Applications in Plant Sciences

SOFTWARE NOTE

nQuack: An R package for predicting ploidal level from sequence data using site-based heterozygosity

Michelle L. Gaynor[1,2] | **Jacob B. Landis**[3] | **Timothy K. O'Connor**[4] |
Robert G. Laport[5] | **Jeff J. Doyle**[3] | **Douglas E. Soltis**[1,2] |
José Miguel Ponciano[2] | **Pamela S. Soltis**[1]

[1]Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, USA

[2]Department of Biology, University of Florida, Gainesville, Florida 32611, USA

[3]School of Integrative Plant Science, Cornell University, Ithaca, New York 14850, USA

[4]Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA

[5]Department of Biology, The College of Idaho, Caldwell, Idaho 83605, USA

Correspondence

Michelle L. Gaynor, Florida Museum of Natural History, University of Florida, Dickinson Hall, 1659 Museum Rd., Gainesville, Florida 32611, USA.
Email: shellyleegaynor@gmail.com

This article is part of the special issue "Twice as Nice: New Techniques and Discoveries in Polyploid Biology."

Abstract

Premise: Traditional methods of ploidal-level estimation are tedious; using DNA sequence data for cytotype estimation is an ideal alternative. Multiple statistical approaches to leverage sequence data for ploidy inference based on site-based heterozygosity have been developed. However, these approaches may require high-coverage sequence data, use inappropriate probability distributions, or have additional statistical shortcomings that limit inference abilities. We introduce nQuack, an open-source R package that addresses the main shortcomings of current methods.

Methods and Results: nQuack performs model selection for improved ploidy predictions. Here, we implement expectation maximization algorithms with normal, beta, and beta-binomial distributions. Using extensive computer simulations that account for variability in sequencing depth, as well as real data sets, we demonstrate the utility and limitations of nQuack.

Conclusions: Inferring ploidy based on site-based heterozygosity alone is difficult. Even though nQuack is more accurate than similar methods, we suggest caution when relying on any site-based heterozygosity method to infer ploidy.

KEYWORDS

copy number variation, expectation maximization, ploidal inference, ploidy, polyploidy

Whole-genome duplication (WGD), or polyploidy, is ubiquitous across the green plant tree of life, with all extant angiosperms exhibiting evidence of at least one ancient WGD event (Jiao et al., 2011; Soltis et al., 2015; Landis et al., 2018; One Thousand Plant Transcriptomes Initiative, 2019). Identifying ploidal diversity is a crucial first step to understanding the impact of WGD on patterns of biodiversity. Direct estimation is achieved through chromosome counting at either mitosis or meiosis. However, indirect estimation (e.g., flow cytometry, stomatal guard cell measurements, pollen size, leaf spectra) can be used for broad surveys of select taxa when complemented with known chromosome numbers and/or ploidal levels (Masterson, 1994; Beaulieu et al., 2008;

Blonder et al., 2020; Sanders, 2021; Sliwinska et al., 2021). The application of flow cytometry to determine ploidal level in naturally occurring populations (Galbraith et al., 1983; Keeler et al., 1987) has been fundamental to understanding evolution and ecology of mixed-ploidy populations. Despite the utility of laboratory-based approaches and the extension of flow cytometry to dried samples (Galbraith et al., 1983; Keeler et al., 1987; Suda and Trávníček, 2006; Sliwinska et al., 2021), the process remains specialized and may involve the use of laboratory equipment that is difficult to access. Therefore, using DNA sequence data for ploidal-level prediction affords a great opportunity to streamline estimation while revolutionizing our understanding of chromosome evolution.

To date, multiple statistical approaches to use DNA sequence data for the prediction of ploidy have been developed based on (1) *k*-mer and (2) site-based heterozygosity. Both of these general methods for ploidal-level prediction require statistical tests to assign ploidal level to a sample; the statistical approach varies among available software.

k-mer-based ploidal-level prediction relies on a *k*-mer profile, which classifies the frequency of each distinct *k*-mer found across the data set. *k*-mers are strings of length *k*, often 21 bases (Vurture et al., 2017), that are composed of a specific sequence of nucleotides. Popular methods for *k*-mer-based ploidal-level prediction are tetmer (Becher et al., 2022) and smudgeplot, which plots minor allele frequency by total coverage to predict copy number variants (Ranallo-Benavidez et al., 2020). These methods have been recently expanded to single-cell ATAC-seq data (Takeuchi and Kato, 2024). However, a limitation of these methods is that at least 15–25× sequence coverage per homolog is required (Ranallo-Benavidez et al., 2020).

Site-based heterozygosity methods rely on biallelic single-nucleotide polymorphisms (SNPs) within an individual and the expected number of copies of each base at that SNP. For example, in a diploid individual, at a biallelic site with alleles A and B, about 50% of all nucleotides sequenced are expected to represent allele A. Comparatively, in a triploid, at a site with alleles A and B, 33% of the nucleotides are expected to be allele A and 67% allele B, or vice versa

(Figure 1). Alleles are called based on sequence data that have been aligned to a single genomic reference and are therefore calculated based on the copies of nucleotides at sites with only two nucleotides. The types of heterozygotes discussed here are not equivalent to heterozygous genotypes, which are known to differ between autopolyploids and allopolyploids (i.e., up to four alleles at a locus vs. up to two alleles at two homeologous loci, respectively). Therefore, the types of heterozygotes considered by these methods do not differ based on the mode of origin; however, the proportion of each type across the genome is expected to differ among individuals of autopolyploid or allopolyploid origin (see Methods, below).

