Bayesian Tests for Random Mating in Polyploids

David Gerard

Department of Mathematics and Statistics, American University, Washington, DC, 20016, USA

Abstract

Hardy-Weinberg proportions (HWP) are often explored to evaluate the assumption of random mating. However, in autopolyploids, organisms with more than two sets of homologous chromosomes, HWP and random mating are different hypotheses that require different statistical testing approaches. Currently, the only available methods to test for random mating in autopolyploids (i) heavily rely on asymptotic approximations and (ii) assume genotypes are known, ignoring genotype uncertainty. Furthermore, these approaches are all frequentist, and so do not carry the benefits of Bayesian analysis, including ease of interpretability, incorporation of prior information, and consistency under the null. Here, we present Bayesian approaches to test for random mating, bringing the benefits of Bayesian analysis to this problem. Our Bayesian methods also (i) do not rely on asymptotic approximations, being appropriate for small sample sizes, and (ii) optionally account for genotype uncertainty via genotype likelihoods. We validate our methods in simulations, and demonstrate on two real datasets how testing for random mating is more useful for detecting genotyping errors than testing for HWP (in a natural population) and testing for Mendelian segregation (in an experimental S1 population). Serendipitously, we further show that our tests for random mating in autopolyploids may also be applied as tests for HWP in allopolyploids. Our methods are implemented in Version 2.0.2 of the `hwep` R package on the Comprehensive R Archive Network https://cran.r-project.org/package=hwep.

1 Introduction

One reason researchers often test for Hardy-Weinberg proportions (HWP) is to evaluate the presence of random mating [Waples, 2015]. At autosomal loci in diploids, one or two generations of random mating, for monoecious and dioecious species, respectively, are sufficient to reach HWP [Gillespie, 2004]. Thus, for diploids, random mating and HWP are effectively equivalent hypotheses, and so use the same statistical procedures. However, in autopolyploids, organisms with more than two sets of homologous chromosomes, HWP is only approached asymptotically [Haldane, 1930], and so random mating and HWP are separate hypotheses. To put notation to these hypotheses, let $\boldsymbol{q} = (q_0, q_1, \ldots, q_K)$ be the genotype frequencies of a K-ploid population at a biallelic locus with allele frequency $r = \frac{1}{K}\sum_{k=0}^{K} kq_k$. That is, q_k is the proportion of the population with k copies of the minor allele. Then the HWP hypothesis states that, in the absence of double reduction [Huang et al., 2019], the genotype frequencies follow binomial proportions,

$$q_k = \binom{K}{k} r^k (1-r)^{K-k}. \tag{1}$$

While, under the hypothesis of random mating, we have

$$q_k = \sum_{i=\max(0,k-K/2)}^{\min(k,K/2)} p_i p_{k-i}, \tag{2}$$

where $\boldsymbol{p} = (p_0, p_1, \ldots, p_{K/2})$ are the *gamete* frequencies of the population. That is, p_k is the proportion of gametes in the population that have k copies of the minor allele. Equation (2) follows since, under random mating, genotypes are the sums of independent gametes, and so $\boldsymbol{q}$ is the discrete linear convolution of $\boldsymbol{p}$ with itself (2). We may concisely denote discrete linear convolution (2) by $\boldsymbol{q} = \boldsymbol{p} * \boldsymbol{p}$. Hypotheses (1) and (2) are clearly distinct, which indicates the necessity to create separate testing procedures to evaluate them.

Keywords and phrases: Bayes factors, genotyping errors, Hardy-Weinberg, polyploids, random mating.

Testing for binomial proportions (1) is a fairly straightforward statistical problem, and several researchers have developed permutation approaches to do so effectively, particularly for small counts [Hardy and Vekemans, 2002, Meirmans and Van Tienderen, 2004]. By contrast, the only approaches that test for random mating (2) heavily rely on asymptotic approximations [Sun et al., 2020, Gerard, 2022b]. (Though, Matoka Nana [2023] developed exact tests for random mating in autopolyploids, but these tests were found to be far too conservative for practical use). Such asymptotic approximations are particularly fraught in autopolyploids where many genotypes are possible, and sample sizes are typically small. Thus, in most applications, there will be genotypes with few/no individuals, limiting the applicability of asymptotic approaches. The current approaches also do not account for genotype uncertainty, a pervasive issue in polyploid data [Gerard et al., 2018, Gerard and Ferrão, 2019], and one that limits the real-world applicability of current methods.

Furthermore, the only approaches that test hypothesis (1) use frequentist procedures. Sun et al. [2020] developed a test for random mating in tetraploids ($K = 4$) that finds the MLEs under model (2) via an EM algorithm, and uses these estimated genotype frequencies to run a chi-squared test. Sun et al. [2020] called this a "A Gamete-Based aHWE [asymptotic Hardy-Weinberg equilibrium] Test", but this is incorrect as it is a test for random mating, not HWP. Sun et al. [2020] also uses an incorrect degrees of freedom for their test, and contains other issues [Gerard, 2022a]. Gerard [2022b] generalized this EM algorithm to every ploidy ($K \geq 4$) and implemented likelihood ratio tests for random mating for every ploidy. One issue with frequentist procedures is that, for increasing sample size, the p-value threshold should decrease to provide consistency under the null [Wakefield, 2009], but this is rarely done.

Bayesian approaches provide solutions to all of the above issues. Small sample sizes are not an issue for Bayesian methods as all inference is conditional on the observed data, not on hypothetical realizations of data which most frequentist approaches approximate via asymptotics. Bayesian methods provide a natural foundation to integrate over genotype uncertainty. And, whereas frequentist p-values should have lower significance levels for larger sample sizes so that both type I and type II error rates decrease to zero, Bayes factors automatically calibrate themselves with increasing sample size [Shoemaker et al., 1998, Wakefield, 2009]. The appeal of the Bayesian approach can

be evidenced by the large number of methods in the diploid literature that implement Bayesian tests for HWP [Altham, 1971, Pereira and Rogatko, 1984, Lindley, 1988, Shoemaker et al., 1998, Montoya-Delgado et al., 2001, Consonni et al., 2008, Wakefield, 2010, Consonni et al., 2011, Puig et al., 2017, 2019].

The goal of this paper is to produce and implement Bayesian tests for hypotheses

$$H_0 : \boldsymbol{q} = \boldsymbol{p} * \boldsymbol{p} \text{ for some } \boldsymbol{p} \text{ s.t. } \sum_{k=0}^{K/2} p_k = 1, \text{ versus} \tag{3}$$

$$H_1 : \boldsymbol{q} \neq \boldsymbol{p} * \boldsymbol{p}. \tag{4}$$

When genotypes are known, we implement a Bayesian test via an efficient Gibbs sampler that uses a novel data augmentation scheme (Section 2.1). However, genotypes are rarely known in polyploid organisms, and so we account for genotype uncertainty using genotype likelihoods (Section 2.2). Using simulations, we explore the behavior of our tests when genotypes are known (Section 3.1), before exploring the performance of our procedures under genotype uncertainty (Section 3.2). We demonstrate the utility of our methods on both a natural population of tetraploid white sturgeon (Section 3.3) and an experimental population of hexaploid sweet potatoes (Section 3.4), where we show that competing methods are less useful, particularly at detecting genotyping errors.

2 Materials and methods

2.1 Bayesian test with known genotypes

In this section, we will create a Bayesian test for random mating when genotypes are known. We will limit ourselves to summaries in the main text, leaving more technical details to Appendix S1. Let x_k be the number of individuals with genotype $k \in \{0, 1, \ldots, K\}$, which we collect into the vector $\boldsymbol{x} = (x_0, x_1, \ldots, x_K)$. Then, given genotype frequencies $\boldsymbol{q}$, we have that

$$[\boldsymbol{x}|\boldsymbol{q}] \sim \text{Multinom}(n, \boldsymbol{q}), \tag{5}$$

where $\text{Multinom}(n, \boldsymbol{q})$ denotes the multinomial distribution with size $n = \sum_{k=0}^{K} x_k$ and probability vector $\boldsymbol{q}$. Our goal is to use likelihood (5) to test hypotheses (3)–(4). We will do so by calculating a Bayes factor [Kass and Raftery, 1995], the ratio of the marginal likelihoods under the null and alternative,

$$\text{BF} = \frac{\Pr(\boldsymbol{x}|H_0)}{\Pr(\boldsymbol{x}|H_1)} = \frac{\int \text{Multinom}(\boldsymbol{x}|n, \boldsymbol{p} * \boldsymbol{p})\pi_0(\boldsymbol{p})\text{d}\boldsymbol{p}}{\int \text{Multinom}(\boldsymbol{x}|n, \boldsymbol{q})\pi_1(\boldsymbol{q})\text{d}\boldsymbol{q}}, \tag{6}$$

where $\pi_0(\cdot)$ and $\pi_1(\cdot)$ denote prior densities under the null and alternative, respectively, and $\text{Multinom}(\boldsymbol{x}|n, \boldsymbol{q})$ is the multinomial likelihood of genotype counts $\boldsymbol{x}$, size n, and probability vector $\boldsymbol{q}$. In this manuscript, we will assume that these prior densities are both Dirichlet (Section 2.3) with concentration hyperparameters $\boldsymbol{\alpha}$ under the null and $\boldsymbol{\beta}$ under the alternative,

$$\boldsymbol{p} \sim \text{Dirichlet}(\boldsymbol{\alpha}), \text{ and} \tag{7}$$

$$\boldsymbol{q} \sim \text{Dirichlet}(\boldsymbol{\beta}). \tag{8}$$

Smaller Bayes factors provide more evidence for H_1, and larger Bayes factors provide more evidence for H_0.

