frontiers | Frontiers in Genetics

TYPE Opinion
PUBLISHED 04 October 2022
DOI 10.3389/fgene.2022.1027209

Check for updates

OPEN ACCESS

EDITED BY
Matthew B. Hamilton,
Georgetown University, United States

REVIEWED BY
Sebastian E. Ramos,
Centre for Research in Agricultural Genomics (CSIC), Spain

*CORRESPONDENCE
David Gerard,
dgerard@american.edu

SPECIALTY SECTION
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

RECEIVED 24 August 2022
ACCEPTED 12 September 2022
PUBLISHED 04 October 2022

CITATION
Gerard D (2022), Comment on three papers about Hardy–Weinberg equilibrium tests in autopolyploids. *Front. Genet.* 13:1027209. doi: 10.3389/fgene.2022.1027209

COPYRIGHT
© 2022 Gerard. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Comment on three papers about Hardy–Weinberg equilibrium tests in autopolyploids

David Gerard*

Department of Mathematics and Statistics, American University, Washington, DC, United States

KEYWORDS

double reduction, equilibrium, Hardy–Weinberg, hypothesis tests, polyploids, random mating

1 Introduction

One goal in the population genetics of autopolyploids, which are organisms with more than two sets of homologous chromosomes, is to model their genotype frequencies. This modeling presents a greater challenge in autopolyploids than in diploids because certain meiotic configurations in autopolyploids can result in a phenomenon known as double reduction is the co-migration of sister chromatid segments to the same gamete (Stift et al., 2010). Double reduction affects both the segregation frequencies of genotypes from individuals to their offspring (Mather, 1936; Fisher and Mather, 1943; Bever and Felber, 1992; Huang et al., 2019) and the equilibrium genotype frequencies of a panmictic population (Geiringer, 1949; Bennett, 1968; Bever and Felber, 1992; Huang et al., 2019). Testing if a population is in equilibrium, or merely exhibits random mating, is important for the same reasons as tests for Hardy–Weinberg equilibrium in diploids, namely, for 1) determining the mating system, 2) detecting segregation distortion, 3) detecting inbreeding, and 4) detecting genotyping errors (Gerard, 2022b).

Three similar articles which attempt to test for equilibrium and random mating were recently released: one for tetraploids (Sun et al., 2021), one for hexaploids (Wang et al., 2022), and one for octoploids (Wang et al., 2021). These three articles have numerous implementation mistakes, confuse random mating and equilibrium, confuse autopolyploids and allopolyploids (organisms with homoeologous subgenomes), and provide suboptimal testing approaches. The objectives of this study are to correct the authors' mistakes (Section 2), provide examples of how random mating and equilibrium differ in autopolyploids (Section 3), provide examples of how allo- and autopolyploids differ (Section 4), and promote the better methods of Gerard (2022b) and Gerard (2022a) (Section 5).

This study requires a little notation before the issues are discussed here. Let $\boldsymbol{q} = (q_0, q_1, \ldots, q_K)$ be the genotype frequencies at a single biallelic locus for an autopolyploid population with ploidy K; that is, q_k is the proportion of individuals in the population with k copies of the minor allele. Let $\boldsymbol{x} = (x_0, x_1, \ldots, x_K)$ be the genotype counts in a random sample of $n = \sum_{k=0}^{K} x_k$ individuals. Then $\boldsymbol{x}$ is multinomially distributed with size n and probability vector $\boldsymbol{q}$. Under random mating, the genotype frequencies are (Gerard, 2022b)

TABLE 1 Segregation frequencies for an autooctoploid when there is no double reduction, either according to Table 1 from the study by Wang et al. (2021) or according to the correct calculation using the hypergeometric distribution (Eq. 4). The two approaches are different, so the general model for meiosis in the study by Wang et al. (2021) is incorrect.

Parent genotype	Method	Gamete genotype				
		4	3	2	1	0
8	Wang et al. (2021)	1	0	0	0	0
8	Correct	1	0	0	0	0
7	Wang et al. (2021)	9/16	3/8	1/16	0	0
7	Correct	1/2	1/2	0	0	0
6	Wang et al. (2021)	225/784	45/98	87/392	3/98	1/784
6	Correct	3/14	8/14	3/14	0	0
5	Wang et al. (2021)	25/196	75/196	285/784	45/392	9/784
5	Correct	1/14	6/14	6/14	1/14	0
4	Wang et al. (2021)	9/196	12/49	41/98	12/49	9/196
4	Correct	1/70	16/70	36/70	16/70	1/70
3	Wang et al. (2021)	9/784	45/392	285/784	75/196	25/196
3	Correct	0	1/14	6/14	6/14	1/14
2	Wang et al. (2021)	1/784	3/98	87/392	45/98	225/784
2	Correct	0	0	3/14	8/14	3/14
1	Wang et al. (2021)	0	0	1/16	3/8	9/16
1	Correct	0	0	0	1/2	1/2
0	Wang et al. (2021)	0	0	0	0	1
0	Correct	0	0	0	0	1

$$q_k = \sum_{i=\max\left(0,k-K/2\right)}^{\min\left(k,K/2\right)} p_i p_{k-i}, \tag{1}$$

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. Suppose that a population is randomly mating, then there exists a function $f(\boldsymbol{q}, \alpha) = (f_0(\boldsymbol{q}, \alpha), \ldots, f_K(\boldsymbol{q}, \alpha))$ that updates the genotype frequencies from the current generation $\boldsymbol{q}$ to the next $f(\boldsymbol{q}, \alpha)$. Here, α is called the double reduction rate, which is a property of meiosis in autopolyploids (Stift et al., 2010). If the population is at equilibrium, then the genotype frequencies follow

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

For each ploidy, there is a $\boldsymbol{q}$ that satisfies Eq. 2, which is called the "equilibrium genotype frequencies" (Huang et al., 2019). These frequencies are a function of the double reduction rate α and the allele frequency $r = \frac{1}{K}\sum_{k=0}^{K} k q_k$, and have been calculated for ploidies less than or equal to ten (Huang et al., 2019). If $\alpha = 0$, then these equilibrium genotype frequencies reduce to binomial proportions (Haldane, 1930),

$$q_k = \binom{K}{k} r^k (1-r)^{K-k}. \tag{3}$$

This study concerns tests for Eqs 1–3.

