Heredity (2019) 123:549–564
https://doi.org/10.1038/s41437-019-0232-0

the geneticssociety

ARTICLE

Bayesian model selection for the study of Hardy–Weinberg proportions and homogeneity of gender allele frequencies

Xavier Puig[1] · Josep Ginebra[1] · Jan Graffelman[1,2]

Received: 6 February 2019 / Revised: 15 April 2019 / Accepted: 2 May 2019 / Published online: 29 May 2019

Abstract

Standard statistical tests for Hardy–Weinberg equilibrium assume the equality of allele frequencies in the sexes, whereas tests for the equality of allele frequencies in the sexes assume Hardy–Weinberg equilibrium. This produces a circularity in the testing of genetic variants, which has recently been resolved with new frequentist likelihood and exact procedures. In this paper, we tackle the same problem by posing it as a Bayesian model comparison problem. We formulate an exhaustive set of ten alternative scenarios for biallelic genetic variants. Using Dirichlet and Beta priors for genotype and allele frequencies, we derive marginal likelihoods for all scenarios, and select the most likely scenario using the posterior probabilities that each of these scenarios is the one in place. Different from the usual frequentist testing approach, the Bayesian approach allows one to compare any number of models, and not just two at a time, and the models compared do not have to be nested. We illustrate our Bayesian approach with genetic data from the 1,000 genomes project and through a simulation study.

Introduction

The Hardy–Weinberg law, independently formulated by Hardy (1908) and Weinberg (1908) more than 100 years ago, is a fundamental genetic principle. It is of importance in many areas of genome research. Hardy–Weinberg equilibrium is typically assumed in haplotype estimation (Single et al. 2002). Classical estimation of genetic relatedness between individuals by maximum likelihood methods also rests on the equilibrium assumption (Thompson 1975). In fact, many statistical models and procedures used in genetic epidemiology make the Hardy–Weinberg assumption. Data produced by genotyping arrays undergoe extensive quality control procedures, of which testing for Hardy–Weinberg proportions (HWP) forms an important part (Laurie et al. 2010). It is well-known, and widely stated in genetic textbooks (Hartl 1980; Hamilton 2009), that a biological population will reach HWP in one generation of random mating.

Recently, Graffelman and Weir (2017) have stressed that this is only true under the assumption of equal allele frequencies (EAF) in the sexes. If such equality does not hold, it will take two generations before equilibrium is achieved. Graffelman and Weir (2017) show that the statistical testing of EAF and HWP by chi-square or exact procedures is intricately linked in assumptions, leading to circularity in the statistical testing, because EAF tests assume HWP, whereas HWP tests assume EAF. The authors propose novel exact and likelihood ratio procedures that avoid this dependence in assumptions, making it possible to test EAF and HWP whether independently or simultaneously.

In this paper, we readdress this issue from a Bayesian perspective. Our approach extends previous Bayesian work for the analysis of X-chromosomal variants (Puig et al. 2017). Here, we enumerate ten possible scenarios (models) for the data, choose prior distributions for the parameters of each one of the models and prior probabilities for them, derive marginal likelihoods, and compute posterior probabilities to identify the most probable model, given the observed data. Six of the scenarios (models) considered here coincide with the ones considered in Graffelman and Weir (2017); they include the scenario with both HWP and EAF in place, together with five scenarios where either

✉ Jan Graffelman
jan.graffelman@upc.edu

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

[2] Department of Biostatistics, University of Washington, Seattle, WA, USA

HWP or EAF or both restrictions fail. The four scenarios considered here for the first time are included to admit the possibility of having one sex in HWP and the other one not in HWP.

Different from the approach based on the usual frequentist testing, which compares only two scenarios at a time with one of the scenarios typically nested into the other one, the Bayesian approach allows one to compare any number of models which can be nested or not nested into each other. Different from the approach based on the use of heuristic model selection criteria, under the Bayesian approach, models are assessed based on their posterior probabilities, adding up to one, which helps to assess the degree of the uncertainty behind the model choice. Prior distributions are chosen in a way such that posterior probabilities can be computed either exactly or through numerical integration, hence avoiding the need to estimate them through more intensive computational methods.

The structure of the paper is as follows. In the section Theory, we develop Bayesian theory for this particular genetic context. In the section Examples, we illustrate the use of the Bayesian approach with data taken from the Japanese population of the 1,000 Genomes project (The 1,000 Genomes Project Consortium et al. 2010). A Discussion section completes the paper.

Theory

Here, we describe the Bayesian framework that enables one to select the most credible scenario among ten alternative scenarios, including the one in which both HWP as well as EAF are in place. In the following, the subsection Notation presents our basic definitions and the subsection Scenarios and priors gives the probabilistic definition of each scenario, including both the corresponding statistical models as well as the prior distributions on their parameters. Subsection Bayesian model selection addresses the way Bayesian model selection works, and finally the subsection Simulation evaluation explores the performance of the model selection procedure through a simulation study.

Notation

We consider a biallelic genetic polymorphism with alleles A and B having population allele frequencies p_{Af} and p_{Bf} in females and p_{Am} and p_{Bm} in males, with $p_{Af}+p_{Bf}=p_{Am}+p_{Bm}=1$. We denote the observed A and B allele counts in females as n_{Af} and n_{Bf}, and in males as n_{Am} and n_{Bm}, and their totals by $n_A=n_{Af}+n_{Am}$ and $n_B=n_{Bf}+n_{Bm}$. When the population has equal allele frequencies for both sexes, denoted by EAF, their ratio, d, is given by

$$d=\frac{p_{Am}}{p_{Af}}=1, \tag{1}$$

and d will be used as a measure of the discrepancy of male and female allele frequencies.

Let (p_{AAf}, p_{ABf}, p_{BBf}), with $p_{AAf}+p_{ABf}+p_{BBf}=1$, be the female genotype frequencies, and let (p_{AAm}, p_{ABm}, p_{BBm}), with $p_{AAm}+p_{ABm}+p_{BBm}=1$, be the male genotype frequencies in the population. We denote the observed genotype counts in females by (n_{AAf}, n_{ABf}, n_{BBf}), and in males by (n_{AAm}, n_{ABm}, n_{BBm}). The total genotype counts are given by the sum of the latter vectors, and indicated by (n_{AA}, n_{AB}, n_{BB}), without the index for sex. The total sample size is $n=n_m+n_f$, where $n_f=n_{AAf}+n_{ABf}+n_{BBf}$ is the total number of females, and $n_m=n_{AAm}+n_{ABm}+n_{BBm}$ is the total number of males. The total allele counts can be obtained as $n_A=n_{Af}+n_{Am}=2n_{AAf}+n_{ABf}+2n_{AAm}+n_{ABm}$ and $n_B=2n_{BBf}+n_{ABf}+2n_{BBm}+n_{ABm}$. One considers the population of females to be in a Hardy–Weinberg equilibrium when their genotype frequencies are such that

$$\rho_f=\frac{p_{AAf}-p_{Af}^2}{p_{Af}(1-p_{Af})}=0, \tag{2}$$

and the population of males to be so when

$$\rho_m=\frac{p_{AAm}-p_{Am}^2}{p_{Am}(1-p_{Am})}=0, \tag{3}$$

where ρ_f and ρ_m are the inbreeding coefficients for males and females, which can be used as measures of the deviation of female and of male genotype frequencies from HWP. The term “inbreeding coefficient” might be regarded as a misnomer because disequilibrium might arise from genotyping error or by chance, instead of from inbreeding, and because here we have a different coefficient for each sex while the two sexes intervene when breeding. Nevertheless, we still use the term “inbreeding coefficient” for historical reasons and because of its widespread use in population genetics.

The range of ρ_f is the interval $[-p_{min}/(1-p_{min}),\ 1]$, where $p_{min}=\min(p_{Af},\ 1-p_{Af})$; in particular, when p_{Af} is 0 or 1, the range for ρ_f is [0, 1], and its range will only be [−1, 1] when p_{Af} is 0, 5. The same applies to the range of ρ_m after replacing p_{Af} by p_{Am}.

When $\rho_m=\rho_f$, we will state that males and females have equal inbreeding coefficients, and it will be denoted by EIC. It is easy to check that the relationship between male genotype and allele frequencies in the population is such that

$$p_{Am}=\frac{2p_{AAm}+p_{ABm}}{2}, \tag{4}$$

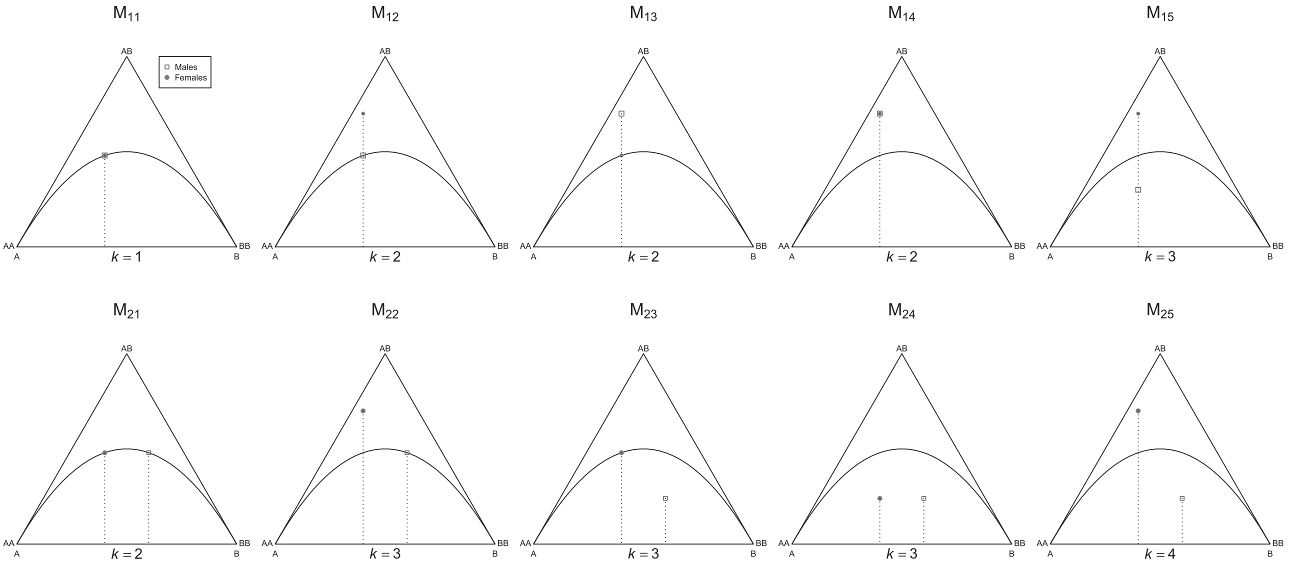

Fig. 1 Ternary diagrams with male and female population genotype frequencies for ten different scenarios. Top row: scenarios with equal allele frequencies in the sexes. Bottom row: scenarios with different allele frequencies for both sexes. Population allele frequencies shown by vertical projections onto the base of the diagram. Symbols on the parabola indicate that the corresponding sexes are in HWP. The dimension of the parameter space for each scenario (k) is given at the bottom of each diagram

