_A, 0, 0)$ and $\hat{\theta}_1 = (\hat{p}_{Am}, \hat{p}_{Af}, \hat{\rho}_m, \hat{\rho}_f)$ are the constrained and unconstrained maximizers of L, respectively. We have $G_A^2 = -2\ln(\Lambda_A)$, and asymptotically $G_A^2 \sim \chi^2_{(3)}$. At a conventional significance threshold of $\alpha = 0.05$ one rejects the null of HWP and EAF if G_A^2 exceeds 7.81.

Scenario B: EAF and EIC

Under this scenario, deviation from Hardy-Weinberg equilibrium is admitted, but the inbreeding coefficient is assumed to be the same in both sexes, such that we have equality of inbreeding coefficients (EIC). The corresponding null hypothesis can be stated as $H_0 : p_{Af} = p_{Am} \cap \rho_f = \rho_m$. Under the null, the sexes are not distinguished, which we parametrize as $p_A = p_{Af} = p_{Am}$ and $\rho = \rho_f = \rho_m$. The ML estimators of p_A and ρ are the usual autosomal estimators given by $\hat{p}_A = (2n_{AA} + n_{AB})/(2n)$ and $\hat{\rho} = (4n_{AA}n_{BB} - n_{AB}^2)/(n_A n_B)$. We can test this hypothesis by using the LR statistic $\Lambda_B = L(\hat{\theta}_0)/L(\hat{\theta}_1)$ with $\hat{\theta}_0 = (\hat{p}_A, \hat{p}_A, \hat{\rho}, \hat{\rho})$ and $\hat{\theta}_1 = (\hat{p}_{Am}, \hat{p}_{Af}, \hat{\rho}_m, \hat{\rho}_f)$ and we have $G_B^2 = -2\ln(\Lambda_B)$, and asymptotically $G_B^2 \sim \chi^2_{(2)}$. At a conventional significance threshold of $\alpha = 0.05$ one rejects the null of EAF and EIC if G_B^2 exceeds 5.99.

Scenario C: EAF Only

This scenario assumes EAF, but possibly different inbreeding coefficients for males and females. The null hypothesis is now simply $H_0 : p_{Af} = p_{Am} = p_A$, with no restrictions on the inbreeding coefficients. No closed form expressions for the ML estimators of the parameters were found. ML estimators were therefore obtained by maximizing the likelihood function numerically, using R-package Rsolnp (Ghalanos & Theussl, 2015). Maximization respected the nonlinear constraint $-\min(p_A, p_B)/(1 - \min(p_A, p_B)) \le \rho_m \le 1$ for males; the analogous constraint was used for females with ρ_m replaced by ρ_f. The LR statistic is $\Lambda_C = L(\hat{\theta}_0)/L(\hat{\theta}_1)$ with $\hat{\theta}_0 = (\hat{p}_A, \hat{p}_A, \hat{\rho}_m, \hat{\rho}_f)$ and $\hat{\theta}_1 = (\hat{p}_{Am}, \hat{p}_{Af}, \hat{\rho}_m, \hat{\rho}_f)$ and we have $G_C^2 = -2\ln(\Lambda_C)$, and asymptotically $G_C^2 \sim \chi^2_{(1)}$. At a conventional significance threshold of $\alpha = 0.05$ one rejects the null of EAF if G_C^2 exceeds 3.84. This test breaks the vicious circle in Figure 1, as it is a test for EAF that is free of the HWP assumption.

Scenario D: HWP in Both Sexes

In this scenario, there is equilibrium in both sexes such that we have $H_0 : \rho_m = \rho_f = 0$, whereas male and female allele frequencies can freely vary. The ML estimators for the allele frequencies are $\hat{p}_{Am} = (2m_{AA} + m_{AB})/(2n_m)$ and $\hat{p}_{Af} = (2f_{AA} + f_{AB})/(2n_f)$. We can test this hypothesis by using the LR statistic $\Lambda_D = L(\hat{\theta}_0)/L(\hat{\theta}_1)$ with $\hat{\theta}_0 = (\hat{p}_{Am}, \hat{p}_{Af}, 0, 0)$ and $\hat{\theta}_1 = (\hat{p}_{Am}, \hat{p}_{Af}, \hat{\rho}_m, \hat{\rho}_f)$. We have $G_D^2 = -2\ln(\Lambda_D)$, and asymptotically $G_D^2 \sim \chi^2_{(2)}$. This test also breaks the vicious circle in Figure 1, as it is a test for HWP that does not make the EAF assumption.

Scenario E: EIC

In this scenario, the inbreeding coefficient is the same in both sexes, and their allele frequencies are not restricted, such that we have $H_0 : \rho_m = \rho_f = \rho$. No closed form expressions for the ML estimators were obtained, and for this scenario, we also maximized the likelihood numerically, using the constraint $-\min(p_{Am}, p_{Bm})/(1 - \min(p_{Am}, p_{Bm})) \le \rho \le 1$ for male allele frequencies. The same constraint was applied to females, replacing p_{Am} and p_{Bm} by p_{Af} and p_{Bf}, respectively. The null hypothesis can be tested with the LR statistic $\Lambda_E = L(\hat{\theta}_0)/L(\hat{\theta}_1)$ with $\hat{\theta}_0 = (\hat{p}_{Am}, \hat{p}_{Af}, \hat{\rho}, \hat{\rho})$ and $\hat{\theta}_1 = (\hat{p}_{Am}, \hat{p}_{Af}, \hat{\rho}_m, \hat{\rho}_f)$. We have $G_E^2 = -2\ln(\Lambda_E)$, and asymptotically $G_E^2 \sim \chi^2_{(1)}$.

Scenario F

Scenario F corresponds to the full model that does not have any additional constraints on the parameters beyond the usual range constraints for inbreeding coefficients and allele frequencies. The ML estimators for this scenario were given in Equation (4).

In the foregoing, we have used the entirely unrestricted scenario F as a reference against which the other scenarios were compared. In particular, an LR test of scenario A against F is a joint test for EAF and HWP, and the LR test of D against F establishes a test for HWP that does not assume the EAF. Likewise, an LR test of scenario C against F is a test for EAF that does not rely on the assumption of HWP. However, many other scenarios have a nested relationship, with one being a particular case of another. For example, A, B, and C are particular instances of D, E, and F, respectively, and thus we could also test A versus D, B versus E, and C versus F by an LR test, all three corresponding LR statistics having a $\chi^2_{(1)}$ distribution under the null. If EAF is assumed, then A can also be tested against B, B against C, and A against C. Likewise, if EAF is not assumed, D can also be tested against E, or E against F, or D against F, for these are all nested models. The degrees of freedom for the corresponding LR statistics are calculated as the difference in number of parameters of the two scenarios involved. The number of free parameters for each model (k) is shown in Figure 2. Note that some scenarios cannot be compared for not being nested.

