bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Version dated: 2 July, 2015

# Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids

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

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

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

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

**Running title**: Genotype uncertainty in autopolyploids

## 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 even though various solutions to circumvent the difficulty of estimating polyploid genotypes have been proposed. 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 autopolyploids 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 a 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

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 extensions to generalize our model and its application to polyploids.

(**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 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).

Polyploids are generally divided into two types based on how they are formed: auto- and allopolyploids. Autopolyploids form when a WGD event occurs within a single evolutionary lineage and typically have polysomic inheritance. Allopolyploids are formed by hybridization between two separately evolving lineages followed by WGD and are thought to have mostly disomic inheritance. Multivalent chromosome pairing during meiosis can occur in allopolyploids, however, resulting in mixed inheritance patterns across loci in the genome [segmental allopolyploids] (Stebbins 1950). Autopolyploids can also undergo double reduction, a product of multivalent chromosome pairing wherein segments from sister chromatids move together during meiosis—resulting in allelic inheritance that breaks away from a strict pattern of polysomy (Haldane 1930). Autopolyploidy was also thought to be far less common than allopolyploidy, but recent studies have concluded that autopolyploidy occurs much more frequently than originally proposed (Soltis *et al.* 2007; Parisod *et al.* 2010).

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 first of these works was Haldane's 1930 paper on autopolyploid inheritance in $2k$-ploid

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

($k = 2, 3, \dots$) 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 Kimura (1964) 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.

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 common 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 many 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 at a given locus. This is the same problem that has been encountered by other codominant markers such as microsatellites, which have been commonly used for population genetic analyses in polyploids. One way of dealing with allelic dosage that has been used for multi-allelic microsatellite markers has been to code alleles as either present or absent based on electropherogram readings (allelic phenotypes) and to analyze the resulting dominant data using a program such as POLYSAT (Clark & Jasieniuk 2011; Dufresne *et al.* 2014). de Silva *et al.* (2005) developed a method for inferring allele frequencies using observed allelic phenotype data and used an expectation-maximization algorithm to deal with the incomplete genotype data resulting from ADU. Attempts to directly infer the genotype of polyploid microsatellite loci have also been successfully completed in some cases by using the

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

relative electropherogram peak heights of the alleles in the genotypes (Esselink *et al.* 2004). The estimation problem would be similar for biallelic SNP data collected using RADseq, where a partially heterozygous polyploid will have high throughput sequencing reads containing both alleles. For a tetraploid, the possible genotypes for a partial heterozygote (alleles A and B) would be 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 K-ploid ($\mathrm{K} = 3, 4, 5, \ldots$) is $\mathrm{K} - 1$. A possible solution to this problem for SNPs would be to try to use existing genotype callers and to rely on the relative number of sequencing reads containing the two alleles (similar to what was done for microsatellites). 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 high throughput sequencing read counts as data (Figure 1). 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. This approach has been used by Buerkle & Gompert (2013) to deal with uncertainty in calling genotypes in diploids and the work we present here builds off of their earlier models. 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 ($\psi$), 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). For our model, a genotype is simply a count of the number of reference alleles at a locus which can range from 0 (a homozygote with no reference alleles in the genotype) to $\psi$ (a homozygote with only reference alleles in the genotype). All whole numbers in between 0 and $\psi$ represent partially heterozygous genotypes. This hierarchical

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 $\psi$ 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: $\boldsymbol{R}$ and $\boldsymbol{T}$. 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 reference allele at each locus for each individual. Then for individual $i$ at locus $\ell$, 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 ($\epsilon$)

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

If we assume conditional independence of the sequencing reads given the genotypes, the joint probability distribution for sequencing reads is given by

$$P(\boldsymbol{R}|\boldsymbol{T}, \boldsymbol{G}, \epsilon) = \prod_{\ell=1}^{L}\prod_{i=1}^{N} P(r_{i\ell}|t_{i\ell}, g_{i\ell}, \epsilon)\,. \tag{2}$$