$$p_{AAm} = p_{Am}^2 + p_{Am}p_{Bm}\rho_m, \tag{5}$$

$$p_{ABm} = 2p_{Am}p_{Bm}(1 - \rho_m), \tag{6}$$

$$p_{BBm} = p_{Bm}^2 + p_{Am}p_{Bm}\rho_m, \tag{7}$$

and similarly between female genotype and allele frequencies. The distribution of the vector of female genotype counts (n_{AAf}, n_{ABf}, n_{BBf}), is assumed to be:

$$\text{Multinomial}(n_f, (p_{AAf}, p_{ABf}, p_{BBf})), \tag{8}$$

and the one for male genotype counts (n_{AAm}, n_{ABm}, n_{BBm}) to be:

$$\text{Multinomial}(n_m, (p_{AAm}, p_{ABm}, p_{BBm})). \tag{9}$$

The precise situation of a biallelic genetic variant with respect to the HWP and EAF hypotheses can be efficiently visualized by plotting the genotype probabilities in a ternary diagram, also known as a de Finetti diagram (de Finetti 1926). A sample can be represented by a single point in the diagram that uniquely defines its genotype and allele frequencies. The base of the diagram is a 0–1 axis for the allele frequencies. Genotype frequencies are represented by the relative length of the three line segments obtained by perpendicular projection of the sample onto the edges of the diagram. Allele frequencies can be read off by projection onto the triangle base. The two vertices at the base of the ternary diagram therefore have a double interpretation: they represent the homozygote genotypes (AA and BB) as well as the two alleles, A and B. Genotype compositions in HWP must satisfy $p_{AB}^2 = (2p_Ap_B)^2 = 4p_{AA}p_{BB}$ and are therefore constrained to be on a parabola in the triangle. EAF in the sexes is indicated by male and female compositions that line up vertically, perpendicularly with respect to the triangle base. For more background on the use of the ternary diagram for genetic data, we refer to Cannings and Edwards (1968), Graffelman and Morales-Camarena (2008), and Graffelman and Weir (2016) for X-chromosomal variants.

The ten ternary diagrams presented in Fig. 1 distinguish ten different possibilities that we consider, which we call *scenarios* or *models*. The diagrams in the first row of Fig. 1 refer to a population with equal allele frequencies for both sexes, whereas the second row in that figure corresponds to populations with heterogeneous gender allele frequencies. The diagrams in the first column in Fig. 1 refer to the two scenarios in which genotype frequencies satisfy the HWP. Scenarios with deviations from Hardy–Weinberg proportions, shown in columns 2–5 in Fig. 1, all have points which are off the HW parabola. When the inbreeding coefficients are positive, indicating a lack of heterozygotes, the points fall below the parabola. When the inbreeding coefficients are negative, indicating an excess of heterozygotes, they will fall above the parabola.

The statistical models that correspond to each one of these ten scenarios, together with the prior distribution assumed for their parameters, are described next in a systematic way.

Scenarios and priors

We label the statistical models behind scenarios with a double subindex, M_{ij}, setting the first subindex i to 1 if the EAF hypothesis holds and to 2 otherwise. We use the second subindex j for the HWP hypothesis, setting it to 1 when the HWP hypothesis holds for both males and females, to 2 when it holds for males but not for females, to 3 when it holds for females but not for males, to 4 when HWP neither hold for males nor for females but their inbreeding coefficients are equal, and to 5 when HWP neither hold for males nor for females and their inbreeding coefficients are different.

Scenarios M_{11}, M_{21}, M_{14}, M_{24}, M_{15}, and M_{25} are the ones considered by Graffelman and Weir (2017). Here, we provide some more detail by explicitly admitting the possibility of having one sex in HWP and the other not, which correspond to the scenarios with models M_{12}, M_{22}, M_{13}, and M_{23}.

Scenario M_{11}: EAF and HWP in both sexes

If there are no disturbing factors operating in the population, one expects EAF to hold together with HWP for both males as well as females. When these three conditions hold, that is, when $d = 1$, together with $\rho_m = \rho_f = 0$, then all male and female genotype frequencies can be written as a function of p_{Af}, (or equivalently p_{Am}), and

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_A \sim \text{Multinomial}\Big(n_f, (p_A^2, 2p_A(1-p_A), (1-p_A)^2), \tag{10}$$

$$(n_{AAm}, n_{ABm}, n_{BBm})|p_A \sim \text{Multinomial}(n_m,(p_A^2, 2p_A(1-p_A), (1-p_A)^2), \tag{11}$$

where $p_A = p_{Af} = p_{Am}$. Under this scenario, male and female allele frequencies, p_{Am} and p_{Af}, are equal, and they will be assumed to be $\text{Beta}(b_1^{11}, b_2^{11})$ distributed, where the superindex ij in b_k^{ij} denotes the model. This prior distribution on allele frequencies univocally determines the prior distribution of all male and female genotype frequencies.

The full equilibrium in M_{11} can be broken because EAF does not hold, and therefore $d \neq 1$, because HWP do not hold for females, and therefore $\rho_f \neq 0$, because HWP do not hold for males, and therefore $\rho_m \neq 0$, or because of the simultaneous occurrence of any two or of all three of these conditions. These disequilibrium situations are the ones covered by the next nine scenarios.

Scenario M_{21}: HWP in both sexes

With HWP in both sexes, we have $\rho_m = \rho_f = 0$, but here male and female allele frequencies are not the same, and hence $d \neq 1$. In that case, all male and all female genotype frequencies can be posed as functions of p_{Af} and p_{Am}, respectively:

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_{Af} \sim \text{Multinomial}\Big(n_f, (p_{Af}^2, 2p_{Af}(1-p_{Af}), (1-p_{Af})^2)\Big), \tag{12}$$

$$(n_{AAm}, n_{ABm}, n_{BBm})|p_{Am} \sim \text{Multinomial}\Big(n_m, (p_{Am}^2, 2p_{Am}(1-p_{Am}), (1-p_{Am})^2)\Big). \tag{13}$$

The allele frequencies, p_{Af} and p_{Am}, are assumed to be independent with $\text{Beta}(b_{1f}^{21}, b_{2f}^{21})$ and $\text{Beta}(b_{1m}^{21}, b_{2m}^{21})$ prior distributions, respectively where, unless one has different information about p_{Af} and about p_{Am}, one will most likely set $b_{1f}^{21} = b_{1m}^{21}$ and $b_{2f}^{21} = b_{2m}^{21}$.

Scenario M_{12}: EAF and HWP in males only

Under this scenario, one assumes that there is EAF and HWP among males, and hence $d = 1$ and $\rho_m = 0$, but HWP do not hold among females, and hence $\rho_f \neq 0$, and in that case, male and female genotype frequencies can be posed just as a function of p_{AAf} and of p_{ABf}, and

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_{AAf}, p_{ABf} \sim \text{Multinomial}(n_f, (p_{AAf}, p_{ABf}, 1-p_{AAf}-p_{ABf})), \tag{14}$$

$$\begin{aligned}(n_{AAm}, n_{ABm}, n_{BBm})|p_{AAf}, p_{ABf} \sim \text{Multinomial}\Big(n_m, \big((p_{AAf}+\tfrac{1}{2}p_{ABf})^2,\\ 2(p_{AAf}+\tfrac{1}{2}p_{ABf})(1-(p_{AAf}+\tfrac{1}{2}p_{ABf})), (1-(p_{AAf}+\tfrac{1}{2}p_{ABf}))^2\big)\Big).\end{aligned} \tag{15}$$

Under this scenario, the prior distribution used for (p_{AAf}, p_{ABf}, $1-p_{AAf}-p_{ABf}$) will be Dirichlet $(a_{1f}^{12}, a_{2f}^{12}, a_{3f}^{12})$, where the superindex ij in a_{kf}^{ij} denotes the model.

Scenario M_{22}: HWP for males only

This scenario is like M_{12} but without EAF, and genotype frequencies can be written as a function of p_{AAf}, p_{ABf}, and p_{Am}, and are such that

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_{AAf}, p_{ABf} \sim \text{Multinomial}(n_f, (p_{AAf}, p_{ABf}, 1-p_{AAf}-p_{ABf})), \tag{16}$$

$$(n_{AAm}, n_{ABm}, n_{BBm})|p_{Am} \sim \text{Multinomial}(n_m, (p_{Am}^2, 2p_{Am}(1-p_{Am}), (1-p_{Am})^2)). \tag{17}$$

The prior distribution for p_{Am} here will be $\text{Beta}(b_{1m}^{22}, b_{2m}^{22})$ and for (p_{AAf}, p_{ABf}, $1-p_{AAf}-p_{ABf}$) will be Dirichlet $(a_{1f}^{22}, a_{2f}^{22}, a_{3f}^{22})$.

Scenario M$_{13}$: EAF and HWP for females only

This scenario is like M_{12} in that EAF holds, but where HWP holds among females and not among males. In that case, genotype frequencies are a function of just p_{AAm} and of p_{ABm}, and

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_{AAm}, p_{ABm} \sim \text{Multinomial}\Big(n_f, (p_{AAm} + \tfrac{1}{2}p_{ABm})^2, 2(p_{AAm} + \tfrac{1}{2}p_{ABm})(1 - (p_{AAm} + \tfrac{1}{2}p_{ABm})), (1 - (p_{AAm} + \tfrac{1}{2}p_{ABm}))^2\Big), \tag{18}$$

$$(n_{AAm}, n_{ABm}, n_{BBm})|p_{AAm}, p_{ABm} \sim \text{Multinomial}(n_m, (p_{AAm}, p_{ABm}, 1 - p_{AAm} - p_{ABm})). \tag{19}$$

Analogously to M_{12}, we now use the Dirichlet $(a^{13}_{1m}, a^{13}_{2m}, a^{13}_{3m})$ prior on (p_{AAm}, p_{ABm}, $1 - p_{AAm} - p_{ABm}$).

Scenario M$_{23}$: HWP for females only

This scenario is like M_{13} but without EAF, and genotype frequencies are a function of p_{Af}, of p_{AAm}, and of p_{ABm}, and are such that

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_{Af} \sim \text{Multinomial}\Big(n_f, (p^2_{Af}, 2p_{Af}(1 - p_{Af}), (1 - p_{Af})^2)\Big), \tag{20}$$

$$(n_{AAm}, n_{ABm}, n_{BBm})|p_{AAm}, p_{ABm} \sim \text{Multinomial}(n_m, (p_{AAm}, p_{ABm}, 1 - p_{AAm} - p_{ABm})). \tag{21}$$

The prior distribution will be Beta$(b^{23}_{1f}, b^{23}_{2f})$ for p_{Af} and Dirichlet $(a^{23}_{1m}, a^{23}_{2m}, a^{23}_{3m})$ for (p_{AAm}, p_{ABm}, $1 - p_{AAm} - p_{ABm}$).

Finally, in the case where neither males nor females are in HWP, we will distinguish the setting in which male and female inbreeding coefficients are equal, EIC, which will be labeled with a 4 as the second index, from the case in which the two inbreeding coefficients are different, which will be labeled with a 5. That, coupled with the possibility of having or not having EAF, leads to the last four possible scenarios.