In order to determine which scenario best describes a marker, successive hypothesis tests can be performed until a model is found that cannot be rejected. The principle of parsimony applies, where we favor, among the models that cannot be rejected, the one that has fewer parameters. Alternatively, model selection can also be performed by calculating Akaike's information criterion (AIC; Akaike, 1973), defined as $2k - 2\ln(L(\hat{\theta}))$ for all six models and choosing the model with the smallest AIC.

4 | TYPE I ERROR RATE AND POWER

In this section, we evaluate the proposed omnibus exact and LR tests of the previous sections with Type I error rate and power calculations.

4.1 | Type I Error Rate

We compare the Type I error rate of the omnibus exact and LR tests of the previous sections as a function of the minor allele frequency and the sex ratio. Type I error rates were calculated by exhaustive enumeration, which is computationally expensive, following the procedure detailed by Graffelman and Moreno (2013), and are shown in Figure 3.

Figure 3 shows that the omnibus exact test has a better Type I error rate than the LR test. The convergence to the nominal rate is faster, and the exact test strictly controls the Type I error rate, never exceeding the nominal level. For low MAF variants, the LR test is more conservative than the exact test, and for higher MAF, the LR test is above the nominal level. The Type I error rate of the exact test is improved, coming closer to the nominal level, by using the mid P-value, which is particularly manifest for low MAF variants. This has also been observed for the usual standard autosomal exact test for HWP (Graffelman & Moreno, 2013). Moderate bias in the sex ratio has little influence on the Type I error rate, and changes only its erratic pattern at low MAF. The better Type I error rate profile of the joint exact test implies the latter also has better power than the LR test at low MAF.

4.2 | Power

Exact power calculations require the distributions of the LR statistic and the joint density of M_{AB} and F_{AB} under the alternative hypothesis. These distributions are not readily available, and we therefore evaluate power by carrying out some simulations. Simulations were designed as follows. Genotype data was simulated under the six different scenarios by sampling males and females separately from multinomial distributions with parameters specified by Equation (3). Scenarios A, B, and C were simulated with $\rho_m = \rho_f = 0$, $\rho_m = \rho_f = -0.1$, and $\rho_m = +0.1, \rho_f = -0.1$, respectively, for varying minor allele frequencies that were equal in males and females. We used 10,000 simulations with $\alpha = 0.05$ and $n_m = n_f = 50$. This sample size corresponds closely to the empirical data analyzed in Section 5. Scenarios D, E, and F were also simulated with $\rho_m = \rho_f = 0$, $\rho_m = \rho_f = -0.1$ and $\rho_m = +0.1, \rho_f = -0.1$, respectively, but with a varying ratio of male and female allele frequencies. Power graphics for the six scenarios are shown in Figure 4. Scenarios D, E, and F were simulated twice, for a low male MAF ($p_{Am} = 0.1$), and for a higher male MAF ($p_{Am} = 0.2$). Figures 4.4, 4.5, and 4.6 correspond to the low MAF simulations, and Figures 4.7, 4.8, and 4.9 to the higher MAF simulations. Power was calculated as the fraction of simulations for which the tests rejected the null hypothesis of the corresponding scenario. We evaluated four tests: the standard exact test for HWP (ignoring sex), a standard exact test for EAF (ignoring HWP), the joint exact test for HWP and EAF, and the joint LR ratio test for HWP and EAF (scenario A against F).

Figure 4 A shows that under EAF, the joint exact test has the best Type I error rate, and confirms the LR test is somewhat liberal. Under EAF and EIC, the standard exact test for HWP has better power than the joint exact test. If EIC cannot be assumed, with different signs for the inbreeding coefficient for males and females, the power of the standard exact test for HWP drops, and the joint procedures outperform the standard exact test. For many scenarios, the joint LR test appears to have slightly better power than the joint exact test, but this is most likely due to the fact that the LR test is somewhat liberal, as it does not control strictly the Type I error α. As expected, power is better for larger minor allele frequencies. In general, under EAF power is low and for the given sample size, and does not exceed 0.20.

For scenarios D, E, and F, with differences in allele frequencies between the sexes, the joint exact, the joint LR, and the standard exact test for EAF have similar power, and their power increases when the ratio of the allele frequencies increases. The good power of the EAF exact test may be considered flattered to some extent, because this test is allele-based and has therefore a doubled sample size. Comparison of Figures 4.4, 4.5, and 4.6 with 4.7, 4.8, and 4.9 shows, as expected, that all tests have better power when the MAF of one sex increases. In scenarios E and F, a standard exact test for HWP that ignores gender has in general low power to detect deviation from equilibrium. At extreme p_{Af}/p_{Am} ratios, this test acquires more power. This can be ascribed to the fact that the overall allele frequency is the average of the allele frequency of both sexes, and at this average allele frequency, overall heterozygosity is reduced with respect to HWP (see Discussion).