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 $\psi$ (i.e., when the genotype is homozygous) but the sequence data collected for the individual contain both alleles, the likelihood will be 0. To correct for this, we include error ($\epsilon$) into the model. The intuition behind including error is that reads sampled from a locus that is truly homozygous can still contain more than one allele due to sequencing errors, giving the false impression that the individual may be a partial heterozygote. If we assume that the probability of a sequencing error is smaller than the probability of being truly heterozygous (i.e., containing at least one reference allele in the genotype), then the probability model we use for reference reads should be able to distinguish between homozygotes with reads containing sequencing errors and partial heterozygotes. When we do this, the probability distribution for $r_{i\ell}$ given $g_{i\ell}$ and $\epsilon$ is split into $\psi + 1$ cases as above in Eq. 1.

The next level in the hierarchy is the conditional prior for genotypes. 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 $\ell$ ($p_\ell$):

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

We also assume that the genotypes of the sampled individuals are conditionally independent given the allele frequencies, which is equivalent to taking a random sample from a population in Hardy-Weinberg equilibrium. Factoring the distribution for genotypes and taking the product across loci and individuals

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

gives us the joint probability distribution of genotypes given the ploidy level of the population and the vector of allele frequencies at each locus ($\boldsymbol{p} = \{p_1, \ldots, p_L\}$):

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

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 $\alpha$ and $\beta$ 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 here 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:

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

This full conditional distribution for $p_\ell$ 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 will not have a strong influence on the posterior when the sample size is large.

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 $\ell$ is

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

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:

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

The full conditional distribution for genotypes can be seen as the product of two quantities: (1) the probability of each of the possible genotypes based on the observed reference reads and (2) the probability of drawing each genotype value based on the current value for the frequency of the reference allele for locus $\ell$ in the population.

We begin our Gibbs sampling algorithm in a random position in parameter space through the use of uniform probability distributions. The genotype matrix is initialized with random draws from a Discrete Uniform distribution from 0 to $\psi$ and the initial allele frequencies are drawn from a Uniform distribution on the interval [0,1].

## Simulation study

Simulations were performed to assess error rates in allele frequency estimation for tetraploid and hexaploid populations ($\psi = 4$ and 6, respectively). 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 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. 1; $\epsilon$ 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 100,000 generations with parameter values stored every 100 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 by calculating the posterior mean of each replicate, followed by subtracting the known value from the calculated means and then calculating their overall standard deviation.

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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). All code is available on GitHub (`https://github.com/pblischak/polyfreqs-ms-data`).

# Results

## Gibbs sampler

Our Gibbs sampling algorithm was able to accurately estimate allele frequencies for a number of simulation settings while simultaneously allowing for genotype uncertainty. There were no indications of a lack of convergence (ESS values $> 200$) for any of the simulation replicates and all trace plots examined also indicated that the Markov chain had reached stationarity. We have aggregated the scripts for our Gibbs sampler as a developmental R package—POLYFREQS—which is available on GitHub (`https://github.com/pblischak/polyfreqs`). Though POLYFREQS is written in R, it deals with the large datasets that are generated by high throughput sequencing platforms in two ways. First, it takes advantage of R's ability to incorporate C++ code via the RCPP and RCPPARMADILLO packages, allowing for a faster implementation of our MCMC algorithm (Eddelbuettel & François 2011; Eddelbuettel 2013; Eddelbuettel & Sanderson 2014). Second, since the model assumes independence among loci, POLYFREQS can facilitate the process of parallelizing analyses by splitting the total read count and reference read count matrices into subsets of loci which can be analyzed at the same time on separate nodes of a computing cluster.

## Simulation study

Increasing the number of individuals sampled had the largest effect on the accuracy of allele frequency estimation (Figures 2 & 3). Since allele frequencies are population level parameters, it is not surprising that sampling more individuals from the population leads to better estimates. This appears to be the case even when sequencing coverage is quite low (5x, 10x), which corroborates the observations made by Buerkle & Gompert (2013). Lower sequence coverage does affect the posterior distribution for allele frequencies even when the number of individuals sequenced is large however, by increasing the posterior standard deviation (Figure 4). An interesting pattern that emerged during the simulation study is the observation that the allele frequencies closer to 0.5 tend to have higher error rates, which is to be expected given that the variance of a Binomial random variable is highest when the probability of success is 0.5.