2 Implementation and coding errors

There are many logical and coding issues in the studies by Sun et al. (2021), Wang et al. (2021), and Wang et al. (2022). In this section, the ones that were found are listed. However, the code from the study by Sun et al. (2021) is not available, and the code from the study by Wang et al. (2021) and Wang et al. (2022) is verbose and sparsely documented, so there might be more implementation errors that were missed. In particular, the following were found: 1) an incorrect model for meiosis for autooctoploids that results in incorrect equilibrium genotype frequencies, 2) two instances of incorrect χ^2 test statistic calculations, 3) five instances of incorrect degrees of freedom calculations, and 4) two instances of using unknown parameters in an estimation scheme.

The model for meiosis in the study by Wang et al. (2021) is incorrect. This leads to incorrect equilibrium genotype frequencies in their "recursive" test for equilibrium, and thus an incorrect test for equilibrium. It can be determined that their model is incorrect by looking at what it implies when $\alpha = 0$. In this case, the distribution of gamete dosages is known to follow a hypergeometric distribution (Table 1 from Haldane, 1930; Huang et al., 2019). If X is the parental genotype and Y is the gamete genotype, the reader can see this result by thinking of the probability of obtaining Y minor alleles out of $K/2$ chosen alleles from an individual with K total

alleles and X total minor alleles. Therefore, the correct segregation frequencies are obtained *via*

$$Pr\left(Y = y|X = x\right) = \frac{\binom{x}{y}\binom{K-x}{K/2-y}}{\binom{K}{K/2}}. \quad (4)$$

Table 1 shows that the model for meiosis from Table 1 of the study by Wang et al. (2021) does not equal the probabilities from Eq. 4 when $\alpha = 0$, indicating that their model for meiosis is incorrect. It can be empirically observed that their equilibrium frequencies also do not equal binomial proportions when $\alpha = 0$ (Supplementary Appendix S2), which they should (Haldane, 1930).

The χ^2 statistics testing hypotheses Eq. 1 and Eq. 2 are implemented incorrectly in the study by Wang et al. (2022). The χ^2 statistic in Eq. 1 of the study by Wang et al. (2022) is correct in the study, but in their code, they left out the N term. This affects both their equilibrium testing results and their random mating results. This is known because this study reproduced their 6.602 and 6.649 values from page five of their article (Supplementary Appendix S3). Thus, their tests are improperly implemented.

The random mating test is implemented incorrectly in the study by Wang et al. (2022), even after the N term is included; that is, the authors calculate it differently than what they state in their article. Particularly, in their code, they first estimate the gamete frequencies *via* maximum likelihood, and then put the resulting genotype frequencies through the recursive formula to come up with equilibrium values. However, using this recursive formula just results in the same genotype frequencies as the equilibrium "recursive" test. So, the 6.602 value and the 6.649 value aforementioned are different merely because the authors ran the recursive relationship for a different number of iterations.

A total of five instances of incorrect degrees of freedom calculations were counted for the chi-squared tests in the studies by Sun et al. (2021), Wang et al. (2022), and Wang et al. (2021). These calculations are described in Supplementary Appendix S6. Thus, most of the test statistics from these three articles are compared to the incorrect null distribution, resulting in incorrect p-values.

The estimates of α are implemented incorrectly in the study by Wang et al. (2022). The authors do not modularize their code into functions, and this led to some logical errors. They have a variable in their simulations called `alpha`, that is, the true double reduction rate. Their code returns `alpha1`, that is, the estimated double reduction rate. However, their EM algorithm uses `alpha`, not the current version of `alpha1`, to update the parental gamete frequencies. Thus, they use the true value of α in their code that estimates α. This clearly results in unwarranted advantages. Their simulations were rerun after that bug was fixed, obtaining very biased estimates of α (Supplementary Appendix S5). This indicates that either their EM algorithm is wrong or their code is incorrect. It is hard to judge if their EM algorithm is wrong since the EM algorithm used to estimate α is neither in the article nor in the Supplementary Materials. It is also noted that when the authors' "estimate_alpha.R" was run, which should produce the simulation results in their Table 3, it was not able to actually reproduce their Table 3.

From what can be understood through their code (in files "table2_power.R" and "LR.R″), Wang et al. (2021) implemented their tests by using the true genotype frequencies when constructing their test statistics. Needless to say, researchers would not have access to the true genotype frequencies in reality. In their "table2_power.R," they set some genotype frequencies $\boldsymbol{q}_1$ and then obtain the underlying true genotype frequencies *via* a perturbation of $\boldsymbol{q} = f(\boldsymbol{q}_1, \alpha)$, where $\alpha = 0$. They obtain two equivalently valued variables called `prob` and `prob1`. They use a perturbation of `prob` to generate the data, and `prob1` to construct the test statistic, but both `prob` and `prob1` are equal to $\boldsymbol{q}$. An annotated version of "table2_power.R" is provided in the Supplementary Material so that it is easier for the reader to see the issue here. Though the reader is warned, their code is rather verbose and spans 49 $8.5 \times 11''$ pages. Because their test statistic is impossible to be calculated in real analyses (because it uses the true genotype frequencies), the simulation results of Wang et al. (2021) are invalid. It is also noted that when "table2_power.R" was run using the authors' original code, it was not able to actually reproduce the power results in their Table 2.

3 Distinction between random mating and equilibrium

Sun et al. (2021), Wang et al. (2022), and Wang et al. (2021) suggest that Eqs 1 and 2 are the same hypothesis, or at least approximately so. In their articles, they have a "recursive" test and a "gamete-based" test that they claim both tests for "asymptotic Hardy–Weinberg equilibrium." Their "recursive" test does indeed evaluate Eq. 2 (assuming α is known). However, the "gamete-based" test actually evaluates Eq. 1.

Since the authors say that Eq. 1 is about the same as Eq. 2 for any choice of α, this is worth some exploration. As an extreme counterexample (Supplementary Appendix S1), let $\boldsymbol{p} = (0, 0, 1, 0)$, then hypothesis 1) states that

$$\boldsymbol{q}_1 = (0, 0, 0, 0, 1, 0, 0). \quad (5)$$

But $\boldsymbol{q}_1$ is not at equilibrium, and one can use $\boldsymbol{q}_1$ as the starting point for many rounds of random mating to reach equilibrium (Eq. 2). When one does, one obtains

$$\boldsymbol{q}_2 = (0.001, 0.016, 0.082, 0.219, 0.329, 0.263, 0.088), \quad (6)$$

when $\alpha = 0$, the lower bound of the double reduction rate. One also obtains

$$\boldsymbol{q}_3 = (0.005, 0.032, 0.098, 0.204, 0.277, 0.251, 0.133), \quad (7)$$

when $\alpha = 0.3$, the upper bound of the double reduction rate (Huang et al., 2019). Clearly, $\boldsymbol{q}_1$, $\boldsymbol{q}_2$, and $\boldsymbol{q}_3$ are very different. But Sun et al. (2021), Wang et al. (2022), and Wang et al. (2021) suggest that they should be about the same.