We will need to separately calculate the integrals in the denominator and numerator of (6). The denominator in (6) is a standard calculation. The integral,

$$\text{DM}(\boldsymbol{x}|n, \boldsymbol{\beta}) = \int \text{Multinom}(\boldsymbol{x}|n, \boldsymbol{q})\,\text{Dirichlet}(\boldsymbol{q}|\boldsymbol{\beta})\text{d}\boldsymbol{q}, \tag{9}$$

results in a Dirichlet-Multinomial distribution [Mosimann, 1962]. This has a known form (S1) and is easily computable.

To calculate the marginal likelihood in the numerator of (6), we will need a data augmentation scheme to represent the distribution of $\boldsymbol{x} \sim \text{Multinom}(n, \boldsymbol{p}*\boldsymbol{p})$, as this will allow us to use conjugacy to derive a closed form for $\Pr(\boldsymbol{x}|H_0)$ and then, for computational reasons, a Gibbs sampler to approximate the distribution of $[\boldsymbol{p}|\boldsymbol{x}]$. Let us begin by considering how random mating arises via the union of two randomly sampled, unobserved gametes. Let $\boldsymbol{y} = (y_0, y_1, \ldots, y_{K/2})$ be the

gamete counts, where y_j is the number of gametes in our sample with genotype j. Random mating implies that these gametes are selected with replacement according to the gamete frequencies, i.e. $\boldsymbol{y} \sim \text{Multinom}(2n, \boldsymbol{p})$. These gametes then randomly pair to form the next generation of offspring, and we denote the counts of pairings by the matrix $\boldsymbol{A} = (a_{ij})$, where a_{ij} is the number of individuals who were formed from gametes with dosages i and j with $i \leq j$. The number of offspring with genotype k, which we denote by x_k, is derived by summing over the cells of $\boldsymbol{A}$ such that $k = i + j$,

$$x_k = \sum_{i \text{ s.t. } i \leq k-i} a_{i,k-i}. \tag{10}$$

It is easy to show that $\boldsymbol{y}$ is also a function of $\boldsymbol{A}$,

$$y_k = \sum_{i \text{ s.t. } i<k} a_{ik} + 2a_{kk} + \sum_{j \text{ s.t. } j>k} a_{kj}. \tag{11}$$

See Figure 1 for a visualization of the connections between $\boldsymbol{A}$, $\boldsymbol{x}$, and $\boldsymbol{y}$.

As we only observe $\boldsymbol{x}$, our goal is to write the probability mass function (PMF) of $\boldsymbol{x}$ in terms of $\boldsymbol{A}$ and $\boldsymbol{y}$. We know the PMF of $[\boldsymbol{y}|\boldsymbol{p}]$ is $\text{Multinom}(\boldsymbol{y}|2n, \boldsymbol{p})$. Further, the PMF of $[\boldsymbol{A}|\boldsymbol{y}]$ is a known form (S6) [Levene, 1949], which we will denote by $f(\boldsymbol{A}|\boldsymbol{y})$. Also, since $\boldsymbol{y}$ is a function of $\boldsymbol{A}$ (11), which we denote by $\boldsymbol{y}(\boldsymbol{A})$, the joint PMF of $\boldsymbol{A}$ and $\boldsymbol{y}$ may be written as $f(\boldsymbol{A}|\boldsymbol{y}(\boldsymbol{A}))\,\text{Multinom}(\boldsymbol{y}(\boldsymbol{A})|2n, \boldsymbol{p})$. We may use this to marginalize over $\boldsymbol{A}$ to derive the following identity,

$$\text{Multinom}(\boldsymbol{x}|n, \boldsymbol{p} * \boldsymbol{p}) = \sum_{\boldsymbol{A} \in \mathcal{T}(\boldsymbol{x})} f(\boldsymbol{A}|\boldsymbol{y}(\boldsymbol{A}))\,\text{Multinom}(\boldsymbol{y}(\boldsymbol{A})|2n, \boldsymbol{p}), \tag{12}$$

where $\mathcal{T}(\boldsymbol{x})$ is the set of $\boldsymbol{A}$'s that are consistent with $\boldsymbol{x}$ (Figure 1).

Identity (12) allows us to derive the marginal likelihood under the null in closed form when using a Dirichlet prior over $\boldsymbol{p}$. Namely.

$$\Pr(\boldsymbol{x}|H_0) = \int \text{Multinom}(\boldsymbol{x}|n, \boldsymbol{p} * \boldsymbol{p})\,\text{Dirichlet}(\boldsymbol{p}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{p} \tag{13}$$

$$= \int \sum_{\boldsymbol{A} \in \mathcal{T}(\boldsymbol{x})} f(\boldsymbol{A}|\boldsymbol{y}(\boldsymbol{A}))\,\text{Multinom}(\boldsymbol{y}(\boldsymbol{A})|2n, \boldsymbol{p})\,\text{Dirichlet}(\boldsymbol{p}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{p} \tag{14}$$

$$= \sum_{\boldsymbol{A} \in \mathcal{T}(\boldsymbol{x})} f(\boldsymbol{A}|\boldsymbol{y}(\boldsymbol{A})) \int \text{Multinom}(\boldsymbol{y}(\boldsymbol{A})|2n, \boldsymbol{p}) \, \text{Dirichlet}(\boldsymbol{p}|\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{p} \tag{15}$$

$$= \sum_{\boldsymbol{A} \in \mathcal{T}(\boldsymbol{x})} f(\boldsymbol{A}|\boldsymbol{y}(\boldsymbol{A})) \, \text{DM}(\boldsymbol{y}(\boldsymbol{A})|\boldsymbol{\alpha}). \tag{16}$$

We describe in Appendix S1 how to sum over (16).

Summation (16) is efficient to calculate for tetraploids, and is possible to calculate for small-to-moderate sample sizes of hexaploids. However, for larger ploidies the summation (16) becomes too computationally intensive for all but the smallest of sample sizes. Therefore, in Appendix S2, we derive a Gibbs sampler [Gelfand and Smith, 1990] to sample from the posterior distribution of $[\boldsymbol{p}|\boldsymbol{x}]$ using the same data augmentation scheme as in Figure 1. In this Gibbs sampler, we iteratively sample $[\boldsymbol{A}|\boldsymbol{p}, \boldsymbol{x}]$ and $[\boldsymbol{p}|\boldsymbol{A}]$, and the resulting marginal samples of $\boldsymbol{p}$ are from the distribution of $[\boldsymbol{p}|\boldsymbol{x}]$. Standard calculations can be used to show that

$$[\boldsymbol{p}|\boldsymbol{A}] \sim \text{Dirichlet}(\boldsymbol{y}(\boldsymbol{A}) + \boldsymbol{\alpha}). \tag{17}$$

To sample $[\boldsymbol{A}|\boldsymbol{p}, \boldsymbol{x}]$ we note that the probability of a pair of gametes adding a count to a_{ij} is proportional to p_i^2 when $i = j$ and $2p_ip_j$ when $i \neq j$. This, combined with the constraint that $\boldsymbol{A} \in \mathcal{T}(\boldsymbol{x})$, allows us to sample from $[\boldsymbol{A}|\boldsymbol{p}, \boldsymbol{x}]$ (Procedure S1). The full Gibbs sampler is summarized in Procedure S2. After obtaining the posterior samples from $[\boldsymbol{p}|\boldsymbol{x}]$, we can estimate the marginal likelihood by the method of Chib [1995].

2.2 Bayesian test using genotype likelihoods

Due to the higher number of possible dosages, polyploids are more susceptible than diploids to data variability and data biases, and thus exhibit higher amounts of genotype uncertainty [Gerard et al., 2018, Gerard and Ferrão, 2019, Gerard, 2021a,b]. A common way to account for this is through the use of genotype likelihoods [Li, 2011]. A nice feature of Bayesian methodology is its ability to easily adapt to additional variability by integrating over it. Here, we develop Bayesian testing procedures for random mating that account for genotype uncertainty using genotype likelihoods.

