Double reduction estimation and equilibrium tests in natural autopolyploid populations

David Gerard

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

Abstract

Many bioinformatics pipelines include tests for equilibrium. Tests for diploids are well studied and widely available, but extending these approaches to autopolyploids is hampered by the presence of double reduction, the co-migration of sister chromatid segments into the same gamete during meiosis. Though a hindrance for equilibrium tests, double reduction rates are quantities of interest in their own right, as they provide insights about the meiotic behavior of autopolyploid organisms. Here, we develop procedures to (i) test for equilibrium while accounting for double reduction, and (ii) estimate double reduction given equilibrium. To do so, we take two approaches: a likelihood approach, and a novel U -statistic minimization approach that we show generalizes the classical equilibrium χ^2 test in diploids. For small sample sizes and uncertain genotypes, we further develop a bootstrap procedure based on our U -statistic to test for equilibrium. Finally, we highlight the difficulty in distinguishing between random mating and equilibrium in tetraploids at biallelic loci. Our methods are implemented in the `hwep` R package on GitHub <https://github.com/dcgerard/hwep>.

1 Introduction

After its discovery, the Hardy-Weinberg law [Hardy, 1908, Weinberg, 1908] emerged as a foundational assumption in much of population genetics [Crow, 1988]. Since the implementation of the classical χ^2 and likelihood ratio tests [Weir, 1996], the literature on testing the Hardy-Weinberg law in diploids has continued to grow [e.g. Levene, 1949, Haldane, 1954, Guo and Thompson, 1992, Wakefield, 2010, Graffelman and Weir, 2016, 2018, for a small sample]. The research in polyploids is much more limited. In allopolyploids, if data from the different homoeologous genomes can be distinguished, then diploid approaches may be used. However, the research in equilibrium tests for autopolyploids is obstructed by the presence of double reduction, the co-migration of sister chromatid segments to the same gamete [Mather, 1935]. Double reduction increases homozygosity [Hardy, 2016] (Figure 1), introducing an inflation of Type I error into tests that ignore it (Section 3.1). It would therefore benefit the field to develop test procedures that account for double reduction.

Estimating double reduction is also of interest in its own right due to its role in evolution [Meir-mans et al., 2018], QTL mapping [Li et al., 2010], and plant breeding [Butruille and Boiteux, 2000], among other tasks. Most methods that estimate double reduction use experimental populations

Keywords and phrases: double reduction, equilibrium, Hardy-Weinberg, polyploid, random mating, U -statistics.

(such as F1 populations). Since experimental populations only require modeling a finite number of meioses for each offspring, this makes modeling much simpler. Using experimental populations, researchers have estimated double reduction either using segregation frequencies [Tai, 1982a,b, Haynes and Douches, 1993, Wu et al., 2001, Stift et al., 2008, Bourke et al., 2015], or they obtain estimates as a byproduct of haplotype reconstruction [Zheng et al., 2016, 2021]. For natural populations, the only manuscript we could find that claims to estimate double reduction is that of Jiang et al. [2021]. However, there is a vital flaw in the assumptions of this paper (Section S1), which results in biased estimates of double reduction (Section 3.1). In brief, Jiang et al. [2021] do not model equilibrium, but only one generation of possible double reduction.

If the equilibrium genotype frequencies, given double reduction, could be derived, then estimation and testing procedures would become feasible. Various manuscripts derive genotype frequencies at equilibrium in autopolyploids, but most without the full generality required to perform inference. Haldane [1930] provide genotype frequencies at equilibrium when there is no double reduction, but possible self-fertilization. Geiringer [1949] and Bennett [1968] provide genotype frequencies at equilibrium for tetraploids and hexaploids in terms of the rate of double reduction and the allele frequency of the population, but do not provide estimation or testing procedures, nor provide these frequencies at higher ploidy levels.

In a useful paper, Huang et al. [2019] layout a strategy for deriving genotype frequencies at equilibrium. One step in this strategy is to solve a complex system of nonlinear equations, a generically difficult task. Solving this system is possible for tetraploids (Section S3), and even for ploidies 6, 8, and 10 when using modern symbolic algebra systems (Section S4). But this approach becomes computationally infeasible for higher ploidies. Furthermore, the resulting equilibrium frequencies, because they are rather involved, provide no insight on any estimation and testing procedures that would use them, making this approach difficult to generalize to other models of meiosis or other evolutionary forces. And though Huang et al. [2019] derive these frequencies, they provide no means to test for equilibrium nor estimate double reduction given equilibrium.

In this manuscript, we will tackle equilibrium testing and double reduction estimation from multiple angles. In Section 2.3 we use the genotype frequencies derived by Huang et al. [2019] to implement a likelihood framework to test for equilibrium and estimate double reduction given equilibrium. However, because of the complexity of these genotype frequencies and the difficulty in obtaining such frequencies for larger ploidies, we develop a different approach for testing in Section 2.4. In this approach, we find a statistic that has expectation zero at equilibrium given the rate of double reduction. This motivates minimizing a norm of this statistic to make it as close to 0 as possible, thereby providing an estimate of the double reduction rate. We show that this minimized norm has an asymptotic χ^2 distribution given equilibrium, which can be used to test for equilibrium. This approach generalizes the classical χ^2 test for equilibrium in diploids to polyploids (Section S8). Finally, to account for small sample sizes and genotype uncertainty, in Section 2.5 we present a bootstrap approach to test for equilibrium.

There also exists a point of confusion in the literature on the difference between random mating and equilibrium. For example, Sun et al. [2020] develop a likelihood framework that they claim tests for equilibrium in tetraploids, but we would say actually tests for random mating (Section 2.2). In brief, we propose that random mating implies that genotype frequencies of the current generation are a convolution of the genotype frequencies of the gametes created by the parental generation, while equilibrium implies a steady state of genotype frequencies given a meiotic process and other population forces. We clarify this distinction in Section 2.1, and show that distinguishing between these hypotheses is rather subtle, particularly for tetraploids. The random mating hypothesis might

be of interest, and so in Section 2.2 we generalize the approach of Sun et al. [2020] to higher ploidy levels.

2 Methods

2.1 Random mating versus equilibrium in tetraploids

In this section, we will highlight in tetraploids the difficulty of distinguishing between random mating and equilibrium while accounting for double reduction. In effect, our results show that it is impossible for any method to test for small deviations from equilibrium in the presence of double reduction using just a biallelic locus, but testing for moderate to large deviations from equilibrium is possible.

Genotype frequencies are structured under random mating. Let $\mathbf{q} = (q_0, \dots, q_K)^\top$ denote the genotype frequencies of the $K + 1$ genotypes at a biallelic locus of a K -ploid species, where K is even. Consider the “gamete frequencies”, $\mathbf{p} = (p_0, \dots, p_{K/2})^\top$, where p_i is the probability that a gamete has dosage i . Since individual genotypes are sums of gamete genotypes, random mating can be represented mathematically by

$$\mathbf{q} = \mathbf{p} * \mathbf{p}, \text{ i.e.} \\ q_\ell = \sum_{\substack{\min(\ell, K/2) \\ \max(0, \ell - K/2)}} p_i p_{\ell - i}. \quad (1)$$

where “*” denotes discrete linear convolution. For example, in the case of diploids, we have $\mathbf{p} = (1 - r, r)$ for some allele frequency r , and $\mathbf{q} = ((1 - r)^2, 2r(1 - r), r^2)^\top$, which is the well-known Hardy-Weinberg law.

