SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data

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

[1] Dept. of Evolution, Ecology, and Organismal Biology, and
[2] Dept. of Statistics, The Ohio State University, Columbus, OH, USA.

*To whom correspondence should be addressed.

Abstract

Motivation: Genotyping and parameter estimation using high throughput sequencing data are everyday tasks for population geneticists, but methods developed for diploids are typically not applicable to polyploid taxa. This is due to their duplicated chromosomes, as well as the complex patterns of allelic exchange that often accompany whole genome duplication (WGD) events. For WGDs within a single lineage (autopolyploids), inbreeding can result from mixed mating and/or double reduction. For WGDs that involve hybridization (allopolyploids), alleles are typically inherited through independently segregating subgenomes.

Results: We present two new models for estimating genotypes and population genetic parameters from genotype likelihoods for auto- and allopolyploids. We then use simulations to compare these models to existing approaches at varying depths of sequencing coverage and ploidy levels. These simulations show that our models typically have lower levels of estimation error for genotype and parameter estimates, especially when sequencing coverage is low. Finally, we also apply these models to two empirical data sets from the literature. Overall, we show that the use of genotype likelihoods is a promising approach for conducting population genomic inferences in polyploids.

Availability: A C++ program, EBG, is provided to perform inference using the models we describe. It is available under the GNU GPLv3 on GitHub: https://github.com/pblischak/polyploid-genotyping.

Contact: blischak.4@osu.edu.

Supplementary information: Supplementary data are available online.

1 Introduction

The discovery and analysis of genetic variation in natural populations is a central task of evolutionary genetics, with applications ranging from the inference of population structure and patterns of historical demography, detecting selection and local adaptation, and performing genetic association studies. The ability to use high throughput sequencing technologies to detect variants across the genome has further advanced our understanding of the impact of evolutionary forces on genetic diversity in populations. However, the nature of data sets collected using high throughput sequencing often require special considerations regarding sequencing error and, especially, the level of sequencing coverage. Common approaches for dealing with low-coverage sequence data use genotype likelihoods to integrate over the uncertainty of inferring genotypes when estimating other parameters [allele frequencies, inbreeding coefficients, population differentiation, etc.] (e.g., Martin *et al.*, 2010; Li, 2011; Nielsen *et al.*, 2011, 2012; Fumagalli *et al.*, 2013; Vieira *et al.*, 2013; Huang *et al.*, 2016, among others). Genotype likelihoods for biallelic SNPs are calculated as the probability of the sequencing read data mapping to a variable site (total number of reads, number of reads with the alternative allele, and probability of sequencing error) given the possible values of the genotypes (e.g., typically 0, 1, or 2 for the number of copies of the alternative allele in diploids). When combined with computationally efficient algorithms for inference, these models are the primary tools used for conducting population genetic analyses from high throughput data.

Although the theory for these models is well established for diploids and even special cases of higher ploidy samples (treated equivalently to pooled samples of multiple diploids), the application of these tools to taxa that have experienced a recent whole genome duplication (WGD) is currently limited (McKenna *et al.*, 2010; DePristo *et al.*, 2011; Li, 2011). This is due in part because of ambiguity in the copy number of each allele in the genotype of a polyploid, a phenomenon referred to as allelic dosage uncertainty (Blischak *et al.*, 2016). Another important aspect of polyploid evolution to consider is that the occurrence of WGD can have an impact on how alleles are exchanged in a population, making the assumption of randomly inherited alleles inappropriate. Together these two factors have limited the widespread application of population genomic tools to gain insights about levels of genetic variation following WGD. Given both the evolutionary and economic importance of many of these organisms (e.g., agricultural crops, farmed fishes), the development of methods that can accommodate more complex patterns of inheritance is critical for the study of polyploids (Stebbins, 1950; Grant, 1971; Otto and Whitton, 2000; Soltis and Soltis, 2000; Soltis *et al.*, 2014).

In this paper we present two new models for SNP genotyping in polyploids using high throughput sequencing data. The models correspond to two different ways in which polyploids can be formed: WGD within a lineage (autopolyploid) or involving hybridization between two lineages (allopolyploid). The former builds off of previous work to relax the assumption of Hardy-Weinberg equilibrium by including inbreeding (Blischak *et al.*, 2016) and the latter provides a framework for separately determining the genotypes within the two genomes that compose the allopolyploid (typically referred to as subgenomes). We test our models using a wide range of simulations and describe our numerical approach for parameter estimation using the expectation maximization (EM) algorithm (Dempster *et al.*, 1977). For comparison, we analyzed our simulated data sets using two additional approaches based on models that assume either Hardy Weinberg or equal genotype probabilities. Finally, we also test the models on empirical data sets collected for a diploid-allotetraploid species pair from the genus *Betula* (birch trees) and a mixed-ploidy grass species, *Andropogon gerardii*. Overall, we demonstrate that genotype uncertainty resulting from both low-coverage sequencing data and allelic dosage uncertainty can be overcome in polyploids using genotype likelihoods.

2 Models

Assumptions: For each of the models below, we assume that SNPs are biallelic, and that loci and individuals are independent. For the autopolyploid model, we do not directly include double reduction (but see **Discussion**). For the allopolyploid model, we assume that subgenomes are independent and that they do not interact during meiosis (i.e., no homoeologous recombination).

Notation for each model is introduced in the descriptions we provide below and is also summarized in Table 1. Throughout the paper, we use boldface letters to denote an array of the respective parameter across either individuals (N), loci (L), or both (e.g., $\boldsymbol{p} := p_1, \ldots, p_L$, $\boldsymbol{F} := F_1, \ldots, F_N$, and $\boldsymbol{G} := g_{11}, g_{12}, \ldots, g_{N(L-1)}, g_{NL}$).