It is possible to modify our Gibbs sampler from Section 2.1 to account for genotype likelihoods

via a data augmentation scheme where we introduce latent genotypes that we sample over (Appendix S3). This approach works well for small ploidies and small sample sizes. However, we have found that this approach produces poor mixing (i.e., it inefficiently explores the posterior) when both the ploidy K and the sample size n is large (Figures S7–S9), and thus requires a large number of samples to produce accurate results. We have therefore opted to use the Stan programming language [Stan Development Team, 2022a,b], which uses Hamiltonian Monte Carlo, to more efficiently explore the posterior distribution in fewer samples by using gradient information [see Betancourt, 2018, for an accessible introduction]. Marginal likelihoods and Bayes factors may then be calculated via bridge sampling [Meng and Wong, 1996, Gronau et al., 2017, 2020].

Let us discuss the model implementation. Let ℓ_{ik} be the genotype likelihood for individual $i = 1, 2, \ldots, n$ for genotype $k = 0, 1, \ldots, K$. That is, ℓ_{ik} is the probability of the data (sequencing, microarray, or otherwise) for individual i given that the genotype for that individual is k. Genotype likelihoods are provided by many genotyping programs [McKenna et al., 2010, Voorrips et al., 2011, Serang et al., 2012, Gerard et al., 2018, Clark et al., 2019, Gerard and Ferrão, 2019]. Then, given these genotype likelihoods, we have the likelihood for these data is

$$f(\boldsymbol{L}|\boldsymbol{q}) = \prod_{i=1}^{n} \sum_{k=0}^{K} \ell_{ik} q_k. \tag{18}$$

Under the alternative, the prior over $\boldsymbol{q}$ is (8), while under the null, the prior over $\boldsymbol{q} = \boldsymbol{p} * \boldsymbol{p}$ is induced from (7). We run two Markov chains, one using prior (8) and one using prior (7), estimating the marginal likelihoods of each via bridge sampling, and from there calculating the Bayes factor. These implementations in Stan are relatively straightforward so we omit the specific details. The Stan files may be found in the source code for our `hwep` R package (`https://cran.r-project.org/package=hwep`).

2.3 Prior elicitation

Here, we describe our default priors. For convenience, as in Sections 2.1 and 2.2, we assume that the gamete frequencies are *a priori* Dirichlet($\boldsymbol{\alpha}$) under the null, and the genotype frequencies are *a priori* Dirichlet($\boldsymbol{\beta}$) under the alternative, and our focus will be on selecting appropriate values of

$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

To motivate our choice for $\boldsymbol{\alpha}$, we note that at HWP, in the absence of double reduction, we have for autopolyploids that the p_i's are binomial proportions for some allele frequency r [Haldane, 1930]. That is,

$$p_i = \binom{K/2}{i} r^i (1-r)^{K/2-i}. \tag{19}$$

For natural populations, it is reasonable to place a uniform prior over r. Matching moments with the Dirichlet, we set α_i to be proportional to the expected value of (19) when r is uniformly distributed. We have that

$$\alpha_i \propto \int_0^1 \binom{K/2}{i} r^i (1-r)^{K/2-i} \mathrm{d}r = \frac{2}{K+2}, \tag{20}$$

where the calculation in (20) uses the kernel of the Beta density. Thus, each α_k should have the same value *a priori*, and this motivates us to take

$$\boldsymbol{\alpha} = \mathbf{1}_{K/2+1}. \tag{21}$$

That is, $\boldsymbol{p}$ is uniformly distributed over the $K/2$ unit simplex. This might appear to be just the naive prior one would choose as "uninformative", but we have also motivated it by placing an uninformative prior on the allele frequency of a randomly mating natural population at HWP. Though we have motivated it using HWP, it is important to note that our methods *do not* assume HWP.

We choose our prior for $\boldsymbol{\beta}$ conditional on our choice of $\boldsymbol{\alpha}$. Our elicitation of $\boldsymbol{\beta}$ makes the following key assumption: *samples of size one are completely uninformative on the status of random mating.* This approach is conceptually similar to that of Smith and Spiegelhalter [1980] and Spiegelhalter and Smith [1982], except they chose their datasets that are most favorable to the null (since they worked in a continuous data paradigm) while we choose all possible datasets where $n = 1$. To specify this prior, let $\boldsymbol{e}_k$ be a 1-of-$(K+1)$ vector that is 1 at position k and is 0 elsewhere (so $\boldsymbol{e}_k$ is

the value of $\boldsymbol{x}$ when $n = 1$), then our assumption is that

$$\mathrm{BF}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{\Pr(\boldsymbol{e}_k | H_0, \boldsymbol{\alpha})}{\Pr(\boldsymbol{e}_k | H_1, \boldsymbol{\beta})} = 1 \text{ for all } k = 0, 1, \ldots, K. \tag{22}$$

We can show that, given $\boldsymbol{\alpha} = \mathbf{1}_{K/2+1}$, for any choice of scaling factor $c > 0$, the following choice of $\boldsymbol{\beta}$ will result in property (22),

$$\beta_k = c \left[\min \left(\left\lfloor \frac{k}{2} \right\rfloor, \left\lfloor \frac{K-k}{2} \right\rfloor \right) + 1 \right]. \tag{23}$$

Equation (23) is merely a constant c multiplied by the number of elements in $\boldsymbol{A}$ that correspond to each x_k (Lemma S1). We prove this result in Corollary S1 in Appendix S4.

By default, we choose a value of c such that the total sum of $\boldsymbol{\beta}$ is the same as $\mathbf{1}_{K+1}$. This intuitively means that we place the same overall weight on the prior as a uniform distribution would. When $\boldsymbol{\alpha} = \mathbf{1}_{K/2+1}$, this results in default values of $\boldsymbol{\beta}$,

$$\beta_k = \frac{(K+1)\left(\min[\lfloor k/2 \rfloor, \lfloor (K-k)/2 \rfloor] + 1\right)}{\binom{K/2+2}{K/2}}. \tag{24}$$

The default values of $\boldsymbol{\beta}$ for even ploidies less than or equal to 12 are presented in Table S1.

It is important to note that priors can impact Bayes factors for even large sample sizes [Kass and Raftery, 1995], and so choosing reasonable default priors is extremely important. However, because our priors are proper, have support over the entire parameter space, and the models are nested, the Bayes factors are guaranteed to be consistent for large sample sizes [O'Hagan, 1994, Section 7.52]. That is, the Bayes factor goes to infinity if the null is true and goes to zero if the alternative is true. Thus, for large sample sizes we will at least make the correct statistical decision, even if the Bayes factors do not accurately describe the odds of the null to the alternative that are given by the data. We evaluate prior sensitivity in Appendix S6.

2.4 HWP tests in allopolyploids

Though this paper has focused on autopolyploids, our tests for random mating in autopolyploids are also tests for HWP in allopolyploids. Since each subgenome of an allopolyploid is diploid, allopolyploids reach HWP after just one or two generations of random mating. This indicates that, in allopolyploids, tests for random mating are essentially tests for HWP. Furthermore, when allopolyploids exhibit random mating, they still produce genotype frequencies of the form $\boldsymbol{q} = \boldsymbol{p} * \boldsymbol{p}$, the null hypothesis for our tests for random mating. However, whereas in Section 2.3 we elicited our prior for $\boldsymbol{p}$ by considering a natural population of autopolyploids at HWP, for allopolyploids we should update this prior by considering a natural population of allopolyploids at HWP. We derive our default priors when analyzing allopolyploid genomes in Appendix S4.1.

3 Results

3.1 Performance when genotypes are known

We can verify that the Bayes factors increase under the null, and decrease under the alternative, as the sample size increases. To do this, we simulated genotype counts for ploidies $K \in \{4, 6, 8\}$ with $n \in \{10, 100, 1000\}$ individuals under two scenarios. The first scenario was where the null was satisfied with $\boldsymbol{p} = \mathbf{1}_{K/2+1}/(K/2+1)$ and $\boldsymbol{q} = \boldsymbol{p} * \boldsymbol{p}$, and the second scenario was where the alternative was satisfied with $\boldsymbol{q} = \mathbf{1}_{K+1}/(K+1)$. For each scenario, we generated genotype counts via $\boldsymbol{x} \sim \text{Multinom}(n, \boldsymbol{q})$. Each replication, we obtained the Bayes factor from Section 2.1 and the p-value from Appendix C of Gerard [2022b] (a likelihood ratio test for random mating). For each combination of parameters, we replicated these simulations 1000 times.

