Version dated: 11 June, 2015

Estimating allele frequencies in non-model polyploids using high throughput sequencing data

Paul D. Blischak[1,†], Laura S. Kubatko[1,2] and Andrea D. Wolfe[1]

[1] *Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 318 W. 12th Avenue, Columbus, OH 43210, USA.*

[2] *Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA.*

[†] **Corresponding author**: Paul Blischak, Dept. of Evolution, Ecology and Organismal Biology, Aronoff Laboratory, The Ohio State University, 318 W. 12th Avenue, Columbus, OH 43210. E-mail: blischak.4@osu.edu.

Running title: Allele frequencies in polyploids

Abstract

Despite the ever increasing opportunity to collect large-scale datasets for population genomic analyses, the use of high throughput sequencing to study populations of polyploids has seen little application. This is due in large part to problems associated with determining allele copy number in the genotypes of polyploid individuals (allelic dosage uncertainty–ADU), which complicates the calculation of important quantities such as allele frequencies. This well known problem has hindered population genetic studies in polyploids for decades, though several tools exist for analyzing genetic data from polyploids by dealing with particular issues of ADU. Additional complications arise because of the mixed inheritance patterns and variable reproductive modes that are characteristic of many polyploid taxa, making the development of population genetic models for polyploids especially difficult. Here we describe a statistical model to estimate biallelic SNP frequencies in a population of polyploids using high throughput sequencing data in the form of read counts. Uncertainty in the number of copies of an allele in an individual's genotype is accounted for by treating genotypes as an intermediate, latent variable in a hierarchical Bayesian model. In this way, we bridge the gap from data collection (using techniques such as restriction-site associated DNA sequencing) to allele frequency estimation in a unified inferential framework by summing over genotype uncertainty. Simulated datasets were generated under various conditions for both tetraploid and hexaploid populations to evaluate the model's performance and to help guide the collection of empirical data. We also discuss potential sources of bias that could influence results, as well as propose model extensions to ameliorate some of these biases.

(**Keywords**: allelic dosage uncertainty, allele frequencies, hierarchical Bayesian modeling, polyploidy, population genomics)

Introduction

Biologists have long been fascinated by the occurrence of whole genome duplication (WGD) in natural populations and have recognized its role in the generation of biodiversity (Clausen *et al.* 1940; Stebbins 1950; Grant 1971; Otto & Whitton 2000). Though WGD is thought to have occurred at some point in nearly every branch of the Tree of Life (plants, animals and fungi), it is a particularly common phenomenon in plants and is regarded by many to be an important factor in plant diversification (Wood *et al.* 2009; Soltis *et al.* 2009; Scarpino *et al.* 2014). The role of polyploidy in plant evolution was originally considered by some to be a "dead-end" (Stebbins 1950; Wagner 1970; Soltis *et al.* 2014) but, since its first discovery in the early twentieth century, polyploidy has been continually studied in nearly all areas of botany (Winge 1917; Winkler 1916; Clausen *et al.* 1945; Grant 1971; Stebbins 1950; Soltis *et al.* 2003, 2010; Soltis & Soltis 2009; Ramsey & Ramsey 2014). Though fewer examples of WGD are currently known for animal systems, groups such as amphibians, fish, and reptiles all exhibit polyploidy (Allendorf & Thorgaard 1984; Gregory & Mable 2005). Ancient genome duplications are also thought to have played an important role in the evolution of both plants and animals, occurring in the lineages preceeding the seed plants, angiosperms and vertebrates (Ohno 1970; Otto & Whitton 2000; Furlong & Holland 2001; Jiao *et al.* 2011). These ancient WGD events during the early history of seed plants and angiosperms have also been followed by several more WGDs in all major plant groups (Cui *et al.* 2006; Scarpino *et al.* 2014; Cannon *et al.* 2014). Recent experimental evidence has also demonstrated increased survivorship and adaptability to foreign environments of polyploid taxa when compared with their lower ploidy relatives (Ramsey 2011; Selmecki *et al.* 2015).