2.1 Autopolyploid Model

The genotype for a biallelic SNP in an autopolyploid with K sets of chromosomes has $K+1$ possible values. For example, using A and a to denote the two alleles, an autotetraploid can have genotypes equal to $AAAA$, $AAAa$, $AAaa$, $Aaaa$, or $aaaa$ (e.g., $g_{i\ell}$ = 0, 1, 2, 3, or 4, if a is the alternative allele; $i = 1, \ldots, N$ and $\ell = 1, \ldots, L$). A simple extension of the typical binomial sampling (Hardy Weinberg; HW) model used for diploids but with larger sample size to accommodate higher ploidy levels has been used previously (Li, 2011; Blischak *et al.*, 2016). However, inbreeding in various forms can bias inferences made when HW equilibrium is assumed. Vieira *et al.* (2013) introduced a genotype prior to include inbreeding either per-site or per-individual for a sample of diploids (implemented in the programs ngsF and ANGSD). This model used a formulation for generalized HW that includes the inbreeding coefficient, F, which is the probability that two alleles are identical by decent (ibd). Instead of using a generalized HW formulation for autopolyploids, we used the Balding-Nichols beta-binomial model (Balding and Nichols, 1995, 1997; Bradburd *et al.*, 2013), which also models the probability of two alleles being ibd but is more easily extended to higher ploidy levels by not directly enumerating all combinations of allele draws for the genotype of an autopolyploid.

Given genotype values at L loci for N individuals each of ploidy m_i, we model individual genotypes at each locus ($g_{i\ell} = 0, \ldots, m_i$ copies of the alternative allele) as a beta-binomial random variable. The log likelihood of the genotype data given the allele frequency at each site (p_ℓ) and the per-individual inbreeding coefficients (F_i) is then

$$\log \mathcal{L}(\boldsymbol{p}, \boldsymbol{F}; \boldsymbol{G}) = \sum_i \sum_\ell \log P(g_{i\ell} | p_\ell, F_i)$$

$$= \sum_i \sum_\ell \log \frac{\mathcal{B}\left(g_{i\ell} + p_\ell \frac{1-F_i}{F_i}, m_i - g_{i\ell} + (1-p_\ell)\frac{1-F_i}{F_i}\right)}{\mathcal{B}\left(p_\ell \frac{1-F_i}{F_i}, (1-p_\ell)\frac{1-F_i}{F_i}\right)}. \tag{1}$$

where $\mathcal{B}(\alpha, \beta)$ represents the beta function with parameters α and β. Since genotypes must be inferred from sequence data ($d_{i\ell}$; see **Methods**), we can also account for this uncertainty by summing over genotypes to get the likelihood of the sequence data given allele frequencies and inbreeding coefficients by including genotype likelihoods:

$$\log \mathcal{L}(\boldsymbol{p}, \boldsymbol{F}; \boldsymbol{D})$$

$$= \sum_i \sum_\ell \log \left[\sum_a P(d_{i\ell} | g_{i\ell} = a) P(g_{i\ell} = a | p_\ell, F_i) \right]. \tag{2}$$

Because maximization of the log likelihood is encumbered by the logarithm of the sum over genotypes, we instead use an expectation conditional maximization algorithm to obtain maximum likelihood (ML) estimates for $\boldsymbol{p}$ and $\boldsymbol{F}$ (Meng and Rubin, 1993). Since an analytical solution for the maximization step is not readily available, we instead employ numerical maximization of the likelihood using Brent's method (Brent, 1973). Then, given the ML parameter estimates, we can calculate the posterior probability of the genotype of each individual at each locus using Bayes' theorem:

$$P(g_{i\ell} = a | d_{i\ell}) = \frac{P(d_{i\ell} | g_{i\ell} = a) P(g_{i\ell} = a | \hat{p}_\ell, \hat{F}_i)}{\sum_{a'=0}^{m_i} P(d_{i\ell} | g_{i\ell} = a') P(g_{i\ell} = a' | \hat{p}_\ell, \hat{F}_i)}, \tag{3}$$

for $a = 0, \ldots, m_i$.

2.2 Allopolyploid Model

Deviations from HW are evident in allopolyploids in that they have two (sometimes more) sets of chromosomes inherited from separate evolutionary lineages. When these sets of chromosomes (called *homoeologs*, or *homoeologous chromosomes*) segregate during meiosis, they are inherited separately from one another and should be treated independently. For example, the genotypes for a biallelic SNP in an allotetraploid could have values $AA|A'A'$, $AA|A'a'$, $AA|a'a'$, $Aa|a'a'$, or $aa|a'a'$. Here the vertical bar '|' denotes separation between the subgenomes and the $'$ indicates homoeologous alleles. With perfect knowledge about which alleles go with each subgenome, determining the genotypes could be done completely independently. However, if separate reference genomes for the homoeologous chromosomes are not available, all reads mapping to a variable position will not be separable into reads coming from one subgenome or the other. Thus, when considering a variable site across the full set of homoeologs, we need to account for the fact that the frequency of the alternative allele may not be the same in each subgenome due to their separate evolutionary histories. When we cannot separate reads, we can instead consider the overall genotype of an allopolyploid with two subgenomes as being a combination of the genotypes within the subgenomes. For example, a tetraploid with two diploid subgenomes can have an overall genotype of $0, \ldots, 4$ copies of the alternative allele, but each of these full genotypes can be found via a different combination of genotypes in the subgenomes: $\{0 = (0,0); 1 = (0,1),(1,0); 2 = (0,2),(2,0),(1,1); 3 = (1,2),(2,1); 4 = (2,2)\}$.