As a less contrived example, S1 populations (a single generation of selfing) are technically random mating populations, but hardly any researcher would claim that an S1 population is at equilibrium (see Gerard (2022b) for details).

The only real data example used in the study by Wang et al. (2022) consists of four markers from an F1 population. This is insufficient to explore their methods, as F1 populations exhibit neither random mating (Eq. 1) nor equilibrium (Eq. 2). Furthermore, they did not apply their test for random mating on these data, but rather a test for binomial frequencies (Eq. 3), which is a standard approach, though an incorrect one for F1 populations.

4 Distinction between allo- and autopolyploids

Wang et al. (2021) stated on page four that "The case of no double reduction in the autopolyploid model reduces to allopolyploids if no preferential pairing is assumed." Sun et al. (2021) stated on page three that "When $\alpha = 0$, the pattern of allelic inheritance reduces from autotetraploids to allotetraploids." Since allopolyploids exhibit disomic inheritance within each subgenome (Stift et al., 2010), this is true only if all subgenomes of an allopolyploid have the exact same allele frequency. This is likely not the case in true allopolyploids. In an extreme example, suppose that there is an allooctoploid population with an allele frequency of 0 in two of its subgenomes, and an allele frequency of one in the other two subgenomes; then the overall allele frequency is 0.5, and the allooctoploid equilibrium genotype frequencies are

$$\boldsymbol{q}_{allo} = (0, 0, 0, 0, 1, 0, 0, 0, 0), \quad (8)$$

because every individual will have two minor alleles each from two subgenomes, and two major alleles each from two subgenomes, and therefore, all individuals will have genotype 4. Compare this to the genotype frequencies of an autooctoploid with allele frequency 0.5 at equilibrium when there is no double reduction

$$\boldsymbol{q}_{auto} = (0.004, 0.031, 0.109, 0.219, 0.273, 0.219, 0.109, 0.031, 0.004). \quad (9)$$

Clearly, $\boldsymbol{q}_{allo}$ and $\boldsymbol{q}_{auto}$ are very different. This is not a contrived example, as it might be the case that some subgenomes have fixed an allele before the polyploidization event ("fixed heterozygosity," Cornille et al., 2016).

The tests created in the studies by Sun et al. (2021), Wang et al. (2022), and Wang et al. (2021) are only applicable to autopolyploids, but the only real data example in the studies by Sun et al. (2021) and Wang et al. (2021) are allopolyploids. So the authors did not adequately evaluate their method on a reasonable dataset.

5 Hypothesis testing strategies

The test for equilibrium (Eq. 2) in the studies by Sun et al. (2021), Wang et al. (2021), and Wang et al. (2022) assumes that the double reduction rate is known. But it would not be clear to the reader that this is the case from a reading of the articles. The double reduction rate is never known in practice.

The "recursive" approach in the studies by Sun et al. (2021), Wang et al. (2021), and Wang et al. (2022) for equilibrium testing is unnecessary. The equilibrium frequencies of tetraploids, hexaploids, and octoploids in the presence of double reduction are well documented in the excellent article of Huang et al. (2019). For example, for hexaploids, the equilibrium gamete frequencies are

$$p_0 = \left(1 - \frac{9(3-\alpha)(6-\alpha)}{(9+\alpha)(9+2\alpha)}r + \frac{27(1-\alpha)(3-\alpha)}{(9+\alpha)(9+2\alpha)}r^2\right)(1-r), \quad (10)$$

$$p_1 = \left(\frac{9(3-\alpha)(9-4\alpha)}{(9+\alpha)(9+2\alpha)} - \frac{81(1-\alpha)(3-\alpha)}{(9+\alpha)(9+2\alpha)}r\right)r(1-r), \quad (11)$$

$$p_2 = \left(\frac{45\alpha(3-\alpha)}{(9+\alpha)(9+2\alpha)} + \frac{81(1-\alpha)(3-\alpha)}{(9+\alpha)(9+2\alpha)}r\right)r(1-r), \text{ and} \quad (12)$$

$$p_3 = \left(\frac{20\alpha^2}{(9+\alpha)(9+2\alpha)} + \frac{45\alpha(3-\alpha)}{(9+\alpha)(9+2\alpha)}r + \frac{27(1-\alpha)(3-\alpha)}{(9+\alpha)(9+2\alpha)}r^2\right)r. \quad (13)$$

The equilibrium genotype frequencies are discrete linear convolutions of these proportions. Eqs 10–13 look complicated, but they are not complicated for a computer. It is easy to implement a likelihood approach to test for equilibrium using these gamete frequencies, and such an approach, advantageously, does not depend on knowing the double reduction rate, which is a huge benefit over the iterative approach of Sun et al. (2021), Wang et al. (2021), and Wang et al. (2022). Indeed, this likelihood approach is what was used in the study by Gerard (2022b).

Genotype uncertainty is a major issue in polyploids (Gerard et al., 2018; Gerard and Ferrão, 2019; Gerard, 2021a,b), and so methods should be adjusted to account for this uncertainty. The standard approach to do so is using genotype likelihoods (Li et al., 2011), and this is what was used in the studies by Gerard (2022b) and Gerard (2022a). However, Sun et al. (2021), Wang et al. (2021), and Wang et al. (2022) approach this by aggregating heterozygous genotypes into a single count, which leaves them

with only enough degrees of freedom to test for binomial frequencies Eq. 3. They thus provide no way to evaluate hypotheses Eq. 1 and Eq. 2 in the presence of genotype uncertainty.

6 Discussion

Here, some implementation mistakes and some misconceptions about the genotype frequencies of autopolyploid organisms presented in the studies by Sun et al. (2021), Wang et al. (2021), and Wang et al. (2022) have been discussed. Examples of how random mating and equilibrium differ in autopolyploids, and also how allo- and autopolyploid equilibrium genotype frequencies differ, have been provided. Finally, it was suggested that users consider the approaches of Gerard (2022b) and Gerard (2022a), which do not assume that the double reduction rate is known, and can account for genotype uncertainty through the use of genotype likelihoods.

