Demographic History Inference and the Polyploid Continuum

Paul D. Blischak[1,2,*], Mathews Sajan[2], Michael S. Barker[1] and Ryan N. Gutenkunst[2,*]

[1]Department of Ecology & Evolutionary Biology and
[2]Department of Molecular & Cellular Biology, University of Arizona, Tucson, AZ 85721, USA

*Corresponding authors: Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721, USA. Email: rgutenk@arizona.edu; Email: paul.blischak@gmail.com.

Abstract

Polyploidy is an important generator of evolutionary novelty across diverse groups in the Tree of Life, including many crops. However, the impact of whole-genome duplication (WGD) depends on the mode of formation: doubling within a single lineage (autopolyploidy) versus doubling after hybridization between two different lineages (allopolyploidy). Researchers have historically treated these two scenarios as completely separate cases based on patterns of chromosome pairing, but these cases represent ideals on a continuum of chromosomal interactions among duplicated genomes. Understanding the history of polyploid species thus demands quantitative inferences of demographic history and rates of exchange between subgenomes. To meet this need, we developed diffusion models for genetic variation in polyploids with subgenomes that cannot be bioinformatically separated and with potentially variable inheritance patterns, implementing them in the dadi software. We validated our models using forward SLiM simulations and found that our inference approach is able to accurately infer evolutionary parameters (timing, bottleneck size) involved with the formation of auto- and allotetraploids, as well as exchange rates in segmental allotetraploids. We then applied our models to empirical data for allotetraploid shepherd's purse (*Capsella bursa-pastoris*), finding evidence for allelic exchange between the subgenomes. Taken together, our model provides a foundation for demographic modeling in polyploids using diffusion equations, which will help increase our understanding of the impact of demography and selection in polyploid lineages.

Keywords: Autopolyploidy; Allopolyploidy; Demography; Homoeologous Exchange; Site Frequency Spectrum

Polyploidy, or whole genome duplication (WGD), is a mechanism for potentially rapid evolutionary change. Many lineages in the Tree of Life have experienced WGD events in their ancient pasts ("paleopolyploids"; Ohno 1970; Furlong and Holland 2001; Cui *et al.* 2006; Jiao *et al.* 2011; Li *et al.* 2018; Leebens-Mack *et al.* 2019; Li and Barker 2020) and the formation of recent polyploids ("neopolyploids") is especially common in plants and some groups of animals (e.g., amphibians Otto and Whitton 2000; Gregory and Mable 2005; Wood *et al.* 2009). The prevalence and persistence of polyploid lineages through deep time and into the present has led many to hypothesize that polyploids are able to adapt to more extreme environments/conditions than their diploid progenitors (Comai 2005; Baduel *et al.* 2018; Baniaga *et al.* 2020) Although selection and adaptation has been quantified in some studies of polyploids (Selmecki *et al.* 2015; Arnold *et al.* 2016; McIntyre and Strauss 2017; Monnahan *et al.* 2019), there remains a lack of consensus surrounding the role of demographic versus selective processes in contributing to the evolutionary trajectories of polyploids after their formation (Blischak *et al.* 2018b).

One reason for this lack of consensus is the difficulty of building demographic models that accommodate the additional set(s) of chromosomes that polyploids possess plus their potential interactions. The mode of formation for a polyploid, either through WGD within a lineage (autopolyploidy) or following hybridization between two or more different lineages (allopolyploidy), has a great impact on how the lineage evolves post-WGD. These differences result from the patterns of chromosomal interactions that occur in autopolyploids versus allopolyploids, with autopolyploids ranging from free recombination among all chromosomes in the genome to allopolyploids with only recombination between chromosomes from the same parental lineage. Previous work on the demography of polyploids has typically assumed that the species under study falls into one of these two categories. However, the existence of intermediate types of polyploids ("segmental allopolyploids"; Stebbins 1950) challenges the placement of polyploids into these discrete categories, suggesting that polyploids may be better described by a continuum (Meirmans and van Tienderen 2013; Mason and Wendel 2020). Building this continuum-like nature into a demographic model has not been explored extensively other than in a few case studies. For example, Roux and Pannell (2015) and Roux *et al.* (2021) used approximate Bayesian computation to infer the mode of origin of a polyploid by simulating data from different types of polyploids, using migration as a means to simulate polyploid data showing a mixed pattern of allelic inheritance, and comparing the simulated genetic data to the empirical data using summary statistics. Approaches such as this offer a promising avenue for further development of more generalized demographic models for different types of polyploids, but are also limited by their reliance on simulations and the comparison of summary statistics rather than on likelihood-based parameter estimation and model comparison.