# Discussion

The inference of population genetic parameters and the demographic history of non-model polyploid organisms has consistently lagged behind that of diploids. The difficulties associated with these inferences present themselves at two levels. The first of these is the widely known inability to determine

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the genotypes of polyploids due to ADU. Even though there have been theoretical developments in the description of models for polyploid taxa as early as the 1930s, a large portion of this population genetic theory relies on knowledge about individuals' genotypes (e.g., Haldane 1930; Wright 1938). The second complicating factor is the complexity of inheritance patterns and changes in mating systems that often accompany WGD events. Polyploid organisms can sometimes mate by both outcrossing or selfing, and can display mixed inheritance patterns at different loci in the genome (Dufresne *et al.* 2014). If genotypes were known, then it might be easier to develop and test models for dealing with and inferring rates of selfing versus outcrossing, as well as understanding inheritance patterns across the genome. However, ADU only compounds the problems associated with these inferences, making the development and application of appropriate models far more difficult (but see list of software in Dufresne *et al.* 2014). The model we have presented here deals with the first of these two issues by not treating genotypes as observed quantities. Almost all other methods of genotype estimation for polyploids treat the genotype as the primary parameter of interest. Our model is different in that we still use the read counts generated by high throughput sequencing platforms as our observed data but instead integrate across genotype uncertainty when inferring other parameters, thus bypassing the problems caused by ADU.

Despite our focus on bypassing ADU, an important consideration for the model we present here is that, because it approximates the joint posterior distribution of allele frequencies and genotypes, it would also be possible to use the marginal posterior distribution of genotypes to make inferences using existing methods. This could be done using the posterior mode as a maximum *a posteriori* (MAP) estimate of the genotype for downstream analyses, followed by analyzing random samples from the posterior distribution of genotypes. The resulting set of estimates would not constitute a "true" posterior distribution of downstream parameters but would allow researchers to interpret their results based on the MAP estimate of the genotypes while still getting a sense for the amount of variation in their estimates. Using the posterior distribution of genotypes in this way could technically be applied to any type of polyploid, but is only really appropriate for autopolyploids due to the model of inheritance that is used. Other methods for estimating SNP genotypes from high throughput sequencing data include the program SuperMASSA, which models the relative intensity of the two alternative alleles using Normal densities (Serang *et al.* 2012).

A second important factor for using our model is that, although estimates of allele frequencies can be accurate when sequencing coverage is low and sample sizes are large, the distribution for genotypes is likely going to be quite diffuse. For analyses that treat genotypes as a nuisance parameter, this is not an issue since we can integrate across genotype uncertainty. However, if the genotype *is* of primary interest, then the experimental design of the study will need to change to acquire higher coverage at each locus for more accurate genotype estimation. Therefore, the decision between sequencing more individuals with lower average coverage versus sequencing fewer individuals with higher average coverage depends primarily on whether the genotypes will be used or not.

## Model adequacy

As noted earlier, the probability model that we use for the inheritance of alleles is one of polysomy without double reduction. In some cases, this model may be inappropriate but it can still be informative to check for loci that do or do not follow the model that we assume. Below we describe a simple procedure for rejecting our model of inheritance on a per locus basis using comparisons with the posterior predictive distribution of sequencing reads. Model checking is an important part of

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

making statistical inferences and can play a role in understanding when a model adequately describes the data being analyzed. In the case of our model, it can serve as a basis for understanding the inheritance patterns of the organism being studied by determining which loci adhere to a simple pattern of polysomic inheritance.