In general, for an allopolyploid that has two subgenomes with ploidy levels equal to m_{1i} and m_{2i}, there are a total of $(m_{1i}+1)\times(m_{2i}+1)$ genotype combinations to consider. The probabilities of these genotypes are then determined using the allele frequencies for the alternative allele in the subgenomes.

An obvious complication of not being able to separate the sequencing reads into sets coming from each subgenome is that it makes independently estimating the allele frequencies and genotypes impossible. However, it is sometimes the case that the parental species of the allopolyploid are known, which can help with inferring genotypes by providing an outside estimate of the allele frequencies within the subgenomes. For our model, we relax this use of outside knowledge further and assume that only a single parent has been identified. Arbitrarily designating the known parent as subgenome one, we treat the allele frequencies at each locus estimated in the parental population to be known ($\boldsymbol{p}_1^*$) and require only the estimation of the allele frequencies in subgenomes two ($\boldsymbol{p}_2$). We then model the full genotype in the allopolyploid as the sum of the two independent subgenomes with separate, and potentially unequal, allele frequencies. Using binomial distributions to model the genotype in each subgenome ($\boldsymbol{G}_1$ and $\boldsymbol{G}_2$), the log likelihood for known genotype data is given by

$$\log\mathcal{L}(\boldsymbol{p}_2;\boldsymbol{p}_1^*,\boldsymbol{G}_1,\boldsymbol{G}_2) = \sum_i\sum_\ell\Bigg[\log\binom{m_{1i}}{g_{1i\ell}}(p_{1\ell}^*)^{g_{1i\ell}}(1-p_{1\ell}^*)^{(m_{1i}-g_{1i\ell})} + \log\binom{m_{2i}}{g_{2i\ell}}(p_{2\ell})^{g_{2i\ell}}(1-p_{2\ell})^{(m_{2i}-g_{2i\ell})}\Bigg]. \quad (4)$$

The inclusion of genotype likelihoods is done in a similar way to the autopolyploid model, only now we are summing over the values of the genotypes in both subgenomes one and two. The log likelihood for observed sequence data given the allele frequencies in each of the subgenomes is

$$\log\mathcal{L}(\boldsymbol{p}_2;\boldsymbol{p}_1^*,\boldsymbol{D}) = \sum_i\sum_\ell\log\Bigg[\sum_{a_1}\sum_{a_2}P(d_{i\ell}|g_{i\ell}=g_{1i\ell}+g_{2i\ell}) \times P(g_{1i\ell}=a_1|p_{1\ell}^*)P(g_{2i\ell}=a_2|p_{2\ell})\Bigg]. \quad (5)$$

Because maximizing the log likelihood involves the logarithm of a double sum, we turn once again to the expectation maximization algorithm to obtain a ML estimate for the allele frequency at each locus in subgenome two (Dempster *et al.*, 1977). An analytical solution for the maximization step of the EM algorithm is given by

$$p_{2\ell}^{(t+1)} = \frac{\sum_i\sum_{a_1}\sum_{a_2}a_2P(g_{i\ell}=a_1+a_2|d_{i\ell},p_{1\ell}^*,p_{2\ell}^{(t)})}{\sum_i m_{2i}}. \quad (6)$$

Using these ML estimates, an empirical Bayes estimate of the genotypes within each of the subgenomes can be found using their joint posterior probability (note that subscripts i and ℓ are dropped for readability)

$$P(g_1=a_1,g_2=a_2|d) = \frac{P(d|g=g_1+g_2)P(g_1=a_1|p_1^*)P(g_2=a_2|\hat{p}_2)}{\sum_{a_1'}\sum_{a_2'}P(d|g=g_1+g_2)P(g_1=a_1'|p_1^*)P(g_2=a_2'|\hat{p}_2)}, \quad (7)$$

where $a_1=0,\ldots,m_{1i}$ and $a_2=0,\ldots,m_{2i}$.

Table 1. A key to the symbols and notation that are used in describing the autopolyploid and allopolyploid models. We use a either a bold or bold-capitalized letter when referring to the collection of parameters together (e.g., $\boldsymbol{G}$ refers to $g_{i\ell}$ for all individuals at all loci). Parameters within subgenomes for the allopolyploid model use the same symbol but with either a 1 or a 2 added as a subscript.

Symbol	Description
N, L	The number of individuals and loci sampled.
m_i	Ploidy level of individual i.
$d_{i\ell}$	Sequence data for individual i at locus ℓ (=$\{t_{i\ell},r_{i\ell},\epsilon_\ell\}$).
$t_{i\ell}$	Total number of reads for individual i at locus ℓ.
$r_{i\ell}$	Number of alternative allele reads for individual i at locus ℓ.
ϵ_ℓ	Average sequencing error at locus ℓ.
$g_{i\ell}$	Genotype for individual i at locus ℓ.
p_ℓ	Allele frequency at locus ℓ.
F_i	Inbreeding coefficient for individual i.

2.3 Other Approaches

We consider two additional approaches that use genotype priors that have been described in previous studies. The first is an implementation of the SAMtools Hardy Weinberg equilibrium prior (Li, 2011) and the second is a flat prior on genotypes that is similar to the model used by the Genome Analysis Toolkit (GATK; McKenna *et al.*, 2010). Other approaches that accommodate polyploids such as the FITTETRA package in R (Voorrips *et al.*, 2011) and the method of Maruki and Lynch (2017) were not considered here because they can only handle specific ploidy levels (triploids and/or tetraploids).