The most commonly used site-based heterozygosity software is nQuire (Weiß et al., 2018), but additional software exists for de novo sequences (Sun et al., 2023). As for *k*-mer-based estimation, sequence coverage per site of at least 20–25× is recommended for the use of nQuire (Weiß et al., 2018). In addition, the performance and limitations of nQuire are poorly understood in terms of accuracy. Combining nQuire's model inference with additional data, such as genome size estimates, and with goodness-of-fit tests has been suggested (Viruel et al., 2019). Notably, nQuire's accuracy and limitations were assessed using only genome resequencing data for only five samples representing two taxonomic groups (Weiß et al., 2018). Numerous studies have since identified inconsistencies between nQuire's estimates and indirect or direct ploidal estimates (Jantzen et al., 2022; Landis and Doyle, 2023; Folk et al., 2024).
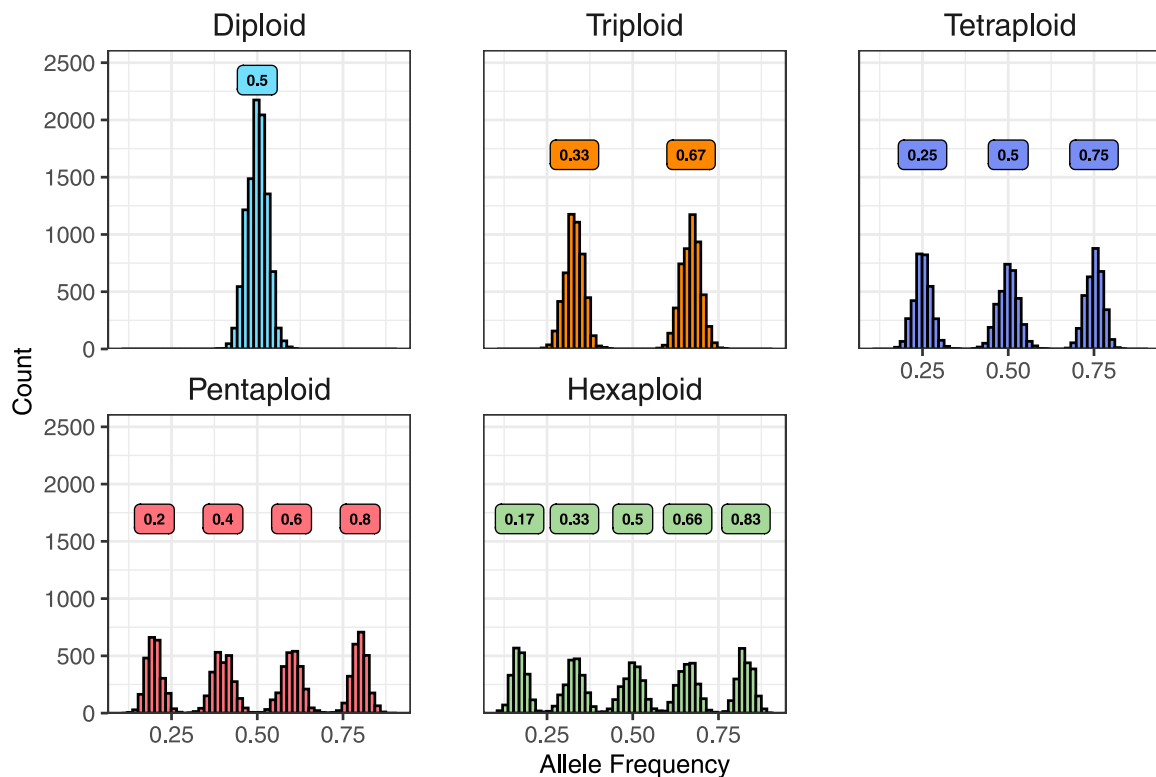
FIGURE 1 Expected site-based allele frequencies at biallelic sites for diploid, triploid, tetraploid, pentaploid, and hexaploid individuals.

Moreover, in regard to concerns of accuracy, guidelines for data preparation are limited, as it is unknown how nQuire predictions are influenced by the number of sites, sequencing coverage, and amount of variance or noise in a data set. In real data sets, this noise can be introduced through sequencing error or general mapping error, as well as through the inclusion of non-single-copy loci.

Here we introduce nQuack, an R package that (1) provides expanded tools and implementations to improve site-based heterozygosity inferences of ploidal level, and (2) rigorously evaluates the accuracy of ploidy inference from site-based heterozygosity data with comparison to nQuire. Specifically, nQuack implements expectation maximization algorithms with normal, beta, and beta-binomial distributions to identify the ploidal level (ranging from diploid to hexaploid) of samples based on DNA sequence data, building upon the framework proposed by nQuire. We designed three new implementations of the expectation maximization algorithm that allow additional distributions to be tested. Although we implement the normal distribution, similar to nQuire, this distribution may be ill-suited for site-based allele frequencies that range from 0 to 1, as the normal distribution ranges from negative infinity to infinity. Our second implementation uses a beta distribution to match the constrained range of allele frequencies. Because sequence data provide allele counts, frequencies represent transformed data, which may lack original data attributes and misrepresent sampling variances and one or more sources of heterogeneity. Therefore, our final implementation includes the beta-binomial distribution, which allows raw allele counts to be leveraged.

Our goal was to identify the best approach to infer ploidy from DNA sequence data from a variety of species exhibiting a range of life histories and origins of polyploidy. We therefore rigorously tested our new implementations to identify limitations to these new methods and provide guidance for users. We examine the same five samples as used in the development of nQuire (Weiß et al., 2018), plus 477 samples representing three additional taxonomic groups and three additional sequence data types (genotype-by-sequencing, target enrichment, and RADcap). To provide recommendations regarding coverage and the number of sites needed for each implementation and model type, we also test our model on 355 simulated samples, representing two simulation approaches that vary in the amount of variance introduced.

METHODS AND RESULTS

Likelihood calculations and model selection

The basis of our models is the expected site-based allele frequency at variable biallelic sites within an individual for each ploidal level, including diploid (0.5), triploid (0.33, 0.67), tetraploid (0.25, 0.5, 0.75), pentaploid (0.2, 0.4, 0.6, 0.8), and hexaploid (0.17, 0.33, 0.5, 0.67, 0.83), as introduced above

(Figure 1, Appendix S1; see Supporting Information with this article). To use the expected allele frequencies to determine the most likely ploidal level given a set of allele frequencies or allele counts representing biallelic sites from a single individual, we developed three implementations of expectation maximization algorithms with the normal, beta, and beta-binomial distributions, each with and without a uniform distribution to capture uniform noise components. The normal distribution implemented here differs from that of nQuire in our augmented-likelihood calculation (Appendix S1: Equation 5); however, all model comparisons were investigated with both the nQuire-style implementation and our implementation of the normal distribution (Appendix S1). We found our implementation to have lower confidence in incorrect models compared to nQuire's implementation, and therefore we focus only on our implementation of the normal distribution here.

The details of our implementations are provided in Appendix S1, but are summarized here. Given the expected frequencies, the likelihood for each ploidal level based on a set of observed allele frequencies (or allele counts) is defined as the sum of the product of the mixture proportion (alpha) and the relative likelihood of the observations, or probability density function, based on the expected frequency (mean) and variance of that mixture and the given distribution (Figure 2). To maximize the likelihood for a set of mixtures, values of alpha, variance, and mean can be modified through the expectation maximization algorithm and optimized with the Nelder–Mead simplex optimization algorithm (Nelder and Mead, 1965). Furthermore, to allow model selection via information criteria, where divergence among models can be estimated by calculating the log-likelihood ratio, we allow "free" and "fixed" models, where all "fixed" models are nested in a "free" model. In our free model, all parameter values (alpha, variance, and mean) are estimated for a mixture of all potential ploidal levels. Although we have an expected value for the mean of each mixture, the expected values of alpha, as well as the variance, are not well defined. Because the proportions of each type of heterozygote may differ for an
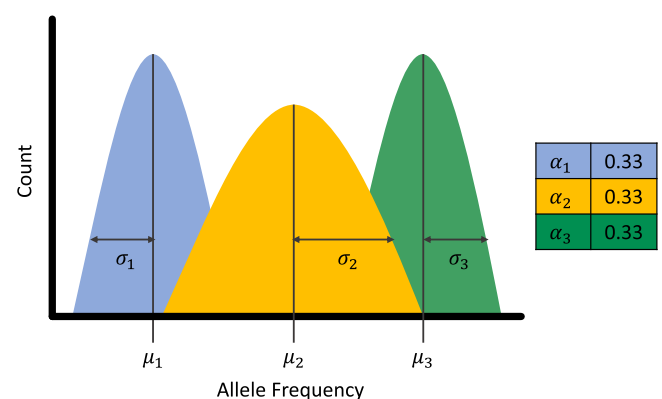
FIGURE 2 The basic components of a mixture model include mean (μ), variance (σ), and proportion (or alpha, α). The expected distributions for an autotetraploid, as defined by Lloyd and Bomblies (2016), are shown here.

allopolyploid compared to an autopolyploid (reviewed in Lloyd and Bomblies, 2016), we were interested in exploring models where alpha is free (see caveats in the Conclusions, below). Therefore, we tested three "fixed" models: (1) where only alpha is free, (2) where only variance is free, and (3) where both alpha and variance are free. Therefore, for each implementation, we provide 32 model types, including three fixed models, at each of the five ploidal levels examined here and one "free" model, all of which can be examined with and without a uniform distribution.