On the other hand, equilibrium is a phenomenon that occurs after successive generations. To define equilibrium, we must first posit some process of updating genotype frequencies given the current genotype frequencies. Let $f(\mathbf{q}, \boldsymbol{\alpha})$ be the updated genotype frequencies in the next generation given a rate of double reduction $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{\lfloor K/4 \rfloor})^\top$, where α_i is the frequency that an individual contains i pairs of identical-by-double-reduction (IBDR) alleles (Section S5). Equilibrium occurs when the updated genotype frequencies are unchanged from one generation to the next, that is

$$\mathbf{q} = f(\mathbf{q}, \boldsymbol{\alpha}). \quad (2)$$

This update, $f(\cdot, \cdot)$, depends on the model for meiosis and other population genetic forces, random mating being one such possible force. The process we will use in this paper uses the same assumptions as used in deriving Hardy-Weinberg equilibrium in diploids [Waples, 2015]: (i) random mating, (ii) no selection, migration, gene flow, admixture, or mutation, (iii) only sexual selection, (iv) an infinite population size, (v) generations are non-overlapping, and (vi) allele frequencies are equal between sexes. The only generalization that we allow is double reduction.

As defined here, random mating is not the same as equilibrium. As noted by many authors, in polyploids it takes multiple generations of random mating to reach equilibrium [Haldane, 1930]. One clarifying example is that in an S1 population (one generation of selfing), the random mating hypothesis is satisfied, since the distribution of offspring genotypes is a convolution of the gamete probabilities within a single individual, but hardly any researcher would claim that the population is at equilibrium. However, Sun et al. [2020] conflated these hypotheses: they use null hypothesis

(1) in tetraploids and claim that their resulting test is one against the null of equilibrium, when theirs in actuality is a test against the null of random mating.

Actually distinguishing between the random mating hypothesis and the equilibrium hypothesis is quite subtle. In tetraploids, there are two parameters under the random mating hypothesis (since $p_0 + p_1 + p_2 = 1$), and there are two parameters under the equilibrium hypothesis (the double reduction rate, α , and the allele frequency, r). Thus, at equilibrium, there is a one-to-one correspondence between these two parameterizations (Section S3). In terms of $\mathbf{p}$, we have

$$r = \frac{1}{2}(p_1 + 2p_2), \text{ and} \quad (3)$$

$$\alpha = \frac{(1 - 2p_1) - (p_2 - p_0)^2}{(1 + p_1) - (p_2 - p_0)^2}, \quad (4)$$

and in terms of (α, r) , we have

$$p_0 = \left(1 - 2\frac{1 - \alpha}{2 + \alpha}r\right)(1 - r), \quad (5)$$

$$p_1 = \frac{4(1 - \alpha)}{2 + \alpha}r(1 - r), \text{ and} \quad (6)$$

$$p_2 = \left(\frac{3\alpha}{2 + \alpha} + 2\frac{1 - \alpha}{2 + \alpha}r\right)r. \quad (7)$$

Because of this one-to-one correspondence, if we had random mating but *not* equilibrium, then the gamete frequencies $\mathbf{p}$ would correspond to *some* (α, r) at equilibrium, but we would not be able to tell if these are the actual values at equilibrium. Stated another way, because of this one-to-one correspondence, there are no left over degrees of freedom to distinguish between random mating and equilibrium.

Though there is a one-to-one correspondence between $\mathbf{p}$ and (α, r) , not all values of $\mathbf{p}$ result in theoretically sound values of α . Other than constraining α to be non-negative, it is well known that under different theoretical models of meiosis the rate of double reduction is bounded above. E.g., under a complete equational segregation model, the maximum rate of double reduction is 1/6 [Mather, 1935]. If under equilibrium we have $0 \leq \alpha \leq c$ for some $c \in [0, 1]$, then this corresponds to the following gamete frequencies (derivation in Section S6).

$$0 \leq \alpha \leq c \Leftrightarrow 1 - \frac{2 + c}{1 - c}p_1 \leq (p_2 - p_0)^2 \leq 1 - 2p_1. \quad (8)$$

We can use this fact to test for extreme deviations from equilibrium, by testing if the gamete frequencies satisfy (8). However, even if the gamete frequencies satisfy (8), we are still unable to distinguish between random mating and equilibrium. It is only possible to test for moderate to major deviations from equilibrium.

A visualization of the region of gamete frequencies consistent with equilibrium is presented in Figure 2. If the gamete frequencies lie within this region, then they are consistent with *some* equilibrium, but we are not able to claim that they are *at* equilibrium. A summary of the nested hypotheses that can be tested at biallelic loci is presented in Table 1. We develop likelihood ratio tests to distinguish between these hypotheses in Section 2.3.

An example of the difficulty in testing for equilibrium is graphically displayed in Figure 3, where

we plot gamete frequencies along multiple generations of random mating. Generation 1 is an S1 population where the parental genotype was 2. The true level of α is 1/6. However, each iteration yields genotype frequencies that satisfy the random mating assumption (1), and thus correspond to *some* level of α that would maintain the gamete frequencies at their current levels. Only in generation 1 would this level of α be unreasonable, while the genotype frequencies are close to equilibrium only at around generation 5.

2.2 Likelihood inference for random mating

In this section, we will generalize the likelihood framework of Sun et al. [2020] from tetraploids to general ploidy levels. Though Sun et al. [2020] claim that their likelihood approach is a test for equilibrium, as we have shown in Section 2.1, this is really a test for random mating. However, testing this hypothesis might be useful in its own right to clarify the mating behavior in a system.

To define a test against the null of random mating, let $\mathbf{y} = (y_0, \dots, y_K)^\top$ denote the observed number of individuals with each genotype, so y_i is the number of the individuals with genotype i . Let $n = \sum_{i=0}^K y_i$ denote the total number of individuals. Then the likelihood is

$$\mathcal{L}(\mathbf{q}) = \text{Multinom}(\mathbf{y}|n, \mathbf{q}), \quad (9)$$

where $\text{Multinom}(\mathbf{y}|n, \mathbf{q})$ denotes the probability mass function of the multinomial distribution with size n and probability vector $\mathbf{q} = (q_0, \dots, q_K)^\top$, evaluated at $\mathbf{y}$. Under the unconstrained model (where the population may or may not exhibit random mating), the maximum likelihood estimate of $\mathbf{q}$ is $\hat{\mathbf{q}} = \mathbf{y}/n$. Under the constrained model where genotype frequencies are a result of random mating (1), we developed an expectation-maximization algorithm [Dempster et al., 1977] to maximize (9) using constraint (1) to obtain $\hat{\mathbf{p}}$ (Section S2). Under the null hypothesis of (1), the likelihood ratio test statistic,

$$-2[\log \text{Multinom}(\mathbf{y}|n, \hat{\mathbf{p}} * \hat{\mathbf{p}}) - \log \text{Multinom}(\mathbf{y}|n, \hat{\mathbf{q}})], \quad (10)$$

follows an asymptotic $\chi^2_{K/2}$ distribution, which we may use to calculate a p -value against the null of random mating.

2.3 Likelihood inference for equilibrium

In order to develop a likelihood framework to test against the null of equilibrium, we need to have access to the genotype frequencies at equilibrium given our meiotic process. Huang et al. [2019] developed a strategy to derive these genotype frequencies by setting up a system of non-linear equations. This system of equations is possible to solve in tetraploids (Section S3), and even in ploidies 6, 8, and 10 using modern symbolic algebra systems (Section S4). We will use these genotype frequencies to develop a likelihood ratio test for equilibrium. In tetraploids, this test is actually one against the null of at most minor deviations from equilibrium (Section 2.1).