Sun et al. (2021), Wang et al. (2021), and Wang et al. (2022) could have averted many of their issues if they would have adhered to standard practices in the validation of numeric analysis and coding. In the future, the authors are encouraged to 1) apply unit testing (Wickham, 2011), 2) set up continuous integration (Hilton et al., 2016), 3) implement code review (Vable et al., 2021), 4) modularize their code into functions, ideally, in a package (Wickham, 2015), 5) use a workflow management software to aid in reproducibility and decrease the chance for coding errors (Blischak et al., 2019), 6) provide instructions (ideally automation) on specifically how to reproduce their methods (Heil et al., 2021), and 7) post their code on a repository that is committed to permanency and produces DOI's, such as Zenodo or Figshare, as this extends the lifetime of a work's reproducibility. It is also recommended to encourage greater validation checks, such as demonstrating that the authors' test statistic produce p-values that are uniform under the null. This alone could have detected the test statistic and degrees of freedom issues discussed in Section 2 (Supplementary Appendix S4).

Data availability statement

Additional materials related to this work is available on Zenodo: https://doi.org/10.5281/zenodo.7019205.

- The file "hwesupp.Rmd" is an R Markdown file that contains Appendices S1–S6, and is sufficient to reproduce all of the results of this paper. It has been knitted into "hwesupp.pdf".
- The file "sims.csv" contains the simulation output from Appendix S5 of "hwesupp.Rmd".
- The file "table 2_power.Rmd" contains one iteration of "table 2_power.R" from Wang et al., 2021, annotated to demonstrate the mistakes here. It has been knitted into "table 2_power.pdf".

Much of the code from Wang et al., 2021 and Wang et al., 2022 was packaged by me in the `hexocto` package on Zenodo https://doi.org/10.5281/zenodo.7019230.

A fork of the original code from Wang et al., 2021 and Wang et al., 2022 may be found at at https://github.com/dcgerard/hexaploid and https://github.com/dcgerard/OctoploidDeer.

Author contributions

DG conceived of the study, implemented the study, and wrote the manuscript.

Funding

This material is based upon the work supported by the National Science Foundation under grant no. 2132247.

Acknowledgments

Most analyses were performed using the R statistical language (R Core Team, 2022).

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1027209/full#supplementary-material

References

Bennett, J. H. (1968). Mixed self- and cross-fertilization in a tetrasomic species. *Biometrics* 24, 485–500. doi:10.2307/2528313

Bever, J. D., and Felber, F. (1992). The theoretical population genetics of autopolyploidy. *Oxf. Surv. Evol. Biol.* 8, 185–217.

Blischak, J. D., Carbonetto, P., and Stephens, M. (2019). Creating and sharing reproducible research code the workflowr way. *F1000Res*, 8, 1749. doi:10.12688/f1000research.20843.1

Cornille, A., Salcedo, A., Kryvokhyzha, D., Glémin, S., Holm, K., Wright, S. I., et al. (2016). Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd's purse (Capsella bursa-pastoris). *Mol. Ecol.* 25, 616–629. doi:10.1111/mec.13491

Fisher, R. A., and Mather, K. (1943). The inheritance of style length in *Lythrum salicaria*. *Ann. Eugen.* 12, 1–23. doi:10.1111/j.1469-1809.1943.tb02307.x

Geiringer, H. (1949). Chromatid segregation of tetraploids and hexaploids. *Genetics* 34, 665–684. doi:10.1093/genetics/34.6.665

Gerard, D. (2022a). Bayesian tests for random mating in autopolyploids. *bioRxiv*. doi:10.1101/2022.08.11.503635

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

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

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

Gerard, D. (2021a). Pairwise linkage disequilibrium estimation for polyploids. *Mol. Ecol. Resour.* 21, 1230–1242. doi:10.1111/1755-0998.13349

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

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

Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S.-I., Greene, C. S., and Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nat. Methods* 18, 1132–1135. doi:10.1038/s41592-021-01256-7

Hilton, M., Tunnell, T., Huang, K., Marinov, D., and Dig, D. (2016). Usage, costs, and benefits of continuous integration in open-source projects. *2016 31st IEEE/ACM Int. Conf. Automated Softw. Eng. (ASE) (IEEE)*, 426–437. doi:10.1145/2970276.2970358

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

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., and Abecasis, G. R. (2011). Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.* 21, 940–951. doi:10.1101/gr.117259.110

Mather, K. (1936). Segregation and linkage in autotetraploids. *Journ. Genet.* 32, 287–314. doi:10.1007/BF02982683

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Sun, L., Gan, J., Jiang, L., and Wu, R. (2021). Recursive test of Hardy-Weinberg equilibrium in tetraploids. *Trends Genet.* 37, 504–513. doi:10.1016/j.tig.2020.11.006

Vable, A. M., Diehl, S. F., and Glymour, M. M. (2021). Code review as a simple trick to enhance reproducibility, accelerate learning, and improve the quality of your team's research. *Am. J. Epidemiol.* 190, 2172–2177. doi:10.1093/aje/kwab092

Wang, J., Feng, L., Mu, S., Dong, A., Gan, J., Wen, Z., et al. (2022). Asymptotic tests for Hardy-Weinberg equilibrium in hexaploids. *Hortic. Res.* 9. doi:10.1093/hr/uhac104

Wang, J., Lv, X., Feng, L., Dong, A., Liang, D., and Wu, R. (2021). A tracing model for the evolutionary equilibrium of octoploids. *Front. Genet.* 12. doi:10.3389/fgene.2021.794907

Wickham, H. (2015). R packages: Organize, test, document, and share your code *(O'reilly media)*.

Wickham, H. (2011). testthat: Get started with testing. *R J.* 3, 5–10. doi:10.32614/RJ-2011-002

Supplement to: Comment on Three Papers about Hardy-Weinberg Equilibrium Tests in Autopolyploids

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

Abstract

This supplementary document contains additional simulations, coding examples, and other supporting material for "Comment on Three Papers about Hardy-Weinberg Equilibrium Tests in Autopolyploids".

This document was written in R Markdown and may be explored interactively. All code chunks are executable in the order given. You can access the R Markdown file at https://doi.org/10.5281/zenodo.7019205.