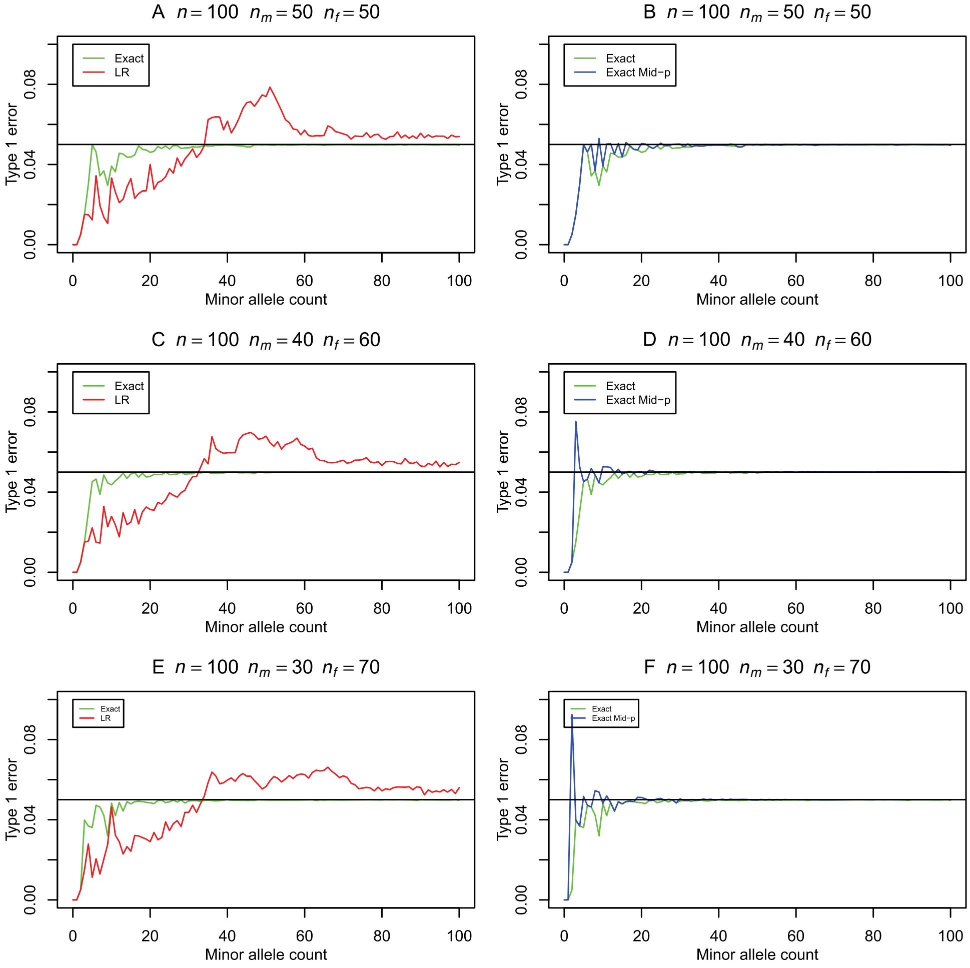

FIGURE 3 Type I error rates for the omnibus exact and LR tests as a function of the MAF and the sex ratio

Note: (A), (C), and (E) show the Type I error rate for exact and LR tests. (B), (D), and (F) show the Type I error rate for the exact test using the standard and mid *P*-value.

5 | EMPIRICAL EXAMPLES

In this section, we apply the previously developed methodology to single nucleotide polymorphisms (SNPs) from the JPT sample of the 1000 Genomes project (The 1000 Genomes Project Consortium, 2015), consisting of 104 individuals, 56 males and 48 females. We first illustrate the methodology by analyzing some individual SNPs, followed by an analysis of some larger genomic areas of the same sample.

5.1 | Single SNPs

We comment on the analysis of six SNPs that correspond to the different scenarios, all represented in Figure 5. In these ternary plots, the acceptance region of a Chi-square test with $n = 52$ (the average of the male and female sample size) and $\alpha = 0.05$ has been indicated (Graffelman & Morales-Camarena, 2008). This makes it possible to judge graphically the significance of the males and the females in separate tests for HWP. We calculated the AIC for all models in order to compare this with the final model obtained by successive hypothesis testing. AIC statistics for all six markers considered are given in Table 2 C.

We first reanalyze a single SNP, rs147120681, previously presented in the Introduction, adopting a significance level $\alpha = 0.05$. If we jointly test HWP and EAF for this marker with the exact test presented in Section 2, we obtain *P*-value 0.0031. Using the LR approach, the joint test (A against F) is also significant ($P = 0.0029$). A test for EAF without assuming HWP (C against F), shows EAF cannot be rejected ($P = 0.1801$). Testing common inbreeding coefficients (B vs. C) gives *P* values 0.0005, indicating the marker is best described by scenario C. Genotype counts and exact test results are summarized in the third row of Table 2 A. This example shows one cannot blindly rely on separate HWP and EAF tests. A ternary diagram of this marker shown in Figure 5 C shows that if a standard HWP test is applied, the differences in inbreeding coefficients between males and females are averaged out, and disequilibrium goes unnoticed. Note that Figure 5 C actually shows HWP has to be rejected when

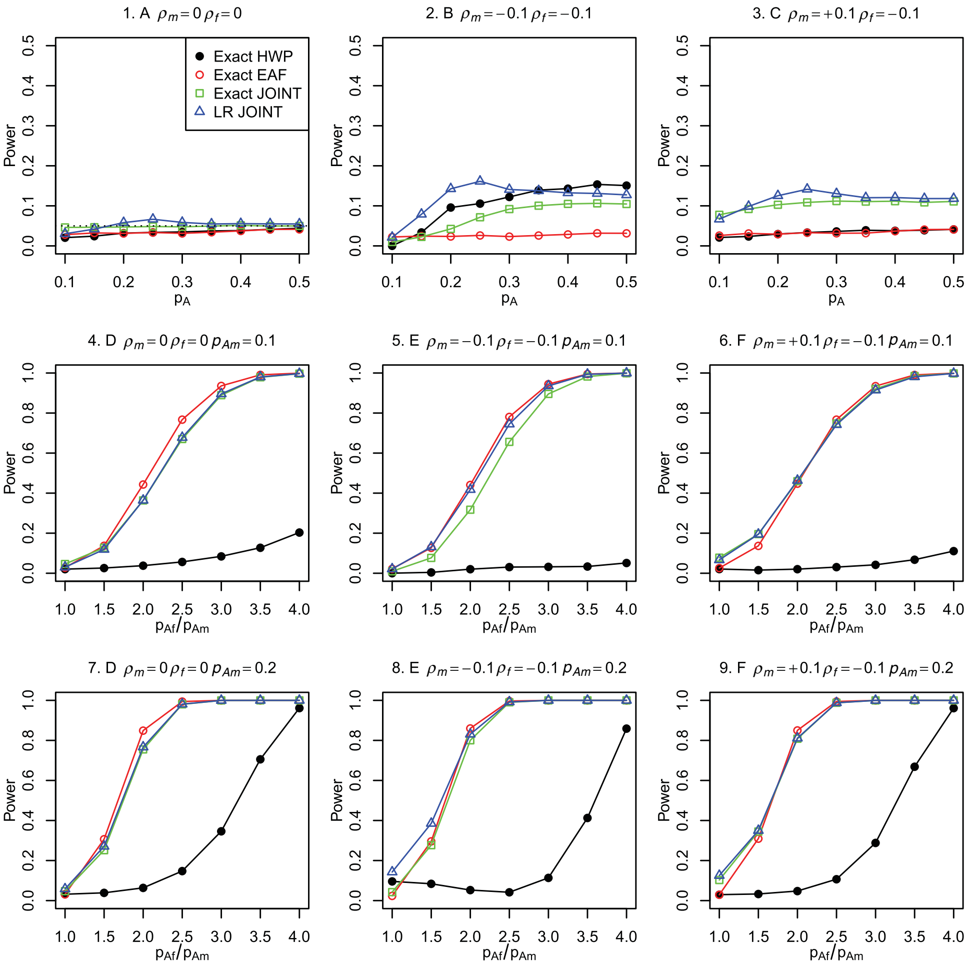

FIGURE 4 Power comparison for the omnibus exact test, the joint likelihood-ratio test, and standard exact procedures for HWP and EAF
Notes: Panel letters A, B,..., F correspond to the theoretical scenarios in Figure 2. Panels 1, 2, and 3 show power as a function of the allele frequency. Panels 4 through 9 show power as a function of the ratio between female and male allele frequencies. For panels 4, 5, and 6, male A allele frequency was set to 0.1, and for panels 7, 8, and 9, male A allele frequency was set to 0.2.

males and females are tested separately. The proposed joint exact test detects disequilibrium and with the LR approach, the most appropriate scenario can be determined. Model C has the smallest value for the AIC statistic.