3 Methods

Genotype likelihoods were calculated using a simplified version of the SAMtools model by using average sequencing error values at each locus, ϵ_ℓ, across reads and individuals (Li, 2011). Then for the possible values of the genotype ($a=0,\ldots,m_i$), the probability of the read data, $d_{i\ell}=\{t_{i\ell},r_{i\ell},\epsilon_\ell\}$ ($t_{i\ell}=$ total read count, $r_{i\ell}=$ alternative allele read count), given the genotype, $g_{i\ell}$, is

$$P(d_{i\ell}|g_{i\ell}=a) = \binom{t_{i\ell}}{r_{i\ell}}f_\epsilon(a,m_i,\epsilon_\ell)^{r_{i\ell}} \times[1-f_\epsilon(a,m_i,\epsilon_\ell)]^{(t_{i\ell}-r_{i\ell})}, \quad (8)$$

where

$$f_\epsilon(x,y,e) = \frac{x}{y}(1-e)+\left(1-\frac{x}{y}\right)e. \quad (9)$$

3.1 Simulations

We generated sequencing read data with mean coverage per individual, per locus equal to 2x, 5x, 10x, 20x, 30x, and 40x, simulated from a Poisson distribution for 10 000 sites. The number of individuals was set to 25, 50, or 100, and we tested ploidy levels equal to 4, 6, and 8 (4=2+2, 6=2+4, and 8=4+4 for allopolyploids). Sequencing errors were drawn from a beta distribution with parameters $\alpha=1$ and $\beta=200$ (mean error ≈ 0.005)]. Allele frequencies were drawn from a truncated beta distribution with a minimum minor allele frequency of 5% and parameters $\alpha=\beta=0.01$. For the autopolyploid model, the values of the inbreeding coefficient were set to 0.1, 0.25, 0.5, 0.75, and 0.9. For the allopolyploid model, the allele frequencies simulated for subgenome one were treated as the reference panel. Genotypes were drawn according to their respective generating models (autopolyploid or allopolyploid), and the number of alternative reads for each individual at each locus was drawn from the binomial distribution in Eq. (8) given the total read count, genotype, and level of sequencing error. For each simulation, we evaluated estimation error using the root mean squared deviation (RMSD)

$$\text{RMSD} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(X_{est.}^{[r]} - X_{true})^2}\,, \qquad (10)$$

where R represents the number of replicates, $X_{est.}^{[r]}$ is the estimated value for replicate r, and X_{true} is the original value used to simulate the data.

To compare our models with other methods, we reused these simulated data as input for the estimation of genotypes and model parameters using priors that assume either Hardy Weinberg equilibrium or equal genotype probabilities (GATK-like). For the allopolyploid model, this also equates to ignoring the fact that genotypes are drawn from two independent subgenomes. Inference for the Hardy Weinberg model used the EM algorithm described in Li (2011). Genotyping based on the GATK-like model were calculated based on normalized genotype likelihoods as described in McKenna *et al.* (2010).

Comparisons for the autopolyploid model were based on the RMSD of four estimates of the inbreeding coefficient. The first of these was the estimate obtained by our ECM algorithm, which is built directly into the model. The other three estimates were calculated as a summary statistic from estimated genotypes for the three models (**Supplemental Materials**, Supplemental Text). We then also compared RMSD values of the estimated genotype values for the three methods. For the allopolyploid model, direct comparisons with models that assume Hardy Weinberg or uniform genotype priors are more difficult because they do not share the assumption of two subgenomes. Therefore, we focused on the accuracy of the models to infer the full genotype by again comparing RMSD values.

3.2 Empirical Data Analysis

3.2.1 *Andropogon gerardii*

We tested our autopolyploid model on an empirical data set collected in the grass species *Andropogon gerardii*. SNP data from McAllister and Miller (2016) were downloaded from Dryad as a VCF file (http://datadryad.org/resource/doi:10.5061/dryad.05qs7). The data were filtered using VCFtools with the following criteria: biallelic SNPs only, no more than 50% missing data per site, one SNP per 10 000 base pair window, and a minimum sequencing depth of five reads (Danecek *et al.*, 2011). The output from VCFtools was then converted to a plain text format containing the number of total reads and alternative allele reads per individual per site using a Perl script (`read-counts-from-vcf.pl`; available on GitHub). We then also removed any individuals with more than 50% missing data using an R script (`filter-inds.R`; available on GitHub). Since *A. gerardii* has two cytotypes (6N and 9N), we analyzed the hexaploid and nonaploid individuals separately and compared the estimates of the inbreeding coefficients across ploidy levels.

3.2.2 *Betula pubescens* and *B. pendula*

To test the allopolyploid model, biallelic SNP genotypes from Zohren *et al.* (2016) for the allotetraploid *Betula pubescens* and its putative diploid progenitor, *B. pendula*, were downloaded from Dryad (http://datadryad.org/resource/doi:10.5061/dryad.815rj). Treating the genotypes as known, we simulated read data and error values as before using Eq. (8) with beta distributed error values. We varied the level of sequencing coverage (5x, 10x, 20x) but did not alter the amount of missing data. Allele frequencies for *B. pendula* were estimated under the assumption of Hardy Weinberg equilibrium and disequilibrium to assess which was a better fit. These allele frequency estimates were then used as the reference panel for genotype estimation in *B. pubescens* using the allopolyploid model.