The theoretical treatment of population genetic models in polyploids has it origins in the Modern Synthesis with Fisher, Haldane and Wright each contributing to the development of some of the earliest mathematical models for understanding the genetic patterns of inheritance in polyploids. Among the earliest of these works was Haldane's 1930 paper on autopolyploid inheritance in $2k$-ploid ($k = 2, 3, \dots$) organisms. Influenced in part by the works of Hermann J. Muller in tetraploid species of *Primula* (1914) and W. J. C. Lawrence in octoploid species of *Dahlia* (1929), Haldane generalized the combinatorial formulas for determining the frequencies of the different possible gametes formed from all genotype combinations for a $2k$-ploid. He also considered additional factors influencing gamete frequencies such as double reduction and the effects of partial selfing (Haldane 1930). Fisher's interest in polyploidy stemmed largely from observations made in the plant genus *Lythrum*, which exhibited conspicuous patterns of trimorphic heterostyly (Fisher 1941). Empirical works by Nora Barlow (1913, 1923), as well as initial investigations into the inheritance patterns of the three style types (Short, Mid, Long) by E. M. East (1927) formed the basis for Fisher's formulation of a model for the inheritance patterns of the Mid length style form in *Lythrum salicaria* (Fisher 1941). He later added to this work by considering double reduction in the inheritance of the Mid length style and complemented his theoretical work through a collaboration with Kenneth Mather to complete crossing experiments (Fisher 1943; Fisher & Mather 1943). Wright's contributions were concerned with the calculation of the distribution of allele frequencies in a $2k$-ploid and were largely an extension of his classic 1931 paper, *Evolution in Mendelian populations*, and a previously published manuscript describing similar processes in diploids (Wright 1931, 1937, 1938). Wright was among the first to consider mutation, migration, selection and inbreeding in his formulation of the distribution of gene frequencies, which helped to establish future ideas about modeling allelic diffusion in a population. For example, it was noted by Motoo Kimura that much of the work on diffusion equations in population genetics could be applied to polyploids in a manner similar to Wright's derivation of the allele frequency distribution in polyploids (Kimura 1964).

The foundation laid down by these early papers has led to the continuing development of population genetic models for polyploids, including models for understanding the rate of loss of genetic diversity and extensions of the coalescent in autotetraploids, as well as modifications of the multispecies coalescent for the inference of species networks containing allotetraploids (Moody *et al.* 1993; Arnold *et al.* 2012; Jones *et al.* 2013). Much of this progress was described in a review by Dufresne *et al.* (2014), who outlined the current state of population genetics in polyploids regarding both molecular techniques and statistical models. Not surprisingly, one of the most promising developments for the future of population genetics in polyploids is the advancement of sequencing technologies. A particularly popular method of gathering large datasets for genome scale inferences is restriction-site associated DNA sequencing [RADseq] (Miller *et al.* 2007; Baird *et al.* 2008; Puritz *et al.* 2014). However, despite its popularity for population genetic inferences at the diploid level, there are much fewer examples of RADseq experiments conducted on polyploid taxa (but see Ogden *et al.* 2013; Wang *et al.* 2013; Logan-Young *et al.* 2015). Among the primary reasons for the dearth in applying RADseq to polyploids is the issue of allelic dosage uncertainty (ADU), or the inability to fully determine the genotype of a polyploid organism when it is partially heterozygous. For a biallelic locus (allele A or B), a partially heterozygous polyploid will have high throughput sequencing reads containing both the A and the B allele. For a tetraploid, the possible genotypes are AAAB, AABB and ABBB. For a hexaploid they are AAAAAB, AAAABB, AAABBB, AABBBB and ABBBBB. In general, the number of possible genotypes for a biallelic locus of a partially heterozygous N-ploid ($\mathrm{N} = 3, 4, 5, \ldots$) is $\mathrm{N} - 1$. A possible solution to this problem would be to develop . However, this could lead to erroneous inferences when genotypes are simply fixed at point estimates based on read proportions without considering estimation error. Furthermore, when sequencing coverage is low, the number of genotypes that will appear to be equally probable increases with ploidy, making it difficult to distinguish among the possible partially heterozygous genotypes.