Given $M$ posterior samples for the allele frequencies at locus $\ell$, $\left\{p_\ell^{[1]}, p_\ell^{[2]}, \ldots, p_\ell^{[M]}\right\}$, we will simulate new values for the genotypes ($\tilde{g}_{i\ell}$) and reference read counts ($\tilde{r}_{i\ell}$) for all individuals and use the ratio of simulated reference read counts to observed total read counts $\left(\frac{\tilde{r}_{i\ell}}{t_{i\ell}}\right)$ as a summary statistic for comparing the observed read count ratios to the distribution of the predicted read count ratios. The use of the likelihood (or similar quantities) as a summary statistic has been a common practice in posterior predictive comparisons of nucleotide substitution models, and more recently for comparative phylogenetics (Ripplinger & Sullivan 2010; Reid *et al.* 2014; Pennell *et al.* 2015). We use the ratio of reference to total read counts here because it is the maximum likelihood estimate of the probability of success for a Binomial random variable and because it is a simple quantity to calculate. The use of other summary statistics, or a combination of multiple summary statistics, would also possible. The procedure for our posterior predictive model check is as follows:

1. For locus $\ell = 1, \ldots, L$:
    1.1. For posterior sample $m = 1, \ldots, M$:
        1.1.1. Simulate new genotype values $\left(\tilde{g}_{i\ell}^{[m]}\right)$ for all individuals ($i = 1, \ldots, N$) by drawing from a Binomial $\left(\psi, p_\ell^{[m]}\right)$.
        1.1.2. Simulate new reference read counts $\left(\tilde{r}_{i\ell}^{[m]}\right)$ from each new genotype for all individuals by drawing from Eq. 1.
        1.1.3. Calculate the reference read ratio for the simulated data for sample $m$ and sum across individuals: $\tilde{\mathcal{S}}_\ell^{[m]} = \sum_{i=1}^{N}\left(\frac{\tilde{r}_{i\ell}^{[m]}}{t_{i\ell}}\right)$.
        1.1.4. Calculate the reference read ratio for the observed data and sum across individuals: $\mathcal{S}_\ell = \sum_{i=1}^{N}\left(\frac{r_{i\ell}}{t_{i\ell}}\right)$.
    1.2. Calculate the difference between the observed reference read ratio and the $M$ simulated reference read ratios: $\left\{\mathcal{S}_\ell - \tilde{\mathcal{S}}_\ell^{[1]}, \ldots, \mathcal{S}_\ell - \tilde{\mathcal{S}}_\ell^{[M]}\right\}$.
2. Determine if the 95% highest posterior density (HPD) interval of the distribution of re-centered reference read ratios contains 0.

When the distribution of the differences in ratios between the observed and simulated datasets does not contain 0 in the 95% HPD interval, it provides evidence that the locus being examined does not follow a pattern of strict polysomic inheritance. A similar approach could be used on an individual basis by comparing the observed ratio of reference reads to the predicted ratios for each individual at each locus. We implement the per locus version of this posterior predictive model checking procedure in the POLYFREQS package with the function `polyfreqs_pps`.

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## Extensibility

The modular nature of our hierarchical model allows for the addition and modification of nodes in the model graph (Figure 1). One of the simplest extensions to the model that can build directly on the current setup would be to consider loci with more than two alleles. This can be done using Multinomial distributions for sequencing reads and genotypes and a Dirichlet prior on allele frequencies (the Multinomial and Dirichlet distributions form a conjugate family; Gelman *et al.* 2014). We could also model populations of mixed ploidy by using a vector of individually assigned ploidy levels instead of assuming a single value for the whole population ($\boldsymbol{\psi} = \{\psi_1, \ldots, \psi_N\}$). However, this would assume random mating among ploidy levels.