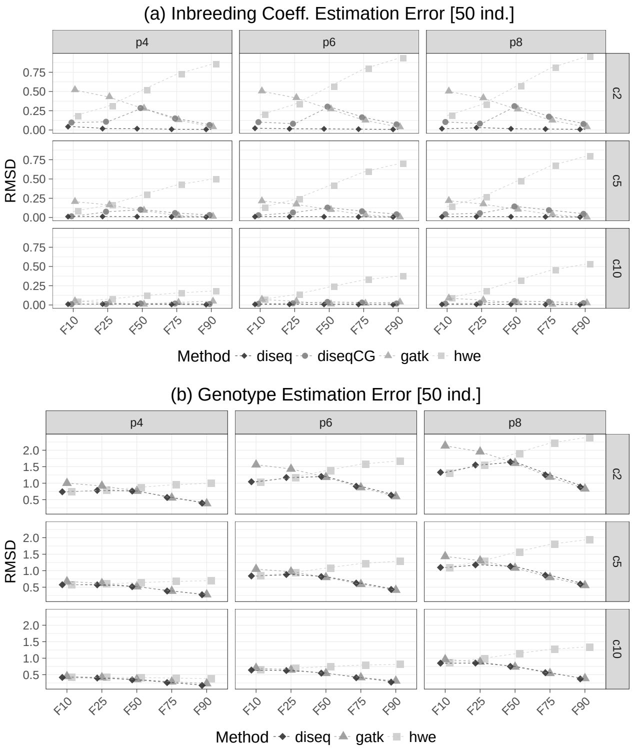

Fig. 1. RMSD values for simulations under the autopolyploid model with inbreeding for (a) estimated inbreeding coefficients and (b) estimated genotypes. Each individual plot within (a) and (b) displays the RMSD on the y-axis and inbreeding coefficients on the x-axis. Rows correspond with the depth of sequencing coverage (2x, 5x, 10x) and the columns correspond to the ploidy level (4, 6, 8). The different estimation methods (diseq, diseqCG, gatk, hwe) are represented by different shapes within each plot. (a) The inbreeding coefficient estimated by our model (diseq) is consistently the lowest across all depths of sequencing coverage, ploidy level, and level of inbreeding. (b) Genotypes estimated by our model are at least as accurate as the other methods and are not as affected by high or low levels of inbreeding.

3.3 Software and reproducibility

We have packaged our code for the EM/ECM algorithms in a command line C++ program called EBG, which we have included as part of a GitHub repository for this manuscript (doi:10.5281/zenodo.195779). This software includes our implementations of the autopolyploid (diseq), allopolyploid (alloSNP), Hardy Weinberg (hwe), and GATK-like (gatk) models for genotyping in polyploids. Code for the simulation study and empirical data analyses was written using a combination of the R statistical language and C++ through the use of the RCPP package (Eddelbuettel and François, 2011; Eddelbuettel, 2013; R Core Team, 2014). Figures were generated using the GGPLOT2 package in R (Wickham, 2009). Additional figure manipulations were done using Inkscape (https://inkscape.org/).

4 Results

4.1 Simulations

4.1.1 Autopolyploid model

Simulated read count data were generated to assess the impact of sequencing coverage and ploidy level on estimation error in autopolyploids using an expectation conditional maximization (ECM) algorithm. Convergence of the ECM algorithm depended on the number of individuals sampled, sequencing coverage, and ploidy. Each iteration of the algorithm employs Brent's method, itself an iterative maximization algorithm, resulting in slower M-steps than the other EM algorithms we describe. However, overall convergence was reached before the maximum number of allowed iterations (1000) in all cases, with analyses typically employing between 50–100 iterations.

For the estimation of individual inbreeding coefficients (F_i), Figure 1a shows the root mean squared deviation (RMSD) for estimated inbreeding coefficients for the four different estimation methods across ploidy levels and the three lowest levels of sequencing coverage (sample size of 50 individuals). Compared with the other methods that used called genotypes (diseqCG, hwe, gatk), the level of sequencing coverage and ploidy level had virtually no effect on estimation error using our model (diseq). For the other estimates, increasing sequencing coverage lowered estimation error as expected, and higher ploidy levels showed higher levels of error. However, inbreeding coefficients estimated from genotypes called from our model (diseqCG) did have lower RMSD values than the other methods, except when the inbreeding coefficient was 0.5, when the level of error was about the same. All of the methods except for Hardy Weinberg showed low levels of estimation error once the depth of sequencing reached 10x. Figures S1–S3 show the results for all simulated depths of sequencing (2x to 40x) and sample sizes (25, 50, and 100 individuals).

Our empirical Bayes approach for maximum a posteriori (MAP) genotype estimation resulted in a similar overall pattern of lower estimation error for increased sequencing coverage (Figure 1b). Interestingly, the other two methods for genotyping (gatk, hwe) showed opposing patterns of accuracy: the GATK-like model increased in accuracy with increasing levels of inbreeding but the Hardy Weinberg model had decreasing accuracy. Genotypes called by our method showed some dependence on the level of inbreeding with intermediate values having the most error. However, our method was still the most accurate across the range of inbreeding values simulated. Ploidy also had an impact on genotyping with higher ploidy levels having higher levels of estimation error. This is largely due to the fact that higher ploidy individuals have a larger number of possible values for the genotype and that the average sequencing coverage per allele (chromosome) is lower (e.g., 10x coverage in a tetraploid is on average 2.5x per allele but is 1.25x in an octoploid). Once the depth of sequencing reached 10x, the only model that still showed a higher level of error was the Hardy Weinberg model. Figures S4–S6 show the results for all simulated depths of sequencing (2x to 40x) and sample sizes (25, 50, and 100 individuals).