Scenario M$_{14}$: EAF and EIC

With this scenario, one has EAF, and neither males nor females are in HWP, but male and female inbreeding coefficients are equal, $\rho_m = \rho_f \neq 0$, in which case male and female genotype frequencies are a function of p_{AAf} and of p_{ABf} and are such that

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_{AAf}, p_{ABf} \sim \text{Multinomial}(n_f, (p_{AAf}, p_{ABf}, 1 - p_{AAf} - p_{ABf})), \tag{22}$$

$$(n_{AAm}, n_{ABm}, n_{BBm})|p_{AAf}, p_{ABf} \sim \text{Multinomial}(n_m, (p_{AAf}, p_{ABf}, 1 - p_{AAf} - p_{ABf})). \tag{23}$$

Under this scenario, the prior distribution for (p_{AAf}, p_{ABf}, $1 - p_{AAf} - p_{ABf}$) will be Dirichlet $(a^{14}_{1f}, a^{14}_{2f}, a^{14}_{3f})$.

Scenario M$_{24}$: EIC only

This scenario is like M_{14} but without EAF, and in that case, genotype frequencies are a function of p_{AAf}, of p_{ABf}, and of p_{Am}. Their distribution is

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_{AAf}, p_{ABf} \sim \text{Multinomial}(n_f, (p_{AAf}, p_{ABf}, 1 - p_{AAf} - p_{ABf})) \tag{24}$$

$$\begin{aligned}(n_{AAm}, n_{ABm}, n_{BBm})|p_{Am}, p_{AAf}, p_{ABf} \sim \text{Multinomial} \\ \Big(n_m, \Big(p^2_{Am} + p_{Am}(1 - p_{Am})\frac{p_{AAf} - (p_{AAf} + \frac{1}{2}p_{ABf})^2}{(p_{AAf} + \frac{1}{2}p_{ABf})(1 - p_{AAf} - \frac{1}{2}p_{ABf})}, \\ 2p_{Am}(1 - p_{Am})(1 - \frac{p_{AAf} - (p_{AAf} + \frac{1}{2}p_{ABf})^2}{(p_{AAf} + \frac{1}{2}p_{ABf})(1 - p_{AAf} - \frac{1}{2}p_{ABf})}), \\ (1 - p_{Am})^2 + p_{Am}(1 - p_{Am})\frac{p_{AAf} - (p_{AAf} + \frac{1}{2}p_{ABf})^2}{(p_{AAf} + \frac{1}{2}p_{ABf})(1 - p_{AAf} - \frac{1}{2}p_{ABf})}\Big)\Big).\end{aligned} \tag{25}$$

The prior distribution will be Dirichlet $(a^{24}_{1f}, a^{24}_{2f}, a^{24}_{3f})$ for (p_{AAf}, p_{ABf}, $1 - p_{AAf} - p_{ABf}$) and it will be Beta$(b^{24}_{1m}, b^{24}_{2m})$ for p_{Am}, but truncated to the set of feasible values for that parameter, the way indicated in Appendix 1.

Scenario M$_{15}$: EAF only

With this scenario, one has EAF, and neither males nor females are in HWP, but different from scenario M_{14}, male and female inbreeding coefficients are assumed to be different. In that case, genotype frequencies are a function of p_{AAf}, p_{ABf}, and p_{AAm}, and

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_{AAf}, p_{ABf} \sim \text{Multinomial}(n_f, (p_{AAf}, p_{ABf}, 1 - p_{AAf} - p_{ABf})), \tag{26}$$

$$\begin{aligned}(n_{AAm}, n_{ABm}, n_{BBm})|p_{AAm}, p_{AAf}, p_{ABf} \sim \text{Multinomial}(n_m, (p_{AAm}, \\ 2p_{AAf} + p_{ABf} - 2p_{AAm}, 1 + p_{AAm} - 2p_{AAf} - p_{ABf})).\end{aligned} \tag{27}$$

Here, the prior for (p_{AAf}, p_{ABf}, $1 - p_{AAf} - p_{ABf}$) will be Dirichlet $(a^{15}_{1f}, a^{15}_{2f}, a^{15}_{3f})$ and the prior for p_{AAm} will be Beta $(b^{15}_{1m}, b^{15}_{2m})$, but truncated to the interval of feasible values for that parameter, the way indicated in Appendix 1.

Scenario M$_{25}$: neither EAF nor HWP nor EIC

Finally, under M_{25}, here, neither the EAF nor the HWP hypotheses for males and females hold, and male and female inbreeding coefficients are different, and we deal with the general unrestricted full four-dimensional parameter space model, with

$$(n_{AAf}, n_{ABf}, n_{BBf})|p_{AAf}, p_{ABf} \sim \text{Multinomial}(n_f, (p_{AAf}, p_{ABf}, 1 - p_{AAf} - p_{ABf})), \tag{28}$$

$$(n_{AAm}, n_{ABm}, n_{BBm})|p_{AAm}, p_{ABm} \sim \text{Multinomial}(n_m, (p_{AAm}, p_{ABm}, 1 - p_{AAm} - p_{ABm})). \tag{29}$$

Under this scenario, the prior for (p_{AAf}, p_{ABf}, $1 - p_{AAf} - p_{ABf}$) is Dirichlet $(a^{25}_{1f}, a^{25}_{2f}, a^{25}_{3f})$ while the prior for (p_{AAm}, p_{ABm}, $1 - p_{AAm} - p_{ABm}$) is Dirichlet $(a^{25}_{1m}, a^{25}_{2m}, a^{25}_{3m})$.

In frequentist inference, this last entirely unrestricted scenario M_{25} is used as the reference (alternative) hypothesis against which all other scenarios are tested. In the Bayesian setting, it becomes just another model, treated on the same level as the other nine.

Depending on the values picked for (a_1, a_2, a_3), the Dirichlet(a_1, a_2, a_3) distribution will be more or less informative, and it will capture different information about male or female genotype frequencies. In particular, its expected value is $(a_1, a_2, a_3)/(\sum a_j)$, and one can choose the a_j's to reflect the fact that one expects some genotypes to have larger probabilities than others. Also, the larger $\sum a_j$, the smaller the variances of the components of the Dirichlet random variable, and the more informative that prior distribution will be. When one is not willing to use subjective information about genotype frequencies, Bernardo and Tomazella (2010) and Berger et al. (2015) recommend using a Dirichlet prior with $a_1 = a_2 = a_3 = 1/3$. We will use this reference prior, which is like assuming that what you know about genotype frequencies is worth as much as what you learn from a sample with n_m or n_f equal to one. Given that the actual sample sizes in our setting will typically be a lot larger than one, the impact of this Dirichlet prior on the posterior distribution for the genotype frequencies will be negligible.

An analogous argument can be made for choosing the parameters of the Beta(b_1, b_2) distribution to model the prior information about allele frequencies in those scenarios where that is needed. In that case, in the absence of subjective information, one often chooses Beta(b_1, b_2) with $b_1 = b_2 = 1/2$.

Note that our choice of prior for the genotype frequencies determines that the prior distributions for ρ_f and for ρ_m will have two modes at the extremes of the range of values that they take, and one mode at 0, which are features that one expects from reference priors for parameters with finite support and a singular point in its interior. Even though the prior probability that each one of these coefficients is larger than 0 is 0.55, and not 0.5, due to the asymmetry of the support for these parameters, the prior is vague enough to avoid having that much impact on the posterior distribution for these coefficients. Our choice of prior for genotype frequencies determines that the prior distributions for allele frequencies will have modes at 0, at 0.5, and at 1.

Different parameterizations for the statistical models allow for different ways of capturing what one knows about the parameters of the model through a prior distribution for them. Adopting the parameterizations used above allows for a choice of priors that leads to simple closed-form expressions for most of the posterior probabilities of the models considered. Alternative ways of choosing prior distributions for testing for HWP under the usual autosomal data, often involving different parameterizations of the statistical models, can be found in Lindley (1988), Shoemaker et al. (1998), Consonni et al. (2008), and Wakefield (2010). All their proposals could be adapted here, but if one chose these priors to have a small effective sample size, choosing them instead of the one we chose would make a small difference at a considerable extra computational cost, because they would require the use of Markov chain Monte Carlo methods to estimate posterior probabilities.

Bayesian model selection

In the frequentist literature, one usually tests for EAF and for HWP separately, assuming that the other hypothesis holds. Graffelman and Weir (2017) propose an omnibus exact test to test both hypotheses jointly against a specific alternative scenario. They also propose a likelihood ratio approach to pick up the scenario that is most parsimonious among all scenarios that cannot be rejected. Alternatively, they also suggest doing model selection among the six models they consider, based on Akaike's information criterion (AIC).

Instead, in the Bayesian setting, one can tackle the same problem by classifying a genetic variant into one scenario among the ten alternative scenarios described above, which is equivalent to selecting one model among ten. This is done by choosing a prior distribution on the parameter space for each model, together with a prior distribution on the model space, and then computing the posterior probability of each one of the ten models. Then, one selects the model that best represents the variant by picking up the model with the largest posterior probability. In that setting, one treats all ten models involved on the same level, without assigning a special role to the full equilibrium model, M_{11}.

The posterior probability of each model, $P(M_{ij}|y)$, which is the probability that M_{ij} is the model generating the data, $y = (n_{AAf}, n_{ABf}, n_{BBf}, n_{AAm}, n_{ABm}, n_{BBm})$, assessed after the data have been observed, can be computed through Bayes theorem:

$$P(M_{ij}|y) = \frac{P(M_{ij})P(y|M_{ij})}{\sum_{i=1}^{2}\sum_{j=1}^{5} P(M_{ij})P(y|M_{ij})}, \tag{30}$$