Given a rate of double reduction α and an allele frequency r , we have access to the gamete frequencies at equilibrium $\mathbf{p}$. We will denote these gamete frequencies by $\mathbf{p}(\alpha, r)$ to emphasize their dependence on α and r . We can thus maximize the following likelihood (9) over $0 \leq \alpha_i \leq c_i$ and r to obtain the MLEs,

$$(\hat{\alpha}, \hat{r}) = \arg \max_{\alpha, r} \text{Multinom}(\mathbf{y}|n, \mathbf{p}(\alpha, r) * \mathbf{p}(\alpha, r)). \quad (11)$$

Then the likelihood ratio test statistic against the null of equilibrium is

$$-2[\log \text{Multinom}(\mathbf{y}|n, \mathbf{p}(\hat{\alpha}, \hat{r}) * \mathbf{p}(\hat{\alpha}, \hat{r})) - \log \text{Multinom}(\mathbf{y}|n, \hat{\mathbf{q}})]. \quad (12)$$

Equation (12) follows an asymptotic $\chi^2_{K-\lfloor K/4 \rfloor - 1}$ distribution under the null.

If one wanted to compare the null hypothesis of equilibrium against the alternative of random mating, one could calculate the following test statistic,

$$-2[\log \text{Multinom}(\mathbf{y}|n, \mathbf{p}(\hat{\alpha}, \hat{r}) * \mathbf{p}(\hat{\alpha}, \hat{r})) - \log \text{Multinom}(\mathbf{y}|n, \hat{\mathbf{p}} * \hat{\mathbf{p}})], \quad (13)$$

where $\hat{\mathbf{p}}$ is calculated as in Section S2. The likelihood ratio test statistic (13) can be compared to a $\chi^2_{K/2-\lfloor K/4 \rfloor - 1}$ distribution, for ploidies greater than 4, to obtain a p -value. For tetraploids, we use a χ^2_1 distribution, which is the asymptotic distribution of the test statistic under a fixed null value of $\alpha = 0$ or $\alpha = c$, a worst-case scenario for the test. Again, the null for tetraploids is one of only small deviations from equilibrium.

We can test against the null of no double reduction, given the population is in equilibrium. When the population is in equilibrium, but there is no double reduction, the genotypes follow a Binomial distribution with size K and success probability r [Haldane, 1930]. The MLE of r under the null is

$$\hat{r} = \frac{1}{Kn} \sum_{\ell=0}^K \ell y_{\ell}. \quad (14)$$

Let $\hat{\mathbf{q}}_0$ be the MLE of genotype frequencies at equilibrium given no double reduction, that is

$$\hat{q}_{0i} = \binom{K}{i} \hat{r}^i (1 - \hat{r})^{K-i}. \quad (15)$$

The test statistic is the following log-likelihood ratio

$$-2[\log \text{Multinom}(\mathbf{y}|n, \hat{\mathbf{q}}_0) - \log \text{Multinom}(\mathbf{y}|n, \mathbf{p}(\hat{\alpha}, \hat{r}) * \mathbf{p}(\hat{\alpha}, \hat{r}))], \quad (16)$$

which under the null follows an asymptotic $\chi^2_{\lfloor K/4 \rfloor}$ distribution, which may be used to calculate a p -value.

The asymptotics for all of these tests can be rather poor as many genotypes might exist with few or no individuals. To improve the asymptotic approximation (at the risk of losing power), we apply an aggregation strategy. Specifically, we aggregate genotypes that have no individuals, adjusting the resulting degrees of freedom of the χ^2 distributions accordingly.

Though we were able to use the genotype frequencies at equilibrium, these formulas are rather complicated, sometimes spanning pages. They thus provide no insight behind the methods used to test for equilibrium. This makes it difficult to generalize the likelihood approach to other models of meiosis or population forces. Furthermore, for higher ploidies, the system of non-linear equations becomes computationally infeasible, at least given the author's computational resources, and so these methods are difficult to generalize to higher ploidy individuals. All of this motivates the development of an alternative procedure that is more intuitive. We will present such an approach in Section 2.4.

2.4 A U -statistic approach to estimate double reduction and test for equilibrium given a meiotic process

In this Section we will develop a novel approach to test for equilibrium and to estimate double reduction given equilibrium, implemented for every ploidy level. We derive a statistic, called a “ U -statistic” [Hoeffding, 1948], that has expectation zero at equilibrium given the rate of double reduction. To obtain an estimate of this rate, we minimize a norm of the U -statistic over the double reduction rate. We can show that, at equilibrium, this minimized norm has an asymptotic χ^2 distribution, which we can use to obtain a p -value for a test against the null of equilibrium. This method is a generalization to polyploids, in the presence of double reduction, of the classical χ^2 test for equilibrium in diploids (Section S8). Formal details of our method are presented in Section S7. Here, we will provide a more intuitive description.

Let us begin by deriving a statistic that has expectation zero at equilibrium. Let $Pr(g|G_1, G_2, \theta)$ be the probability of two parents with genotypes $G_1, G_2 \in \{0, 1, \dots, K\}$ producing an offspring with genotype $g \in \{0, 1, \dots, K\}$. This probability depends on the model of meiosis, and we denote the parameters of this meiotic model by θ . For this manuscript, we use the model for meiosis in Section S5, and so $\theta = \alpha$, the double reduction rate. But our methods in this section would be applicable to different models of meiosis. Let x_i, x_j , and x_ℓ be the genotypes for individuals i, j , and ℓ . Consider the following statistic

$$h_g(\theta, x_i, x_j, x_\ell) := 1(x_i = g) - Pr(g|G_1 = x_j, G_2 = x_\ell, \theta), \quad (17)$$

where $1(\cdot)$ is the indicator function. The idea behind (17) is that we are choosing two individuals (j and ℓ) to act as the “parents” to the “child” (i). We then compare the probability of child i having genotype g to whether or not child i actually has genotype g .

Since at equilibrium individuals j and ℓ will be identically distributed to the actual parents of individual i , it is easy to show that (17) has expectation zero at equilibrium,

$$E[h_g(\theta, x_i, x_j, x_k)] = E[1(x_i = g)] - E[Pr(g|G_1 = x_j, G_2 = x_\ell, \theta)] \quad (18)$$

$$= Pr(g|\theta) - Pr(g|\theta) = 0, \quad (19)$$

where $Pr(g|\theta)$ is the frequency of genotype g at equilibrium. The $h_g(\theta, x_i, x_j, x_k)$ statistic (17) only involves three individuals, but we can average over all possible triples of individuals to come up with a statistic that still has (approximately) expectation zero.

$$h_g(\theta) := \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n h_g(\theta, x_i, x_j, x_k). \quad (20)$$

Equation (20) is our “ U -statistic” [Hoeffding, 1948]. (Technically, (20) is actually a “ V -statistic” [von Mises, 1947], but these have the same asymptotic behavior as U -statistics, and since U -statistics are more familiar to the Statistics community, we will stick to this terminology). Equation (20) actually represents $K+1$ U -statistics for $g = 0, 1, \dots, K$, each having expectation zero at equilibrium. We’ll let the vector of these U -statistics be represented by

$$\mathbf{h}(\theta) := [h_0(\theta), h_1(\theta), \dots, h_K(\theta)]^\top. \quad (21)$$

Due to sampling variability, the $h_g(\theta)$ ’s will not actually equal zero. But because they should

equal zero in expectation, one intuitive way to estimate θ is to find a θ that makes (21) as close to 0 as possible. We do this by minimizing a norm of $\mathbf{h}(\theta)$,

$$\hat{\theta} := \arg \min_{\theta} n\mathbf{h}(\theta)^{\top} \Omega \mathbf{h}(\theta), \quad (22)$$

where Ω is some positive semi-definite matrix. Such estimators are sometimes called “ U -process minimizers” or “ M_m estimators” in the literature [Honoré and Powell, 1994, Bose, 2002, Bose and Chatterjee, 2018]. It turns out that, for a properly chosen value of Ω , the norm evaluated at $\hat{\theta}$,

$$T(\hat{\theta}) := n\mathbf{h}(\hat{\theta})^{\top} \Omega \mathbf{h}(\hat{\theta}), \quad (23)$$

has an asymptotic χ^2 distribution under the null of equilibrium. This result is derived from the theory of generalized method of moments from the econometrics literature [Hansen, 1982, Newey and West, 1987], and is an application of the Sargan-Hansen J test. We can use this result to explicitly test for equilibrium.