To evaluate each model, we examined the log-likelihood ratio and the Bayesian information criterion, or BIC score. The BIC score is the log-likelihood of a model penalized by both sample size and the number of parameters included, which leads to less error in model selection (Taper et al., 2021). We examined both the log-likelihood ratio and BIC score for all models and determined that BIC identified the correct ploidal level of more samples than the log-likelihood ratio; thus, we focused on BIC scores in all model comparisons. The BIC difference between the best and second-best model can be leveraged as an information criterion to assess confidence in model selection (Jerde et al., 2019; Taper et al., 2021).

Model evaluation

To evaluate our models and determine guidelines for implementing these models, we examined 513,792 models based on both simulated and real samples. Simulated data representing all five ploidal levels varied in sequence coverage and number of sites, as well as the amount of random noise. Real samples include 482 samples of known ploidy (Table 1), inferred via indirect and direct estimates, and represent five taxonomic groups and four types of sequence data.

Simulated data

We simulated samples based on two approaches that represent two sampling scenarios: a "simplistic" one and a "realistic" scenario where the sampling is done at various levels of DNA sequence coverage (3–120×). Simplistic simulated samples are simple, with little to no variance introduced during the simulation process. The simplistic approach simulates heterozygous biallelic sites based on a binomial distribution where coverage among sites is equal and all expected frequencies have an equal probability of being sampled. For each ploidal level, we simulated 11 samples that differed in coverage per site (5×, 10×, 20×, 30×, 40×, 50×, 60×, 70×, 80×, 90×, or 100×). For the 55 simulated samples, models were evaluated at six different numbers of sites, or the total number of SNPs (1250, 2500, 5000, 10,000, 20,000, 30,000).

For our realistic simulations, we simulated samples where coverage across sites was variable and allele frequencies had higher variance than the simplistic simulations. The variance introduced in these simulations is meant to resemble noise introduced by sequencing errors and data processing errors (e.g., mapping errors). We simulated 60 different coverage amounts for each ploidal level; these simulations varied in the minimum and maximum coverage, as well as the expected number of samples within an interval, or lambda. Based on the minimum and maximum coverage, as well as the expected number of events (lambda), the total coverage for each site is sampled from a truncated Poisson distribution, as coverage across a genome resembles a Poisson distribution with multiple peaks (Pfenninger et al., 2022). For each of our 60 simulations, we set the minimum coverage as i, maximum coverage as $(i + 1) \times 3$, and lambda as half of the sum of the minimum and maximum coverage (Appendix S2: Figure S1). The resulting mean coverage simulated by this method ranged from 3 to 120×. Given a randomly selected proportion (i.e., mean and associated variance), the copies of allele A were then defined with a binomial sample with the probability defined by the beta distribution (i.e., a beta-binomial) and the copies of allele B are equal to the remainder. We then followed the data processing steps applied to real data. First, the simulated data were filtered to remove any sites where only one allele was sampled by chance. Next, we filtered the sites based on the total coverage and sequencing coverage of each allele. This function can also filter sites based on truncated allele frequencies. Finally, we randomly sampled an allele with

TABLE 1 An overview of all included sample sets including the species, total number of samples, ploidal levels included, and sequencing approach.

Sample set	Species	Total	Ploidal levels	Sequencing approach
Yeast	*Saccharomyces cerevisiae*	3	2x, 3x, 4x	Whole-genome resequencing
Oomycete	*Phytophthora infestans*	2	2x, 3x	Whole-genome resequencing
Glycine spp.	*G. albicans* (2x), *G. arenaria* (2x), *G. falcata* (2x), *G. hirticaulis* (2x), *G. tomentella* (2x, 4x), *G. stenophita* (2x), *G. syndetika* (2x), *G. tabacina* (4x)	17	2x, 4x	Genotype-by-sequencing
Galax	*Galax urceolata*	190	2x, 3x, 4x	Target enrichment
Larrea	*Larrea tridentata*	270	2x, 3x, 4x, 5x, 6x	RADcap (Hoffberg et al., 2016; Bayona-Vásquez et al., 2019)

Note: Additional information, including available accessions, can be found in Appendices S3 and S4.

equal probability at each site. The resulting data set includes the total coverage per site and the coverage associated with a randomly sampled allele. For the 300 simulated samples, models were evaluated at six different numbers of sites, or the total number of SNPs (1250, 2500, 5000, 10,000, 20,000, 25,000).

Organismal data

We applied our model to available data sets with DNA sequence data and known ploidy for samples of *Saccharomyces cerevisiae* Meyen ex E. C. Hansen, *Phytophthora infestans* (Mont.) de Bary, *Glycine* Willd. spp., *Larrea tridentata* (DC.) Coville, and *Galax urceolata* (Poir.) Brummitt; for simplicity, we refer to these as yeast, oomycete, *Glycine* spp., *Larrea*, and *Galax*, respectively (Table 1). Both the yeast and oomycete sample sets were used to develop and test nQuire (Weiß et al., 2018); thus, we chose to investigate these samples with nQuack. Of the remaining data sets, *Galax urceolata* and *Larrea tridentata* represent likely autopolyploids (Hunziker et al., 1977; Lewis, 1980; Nesom, 1983; Servick et al., 2015), while *Glycine* spp. are allopolyploids with likely disomic inheritance (Sherman-Broyles et al., 2014; Zhuang et al., 2022; Landis and Doyle, 2023). The type of DNA sequence data varied across these samples, including whole-genome resequencing, genotype-by-sequencing (Elshire et al., 2011), target enrichment, and RADcap data (Hoffberg et al., 2016; Bayona-Vásquez et al., 2019) (Table 1). RADcap (Hoffberg et al., 2016; Bayona-Vásquez et al., 2019) combines reduced-representation 3RAD library preparation (Hoffberg et al., 2016; Bayona-Vásquez et al., 2019) with probe-based target capture. These sample sets also vary in the number of samples, diversity in ploidal level, taxonomic diversity, and quality of the reference genome (Table 1, Appendices S3 and S4).

We aligned reads from each sample to the associated reference genome for that species (Appendix S3) with BWA-MEM2 version 2.2.1 (Vasimuddin et al., 2019), converted the SAM file to a BAM file, and sorted the results with SAMtools version 1.15 (Danecek et al., 2021). We identified and masked repeat regions with RepeatModeler version 2.0 (Flynn et al., 2020) and RepeatMasker version 4.1.1 (Smit et al., 2015). Repetitive regions should be removed from alignments before the estimation of ploidal level, as these regions will have high coverage and will likely not represent the copy number variation found in coding or single-copy regions. Based on the masked genomes, we then created databases of repeat regions that were removed from each sample alignment. We also removed poorly mapped reads and any sites that had a 10% chance or more of being mapped to the wrong location (-q 10).