Another important issue for polyploids is the additional com-

plexity of standard bioinformatic procedures, such as read mapping and variant calling, in the presence of duplicated chromosomes. Because the high-throughput sequencing reads collected for reference genome-based SNP calling are often short (∼50–250 bp depending on the platform), reads can potentially map to multiple chromosomes, leading either to large amounts of discarded data if only uniquely mapped reads are kept or to the identification of erroneous SNPs due to read mismapping. One way to help alleviate this issue is to focus on calling SNPs at the correct ploidy level of the sequenced individual, rather than trying to ensure that reads are mapping to the appropriate subgenome and calling SNPs separately at the ploidy level of each subgenome (e.g., calling tetraploid genotypes instead of calling genotypes in two diploid subgenomes). For autopolyploids, this method of SNP calling is simpler since their homoeologous chromosomes are derived from the same lineage and several methods for genotyping in this scenario exist (Serang *et al.* 2012; Blischak *et al.* 2018a; Gerard *et al.* 2018; Clark *et al.* 2019). However, for allopolyploid and segmental allopolyploid lineages, divergence between the parental subgenomes means that identifying homoeologous positions for genotyping is more difficult, and it is further complicated by having to distinguish between SNPs within a subgenome and fixed differences between subgenomes. Models to separately estimate SNPs within allopolyploid subgenomes have been proposed and used in a variety of crop species (Blischak *et al.* 2018a; Clevenger and Ozias-Akins 2015; Clevenger *et al.* 2018; Korani *et al.* 2019; Clark *et al.* 2019), but they typically require knowledge about all parental subgenomes. If the nature of a polyploid's formation mode is unknown or if parental information is not available for an allopolyploid, then these approaches are not able to be used. One possible solution would be to simply genotype a polyploid at its ploidy level without any attempt to separate out lower-ploidy subgenomes. With this approach, the task of determining the mode of polyploid formation could be incorporated into any downstream analyses rather than needing to be decided up front.

In this paper we develop a method for inferring the demographic history of a single polyploid population using a diffusion framework that includes homoeologous exchanges and a modified version of the site frequency spectrum (SFS) that collapses data across polyploid subgenomes into a single SFS, removing the need to classify the polyploid as an autopolyploid, allopolyploid, or in between during SNP calling. After describing the model, we compare the SFS generated by this diffusion approximation with frequency spectra generated by forward simulations. We then use forward simulations to assess our ability to infer demographic parameters under various combinations of population bottleneck sizes, bottleneck durations, population divergence times, and homoeologous exchange rates. We then use our model to infer the demographic history of allotetraploid shepherds purse (*Capsella bursa-pastoris*). Finally, we conclude with guidance about the model's use and interpretation, as well as possible extensions and future directions for work in the area of demographic inference in polyploid species.

Materials and Methods

Model description

Our model is inspired by the work of Meirmans and van Tienderen (2013). They focused on the effects of homoeologous exchanges on measures of genetic diversity and population structure in tetraploids by parameterizing an exchange rate, Θ, that determines how frequently alleles are inherited across subgenomes due to homoeologous crossovers and non-disomic inheritance. This exchange rate parameter ranges between 0 and 1, with $\Theta = 0$ corresponding to no allelic exchange and $\Theta = 1$ corresponding to free allelic exchange between subgenomes. Within this framework, the standard categorizations of allopolyploid and autopolyploid fit naturally and represent the extremes in the amount of expected homoeologous crossover. Values of Θ between 0 and 1 allow for intermediate amounts of homoeologous exchange and most closely align with what we would consider to be segmental allopolyploids. We incorporate this range of exchange rates into a diffusion framework, parameterizing homoeologous exchanges akin to migration, much like Roux and Pannell (2015), to provide a generalized model for polyploid demography.