The results are in Figure 2. There, we see that the Bayes factor is consistent under both the alternative and null scenarios. We also plot the computation time in Figure S3 under these scenarios. We see that, when using simulation, the Bayesian approach has about the same computation time for any ploidy and sample size. The Bayesian test (which is $\mathcal{O}(1)$) is even slightly faster than the likelihood ratio test from Appendix C of Gerard [2022b], but both are still fast enough for genome-wide applications. Perhaps the biggest disadvantage of the likelihood ratio test is that the p-values,

under the null, are not uniformly distributed for small sample sizes (Figure 3) because of its reliance on asymptotics. The Bayesian approach does not depend on asymptotics and so may be used at any sample size.

In Appendix S5, we consider the behavior of our tests when the sample size is very small. We show that the range of possible Bayes factors is limited when samples are ~ 5. At such sample sizes, only moderate evidence is possible either for or against the random mating hypothesis. Of course, this is a property of the data and this limitation also applies to frequentist approaches.

To evaluate the sensitivity of our methods to prior selection, we ran a prior sensitivity analysis in Appendix S6. There, we confirm that the Bayes factors are consistent for various prior selections, and we demonstrate that the Bayes factors are relatively stable to moderate deviations from our default priors. This sensitivity analysis also includes our default priors for autopolyploids (Section 2.3) and allopolyploids (Section 2.4), both of which perform similarly under the simulation conditions studied.

3.2 Performance under genotype uncertainty

To explore the performance of our approaches under genotype uncertainty, we incorporated genotype likelihoods into the simulation study from Section 3.1. For each replication, using the same simulation settings as in Section 3.1, we simulated read counts using `rflexdog()` from the `updog` R package [Gerard et al., 2018, Gerard and Ferrão, 2019] at a read depth of 10, a sequencing error rate of 0.01, an overdispersion parameter of 0.01, and no allelic bias. From these read counts, we then used `updog` (with the general categorical prior class) to generate genotype likelihoods and genotype posteriors. We used these genotype likelihoods to run the Bayesian test from Section 2.2. To compare this approach to that of Section 2.1, we need an estimate of $\boldsymbol{x}$. We did this by two methods: (i) summing up the posterior probabilities for each genotype across individuals and then rounding these sums to the nearest integers, and (ii) obtaining the posterior mode genotype for each individual and tabulating the genotype counts.

The results are presented in Figure S4. We see there that only the genotype likelihood approach from Section 2.2 is consistent, with the Bayes factor increasing with the sample size when the null

is satisfied. Assuming known genotypes, the method of Section 2.1 appears to be anti-conservative in the presence of genotype uncertainty, generally providing Bayes factors that are too small, when using either approach to generate $\boldsymbol{x}$. This indicates the need to account for genotype uncertainty in this testing situation to provide accurately calibrated test statistics.

To evaluate the sensitivity of our methods to the calibration of the genotype likelihoods, we altered the conditions of `updog` when genotyping the simulated read-counts. We either estimated the likelihood parameters (sequencing error rate, overdispersion, and allelic bias) or fixed them at their known values. When estimating the likelihood parameters, the prior class can greatly affect the estimates [Gerard and Ferrão, 2019] (the prior class has no effect on the genotype likelihoods when the likelihood parameters are known), and so when estimating these parameters we either used the general categorical prior class, the proportional normal prior class [Gerard and Ferrão, 2019], or the discrete uniform prior. The ideal scenario is using known likelihood parameters. The typical scenarios are using either the proportional normal prior or the general categorical prior and estimating the likelihood parameters. The most poorly calibrated scenario is using the discrete uniform prior while estimating the likelihood parameters [Gerard et al., 2018]. The results are in Figure S5. There, we see that all approaches to generating the genotype likelihoods yield consistent results, indicating that our methods are very robust to genotype likelihood miscalibration. Though, using the uniform prior or the general categorical prior class does perform worse.

We also evaluated the Gibbs sampler from Appendix S3 in these simulations. The number of iterations in the Gibbs sampler was chosen to have the same run-time as the Stan approach when $K = 8$ and $n = 1000$ (Figure S6). The Bayes factors calculated using the Gibbs sampler and those calculated using Stan are mostly the same except when both the sample size and ploidy are large (Figures S7–S8). The differences are due to poor mixing in the Gibbs sampler, as this approach is not consistent for $K = 8$ (Figure S9), at least given the number of sampling iterations used. This demonstrates the benefits of using Stan, at least for large sample sizes and large ploidies, to improve mixing.

3.3 A natural population

In this section, we analyze the tetraploid white sturgeon (*Acipenser transmontanus*) data from Delomas et al. [2021]. These data were collected from a central California caviar farm with broodstock from the Sacramento river, and so likely exhibit random mating. We genotyped the $n = 19$ individuals at 325 loci using `updog` [Gerard et al., 2018, Gerard and Ferrão, 2019]. These data were sequenced to a very high average read depth of 5342, and so it is reasonable to take genotypes at most loci as known. We ran our Bayesian test (Section 2.1) and the likelihood ratio test for random mating from Gerard [2022b] on these loci.

Figure 4 contains a histogram of the log Bayes factors. We see there that the mode of the log Bayes factors is above 0, with most log Bayes factors near or slightly above zero. This provides strong evidence that the population is undergoing random mating. There are just three loci that appear to deviate from this trend (Table S3).

The test for random mating immediately reveals its use as, looking deeper at the three loci with the smallest Bayes factors, it appears that all three are a result of genotyping errors. Let us take, as an example, the SNP with the lowest Bayes factor, Atr_20529-52. When we fit `updog` on these data, there are two modes of the log-likelihood (Figure 5). The maximum likelihood estimate results in $\boldsymbol{x} = (0, 6, 2, 1, 10)$ (log-likelihood of -90) while a smaller local mode results in $\boldsymbol{x} = (0, 0, 6, 3, 10)$ (log-likelihood of -105). The local mode was found by starting the algorithm of `updog` at a different value of the bias parameter. This local mode estimates the allelic bias (the ratio of mapping probabilities of the minor and major alleles) to be 2.2, which is large but not too out of the ordinary for these data (Figure S13). Interestingly, when we run our Bayesian test for random mating on this local mode with $\boldsymbol{x} = (0, 0, 6, 3, 10)$, we get a log Bayes factor of 1.1, which is in alignment with the rest of these data (Figure 4). Thus, it is likely that this SNP was a genotyping error. When we do the same analysis with the SNPs with the next two smallest Bayes factors, Atr_33123-54 and Atr_26310-63, we also see that the global modes are likely genotyping errors and local modes likely provide the correct genotype estimates (Figures S10–S11). This demonstrates the practical usefulness of our approach.

The likelihood ratio test for random mating from Gerard [2022b] produces log p-values that

are linearly associated with the log Bayes factors, particularly in the tail (Figure S14). However, standard frequentist approaches would indicate that no loci exhibit strong deviations from random mating. For example, the smallest three q-values [Storey, 2003] are 0.06, 0.17, and 0.17, which many researchers would not raise an eyebrow to since they are all above 0.05.

Tests for HWP are less useful for these data. The p-values, using the U-statistic based test from Gerard [2022b], are highly non-uniform (Figure S12). Indeed, using these p-values, the `qvalue` package [Storey, 2003] estimates that 53% of loci do not satisfy the HWP hypothesis, with 14 loci being rejected at the 0.05 FDR level. This indicates that, though these data exhibit random mating, there are some severe deviations from HWP, likely caused by unobserved relatedness between individuals. This limits the usefulness of this approach for detecting genotyping errors.

3.4 An S1 population

Here, we analyze the S1 population of 142 autohexaploid sweet potatoes (*Ipomoea batatas*) from Shirasawa et al. [2017]. An S1 population consists of a single generation of selfing from the same parent. It might seem strange, but random mating is fulfilled in S1 populations because (i) all offspring gametes are drawn independently and identically distributed from the distribution of gametes produced by the parent, and (ii) these gametes are randomly paired during fertilization. Though, random mating would be violated at loci that affect fitness (e.g., inbreeding depression), and tests for random mating would be one way to detect such cases.

Data were prepared and filtered as in Gerard [2022b]. That is, SNPs were selected that had allele frequencies between 0.1 and 0.9, average read depths of at least 100, and no more than 50% missing data. This resulted in 5074 SNPs. Genotype likelihoods were obtained by `updog` [Gerard et al., 2018, Gerard and Ferrão, 2019]. We used these genotype likelihoods to fit the Bayesian test from Section 2.2. The numbers of individuals with each dosage were estimated by summing posterior posterior genotype probabilities across individuals and then rounding these sums. We used these estimated counts to fit the Bayesian test from Section 2.1 on these loci.

The Bayes factors calculated using either genotype likelihoods (Section 2.2) or assuming known genotypes (Section 2.1) are only moderately correlated (Figure S15), indicating that there is enough

genotype uncertainty that we should only use the genotype likelihood approach for these data (Section 3.2). The log Bayes factors for these loci are almost all positive (Figure S16) indicating strong evidence of random mating, as it should for these data since random mating is indeed fulfilled.