The place where we believe our model could have the greatest impact is through modifications and extensions of the probability model used for the inheritance of alleles. These models have been difficult to apply in the past as a result of genotype uncertainty. However, using our model as a starting point, it could be possible to infer patterns of inheritance (polysomy, disomy, heterosomy) and other demographic parameters (e.g., effective population size, population differentiation) without requiring direct knowledge about the genotypes of the individuals in the population. For example, Haldane's (1930) model of genotype frequencies for autopolyploids that are partially selfing could be used to infer the prevalence of self-fertilization within a population. Similarly, Fisher's (1943) model for double reduction in the inheritance of style lengths for *Lythrum* could be generalized and used alone or together with a model for partial selfing to better understand how these processes affect the genetic diversity of a population. A more recent model described by Stift *et al.* (2008) used microsatellites to infer the different inheritance patterns (disomic, tetrasomic, intermediate) for tetraploids in the genus *Rorippa* (Brassicaceae) following crossing experiments. The reformulation of such a model for biallelic SNPs gathered using high throughput sequencing could provide a suitable framework for understanding inheritance patterns across the genome. An ideal model would be one that could help to understand inheritance patterns without the need conduct additional experiments. However, to our knowledge, such a model does not currently exist and may not even be possible to implement due to the complexity of possible inheritance patterns that might need to be considered without the addition of information from crosses.

# Conclusions

The recent emergence of models for genotype uncertainty in diploids has introduced a theoretical framework for dealing with the fact that genotypes are unobserved quantities (Gompert & Buerkle 2012; Buerkle & Gompert 2013). Our extension of this theory to cases of higher ploidy (specifically to autopolyploids) progresses naturally from the original work but also serves to alleviate the deeper issue of ADU. The power and flexibility of these models as applied at the diploid level has the potential to be replicated for polyploid organisms with the addition of suitable models for allelic inheritance. The construction of hierarchical models containing suitable probability models for ADU, allelic inheritance and perhaps even additional levels for important parameters such as F statistics or the allele frequency spectrum also have the potential to provide key insights into the population genetics of polyploids (Gompert & Buerkle 2011; Buerkle & Gompert 2013). Future work on such models will help to progress the study of polyploid taxa and could eventually lead to more generalized models for understanding the processes that have shaped their evolutionary histories.

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

## 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. We would also like to thank *Associate Editor*, Aaron Wenzel and # anonymous reviewers for their comments on the manuscript. This work was partially funded through a grant from the National Science Foundation (DEB-1455399) to ADW and LSK.

## 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.

Clark LV, Jasieniuk M (2011) POLYSAT: an R package for polyploid microsatellite analysis. *Molecular Ecology Resources*, **11**, 562–566.

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.

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Eddelbuettel D (2013) *Seamless R and C++ integration with Rcpp*. Springer, New York.

Eddelbuettel D, François R (2011) Rcpp: seamless R and C++ integration. *Journal of Statistical Software*, **40**, 1–18.

Eddelbuettel D, Sanderson C (2014) RcppArmadillo: accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, **71**, 1054–1063.

Esselink GD, Nybom H, Vosman B (2004) Assignment of allelic configuration in polyploids using the MAC-PR (microsatellite DNA allele counting–peak ratios) method. *Theoretical and Applied Genetics*, **109**, 402–408.

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.

Gompert Z, Buerkle CA (2011) A hierarchical Bayesian model for next-generation population genomics. *Genetics*, **187**, 903–917.

Gompert Z, Buerkle CA (2012) bgc: software for Bayesian estimation of genomic clines. *Molecular Ecology Resources*, **12**, 1168–1176.

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.

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

Parisod C, Holderegger R, Brochmann C (2010) Evolutionary consequences of autopolyploidy. *New Phytologist*, **186**, 5–17.

Pennell MW, FitzJohn RG, Cornwell WK, Harmon LJ (2015) Model adequacy and the macroevolution of angiosperm functional traits. *American Naturalist*, **186**, E100.

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.

Reid NM, Hird SM, Brown JM, *et al.* (2014) Poor fit to the multispecies coalescent is widely detectable in empirical data. *Systematic Biology*, **63**, 322–333.

Ripplinger J, Sullivan J (2010) Assessment of substitution model adequacy using frequentist and Bayesian methods. *Molecular Biology and Evolution*, **27**, 2790–2803.

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.

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Serang O, Mollinari M, Garcia AAF (2012) Efficient exact maximum *a posteriori* computation for Bayesian SNP genotyping in polyploids. *PloS ONE*, **7**, e30906.

de Silva H, Hall A, Rikkerink E, McNeilage M, Fraser L (2005) Estimation of allele frequencies in polyploids under certain patterns of inheritance. *Heredity*, **95**, 327–334.