A diffusion approximation for polyploids: We considered a single population of a K-ploid organism with S subgenomes, each containing $k_1, k_2, \ldots, k_S$ chromosomes ($\sum_i k_i = K$). Our diffusion model tracks the joint density of derived mutations across each of the S subgenomes, in an infinite sites model. The density of derived mutations at relative frequencies $x_1, \ldots, x_S$ at time t is denoted as $\phi(x_1, \ldots, x_S, t)$. In general, the relevant diffusion equation is (Kimura 1964)

$$\frac{\partial}{\partial \tau}\phi = \frac{1}{2}\sum_{i=1,\ldots,S}\frac{\partial^2}{\partial^2 x_i}\left\{V_i\phi\right\} + \sum_{i=1\ldots,S}\sum_{j\neq i}\frac{\partial^2}{\partial x_i \partial x_j}\left\{W_{ij}\phi\right\} - \sum_{i=1\ldots,S}\frac{\partial}{\partial x_i}\left\{M_i\phi\right\}. \quad (1)$$

Here, M_i represents the per-generation mean change in the frequency of an allele in subgenome i, V_i the variance in that change, and W_{ij} the covariance between changes in subgenomes i and j.

Let N be some reference number of individuals (often the ancestral population size) and $\nu(t)$ be the current relative size of the population. In a population of N individuals, there are $N\nu k_i x_i$ chromosomes in subgenome i that carry the derived allele. For a Wright-Fisher model, binomial sampling results in a variance in the number of carriers in the next generation of $N\nu k_i x_i(1-x_i)$. The variance in the allele frequency in the next generation is thus $x_i(1-x_i)/N\nu k_i$. Because sampling is independent among subgenomes, all the covariances W are 0.

Let $e_{i\leftrightarrow j}$ be the probability that a meiosis results in an exchange of genetic material between subgenomes i and j. The expected change in the number of chromosomes in subgenome i carrying the derived allele in one generation is then $-N\nu e_{i\leftrightarrow j}x_i + N\nu e_{i\leftrightarrow j}x_j$. The first term is the loss of alleles due to exchange out of subgenome i and the second is the gain due to exchange out of subgenome j and into subgenome i. The change in the derived allele frequency within subgenome i is then $\left(-N\nu e_{i\leftrightarrow j}x_i + N\nu e_{i\leftrightarrow j}x_j\right)/(N\nu k_i) = e_{i\leftrightarrow j}(x_j - x_i)/k_i$.

For compatibility with the widely-used diploid equations, we rescaled time to measure in units of $2N$ generations. Defining $E_{i\leftrightarrow j} = 2Ne_{i\leftrightarrow j}$ and $\Delta t = \Delta\tau/(2N)$, plugging in our results for the mean and variance terms, and multiplying both sides of Eq. 1 by $2N$, we obtained:

$$\frac{\partial}{\partial t}\phi = \frac{1}{2}\sum_{i=1,\ldots,S}\frac{2}{k_i}\frac{\partial^2}{\partial^2 x_i}\frac{x_i(1-x_i)}{\nu}\phi - \sum_{i=1\ldots,S}\frac{2}{k_i}\frac{\partial}{\partial x_i}\left(\sum_{j=1\ldots,S}E_{i\leftrightarrow j}(x_j - x_i)\right)\phi. \quad (2)$$