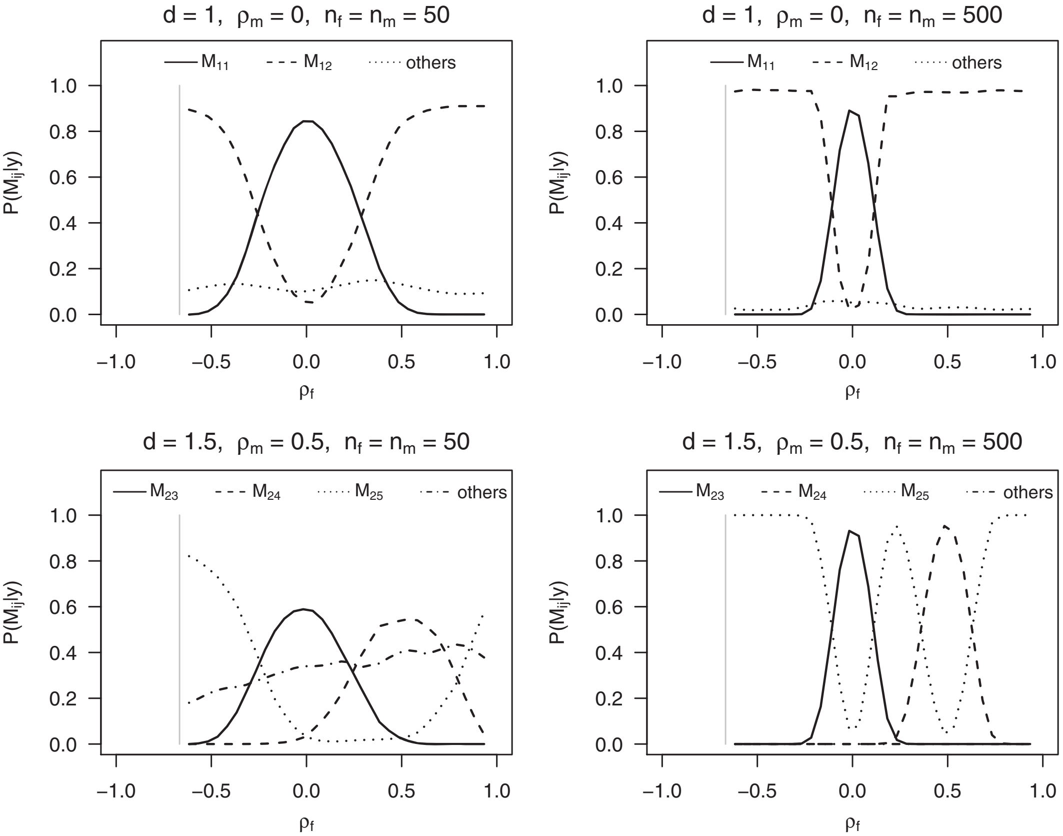

Fig. 2 Expected value of $P(M_{ij}|y)$ for M_{11} and M_{12} for two different sample sizes in the top two panels, and for M_{23}, M_{24}, and M_{25} for the same two sample sizes in the bottom panels, and of the sum of the posterior probabilities of the remaining models, as a function of ρ_f in its whole support, when $p_{Af} = 0.4$

where $P(M_{ij})$ is the prior probability assigned to M_{ij}, and where $P(y|M_{ij})$ is the marginal likelihood of M_{ij}. With everything else staying constant, the larger $P(y|M_{ij})$, the larger $P(M_{ij}|y)$. If all models were considered equally likely a priori, with $P(M_{ij}) = 1/10$, then $P(M_{ij}|y)$ is proportional to $P(y|M_{ij})$.

Most often, computing $P(y|M_{ij})$ exactly is too complicated, and the marginal likelihoods need to be estimated through MCMC simulation. In our multinomial setting with Beta and Dirichlet priors though, there are closed-form expressions for $P(y|M_{ij})$ which allow one to compute these marginal likelihoods exactly in the case of M_{11}, M_{14}, M_{21}, M_{22}, M_{23}, and M_{25}, and to evaluate them through numerical integration in the four remaining cases. The expressions for the marginal likelihoods, $P(y|M_{ij})$, under our choice of prior distribution can be found in Appendix 1.

Simulation evaluation of the performance of the method

To evaluate the performance of this Bayesian model selection procedure and of the inferences that follow from it, here, it is used under a large set of known scenarios through a simulation study. In particular, the method is tried on SNPs from populations where the female inbreeding coefficient, ρ_f, is known and it takes values in its whole range, where ρ_m is either 0 or 0.5, where d is either 1 or 1.5, and where p_{Af} is either 0.2 or 0.4. Sample sizes, n_f and n_m, are assumed to be equal and either 50 or 500.

For each set of values for (ρ_f, ρ_m, d, p_{Af}, n_f, n_m) considered, we have simulated 1,000 SNPs from a population with the corresponding values for (ρ_f, ρ_m, d, p_{Af}), and for each one of the samples, we have computed $P(M_{ij}|y)$ for the ten models considered, and the posterior expected value of ρ_f, $\hat{\rho}_f = E(\rho_f|y)$ under the fully unrestricted model, M_{25}.

When simulations are set for populations with $d = 1$ and $\rho_m = 0$, as in the top two panels of Fig. 2, the true model is bound to be either M_{11}, when $\rho_f = 0$, or M_{12} when $\rho_f \neq 0$. These two panels present the average of the 1,000 values obtained for $P(M_{11}|y)$, for $P(M_{12}|y)$, and for the sum of the posterior probabilities of the other eight models, as a function of ρ_f. These averages estimate the expected value of $P(M_{11}|y)$, of $P(M_{12}|y)$, and of the sum of the remaining posterior probabilities for each given (ρ_f, ρ_m, d, p_{Af}). As desirable, the expected value of $P(M_{11}|y)$ and of $P(M_{12}|y)$ peak for the values of ρ_f where the corresponding model

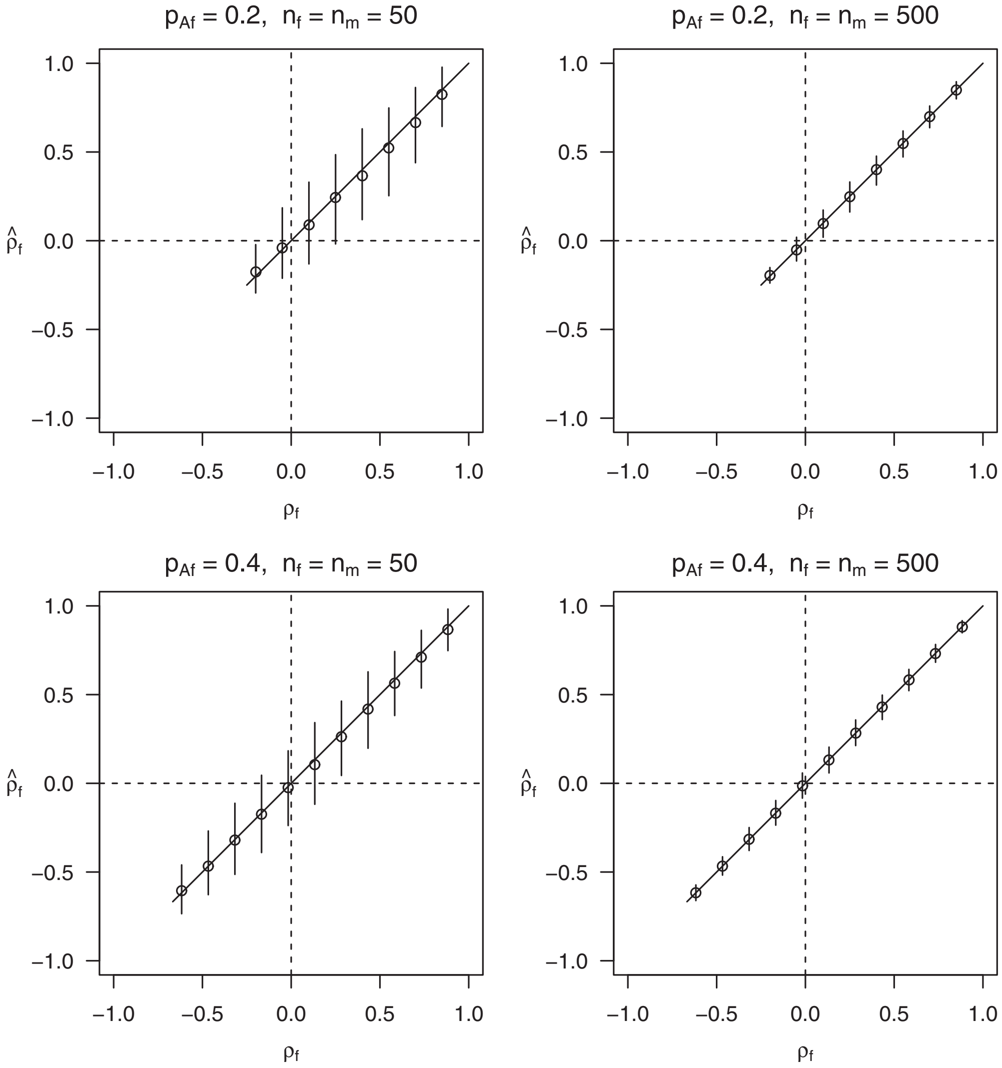

Fig. 3 Expected value of $\hat{\rho}_f = E(\rho_f|y)$ (circles), and 90% posterior-credible intervals for $\hat{\rho}_f = E(\rho_f|y)$ (line segments), as a function of ρ_f, when $d = 1$ and $\rho_m = 0$

holds true. One also observes that for n_f and n_m as small as 50, the sum of the posterior probabilities for the eight models known to be wrong is already negligible, and that the larger the sample sizes, the more peaked the expected values of $P(M_{11}|y)$ and of $P(M_{12}|y)$ are as a function of ρ_f, and hence the larger the power of the model selection procedure.

When simulations are set for populations with $d = 1.5$ and $\rho_m = 0.5$, as in the bottom two panels in Fig. 2, the true models are M_{23}, when $\rho_f = 0$, M_{24}, when $\rho_f = \rho_m = 0.5$, or M_{25}, when $\rho_f \neq 0$ and $\rho_f \neq \rho_m$. These two panels show that when $n_f = n_m = 500$, these three models are indeed the ones with the largest $P(M_{ij}|y)$ around the values for ρ_f for which they are the models in place, and they also show that the sum of the posterior probabilities of the other seven models, which are known to be wrong, is negligible. Instead, when $n_f = n_m = 50$, there is not enough power to tell M_{25} apart from M_{23} and M_{24} when M_{25} is the correct model and ρ_f is between 0 and 0.5. Also, when sample sizes are small, the posterior probabilities of the other seven models are not negligible anymore, even though none of these alternative models ever comes as the winner.

To evaluate the performance of our approach when it comes to estimating inbreeding coefficients, we explore how does the posterior expected value of ρ_f under M_{25} perform as an estimate for ρ_f. Figure 3 presents the 90% posterior-credible interval for $E(\rho_f|y)$ together with an average of the sample of 1,000 values obtained for $E(\rho_f|y)$ for known values of ρ_f in its whole range. Observe that the average of the $E(\rho_f|y)$ is extremely close to the true ρ_f of the population from which SNPs have been sampled from, which is an indication that $\hat{\rho}_f = E(\rho_f|y)$ is practically an unbiased estimate of ρ_f even for sample sizes as small as 50. The accuracy of the estimation of ρ_f grows with sample size and it is larger for $p_{Af} = 0.4$. Note also that the prior mode at 0 shrinks inferences about ρ toward 0 and hence, for very small n_f's and positive ρ_f's, there will be a small downward bias and not the upward bias that one might have anticipated from the fact that the prior expected value for ρ_f is positive.

Examples

To illustrate our approach to testing for HWP and EAF through Bayesian model selection, we analyze markers from chromosome 22 using data from the Japanese population of the 1,000 Genomes project (The 1,000 Genomes Project Consortium et al. 2015) consisting of $n_m = 56$ males and $n_f = 48$ females. Genetic variants were extracted with the PLINK program (Purcell et al. 2007), using only variants that had no missing values. Variants were LD pruned (using PLINK option –indep-pairwise 50 5 0.5) in order to produce an approximately uncorrelated subset.

Classification of ten single SNPs

To illustrate the use of our method, here, we report the posterior probabilities for the ten alternative scenarios described in the section Scenarios and priors for the ten SNPs presented in Table 1. These ten SNPs were chosen so that there is one with a largest posterior probability for each one of the ten scenarios considered.