We explored the 5 SNPs with log Bayes factors less than -5. Genotype plots for the `updog` fits of these SNPs are provided in Figure S17. `Updog` has its own measure of genotyping performance, the estimated proportion of individuals mis-genotyped, and yet all but one of these SNPs have reasonable values of this measure (Table S4), so would likely pass filters using it. The numbers of individuals with each genotype deviate strongly from that expected under Mendelian segregation (Table S4), indicating that these are, indeed, problematic SNPs.

Since this is an S1 population, it is possible to use this knowledge to specifically test for genotype frequencies that would result from Mendelian segregation. We could then compare this approach with our more general test for random mating. We created a Bayesian test for segregation distortion in S1 populations in Appendix S7 and applied this test to the sweet potato data. There was significant agreement on which SNPs satisfy the random mating assumption (Table S5). For SNPs that support the random mating hypothesis, the S1 test tends to produce larger Bayes factors, indicating greater discriminative ability, as expected (Figure S18). However, the S1 test provides many more SNPs that support the alternative hypothesis. Genotype plots of a random sample of these SNPs, where the S1 test indicates violations in the null and the random mating test indicates no violations, are presented in Figure S19. These SNPs appear to be relatively well behaved, except for two or three individuals in each SNP, where Mendelian segregation would predict those genotypes to be impossible. Indeed, those problematic individuals are the reasons why the S1 test is so keen on the alternative hypothesis, and if we remove them then the S1 test produces positive log Bayes factors (Table S6). If the goal of a test is to find SNPs where *any* individual might be mis-genotyped, then the S1 test might be appropriate. But it is very sensitive to minor violations in the genotype frequencies predicted by an S1 population, and so might be too finicky for applied work. By contrast, the random mating test only detects broad genotyping issues. We note, though, that these problematic individuals could possibly be a result of double reduction [Stift et al., 2010], in which case a more sophisticated method to test for S1 genotype frequencies in the presence of

double reduction might produce better results.

4 Discussion

In this manuscript, we developed Bayesian tests for random mating in polyploids. Our methods are applicable both when genotypes are known and when there is high genotype uncertainty (e.g., due to low sequencing depth) using genotype likelihoods. The hypothesis of random mating is less strict than that of HWP in autopolyploids, and is equivalent to HWP in allopolyploids. Therefore, regardless of the ploidy type, one should prefer tests for random mating over those for autopolyploid HWP. We proposed using tests for random mating to detect genotyping errors, as is common practice with HWP tests in diploids [Hosking et al., 2004, Anderson et al., 2010, Winkler et al., 2014]. We demonstrated this application on two real datasets, in which the SNPs broadly follow random mating but not HWP, by detecting likely genotyping errors that were missed by other methods (including current frequentist approaches). Our Bayesian approaches are applicable for both small sample sizes (as they do not depend on asymptotics) and large sample sizes (as they are guaranteed to be consistent decision procedures).

In addition to genotyping errors, random mating can be violated through various mechanisms, such as selection, inbreeding, and population structure. Both inbreeding and population structure can lead to widespread deviations from random mating across many SNPs, similar to the effects of inbreeding [Vieira et al., 2013] and population structure [Hao and Storey, 2019] on diploid populations, which cause broad deviations from HWP. On the other hand, selection, similar to genotyping errors, can result in localized deviations from random mating. Tests for random mating can be valuable in detecting any of these issues, just as tests for HWP are useful in detecting them in diploids [Waples, 2015].

Our methods assume that loci are biallelic. This is the most common scenario in practice, particularly in polyploid studies where reduced representation genotyping is often used [Baird et al., 2008, Elshire et al., 2011]. Extending random mating tests to multiallelic loci would greatly increase the complexity of the problem, and so we leave it for future work. However, we can detail here how to mathematically represent the hypothesis of random mating at a multiallelic locus, and

discuss how one would approach testing for random mating. Consider a locus with $m+1$ alleles, $A_0, A_1, \ldots, A_m$, in a K-ploid species. We collect the gamete frequencies for alleles $A_1, A_2, \ldots, A_m$ into a m-dimensional array $\mathcal{P} = (p_{\boldsymbol{k}}) \in \mathbb{R}^{(K/2+1)\times(K/2+1)\times\cdots\times(K/2+1)}$, where for the multi-index $\boldsymbol{k} = (k_1, k_2, \ldots, k_m)$, $p_{\boldsymbol{k}}$ is the proportion of gametes with k_1 copies of A_1, k_2 copies of A_2, $\ldots$, k_m copies of A_m, and, thus, $K/2 - \sum_{j=1}^m k_j$ copies of A_0. Each k_j ranges from $0, 1, \ldots K/2$. Because each gamete has $K/2$ genome copies, this implies that $p_{\boldsymbol{k}} = 0$ when $\sum_{j=1}^m k_j > K/2$. Let $\mathcal{Q} = (q_{\boldsymbol{k}}) \in \mathbb{R}^{(K+1)\times(K+1)\times\cdots(K+1)}$ be the m-dimensional array of genotype frequencies, defined similarly as $\mathcal{P}$ but with each k_j ranging from $0, 1, \ldots, K$. Then random mating implies that

$$\mathcal{Q} = \mathcal{P} \overset{m}{*} \mathcal{P}, \text{ or} \tag{25}$$

$$q_{\boldsymbol{k}} = \sum_{\boldsymbol{i}} p_{\boldsymbol{i}} p_{\boldsymbol{k}-\boldsymbol{i}} \tag{26}$$

where $\overset{m}{*}$ denotes m-dimensional discrete linear convolution [Rakhuba and Oseledets, 2015], represented in (26) where the multi-index $\boldsymbol{i}$ sums over $0 \leq i_1, i_2, \ldots, i_m \leq K/2$. For example, suppose we have a tri-allelic locus in a tetraploid population, then

$$\mathcal{P} = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & 0 \\ p_{20} & 0 & 0 \end{pmatrix}, \tag{27}$$

where p_{ij} is the frequency of gametes with i copies of A_1, j copies of A_2, and $2-i-j$ copies of A_0. Under random mating, we have

$$\mathcal{Q} = \mathcal{P} \overset{m}{*} \mathcal{P} = \begin{pmatrix} p_{00}^2 & 2p_{00}p_{01} & p_{01}^2 + 2p_{00}p_{02} & 2p_{01}p_{02} & p_{02}^2 \\ 2p_{00}p_{10} & 2p_{00}p_{11} + 2p_{01}p_{10} & 2p_{01}p_{11} + 2p_{02}p_{10} & 2p_{02}p_{11} & 0 \\ p_{10}^2 + 2p_{00}p_{20} & 2p_{10}p_{11} + 2p_{01}p_{20} & p_{11}^2 + 2p_{02}p_{20} & 0 & 0 \\ 2p_{10}p_{20} & 2p_{11}p_{20} & 0 & 0 & 0 \\ p_{20}^2 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{28}$$

Clearly, testing (25) is much more complicated than testing (3). It might be easiest to avoid a data

augmentation scheme like in Section 2.1 and implement a test for (25) using Stan. However, since (25) contains many more parameters than (3) one would have to be much more cautious about prior sensitivity, as any developed Bayes factor might not be as robust as our methods (Appendix S6).

In Section 2.3 we motivated our prior over $\boldsymbol{p}$ by considering a natural population at equilibrium, in the absence of double reduction. There, we assumed that such a natural population would have a uniform allele frequency distribution. The distribution of allele frequencies is called the "site frequency spectrum" and there has been significant work on deriving its properties [Fu, 1995, Griffiths and Tavaré, 1998, Wooding and Rogers, 2002, Polanski et al., 2003, Polanski and Kimmel, 2003, Birkner et al., 2013]. The distribution of the site frequency spectrum is a complicated function of population size and various aspects of data collection [Polanski and Kimmel, 2003]. If this information is known about a population, it might be possible to incorporate the theoretical expectations of the sight frequency spectrum by modifying the expectation in (20), and thereby improve the prior elicitation. One would have to be careful about ascertainment bias caused, e.g., by using variants called from a previous sample, or the standard practice of filtering out SNPs with small minor allele frequencies. We do not explore these ideas further because our results in Appendix S6 indicate that any improvements from modifying the prior would be modest.

The random mating hypothesis could be used to improve genotyping methods. Many genotyping programs implement empirical Bayes approaches that estimate the genotype frequencies from a class of prior distributions before implementing a full Bayesian approach to genotyping [Voorrips et al., 2011, Serang et al., 2012, Blischak et al., 2018, Gerard et al., 2018, Clark et al., 2019, Gerard and Ferrão, 2019]. The choice of the class of prior distribution can significantly improve or harm the genotype calling performance [Gerard and Ferrão, 2019]. Our analysis in Section 3.3 suggests that we could improve genotyping performance by using (2) as the class of prior distributions, and doing so might have corrected the three genotyping errors we detected. Our methods here are intended to be agnostic to the genotyping method used, and so we leave implementing (2) as a prior class for genotyping as future work.