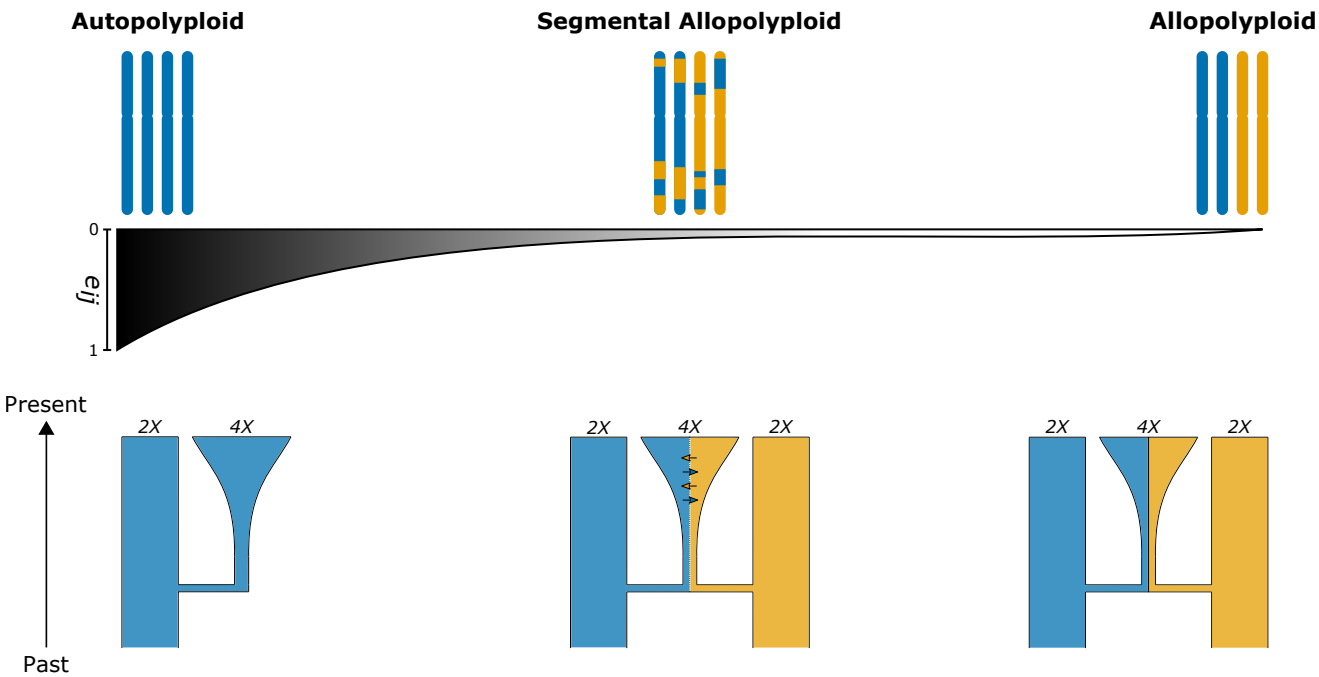

Figure 1 Conceptual representation of the polyploid continuum and corresponding demographic models for polyploid formation.

We noted the close analogy between Eq. 2 and the diffusion equation governing alleles within multiple diploid populations (Gutenkunst *et al.* 2009). The first terms in Eq. 2 model genetic drift and contain an additional constant scaling factor of $2/k_i$ compared to the typical term. Drift is thus slower in subgenomes with higher ploidy; a tetraploid subgenome ($k_i = 4$) experiences half the genetic drift of a diploid. This is simply because stochasticity is reduced when there are more chromosomes in the population. This same scaling factor also applies to the rate of mutation influx into each subgenome, compared to the diploid case. The second terms model exchange between subgenomes, which is analogous to migration between populations. The effect on subgenome i of exchange with subgenome j depends on $2E_{i\leftrightarrow j}/k_i$ compared to the typical migration term $M_{i\leftarrow j}$. The additional factor of $2/k_i$ implies that the effect on the allele frequency within a subgenome of a given influx of alleles decreases as the ploidy increases. Symmetric exchanges between genomes of different ploidy thus have asymmetric effects on allele frequencies. For example, an exchange between a diploid subgenome and a tetraploid subgenome, while biologically rare, could only lead to a $\frac{1}{4N}$ change in allele frequency in a single generation in the tetraploid subgenome versus a potential $\frac{1}{2N}$ change for the diploid subgenome.

With a distribution for the expected frequency of derived mutations in each subgenome at time t, we can generate the expected SFS for a sample of n individuals by integrating over the distribution of allele frequencies and calculating the probability of observing d_i derived alleles using a binomial distribution (Sawyer and Hartl 1992):

$$\mathbb{E}[d_1, \ldots, d_S] = \int_0^1 \cdots \int_0^1 \prod_{i=1,\ldots,S} \binom{nk_i}{d_i} x_i^{d_i} (1-x_i)^{nk_i - d_i} \phi(x_1, \ldots, x_S) dx_i. \quad (3)$$

Here each dimension of the SFS corresponds to a different subgenome.