In this paper we describe a model that aims to address the problems associated with ADU by treating genotypes as a latent variable in a hierarchical Bayesian model and using read counts as data. In this way, we preserve the uncertainty that is inherent in the genotypes of partially heterozygous polyploids by inferring a probability distribution across all possible values of the genotype, rather than treating the genotypes as being directly observed. Our model assumes that the ploidy level of the population is known and that the genotypes of individuals in the population are drawn from a single underlying allele frequency for each locus. These assumptions imply that alleles in the population are undergoing polysomic inheritance without double reduction, which most closely adheres to the inheritance patterns of an autopolyploid. We acknowledge that the model in its current form is an oversimplification of biological reality and realize that it does not apply to a large portion of polyploid taxa. Nevertheless, we believe that accounting for ADU by modeling genotype uncertainty has the potential to be applied more broadly via modifications of the probability model used for the inheritance of alleles, which could lead to more generalized population genetic models for polyploids (see the **Extensibility** section of the **Discussion**).

Materials and Methods

Our goal is to estimate the frequency of a reference allele for each locus sampled from a population of known ploidy (ψ), where the reference allele can be chosen arbitrarily between the two alleles at a given biallelic SNP. To do this we extend the population genomic models of Buerkle & Gompert (2013), which employ a Bayesian framework to model high throughput sequencing reads ($\boldsymbol{T}, \boldsymbol{R}$), genotypes ($\boldsymbol{G}$) and allele frequencies ($\boldsymbol{p}$), to the case of arbitrary ploidy. The idea behind the model is to view the sequencing reads gathered for an individual as a random sample from the unobserved genotype at each locus. Genotypes can then be treated as a parameter in a probability model that governs how likely it is that we see a particular number of sequencing reads carrying the reference allele. Similarly, we can treat genotypes as a random sample from the underlying allele frequency in the population (assuming Hardy-Weinberg equilibrium). This hierarchical setup addresses the problems associated with ADU by treating genotypes as a latent variable that can be integrated out using Markov chain Monte Carlo (MCMC).

Model setup

Here we consider a sample of N individuals from a single population of ploidy level ψ ($\psi \geq 2$) sequenced at L unlinked SNPs. The data for the model consist of two matrices containing counts of high throughput sequencing reads mapping to each locus for each individual: $\boldsymbol{T}$ and $\boldsymbol{R}$. The $N \times L$ matrix $\boldsymbol{T}$ contains the total number of reads sampled at each locus for each individual. Similarly, $\boldsymbol{R}$ is an $N \times L$ matrix containing the number of sampled reads with the chosen reference allele at each locus for each individual. Assuming conditional independence of the sequencing reads given genotypes, the probability distribution for sequencing reads can be factored as

$$P(\boldsymbol{R}|\boldsymbol{T}, \boldsymbol{G}, \epsilon) = \prod_{\ell=1}^{L} \prod_{i=1}^{N} P(r_{i\ell}|t_{i\ell}, g_{i\ell}, \epsilon)\,. \tag{1}$$

For an individual, i, at a particular locus, ℓ, we model the number of sequencing reads containing the reference allele ($r_{i\ell}$) as a Binomial random variable conditional on the total number of sequencing reads ($t_{i\ell}$), the underlying genotype ($g_{i\ell}$) and a constant level of sequencing error (ϵ)

$$P(r_{i\ell}|t_{i\ell}, g_{i\ell}, \epsilon) = \binom{t_{i\ell}}{r_{i\ell}} \begin{cases} \epsilon^{r_{i\ell}}(1-\epsilon)^{t_{i\ell}-r_{i\ell}} & \text{if } g_{i\ell} = 0, \\ \left(\frac{g_{i\ell}}{\psi}\right)^{r_{i\ell}} \left(1 - \frac{g_{i\ell}}{\psi}\right)^{t_{i\ell}-r_{i\ell}} & \text{if } g_{i\ell} = 1, \ldots, \psi - 1, \\ (1-\epsilon)^{r_{i\ell}}\epsilon^{t_{i\ell}-r_{i\ell}} & \text{if } g_{i\ell} = \psi\,. \end{cases} \tag{2}$$