The posterior probabilities, presented in Table 2, are computed through Eq. (30), assuming equal prior probabilities for the ten models, and hence $P(M_{ij}) = 1/10$, and using the expressions for the marginal likelihoods, $P(y|M_{ij})$ in Appendix 1 with $a_j = 1/3$ for all the Dirichlet priors on population genotype frequencies, and with $b_j = 1/2$ for the beta priors on population allele frequencies. The only exception will be in Scenario M_{15}, where we will use a Beta $(b_1^{15} = 1/3, b_2^{15} = 2/3)$ prior for the population genotype frequency, p_{AAm}. Given that each one of these priors corresponds to an effective sample size of only one, and data involve a sample size of $n = 104$, the role played by prior distributions will be negligible. Sample sizes will most often be larger than in this example, and hence in practice, the choice of a prior will most often be even less relevant than here.

The first marker in Table 1, rs566641289, has a posterior probability of 0.503 of being both in HWP and having EAF, and hence one rejects all the nine disequilibrium scenarios, with posterior probabilities of 0.215 or smaller. The second marker in Table 1, has a posterior probability smaller than .001 of being both in HWP and having EAF, but it has instead a posterior probability of 0.741 of being in the M_{12} disequilibrium scenario, with $d = 1$ and $\rho_m = 0$ but with $\rho_f \neq 0$, and hence where EAF holds but where HWP fails for females. For the third marker, the equilibrium scenario M_{11} is also rejected, because it has a posterior probability smaller than 0.001, and one settles with the M_{13} scenario,

Table 1 Genotype counts of ten single-nucleotide polymorphisms for males and females from a sample of the Japanese population study with $n_m = 56$ and $n_f = 48$

	SNP	n_{AAm}	n_{ABm}	n_{BBm}	n_{AAf}	n_{ABf}	n_{BBf}
1	rs566641289	46	10	0	40	8	0
2	rs67982243	19	29	8	10	37	1
3	rs398040486	12	42	2	17	25	6
4	rs11913608	22	34	0	18	30	0
5	rs5761644	33	23	0	35	8	5
6	rs2071891	18	29	9	30	15	3
7	rs137960	37	18	1	17	30	1
8	rs546414427	29	27	0	16	23	9
9	rs9798787	48	6	2	29	13	6
10	rs112334000	6	50	0	0	47	1

Table 2 Posterior probabilities, $P(M_{ij}|y)$, of the ten scenarios considered in the section Scenarios and priors for the ten SNPs presented in Table 1

SNP	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{21}	M_{22}	M_{23}	M_{24}	M_{25}
rs566641289	**0.503**	0.078	0.072	0.215	0.021	0.056	0.019	0.020	0.008	0.007
rs67982243	0.000	**0.741**	0.000	0.004	0.170	0.000	0.071	0.000	0.001	0.012
rs398040486	0.000	0.000	**0.775**	0.021	0.098	0.000	0.000	0.083	0.006	0.016
rs11913608	0.000	0.000	0.000	**0.899**	0.033	0.000	0.000	0.000	0.041	0.027
rs5761644	0.030	0.280	0.051	0.004	**0.434**	0.003	0.036	0.014	0.001	0.147
rs2071891	0.043	0.013	0.006	0.005	0.003	**0.605**	0.123	0.095	0.088	0.019
rs137960	0.009	0.087	0.002	0.034	0.028	0.028	**0.597**	0.007	0.058	0.151
rs546414427	0.007	0.001	0.036	0.001	0.004	0.046	0.007	**0.735**	0.053	0.111
rs9798787	0.002	0.005	0.011	0.050	0.035	0.067	0.076	0.105	**0.529**	0.120
rs112334000	0.000	0.000	0.000	0.102	0.037	0.000	0.000	0.000	0.427	**0.433**

The largest value for each variant is marked in bold with $d = 1$ and $\rho_f = 0$ but $\rho_m \neq 0$, because it has a posterior probability of 0.775. For the last marker, the most probable scenario is the saturated M_{25} disequilibrium scenario with $d \neq 1$, $\rho_f \neq 0$, and $\rho_m \neq 0$, even though $P(M_{25}|y) = 0.433$ is smaller than 0.5.

One of the advantages of the Bayesian approach, is that one can easily simulate samples from the marginal posterior distribution of any function of the parameters, and that is very useful when it comes to present and interpret the results. Figures 4 and 5, for example, present samples from the marginal posterior distributions of (p_{Af}, p_{Am}), of (ρ_f, ρ_m), of $\log(d)$, and of $\rho_m - \rho_f$ for the ten markers in Table 1. These marginal posterior distributions are computed assuming the fully unrestricted model, M_{25}, as described in Appendix 2. These two figures also present 90% highest posterior density (hpd) credible intervals/ regions for these parameter values or pairs of parameter values.

Note that in Figs. 4 and 5, all posterior distributions for a given SNP are coherent with the characteristics of the model with the largest posterior probability. In particular, observe that the posterior-credible intervals for $\log(d)$ in Fig. 4, corresponding to the five SNPs with a most probable model with $d = 1$, all include 0 in the intervals for $\log(d)$, while the opposite happens in Fig. 5, where all SNPs are from scenarios with $d \neq 1$. Also, note that for all the SNPs classified into one of the four models having EIC, the credible intervals for $\rho_m - \rho_f$ include 0 and the credible regions for (ρ_f, ρ_m) include a substantial part of the diagonal $\rho_f = \rho_m$. The same coherence is observed in the samples from the posterior of (p_{Af}, p_{Am}).

We analyzed the same set of SNPs using Akaike's information criterion as proposed by Graffelman and Weir (2017). In order to do so, the maximum likelihood (ML) estimators of scenarios M_{12}, M_{13}, M_{22}, and M_{23}, not covered by that paper, were derived. Appendix 3 provides the details on the computation of the MLE for all ten scenarios. Table 3 presents the value taken by AIC for all ten SNPs and all ten scenarios; note that the AIC is always the smallest for the same scenario chosen by our Bayesian procedure, and that AIC and posterior probabilities provide very similar rankings of the models for all ten SNPs considered.

Simultaneous analysis of multiple SNPs

In this section, we illustrate the Bayesian model selection approach to testing for HWP and EAF by carrying out the simultaneous analysis of the set of all 107.261 complete polymorphic SNPs with RS identifiers on chromosome 22 of the Japanese population of the 1,000 Genomes project.

Figure 6 presents the model with the largest posterior probability for each one of these SNPs, presented in the order in which these SNPs appear on the chromosome. Consecutive sequences of markers being systematically classified to the same scenario might be an indication of quality control problems in the SNP measurements. Too few SNPs being classified as being in HW equilibrium would also be an indication of either a problem in the measurements or of the fact that the population under scrutiny is actually in disequilibrium.

Figure 6 also indicates the proportion of these 107.261 SNPs that have been classified into each one of the ten alternative scenarios described in the section Scenarios and priors. Scenario M_{11}, representing the setting where both EAF as well as HWP for both males and females are in place, is the scenario with the largest posterior probability for 92.53% of all the SNPs considered. Scenario M_{14} is the second most frequent scenario among all SNPs, because it is the one with the largest posterior probability in 2.67% of the cases considered. Scenarios M_{21} and M_{12} are the third and fourth most frequent ones, because they are the ones with the largest probability in 1.66% and in 1.46% of the cases, respectively. The least frequent scenario among the SNPs from chromosome 22 is the saturated model, M_{25}, which is the most probable model in only 0.01% of the cases.

Figure 6 shows a dense stripe for the most common scenario M_{11}, and also reflects that EAF scenarios M_{1i} are far more common than the corresponding M_{2i} scenarios that have different allele frequencies in the sexes. There is overall, little evidence for deviations from EAF, and if it is found, then HWP still mostly hold for both sexes separately. Chromosome 22 is acrocentric, having its centromere in the interval 12.2–17.9 Mb (hg 19). It is known that the centromere region is hard to genotype, and that deviation from HWP is more often found for variants inside and flanking the centromere (Graffelman et al. 2017). Figure 6 also shows that scenario M_{15} is more common in the centromere in comparison with the rest of the chromosome. In fact, all disequilibrium scenarios with EAF were found to be more common in the centromere, though this is hard to perceive in Fig. 6. Outside the centromere, 92.7% of the variants are assigned to the M_{11} scenario, whereas this drops to 88.5% in the centromere region.

Discussion

We have presented a Bayesian method for joint inference on Hardy–Weinberg equilibrium and on equality of allele frequencies for biallelic markers. Disequilibrium might be due to a difference in allele frequencies between the sexes, to males or females not satisfying the HW proportions, or any combination of these situations simultaneously. By computing the posterior probability for each scenario, one can classify each SNP into its most probable scenario.