There is another intuitive way to describe (21). Let $\mathbf{q} = (q_0, q_1, \dots, q_K)^{\top}$ represent the genotype frequencies of a population. Then, under random mating, and given a model for meiosis, there is a function $f(\cdot, \cdot)$ that updates the genotype frequencies from one generation to the next. That is, the genotype frequencies of the following generation would be $f(\mathbf{q}, \theta)$. The specific form of $f(\cdot, \cdot)$ used in our manuscript is described in Section S5, but our methods are more general. If we let $\hat{\mathbf{q}}$ be the empirical genotype frequencies of a sample, then one can show that

$$\mathbf{h}(\theta) = \hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \theta). \quad (24)$$

That is, our vector of U -statistics is the difference between the empirical genotype frequencies and one update of the empirical genotype frequencies to the next generation. This is very intuitive, as at equilibrium we have $\mathbf{q} = f(\mathbf{q}, \theta)$ (2), and so we would expect $\hat{\mathbf{q}} \approx f(\hat{\mathbf{q}}, \theta)$. This formulation also shows that we can calculate the U -statistic in linear time, instead of cubic time, since the most computationally intensive part is calculating the sample frequencies $\hat{\mathbf{q}}$. However, framing (24) in terms of U -statistics allows us to wield the asymptotic machinery developed by the field of Statistics, and derive a null distribution given equilibrium.

The full estimation algorithm is presented in Procedure 1. See Section S7 for a formal derivation of our results, Section S7.1 for some practical methodological considerations, and Section S8 for a proof that our U -statistic approach generalizes the classical χ^2 test from diploids to polyploids.

2.5 A bootstrap approach for small samples and uncertain genotypes

Two deficiencies arise from the methods of Sections 2.3 and 2.4. First, these approaches require known genotypes. Genotype uncertainty is a severe issue in polyploid genetics research [Gerard et al., 2018, Gerard and Ferrão, 2019], motivating approaches to account for it, such as when estimating linkage disequilibrium [Gerard, 2021a,b] or inbreeding [Blischak et al., 2018]. The ubiquity of genotype uncertainty limits the applicability of the methods in Sections 2.3 and 2.4 (though, these methods appear to be conservative when using posterior mean genotype counts, see Section S11). Second, these approaches are heavily dependent on asymptotic approximations, so might not be accurate for small sample sizes. Exact tests (or Monte Carlo approximations to exact tests) are generally preferable to those requiring asymptotic approximations [Rohlf and Weir, 2008]. In polyploids, permutation approaches have been created in the SPAGeDi [Hardy and Vekemans,

2002] and GENODIVE [Meirmans and Van Tienderen, 2004] softwares. However, because the null distribution these programs use is that under random segregation of alleles, this approach effectively tests for the presence of binomial genotype frequencies, being unable to account for double reduction.

To overcome both of these deficiencies, we propose a bootstrap approach [Efron, 1979] applied to our U -statistic. Let c_{ik} denote the posterior probability that individual $i = 1, 2, \dots, n$ has genotype $k = 0, 1, \dots, K$. These posterior probabilities are obtainable from many genotyping programs [Voorrips et al., 2011, Serang et al., 2012, Gerard et al., 2018, Gerard and Ferrão, 2019]. If there were no genotype uncertainty, then c_{ik} would be an indicator function for the true genotype for individual i . One bootstrap iteration consists of first sampling n individuals *with replacement*, then sampling each sampled individual's genotype using the c_{ik} 's, then estimating α by minimizing norm (22), then finally calculating a test statistic. If $\mathbf{h}(\hat{\alpha})$ is the vector of observed U -statistics (21), and $\mathbf{h}_b(\hat{\alpha})$ is the b th bootstrap replication of this vector of U -statistics, then we propose to use the following test statistic

$$\hat{T}^{(b)} = n[\mathbf{h}_b(\hat{\alpha}) - \hat{\mathbf{h}}_{boot}]^T [\mathbf{h}_b(\hat{\alpha}) - \hat{\mathbf{h}}_{boot}], \quad (25)$$

where $\hat{\mathbf{h}}_{boot}$ is the empirical mean of the bootstrap $\mathbf{h}_b(\hat{\alpha})$'s. We repeat this process for B bootstrap replications. To obtain a p -value against the null of equilibrium, we calculate the proportion of $\hat{T}^{(b)}$'s that are larger than

$$\hat{T} = n\mathbf{h}(\hat{\alpha})^T \mathbf{h}(\hat{\alpha}). \quad (26)$$

The intuition is that the bootstrap samples of (25) approximate the distribution of the distance of $\mathbf{h}(\hat{\alpha})$ from its mean, so we can compare the distance of $\mathbf{h}(\hat{\alpha})$ from $\mathbf{0}$, calculated in (26), to this bootstrap distribution to see if $\mathbf{0}$ is a reasonable value for the mean. Details of the algorithm are presented in Procedure 2.

3 Results

3.1 Simulations

We ran simulations to study the accuracy of the double reduction estimators, and to evaluate the properties of the proposed tests for equilibrium. Each replication, we generated genotypes of $n \in \{25, 100, 1000\}$ individuals of ploidy $K \in \{4, 6, 8\}$, given the major allele frequency $r \in \{0.1, 0.25, 0.5\}$ and the double reduction parameters $\alpha \in \{\mathbf{0}, \alpha_m/2, \alpha_m\}$, where α_m contains the maximum values of the double reduction parameters under the complete equational segregation model (S76). Given r and α we generated allele frequencies $\mathbf{q}$ after several rounds of random mating using Procedure 3. The number of rounds of random mating was used as a proxy for deviation from equilibrium, where more rounds of random mating result in frequencies closer to equilibrium. If m is the maximum number of rounds of random mating in Procedure 3, we set $m \in \{1, 2, 3, \infty\}$, where $m = 1$ results in the furthest deviation from equilibrium and $m = \infty$ results in the frequencies at equilibrium. Each replication, given the generated $\mathbf{q}$, we simulated the number of individuals with each genotype $\mathbf{y} \sim \text{Multinom}(n, \mathbf{q})$. We ran 1000 replications for each unique combination of parameter settings.

Given these simulated counts, estimates for α were obtained using the likelihood approach (Section 2.3), the U -statistic approach (Section 2.4), and (for tetraploids) the method of Jiang et al.

[2021]. These simulated counts were also used to run the U -statistic test for equilibrium (Section 2.4), the likelihood ratio test for equilibrium (Section 2.3), the likelihood ratio test for random mating (Section 2.2), the bootstrap test for equilibrium (Section 2.5), as well as a likelihood ratio test for equilibrium that does not account for double reduction (Section S10). Due to computational demands, the bootstrap method was only run for a subset of the simulations for only 200 replications each, specifically for $n = 25$, $K = 8$, $r \in \{0.1, 0.5\}$, $\alpha \in \{0, \alpha_m/2, \alpha_m\}$, and $m \in \{1, 3, \infty\}$.

The results for estimating α are presented in Figures S2–S4. There, we see that the method of Jiang et al. [2021] is highly biased. Our methods are somewhat biased for small sample sizes, but appear to be consistent for larger sample sizes. The likelihood and the U -statistic approaches appear to behave very similarly, with the likelihood approach being slightly biased up and the U -statistic approach being slightly biased down for small sample sizes. The variability of the estimators is quite large, even for larger sample sizes, highlighting the difficulty in estimating the rate of double reduction using just a single biallelic locus.