Since the $r_{i\ell}$'s are the data that we observe, the product of $P(r_{i\ell}|t_{i\ell}, g_{i\ell}, \epsilon)$ across loci and individuals will form the likelihood in the model. An important consideration here is that when $g_{i\ell}$ is equal to 0 or ψ (i.e., when the genotype is homozygous) the likelihood will always be 0. To correct for this, we include error (ϵ) into the model. The intuition behind including error is that we may have just not sampled the alternative allele due to lack of sequencing depth or due to a sequencing error. When we do this, the probability distribution for $r_{i\ell}$ given $g_{i\ell}$ and ϵ is split into $\psi + 1$ cases as above in Eq. 2.

The next level in the hierarchy is the conditional prior for genotypes. We assume that the genotypes of the sampled individuals are conditionally independent given the allele frequencies. Factoring the distribution for genotypes and taking the product across loci and individuals gives us the joint probability distribution of genotypes given the ploidy level of the population and the allele frequencies:

$$P(\boldsymbol{G}|\psi, \boldsymbol{p}) = \prod_{\ell=1}^{L} \prod_{i=1}^{N} P(g_{i\ell}|\psi, p_\ell)\,. \tag{3}$$

We model each $g_{i\ell}$ as a Binomial random variable conditional on the ploidy level of the population and the frequency of the reference allele for locus ℓ:

$$P(g_{i\ell}|\psi, p_\ell) = \binom{\psi}{g_{i\ell}} p_\ell^{g_{i\ell}} (1 - p_\ell)^{\psi - g_{i\ell}}\,.$$

The final level of the model is the prior distribution on allele frequencies. Assuming *a priori* independence across loci, we use a Beta distribution with parameters α and β both equal to 1 as our prior distribution for each locus. A Beta(1,1) is equivalent to a Uniform distribution over the interval $[0, 1]$, making our choice of prior uninformative. The joint posterior distribution of allele frequencies and genotypes is then equal to the product across all loci and all individuals of the likelihood, the conditional prior on genotypes and the prior distribution on allele frequencies up to a constant of proportionality

$$\begin{aligned} P(\boldsymbol{p}, \boldsymbol{G}|\boldsymbol{R}, \epsilon) &\propto P(\boldsymbol{R}|\boldsymbol{T}, \boldsymbol{G}, \epsilon) P(\boldsymbol{G}|\psi, \boldsymbol{p}) P(\boldsymbol{p}) \\ &= \prod_{\ell=1}^{L} \prod_{i=1}^{N} P(r_{i\ell}|t_{i\ell}, g_{i\ell}, \epsilon) P(g_{i\ell}|\psi, p_\ell) P(p_\ell)\,. \end{aligned} \tag{4}$$

The marginal posterior distribution for allele frequencies can be obtained by summing over genotypes

$$P(\boldsymbol{p}|\boldsymbol{R}, \epsilon) \propto \sum_{\boldsymbol{G}} P(\boldsymbol{p}, \boldsymbol{G}|\boldsymbol{R}, \epsilon)\,. \tag{5}$$

It would also be possible to examine the marginal posterior distribution of genotypes but for now we will only focus on allele frequencies (see **Discussion** for potential applications of the marginal distribution of genotypes).