4.1.2 Allopolyploid model

Using the same general parameter settings as the simulations for the autopolyploid model (except for inbreeding), we calculated genotype likelihoods by simulating read data from genotypes generated under the model from Eq. (4). The ploidy of each subgenome was as follows: tetraploids = diploid + diploid, hexaploid = diploid + tetraploid, and octoploid = tetraploid + tetraploid. Our expectation maximization algorithm for this model was slow to converge, despite each maximization step taking less time when compared with the autopolyploid model. Analyses never reached the upper limit on the number of iterations (again 1000) but some analyses did not reach convergence until over 900 iterations had been run. To make analyses with this model more practical, we reanalyzed all simulated data sets using only 100 EM iterations followed by direct maximization of the observed data log likelihood function in Eq. (5) using Brent's method (EM+Brent).

Comparing our model with other genotype priors (Hardy Weinberg, GATK) only allowed us to consider the full genotype estimates from the different methods. Figure 2 shows the level of estimation error for each of the three genotyping methods for each ploidy level across all depths of sequencing coverage. For low depths of sequencing, genotyping with the GATK-like model resulted in high levels of error. As the depth of coverage increased, the three methods converged. However, this was dependent on the ploidy level: octoploids required a higher depth of sequencing for the GATK model than tetraploids or hexaploids to achieve the same level of accuracy. The Hardy Weinberg prior performed almost identically to our allopolyploid model, most likely as a result of our assuming Hardy Weinberg within the subgenomes of the allopolyploid.

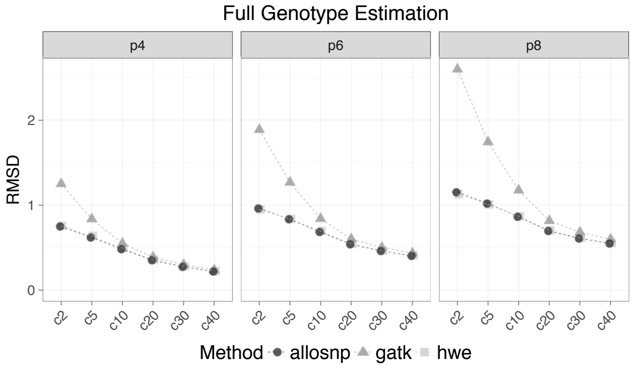

Fig. 2. RMSD values for full genotype estimation. Sequencing coverage is on the x-axis and RMSD values are on the y-axis. Each column represents a different ploidy level and the three methods used (allosnp, gatk, hwe) are represented by different shapes. For low levels of sequencing coverage, the allosnp and hwe models have much lower levels of estimation error when compared with the gatk model. The level of sequencing coverage required for the three methods to converge in error rate depends on the ploidy level, with tetraploids needing less coverage and octoploids needing more.

We also assessed the accuracy of the model for estimating parameters based on the true values used for the simulations. Allele frequency estimates for subgenome two improved as the number of individuals and sequencing coverage were increased (Figure S7). Tetraploids showed the highest estimation error for subgenome two (diploid), followed by octoploids and hexaploids (tetraploid subgenomes), respectively. This pattern with hexaploids and octoploids is counterintuitive considering that higher ploidy levels typically result in better estimates of allele frequencies since more alleles are sampled from the population (Blischak *et al.*, 2016). However, the tetraploid subgenomes in the hexaploid and octoploid individuals do not show similar levels of error as would be expected. This is likely a result of subgenome one having higher ploidy in the octoploid simulations, resulting in a larger number of possible genotype combinations and therefore higher estimation error (octoploid: $5 \times 5 = 25$ vs. hexaploid: $3 \times 5 = 15$). Figures S8 and S9 show the error in genotype estimation in subgenome one and two, respectively. Here we again observe that higher ploidy levels have higher levels of estimation error for genotypes. Overall, genotype estimates were inferred with higher error for subgenome two. This result makes sense given that we treat the allele frequencies for subgenome one as known but have to estimate them in subgenome two.

4.2 Empirical Data Analysis

4.2.1 *Andropogon gerardii*

Analyzing and filtering the data sets for hexaploid and nonaploid *A. gerardii* separately resulted in slightly different numbers of loci (6N: 83 individuals, 6 928 loci; 9N: 70 individuals, 6 887 loci). The average depth of sequencing coverage was 10.9x for hexaploids and 10.8x for nonaploids. Though levels of inbreeding for both cytotypes were low, nonaploids showed significantly higher levels of inbreeding than hexaploids (Figure 3a; $F_{1,151} = 36.14$, $p = 1.3 \times 10^{-8}$).