Test statistics of the U -statistic approach and the likelihood approach both generally follow their theoretical distributions under the null, except when the allele frequency is small for tetraploids (Figures S5–S10). The bootstrap approach also produces uniform p -values when the null is satisfied (Figure S11). All three tests are mostly able to control type I error at the 0.05 level, being only slightly anti-conservative in some settings, while the classical approach (Section S10) is completely unable to control type I error except when there is no double reduction (Figure S12–S14). The U -statistic, likelihood, and bootstrap approaches behave very similarly in terms of power, appearing to be asymptotically equivalent (Figures S15–S17). All three can distinguish between the null and alternative hypotheses, but have small power for only small deviations from equilibrium. To understand the power results for tetraploids in S15–S17, recall from Section 2.1 that all sets of gamete frequencies have an “inferred” double reduction rate at which these gamete frequencies would indeed be at equilibrium. We plot this inferred rate in Figure S18 and we see that after three generations, the $\alpha = 0$ and $\alpha = \alpha_m/2$ scenarios have inferred rates within the bounds of theoretically sound double reduction rates. Thus, after three generations of random mating, these are also null scenarios, as they only exhibit small deviations from equilibrium.

All scenarios we simulated under satisfy the random mating assumption, and so we also explored the ability of the random mating test of Section 2.2 to control Type I error in Figures S19–S21. The test is generally able to control for Type I error, being only somewhat anti-conservative for small sample sizes.

See Section S11 for an initial exploration on the effects of genotype uncertainty.

3.2 Analysis of a hatchery population of *Acipenser transmontanus*

In this section, we analyze the dataset of white sturgeon (*Acipenser transmontanus*) from Delomas et al. [2021]. These data consist of ancestral octoploids that likely exhibit tetrasomic inheritance, with disomic inheritance being likely rare in sturgeon [Drauch Schreier et al., 2011]. These data were collected from a Central California caviar farm with broodstock from the Sacramento River, and so likely exhibit random mating and possibly only minor deviations from equilibrium caused, for example, by unobserved relatedness.

We took the $n = 19$ tetraploid individuals from these data and applied updog [Gerard et al., 2018, Gerard and Ferrão, 2019] on the sequencing data to obtain genotype estimates. These data were sequenced to an exceptionally high depth, with a mean average read-depth of 5342, and so it is reasonable to take these genotypes as known. We then tested for equilibrium using the likelihood

approach (Section 2.3), the U -statistic approach (Section 2.4), the bootstrap approach (Section 2.5), and the approach that assumes no double reduction (Section S10). We also tested for random mating (Section 2.2).

The random mating test generally produced more uniform p -values than the other tests, while the equilibrium approaches that allow for double reduction generally produced more uniform-looking p -values than the likelihood ratio test that does not account for double reduction (Figures S22–S23). One can partially explain this by the excess homozygosity we observe in some loci. For example, Figure S24 explores two loci where the p -values assuming no double reduction were 0.058 and 0.056 while the p -values using the U -statistic approach were 0.32 and 0.33. There, we plot the observed genotype counts against the expected genotype counts while either allowing for double reduction, or assuming no double reduction. We see that there is an excess of homozygosity, which can be partially accounted for by modeling double reduction.

We do not display the resulting estimates of double reduction as simulations indicate that such estimates are highly unreliable for such a small sample size (Section 3.1).

3.3 Analysis of an S1 population of *Ipomoea batatas*

We evaluated our methods on an S1 population autohexaploid sweetpotato (*Ipomoea batatas*, $2n = 6x = 90$), obtained from Shirasawa et al. [2017]. SNPs were filtered using the VariantAnnotation package [Obenchain et al., 2014] using the following criteria: allele frequencies between 0.1 and 0.9, average read depth at least 100, and no more than 50% missing data. We obtained genotype estimates for each individual at each locus with updog [Gerard et al., 2018, Gerard and Ferrão, 2019]. Genotype counts were then fed into hwep to test for equilibrium using the likelihood approach (Section 2.3), the U -statistic approach (Section 2.4), and the naive approach (Section S10). We did not run the bootstrap approach due to the computational burden of this method. These data are not at equilibrium, and so each method is expected to produce small p -values. We also tested for random mating (Section 2.2), which should be fulfilled in these data.

Our results are presented in Figures S25–S27. We see that all equilibrium tests behave similarly and result in small p -values (Figure S25). The test for random mating is very conservative (Figure S26), likely because each locus has a large number of genotypes with no or few individuals, worsening the asymptotic approximation of this approach. However, the random mating method provides accurate estimates of gamete frequencies, which we compared to the theoretical hypergeometric distribution of gamete frequencies given the estimated parental genotype at each locus (Figure S27).

4 Discussion

In this manuscript, we clarified the distinction between random mating and equilibrium, noting subtleties in distinguishing between these two hypotheses in the presence of double reduction. We then developed likelihood approaches to both test for random mating and test for equilibrium. Our likelihood approach for equilibrium also yields an estimate of double reduction given equilibrium. However, because of the complexity of genotype frequencies at equilibrium, we developed an intuitively simpler approach based on a statistic that is, on average, zero at equilibrium. This simpler approach reduces down to the classical χ^2 test for Hardy-Weinberg equilibrium in diploids. For small samples and uncertain genotypes, we developed a bootstrap approach to test for equilibrium that uses our U -statistic formulation. We verified our methods in simulations and on real data.

Double reduction has a similar effect on equilibrium as inbreeding [Hardy, 2016]. Thus, it is possible that any double reduction estimates might be picking up the effects of inbreeding. However, one benefit of our approach is that if one rejects the null of equilibrium, then this might suggest that greater inbreeding has occurred than can be accounted for by double reduction, while an approach that does not account for double reduction would not be able to make this claim. It might be possible to adjust our U -statistic approach to incorporate inbreeding. For example, instead of choosing individuals uniformly to be the “parents” of “child” i in (17), one could incorporate a parameter that makes individual i more likely to be its own “parent”. Having an additional parameter would make the model unidentified, at least in tetraploids. However, because inbreeding is likely a individual-specific phenomenon [Blischak et al., 2018] while double reduction is a locus-specific phenomenon, this might inspire an estimation strategy for these parameters. We do not pursue these ideas further in this manuscript.

Our methods do not allow for mixed ploidy populations. Such populations are of interest, as they allow for the study of the evolutionary consequences of different ploidy levels [McAllister and Miller, 2016, Monnahan et al., 2019, Wos et al., 2019], or they exist as a consequence of fish management [Delomas, 2019]. One benefit of the U -statistic approach is that it would be feasible to modify it for mixed ploidy populations with special forms of meiosis, such as those found in *Andropogon gerardii* [Norrman et al., 1997]. For example, if individual i had ploidy K_i , and we had a model for meiosis given parental genotypes and ploidies, say $Pr(g|G_1 = x_j, G_2 = x_\ell, K_j, K_\ell, \theta)$, then we could develop a U -statistic for evaluating equilibrium in this mixed-ploidy population, based on the following,

$$1(x_i = g) - Pr(g|G_1 = x_j, G_2 = x_\ell, K_j, K_\ell, \theta). \quad (27)$$

We leave studying and implementing this extension for future work.

Our methods also do not allow for odd ploidy levels. Though rare, all-triploid bisexual reproduction is possible [Stöck et al., 2002]. The equilibrium frequencies for such species have not yet been derived, limiting the ability to apply a likelihood approach to test for equilibrium. However, our U -statistic approach would potentially provide a strategy to test for equilibrium, as long as a model for meiosis was provided.

Acknowledgments

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

We express our thanks to Dr. Dennis Leung, University of Melbourne, and Dr. Y. Samuel Wang, Cornell University, for insightful discussions. Many thanks to Dr. Thomas Delomas, Pacific States Marine Fisheries Commission, for help with the *Acipenser transmontanus* data.

Most analyses were performed using the R statistical language [R Core Team, 2020].