To allow for investigation of multiple filtering approaches, we first prepared a text file of the alignment. After preparing text files with our function *prepare_data()*, we manually inspected each data set and specified the minimum filtering settings accordingly. Filtering strategies differed in minimum coverage and maximum coverage quantile, as well as the lower bound (C_L) and upper bound (C_U) for allele frequency truncation. For all filtering strategies, sequencing depth per allele was filtered based on a sequencing error rate of 0.01, where the coverage of each allele must be more than the total coverage multiplied by the error rate, but less than the total coverage multiplied by one minus the error rate. To avoid enhancement of signal from data duplication, we randomly sampled an allele with equal probability at each site. After filtering, the resulting data set includes the total coverage per site and the coverage associated with a randomly sampled allele.

We examined four filtering strategies across sample sets, with at least two examined per set. For all sample sets, we examined the minimum filtering approach (D1) and the maximum filtering approach (D4). Because hexaploid samples are expected to have mixtures with means equal to 0.17 and 0.83, we investigated filtering approaches that differed in C_L and C_U, to ensure we did not remove these peaks in our filtering process. The minimum filtering approach (D1) settings differed per sample set, with three groups of settings: yeast and oomycete, *Galax* and *Glycine* spp., and *Larrea*. Respectively, the settings for the minimum filtering approach were minimum coverage equal to 10, 2, and 3; maximum coverage quantile equal to 0.90, 0.90, and 1; C_L equal to 0.11, 0.1, and 0.11; and C_U equal to 0.89, 0.9, and 0.89. The maximum filtering approach (D4) represents nQuire's default settings, where minimum coverage is 10, C_L is 0.15, C_U is 0.85, and there is no maximum coverage cutoff. The maximum filtering approach (D4) was applied with nQuire's create function on all samples except for the *Larrea* sample set, which was prepared with a maximum depth quantile of 0.9 and error correction of 0.01. For *Galax* and *Larrea*, we examined two additional filtering approaches to examine the intermediate between the minimum and maximum filtering approaches. First, we increased the minimum coverage to 10, but retained the C_L and C_U in the minimum filtering approach (D2). Second, we increased our allele truncation with C_L as 0.15 and C_U as 0.85, with the minimum coverage retained from the minimum filtering approach (D3). After filtering, the resulting data set included the total coverage per site and the coverage associated with a randomly sampled allele.

Model performance on simulated data

Overall, we found that no single model correctly assigned ploidal levels to all simulated samples (Figure 3). The amount of random noise in simulated data influenced which model correctly predicted the most simulated samples, with the best model differing for the simplistic and realistic simulated data (Appendix S2: Figures S2–S55). When considering all five potential ploidal levels, the most accurate model for the simplistic simulated samples was the beta distribution with variance free and a uniform mixture. For this model, the first three ploidal levels can be
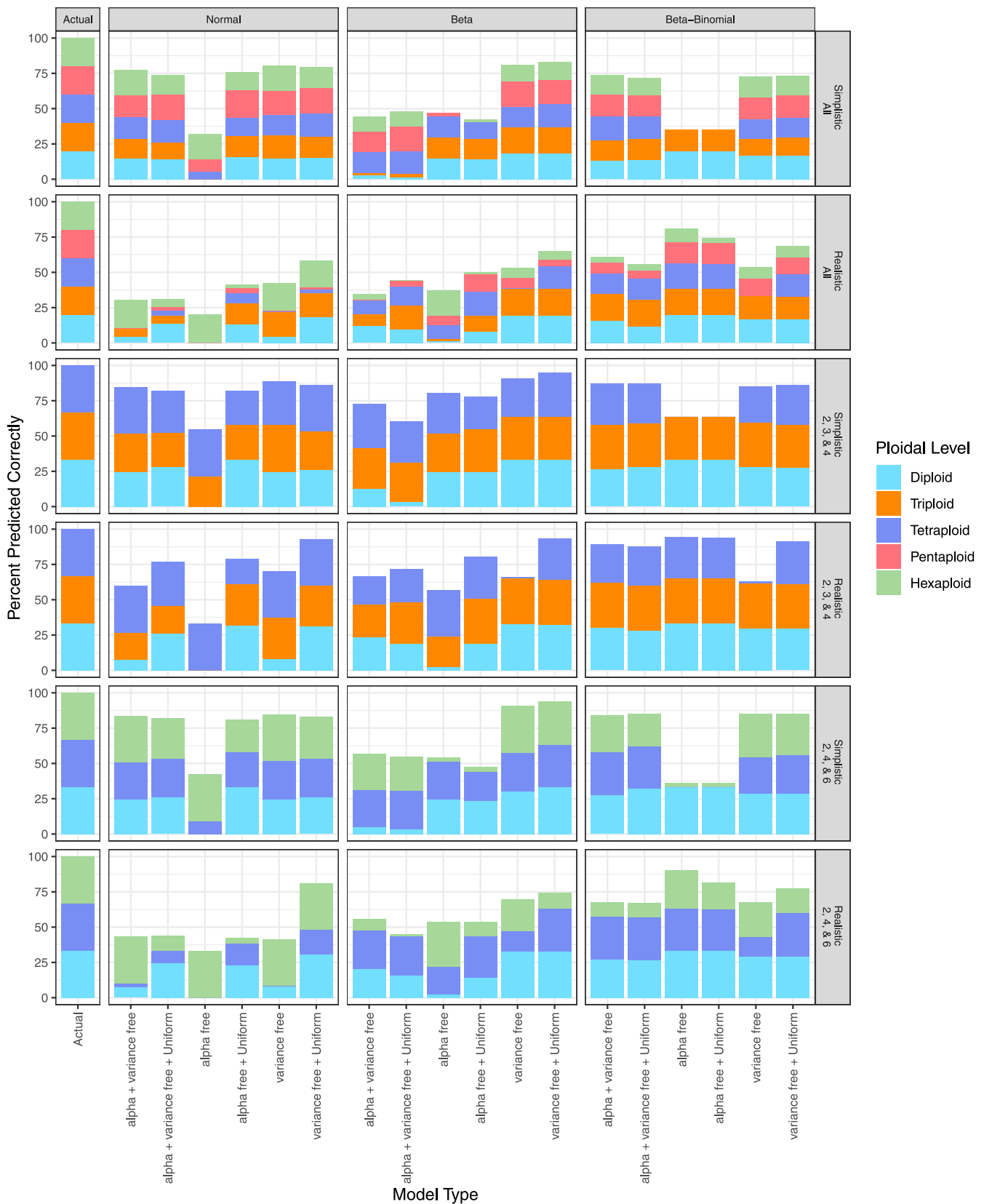
FIGURE 3 Variance in simulated data led to a higher rate of incorrect ploidal-level assignment. A larger percentage of samples was properly assigned ploidal level when the number of mixtures examined was reduced. Some models are unsuitable for assigning specific ploidal levels, for example, diploids are not identified under the normal distribution when alpha is free.

differentiated at about 20× coverage; however, pentaploid and hexaploid samples could not be differentiated until about 70× coverage. For the realistic simulated samples, when considering all five potential ploidal levels, the most accurate model was the beta-binomial with alpha free. For this model, diploids, triploids, tetraploids, and pentaploids can be differentiated at about 30× coverage, but hexaploids cannot be accurately identified until about 70× coverage.

