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.

Keywords: polyploidy, population genomics, hierarchical Bayesian modeling, RAD sequencing, POLYFREQS

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 parameter 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.

Introduction

The process of whole genome duplication (WGD) has come to be acknowledged as an important evolutionary force in the generation of potentially adaptive genetic variation (Otto & Whitton 2000; Soltis & Soltis 2000; Soltis *et al.* 2003, 2009, 2010, 2014; Selmecki *et al.* 2015). Though its effects were initially debated (Wagner 1970; Stebbins 1950), a reevaluation of the prevalence and impact of WGD has solidified its influence at evolutionary time scales both deep and shallow (Soltis *et al.* 2004;

Comai 2005; Cui *et al.* 2006; Jiao *et al.* 2011; Cannon *et al.* 2014; Douglas *et al.* 2015).

The theoretical treatment of population genetics 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 allele frequencies in polyploids (Haldane 1930; Wright 1938; **?**). Among the earliest of these papers was Haldane's extension of the Hardy-Weinberg theorem to arbitrary ploidy, as well as his treatment of autotetraploid inhertance. Fisher's interest in polyploidy stemmed largely from observations made in the plant genus , which exhibited differing levels of heterostyly. The further observation of mixed ploidy-levels within the populations led Fisher to examine patterns of inheritance of polyploids. Wright's contributions were concerned mostly with the calculation of changes in allele frequencies through mutation, migration and to some degree selection, and can be considered a simple extension of his previous work on allele frequencies to the case of a $2k$-ploid. Moody *et al.* (1993) It was also noted by Moto Kimura in his papers on diffusion processes for allele frequencies that models such as Wright's ϕ could potentially be applied to polyploid organisms (Kimura 1964). All of these models depend on the accurate estimation of genotypes, which has been the Achilles' heel of polyploid population genetics. For a more contemporary review on population genetics in polyploids see Dufresne *et al.* (2014) and references within.

There are few examples of RADseq experiments (but see (Logan-Young *et al.* 2015)).

Materials and methods

Our aim is to estimate the frequency of a reference allele for each locus sampled from a population of known ploidy, where the reference allele can be chosen arbitrarily between the two alleles at a given SNP. To do this we extend the population genomic models of Buerkle & Gompert (2013), which employ a Bayesian framework to model next-generation sequencing reads, genotypes and allele frequencies, to the case of arbitrary ploidy. The basic idea is to view the sequencing reads gathered for an individual as a random sample from the unobserved genotype at a given locus. Genotypes can then be used 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 nuisance parameter that we integrate out using Markov chain Monte Carlo (MCMC). . Figure 1 gives an illustration of the model for a tetraploid ($\psi = 4$) and Table 1 provides a reference for the notation and symbols used in the description of the model below.

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 reads typically generated by high throughput sequencing platforms: $\boldsymbol{T}$ 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

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

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 model for genotypes will also be Binomial with parameter

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

The joint posterior distribution of allele frequencies and genotypes is 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

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

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

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

It would also be possible to examine the marginal posterior distribution of genotypes but here we focus on allele frequencies only.

Full conditionals and MCMC using Gibbs sampling

We estimate the joint posterior distribution for allele frequencies and genotypes in Eq. 4 using Markov chain Monte Carlo (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 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 $\frac{\alpha}{\alpha+\beta}$. 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

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

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

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

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 ($p = 0.2, 0.4, 0.6, 0.8$). Total read counts per individual were simulated for a single locus using a Poisson distribution with mean coverage equal to λ ($\lambda = 5, 10, 15, 20$), followed by the sampling of sequencing reads containing the reference allele from a binomial distribution conditional on the number of total reads, the genotype and sequencing error (ϵ fixed to 0.01). 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 12,800 simulation runs. MCMC analyses using Gibbs sampling were run for 500,000 generations with parameter values stored every 500 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 calculated using the posterior mean of each of the 100 replicates for a given simulation set and comparing it to the known frequency using the root mean squared error (RMSE).

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 add-on 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). Scripts for simulating and analyzing the datasets, as well as for reproducing all figures, are available on GitHub (`https://github.com/pblischak/polyfreqs-ms-data`).

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 descrigin biological reality of polyploids. Emphasize the fact that the model directly contributes to ameliorate the issues associated with ADU.

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.

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

R and bash 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 wiki for more details (`https://github.com/pblischak/polyfreqs/wiki`).

References

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.

Comai L (2005) The advantages and disadvantages of being polyploid. *Nature Reviews. Genetics*, **6**, 836–846.

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

Douglas G, Gos G, Steige K, *et al.* (2015) Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *bioRxiv*, **doi**, http://dx.doi.org/10.1101/006783.

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.

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

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.

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

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.

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

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.

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

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, Pires JC, Kovarik A, Tate JA, Mavrodiev E (2004) Recent and recurrent polyploidy in tragopogon (asteraceae): cytogenetic, genomic and genetic comparisons. *Biological Journal of the Linnean Society*, **82**, 485–501.

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 (2000) The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences*, **97**, 7051–7057.

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

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

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.

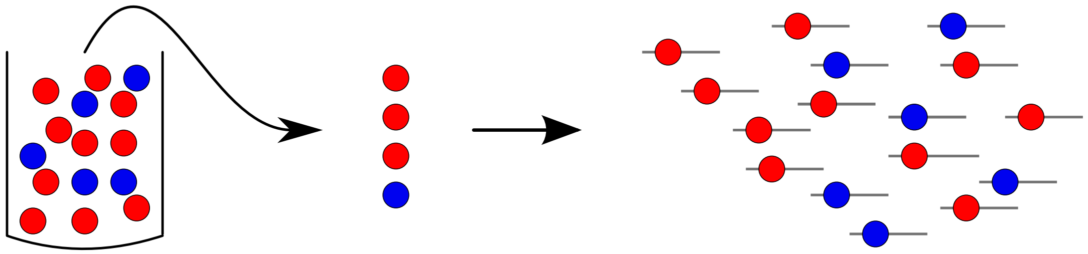
Wickham H (2011) The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, **40**, 1–29.