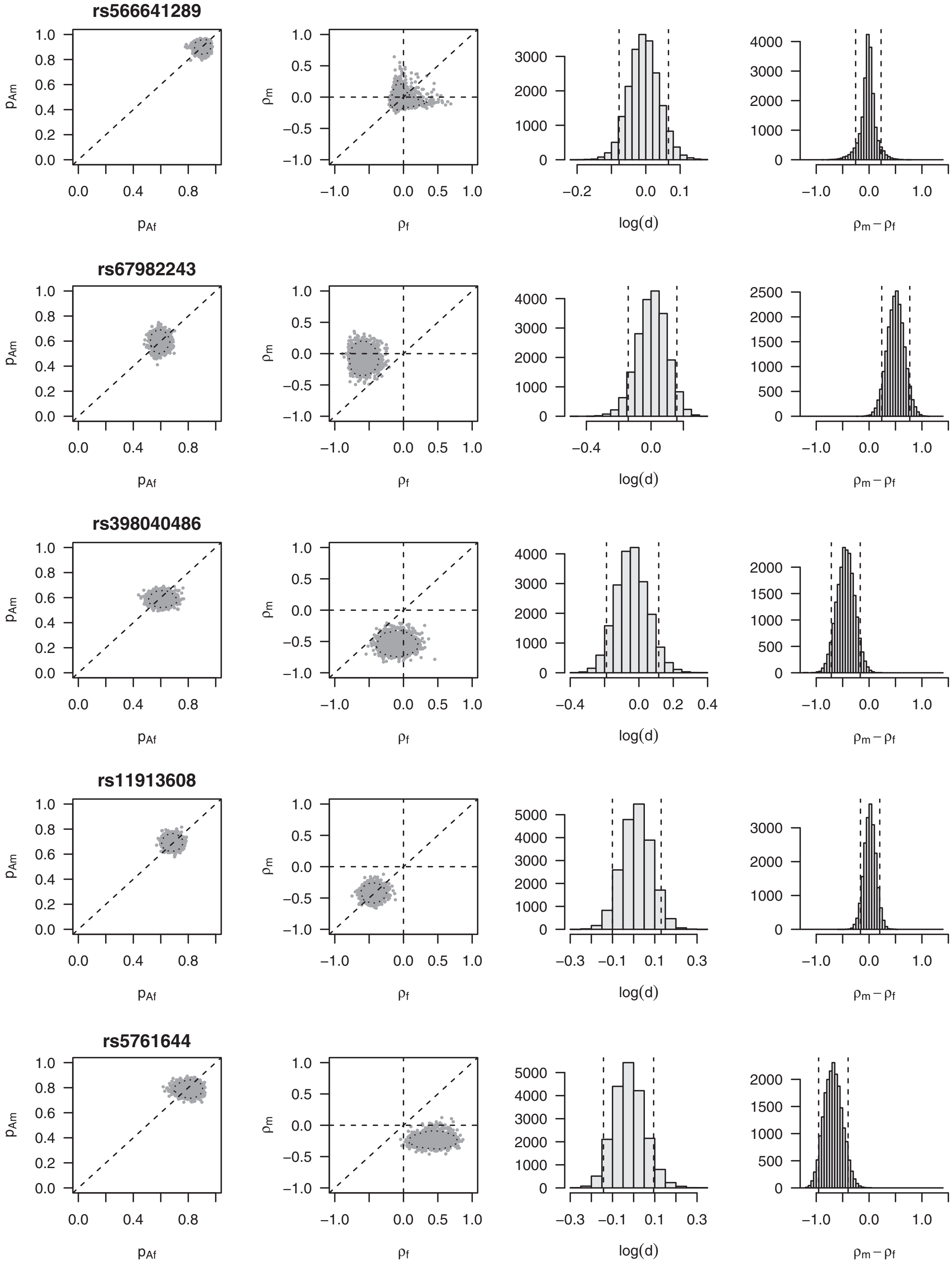

Fig. 4 Samples from the marginal posterior distributions of (p_{Af}, p_{Am}), of (ρ_f, ρ_m), of $\log(d)$, and of $\rho_m - \rho_f$, and 90% hpd posterior-credible regions, all for the first five SNPs in Table 1

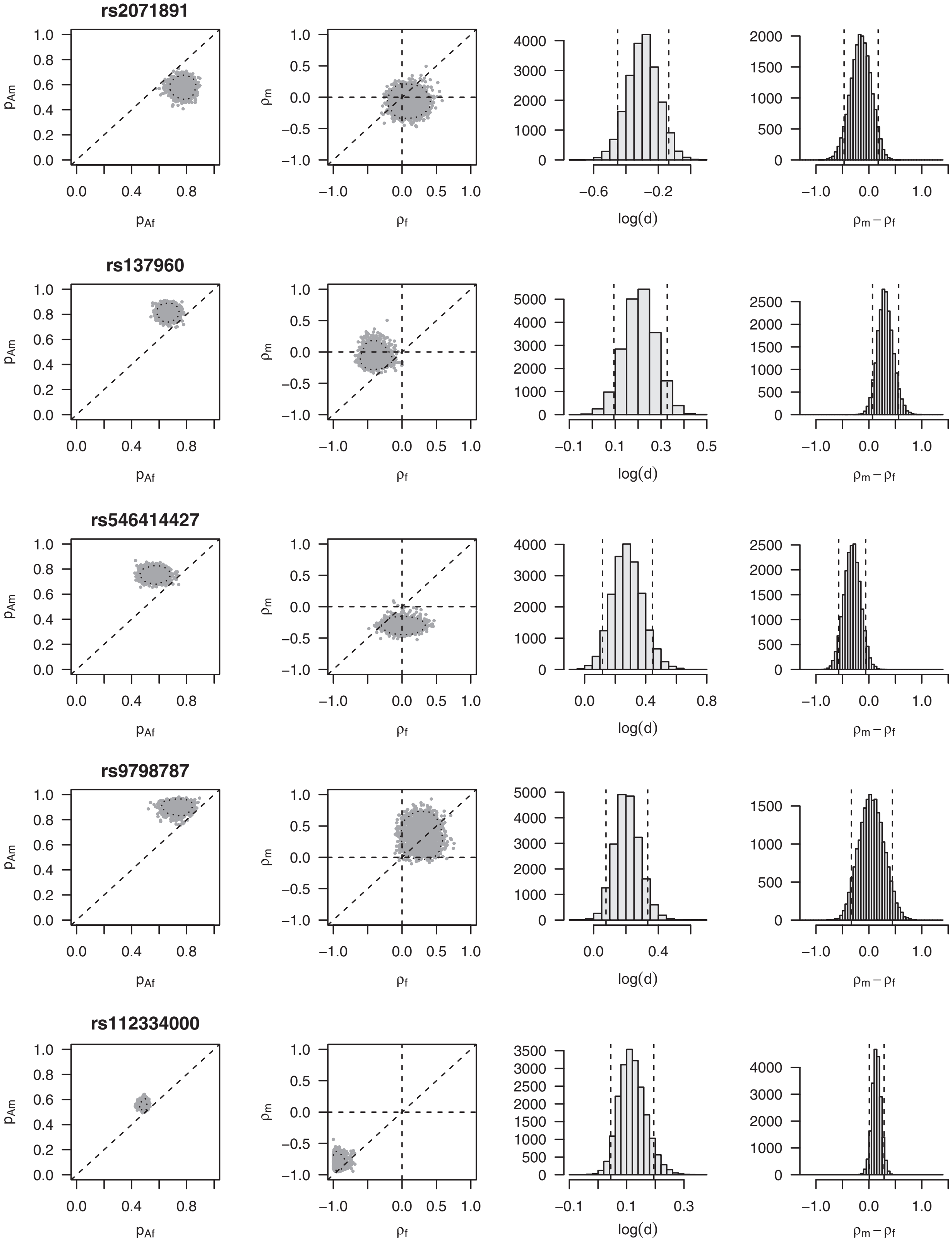

Fig. 5 Samples from the marginal posterior distributions of (p_{Af}, p_{Am}), of (ρ_f, ρ_m), of log(d). and of $\rho_m - \rho_f$, and 90% hpd posterior-credible regions, all for the last five SNPs in Table 1

Table 3 AIC of the ten scenarios considered in the section Scenarios and priors for the ten SNPs presented in Table 1

SNP	M_{11}	M_{12}	M_{13}	M_{14}	M_{15}	M_{21}	M_{22}	M_{23}	M_{24}	M_{25}
rs566641289	**99.54**	100.81	100.56	99.83	101.83	101.52	102.79	102.53	101.83	103.81
rs67982243	191.11	**173.09**	192.78	182.15	174.76	193.11	175.09	194.77	183.99	176.75
rs398040486	188.93	190.46	**171.99**	178.03	173.52	190.79	192.31	173.80	180.03	175.33
rs11913608	170.05	157.91	156.73	**142.59**	144.59	172.03	159.88	158.70	144.59	146.55
rs5761644	165.52	159.33	161.52	167.18	**155.32**	167.41	161.24	163.40	168.91	157.23
rs2071891	203.91	204.88	205.86	205.75	206.82	**196.22**	197.89	198.00	198.22	199.66
rs137960	169.39	163.35	171.15	164.82	165.21	164.77	**157.39**	166.20	161.12	158.81
rs546414427	195.59	197.41	190.80	196.68	192.47	189.44	191.42	**182.71**	187.55	184.69
rs9798787	164.16	160.94	159.89	155.93	157.12	155.17	153.22	153.02	**149.10**	151.07
rs112334000	155.40	100.15	111.40	61.04	62.99	156.55	100.38	112.02	57.30	**55.86**

The minimal value for each variant is marked in bold

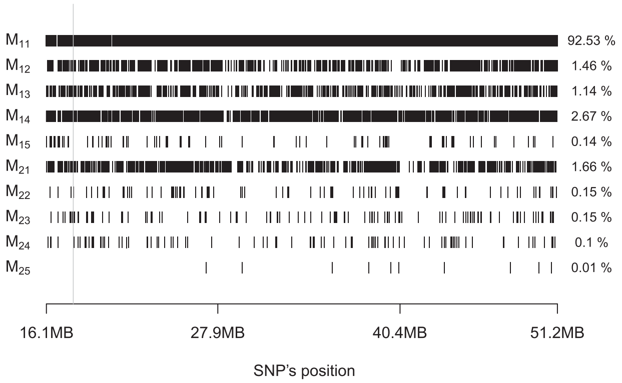

Fig. 6 Model with the largest posterior probability for the SNPs selected from the Japanese population, presented in the position where they are placed on the chromosome. The vertical gray line indicates the boundary of the centromere, and the numbers on the right extreme correspond to the proportion of SNPs classified into each one of the scenarios considered

Frequentist tests compare models (scenarios) in pairs, and when choosing between equilibrium, (i.e., M_{11}), and disequilibrium, (i.e., any of the other nine models), one can pair equilibrium with disequilibrium in nine different ways. In order to precisely determine the disequilibrium scenario that is in place with a frequentist approach, several statistical tests are necessary, because the number of ways in which scenarios can be paired increases a lot, and many of these pairs of scenarios are not nested. Instead, by assigning a posterior probability to each one of the ten scenarios, with the ten probabilities adding up to one, our Bayesian approach provides a simple way of selecting the most probable scenario in the light of the data. Moreover, the Bayesian approach allows the simultaneous comparison of all possible models, whereas the likelihood ratio approach is restricted to compare models that are nested.

Akaike's information criterion also offers an easy way to select the best fitting model among all available models. However, if two models have a similar AIC, then it may be hard to tell how strong the evidence is in favor of the better fitting model. In the Bayesian approach, the scenarios are compared in a probability scale, and this gives a better idea of the extent to which the best fitting model outperforms its competitors.

One nice feature of the Bayesian approach is that, on top of providing posterior probabilities for the scenarios, it also yields the posterior distribution of the parameters of interest, the way described in Appendix 2 and illustrated in the section Classification of ten single SNPs.

We have found it convenient to parameterize disequilibrium by using the inbreeding coefficient and the ratio of male to female allele frequencies, using a Dirichlet prior on the genotype frequencies. Alternatively, priors specified directly on the disequilibrium measures might also be considered.

The Bayesian procedures described here do not require one to implement MCMC methods, as it is usual in most Bayesian applications, and that simplifies computations a lot. If the integration required for the computation of some of the posterior probabilities is carried out efficiently, there will not be any problem in using this method for complete chromosomes.

The analysis of variants on chromosome 22 by our proposed Bayesian procedure reveals more deviation from the equilibrium scenario in the centromere region, in comparison with the rest of the chromosomes, which is consistent with previous disequilibrium studies. A more detailed analysis of centrometric regions could be of interest, given the role this region has in human diseases like cancer (Barra and Fachinetti 2018). The analysis also shows that deviation from HWP is far more frequent than deviation from EAF. This is in agreement with the population-genetic principle that it takes only one generation to achieve EAF, but two to achieve HWP.

Software

The Bayesian methods presented here have been programmed in R by Xavier Puig. Bayesian model selection can be carried out using function HWPosterior of version

1.6.2 of the Hardy–Weinberg package (Graffelman 2015). Function HWAIC does the AIC calculations.