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, Schemske DW, *et al.* (2007) Autopolyploidy in angiosperms: have we grossly underestimated the number of species? *Taxon*, **56**, 13–30.

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.

Stift M, Berenos C, Kuperus P, van Tienderen PH (2008) Segregation models for disomic, tetrasomic and intermediate inheritance in tetraploids: a general procedure applied to *Rorippa* (yellow cress) microsatellite data. *Genetics*, **179**, 2113–2123.

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.

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 initial draft of the manuscript. PDB, LSK and ADW reviewed all parts of the manuscript and all authors approved of the final version.

## Data Accessibility

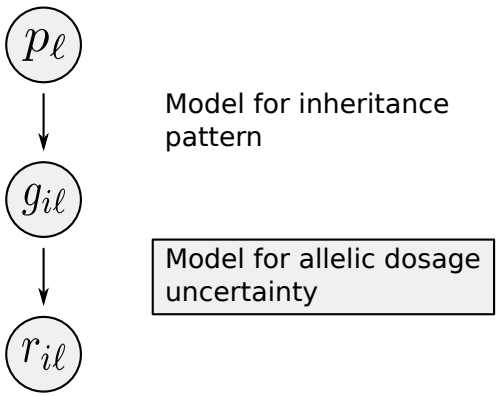
Scripts for simulating the datasets, analyzing them using Gibbs sampling and producing the figures from the resulting output can all be found on GitHub (`https://github.com/pblischak/polyfreqs-ms-data`). We also provide an implementation of the Gibbs sampler for estimating allele frequencies in the developmental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

bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

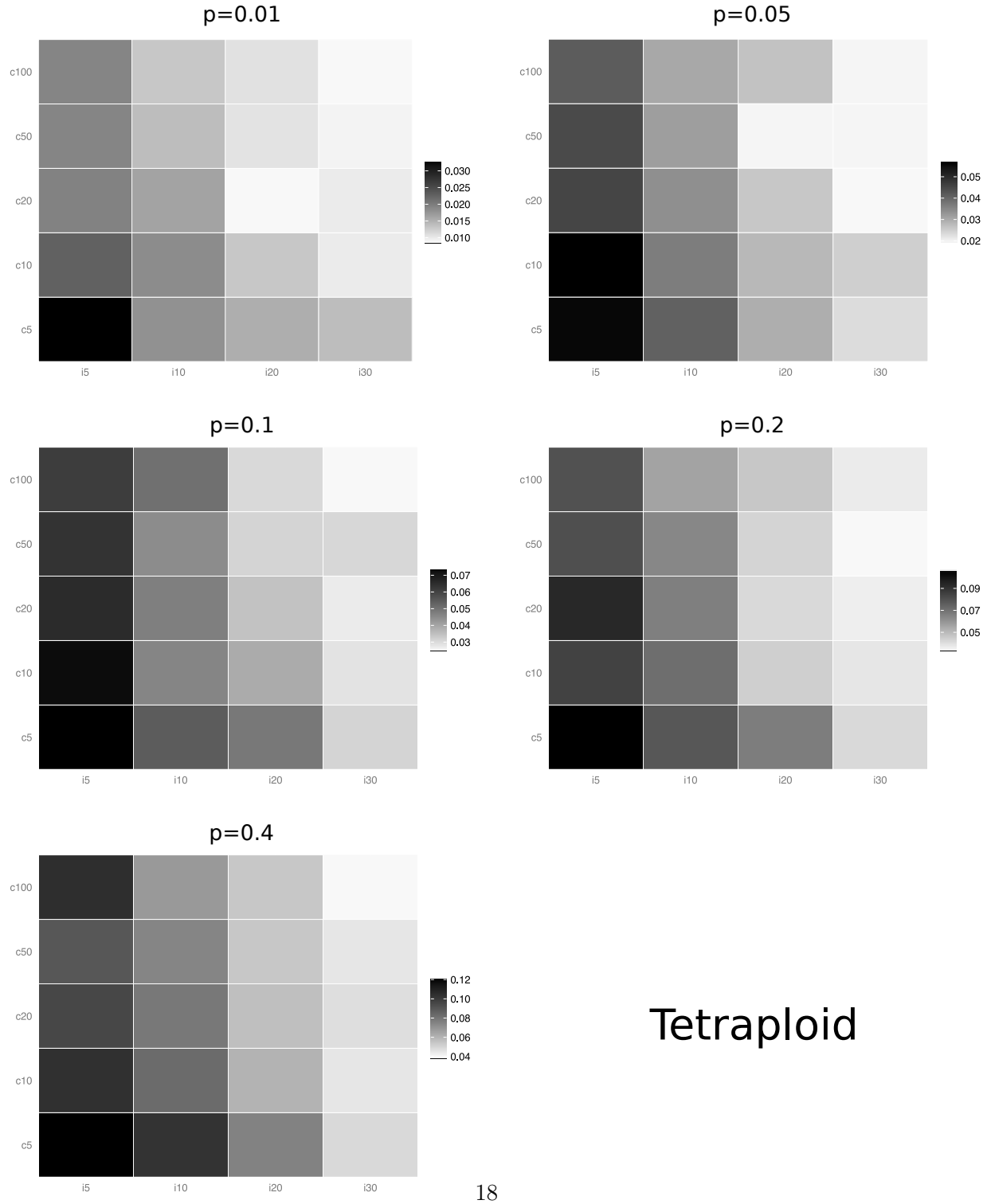
| Symbol | Description |
|---|---|
| $L$ | The number of loci. |
| $\ell$ | Index for loci ($\ell \in \{1, \ldots, L\}$). |
| $N$ | Total number of individuals sequenced. |
| $i$ | Index for individuals ($i \in \{1, \ldots, N\}$). |
| $\psi$ | The ploidy level of individuals in the population (e.g., tetraploid: $\psi$=4). |
| $p_\ell$ | Frequency of the reference allele at locus $\ell$. [$\boldsymbol{p}$] |
| $g_{i\ell}$ | The number of copies of the reference allele for individual $i$ at locus $\ell$. [$\boldsymbol{G}$] |
| $\tilde{g}_{i\ell}$ | Simulated genotype for posterior predictive model checking. |
| $t_{i\ell}$ | The total number of reads for individual $i$ at locus $\ell$. [$\boldsymbol{T}$] |
| $r_{i\ell}$ | The number of reads with the reference allele for individual $i$ at locus $\ell$. [$\boldsymbol{R}$] |
| $\tilde{r}_{i\ell}$ | Simulated reference read count for posterior predictive model checking. [$\tilde{\boldsymbol{R}}$] |
| $\epsilon$ | Sequencing error. |

Figure 1: Graphical representation of a hierarchical Bayesian model for estimating allele frequencies. The two levels (allelic dosage and inheritance) represent the probability models that are used for inference from one graph node to another. The model we present here focuses primarily on allelic dosage.