Polymorphism rs1574243 is nonsignificant in all three exact tests. The LR procedures does not reject EAF (without assuming HWP, $P = 0.4698$), neither rejects EIC (B vs. C, $P = 0.3513$), and finally neither rejects HWP (B vs. A, $P = 0.5555$). This polymorphism corresponds to scenario A, which is the generally expected scenario. AIC identifies model A as the best fitting model.

SNP rs200455936 is significant in the HWP and in the joint exact tests, but not in an exact test for EAF. With the LR approach, EAF could not be rejected (C vs. F, $P = 0.5079$), a common inbreeding coefficient could neither be rejected (B vs. C, $P = 0.2054$), but HWP are rejected (A vs. B, $P < 0.0001$). Correspondingly, model B has the lowest AIC.

SNP rs809600 is not significant in an exact test for HWP, but is significant in an exact test for EAF, and consequently also significant in the joint test. The LR procedure rejects EAF (C vs. F, $P = 0.0005$). Despite differences in allele frequencies, EIC ($P = 0.9733$) and HWP ($P = 0.6690$) are not rejected, and the marker is best described by scenario D. Model D also clearly has the lowest AIC.

SNP rs536079471 is significant in all exact tests. With an LR approach, EAF is rejected (C vs. F, $P = 0.0074$), but a common inbreeding coefficient is not rejected (E vs. F, $P = 0.6433$). Finally, HWP are rejected (D vs. E, $P < 0.0001$). The marker is best described by scenario E.

SNP rs536987805 is significant in all exact tests. Using the LR approach, EAF is rejected (C vs. F, $P = 0.0014$), and EICs too ($P = 0.0002$). Not surprisingly, HWP are rejected too (D vs. F, $P = 0.0002$). If, like in this case, EICs cannot be assumed, then implicitly HWP are rejected for this would imply both coefficients to be equal (and zero). If, at any rate, an additional test for HWP is desired, then it seems more appropriate to test D against F and not D against E. The ternary diagram in Figure 5 F shows that

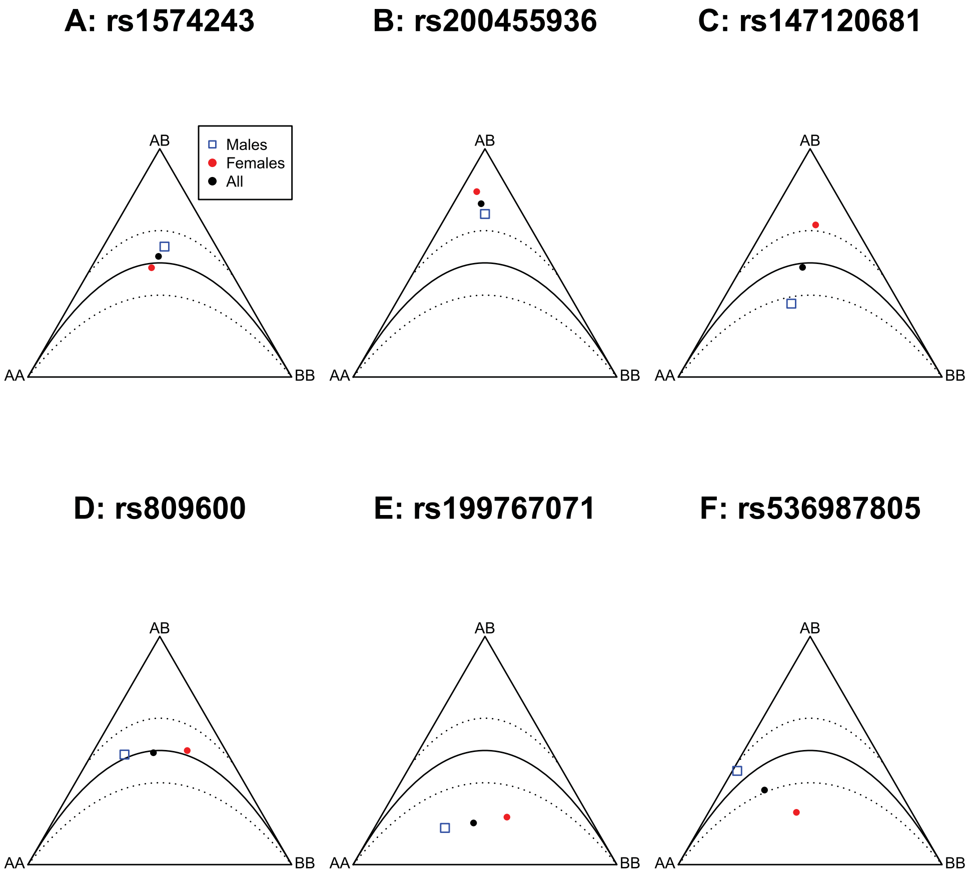

FIGURE 5 Ternary diagrams for six SNPs on chromosome 1 of the JPT sample
Notes: Dotted curves delimit the acceptance region of a chi-square test with a sample size of $n = 52$ and $\alpha = 0.05$

disequilibrium is principally due to females being out of HWP.

The six SNPs studied in Table 2 illustrate that all scenarios theoretically envisioned in Figure 2 do actually occur in practice. The question arises which scenarios are more common, and which are improbable. This is addressed by studying larger genomic areas in the next section.

5.2 | Genomic Areas

We analyzed all 965.458 complete nonmonomorphic SNPs with RS identifier on chromosome 1 of the JPT sample with the HWP, EAF, and joint exact tests described in this paper. The degree of (dis)agreement of the exact test procedures is shown in the Venn diagram in Figure 6. This shows that there are more significant markers in the HWP test (0.65%) than in the EAF tests (0.02%). The percentage of significant markers for both tests is larger than what is expected by chance alone, if the variants are assumed to be independent and using the HapMap significance threshold, $\alpha = 0.001$. The joint exact test uncovers 0.05% of the markers as significant that were not significant when tested separately for HWP and EAF. Not surprisingly, this subset of markers almost exclusively pertains to scenarios F, E, and C, being F the most frequent. Most of them have considerable, but statistically nonsignificant, differences in inbreeding coefficients and allele frequencies between the sexes. In the joint test, which considers both differences, such variants then appear significant.

Of the small subset (six variants) significant in all exact tests, several map to the same area and most likely correspond to the same haplotype. All of these variants had a deficiency of heterozygotes. A considerable set of variants does not appear as significant in the joint exact test, but is significant in a HWP or EAF test only. The first group mainly concerns variants corresponding to scenario B, whereas the second group mainly corresponds to variants with scenario D.