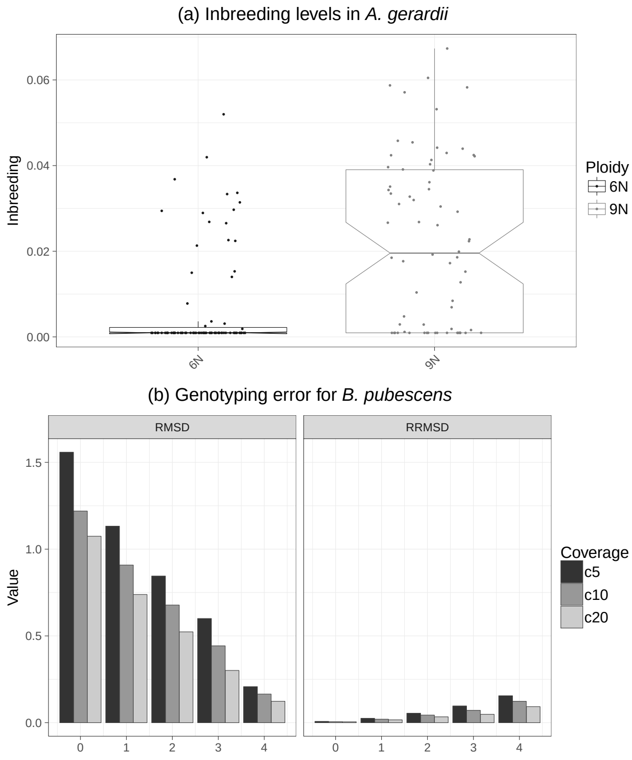

Fig. 3. Results of empirical data analyses. (a) Levels of inbreeding in *Andropogon gerardii*. Inbreeding in the two cytotypes of *A. gerardii* is generally low, but the nonaploid (9N) samples have higher levels of inbreeding on average. (b) Genotype estimation error in *Betula pubescens*. The left panel shows the RMSD values for each of the possible genotypes (0–4). The right panel shows a relative measure of the RMSD where each value is weighted the occurrence of the particular genotype in the data set (see text for details).

4.2.2 *Betula pubescens* and *B. pendula*

The data set for the species of *Betula* consisted of 130 individuals for *B. pubescens* and 34 individuals for *B. pendula* with genotype data for 49 021 loci. For *B. pendula*, we inferred allele frequencies and genotypes assuming Hardy Weinberg (HW), as well as using our model for individual inbreeding coefficients. The log likelihoods of the two models were very similar and most of the inbreeding coefficients were estimated to be close to 0, so we used the allele frequency estimates from the HW model as the reference panel for the allopolyploid model. After estimating the parameters of this model for *B. pubescens* using the EM+Brent method, we assessed the accuracy of our empirical Bayes genotype estimates by comparing them to the original data set using the root mean squared deviation. This comparison is shown for each of the possible genotype values (0–4) in Figure 3b. The left panel shows the RMSD for each genotype value and the right panel shows a weighted measure of the RMSD that corresponds to the relative amount of error based on the frequency of that genotype in the original data set. For example, we do a poor job of estimating the genotype when the true value is 0, but very few of the true genotypes have that value (∼0.5%), so the relative contribution to the overall error is much less. In contrast, roughly 75% of the true genotypes have a value of 4, which is the value that we estimate the best. In addition, many of the genotypes in *B. pendula* were equal to 2 (∼88%), so the estimates of the allele frequencies were very close to 1.0, which could have led to more error prone estimates of the genotypes in *B. pubescens* when using them as the reference panel.

5 Discussion

The ability to genotype individuals in a population can be an underappreciated task, even though it is typically the first step of any population genetic analysis. This is especially true for populations of polyploids, where genotyping is further complicated by duplicated chromosomes and their subsequent genome evolution. Until recently, genotyping polyploids using high throughput sequencing data was only possible in model organisms with reference genomes and/or subgenomes. However, recent studies have begun genotyping SNPs in both model and non-model organisms using whole genome resequencing and reduced representation methods such as restriction-site associated DNA sequencing (RADseq) and its variants (e.g., Arnold *et al.*, 2015; Douglas *et al.*, 2015; Cornille *et al.*, 2016; Zohren *et al.*, 2016). Most of these studies used already existing pieces of software to perform SNP calling and genotyping [e.g., Genome Analysis Toolkit (McKenna *et al.*, 2010), UNEAK (Lu *et al.*, 2012), TASSEL-GBS (Glaubitz *et al.*, 2014)] but others used novel approaches for estimating genotypes (e.g., Voorrips *et al.*, 2011; Zohren *et al.*, 2016; Maruki and Lynch, 2017). Using these tools to identify variants in combination with our models for genotyping and parameter estimation will be especially helpful for studies with low-coverage data.

The use of genotype likelihoods for arbitrary ploidy levels using EM algorithms and the inclusion of deviations from Hardy Weinberg is a much needed addition to allow for the analysis of low-coverage sequencing data in autopolyploids and allopolyploids. We have also written algorithms for genotype and parameter estimation using Hardy Weinberg and flat genotype priors. The use of EM algorithms improves greatly on our previous approach, which used Gibbs sampling for a model that assumed Hardy Weinberg equilibrium for autopolyploids (Blischak *et al.*, 2016). The computational burden of inferences based on Markov chain Monte Carlo (MCMC) rendered this approach impractical for most reasonably sized population genomic data sets. Our new implementation of the Hardy Weinberg model, plus the others, now run in a matter of minutes, rather than hours or possibly even days using MCMC.

Though our models were accurate for many of our simulations and outperformed comparable methods at low depths of sequencing coverage, it is important to consider scenarios when their assumptions are inappropriate. One concern for autopolyploids is the occurrence of double reduction, a process by which alleles in the genotype are identical by decent due to the segregation of sister chromatids to the same gamete during meiosis (Haldane, 1930). As we mentioned before, our model does not directly estimate rates of double reduction. However, because double reduction leads to identity by descent, it contributes to deviations from Hardy Weinberg that are similar to inbreeding. Therefore, our model for individual inbreeding coefficients should be able to accommodate, but not specifically estimate, double reduction.

Allopolyploids present a different set of challenges that are a result of their hybrid origins. In our model, we assume that the two subgenomes of the allopolyploid are completely independent. However, homoeologous recombination can make this assumption inappropriate. Future work that models this exchange of alleles between subgenomes will be an important extension of the model we presented here. Another potential avenue would be to develop ways to use more parental information, as well as demographic parameters to account for the amount of divergence between the allopolyploid and its parents. Models that help to identify parental taxa will also be an important contribution for future research on allopolyploids.