Full conditionals and MCMC using Gibbs sampling

We estimate the joint posterior distribution for allele frequencies and genotypes in Eq. 4 using MCMC. This is done using Gibbs sampling of the states $(\boldsymbol{p}, \boldsymbol{G})$ in a Markov chain by alternating samples from the full conditional distributions of $\boldsymbol{p}$ and $\boldsymbol{G}$. Given the setup for our model using Binomial and Beta distributions (which form a conjugate family), analytical solutions for these distributions can be readily acquired (Gelman *et al.* 2014). The full conditional distribution for allele frequencies is Beta distributed and is given by Eq. 6 below:

$$P(\,p_\ell | g_{i\ell}, r_{i\ell}, \epsilon) = \text{Beta}\left(\alpha = \sum_{i=1}^{N} g_{i\ell} + 1,\ \beta = \sum_{i=1}^{N} (\psi - g_{i\ell}) + 1\right), \quad \text{for } \ell = 1, \ldots, L. \tag{6}$$

This full conditional distribution for p_ℓ has a natural interpretation as it is roughly centered at the proportion of sampled alleles carrying the reference allele divided by the total number of alleles sampled given the current state of $\boldsymbol{G}$ in the Markov chain. The "+1" comes from the prior distribution and won't have a strong influence on the posterior distribution when the sample size is large.

The full conditional distribution for genotypes is split into $\psi + 1$ cases (similar to the conditional prior), making it a discrete categorical distribution over the possible values for the genotypes $(0, \ldots, \psi)$. Using k as a generic index, the distribution for individual i at locus ℓ is

$$P(g_{i\ell} | g_{(\text{-}i)\ell}, p_\ell, r_{i\ell}, \epsilon) = \frac{1}{\mathcal{C}_{i\ell}} \begin{cases} \epsilon^{r_{i\ell}} (1-\epsilon)^{t_{i\ell} - r_{i\ell}} (1-p_\ell)^{\psi} & \text{for } k = 0, \\ \left(\frac{k}{\psi}\right)^{r_{i\ell}} \left(1 - \frac{k}{\psi}\right)^{t_{i\ell} - r_{i\ell}} \binom{\psi}{k} p_\ell^k \, (1-p_\ell)^{\psi - k} & \text{for } k = 1, \ldots, \psi - 1, \\ (1-\epsilon)^{r_{i\ell}} \epsilon^{t_{i\ell} - r_{i\ell}} p_\ell^{\psi} & \text{for } k = \psi\,, \end{cases} \tag{7}$$

where $g_{(\text{-}i)\ell}$ is the value of the genotypes for all sampled individuals excluding individual i and $\mathcal{C}_{i\ell}$ is a normalizing constant equal to the sum of all of the terms:

$$\mathcal{C}_{i\ell} = \epsilon^{r_{i\ell}} (1-\epsilon)^{t_{i\ell} - r_{i\ell}} (1-p_\ell)^{\psi} + (1-\epsilon)^{r_{i\ell}} \epsilon^{t_{i\ell} - r_{i\ell}} p_\ell^{\psi} + \sum_{k=1}^{\psi - 1} \left(\left(\frac{k}{\psi}\right)^{r_{i\ell}} \left(1 - \frac{k}{\psi}\right)^{t_{i\ell} - r_{i\ell}} \binom{\psi}{k} p_\ell^k (1-p_\ell)^{\psi - k} \right).$$

More stuff . . .

Simulation study

Simulations were performed to assess error rates in allele frequency estimation for tetraploid and hexaploid populations. Data were generated under the model by sampling genotypes from a binomial distribution conditional on a fixed, known allele frequency ($\,p_\ell = 0.01, 0.05, 0.1, 0.2, 0.4$). Total read counts per individual were simulated for a single locus using a Poisson distribution with mean coverage equal to 5, 10, 20, 50 or 100 reads per individual. We then sampled the number of sequencing reads containing the reference allele from a binomial distribution conditional on the number of total reads, the genotype and sequencing error (Eq. 2; ϵ fixed to 0.01). Finally, we varied the number of individuals sampled per population ($N = 5, 10, 20, 30$) and ran all possible combinations of the simulation settings. Each combination of sequencing coverage, individuals sampled and allele frequency was analyzed using 100 replicates for both tetraploid and hexaploid populations for a total of 20,000 simulation runs. MCMC analyses using Gibbs sampling were run for 50,000 generations with parameter values stored every 50 samples. The first 25% of the posterior was discarded as burn-in, resulting in 750 posterior samples for each replicate. Convergence on the stationary distribution, $P(\boldsymbol{p}, \boldsymbol{G}|\boldsymbol{R}, \epsilon)$, was assessed by examining trace plots for a subset of runs for each combination of settings and ensuring that the effective sample sizes (ESS) were greater than 200. Deviations from the known underlying allele frequency used to simulate each data set were assessed using the standard deviation of the posterior means of each replicate subtracted from the known value.