Acknowledgements This work was supported by grants MTM2015-65016-C2-2-R (MINECO/FEDER) and MTM2013-43992-R of the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund, and by Grant R01 GM075091 from the United States National Institutes of Health. Grant numbers: MTM2015-65016-C2-2-R, MTM2013-43992-R (MINECO/FEDER), and R01 GM075091 (US NIH).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix 1: marginal likelihoods

Here, we present the marginal likelihoods, $P(y|M_{ij})$ for the ten models needed to compute the posterior probabilities, $P(M_{ij}|y)$, through Eq. (30). The priors assumed are the ones described in the section Scenarios and priors, and $y=(n_{AAf}, n_{ABf}, n_{BBf}, n_{AAm}, n_{ABm}, n_{BBm})$.

For convenience, we define constant K as the product of the multinomial coefficients involving the male and female genotype counts:

$$K=\frac{n_f!}{n_{AAf}!n_{ABf}!n_{BBf}!}\frac{n_m!}{n_{AAm}!n_{ABm}!n_{BBm}!}.$$

The marginal likelihood under M_{11}, is

$$P(y|M_{11})=K\frac{\Gamma(\sum_{j=1}^{2}b_j^{11})}{\prod_{j=1}^{2}\Gamma(b_j^{11})}\frac{\Gamma(b_1^{11}+n_A)\Gamma(b_2^{11}+n_B)}{\Gamma(\sum_{j=1}^{2}b_j^{11}+2n)}2^{n_{AB}},$$

The marginal likelihood under M_{21}, with $d\neq 1$ and $\rho_f=\rho_m=0$, can be computed through

$$P(y|M_{21})=K\frac{\Gamma\left(\sum_{j=1}^{2}b_{jf}^{21}\right)\Gamma\left(\sum_{j=1}^{2}b_{jm}^{21}\right)\Gamma(b_{1f}^{21}+n_{Af})\Gamma(b_{2f}^{21}+n_{Bf})\Gamma(b_{1m}^{21}+n_{Am})\Gamma(b_{2m}^{21}+n_{Bm})}{\prod_{j=1}^{2}\Gamma(b_{jf}^{21})\ \prod_{j=1}^{2}\Gamma(b_{jm}^{21})\ \Gamma\left(\sum_{j=1}^{2}b_{jf}^{21}+2n_f\right)\ \Gamma\left(\sum_{j=1}^{2}b_{jm}^{21}+2n_m\right)}2^{n_{AB}}$$

The marginal likelihood under M_{12}, with $d=1$, $\rho_f\neq 0$, and $\rho_m=0$, is

$$\begin{aligned}P(y|M_{12})=K&\frac{\Gamma\left(\sum_{j=1}^{3}a_{jf}^{12}\right)}{\prod_{j=1}^{3}\Gamma\left(a_{jf}^{12}\right)}2^{n_{ABm}}\\&\times\int_0^1\int_0^{1-p_{ABf}}p_{AAf}^{a_{1f}^{12}+n_{AAf}-1}\\&\times p_{ABf}^{a_{2f}^{12}+n_{ABf}-1}(1-p_{AAf}-p_{ABf})^{a_{3f}^{12}+n_{BBf}-1}\\&\times\left(\tfrac{2p_{AAf}+p_{ABf}}{2}\right)^{n_{Am}}\left(1-\tfrac{2p_{AAf}+p_{ABf}}{2}\right)^{n_{Bm}}dp_{AAf}dp_{ABf}.\end{aligned}$$

The marginal likelihood under M_{22}, with $d\neq 1$, $\rho_f\neq 0$, and $\rho_m=0$, is

$$\begin{aligned}&P(y|M_{22})=\\&K\frac{\Gamma\left(\sum_{j=1}^{3}a_{jf}^{22}\right)\Gamma\left(\sum_{j=1}^{2}b_{jm}^{22}\right)\Gamma(a_{1f}^{22}+n_{AAf})\Gamma(a_{2f}^{22}+n_{ABf})\Gamma(a_{3f}^{22}+n_{BBf})\Gamma(b_{1m}^{22}+n_{Am})\Gamma(b_{2m}^{22}+n_{Bm})}{\prod_{j=1}^{3}\Gamma(a_{jf}^{22})\ \prod_{j=1}^{2}\Gamma(b_{jm}^{22})\ \Gamma\left(\sum_{j=1}^{3}a_{jf}^{22}+n_f\right)\ \Gamma\left(\sum_{j=1}^{2}b_{jm}^{22}+2n_m\right)}2^{n_{ABm}}\end{aligned}$$

The marginal likelihood under M_{13}, with $d=1$, $\rho_f=0$, and $\rho_m\neq 0$, is

$$\begin{aligned}P(y|M_{13})=K&\frac{\Gamma\left(\sum_{j=1}^{3}a_{jm}^{13}\right)}{\prod_{j=1}^{3}\Gamma(a_{jm}^{13})}2^{n_{ABf}}\\&\times\int_0^1\int_0^{1-p_{ABm}}p_{AAm}^{a_{1m}^{13}+n_{AAm}-1}\\&\times p_{ABm}^{a_{2m}^{13}+n_{ABm}-1}(1-p_{AAm}-p_{ABm})^{a_{3m}^{13}+n_{BBm}-1}\\&\times\left(\tfrac{2p_{AAm}+p_{ABm}}{2}\right)^{n_{Af}}\left(1-\tfrac{2p_{AAm}+p_{ABm}}{2}\right)^{n_{Bf}}dp_{AAm}dp_{ABm}.\end{aligned}$$

The marginal likelihood under M_{23}, with $d\neq 1$, $\rho_f=0$, and $\rho_m\neq 0$, is

$$\begin{aligned}P(y|M_{23})=K&\frac{\Gamma\left(\sum_{j=1}^{2}b_{jf}^{23}\right)}{\prod_{j=1}^{2}\Gamma(b_{jf}^{23})}\frac{\Gamma\left(\sum_{j=1}^{3}a_{jm}^{23}\right)}{\prod_{j=1}^{3}\Gamma(a_{jm}^{23})}\frac{\Gamma(b_{1f}^{23}+n_{Af})\Gamma(b_{2f}^{23}+n_{Bf})}{\Gamma\left(\sum_{j=1}^{2}b_{jf}^{23}+2n_f\right)}\\&\times\frac{\Gamma(a_{1m}^{23}+n_{AAm})\Gamma(a_{2m}^{23}+n_{ABm})\Gamma(a_{3m}^{23}+n_{BBm})}{\Gamma\left(\sum_{j=1}^{3}a_{jm}^{23}+n_m\right)}2^{n_{ABf}}.\end{aligned}$$

The marginal likelihood under M_{14}, with $d=1$, $\rho_f\neq 0$, and $\rho_m\neq 0$ but $\rho_f=\rho_m$, is

$$P(y|M_{14})=K\frac{\Gamma(\sum_{j=1}^{3}a_{jf}^{14})}{\prod_{j=1}^{3}\Gamma(a_{jf}^{14})}\frac{\Gamma(a_{1f}^{14}+n_{AA})\Gamma(a_{2f}^{14}+n_{AB})\Gamma(a_{3f}^{14}+n_{BB})}{\Gamma(\sum_{j=1}^{3}a_{jf}^{14}+n)}$$

The marginal likelihood under M_{24}, with $d\neq 1$, $\rho_f\neq 0$, and $\rho_m\neq 0$ but $\rho_f=\rho_m$, is

$$\begin{aligned}&P(y|M_{24})=\\&K\frac{\Gamma\left(\sum_{j=1}^{3}a_{jf}^{24}\right)}{\prod_{j=1}^{3}\Gamma(a_{jf}^{24})}2^{n_{ABm}}\int_0^1\int_0^{1-p_{ABf}}\int_{max(0,t_1)}^{min(t_2,1)}p_{AAf}^{a_{1f}^{24}+n_{AAf}-1}p_{ABf}^{a_{2f}^{24}+n_{ABf}-1}\\&\quad\times(1-p_{AAf}-p_{ABf})^{a_{3f}^{24}+n_{BBf}-1}\\&\quad\times p_{Am}^{b_{1m}^{24}+n_{ABm}-1}(1-p_{Am})^{b_{2m}^{24}+n_{ABm}-1}\\&\quad\times\left(p_{Am}^2+p_{Am}(1-p_{Am})\tfrac{p_{AAf}-(p_{AAf}+\frac{1}{2}p_{ABf})^2}{(p_{AAf}+\frac{1}{2}p_{ABf})(1-p_{AAf}-\frac{1}{2}p_{ABf})}\right)^{n_{AAm}}\\&\quad\times\left(1-\tfrac{p_{AAf}-(p_{AAf}+\frac{1}{2}p_{ABf})^2}{(p_{AAf}+\frac{1}{2}p_{ABf})(1-p_{AAf}-\frac{1}{2}p_{ABf})}\right)^{n_{ABm}}\frac{1}{\int_{max(0,t_1)}^{min(t_2,1)}p_{Am}^{b_{1m}^{24}-1}(1-p_{Am})^{b_{2m}^{24}-1}dp_{Am}}\\&\quad\times\left((1-p_{Am})^2+p_{Am}(1-p_{Am})\tfrac{p_{AAf}-(p_{AAf}+\frac{1}{2}p_{ABf})^2}{(p_{AAf}+\frac{1}{2}p_{ABf})(1-p_{AAf}-\frac{1}{2}p_{ABf})}\right)^{n_{BBm}}dp_{Am}dp_{AAf}dp_{ABf},\end{aligned}$$

where

$$t_1=(4p_{AAf}^2+4p_{AAf}p_{ABf}-4p_{AAf}+p_{ABf}^2)/(2p_{ABf}),$$

$$t_2=(-4p_{AAf}^2-4p_{AAf}p_{ABf}+4p_{AAf}-p_{ABf}^2+2p_{ABf})/(2p_{ABf}).$$

The marginal likelihood under M_{15}, with $d = 1$, $\rho_f \neq 0$, and $\rho_m \neq 0$, is

$$P(y|M_{15}) = K\frac{\Gamma\left(\sum_{j=1}^{3} a_{jf}^{15}\right)}{\prod_{j=1}^{3}\Gamma(a_{jf}^{15})}\int_0^1\int_0^{1-p_{ABf}}\int_{max(0,2p_{AAf}+p_{ABf}-1)}^{p_{AAf}+0.5p_{ABf}} p_{AAf}^{a_{1f}^{15}+n_{AAf}-1} p_{ABf}^{a_{2f}^{15}+n_{ABf}-1}$$

$$\times(1-p_{AAf}-p_{ABf})^{a_{3f}^{15}+n_{BBf}-1} p_{AAm}^{b_{1m}^{15}+n_{AAm}-1}$$

$$\times(1-p_{AAm})^{b_{2m}^{15}-1}(2p_{AAf}+p_{ABf}-2p_{AAm})^{n_{ABm}}$$

$$\times(1+p_{AAm}-2p_{AAf}-p_{ABf})^{n_{BBm}}\frac{1}{\int_{max(0,2p_{AAf}+p_{ABf}-1)}^{p_{AAf}+0.5p_{ABf}} p_{AAm}^{b_{1m}^{15}-1}(1-p_{AAm})^{b_{2m}^{15}-1}dp_{AAm}}dp_{AAm}dp_{AAf}dp_{ABf}.$$

The marginal likelihood under M_{25}, with $d \neq 1$, $\rho_f \neq 0$, and $\rho_m \neq 0$, is

$$P(y|M_{25}) = K\frac{\Gamma\left(\sum_{j=1}^{3} a_{jf}^{25}\right)\Gamma\left(\sum_{j=1}^{3} a_{jm}^{25}\right)\Gamma(a_{1f}^{25}+n_{AAf})\Gamma(a_{2f}^{25}+n_{ABf})\Gamma(a_{3f}^{25}+n_{BBf})}{\prod_{j=1}^{3}\Gamma(a_{jf}^{25})\ \prod_{j=1}^{3}\Gamma(a_{jm}^{25})\qquad\Gamma\left(\sum_{j=1}^{3} a_{jf}^{25}+n_f\right)}$$

$$\times\frac{\Gamma(a_{1m}^{25}+n_{AAm})\Gamma(a_{2m}^{25}+n_{ABm})\Gamma(a_{3m}^{25}+n_{BBm})}{\Gamma\left(\sum_{j=1}^{3} a_{jm}^{25}+n_m\right)}$$

Note that the only models that require integration are M_{12}, M_{13}, M_{24}, and M_{15}. However, it can be carried out numerically without any problem because the integration region is compact, and grid size can be set to be as small as needed for the precision required.