The package `{hexocto}` contains the code from Wang et al. (2021) and Wang et al. (2022), formatted in package form by me so that it is easier to compare. The original repos with the original code are https://github.com/CCBBeijing/hexaploid and https://github.com/CCBBeijing/OctoploidDeer. You can install this package using `{devtools}`:

```
# install.packages("devtools")
devtools::install_github("dcgerard/hexocto")
```

The package `{hwep}` contains the code from Gerard (2022). I use it for comparison purposes at times. You can install the development version via:

```
# install.packages("devtools")
devtools::install_github("dcgerard/hwep")
```

I will load these packages into R now:

```
library(hexocto)
library(hwep)
```

S1 Difference between random mating and equilibrium

Here, I numerically demonstrate the difference between random mating and equilibrium in autohexaploids.

For illustration, let's make an extreme example. Suppose the gamete frequencies for a hexaploid are

```
p <- c(0, 0, 1, 0)
p
```

```
## [1] 0 0 1 0
```

Then the genotype frequencies under random mating are

```
q <- convolve(p, rev(p), type = "open")
round(q, digits = 3)
```

```
## [1] 0 0 0 0 1 0 0
```

The allele frequency is

```
r <- sum(0:6 * q) / 6
r
```

```
## [1] 0.667
```

This results in equilibrium frequencies of the following when $\alpha = 0$, the lower bound,

```
hwep::hwefreq(r = r, alpha = 0, ploidy = 6, niter = Inf)
```

```
## [1] 0.00137 0.01646 0.08230 0.21948 0.32922 0.26337 0.08779
```

I can verify this by iterating the recursive scheme from Wang et al. (2022).

```
qw <- q
for (i in 1:20) {
  qw <- hexocto::hex_onegen(yww = qw, alpha = 0)
}
qw
```

```
## [1] 0.00137 0.01646 0.08230 0.21948 0.32922 0.26337 0.08779
```

Equilibrium frequencies when $\alpha = 0.3$, the upper bound (Huang et al. 2019), are

```
hwep::hwefreq(r = r, alpha = 0.3, ploidy = 6, niter = Inf)
```

```
## [1] 0.00537 0.03190 0.09792 0.20350 0.27748 0.25115 0.13268
```

I can also verify this by iterating the recursive scheme from Wang et al. (2022).

```
qw <- q
for (i in 1:20) {
  qw <- hexocto::hex_onegen(yww = qw, alpha = 0.3)
}
qw
```

```
## [1] 0.00537 0.03190 0.09792 0.20350 0.27748 0.25115 0.13268
```

S2 Incorrect equilibrium genotype frequencies from Wang et al. (2021)

I will begin at the same example genotype frequencies as Wang et al. (2021).

```
yww <- c(0.1, 0.1, 0.15, 0.1, 0.2, 0.1, 0.05, 0.1, 0.1)
```

Then I apply their recursive approach to obtain their equilibrium genotype frequencies

```
hexocto::octo_recursive(yww = yww, niter = 20, alpha = 0)
```

```
## [1] 0.0186 0.0677 0.1424 0.2064 0.2234 0.1793 0.1076 0.0443 0.0104
```

These are different from the theoretical binomial proportions Haldane (1930)

```
r <- sum(0:8 * yww) / 8
dbinom(x = 0:8, size = 8, prob = r)
```

```
## [1] 0.00577 0.04177 0.13228 0.23937 0.27071 0.19594 0.08864 0.02291 0.00259
```

My {hwep} package, on the other hand, correctly calculates these using my recursive formula

```
qcurrent <- yww
for (i in seq_len(20)) {
  qcurrent <- hwep::freqnext(freq = qcurrent, alpha = c(0, 0))
```

```
}
qcurrent
```

```
## [1] 0.00577 0.04177 0.13228 0.23937 0.27071 0.19594 0.08864 0.02291 0.00259
```

S3 Coding errors for χ^2 statistics

Wang et al. (2022) use the following as an example for their tests for equilibrium and random mating on page 5 of their manuscript.

```
nvec <- c(29, 21, 17, 10, 10, 10, 23)
nind <- sum(nvec)
```

Here, I will reproduce those tests, and demonstrate that they implemented their χ^2 test statistics incorrectly.

Their recursive test gets a chi-squared value of 6.602, which I can get here.

```
hex_chisq(yww = nvec / nind,
          nind = nind,
          niter = 8,
          alpha = 0,
          method = "incorrect")
```

```
## $chisq
## [1] 6.6019
##
## $df
## [1] 6
##
## $p
## [1] 0.35924
```

This is the "incorrect" way because they forgot to account for the number of individuals in the chi-squared test. It should be 120 times 6.602.

```
# generate their equilibrium frequencies
qhat <- hex_recursive(yww = nvec / nind, niter = 8, alpha = 0)
# does not use nind
sum((qhat - (nvec / nind))^2 / (qhat))
```

```
## [1] 6.6019
```

```
# correct way
nind * sum((qhat - (nvec / nind))^2 / (qhat))
```

```
## [1] 792.23
```

For the "gamete based test", they get 6.649, but this is not correct. They were just calculating the same test statistic as the 6.602 value, but ran it for a different number of iterations.

```
# Estimate gamete frequencies
hout <- hex_em(yww = nvec / nind, niter = 30)
# Feed those into recursive algorithm
rvec <- hex_recursive(yww = hout$q, niter = 8, alpha = 0)
# Incorrect way
sum((nvec/nind - rvec)^2 / rvec)
```

```
## [1] 6.6487
```

Here is the value they were trying to get.

```
# Incorrectly does not multiply by nind
sum((nvec/nind - hout$q)^2 / hout$q)
```

```
## [1] 0.30123
```

```
# Correctly multiplies by nind
nind * sum((nvec/nind - hout$q)^2 / hout$q)
```

```
## [1] 36.147
```

The authors' two procedures would produce the same values if you ran them for long enough.

```
# "recursive test" from Wang et al. (2021)
hex_chisq(yww = nvec / nind,
          nind = nind,
          niter = 20,
          alpha = 0,
          method = "incorrect")$chisq
```

```
## [1] 6.7014
```

```
# Implementation of "gamete-based" test from Wang et al. (2021)
hout <- hex_em(yww = nvec / nind, niter = 30)
rvec <- hex_recursive(yww = hout$q, niter = 20, alpha = 0)
sum((nvec/nind - rvec)^2 / rvec)
```

```
## [1] 6.7014
```

This is the exact same as just testing for binomial frequencies, but calculating the χ^2 statistic incorrectly.

```
rhat <- sum(nvec / nind * 0:6) / 6
qhat <- dbinom(x = 0:6, size = 6, prob = rhat)
sum((nvec/nind - qhat)^2 / qhat)
```

```
## [1] 6.7014
```

S4 Correct degrees of freedom

Here, I show that the method Wang et al. (2022) does not produce uniform p-values under the null of equilibrium. I also show that my correct version, including the correct degrees of freedom of 5, not 6, does produce uniform p-values under the null of equilibrium. I also find the correct degrees of freedom for the recursive test in Wang et al. (2021) to be 7, not 8.