Decreasing the number of ploidal levels considered may allow the proper assignment of ploidal levels to both simplistic and realistic samples (Figure 3). For example, when considering all ploidal levels with a normal distribution with variance free and uniform mixture, tetraploid realistic samples were identified incorrectly as hexaploids (Appendix S2: Figure S7). However, when a subset of mixtures was considered, tetraploids could be properly assigned as tetraploids for both simulation types (see Appendix S2: Figures S25 and S43). The impact on sequence coverage requirements was minimal (see Appendix S2).

In some instances, we found the probability of the correct model choice to increase with the BIC difference between the best and second-best models; however, accuracy and BIC score difference often did not have a linear relationship (Appendix S2: Figures S56–S58). We therefore caution against interpreting the difference in BIC scores between the best and second-best models as a measure of confidence or accuracy.

Model performance on sample sets

As found with our simulated data, a single model was not ideal for all real samples. However, we were able to identify models that assigned ploidal level correctly to all samples or a large subset of samples for all data sets, with the best model for each sample set having at least 78% accuracy (Figure 4, Table 2). For those sample sets without pentaploid or hexaploid samples, we considered only diploid, triploid, and tetraploid mixtures, as this reduced assignment error. Our implementation of nQuire, as well as the best model identified with nQuack, had equal or greater accuracy than the original nQuire model (Table 2).

We were able to properly assign ploidal levels to all five samples originally investigated by nQuire. For the yeast sample set, all three distributions had multiple model types that were able to properly assign ploidal level to all samples under both filtering approaches; the model type implemented in nQuire, variance free with a uniform mixture, was also able to accurately assign ploidal level to all samples with all three distributions. Notably, the normal distribution with alpha and variance free and a uniform mixture was only suitable when the allele truncation was the least constrained (D1). For the oomycete sample set, only one model was suitable when allele truncation was the least constrained: normal distribution with alpha free and a uniform mixture. Surprisingly, for the oomycete sample set, the nQuire model type (variance free with a uniform mixture) was unable to properly assign ploidal

level to the diploid sample when filtering did not match the filtering approach of nQuire. Additionally, the nQuire filtering approach (D4) allowed the proper assignment of both oomycete samples by at least two models from each distribution. Unlike all other sample sets, the maximum filtering approach (D4) increased the number of sites for both oomycete and yeast sample sets (Appendix S2: Figure S59); this is likely due to an excess of sites with high sequencing depth.

For *Glycine* spp., the nQuire filtering approach had low accuracy for all models (<60%); however, the minimum filtering approach allowed 16 of 17 samples to be assigned the correct ploidal level based on the beta-binomial distribution and the alpha- and variance-free model with a uniform mixture. We expected the alpha-free model to be the best model for *Glycine* spp. samples due to the history of ancient polyploidization in *Glycine* spp. (Walling et al., 2006), thus the proportions of each different heterozygote should be unequal. As expected, alpha as a free variable was informative for tetraploids; however, without a uniform mixture, diploids were incorrectly identified. Under the best model, the single incorrectly assigned diploid was an individual of *Glycine tomentella* (D5Bb), which is known to have a history of introgression (see Landis and Doyle, 2023). Hybridization can lead to an increased gene copy number; therefore, a more conserved filtering approach to only retain single-copy loci may be necessary to improve accuracy.

The best model for *Glycine* spp. also had high accuracy for *Galax* samples under the minimum filtering approach, with 185 of 190 samples with properly assigned ploidal levels with only two tetraploids and three triploids misidentified. The tetraploid samples that were incorrectly identified had weak support; the absolute difference between the BIC score of the best model relative to the second-best model was less than 10, and these values were less than the BIC score difference of all accurate estimates. Although we caution against the interpretation of BIC score difference as a measure of accuracy generally, evaluating this method on samples with known ploidal level identified this potential usage for a set of unknown samples. When sample sequence data are more similar to the modeled data-generating process, these criteria may be informative. Here, we targeted single-copy loci with capture-based sequencing, thus avoiding variance among loci that would skew these models. However, BIC score differences were not informative for the incorrectly assigned triploid samples. Two of these three triploid samples were incorrectly identified by all models; both samples have a high abundance of sites with an allele frequency of approximately 0.5, suggesting potential unequal locus loss and retention across targeted sites, which may be due to their origin from diploid–autotetraploid hybridization. When low-coverage sites remained (D1 and D3), the distribution with the best model remained the beta-binomial with 184 of 190 samples correctly predicted under the variance-free with uniform mixture model. When low-coverage sites were removed (D2 and D4), the best model
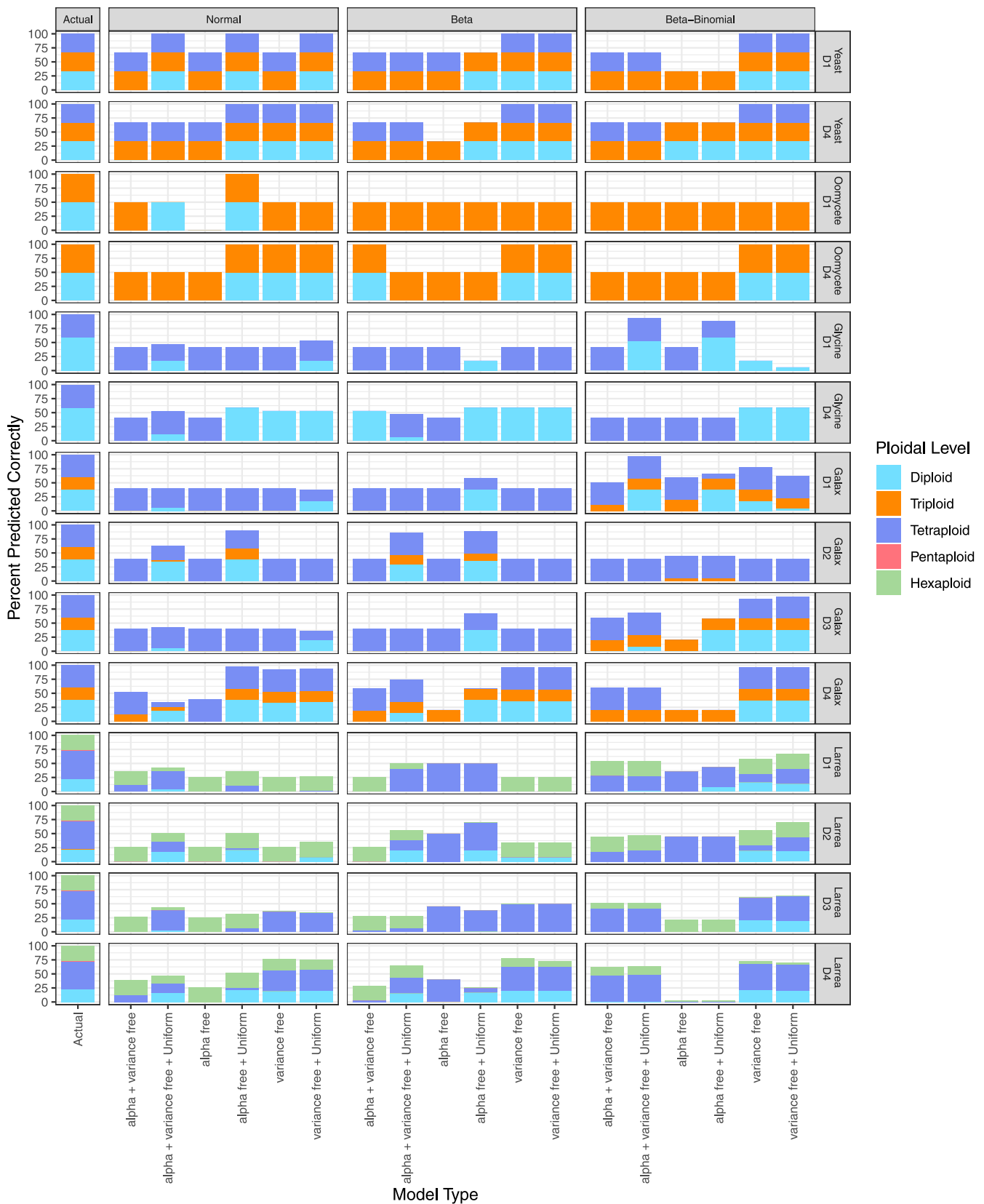
FIGURE 4 A large proportion of samples was properly assigned ploidal level when only considering a subset of mixtures (2*x*, 3*x*, and 4*x* for yeast, oomycete, *Glycine*, and *Galax*; 2*x*, 4*x*, and 6*x* for *Larrea*). All samples were properly identified by at least one model for both yeast and oomycete sample sets. For *Glycine* and *Galax*, the best model identified 16 out of 17 samples and 186 out of 190 samples, respectively. For *Larrea*, the best model was unable to identify 60 samples, for a total of 210 out of 270 samples correctly identified.