Most researchers are not used to the Bayesian approach to hypothesis testing, and so a common question is how to choose the threshold of the Bayes factor before one concludes violations in the

random mating hypothesis? The canonical solution [Section 4.4.3 of Berger, 2013, e.g.] is for a researcher to elicit prior probabilities π_0 and $(1 - \pi_0)$ for the null and alternative hypotheses, respectively, and losses L_I and L_{II} for type I and type II errors, respectively. One would then reject H_0 if the Bayes factor is less than $(1 - \pi_0)L_{II}/(\pi_0 L_I)$. Eliciting these quantities can be difficult for most researchers. So our recommendations would be to (i) plot the Bayes factors and see if any loci are anomalous, (ii) in settings with few loci (~ 100) use the thresholds from Kass and Raftery [1995] as a guide, and (iii) in GWAS settings with many loci use the threshold from Wakefield [2010] as a guide. For (ii) this would correspond to a log Bayes factor being less than -3 indicating strong evidence, and a log Bayes factor less than -5 indicating very strong evidence for violations of random mating. For (iii) this would correspond to a log Bayes factor less than -16 indicating very strong evidence for violations of random mating.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2132247.

Most analyses were performed using the R statistical language [R Core Team, 2022].

References

S. Aldor-Noiman, L. D. Brown, A. Buja, W. Rolke, and R. A. Stine. The power to see: A new graphical test of normality. *The American Statistician*, 67(4):249–260, 2013. doi: 10.1080/00031305.2013.847865.

P. M. E. Altham. Exact Bayesian analysis of an intraclass 2×2 table. *Biometrika*, 58(3):679–680, 1971. doi: 10.1093/biomet/58.3.679.

C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010. doi: 10.1038/nprot.2010.116.

N. A. Baird, P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLOS ONE*, 3(10):1–7, 2008. doi: 10.1371/journal.pone.0003376.

J. Berger. *Statistical Decision Theory: Foundations, Concepts, and Methods*. Springer Series in Statistics. Springer New York, 2013. ISBN 9781475717273. URL `https://books.google.com/books?id=TT_jBwAAQBAJ`.

M. Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv*, 2018. doi: 10.48550/arXiv.1701.02434. preprint.

M. Birkner, J. Blath, and B. Eldon. Statistical properties of the site-frequency spectrum associated with Λ-coalescents. *Genetics*, 195(3):1037–1053, 2013. doi: 10.1534/genetics.113.156612.

P. D. Blischak, L. S. Kubatko, and A. D. Wolfe. SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics*, 34(3):407–415, 2018. doi: 10.1093/bioinformatics/btx587.

S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995. doi: 10.1080/01621459.1995.10476635.

L. V. Clark, A. E. Lipka, and E. J. Sacks. polyRAD: Genotype calling with uncertainty from sequencing data in polyploids and diploids. *G3: Genes |Genomes |Genetics*, 9(3):663–673, 2019. doi: 10.1534/g3.118.200913.

G. Consonni, E. Gutiérrez-Peña, and P. Veronese. Compatible priors for Bayesian model comparison with an application to the Hardy-Weinberg equilibrium model. *Test*, 17(3):585–605, 2008. doi: 10.1007/s11749-007-0057-7.

G. Consonni, E. Moreno, and S. Venturini. Testing Hardy–Weinberg equilibrium: An objective Bayesian analysis. *Statistics in Medicine*, 30(1):62–74, 2011. doi: 10.1002/sim.4084.

T. A. Delomas, S. C. Willis, B. L. Parker, D. Miller, P. Anders, A. Schreier, and S. Narum. Genotyping single nucleotide polymorphisms and inferring ploidy by amplicon sequencing for polyploid, ploidy-variable organisms. *Molecular Ecology Resources*, 21(7):2288–2298, 2021. doi: 10.1111/1755-0998.13431.

R. J. Elshire, J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLOS ONE*, 6(5):1–10, 2011. doi: 10.1371/journal.pone.0019379.

Y.-X. Fu. Statistical properties of segregating sites. *Theoretical population biology*, 48(2):172–197, 1995. doi: 10.1006/tpbi.1995.1025.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990. doi: 10.1080/01621459.1990.10476213.

D. Gerard. Pairwise linkage disequilibrium estimation for polyploids. *Molecular Ecology Resources*, 21(4):1230–1242, 2021a. doi: 10.1111/1755-0998.13349.

D. Gerard. Scalable bias-corrected linkage disequilibrium estimation under genotype uncertainty. *Heredity*, 127(4):357–362, 2021b. doi: 10.1038/s41437-021-00462-5.

D. Gerard. Comment on three papers about Hardy–Weinberg equilibrium tests in autopolyploids. *Frontiers in Genetics*, 13, 2022a. doi: 10.3389/fgene.2022.1027209.

D. Gerard. Double reduction estimation and equilibrium tests in natural autopolyploid populations. *Biometrics*, 2022b. doi: 10.1111/biom.13722. (in press).

D. Gerard and L. F. V. Ferrão. Priors for genotyping polyploids. *Bioinformatics*, 36(6):1795–1800, 2019. doi: 10.1093/bioinformatics/btz852.

D. Gerard, L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens. Genotyping polyploids from messy sequencing data. *Genetics*, 210(3):789–807, 2018. doi: 10.1534/genetics.118.301468.

J. Gillespie. *Population Genetics: A Concise Guide*. Population Genetics. Johns Hopkins University Press, 2004. ISBN 9781421401706. URL `https://books.google.com/books?id=KAcAfiyHpcoC`.

R. C. Griffiths and S. Tavaré. The age of a mutation in a general coalescent tree. *Stochastic Models*, 14(1-2):273–295, 1998. doi: 10.1080/15326349808807471.

Q. F. Gronau, A. Sarafoglou, D. Matzke, A. Ly, U. Boehm, M. Marsman, D. S. Leslie, J. J. Forster, E.-J. Wagenmakers, and H. Steingroever. A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97, 2017. doi: 10.1016/j.jmp.2017.09.005.

Q. F. Gronau, H. Singmann, and E.-J. Wagenmakers. bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10):1–29, 2020. doi: 10.18637/jss.v092.i10.

J. Haldane. Theoretical genetics of autopolyploids. *Journal of Genetics*, 22(3):359–372, 1930. doi: 10.1007/BF02984197.

W. Hao and J. D. Storey. Extending tests of Hardy–Weinberg equilibrium to structured populations. *Genetics*, 213(3):759–770, 2019. doi: 10.1534/genetics.119.302370.

O. J. Hardy and X. Vekemans. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, 2(4):618–620, 2002. doi: 10.1046/j.1471-8286.2002.00305.x.

L. Hosking, S. Lumsden, K. Lewis, A. Yeo, L. McCarthy, A. Bansal, J. Riley, I. Purvis, and C.-F. Xu. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics*, 12(5):395–399, 2004. doi: 10.1038/sj.ejhg.5201164.

K. Huang, T. Wang, D. W. Dunn, P. Zhang, X. Cao, R. Liu, and B. Li. Genotypic frequencies at equilibrium for polysomic inheritance under double-reduction. *G3: Genes |Genomes |Genetics*, 9(5):1693–1706, 2019. doi: 10.1534/g3.119.400132.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90 (430):773–795, 1995. doi: 10.1080/01621459.1995.10476572.

H. Levene. On a matching problem arising in genetics. *The Annals of Mathematical Statistics*, 20 (1):91–94, 1949. doi: 10.1214/aoms/1177730093.

H. Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987, 2011. doi: 10.1093/bioinformatics/btr509.

D. Lindley. Statistical inference concerning Hardy-Weinberg equilibrium. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, editors, *Bayesian Statistics 3: Proceedings of the Third Valencia International Meeting, June 1-5, 1987*, pages 307–326. Oxford University Press, 1988. ISBN 9780198522201. URL https://books.google.com/books?id=2s4x-ZeMTkIC.

K. Matoka Nana. Exact tests for random mating in autotetraploids. Master's thesis, American University, 2023. Advisor: David Gerard.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. DePristo. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, 2010. doi: 10.1101/gr.107524.110.

P. G. Meirmans and P. H. Van Tienderen. GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, 4(4):792–794, 2004. doi: 10.1111/j.1471-8286.2004.00770.x.

X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6(4):831–860, 1996. doi: 10.2307/24306045.

L. E. Montoya-Delgado, T. Z. Irony, C. A. de B. Pereira, and M. R. Whittle. An unconditional exact test for the Hardy-Weinberg equilibrium law: Sample-space ordering using the Bayes factor. *Genetics*, 158(2):875–883, 2001. doi: 10.1093/genetics/158.2.875.