S4.1 Hexaploids

I generate data under the null of equilibrium. I then fit the incorrect method Wang et al. (2022), my corrected version, and the likelihood ratio test from Gerard (2022).

```
qvec <- hwep::hwefreq(r = 0.5, alpha = 0.1, ploidy = 6)
nrep <- 1000
nsize <- 100000
pout_wang <- rep(NA_real_, length.out = nrep)
pout_correct <- rep(NA_real_, length.out = nrep)
pout_hwep <- rep(NA_real_, length.out = nrep)
for (i in seq_len(nrep)) {
  nvec <- c(rmultinom(n = 1, size = nsize, prob = qvec))
  pout_wang[[i]] <- hex_chisq(yww = nvec / sum(nvec),
```

```
                                nind = sum(nvec),
                                alpha = 0.1,
                                method = "incorrect")$p
  pout_correct[[i]] <- hex_chisq(yww = nvec / sum(nvec),
                                 nind = sum(nvec),
                                 alpha = 0.1,
                                 method = "correct")$p
  pout_hwep[[i]] <- hwep::hwelike(nvec = nvec, thresh = 0)$p_hwe
}
```

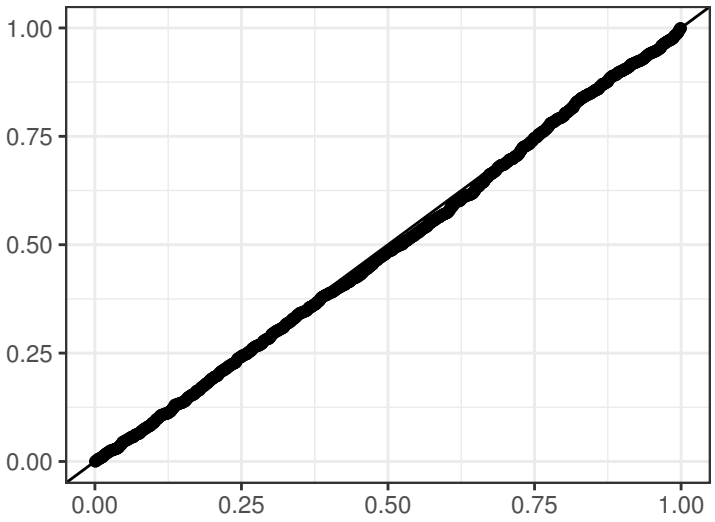
All of the p-values from Wang et al. (2022) are 1, so do not follow a uniform distribution.

```
summary(pout_wang)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1       1       1       1       1       1
```

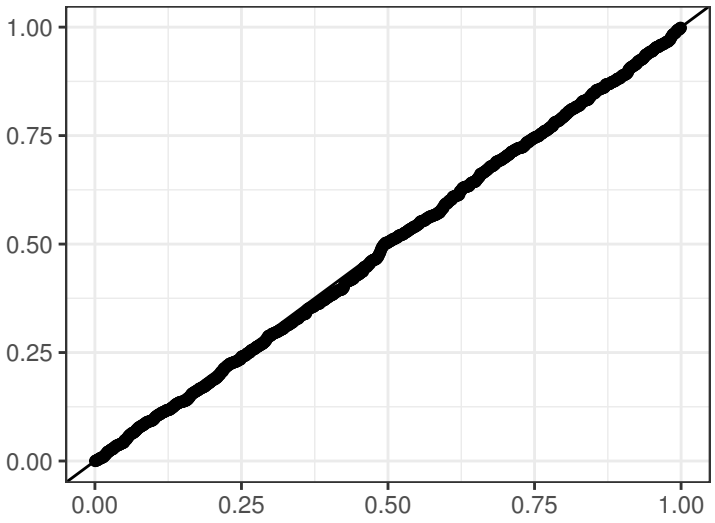
The QQ-plot of the correct p-values follow a uniform distribution.

```
library(ggplot2)
qplot(sample = pout_correct, geom = "qq", distribution = qunif) +
  geom_abline()
```



The QQ-plot of the {hwep} p-values follow a uniform distribution.

```
qplot(sample = pout_hwep, geom = "qq", distribution = qunif) +
  geom_abline()
```



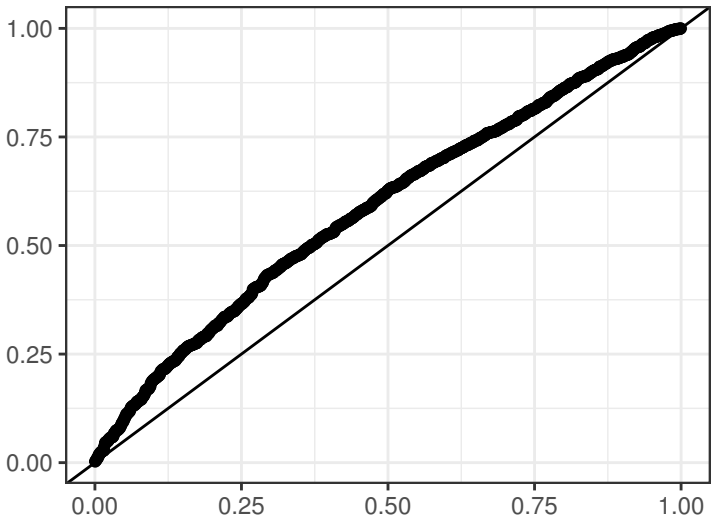
S4.2 Octoploids

I generate data under the incorrect model of Wang et al. (2021), calculating the χ^2 statistic each iteration.

```
yww1 <- c(0, 0, 0, 0, 1, 0, 0, 0, 0)
qvec <- octo_recursive(yww = yww1)
nrep <- 1000
nsize <- 100000
chstat_octo <- rep(NA_real_, length.out = nrep)
for (i in seq_len(nrep)) {
  nvec <- c(rmultinom(n = 1, size = nsize, prob = qvec))
  qemp <- nvec / sum(nvec)
  qnew <- octo_recursive(yww = qemp)
  chstat_octo[[i]] <- sum((qnew - qemp)^2 / qnew) * sum(nvec)
}
```