All simulations were performed using the R statistical package (R Core Team 2014) on the Oakley cluster at the Ohio Supercomputer Center (`https://osc.edu`). Figures were generated using the R packages GGPLOT2 (Wickham 2009), RESHAPE (Wickham 2011) and PLYR (Wickham 2007) with additional figure manipulation completed using Inkscape (`https://inkscape.org`). MCMC diagnostics were done using the CODA package (Plummer *et al.* 2006).

Results

Varying the level of coverage and the number of individuals (Figure #).

It is important to note that we are generating data under the model in the first place, which . However, the high levels of accuracy reported by some of the simulation conditions should not be interpreted as our being selective of settings that show that the model works well, but as validation that the model works when it is supposed to. Analyses of empirical data will obviously not have known values against which the estimated allele frequencies can be compared, but data collection can be informed by these simulations to optimize the number of individuals sequenced per population and the average coverage per locus per individual to obtain good results based on financial resources. Assessments of model adequacy would be also be straightforward to employ to

Discussion

Talk about general discussion points here and maybe include stuff about sources of bias here rather than dedicating an entire section to it.

Potential sources of bias

There are two places where bias can be introduced into the estimation of allele frequencies using this model.

At the level of read counts, unsampled alleles will obviously bias the resulting

Another potential source of bias is treating genotypes among individuals in the population as independent. Polyploid taxa often become apomictic upon genome duplication, and asexual reproduction will generate clonal individuals with identical genotypes. This correlation

Model adequacy

Talk about the adequacy of the model for describing biological reality of polyploids. Emphasize the fact that the model directly contributes to ameliorate the issues associated with ADU.

$$P(\tilde{\boldsymbol{R}}|\boldsymbol{R},\epsilon) = \int \left(\sum_{\boldsymbol{G}} P(\tilde{\boldsymbol{R}}|\boldsymbol{p},\boldsymbol{G}) P(\boldsymbol{p},\boldsymbol{G}|\boldsymbol{R},\epsilon) \right) \mathrm{d}\boldsymbol{p}\,. \tag{8}$$

Extensibility

Talk about how the model can serve as a jumping off point for new models that now that ADU may no longer be as much of an issue. Examples are below.

Replace ψ with a vector of values representing the ploidy of each individual ($\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_N\}$).

Acknowledgements

The authors would like to thank the Ohio Supercomputer Center for access to computing resources and Nick Skomrock for assistance with deriving the full conditional distributions of the model in the diploid case. This manuscript was also greatly improved by comments from the Associate Editor, *editor name*, and # anonymous reviewers.

References

Allendorf FW, Thorgaard GH (1984) *Tetraploidy and the evolution of salmonid fishes. In: Evolutionary genetics of fishes.* Edited by B. J. Turner. Plenum Press, pp. 1–53.

Arnold B, Bomblies K, Wakeley J (2012) Extending coalescent theory to autotetraploids. *Genetics*, **192**, 195–204.

Baird NA, Etter PD, Atwood TS, *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS ONE*, **3**, e3376.

Barlow N (1913) Preliminary note on heterostylism in *Oxalis* and *Lythrum*. *Journal of Genetics*, **3**, 53–65.

Barlow N (1923) Inheritance of the three forms in trimorphic plants. *Journal of Genetics*, **13**, 133–146.

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028–3035.