Data availability

The `hwep` R package implementing the methods described in this paper is available on GitHub (<https://github.com/dcgerard/hwep>). All scripts required to reproduce the results of this manuscript are available at <https://github.com/dcgerard/hwesims>. 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/>.

5 Tables, figures, and procedures

Hypothesis	Description	Parameter Constraints	Number of parameters
H_0	Any genotype frequencies	$\sum_{i=0}^K q_i = 1$	K
H_1	Random mating	$\mathbf{q} = \mathbf{p} * \mathbf{p}$	$K/2$
H_2	Small deviations from equilibrium	$\mathbf{q} = \mathbf{p} * \mathbf{p}$ and $1 - \frac{2+c}{1-c}p_1 \leq (p_2 - p_0)^2 \leq 1 - 2p_1$	$\lfloor K/4 \rfloor + 1$
H_3	Equilibrium and no double reduction	$\mathbf{q} = \mathbf{p} * \mathbf{p}$ and $\mathbf{p} = ((1-r)^2, 2r(1-r), r^2)$	1

Table 1: Summary of hypotheses considered for tetraploids at biallelic loci.

Procedure 1 Procedure to estimate α given equilibrium, and test for equilibrium while accounting for double reduction. $\mathcal{A}(\alpha)$ is a three-dimensional array such that $\mathcal{A}(\alpha)_{plm}$ is the probability an offspring will have genotype p given parental genotypes ℓ and m (Section S5), $\mathcal{A}(\alpha)_{(2)}$ is the second-mode matricization of $\mathcal{A}(\alpha)$ [Kolda and Bader, 2009], $\otimes$ is the Kronecker product, and $\hat{\Sigma}^+$ is the Moore-Penrose inverse of $\hat{\Sigma}$. See Section S7 for a derivation.

- 1: **Input:** The sample genotype frequencies $\hat{\mathbf{q}}$, and the sample size n .
- 2: $\tilde{\alpha} \leftarrow \arg \min_{\alpha} n[\hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \alpha)]^T [\hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \alpha)]$.
- 3: $\hat{\mathbf{f}} \leftarrow f(\hat{\mathbf{q}}, \tilde{\alpha})$ (see Section S5 for the definition of $f(\cdot, \cdot)$).
- 4: $\hat{\Sigma} \leftarrow [\mathbf{I}_{K+1} - 2\mathcal{A}(\tilde{\alpha})_{(2)}(\mathbf{I}_{K+1} \otimes \hat{\mathbf{f}})]^T [\text{diag}(\hat{\mathbf{f}}) - \hat{\mathbf{f}}\hat{\mathbf{f}}^T][\mathbf{I}_{K+1} - 2\mathcal{A}(\tilde{\alpha})_{(2)}(\mathbf{I}_{K+1} \otimes \hat{\mathbf{f}})]$.
- 5: $\nu \leftarrow \text{rank}(\hat{\Sigma}^+)$
- 6: $\hat{\alpha} \leftarrow \arg \min_{\alpha} n[\hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \alpha)]^T \hat{\Sigma}^+ [\hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \alpha)]$
- 7: $T(\hat{\alpha}) \leftarrow n[\hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \hat{\alpha})]^T \hat{\Sigma}^+ [\hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \hat{\alpha})]$
- 8: Compare $T(\hat{\alpha})$ to a $\chi^2_{\nu - \lfloor K/4 \rfloor}$ distribution to obtain a p -value, say p .
- 9: **Return:** The estimated double reduction rate given equilibrium $\hat{\alpha}$, the test statistic for equilibrium $T(\hat{\alpha})$, and the p -value for the test against the null of equilibrium p .

Procedure 2 Bootstrap procedure for calculating a p -value against the null of equilibrium while accounting for both genotype uncertainty and small sample sizes.

```

1: Input:  $c_{ik}$ , the posterior probability that individual  $i = 1, 2, \dots, n$  has genotype  $k = 0, 1, \dots, K$ ,
   and the number of bootstrap iterations  $B$ .
2:  $\hat{\mathbf{q}}_k^{(b)} \leftarrow \frac{1}{n} \sum_{i=1}^n c_{ik}$  for  $k = 0, 1, \dots, K$ , the empirical genotype frequencies.
3:  $\hat{\alpha} \leftarrow \arg \min_{\alpha} n[\hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \alpha)]^{\top} [\hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \alpha)]$ , the estimated double reduction rate.
4:  $\hat{\mathbf{u}} \leftarrow \sqrt{n}[\hat{\mathbf{q}} - f(\hat{\mathbf{q}}, \hat{\alpha})]$ , the observed  $U$ -statistics (24).
5:  $\hat{T} \leftarrow \hat{\mathbf{u}}^{\top} \hat{\mathbf{u}}$ , the observed test statistic.
6: for  $b = 1, \dots, B$  do
7:   Sample  $j_1, j_2, \dots, j_n$  from  $\{1, 2, \dots, n\}$  with replacement.
8:   for  $i = 1, 2, \dots, n$  do
9:     Sample  $x_i^*$  from  $\{0, 1, \dots, K\}$ , where the probability of sampling  $k$  is  $c_{j_i k}$ .
10:  end for
11:   $\hat{\mathbf{q}}_k^{(b)} \leftarrow \frac{1}{n} \sum_{i=1}^n 1(x_i^* = k)$  for  $k = 0, 1, \dots, K$ .
12:   $\hat{\alpha}^{(b)} \leftarrow \arg \min_{\alpha} n[\hat{\mathbf{q}}^{(b)} - f(\hat{\mathbf{q}}^{(b)}, \alpha)]^{\top} [\hat{\mathbf{q}}^{(b)} - f(\hat{\mathbf{q}}^{(b)}, \alpha)]$ 
13:   $\hat{\mathbf{u}}^{(b)} \leftarrow \sqrt{n}[\hat{\mathbf{q}}^{(b)} - f(\hat{\mathbf{q}}^{(b)}, \hat{\alpha}^{(b)})]$ .
14: end for
15:  $\bar{\mathbf{u}} \leftarrow \frac{1}{B} \sum_{b=1}^B \hat{\mathbf{u}}^{(b)}$ .
16: for  $b = 1, 2, \dots, B$  do
17:    $\hat{T}^{(b)} \leftarrow (\hat{\mathbf{u}}^{(b)} - \bar{\mathbf{u}})^{\top} (\hat{\mathbf{u}}^{(b)} - \bar{\mathbf{u}})$ .
18: end for
19:  $p \leftarrow \#\{\hat{T}^{(b)} \geq \hat{T}\}/B$ 
20: Return: The  $p$ -value,  $p$ .

```

Procedure 3 Procedure to generate allele frequencies after several rounds of random mating given the allele frequency r , double reduction parameter α , and ploidy K . Allele frequencies are typically near equilibrium after five iterations.

```

1: Input: Allele frequency  $r$ , ploidy  $K$ , double reduction parameter  $\alpha \in \mathbb{R}^{[K/4]}$ , stopping criteria
    $\tau > 0$ , and maximum number of iterations  $m$ .
2:  $\mathbf{q} \leftarrow (1 - r, 0, \dots, 0, r)$ .
3:  $\epsilon \leftarrow \tau + 1$ 
4:  $i \leftarrow 1$ 
5: while  $\tau < \epsilon$  and  $i \leq m$  do
6:    $\epsilon \leftarrow \|\mathbf{q} - f(\mathbf{q}, \alpha)\|$ 
7:    $i \leftarrow i + 1$ 
8:    $\mathbf{q} \leftarrow f(\mathbf{q}, \alpha)$ 
9: end while
10: Return:  $\mathbf{q}$ .