We calculated the AIC for each SNP on chromosome 1, excluding SNPs with missing data and monomorphic in at least one of the two sexes, and assigned each SNP to the scenario for which it had minimal AIC. Figure 7 A shows the prevalence of the different scenarios according to the AIC, and reveals that 70% of the SNPs are classified as having the expected scenario of HWP and EAF. A considerable part, 30%, has a different scenario, being scenario D with HWP and different allele frequencies the second most prevalent (11.3%). We stratified prevalence by the sign of the inbreeding

TABLE 2 (A) Genotype counts and exact test *P* values for six SNPs of the JPT sample. (B) *P* values of the most relevant LR tests for comparing scenarios. (C) Akaike's information criterion (AIC) for the six scenarios for six SNPs of the JPT sample. Models are labeled in accordance with Figure 2. Best fitting models are marked in bold

	(A) Genotype Counts and Exact Test *P* Values								
	Males			**Females**			**Exact *P* Values**		
SNP	**AA**	**AB**	**BB**	**AA**	**AB**	**BB**	**HWP**	**EAF**	**JOINT**
rs1574243	11	32	13	14	23	11	0.6947	0.4904	0.6553
rs200455936	8	40	8	6	39	3	0.0000	0.6782	0.0000
rs147120681	23	18	15	7	32	9	0.6981	0.2107	0.0031
rs809600	22	27	7	7	24	17	0.8461	0.0008	0.0082
rs199767071	32	9	15	15	10	23	0.0000	0.0008	0.0000
rs536987805	32	23	1	21	11	16	0.0130	0.0007	0.0000

	(B) *P* Values of the Most Relevant LR Tests						
	JOINT	**EAF**	**HWP**			**EIC**	
	A-F	**C-F**	**D-F**	**B-C**	**A-B**	**E-F**	**D-E**
rs1574243	0.6284	0.4698	0.5380	0.3513	0.5555	0.3530	0.5391
rs200455936	0.0000	0.5079	0.0000	0.2054	0.0000	0.2194	0.0000
rs147120681	0.0029	0.1801	0.0022	0.0005	0.7192	0.0005	0.7595
rs809600	0.0073	0.0005	0.9122	0.9633	0.8627	0.9733	0.6690
rs199767071	0.0000	0.0074	0.0000	0.6109	0.0000	0.6433	0.0000
rs536987805	0.0000	0.0014	0.0002	0.0006	0.0096	0.0002	0.0867

	C: AIC of Each Model					
SNP	**A**	**B**	**C**	**D**	**E**	**F**
rs1574243	**214.08**	215.74	216.87	215.58	217.21	218.35
rs200455936	180.66	**153.01**	153.41	182.46	154.48	154.97
rs147120681	220.34	222.21	**212.14**	220.57	222.48	212.35
rs809600	219.17	221.14	223.14	**209.32**	211.13	213.13
rs199767071	262.45	219.77	221.52	252.84	**214.56**	216.35
rs536987805	217.77	213.06	203.27	207.84	206.90	**195.09**

coefficient, and this shows that variants of all scenarios except C mostly have negative inbreeding coefficients. We stratified the variants assigned to each scenario by MAF (Fig. 7 B, MAF ≤ 0.05 or > 0.05), overall HW exact test *P*-value (Fig. 7 C, p-value ≤ 0.05 or > 0.05) and EAF exact test *P*-value (Fig. 7 D; *P*-value ≤ 0.05 or > 0.05). These figures show that low MAF markers are more common among variants with homogeneous allele frequencies, and relatively more common, as expected, among variants classified in the equilibrium scenarios (A and D). For low MAF markers, there is less power to detect Hardy-Weinberg disequilibrium (HWD), and therefore they prevail in the equilibrium scenarios. The largest portions of markers with significant HWD are found in scenarios B and E, suggesting that most HWD is due to variants with a common inbreeding coefficient for males and females. Figure 7 D confirms, not surprisingly, that significant deviation from EAF is observed only in the variants assigned to scenarios D, E, and F. The analysis presented in Figure 7 was repeated for chromosome 2, and very similar barplots were obtained (results not shown).

6 | DISCUSSION

It is very well-known that an autosomal marker is expected to reach HWP in one single generation of random mating to the point that most genetic textbooks state this. We stress that this is contingent upon EAF in the sexes, and if these are different, then it does not take one but *two generations* to reach Hardy-Weinberg equilibrium. In the first generation, the new *A* allele frequency is the average of the male and female allele frequencies of the previous generation, but genotype frequencies are *not* in HWP. The allele frequency in the second generation will now remain unaltered, and this generation will have its genotype frequencies in the HWP. Supposing that the other usual assumptions (absence of migration, mutation, selection, etc.) are met, it may thus be more adequate to state that it will take *at most two generations* to reach Hardy-Weinberg equilibrium.

Standard statistical procedures for testing HWP and EAF (chi-square tests, exact tests) do, in theory, not allow us to adequately test these phenomena because of a mutual dependence

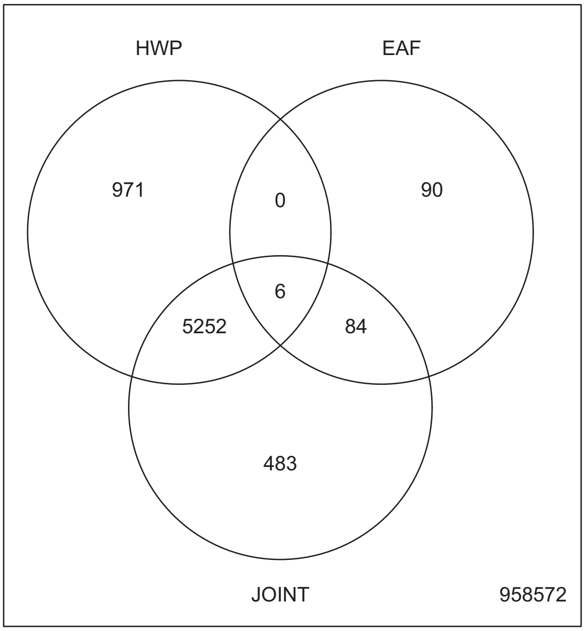

FIGURE 6 Venn diagrams of HWP, EAF, and joint exact test results for all nonmonomorphic complete SNPs on chromosome 1 of the JPT sample

Notes: Circles enclose the number of significant SNPs (at $\alpha = 0.001$) for the different tests.

in their assumptions. An exact test that can test HWP and EAF jointly has been developed, and it has been shown that this test can uncover additional potentially problematic markers (e.g., see rs147120681 in the previous section). The EAF assumption may be avoided by testing HWP in males and females separately, but that brings about an unnecessary reduction in the sample size with a corresponding loss of power.