Combining polyploid subgenomes: In practice, it may be difficult to partition allele counts between two or more subgenomes. In that case, two or more dimensions of the model SFS must be collapsed down to a single dimension for comparison with the observed SFS. This problem can arise due to issues with phasing, unknown or unsampled parental lineages, or simply unknown origin for the polyploid. As an example, consider a sample of n polyploid individuals with two subgenomes with ploidal levels k_1 and k_2. The original SFS for this population, $\mathbb{E}[d_1, d_2]$, would be a two-dimensional array of size $(nk_1 + 1) \times (nk_2 + 1)$. We collapse the two dimensions together using the following equation:

$$\mathbb{E}_{comb}[d] = \sum_{d_1=0}^{nk_1} \mathbb{E}[d_1, d - d_1]. \quad (4)$$

In words, the d entry of the reduced SFS is the summation of all entries for which the total allele count in the removed dimensions equals d. The resulting dimension of the new SFS is of size $nk_1 + nk_2 + 1$ and is what we use for comparison with the observed SFS when performing demographic inference with a polyploid population. If more than two subgenomes are indistinguishable, then this reduction process can be iterated to collapse all indistinguishable genomes into a single dimension of the SFS.

Validating the diffusion approximation for polyploids

To validate this diffusion approximation, we conducted simulations using SLiM v3.4 (Haller and Messer 2019) under various demographic models and parameter combinations for autotetraploids, allotetraploids, and segmental allotetraploids. For each simulation scenario, we simulated 1,000 polyploid individuals each with a single chromosome 1 Mb in length. Because polyploids have multiple subgenomes, a polyploid individual is composed of multiple SLiM individuals, and therefore subgenomes within individuals were simulated by treating each

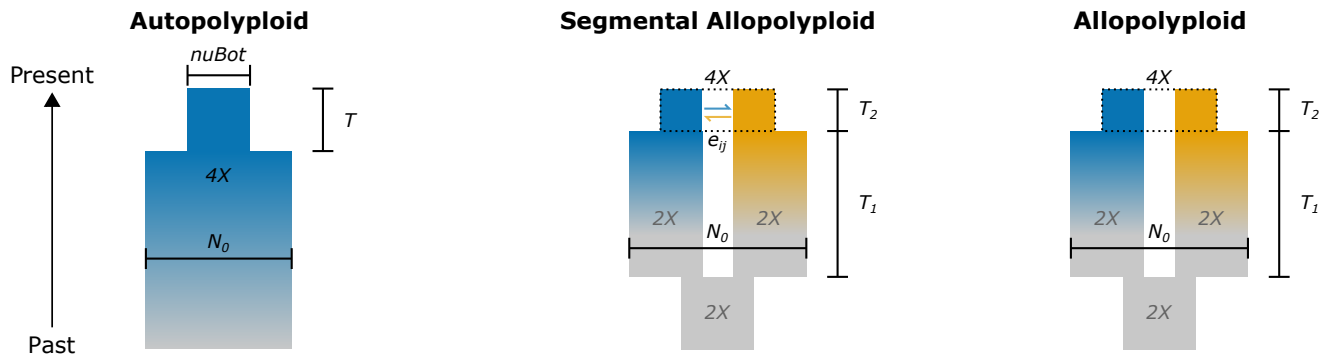

Figure 2 Graphical representations of the models used for validating the diffusion approximation for autopolyploids (left), segmental allopolyploids (middle), and allopolyploids (right). The main parameters used across the models are: N_0 (parental/ancestral population size), *nuBot* (ν_{bot}: proportion of population remaining after bottleneck), T (bottleneck duration in autopolyploid model), T_1 (duration of parental divergence before polyploid formation), T_2 (time before sampling for allo- and segmental allopolyploids), and e_{ij} ($e_{i \leftrightarrow j}$: per generation probability of homoeologous exchange).