TABLE 2 Accuracy of nQuire compared to nQuack's implementation of nQuire (normal distribution with variance free and a uniform mixture with the maximum filtering approach, D4) and the best model by nQuack.

Sample set	Total	nQuire	nQuack's nQuire	nQuack's best model
Yeast	3	3 (100%)	3 (100%)	3 (100%)
Oomycete	2	2 (100%)	2 (100%)	2 (100%)
Glycine spp.	17	9 (53%)	9 (53%)	16 (94%)
Galax	190	172 (91%)	179 (94%)	186 (97%)
Larrea	270	189 (70%)	205 (76%)	210 (78%)

Note: The percent of total samples accurately assigned is given in parentheses. nQuire was run on alignments before our recommended preprocessing steps.

shifted to the normal distribution with alpha free and a uniform mixture. The highest accuracy was found under the D4 filtering approach with the normal distribution with alpha free and a uniform mixture; this model accurately assigned ploidy to 186 of the 190 individuals, only failing to identify a single tetraploid and three triploids.

For the *Larrea* data set, we were able to identify all triploids, tetraploids, pentaploids, and hexaploids under at least one model; however, the best model and filtering approach for each ploidal level differed. Based on the 18 different models and four different filtering approaches investigated for all cytotypes or only a subset of ploidal levels (2x, 4x, and 6x), we identified 22 and 39 instances, respectively, where all 70 hexaploids were assigned the correct ploidal level. For tetraploid samples, all 135 individuals were correctly identified in two instances for all cytotypes and two instances for only a subset of ploidal levels. Similar to the triploids in the *Galax* sample set, there were multiple diploid samples that our implemented models failed to identify correctly. These diploid samples were found to occur in mixed ploidal sites or at the edge of the species range, suggesting that ongoing mixed-ploidy introgression or divergence from the reference may skew the models' ability to accurately assign ploidal levels due to increased gene copy number or mapping error, respectively. When considering all five potential ploidal levels, the best model was the beta distribution with alpha free and a uniform mixture, with 189 of 270 samples correctly assigned ploidal level under the D2 filtering approach; this prediction misidentified all hexaploids and pentaploids, as well as three tetraploids and four diploids. When we reduced the mixture of ploidal levels considered to include only diploids, tetraploids, and hexaploids, the best model shifted to the beta distribution with variance free under the maximum filtering approach with 210 samples correctly identified; the misidentified samples included all triploids and pentaploids, six diploids, 20 tetraploids, and 29 hexaploids. The original nQuire model incorrectly estimated the ploidal level for only six diploids and one tetraploid from the diploid, triploid, and tetraploid *Larrea* samples; comparatively, our implementation of nQuire incorrectly assigned ploidal level

to an increased number of tetraploid samples due to the inclusion of a hexaploid mixture model, which was identified as more likely for these samples. Although reducing the number of ploidal levels considered can increase the number of correctly assigned samples, we do not advise ignoring the presence of triploid, pentaploid, or hexaploid cytotypes in a system to improve model accuracy. Overall, our approach increased the *Larrea* sample set accuracy compared to nQuire by 8% (Table 2).

CONCLUSIONS

Here, we provide expanded tools and implementations to improve site-based heterozygosity inferences of ploidal level. Our program nQuack provides data preparation guidance and tools to decrease noise in input data. These tools include a maximum sequence coverage quantile filter and sequence error–based filter to remove biallelic sites that are likely not representative of copy number variance in the nuclear genome. We also consider only the frequency of allele A or B at each site, instead of both, as implemented in nQuire, as this would inflate the observation by enhancing the signal or noise found in the data. Our model builds upon, and improves, the nQuire framework by extending it to higher ploidal levels (pentaploid and hexaploid), correcting the augmented likelihood calculation, implementing more suitable probability distributions, and allowing additional "fixed" models. We also decrease model selection errors by relying on BIC rather than likelihood ratio tests (Taper et al., 2021). Overall, nQuack facilitates model exploration, as demonstrated by the 513,792 models explored in this study, and improves the accuracy of ploidal estimation based on site-based heterozygosity.

Through the intensive testing of the methods presented here, we found that many variables influence model accuracy. Based on our simulated data, we observed that each model implementation and model type can be influenced by the number of sites, sequencing coverage, and amount of variance or noise in a data set. In real data sets, this noise can be introduced through sequencing error or general sequence mapping error. In addition, although we attempted to retain only single-copy loci by removing repetitive regions, additional filtering may increase accuracy to ensure estimates are not conflated by variation among loci at non-single-copy sites. By examining a large amount of empirical sequence data from diverse organisms with different life histories, we determined that the most accurate model for each data set differed, suggesting that both filtering strategies and model selection must be explored on a set of known samples before applying these models to any sample with an unknown ploidal level to achieve accurate ploidy assignment. Although we attempted to connect model inference errors to the known biology of our sample sets, it is difficult to untangle the biological limitations of a model that is based on pattern. Furthermore, model accuracy will be impacted by any process that modifies

biallelic allele frequencies or proportions relative to the expectation for each ploidal level. As described above, allele frequencies and proportions may be modified by many biological processes including, but not limited to, the mode and timing of WGD, gene duplication or loss, and chromosomal behavior during meiosis. We do not expect inferences to be inhibited by processes that lead to excess homozygosity, such as double reduction, as these would only decrease the number of biallelic sites compared in a model, which is accounted for in the BIC score.