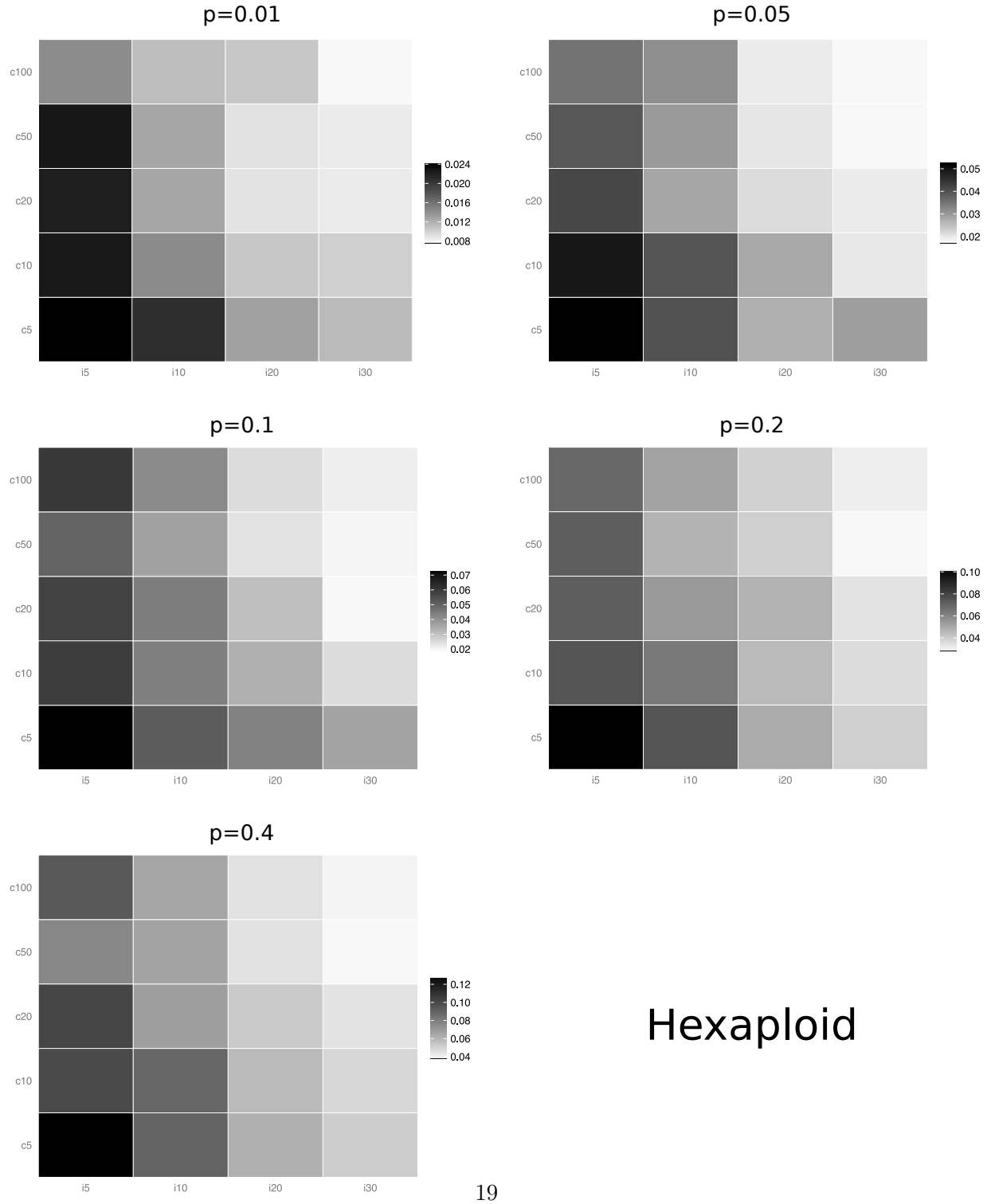
bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2: Heat maps of error rates (standard deviation from the true value) for allele frequency estimation in tetraploids. The x-axis shows the number of individuals (i5, i10, i20, i30) and the y-axis represents the sequencing coverage (c5, c10, c20, c50, c100) for each simulation. Note that the scales for each heat map are not the same, but the overall pattern of increased accuracy as the number of individuals and sequencing coverage increases for all allele frequencies.



bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3: Heat maps of error rates (standard deviation from the true value) for allele frequency estimation in hexaploids. The x-axis shows the number of individuals (i5, i10, i20, i30) and the y-axis represents the sequencing coverage (c5, c10, c20, c50, c100) for each simulation. Note that the scales for each heat map are not the same, but the overall pattern of increased accuracy as the number of individuals and sequencing coverage increases is the same for all allele frequencies.



bioRxiv preprint doi: https://doi.org/10.1101/021907; this version posted July 2, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4: The posterior standard deviation for allele frequencies decreases with increased sequencing coverage. This plot provides a comparison of the distribution of posterior standard deviations of the 100 replicates performed for each level of sequencing coverage for the hexaploid simulation with 30 individuals and an allele frequency of 0.2.