```

Figure 1: Over abundance of genotypes under the complete equational segregation model. The y -axis contains the equilibrium genotype frequencies under no double reduction subtracted from the equilibrium genotype frequencies under the complete equational segregation model [Mather, 1935]. The x -axis contains the allele frequency. Data are stratified by ploidy (column facets) and dosage (row facets). When the curves are above 0 (horizontal dashed line), this indicates that the dosage is overrepresented in the complete equational segregation model. Homozygosity is typically overrepresented, but the result for other dosages is more complex.

Figure 2: Isometric view of the unit 2-simplex, where each point on the triangle represents some combination of (p_0, p_1, p_2) . Colored bands are gametic frequencies consistent with some double reduction rate $0 \leq \alpha \leq 1/6$. Grey regions are inconsistent with double reduction rates within this interval. The dashed black line corresponds to $\alpha = 0$. H_1 (see Table 1 for a description of hypotheses) allows the gamete frequencies to lie anywhere in this simplex, H_2 constrains the gamete frequencies to lie within the colored bands, and H_3 constrains the gamete frequencies to lie on the dashed black line.

Figure 3: The inferred double reduction rate (y -axis) that would yield equilibrium for given gametes (tuples on plot) in successive generations (x -axis) of random mating when Generation 1 is an S1 population and the true double reduction rate is $1/6$ (red dashed line).

References

- J. H. Bennett. Mixed self- and cross-fertilization in a tetrasomic species. *Biometrics*, 24(3):485–500, 1968. doi: [10.2307/2528313](https://doi.org/10.2307/2528313).
- 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](https://doi.org/10.1093/bioinformatics/btx587).
- A. Bose. U statistics and M_m estimates. In J. Misra, editor, *Uncertainty and Optimality: Probability, Statistics and Operations Research*, pages 257–292. World Scientific Publishing Co. Pte. Ltd., 2002. doi: [10.1142/9789812777010_0007](https://doi.org/10.1142/9789812777010_0007).
- A. Bose and S. Chatterjee. *U-statistics, M_m -estimators and Resampling*. Springer, 2018. ISBN 978-981-13-2248-8. doi: [10.1007/978-981-13-2248-8](https://doi.org/10.1007/978-981-13-2248-8).
- P. M. Bourke, R. E. Voorrips, R. G. F. Visser, and C. Maliepaard. The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*, 201(3):853–863, 09 2015. ISSN 1943-2631. doi: [10.1534/genetics.115.181008](https://doi.org/10.1534/genetics.115.181008).
- D. V. Butruille and L. S. Boiteux. Selection–mutation balance in polysomic tetraploids: Impact of double reduction and gametophytic selection on the frequency and subchromosomal localization of deleterious mutations. *Proceedings of the National Academy of Sciences*, 97(12):6608–6613, 2000. ISSN 0027-8424. doi: [10.1073/pnas.100101097](https://doi.org/10.1073/pnas.100101097).
- J. F. Crow. Eighty years ago: the beginnings of population genetics. *Genetics*, 119(3):473–476, 07 1988. ISSN 1943-2631. doi: [10.1093/genetics/119.3.473](https://doi.org/10.1093/genetics/119.3.473).
- T. A. Delomas. Differentiating diploid and triploid individuals using single nucleotide polymorphisms genotyped by amplicon sequencing. *Molecular ecology resources*, 19(6):1545–1551, 2019. doi: [10.1111/1755-0998.13073](https://doi.org/10.1111/1755-0998.13073).
- T. A. Delomas, S. C. Willis, B. L. Parker, D. Miller, P. Anders, A. Schreier, and S. Narum. Genotyping SNPs and inferring ploidy by amplicon sequencing for polyploid, ploidy-variable organisms. *Molecular Ecology Resources*, page In Press, 2021. doi: [10.1111/1755-0998.13431](https://doi.org/10.1111/1755-0998.13431).
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: [10.1111/j.2517-6161.1977.tb01600.x](https://doi.org/10.1111/j.2517-6161.1977.tb01600.x).
- A. Drauch Schreier, D. Gille, B. Mahardja, and B. May. Neutral markers confirm the octoploid origin and reveal spontaneous autopolyploidy in white sturgeon, *Acipenser transmontanus*. *Journal of Applied Ichthyology*, 27(s2):24–33, 2011. doi: [10.1111/j.1439-0426.2011.01873.x](https://doi.org/10.1111/j.1439-0426.2011.01873.x).
- B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979. doi: [10.1214/aos/1176344552](https://doi.org/10.1214/aos/1176344552).
- H. Geiringer. Chromatid segregation of tetraploids and hexaploids. *Genetics*, 34(6):665–684, 11 1949. ISSN 1943-2631. doi: [10.1093/genetics/34.6.665](https://doi.org/10.1093/genetics/34.6.665).
- D. Gerard. Pairwise linkage disequilibrium estimation for polyploids. *Molecular Ecology Resources*, 21(4):1230–1242, 2021a. doi: [10.1111/1755-0998.13349](https://doi.org/10.1111/1755-0998.13349).
- D. Gerard. Scalable bias-corrected linkage disequilibrium estimation under genotype uncertainty. *Heredity*, in press, 2021b. doi: [10.1038/s41437-021-00462-5](https://doi.org/10.1038/s41437-021-00462-5).
- D. Gerard and L. F. V. Ferrão. Priors for genotyping polyploids. *Bioinformatics*, 36(6):1795–1800, 11 2019. ISSN 1367-4803. doi: [10.1093/bioinformatics/btz852](https://doi.org/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. ISSN 0016-6731. doi: [10.1534/genetics.118.301468](https://doi.org/10.1534/genetics.118.301468).

- J. Graffelman and B. Weir. Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome. *Heredity*, 116(6):558–568, 2016. doi: [10.1038/hdy.2016.20](https://doi.org/10.1038/hdy.2016.20).
- J. Graffelman and B. S. Weir. Multi-allelic exact tests for Hardy-Weinberg equilibrium that account for gender. *Molecular Ecology Resources*, 18(3):461–473, 2018. doi: [10.1111/1755-0998.12748](https://doi.org/10.1111/1755-0998.12748).
- S. W. Guo and E. A. Thompson. Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics*, 48(2):361–372, 1992. ISSN 0006341X, 15410420. doi: [10.2307/2532296](https://doi.org/10.2307/2532296).
- J. Haldane. Theoretical genetics of autopolyploids. *Journal of genetics*, 22(3):359–372, 1930. doi: [10.1007/BF02984197](https://doi.org/10.1007/BF02984197).
- J. S. Haldane. An exact test for randomness of mating. *Journal of Genetics*, 52(1):631–635, 1954. doi: [10.1007/BF02981502](https://doi.org/10.1007/BF02981502).
- L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. ISSN 00129682, 14680262. doi: [10.2307/1912775](https://doi.org/10.2307/1912775).
- G. H. Hardy. Mendelian proportions in a mixed population. *Science*, 28(706):49–50, 1908. ISSN 0036-8075. doi: [10.1126/science.28.706.49](https://doi.org/10.1126/science.28.706.49).
- O. J. Hardy. Population genetics of autopolyploids under a mixed mating model and the estimation of selfing rate. *Molecular Ecology Resources*, 16(1):103–117, 2016. doi: [10.1111/1755-0998.12431](https://doi.org/10.1111/1755-0998.12431).
- 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](https://doi.org/10.1046/j.1471-8286.2002.00305.x).
- K. Haynes and D. Douches. Estimation of the coefficient of double reduction in the cultivated tetraploid potato. *Theoretical and applied genetics*, 85(6-7):857–862, 1993. doi: [10.1007/BF00225029](https://doi.org/10.1007/BF00225029).
- W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. ISSN 00034851. doi: [10.2307/2235637](https://doi.org/10.2307/2235637).
- B. E. Honoré and J. L. Powell. Pairwise difference estimators of censored and truncated regression models. *Journal of Econometrics*, 64(1):241–278, 1994. ISSN 0304-4076. doi: [10.1016/0304-4076\(94\)90065-5](https://doi.org/10.1016/0304-4076(94)90065-5).
- 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, 05 2019. ISSN 2160-1836. doi: [10.1534/g3.119.400132](https://doi.org/10.1534/g3.119.400132).
- L. Jiang, X. Ren, and R. Wu. Computational characterization of double reduction in autotetraploid natural populations. *The Plant Journal*, 105(6):1703–1709, 2021. doi: [10.1111/tpj.15126](https://doi.org/10.1111/tpj.15126).
- T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009. ISSN 0036-1445. doi: [10.1137/07070111X](https://doi.org/10.1137/07070111X).
- H. Levene. On a matching problem arising in genetics. *The Annals of Mathematical Statistics*, 20(1):91–94, 1949. doi: [10.1214/aoms/1177730093](https://doi.org/10.1214/aoms/1177730093).
- J. Li, K. Das, G. Fu, C. Tong, Y. Li, C. Tobias, and R. Wu. EM algorithm for mapping quantitative trait loci in multivalent tetraploids. *International journal of plant genomics*, 2010, 2010. doi: [10.1155/2010/216547](https://doi.org/10.1155/2010/216547).
- K. Mather. Reductional and equational separation of the chromosomes in bivalents and multivalents. *Journal of Genetics*, 30(1):53–78, 1935. doi: [10.1007/BF02982205](https://doi.org/10.1007/BF02982205).
- C. A. McAllister and A. J. Miller. Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in *Andropogon gerardii*. *American Journal of Botany*, 103(7):1314–1325, 2016. doi: [10.3732/ajb.1600146](https://doi.org/10.3732/ajb.1600146).