J. E. Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82, 1962. doi: 10.2307/2333468.

A. O'Hagan. *Kendall's Advanced Theory of Statistics, Volume 2B, Bayesian Inference*. Oxford University Press, 1994. ISBN 0340529229. URL https://books.google.com/books?id=ZWa7cQAACAAJ.

C. Pereira and A. Rogatko. The Hardy-Weinberg equilibrium under a Bayesian perspective. *Revista Brasileira de Genética*, 7(4):689–707, 1984.

A. Polanski and M. Kimmel. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. *Genetics*, 165(1): 427–436, 2003. doi: 10.1093/genetics/165.1.427.

A. Polanski, A. Bobrowski, and M. Kimmel. A note on distributions of times to coalescence, under time-dependent population size. *Theoretical population biology*, 63(1):33–40, 2003. doi: 10.1016/S0040-5809(02)00010-2.

X. Puig, J. Ginebra, and J. Graffelman. A Bayesian test for Hardy-Weinberg equilibrium of biallelic X-chromosomal markers. *Heredity*, 119(4):226–236, 2017. doi: 10.1038/hdy.2017.30.

X. Puig, J. Ginebra, and J. Graffelman. Bayesian model selection for the study of Hardy-Weinberg proportions and homogeneity of gender allele frequencies. *Heredity*, 123(5):549–564, 2019. doi: 10.1038/s41437-019-0232-0.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

M. V. Rakhuba and I. V. Oseledets. Fast multidimensional convolution in low-rank tensor formats via cross approximation. *SIAM Journal on Scientific Computing*, 37(2):A565–A582, 2015. doi: 10.1137/140958529.

O. Serang, M. Mollinari, and A. A. F. Garcia. Efficient exact maximum a posteriori computation for Bayesian SNP genotyping in polyploids. *PLOS ONE*, 7(2):1–13, 2012. doi: 10.1371/journal.pone.0030906.

K. Shirasawa, M. Tanaka, Y. Takahata, D. Ma, Q. Cao, Q. Liu, H. Zhai, S.-S. Kwak, J. C. Jeong, U.-H. Yoon, H.-U. Lee, H. Hirakawa, and S. Isobe. A high-density SNP genetic map consisting of a complete set of homologous groups in autohexaploid sweetpotato (*Ipomoea batatas*). *Scientific Reports*, 7, 2017. doi: 10.1038/srep44207.

J. Shoemaker, I. Painter, and B. S. Weir. A Bayesian characterization of Hardy-Weinberg disequilibrium. *Genetics*, 149(4):2079–2088, 1998. doi: 10.1093/genetics/149.4.2079.

A. F. Smith and D. J. Spiegelhalter. Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):213–220, 1980. doi: 10.1111/j.2517-6161.1980.tb01122.x.

D. J. Spiegelhalter and A. F. Smith. Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3):377–387, 1982. doi: 10.1111/j.2517-6161.1982.tb01217.x.

Stan Development Team. *RStan: the R interface to Stan*, 2022a. URL `https://mc-stan.org/`. R package version 2.21.5.

Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual*, 2022b. URL `https://mc-stan.org`. Version 2.29.

M. Stift, R. Reeve, and P. H. Van Tienderen. Inheritance in tetraploid yeast revisited: segregation patterns and statistical power under different inheritance models. *Journal of Evolutionary Biology*, 23(7):1570–1578, 2010. doi: 10.1111/j.1420-9101.2010.02012.x.

J. D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. doi: 10.1214/aos/1074290335.

L. Sun, J. Gan, L. Jiang, and R. Wu. Recursive test of Hardy-Weinberg equilibrium in tetraploids. *Trends in Genetics*, 2020. doi: 10.1016/j.tig.2020.11.006.

F. G. Vieira, M. Fumagalli, A. Albrechtsen, and R. Nielsen. Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. *Genome research*, 23(11): 1852–1861, 2013. doi: 10.1101/gr.157388.113.

R. E. Voorrips, G. Gort, and B. Vosman. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics*, 12(1):172, 2011. doi: 10.1186/1471-2105-12-172.

J. Wakefield. Bayes factors for genome-wide association studies: comparison with p-values. *Genetic Epidemiology*, 33(1):79–86, 2009. doi: 10.1002/gepi.20359.

J. Wakefield. Bayesian methods for examining Hardy–Weinberg equilibrium. *Biometrics*, 66(1): 257–265, 2010. doi: 10.1111/j.1541-0420.2009.01267.x.

R. S. Waples. Testing for Hardy-Weinberg proportions: have we lost the plot? *Journal of Heredity*, 106(1):1–19, 2015. doi: 10.1093/jhered/esu062.

T. W. Winkler, F. R. Day, D. C. Croteau-Chonka, A. R. Wood, A. E. Locke, R. Mägi, T. Ferreira, T. Fall, M. Graff, A. E. Justice, et al. Quality control and conduct of genome-wide association meta-analyses. *Nature protocols*, 9(5):1192–1212, 2014. doi: 10.1038/nprot.2014.071.

S. Wooding and A. Rogers. The matrix coalescent and an application to human single-nucleotide polymorphisms. *Genetics*, 161(4):1641–1650, 2002. doi: 10.1093/genetics/161.4.1641.

Data accessibility

All methods discussed in this manuscript are implemented Version 2.0.2 of the `hwep` R package on the Comprehensive R Archive Network:

https://cran.r-project.org/package=hwep.

Scripts to reproduce all of the results of this manuscript are available on Zenodo:

https://doi.org/10.5281/zenodo.6993722.

All datasets used in this manuscript are open and available at:

https://www.doi.org/10.5061/dryad.crjdfn33r, and

http://sweetpotato-garden.kazusa.or.jp/

Supplementary material

Additional details, figures, and tables are available in the Supplementary Material online.

Author contributions

DG developed the methodology, wrote the software, implemented the study, and wrote the manuscript.

5 Figures

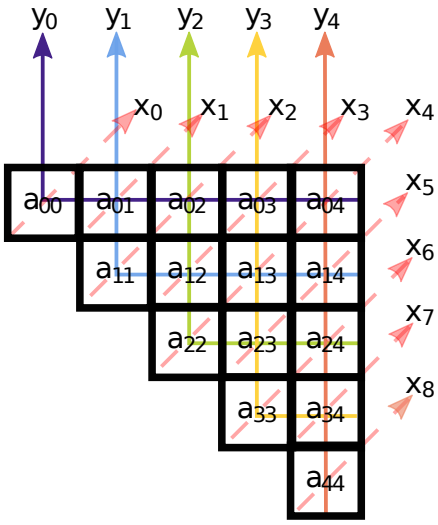

Figure 1: Visualization of data augmentation from (10) and (11) for an octoploid species ($K = 8$). The elements of $\boldsymbol{A}$ are in each cell. The x_k's are the sums of the anti-diagonal elements, indicated by the dashed red lines. The y_k's are the sums of the color-coded solid lines. We sample $[\boldsymbol{A}|\boldsymbol{x}, \boldsymbol{p}]$ by sampling each anti-diagonal by a multinomial with size x_k and probabilities proportional to the cell probabilities of the a_{ij}'s (S25). The sampled $\boldsymbol{y}$ is then calculated from this $\boldsymbol{A}$.

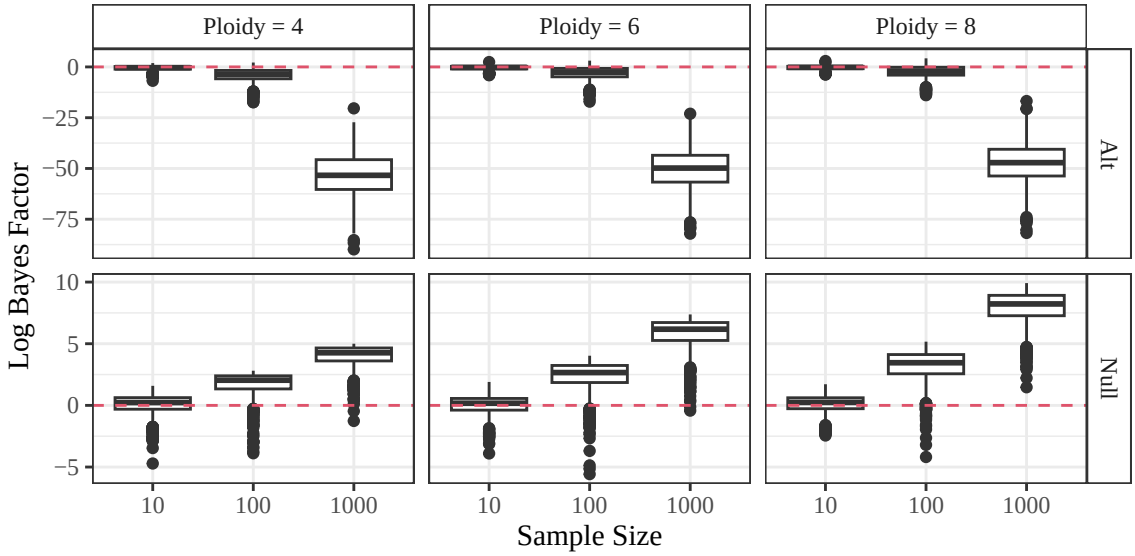

Figure 2: Log Bayes factors (y-axis) at different sample sizes (x-axis) at different ploidies (column facets) when either the null is true ("Null") or the alternative is true ("Alt"). The Bayes factor increases gradually with sample size when the null is true and decreases rapidly with sample size when the alternative is true.