The LR approach creates a whole family of tests that can compare many nested scenarios. In order to avoid doing all tests, we adopted the following strategy. First scenario C is compared with scenario F. This is a test for EAF without any assumptions regarding HWP. If this test is significant, EAF is rejected and one can proceed to test E against F. If the latter turns out significant, E is rejected and F assumed. If not, D can be tested against E to finally decide upon the scenario. If homogeneity of allele frequencies (C vs. F) cannot be rejected, then, in a similar manner, B might be compared with C, eventually followed by A against B. This inevitably brings about multiple statistical tests, and some correction for multiple testing may be considered in the process. The situation is akin to model building in general (e.g., regression modeling), where different models are used successively, and multiple tests for significance or goodness-of-fit are carried out before one or some final models are selected. The LR approach relies on the asymptotic χ^2 distribution of the LR statistic, and therefore requires large samples. The Type I error rate calculations shows that the joint LR test can be too conservative or too liberal, depending on the MAF of the marker, and that exact procedures are more adequate. Additional exact procedures could be further developed in order to cover all possible scenarios outlined in this paper.

The examples given in Section 5 show that spurious significant and spurious nonsignificant results can arise if the standard exact HW test is applied without stratifying for sex. The example in Figure 5 C suggests the overall HWP test is spuriously nonsignificant due to the fact that male and female inbreeding coefficients have a different sign, and therefore tend to average out when sex is ignored. In fact, the marker deserves close inspection for having highly unexpected opposite signs for male and female inbreeding coefficients. Under scenario D, a spuriously significant overall HWP test result can arise, in particular if the minor allele is a different allele for each sex. When the sexes are analyzed separately, their proportions can correspond to HWP, whereas if they are analyzed jointly by a standard exact test, disequilibrium can be found. This is reminiscent of the well-known Wahlund effect, where reduced heterozygosity is found due to population substructure. It is well known that stratified populations with different subgroups hamper statistical inference on HWP (Laird & Lange, 2011) as well as inference on disease association. In our context, an apparently reduced overall heterozygosity can be found due to allele frequency differences between the sexes.

For the six example SNPs discussed, the final model chosen for each SNP by means of successive hypothesis testing coincided with the model suggested by their AIC. It should be noticed that in practice this is not always the case, in particular if there are only small differences in AIC for two models. At the 5% significance level, we found that the two procedures select the same scenario for 81% of the studied complete nonmonomorphic variants. For the remaining 19%, AIC selected generally more complex models, mostly scenarios B and D instead of A. The AIC approach allows the comparison of all models, whereas the LR approach can only compare nested models. The AIC approach is computationally more demanding because all models are estimated, including the ones for which we have no closed-form estimators (models C and E). For a discussion on using a hypothesis testing or an information-theoretic approach (AIC) in model selection, we refer to Burnham & Anderson (2002) and Murtaugh (2014). Alternatively, Bayesian model selection procedures could also be used in this context.

Variants that were assigned to scenario F often had their male and female genotype compositions lining up almost perpendicularly with respect to the AA or BB angle bisector, the sexes thus having almost identical frequencies for one of the two homozygotes. This suggests confounding of the heterozygote with only one of the homozygotes, though it remains unclear why such confounding appears related to gender.

We emphasize that the analysis in Section 5 refers to complete variants (genotypes observed for all individuals) with a

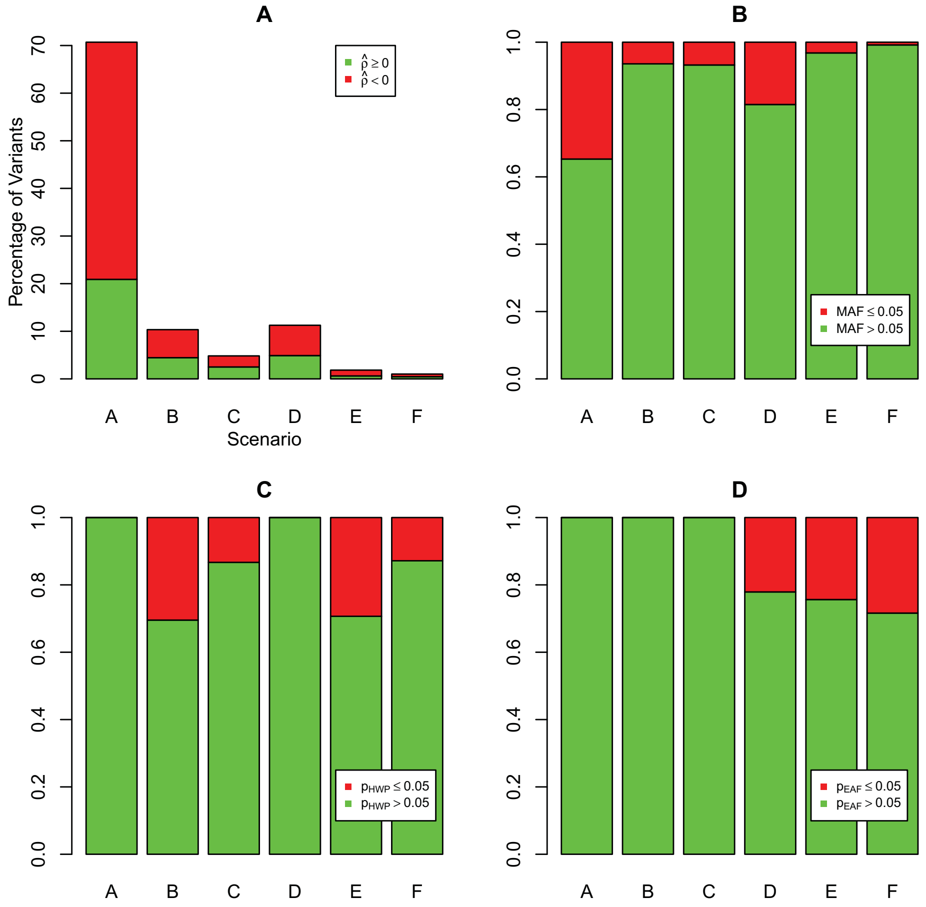

FIGURE 7 Barplots of scenarios based on AIC for all complete SNPs, polymorphic in both sexes, on chromosome 1 of the JPT population
Notes: (A) Prevalence of scenarios stratified by sign of the inbreeding coefficient. (B) Variants in each scenario stratified by MAF. (C) Variants in each scenario stratified by HWD. (D) Variants in each scenario stratified by EAF.

RS identifier, and that this subset should not be considered as representative for the studied chromosome. In particular, variants with missing data typically present more disequilibrium (Graffelman, Nelson, Gogarten, & Weir, 2015).

Of all theoretically possible models, A is the generally expected scenario, as EAF will be reached in one generation, and HWP in at most two, if we admit the initial allele frequencies to differ between the sexes. EAF is thus established *prior to* Hardy-Weinberg equilibrium. The JPT data confirm this since the equal allele frequency models A, B, and C are all more prevalent than their corresponding heterogeneous allele frequency counterparts (see Fig. 7 A), and also there is much more evidence for deviation from HWP than for differences in allele frequencies (see Fig. 6), despite the fact that latter can be expected to have better power because the sample size is doubled ($2n$ alleles instead of n individuals). If systematic deviation from HWP does exist, then we expect males and females to be equally affected. The JPT data confirm this too, with B and E the second most plausible scenarios given the hetero- or homogeneity of allele frequencies.