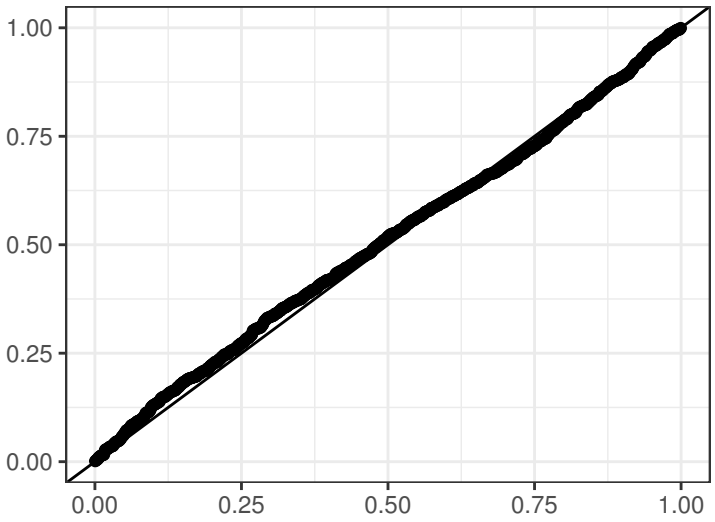
The degrees of freedom is not 8:

```
pocto <- pchisq(q = chstat_octo, df = 8, lower.tail = FALSE)
qplot(sample = pocto, geom = "qq", distribution = qunif) +
  geom_abline()
```



The degrees of freedom *is* 7:

```
pocto <- pchisq(q = chstat_octo, df = 7, lower.tail = FALSE)
qplot(sample = pocto, geom = "qq", distribution = qunif) +
  geom_abline()
```



S5 Simulation study to estimate α

The model Wang et al. (2022) used to create an estimator for α is actually different from (1) and (2). Their model to estimate double reduction says that (i) parent genotypes frequencies satisfy $\tilde{\boldsymbol{q}} = \boldsymbol{p} * \boldsymbol{p}$ for some $\boldsymbol{p}$, and (ii) the current genotype frequencies are $\boldsymbol{q} = f(\tilde{\boldsymbol{q}}, \alpha)$. So this indicates random mating for parents, and one update of random mating for children given the double reduction rate.

I ran simulations with $\boldsymbol{p} = (1,1,1,1)/4$ or $\boldsymbol{p} = (0.1, 0.2, 0.3, 0.4)$, $n \in \{100, 200, 400\}$, and $\alpha \in \{0, 1/7, 1/5, 3/11\}$. This mimics the simulation settings from Wang et al. (2022). I ran each unique combination of parameter settings for 100 replications. Each replication, I generated data according to the assumed model from Wang et al. (2022), then used their code to obtain estimates of $\boldsymbol{p}$ and α. I always initialized the algorithm at $\alpha = 0$ and $\boldsymbol{p} = (1,1,1,1)/4$.

Below is my simulation code.

```
## Parameter settings ----
pvec1 <- rep(1, 4) / 4
qvec1 <- convolve(pvec1, rev(pvec1), type = "open")
pvec2 <- c(0.1, 0.2, 0.3, 0.4)
qvec2 <- convolve(pvec2, rev(pvec2), type = "open")

niter <- 100
paramdf <- expand.grid(seed = seq_len(niter),
                       n = c(100, 200, 400),
                       alpha = c(0, 1/7, 1/5, 3/11),
                       truth = c("A", "B"))

## Estimates to fill in ----
paramdf$alphahat <- NA_real_
paramdf$p0hat <- NA_real_
paramdf$p1hat <- NA_real_
paramdf$p2hat <- NA_real_
paramdf$p3hat <- NA_real_
```

```
## Simulations ----
for (i in seq_len(nrow(paramdf))) {
  set.seed(paramdf$seed[[i]])
  ## offspring genotype frequencies
  if (paramdf$truth[[i]] == "A") {
    qoff <- hex_onegen(yww = qvec1, alpha = paramdf$alpha[[i]])
  } else {
    qoff <- hex_onegen(yww = qvec2, alpha = paramdf$alpha[[i]])
  }

  ## sample of offspring
  nvec <- c(rmultinom(n = 1, size = paramdf$n[[i]], prob = qoff))

  ## estimate parameters
  hout <- hex_estdr(NN = nvec, niter = 1000, tol = 0)
  paramdf$alphahat[[i]] <- hout$alpha
  paramdf$p0hat[[i]] <- hout$p[[1]]
  paramdf$p1hat[[i]] <- hout$p[[2]]
  paramdf$p2hat[[i]] <- hout$p[[3]]
  paramdf$p3hat[[i]] <- hout$p[[4]]
}
write.csv(x = paramdf, file = "./sims.csv", row.names = FALSE)
```

The estimates of α are very biased (Figure S1), and the estimates of $\boldsymbol{p}$ are somewhat biased (Figure S2).

S6 Degrees of Freedom Calculations

Here, I list out the five instances of incorrect degrees of freedom calculations from Sun et al. (2021), Wang et al. (2022), and Wang et al. (2021).

The degrees of freedom for the both the equilibrium and random mating tests are incorrect in Sun et al. (2021). They list the degrees of freedom to be four in both tests. But there are already four free parameters under the alternative (since $q_0 + q_1 + q_2 + q_3 + q_4 = 1$). Since Sun et al. (2021) assume the double reduction rate is known, under the null of equilibrium there is one free parameter (the allele frequency), and so the degrees of freedom for the test for equilibrium is $4 - 1 = 3$, not 4. Under the null of random mating, there are 2 free parameters (since $p_0 + p_1 + p_2 = 1$), and so the degrees of freedom for the test of random mating is $4 - 2 = 2$, not 4.

The degrees of freedom for the random mating test is incorrect in Wang et al. (2022). On page 4 of Wang et al. (2022), the authors say about their test for random mating that "this test statistic follows the chi-square distribution with an unknown degree of freedom. However, we can empirically determine it as a value between 7 - 1 - 1 = 5 to 7 - 1 = 6." I can theoretically determine the degrees of freedom here. There are 6 free parameters under the alternative (since $q_0 + q_1 + q_2 + q_3 + q_4 + q_5 + q_6 = 1$), and there are 3 free parameters under the null (since $p_0 + p_1 + p_2 + p_3 = 1$), and so the degrees of freedom is 6 - 3 = 3, which is neither 5 nor 6.