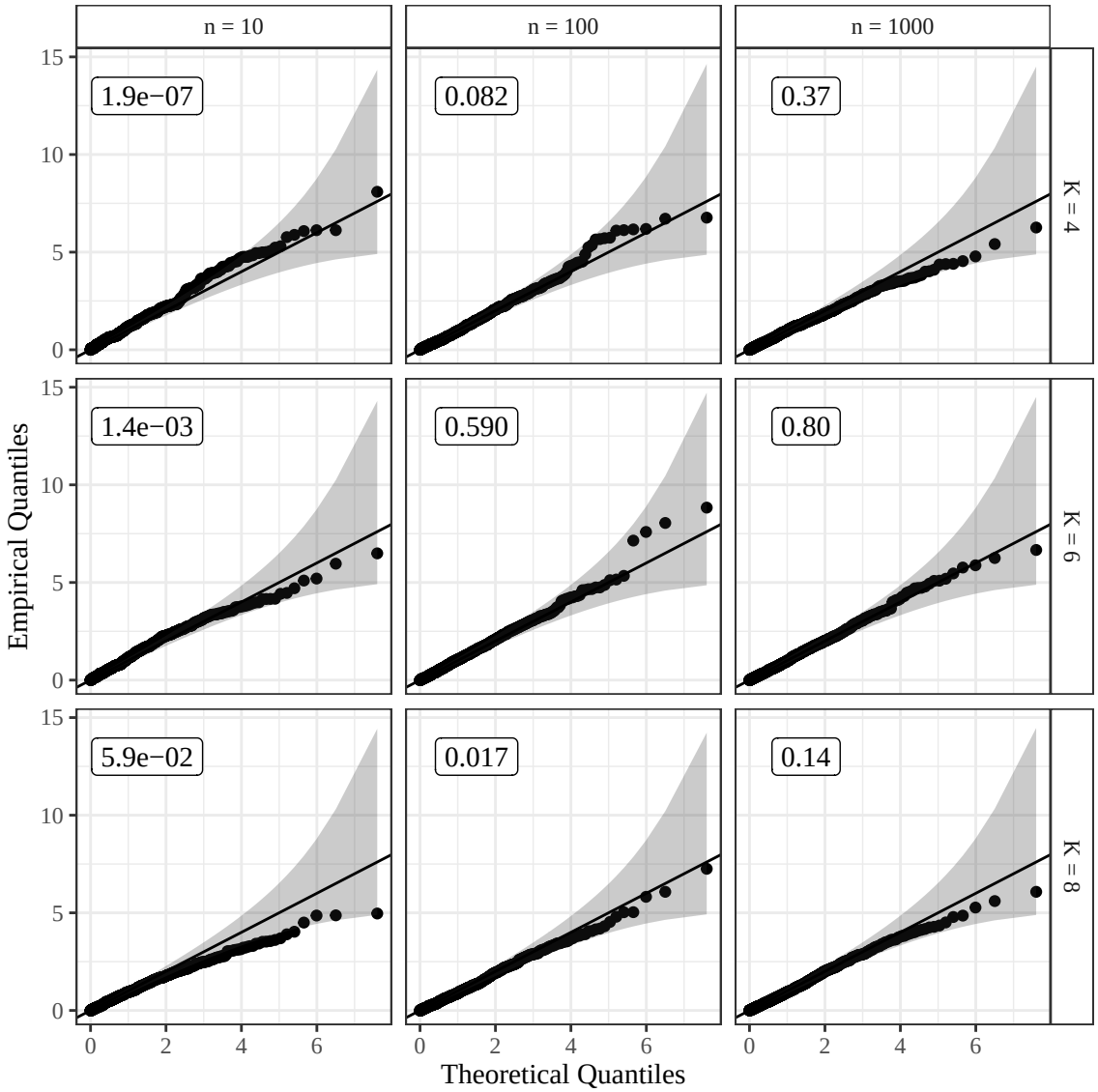

Figure 3: QQ-plots of p-values against the uniform distribution, on the $-\log$ scale, with the 95% simultaneous confidence bands from Aldor-Noiman et al. [2013], stratified by ploidy (row facets) and sample size (column facets). These p-values come from the simulations of Section 3.1 when the null is satisfied, and so should follow a uniform distribution. The numbers in the top-left of each plot are the p-values from a Kolmogorov-Smirnov test against the null of uniformity. Only for larger sample sizes are the p-values appropriately calibrated.

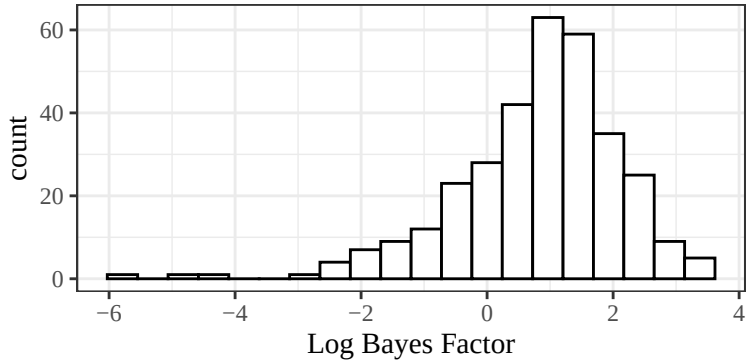

Figure 4: Histogram of log Bayes factors applied to all 325 loci from the white sturgeon data from Section 3.3. We see that most log Bayes factors are near or slightly above zero, indicating strong evidence that the population is undergoing random mating.

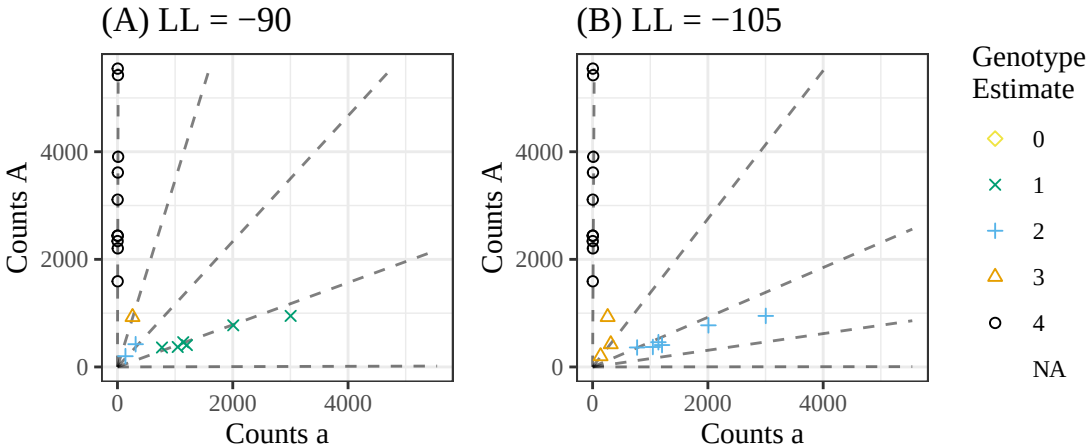

Figure 5: Genotype plots [Gerard et al., 2018] for two fits on SNP Atr_20529-52 from the white sturgeon data of Section 3.3. Each point is an individual, the color and shape of which indicate the estimated genotype under the two fits. We have the number of alternative reads on the x-axis and the number of reference reads on the y-axis. The lines radiating from the origin indicate the expected number of reference/alternative reads of a given genotype under that given fit. The left plot **(A)** is the global mode with a log-likelihood of -90, and the right plot **(B)** is a local mode with a log-likelihood of -105. The local mode exhibits high allelic bias, which can be seen by the lines clustering towards the x-axis. The test for random mating using the global mode of $\boldsymbol{x} = (0, 6, 2, 1, 10)$ yields a log Bayes factor of -5.9, while the test for random mating using the local mode of $\boldsymbol{x} = (0, 0, 6, 3, 10)$ yields a log Bayes factor of 1.1. The global mode is thus likely a genotyping error.