We explored nQuack's performance on an extensive set of simulated data and multiple real-world data sets. These analyses allowed us to benchmark model performance and identify data features that affect nQuack's predictive power. However, the biological data sets we explored cannot represent the full diversity of polyploid systems, and additional tuning is required for real data sets. For example, these models would not be suitable in an allotetraploid with strict disomic inheritance as no AAAB or ABBB loci would occur; therefore, the most likely model could be identified as a diploid, although BIC score parameter corrections would allow the most probable model to be hexaploid or tetraploid. Additional biological systems will likely introduce more complexities and may work best under different filtering conditions. For example, uneven nuclear genome copy number among cells would not modify biallelic allele frequencies; however, processes such as endoreduplication may reduce model accuracy and complicate ploidy assignment based on sequence data alone. To identify the factors that dictate which strategy is the most accurate, multiple mixed-ploidy systems with high-quality reference genomes, well-classified polyploidization events (e.g., mode of formation, timing of polyploidy events, chromosomal segregation patterns), and well-characterized reproductive history should be explored in future model iterations. Regarding summary statistics, non-parametric bootstrapping after model selection would allow for assessing the strength of the evidence in favor of every model and the robustness of model selection results. We provide functions to perform this non-parametric bootstrap sampling; however, completing a full non-parametric bootstrap for all of our real data sets was neither practical nor feasible due to computational limitations. Because all mathematical models are misspecifications of the true data-generating process (Dennis et al., 2019), errors are probable when selecting the model closest to the truth. Therefore, by resampling the data, we can assess the reliability of the model choice. In addition, if analytical-based inferences continue to be pursued, a sliding window approach will likely improve ploidy inferences.

Our results open many interesting avenues for future research. Site-based heterozygosity models such as the ones used here are in essence phenomenological statistical models, which focus on reproducing patterns rather than generating patterns based on a fundamental biological process. Although statistical models embodying fundamental biological processes are common in many areas of biology (for instance, in phylogenetics), in this particular case it is extremely difficult to capture the complexities of nature in an analytical-based inference, and future model exploration utilizing data-based inference to classify ploidal levels is warranted. Alternatively, demographic models such as those we proposed elsewhere (Gaynor et al., 2023) may provide the ecological and evolutionary framework necessary to design process-based predictions for mixed ploidy. These models, however, require rigorous coupling with evolutionary and genomic theory.

Overall, this analysis reveals that it is critical to thoroughly examine proposed methods before inferring biological meaning. Neither nQuack nor nQuire should be used to infer the ploidy in a system for which very little is known, as these models are often positively misleading. As suggested above, both filtering strategies and model selection must be explored on a set of known samples before applying these models to any sample with an unknown ploidal level to achieve accurate ploidy assignment. We also suggest caution when relying on any method using site-based heterozygosity to predict ploidy of a sample even when a known data set is analyzed before applying the method to a sample of unknown ploidy due to the potential impact of various biological processes (e.g., hybridization, divergence) on model inference. Despite the many caveats to this method, nQuack is the best approach currently available and can be easily implemented to leverage sequence data for ploidal estimation.

DATA AVAILABILITY STATEMENT
The R package nQuack is available at https://github.com/mgaynor1/nQuack and https://mlgaynor.com/nQuack/. A full implementation tutorial (https://mlgaynor.com/nQuack/articles/BasicExample.html), as well as detailed tutorials on data preparation (https://mlgaynor.com/nQuack/articles/DataPreparation.html) and model inference (https://mlgaynor.com/nQuack/articles/ModelOptions.html), are available with the package documentation. For three sample sets, reference

genomes and population genetics data are available via open repositories (see Appendix S3 and S4 for accessions). Sequence data for *Galax urceolata* and *Larrea tridentata* will be published in open repositories with future publications. An exemplar data set and processing times required for every step of model implementation (1.46–2.09 s for models with the normal distribution; 6.41–23.16 min for models with the beta distribution; 9.54–46.15 min for models with beta-binomial distribution), as well as the output of each step of our method, are available on our GitHub (https://mlgaynor.com/nQuack/articles/BasicExample.html).

ORCID

Michelle L. Gaynor http://orcid.org/0000-0002-3912-6079
Jacob B. Landis http://orcid.org/0000-0002-5631-5365
Timothy K. O'Connor http://orcid.org/0000-0001-8204-6552
Robert G. Laport http://orcid.org/0000-0001-5672-0929
Jeff J. Doyle http://orcid.org/0000-0003-1579-9380
Douglas E. Soltis http://orcid.org/0000-0001-8638-4137
José Miguel Ponciano http://orcid.org/0000-0001-8457-7840
Pamela S. Soltis http://orcid.org/0000-0001-9310-8659

REFERENCES

Bayona-Vásquez, N. J., T. C. Glenn, T. J. Kieran, T. W. Pierson, S. L. Hoffberg, P. A. Scott, K. E. Bentley, et al. 2019. Adapterama III: Quadruple-indexed, double/triple-enzyme RADseq libraries (2RAD/3RAD). *PeerJ* 7: e7724.

Beaulieu, J. M., I. J. Leitch, S. Patel, A. Pendharkar, and C. A. Knight. 2008. Genome size is a strong predictor of cell size and stomatal density in angiosperms. *The New Phytologist* 179: 975–986.

Becher, H., J. Sampson, and A. D. Twyford. 2022. Measuring the invisible: The sequences causal of genome size differences in eyebrights (*Euphrasia*) revealed by k-mers. *Frontiers in Plant Science* 13: 818410.

Blonder, B., B. J. Graae, B. Greer, M. Haagsma, K. Helsen, R. E. Kapás, H. Pai, et al. 2020. Remote sensing of ploidy level in quaking aspen (*Populus tremuloides* Michx.). *Journal of Ecology* 108(1): 175–188.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10: giab008.

Dennis, B., J. M. Ponciano, M. L. Taper, and S. R. Lele. 2019. Errors in statistical inference under model misspecification: Evidence, hypothesis testing, and AIC. *Frontiers in Ecology and Evolution* 7: 372.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto, E. S. Buckler, and S. E. Mitchell. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6: e19379.

Flynn, J. M., R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences, USA* 117: 9451–9457.

Folk, R. A., J. L. M. Charboneau, M. Belitz, T. Singh, H. R. Kates, D. E. Soltis, P. S. Soltis, et al. 2024. Anatomy of a mega-radiation: Biogeography and niche evolution in *Astragalus*. *American Journal of Botany* 111(3): e16299.

Galbraith, D. W., K. R. Harkins, J. M. Maddox, N. M. Ayres, D. P. Sharma, and E. Firoozabady. 1983. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science* 220: 1049–1051.

Gaynor, M. L., N. Kortessis, D. E. Soltis, P. S. Soltis, and J. M. Ponciano. 2023. Dynamics of mixed-ploidy populations under demographic and environmental stochasticities. *BioRxiv* 2023.03.29.534764 [preprint]. Available at: https://doi.org/10.1101/2023.03.29.534764 [posted 31 March 2023; accessed 15 June 2024].