The degrees of freedom for the recursive test is incorrect in Wang et al. (2022). They say, right after their equation (1) that the degrees of freedom is 6. But there are already 6 free parameters under the alternative. Because Wang et al. (2022) assume the double reduction rate is known, there is only 1 free parameter under the null, the allele frequency. Thus, the true degrees of freedom is 6 - 1 = 5, not 6. See Appendix S4 for an empirical demonstration.

The degrees of freedom for the recursive test is incorrect in Wang et al. (2021). Right after their equation (3), they state that their χ^2 statistic "is thought to follow a chi-square distribution with eight degrees of freedom." But there are already 8 parameters under the alternative (since $\sum_{k=0}^{8} q_k = 1$). The number of parameters

under the null is unclear since they are using a different (incorrect) model for meiosis than I have studied for octoploids, but it likely at least 1 (for the allele frequency). Empirically, it seems the degrees of freedom is 7, not 8 (Appendix S4).

S7 Supplementary Figures

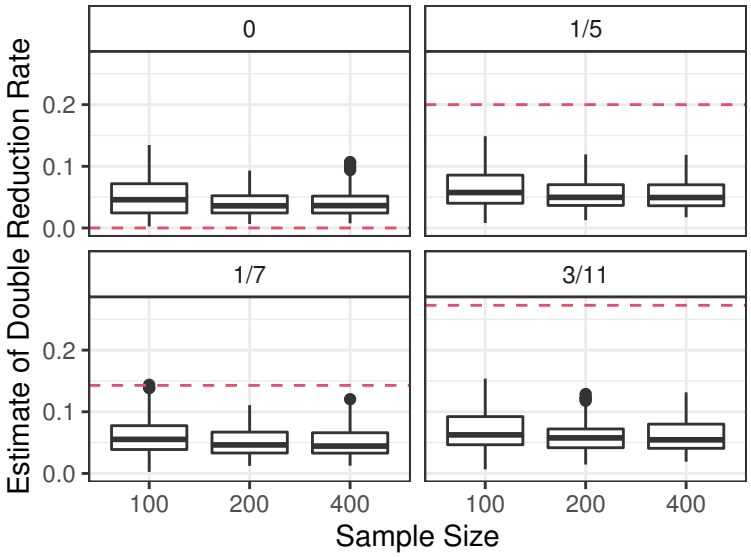


Figure S1: Estimates of α (y-axis) stratified by sample size (x-axis) and true α (facets) using the method of Wang et al. (2022). The red dashed horizontal line is the true α in each facet. The estimates are way off.

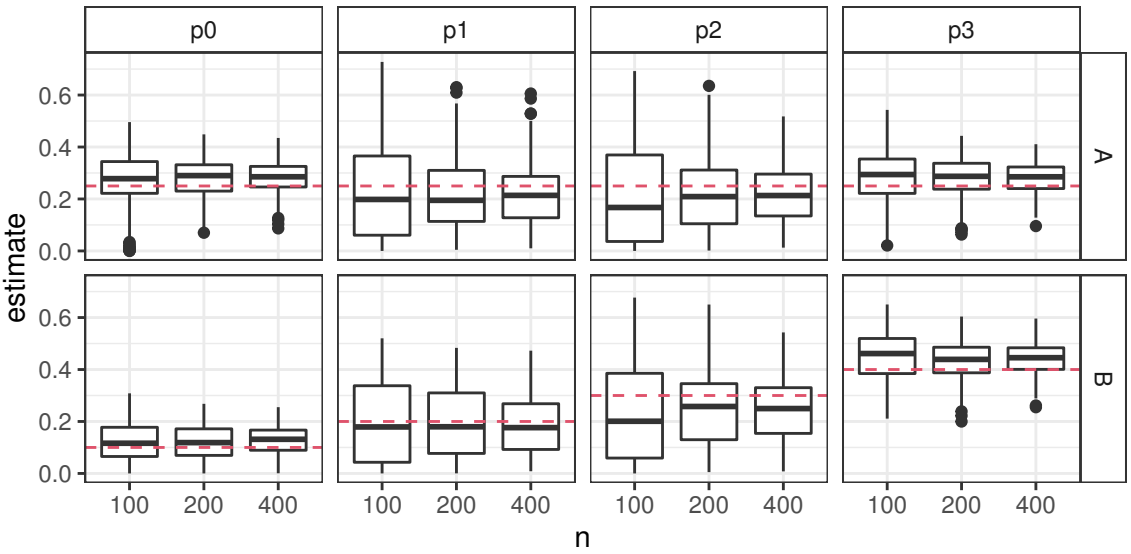


Figure S2: Estimates of p_k (y-axis) for $k = 0, 1, 2, 3$ (row facets) for different sample sizes (x-axis) and different initial values (truth or random) using the method of Wang et al. (2022). The red dashed horizontal line is the true p_k in each facet. The estimates are somewhat biased.

References

Gerard, David. 2022. "Double Reduction Estimation and Equilibrium Tests in Natural Autopolyploid Populations." *Biometrics*. https://doi.org/10.1111/biom.13722.

Haldane, JBS. 1930. "Theoretical Genetics of Autopolyploids." *Journal of Genetics* 22 (3): 359–72. https://doi.org/10.1007/BF02984197.

Huang, Kang, Tongcheng Wang, Derek W Dunn, Pei Zhang, Xiaoxiao Cao, Rucong Liu, and Baoguo Li. 2019. "Genotypic Frequencies at Equilibrium for Polysomic Inheritance Under Double-Reduction." *G3: Genes |Genomes |Genetics* 9 (5): 1693–1706. https://doi.org/10.1534/g3.119.400132.

Sun, Lidan, Jingwen Gan, Libo Jiang, and Rongling Wu. 2021. "Recursive Test of Hardy-Weinberg Equilibrium in Tetraploids." *Trends in Genetics* 37 (6): 504–13. https://doi.org/10.1016/j.tig.2020.11.006.

Wang, Jing, Li Feng, Shuaicheng Mu, Ang Dong, Jinwen Gan, Zhenying Wen, Juan Meng, Mingyu Li, Rongling Wu, and Lidan Sun. 2022. "Asymptotic tests for Hardy-Weinberg equilibrium in hexaploids." *Horticulture Research* 9. https://doi.org/10.1093/hr/uhac104.

Wang, Jing, Xuemin Lv, Li Feng, Ang Dong, Dan Liang, and Rongling Wu. 2021. "A Tracing Model for the Evolutionary Equilibrium of Octoploids." *Frontiers in Genetics* 12. https://doi.org/10.3389/fgene.2021.794907.