Cannon SB, McKain MR, Harkess A, *et al.* (2014) Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. *Molecular Biology and Evolution*, **32**, 193–210.

Clausen J, Keck DD, Hiesey WM (1940) *Experimental studies on the nature of species. I. Effect of varied environments on western American plants.* Carnegie Inst. Washington Publ.

Clausen J, Keck DD, Hiesey WM (1945) *Experimental studies on the nature of species. II. Plant evolution through amphiploidy and autoploidy, with examples from Madiinae.* Carnegie Inst. Washington Publ.

Cui L, Wall PK, Leebens-Mack JH, *et al.* (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Research*, **16**, 738–749.

Dufresne F, Stift M, Vergilino R, Malbe BK (2014) Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, **23**, 40–69.

East EM (1927) The inheritance of heterostyly in *Lythrum salicaria. Genetics*, **12**, 393–414.

Fisher RA (1941) The theoretical consequences of polyploid inheritance for the Mid style form of *Lytrum salicaria. Annals of Eugenics*, **11**, 31–38.

Fisher RA (1943) Allowance for double reduction in the calculation of genotype frequencies with polysomic inheritance. *Annals of Eugenics*, **12**, 169–171.

Fisher RA, Mather K (1943) The inheritance of style length in *Lythrum salicaria. Annals of Eugenics*, **12**, 1–23.

Furlong RF, Holland PWH (2001) Were vertebrates octoploid? *Philosophical Transactions of the Royal Society B: Biological Sciences*, **357**, 531–544.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2014) *Bayesian data analysis.* Chapman & Hall/CRC Press, 3rd edn.

Grant V (1971) *Plant speciation.* Columbia University Press.

Gregory TR, Mable BK (2005) *Polyploidy in animals. In: The evolution of the genome.* Edited by T. R. Gregory. Elsevier, pp. 427–517.

Haldane JBS (1930) Theoretical genetics of autopolyploids. *Journal of Genetics*, **22**, 359–372.

Jiao Y, Wickett NJ, Ayyampalayam S, *et al.* (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 97–100.

Jones G, Sagitov S, Oxelman B (2013) Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Systematic Biology*, **62**, 467–478.

Kimura M (1964) Diffusion models in population genetics. *Journal of Applied Probability*, **1**, 177–232.

Lawrence WJC (1929) The genetics and cytology of *Dahlia* species. *Journal of Genetics*, **21**, 125–158.

Logan-Young CJ, Yu JZ, Verma SK, Percy RG, Pepper AE (2015) SNP discovery in complex allotetraploid genomes (*Gossypium* spp., Malvaceae) using genotyping by sequencing. *Applications in Plant Sciences*, **3**, 1400077.

Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated dna (RAD) markers. *Genome Research*, **17**, 240–248.

Moody ML, Mueller LD, Soltis DE (1993) Genetic variation and random drift in autotetraploid populations. *Genetics*, **134**, 649–657.

Muller HJ (1914) A new mode of segregation in gregory's tetraploid primulas. *American Naturalist*, **48**, 508–512.

Ogden R, Gharbi K, Mugue N, *et al.* (2013) Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology*, **22**, 3112–3123.

Ohno S (1970) *Evolution by gene duplication.* Springer.

Otto SP, Whitton J (2000) Polyploid incidence and evolution. *Annual Review of Genetics*, **34**, 401–437.

Plummer M, Best N, Cowles K, Vines K (2006) CODA: Convergence Diagnostics and Output Analysis for MCMC. *R News*, **6**, 7–11.

Puritz JB, Matz MV, Toonen RJ, Weber JN, Bolnick DI, Bird CE (2014) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.

R Core Team (2014) *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramsey J (2011) Polyploidy and ecological adaptation in wild yarrow. *Proceedings of the National Academy of Sciences*, **108**, 7096–7101.

Ramsey J, Ramsey TS (2014) Ecological studies of polyploidy in the 100 years following its discovery. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **369**, 20130352.