The power study in Section 4.2 shows that the proposed joint tests have relatively good power under all scenarios. It should, however, be kept in mind that the joint tests address a composite, joint null hypothesis, which is different from the null addressed in a standard HWP and a standard EAF test. If EAF strictly holds, the standard HWP exact test has better power to detect HWD, but only if the deviation from equilibrium is in the same direction in both sexes. Deviations with different signs for the sexes are better detected by the joint procedures.

This paper shows that carrying out the Hardy-Weinberg quality control part in an automated numerical way is not without problems. In this context, the ternary diagram, stratified for males and females, is an excellent graphical tool that contributes to a better understanding of a genetic marker. It is not feasible to inspect all ternary diagrams in a genome-wide association study (GWAS), but it may be feasible to calculate all AICs in order to filter out and inspect those SNPs not corresponding to the (most) expected scenario(s). We do not recommend automated elimination of SNPs with unlikely scenarios from GWASs, but we do encourage a thorough inspection of significant GWAS findings with the tools described in this paper.

7 | SOFTWARE

Operational versions of the joint exact test and LR procedures discussed in this paper are made available for the R environment (R Core Team, 2014) in version 1.5.9 of the R-package Hardy-Weinberg (Graffelman, 2015), and are in the process of being optimized for their use in genome-wide studies.

ACKNOWLEDGMENTS

This work was partially supported by grant 2014SGR551 from the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya, by grant MTM2015-65016-C2-2-R (MINECO/FEDER) of the Spanish Ministry of Economy and Competitiveness and European Regional Development Fund, and by grant GM075091 from the United States National Institutes of Health. We thank the editor and the reviewers, whose comments have helped us to improve the article.

The authors declare that there are no conflicts of interest.

ORCID

Jan Graffelman http://orcid.org/0000-0003-3900-0780

REFERENCES

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd ed.). New York: Springer-Verlag.

Cannings, C., & Edwards, A. W. F. (1968). Natural selection and the de Finetti diagram. *Annals of Human Genetics*, *31*(4), 421–428.

Elston, R. C., & Forthofer, R. (1977). Testing for Hardy-Weinberg equilibrium in small samples. *Biometrics*, *33*(3), 536–542.

Ghalanos, A., & Theussl, S. (2015). *Rsolnp: General non-linear optimization using augmented lagrange multiplier method*. R package version 1.16. Retrieved from http://cran.r-project.org/package=Rsolnp

Gomes, I., Collins, A., Lonjou, C., Thomas, N., Wilkinson, J., Watson, J., & Morton, N. (1999). Hardy-Weinberg quality control. *Annals of Human Genetics*, *63*, 535–538.

Graffelman, J. (2015). Exploring diallelelic genetic markers: The Hardy-Weinberg package. *Journal of Statistical Software*, *64*(3), 1–23. Retrieved from http://www.jstatsoft.org/v64/i03/

Graffelman, J., & Morales-Camarena, J. (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Human Heredity*, *65*(2), 77–84.

Graffelman, J., & Moreno, V. (2013). The mid *p*-value in exact tests for Hardy-Weinberg equilibrium. *Statistical Applications in Genetics and Molecular Biology*, *12*(4), 433–448.

Graffelman, J., Nelson, S. C., Gogarten, S. M., & Weir, B. S. (2015). Exact inference for Hardy-Weinberg proportions with missing genotypes: Single and multiple imputation. *G3 (Genes, Genomes, Genetics)*, *5*(11), 2365–2373. https://doi.org/10.1534/g3.115.022111.

Graffelman, J., & Weir, B. S. (2016). Testing for Hardy-Weinberg equilibrium at bi-allelic genetic markers on the X chromosome. *Heredity*, *116*(6), 558–568.

Haldane, J. B. S. (1954). An exact test for randomness of mating. *Journal of Genetics*, *52*(1), 631–635.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., … Xu, C. (2004). Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics*, *12*(5), 395–399.

Laird, N. M., & Lange, C. (2011). *The fundamentals of modern statistical genetics*. New York: Springer.

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., … Weir, B. S. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, *34*(6), 591–602.

Leal, S. M. (2005). Detection of genotyping errors and pseudo-snps via deviations from Hardy-Weinberg equilibrium. *Genetic Epidemiology*, *29*, 204–214.

Levene, H. (1949). On a matching problem arising in genetics. *Annals of Mathematical Statistics*, *20*(1), 91–94.

Li, C. C. (1976). *The first course in population genetics*. California: The Boxwood Press.

Murtaugh, P. A. (2014). In defense of p values. *Ecology*, *95*(3), 611–617.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Stevens, W. L. (1938). Estimation of blood-group gene frequencies. *Annals of Eugenics*, *8*, 362–375.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74.

Weir, B. S. (1996). *Genetic data analysis II*. Sunderland: MA: Sinauer Associates.

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, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, *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*(12), e0145032. https://doi.org/10.1371/journal.pone.0145032

Zheng, G., Joo, J., Zhang, C., & Geller, N. L. (2007). Testing association for markers on the X chromosome. *Genetic Epidemiology*, *31*, 834–843. https://doi.org/10.1002/gepi.20244

How to cite this article: Graffelman J, Weir B. On the testing of Hardy-Weinberg proportions and equality of allele frequencies in males and females at biallelic genetic markers. *Genet Epidemiol*. 2017:1–15. https://doi.org/10.1002/gepi.22079

APPENDIX

OMNIBUS EXACT TEST

In this Appendix, we give a self-contained treatment of exact tests for HWP and EAF, thereby deriving the density given in Equation (1) of this paper. We also derive the Graffelman-Weir X chromosomal exact test for HWP and EAF, and summarize the classical autosomal exact test.