6 Conclusions

As methods for the analysis of polyploid data continue to be developed, we are hopeful that the barriers to more widespread study of these taxa will begin to drop. The prevalence of polyploidy in plants and other groups of eukaryotes, including fish, amphibians, and fungi, make these methods critically important for furthering our understanding of the impact of WGD on genetic diversity (Rogers, 1973; Otto and Whitton, 2000; Gregory and Mable, 2005; Wood *et al.*, 2009). Of the main problems that complicate population genetics in polyploids, we believe that modeling allelic inheritance is the most difficult. It was previously thought that ambiguity in the dosage of alleles in the genotype was also a major complicating factor. As we have shown in this study, dealing with this type of genotype uncertainty can easily be overcome using genotype likelihoods.

Acknowledgements

The authors thank members of the Wolfe lab, M. Fumagalli, and an anonymous reviewer for their helpful comments on this manuscript. We also thank J. Zohren and her colleagues, C. McAllister, and A. Miller for making their data sets publicly available. Early versions of these models were presented to X. He, J. Novembre, M. Stephens, and their lab members, and we thank them for their constructive feedback and advice (especially regarding the use of EM algorithms).

Funding

This work was supported by grants from the National Science Foundation to PDB and ADW (DDIG; DEB-1601096) and to ADW and LSK (DEB-1455399).

References

Arnold, B., Kim, S.-T., and Bomblies, K. (2015). Single geographic origin of a widespread autotetraploid arabidopsis arenosa lineage followed by interploidy admixture. *Molecular Biology and Evolution*, **32**, 1382–1395.

Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differen-tiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Balding, D. J. and Nichols, R. A. (1997). Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity*, **108**, 583–589.

Blischak, P. D., Kubatko, L. S., and Wolfe, A. D. (2016). Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Molecular Ecology Resources*, **16**, 742–754.

Bradburd, G., Ralph, P., and Coop, G. (2013). Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution*, **67**, 3258–3273.

Brent, R. P. (1973). *Algorithms for minimization without derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

Cornille, A., Salcedo, A., Kryvokhyzha, D., Glémin, S., Holm, K., Wright, S. I., and Lascoux, M. (2016). Genomic signature of successful colonization of Eurasia by the allopolyploid shepherd's purse (*Capsella bursa-pastoris*). *Molecular Ecology*, **25**, 616–629.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**, 1–38.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature Genetics*, **43**(5), 491–498.

Douglas, G. M., Gos, G., Steige, K. A., Salcedo, A., Holm, K., Josephs, E. B., Arunkumar, R., Ågren, J. A., Hazzouri, K. M., Wang, W., Platts, A. E., Williamson, R. J., Neuffer, B., Lascoux, M., Slotte, T., and Wright, S. I. (2015). Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid *Capsella bursa-pastoris*. *Proceedings of the National Academy of Sciences USA*, **112**, 2806–2811.

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

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

Fumagalli, M., Vieira, F. G., Korneliussen, T., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., and Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, **195**, 979–992.

Glaubitz, J., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., and Buckler, E. S. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, **9**, e90346.

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

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

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

Huang, G., Wang, S., Wang, X., and You, N. (2016). An empirical Bayes method for genotyping and SNP detection using multi-sample next-generation sequencing data. *Bioinformatics*, **32**, 3240–3245.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., Buckler, E. S., and Costich, D. E. (2012). Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Genetics*, **9**, e1003215.

Martin, E. R., Kinnamon, D. D., Schmidt, M. A., Powell, E. H., Zuchner, S., and Morris, R. W. (2010). SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, **26**, 2803–2810.

Maruki, T. and Lynch, M. (2017). Genotype calling from population-genomic sequencing data. *G3: Genes, Genomes, Genetics*, **Early Online**, DOI: https://doi.org/10.1534/g3.117.039008.

McAllister, C. A. and Miller, A. J. (2016). Single nucleotide polymorphism discovery via genotyping by sequencing to assess population genetic structure and recurrent polyploidization in *Andropogon gerardii*. *American Journal of Botany*, **103**, 1314–1325.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotyping and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.

Nielsen, R., Korneliussen, T., Albrechtsen, A., and Li, Y. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, **7**, e37558.

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

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

Rogers, J. D. (1973). Polyploidy in Fungi. *Evolution*, **27**, 153–160.

Soltis, D. E., Visger, C. J., and Soltis, P. S. (2014). The polyploidy revolution then...and now: Stebbins revisited. *American Journal of Botany*, **101**, 1057–1078.

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

Stebbins, G. L. (1950). *Variation and evolution in plants*. Columbia University Press.

Vieira, F. G., Fumagalli, M., Albrechtsen, A., and Nielsen, R. (2013). Estimating inbreeding coefficients from NGS data: impact on genotype calling and allele frequency estimation. *Genome Research*, **23**, 1852–1861.

Voorrips, R. E., Gort, G., and Vosman, B. (2011). Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC Bioinformatics*, **12**, 172.

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

Wood, T. E., Takebayashi, N., Barker, M. S., Mayrose, I., Greenspoon, P. B., and Rieseberg, L. H. (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences*, **106**, 13875–13879.

Zohren, J., Wang, N., Kardailsky, I., Borrell, J. S., Joecker, A., Nichols, R. A., and Buggs, R. J. A. (2016). Unidirectional diploid–tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. *Molecular Ecology*, **25**, 2413–2426.