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

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.

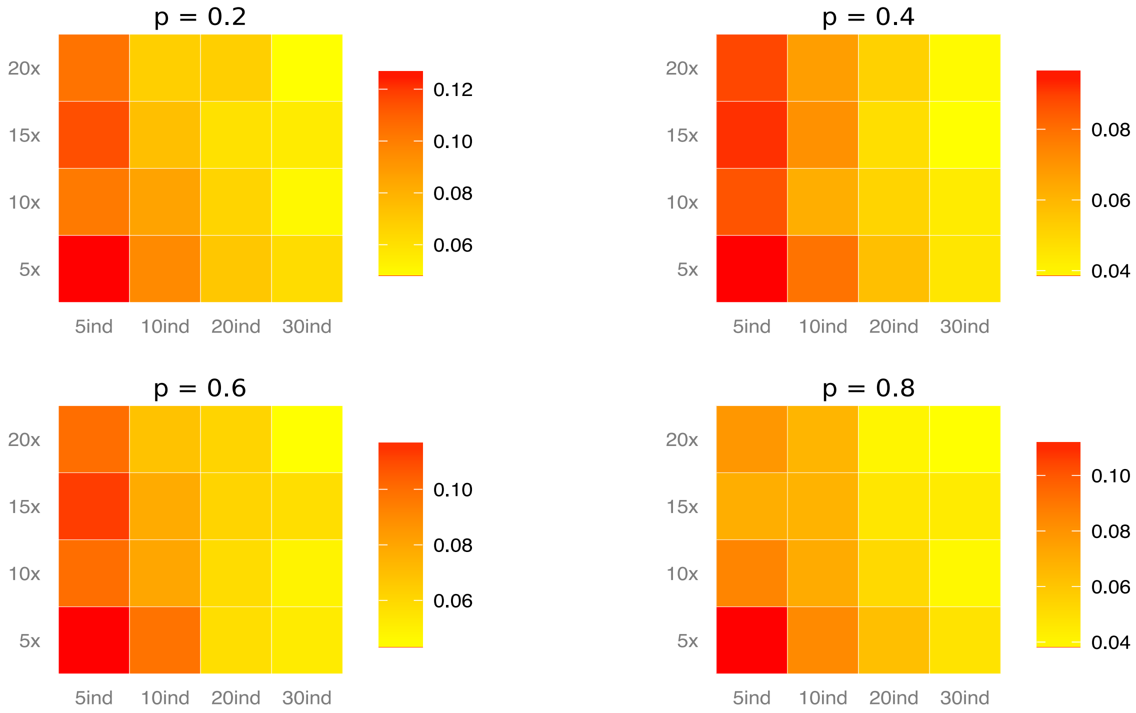
Symbol	**Description**
L	The number of loci.
ℓ	Index for loci ($\ell \in \{1, \ldots, L\}$).
N	Total number of individuals sequenced.
i	Index for individuals ($i \in \{1, \ldots, N\}$).
ψ	The ploidy-level of individuals in the population (e.g., tetraploid: ψ=4).
p_ℓ	Frequency of the reference allele 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.
λ	Average number of reads sequenced for an individual at a given locus (coverage).

Figure 1: The cartoon here illustrates the hierarchical Bayesian model for biallelic SNP frequencies of Buerkle & Gompert (2013). Here it is modified for a tetraploid, but can be used for any ploidy level. We will consider the blue allele to be ancestral (as determined by the outgroup, O) and red will be the derived allele. The ultimate goal is to approximate the posterior distribution $P(p, g|R^b, \epsilon)$, which is the distribution of the frequency of the derived allele in the population at a particular locus and the genotypes at that locus given the number of sequencing reads having the derived allele. The procedure for estimating this distribution is to use Markov chain Monte Carlo methods to draw samples from the posterior distribution of allele frequencies and genotypes given sequencing reads and error. In this example, the genotype (g) is drawn from a Binomial(4,p), where p is modeled by a Beta(1,1) distribution, and is equal to 3. The likelihood is also given by a binomial distribution with the observed number of successes equal to 10 reads given parameters n equal to 16 total reads and *prob* equal to $\frac{3}{4}$. The equations in red boxes indicate how one iteration of the sampling would proceed. First a proposed value for p is randomly selected and is modeled by a Beta(1,1). This value for p is then be used as the *prob* parameter to draw a random genotype from the binomial distribution. This genotype is then divided by the ploidy level to give the value of the *prob* parameter in the binomial likelihood.



$$\boxed{p = rbeta(\alpha = 1, \beta = 1)} \quad \boxed{g = rbinom(n = 4, prob = p)} \quad \boxed{P(R^b = 10|g = 3, \epsilon = 0.01) = \binom{16}{10} \frac{3}{4}^{10} \left(1 - \frac{3}{4}\right)^{6}}$$

Figure 2: Heat maps representing the relative error of the model for estimating allele frequencies across various simulation conditions.


p = 0.2
20x
15x
10x
5x
5ind
10ind
20ind
30ind
p = 0.4
20x
15x
10x
5x
5ind
10ind
20ind
30ind
p = 0.6
20x
15x
10x
5x
5ind
10ind
20ind
30ind
p = 0.8
20x
15x
10x
5x
5ind
10ind
20ind
30ind