Classical Autosomal Exact Test for HWP

Autosomal exact inference is based on the conditional distribution of the number of heterozygotes given the number of A alleles. This distribution is generally ascribed to Levene (1949) and Haldane (1954), but as Wellek (2004) has pointed out, had in fact been posed earlier, without derivation, by Stevens (1938). Haldane (1954) derived the distribution for two alleles by a combinatorial argument. Levene (1949) derived it from the multinomial distribution and using conditioning, and obtained the density for multiple alleles. Under the assumption of HWP, the genotypes counts will follow the multinomial distribution with probability vector $(p^2, 2pq, q^2)$ given by:

$$\begin{aligned} P\left(N_{AA}, N_{AB}, N_{BB}\right) \\ = \binom{n}{n_{AA}, n_{AB}, n_{BB}} \left(p^2\right)^{n_{AA}} (2pq)^{n_{AB}} \left(q^2\right)^{n_{BB}}. \quad \text{(A1)} \end{aligned}$$

Under HWP, all alleles are independent, and the distribution of N_A is given by the binomial distribution:

$$P\left(N_A\right) = \binom{2n}{n_A, n_B} (p)^{n_A} (q)^{n_B}. \quad \text{(A2)}$$

The Stevens-Levene-Haldane distribution is then obtained by:

$$\begin{aligned} & P\left(N_{AB} | N_A\right) \\ & = \frac{P\left(N_{AB} = n_{AB} \cap N_A = n_A\right)}{P\left(N_A = n_A\right)} \\ & = \frac{P\left(N_{AA} = n_{AA} \cap N_{AB} = n_{AB} \cap N_{BB} = n_{BB}\right)}{P\left(N_A = n_A\right)} \\ & = \frac{n_A! n_B! n! 2^{n_{AB}}}{n_{AA}! n_{AB}! n_{BB}! (2n)!}, \quad \text{(A3)} \end{aligned}$$

where the first step follows because $N_{AB} = n_{AB}$ and $N_A = n_A$ imply that $N_{AA} = n_{AA}$, and this in turn implies $N_{BB} = n_{BB}$ because the total sample size is fixed. Fast recursive algorithms for the calculation of (A3) have been developed (Elston & Forthofer, 1977; Wigginton, Cutler, & Abecasis, 2005).

Omnibus X Chromosomal Exact Test for HWP and EAF

Graffelman and Weir (2016) proposed an X chromosomal exact test that takes males into account. For inference with X chromosomal markers, an additional random variable needs to be considered, the number of males carrying the minor allele (M_A). We use the joint distribution of the number of female heterozygotes F_{AB} and M_A, given the total minor allele count N_A and given the number of males observed in the sample. This joint distribution can be factored as:

$$\begin{aligned} & P\left(M_A, F_{AB} \mid n, n_A, n_m\right) \\ & = P\left(F_{AB} \mid M_A, n, n_A, n_m\right) \\ & \times P\left(M_A \mid n, n_A, n_m\right). \quad \text{(A4)} \end{aligned}$$

We note that the conditional probability $P(F_{AB} = f_{AB} \mid M_A, n, n_A, n_m)$ is the same as $P(F_{AB} = f_{AB} \mid F_A, n, n_A, n_f)$ because for a fixed total number of A alleles, conditioning on M_A implies conditioning on F_A since their sum is constant. Because we also condition on the sample size and the observed number of males, the conditioning on n_m is equivalent to conditioning on n_f. We thus have

$$\begin{aligned} & P\left(F_{AB} = f_{AB} \mid F_A, n, n_A, n_f\right) \\ & = \frac{f_A! f_B! n_f! 2^{f_{AB}}}{f_{AA}! f_{AB}! f_{BB}! (2n_f)!}. \quad \text{(A5)} \end{aligned}$$

Equation (A5) is in fact the Stevens-Levene-Haldane distribution for the number of heterozygotes described above in Equation (A3), but applied to the females only. We note that the number of M_A males in a sample of n individuals with n_A alleles that is partitioned into n_m males and n_f females has a hypergeometric distribution given by:

$$P\left(M_A = m_a \mid n, n_A, n_m\right) = \frac{n_A! n_B! n_m! (2n_f)!}{f_A! f_B! m_A! m_B! n_t!}. \quad \text{(A6)}$$

Finally, multiplying (A5) by (A6) we obtain:

$$P\left(M_A, F_{AB} \mid n, n_A, n_m\right) = \frac{n_A! n_B! n_m! n_f! 2^{f_{AB}}}{m_A! m_B! f_{AA}! f_{AB}! f_{BB}! n_t!}, \text{(A7)}$$

which is the hitherto unpublished justification of the result given by Graffelman and Weir (2016).

Omnibus Autosomal Exact Test for HWP and EAF

Under the assumption of HWP and EAF, and given a fixed number of males and females, the genotypes of the two sexes can be described by two separate multinomial distributions that both have probability vector $(p_A^2, 2p_A p_B, p_B^2)$. The joint probability of all six genotypes is given by the product of the two multinomial densities:

$$P\left(M_{AA}, M_{AB}, \dots, F_{BB}\right) = \binom{n_m}{m_{AA}, m_{AB}, m_{BB}} (p_A)^{2m_{AA}} (2p_Ap_B)^{m_{AB}} (p_B)^{2m_{BB}} \times \binom{n_f}{f_{AA}, f_{AB}, f_{BB}} (p_A)^{2f_{AA}} (2p_Ap_B)^{f_{AB}} (p_B)^{2f_{BB}} = \frac{n_m!n_f!}{m_{AA}!m_{AB}!m_{BB}!f_{AA}!f_{AB}!f_{BB}!} p_A^{n_A} p_B^{n_B} 2^{f_{AB}+m_{AB}}. \tag{A8}$$

Again assuming HWP, N_A, the number of A alleles, will have the binomial distribution:

$$P\left(N_A = n_A\right) = \binom{2n}{n_A} (p_A)^{n_A} (p_B)^{n_B}. \tag{A9}$$

Again conditioning on the total number of A alleles, we have

$$P\left(M_{AA}, M_{AB}, \dots, F_{BB} | N_A\right) = \frac{P\left(M_{AA} = m_{AA}, \dots, F_{BB} = f_{BB} \cap N_A = n_A\right)}{P\left(N_A = n_A\right)}. \tag{A10}$$

If N_A is known, and homogeneous allele frequencies are assumed, then F_A and M_A are also known. Joint knowledge of F_A with F_{AB} and M_A with M_{AB} and the number of males and females implies all genotype counts. Consequently, the numerator in (A10) is the product of the multinomial distributions given in (A8). Dividing (A8) by (A9) we obtain:

$$P\left(M_{AA}, M_{AB}, \dots, F_{BB} | N_A\right) = \frac{n_A!n_B!n_m!n_f!}{m_{AA}!m_{AB}!m_{BB}!f_{AA}!f_{AB}!f_{BB}!(2n)!} 2^{m_{AB}+f_{AB}},$$

which is Equation (1) of this paper. We note that this result strongly resembles the density used in the omnibus exact test for X chromosomal markers (A7), the difference being that the hemizygous male genotype counts are replaced by the autosomal homozygote counts, m_{AB} appears as an extra genotype, and that the female heterozygotes in the exponent are replaced by the total number of heterozygotes. Indeed, the classical autosomal distribution and the Graffelman-Weir exact test are special cases of Equation (1). By setting all genotype counts of one sex to zero (e.g., $m_{AA} = m_{AB} = m_{BB} = 0$ and $n_m = 0$), the classical autosomal density is obtained. Setting $m_{AB} = 0$ and m_{AA} to m_A and m_{BB} to m_B the X chromosomal exact distribution is obtained.