Hoffberg, S. L., T. J. Kieran, J. M. Catchen, A. Devault, B. C. Faircloth, R. Mauricio, and T. C. Glenn. 2016. RADcap: Sequence capture of dual-digest RADseq libraries with identifiable duplicates and reduced missing data. *Molecular Ecology Resources* 16: 1264–1278.

Hunziker, J. H., R. A. Palacios, L. Poggio, C. A. Naranjo, and T. W. Yang. 1977. Geographic distribution, morphology, hybridization, cyto-genetics, and evolution. *In* T. J. Mabry, J. H. Hunziker, and D. R. DiFeo Jr. [eds.], Creosote bush: Biology and chemistry of *Larrea* in the New World deserts, 115–134. Dowden, Hutchinson & Ross, Stroudsberg, Pennsylvania, USA.

Jantzen, J. R., P. J. F. Guimarães, L. C. Pederneiras, A. L. F. Oliveira, D. E. Soltis, and P. S. Soltis. 2022. Phylogenomic analysis of *Tibouchina s.s.* (Melastomataceae) highlights the evolutionary complexity of Neotropical savannas. *Botanical Journal of the Linnean Society* 199: 372–411.

Jerde, C. L., K. Kraskura, E. J. Eliason, S. R. Csik, A. C. Stier, and M. L. Taper. 2019. Strong evidence for an intraspecific metabolic scaling coefficient near 0.89 in fish. *Frontiers in Physiology* 10: 1166.

Jiao, Y., N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* 473: 97–100.

Keeler, K. H., B. Kwankin, P. W. Barnes, and D. W. Galbraith. 1987. Polyploid polymorphism in *Andropogon gerardii*. *Genome* 29: 374–379.

Landis, J. B., and J. J. Doyle. 2023. Genomic relationships of *Glycine remota*, a recently discovered perennial relative of soybean, within *Glycine*. *Systematic Botany* 48: 78–87.

Landis, J. B., D. E. Soltis, Z. Li, H. E. Marx, M. S. Barker, D. C. Tank, and P. S. Soltis. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany* 105: 348–363.

Lewis, W. H. 1980. Polyploidy in species populations. *In* W. H. Lewis [ed.], Polyploidy: Biological relevance, 103–144. Springer, Boston, Massachusetts, USA.

Lloyd, A., and K. Bomblies. 2016. Meiosis in autopolyploid and allopolyploid *Arabidopsis*. *Current Opinion in Plant Biology* 30: 116–122.

Masterson, J. 1994. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science* 264: 421–424.

Nelder, J. A., and R. Mead. 1965. A simplex method for function minimization. *The Computer Journal* 7: 308–313.

Nesom, G. L. 1983. *Galax* (Diapensiaceae): Geographic variation in chromosome number. *Systematic Botany* 8: 1–14.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.

Pfenninger, M., P. Schönnenbeck, and T. Schell. 2022. ModEst: Accurate estimation of genome size from next generation sequencing data. *Molecular Ecology Resources* 22: 1454–1464.

Ranallo-Benavidez, T. R., K. S. Jaron, and M. C. Schatz. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11: 1432.

Sanders, H. 2021. Polyploidy and pollen grain size: Is there a correlation? M.S. thesis, University of Nebraska at Kearney, Kearney, Nebraska, USA.

Servick, S., C. J. Visger, M. A. Gitzendanner, P. S. Soltis, and D. E. Soltis. 2015. Population genetic variation, geographic structure, and multiple origins of autopolyploidy in *Galax urceolata*. *American Journal of Botany* 102: 973–982.

Sherman-Broyles, S., A. Bombarely, A. F. Powell, J. L. Doyle, A. N. Egan, J. E. Coate, and J. J. Doyle. 2014. The wild side of a major crop: Soybean's perennial cousins from Down Under. *American Journal of Botany* 101(10): 1651–1665.

Sliwinska, E., J. Loureiro, I. J. Leitch, P. Šmarda, J. Bainard, P. Bureš, Z. Chumová, et al. 2021. Application-based guidelines for best practices in plant flow cytometry. *Cytometry Part A* 101: 749–781.

Smit, A. F. A., R. Hubley, and P. Green. 2015. RepeatMasker. Website: http://www.repeatmasker.org [accessed 1 October 2023].

Soltis, P. S., D. B. Marchant, Y. Van de Peer, and D. E. Soltis. 2015. Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development* 35: 119–125.

Suda, J., and P. Trávníček. 2006. Reliable DNA ploidy determination in dehydrated tissues of vascular plants by DAPI flow cytometry—new prospects for plant research. *Cytometry Part A* 69A(4): 273–280.

Sun, M., E. Pang, W.-N. Bai, D.-Y. Zhang, and K. Lin. 2023. ploidyfrost: Reference-free estimation of ploidy level from whole genome sequencing data based on de Bruijn graphs. *Molecular Ecology Resources* 23: 499–510.

Takeuchi, F., and N. Kato. 2024. Ploidy inference from single-cell data: Application to human and mouse cell atlases. *Genetics* 227(2): iyae061.

Taper, M. L., S. R. Lele, J. M. Ponciano, B. Dennis, and C. L. Jerde. 2021. Assessing the global and local uncertainty of scientific evidence in the presence of model misspecification. *Frontiers in Ecology and Evolution* 9: 679155.

Vasimuddin, M., S. Misra, H. Li, and S. Aluru. 2019. Efficient architecture-aware acceleration of BWA-MEM for multicore systems, 314–324. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), Rio de Janeiro, Brazil.

Viruel, J., M. Conejero, O. Hidalgo, L. Pokorny, R. F. Powell, F. Forest, M. B. Kantar, et al. 2019. A target capture-based method to estimate ploidy from herbarium specimens. *Frontiers in Plant Science* 10: 937.

Vurture, G. W., F. J. Sedlazeck, M. Nattestad, C. J. Underwood, H. Fang, J. Gurtowski, and M. C. Schatz. 2017. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* 33: 2202–2204.

Walling, J. G., R. Shoemaker, N. Young, J. Mudge, and S. Jackson. 2006. Chromosome-level homeology in paleopolyploid soybean (*Glycine max*) revealed through integration of genetic and chromosome maps. *Genetics* 172: 1893–1900.

Weiß, C. L., M. Pais, L. M. Cano, S. Kamoun, and H. A. Burbano. 2018. nQuire: A statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* 19: 122.

Zhuang, Y., X. Wang, X. Li, J. Hu, L. Fan, J. B. Landis, S. B. Cannon, et al. 2022. Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. *Nature Plants* 8(3): 233–244.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Detailed implementation of expectation maximization algorithm and the models available in our method.

Appendix S2. Supplemental figures for "nQuack: An R package for predicting ploidal level from sequence data using site-based heterozygosity."

Appendix S3. Genome statistics including species identity, number of contigs, total length in base pairs, contig minimum length, contig average length, contig maximum length, N50 (Mbp), percent of GC content, BUSCO complete percentage, BUSCO duplicate percentage, BUSCO reference, and any accession information available.

Appendix S4. Extended information on sample sets including the number of samples of each ploidal level, the sequencer used, and any accession information available.