Scarpino SV, Levin DA, Meyers LA (2014) Polyploid formation shapes flowering plant diversity. *American Naturalist*, **184**, doi: 10.1086/677752.

Selmecki AM, Maruvka YE, Richmond PA, *et al.* (2015) Polyploidy can drive rapid adaptation in yeast. *Nature*, **519**, 349–352.

Soltis DE, Albert VA, Leebens-Mack J, *et al.* (2009) Polyploidy and angiosperm diversification. *American Journal of Botany*, **96**, 336–348.

Soltis DE, Buggs RJA, Doyle JJ, Soltis PS (2010) What we still don't know about polyploidy. *Taxon*, **59**, 1387–1403.

Soltis DE, Soltis PS, Tate JA (2003) Advances in the study of polyploidy since plant speciation. *New Phytologist*, **161**, 173–191.

Soltis DE, Visger CJ, Soltis PS (2014) The polyploidy revolution then...and now: Stebbins revisited. *American Journal of Botany*, **101**, 1057–1078.

Soltis PS, Soltis DE (2009) The role of hybridization in plant speciation. *Annual Review of Plant Biology*, **60**, 561–588.

Stebbins GL (1950) *Variation and evolution in plants.* Columbia University Press.

Wagner WH (1970) Biosystematics and evolutionary noise. *Taxon*, **19**, 146–151.

Wang N, Thomson M, Bodles WJA, *et al.* (2013) Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Molecular Ecology*, **22**, 3098–3111.

Wickham H (2007) Reshaping data with the reshape package. *Journal of Statistical Software*, **21**, 1–20.

Wickham H (2009) *ggplot2: elegant graphics for data analysis.* Springer New York.

Wickham H (2011) The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, **40**, 1–29.

Winge Ö (1917) The chromosomes: their number and general importance. *Compt. Rend. Trav. Lab. Carlsberg*, **13**, 131–275.

Winkler H (1916) Über die experimentelle Erzeugung von Pflanzen mit abweichenden Chromosomenzahlen. *Zeitschr. f. Bot.*, **8**, 417–531.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH (2009) The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences*, **106**, 13875–13879.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Wright S (1937) The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences*, **23**, 307–320.

Wright S (1938) The distribution of gene frequencies in populations of polyploids. *Proceedings of the National Academy of Sciences*, **24**, 372–377.

Author Contributions

Conceived of the study: PDB, LSK and ADW. PDB derived the polyploid model, ran the simulations, coded the R package and wrote the manuscript. PDB, LSK and ADW reviewed all parts of the manuscript and all authors approved of the final version.

Data Accessibility

Scripts for simulating the datasets, analyzing them using Gibbs sampling and producing the figures from the resulting output can be all be found on GitHub (`https://github.com/pblischak/polyfreqs-ms-data`). We also provide an implementation of the Gibbs sampler for estimating allele frequencies in the R package POLYFREQS (`https://github.com/pblischak/polyfreqs`). See the package vignette or wiki for more details (`https://github.com/pblischak/polyfreqs/wiki`).

Table 1: Notation and symbols used in the description of the model for estimating allele frequencies in polyploids. Vector and matrix forms of the variables are also provided when appropriate.

Symbol	**Description**
L	The number of loci.
ℓ	Index for loci ($\ell \in \{1, \ldots, L\}$).
N	Total number of individuals sequenced.
i	Index for individuals ($i \in \{1, \ldots, N\}$).
ψ	The ploidy level of individuals in the population (e.g., tetraploid: ψ=4).
p_ℓ	Frequency of the reference allele at locus ℓ. [$\boldsymbol{p}$]
$g_{i\ell}$	The number of copies of the reference allele for individual i at locus ℓ. [$\boldsymbol{G}$]
$t_{i\ell}$	The total number of reads for individual i at locus ℓ. [$\boldsymbol{T}$]
$r_{i\ell}$	The number of reads with the reference allele for individual i at locus ℓ. [$\boldsymbol{R}$]
ϵ	Sequencing error.