Appendix 2: posterior distribution under M_{25}

Under the saturated model M_{25}, (n_{AAf}, n_{ABf}, n_{BBf}) is multinomially (n_f,(p_{AAf}, p_{ABf}, p_{BBf})) distributed and (n_{AAm}, n_{ABm}, n_{BBm}) is multinomially (n_m, (p_{AAm}, p_{ABm}, p_{BBm})) distributed.

If (p_{AAf}, p_{ABf}, p_{BBf}) is Dirichlet $(a_{1f}^{25}, a_{2f}^{25}, a_{3f}^{25})$, and ($p_{AAm}$, p_{ABm}, p_{BBm}) is Dirichlet $(a_{1m}^{25}, a_{2m}^{25}, a_{3m}^{25})$, the posterior distribution for (p_{AAf},p_{ABf},p_{BBf}) is

$$\pi(p_{AAf}, p_{ABf}, p_{BBf}|y) = \text{Dirichlet}(a_{1f}^{25}+n_{AAf}, a_{2f}^{25}+n_{ABf}, a_{3f}^{25}+n_{BBf}), \quad (31)$$

independent of the posterior distribution for (p_{AAm}, p_{ABm}, p_{BBm}), which is

$$\pi(p_{AAm}, p_{ABm}, p_{BBm}|y) = \text{Dirichlet}(a_{1m}^{25}+n_{AAm}, a_{2m}^{25}+n_{ABm}, a_{3m}^{25}+n_{BBm}). \quad (32)$$

The marginal posterior distributions for p_{Af}, p_{Am}, d, ρ_f, and ρ_m follow from the ones for (p_{AAf}, p_{ABf}, p_{BBf}) and for (p_{AAm}, p_{ABm}, p_{BBm}), and they can be easily estimated by simulating large samples of (p_{AAf}, p_{ABf}, p_{BBf}), and of (p_{AAm}, p_{ABm}, p_{BBm}), and by computing the corresponding value of p_{Af}, p_{Am}, d, ρ_f, and ρ_m for each value in the sample.

Appendix 3: maximum likelihood estimators

In this appendix, we give the maximum likelihood estimators for those scenarios for which a closed-form expression has been found. These are needed for the calculation of Akaike's information criterion, $AIC = 2k - 2\ \log L(\widehat{\boldsymbol{\theta}})$, where k is the number of parameters of the model, and $\hat{\theta}$ the vector of ML estimators. Maximum likelihood (ML) estimators for the parameters of the new models M_{12}, M_{13}, M_{22}, and M_{23} were derived for this paper. We also include the estimators for models M_{11}, M_{14}, M_{15}, M_{21}, M_{24}, and M_{25} previously derived by Graffelman and Weir (2017), and labeled in their article as A, B, C, D, E, and F, respectively.

M$_{11}$. This model has $p_A = p_{Af} = p_{Am}$ and $\rho = \rho_f = \rho_m = 0$. There is only one free parameter to be estimated, so $k = 1$. The ML estimator of p_A is just the overall sample A allele frequency $\hat{p}_A = n_A/(2n)$.

M$_{12}$. This model has $p_A = p_{Af} = p_{Am}$, $\rho_m = 0$, no restrictions on ρ_f, and so $k = 2$. Solving the likelihood equations, one finds the relationship $\hat{\rho}_f = -(n_A - 2n\hat{p}_A)/(n_{Am} - 2n_m\hat{p}_A)$. This can be used to solve the likelihood equations numerically in one parameter, after which the other is inferred.

M$_{13}$. This is essentially the same model as M_{12}, but with the sexes interchanged. It has $k = 2$ and the parameters p_A and ρ_m can be estimated by the same numerical procedure outlined for M_{12}.

M$_{14}$. This model has EAF and EIC such that $p_A = p_{Af} = p_{Am}$ and $\rho = \rho_f = \rho_m$, giving $k = 2$. The ML estimator for p_A is the overall sample A allele frequency, and the ML estimator for ρ is $\hat{\rho} = (4n_{AA}n_{BB} - n_{AB}^2)/(n_A n_B)$.

M$_{15}$. This model has only EAF such that $p_A = p_{Af} = p_{Am}$, and different inbreeding coefficients ρ_f and ρ_m for the sexes, and so $k = 3$. No closed-form solution was reported for this model, and its three parameters are estimated by iterative maximization.

M$_{21}$. This model has HWP for both sexes, $\rho_m = \rho_f = 0$, and $k = 2$ because the two allele frequencies p_{Am} and p_{Af} need to be estimated. The ML estimators for the allele frequencies are the male and female sample A allele frequencies, respectively, that is, $\hat{p}_{Am} = n_{Am}/(2n_m)$ and $\hat{p}_{Af} = n_{Af}/(2n_f)$.

M$_{22}$. This model has different allele frequencies for the sexes, p_{Am} and p_{Af}, and one inbreeding coefficient for females only, ρ_f, because $\rho_m = 0$ and so $k = 3$. The ML estimators for the male and female allele frequency are the corresponding sample allele frequencies and the ML estimator for ρ_f is $\hat{\rho}_f = (4n_{AAf}n_{BBf} - n_{ABf}^2)/(n_{Af}n_{Bf})$.

M$_{23}$. This model is essentially the same as model M_{22}, but with the sexes interchanged. It has $\rho_f = 0$, and $k = 3$. The ML estimators for the male and female allele frequencies are the corresponding sample allele frequencies, and ρ_m is estimated by $\hat{\rho}_m = (4n_{AAm}n_{BBm} - n_{ABm}^2)/(n_{Am}n_{Bm})$.

M$_{24}$. This model assumes EIC such that $\rho_m = \rho_f = \rho$, and $k = 3$, requiring estimation of ρ, p_{Am}, and p_{Af}. No closed-form expressions have been found for the ML estimators, and the latter are estimated by iterative maximization, taking into account the usual range constraints for the inbreeding coefficients and allele frequencies.

$\mathbf{M}_{25}$. This is the full model that has no restrictions on the parameters, except the usual range constraints, and has the largest number of free parameters, $k = 4$. The ML estimators are given by the sample allele frequencies, and the sex-specific inbreeding coefficient estimators already given under models M_{22} and M_{23}.

References

Barra V, Fachinetti D (2018) The dark side of centromeres: types, causes and consequences of structural abnormalities implicating centromeric DNA. Nat Commun 9:4340. https://doi.org/10.1038/s41467-018-06545-y

Berger JO, Bernardo JM, Sun D (2015) Overall objective priors. Bayesian Anal 10:189–221. https://doi.org/10.1214/14-BA915

Bernardo J, Tomazella V (2010) Bayesian reference analysis of the Hardy-Weinberg equilibrium. In: Chen MH, Dey DK, Muller P, Sun D, Ye K (eds) Frontiers of Statistical Decision Making and Bayesian Analysis, In Honor of James O. Berger, Springer Verlag. New York, USA. p 31–43

Cannings C, Edwards AWF (1968) Natural selection and the de finetti diagram. Ann Hum Genet 31:421–428. https://doi.org/10.1111/j.1469-1809.1968.tb00575.x

Consonni G, Gutierrez-Pena E, Veronese P (2008) Compatible priors for Bayesian model comparison with an application to the Hardy-Weinberg equilibrium model. Test 17:585–605. https://doi.org/10.1007/s11749-007-0057-7

de Finetti B (1926) Considerazioni matematiche sul l'ereditarieta mendeliana. Metron 6(3):1–41

Graffelman J (2015) Exploring diallelic genetic markers: the Hardy-Weinberg package. J Stat Softw 64:1–23. http://www.jstatsoft.org/v64/i03/paper

Graffelman J, Morales-Camarena J (2008) Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Hum Hered 65:77–84. https://doi.org/10.1159/000108939

Graffelman J, Weir BS (2016) Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome. Heredity 116:558–568. https://doi.org/10.1038/hdy.2016.20

Graffelman J, Weir BS (2017) On the testing of Hardy-Weinberg proportions and equality of allele frequencies in males and females at biallelic genetic markers. Genet Epidemiol 42:34–48. https://doi.org/10.1002/gepi.22079

Graffelman J, Jain D, Weir BS (2017) A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data. Hum Genet 136:727–741. https://doi.org/10.1007/s00439-017-1786-7

Hamilton MB (2009) Population genetics. Wiley-Blackwell, Chichester, UK.

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

Hartl DL (1980) Principles of Population Genetics. Sinauer Associates, Sunderland, Massachusetts

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, Zheng X, Weir BS (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34:591–602. https://doi.org/10.1002/gepi.20516

Lindley D (1988) Statistical inference concerning Hardy-Weinberg equilibrium. In: Bernardo J, DeGroot M, Lindley D, ASmith (eds) Bayesian Statistics 3, Oxford University Press, Oxford, p 307–320

Puig X, Ginebra J, Graffelman J (2017) A Bayesian test for Hardy-Weinberg equilibrium of bi-allelic X-chromosomal markers. Heredity 119:226–236. https://doi.org/10.1038/hdy.2017.30

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559–575. https://doi.org/10.1086/519795

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

Single RM, Meyer D, Hollenbach JA, Nelson MP, Noble JA, Erlich HA, Thomson G (2002) Haplotype frequency estimation in patient populations: the effect of departures from Hardy-Weinberg proportions and collapsing over a locus in the HLA region. Genet Epidemiol 22:186–195. https://doi.org/10.1002/gepi.0163

The 1,000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073. https://doi.org/10.1038/nature09534

The 1,000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM et al. (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393

Thompson EA (1975) The estimation of pairwise relationships. Ann Hum Genet 39:173–188

Wakefield J (2010) Bayesian methods for examining Hardy-Weinberg equilibrium. Biometrics 66:257–265. https://doi.org/10.1111/j.1541-0420.2009.01267.x

Weinberg W (1908) On the demonstration of heredity in man. In: Boyer SH (ed.) Papers on human genetics, Prentice Hall, Englewood Cliffs, NJ. Translated, 1963, p 4–15