- 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](https://doi.org/10.1111/j.1471-8286.2004.00770.x).
- P. G. Meirmans, S. Liu, and P. H. van Tienderen. The analysis of polyploid genetic data. *Journal of Heredity*, 109(3):283–296, 01 2018. ISSN 0022-1503. doi: [10.1093/jhered/esy006](https://doi.org/10.1093/jhered/esy006).
- P. Monnahan, F. Kolář, P. Baduel, C. Sailer, J. Koch, R. Horvath, B. Laenen, R. Schmickl, P. Paaanen, G. Šrámková, M. Bohutínská, B. Arnold, C. M. Weisman, K. Marhold, T. Slotte, K. Bomblies, and L. Yant. Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. *Nature ecology & evolution*, 3(3):457–468, 2019. doi: [10.1038/s41559-019-0807-4](https://doi.org/10.1038/s41559-019-0807-4).
- W. K. Newey and K. D. West. Hypothesis testing with efficient method of moments estimation. *International Economic Review*, 28(3):777–787, 1987. ISSN 00206598, 14682354. doi: [10.2307/2526578](https://doi.org/10.2307/2526578).
- G. A. Norrmann, C. L. Quarín, and K. H. Keeler. Evolutionary implications of meiotic chromosome behavior, reproductive biology, and hybridization in 6x and 9x cytotypes of *Andropogon gerardii* (Poaceae). *American Journal of Botany*, 84(2):201–207, 1997. doi: [10.2307/2446081](https://doi.org/10.2307/2446081).
- V. Obenchain, M. Lawrence, V. Carey, S. Gogarten, P. Shannon, and M. Morgan. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, 30(14):2076–2078, 2014. doi: [10.1093/bioinformatics/btu168](https://doi.org/10.1093/bioinformatics/btu168).
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL <https://www.R-project.org/>.
- R. V. Rohlf and B. S. Weir. Distributions of Hardy–Weinberg equilibrium test statistics. *Genetics*, 180(3):1609–1616, 11 2008. ISSN 1943-2631. doi: [10.1534/genetics.108.088005](https://doi.org/10.1534/genetics.108.088005).
- 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, 02 2012. doi: [10.1371/journal.pone.0030906](https://doi.org/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](https://doi.org/10.1038/srep44207).
- M. Stift, C. Berenos, P. Kuperus, and P. H. van Tienderen. Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to *Rorippa* (yellow cress) microsatellite data. *Genetics*, 179(4):2113–2123, 2008. ISSN 0016-6731. doi: [10.1534/genetics.107.085027](https://doi.org/10.1534/genetics.107.085027).
- M. Stöck, D. K. Lamatsch, C. Steinlein, J. T. Epplen, W.-R. Grosse, R. Hock, T. Klapperstück, K. P. Lampert, U. Scheer, M. Schmid, and M. Schartl. A bisexually reproducing all-triploid vertebrate. *Nature genetics*, 30(3):325–328, 2002. doi: [10.1038/ng839](https://doi.org/10.1038/ng839).
- L. Sun, J. Gan, L. Jiang, and R. Wu. Recursive test of Hardy-Weinberg equilibrium in tetraploids. *Trends in Genetics*, 2020. ISSN 0168-9525. doi: [10.1016/j.tig.2020.11.006](https://doi.org/10.1016/j.tig.2020.11.006).
- G. Tai. Estimation of double reduction and genetic parameters in autotetraploids based on 4x–2x and 4x–4x matings. *Heredity*, 49(3):331–335, 1982a. doi: [10.1038/hdy.1982.106](https://doi.org/10.1038/hdy.1982.106).
- G. Tai. Estimation of double reduction and genetic parameters of autotetraploids. *Heredity*, 49(1): 63–70, 1982b. doi: [10.1038/hdy.1982.65](https://doi.org/10.1038/hdy.1982.65).
- R. von Mises. On the asymptotic distribution of differentiable statistical functions. *The Annals of Mathematical Statistics*, 18(3):309–348, 1947. doi: [10.1214/aoms/1177730385](https://doi.org/10.1214/aoms/1177730385).

- 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. ISSN 1471-2105. doi: [10.1186/1471-2105-12-172](https://doi.org/10.1186/1471-2105-12-172).
- J. Wakefield. Bayesian methods for examining Hardy–Weinberg equilibrium. *Biometrics*, 66(1): 257–265, 2010. doi: [10.1111/j.1541-0420.2009.01267.x](https://doi.org/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](https://doi.org/10.1093/jhered/esu062).
- W. Weinberg. Über den nachweis der vererbung beim menschen. *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg*, 64:369–382, 1908. URL <https://wellcomecollection.org/works/ek48psze>.
- B. Weir. *Genetic data analysis II*. Sinauer Associates, Inc., 1996. ISBN 0-87893-902-4.
- G. Wos, J. Mořkovská, M. Bohutínská, G. Šrámková, A. Knotek, M. Lučanová, S. Španiel, K. Marhold, and F. Kolář. Role of ploidy in colonization of alpine habitats in natural populations of *Arabidopsis arenosa*. *Annals of Botany*, 124(2):255–268, 06 2019. ISSN 0305-7364. doi: [10.1093/aob/mcz070](https://doi.org/10.1093/aob/mcz070).
- R. Wu, M. Gallo-Meagher, R. C. Littell, and Z.-B. Zeng. A general polyploid model for analyzing gene segregation in outcrossing tetraploid species. *Genetics*, 159(2):869–882, 10 2001. ISSN 1943-2631. doi: [10.1093/genetics/159.2.869](https://doi.org/10.1093/genetics/159.2.869).
- C. Zheng, R. E. Voorrips, J. Jansen, C. A. Hackett, J. Ho, and M. C. Bink. Probabilistic multi-locus haplotype reconstruction in outcrossing tetraploids. *Genetics*, 203(1):119–131, 2016. doi: [10.1534/genetics.115.185579](https://doi.org/10.1534/genetics.115.185579).
- C. Zheng, R. R. Amadeu, P. R. Munoz, and J. B. Endelman. Haplotype reconstruction in connected tetraploid F1 populations. *Genetics*, 07 2021. ISSN 1943-2631. doi: [10.1093/genetics/iyab106](https://doi.org/10.1093/genetics/iyab106).