subgenome as a separate SLiM population. Mutation and recombination rates were set such that $\theta = 2KN\mu$ was always equal to 5,000. At the end of each simulation, we generated 50 samples of 10 polyploid individuals by randomly sampling 10 diploid SLiM individuals from each of the SLiM populations and combining them into polyploid individuals to record the SFS. This was repeated 100 times for a grand total of 5,000 simulated frequency spectra for each scenario. We then constructed comparable models using the diffusion approximation implemented in dadi v2.2.0 (Gutenkunst *et al.* 2009) and compared the expected SFS returned by dadi with the mean SFS across the 5,000 replicates from SLiM to assess how well the two methods corresponded with each other. Specific simulation details for each category of polyploid are given in the following paragraphs.

Autopolyploids: For autotetraploids, we constructed a model in SLiM with two populations, representing the two diploid subgenomes that are assumed within the Meirmans and van Tienderen (2013) framework, each containing $N = 1,000$ diploid SLiM individuals. We set the mutation and recombination rates equal to 6.75×10^{-7} and included a symmetric per generation probability of migration equal to 0.5 ($e_{1 \leftrightarrow 2} = 1.0$) to allow free exchange of chromosomes between the subgenomes. Note that this deviates slightly from our original definition of homoeologous exchanges in that pairs of chromosomes (SLiM individuals), rather than single chromosomes, are moving between subgenomes. We return to this point in the Results and Discussion. An initial burn-in of 40,000 generations was used to reach an approximate state of equilibrium in genetic diversity. The first set of frequency spectra were sampled immediately after this burn-in period to obtain the mean SFS in an equilibrium population. We then simulated bottlenecks of three different sizes ($0.2 \times 2N$, $0.5 \times 2N$, and $1.0 \times 2N$), each one lasting for four different lengths of time ($1.0 \times 2N$, $2.0 \times 2N$, $3.0 \times 2N$, and $4.0 \times 2N$ generations) for a total of 12 parameter combinations.

After simulating data in SLiM, we specified a comparable model in dadi by assuming a single panmictic population with a sample size of 40 chromosomes. Models in dadi start with an equilibrium population, allowing us to immediately generate the expected SFS for a standard neutral model. For the models with bottlenecks, we obtained the expected SFS by including the bottleneck size and duration as parameters to the model and integrating the distribution of allele frequencies forward in time before sampling the resulting frequency spectrum. These expected frequency spectra generated using the diffusion approximation were then compared to the averaged spectra from SLiM to assess their level of agreement by plotting the difference in fit (residuals) between the two methods.

Allopolyploids: For allotetraploids, we built our model in SLiM with a initial diploid population of 1,000 individuals representing the ancestral reference population. We set the mutation and recombination rates to 1.25×10^{-6} and simulated 20,000 burn-in generations to reach approximate equilibrium. After the burn-in period, the ancestral population was split into two populations each containing 1,000 diploid individuals. These separate populations, which represent the ancestors of the allotetraploid subgenomes, were then simulated forward in time at a constant population size for varying numbers of generations ($T_1 = 0.5 \times 2N$, $1.0 \times 2N$, $1.5 \times 2N$, and $2.0 \times 2N$) to allow for divergence to develop between the two populations. For the first set of simulations, frequency spectra were sampled immediately after this period of divergence to emulate a newly formed allotetraploid. Within the framework we are proposing here, it is important to note that after this point of polyploid formation the SLiM populations are conceptually a single allotetraploid population. To emulate allopolyploid formation, we also conducted a set of simulations where, after the initial period of divergence between the parents, we introduced bottlenecks of different sizes ($\nu = 0.1 \times 2N$, $0.25 \times 2N$, and $0.5 \times 2N$), sampling the SFS after differing periods of time ($T_2 = 0.25 \times 2N$, $0.5 \times 2N$, and $1.0 \times 2N$ in generations).

The corresponding model specified in dadi began with a single diploid population that was split into two populations. For the models without bottlenecks, the populations were integrated forwards in time at a constant population size for the same amount of time as in SLiM (T_1) before sampling the two-dimensional SFS (20 chromosomes per SLiM population) and combining it into a one-dimensional frequency spectrum using Eq. 4. For models with bottlenecks, the two populations were once again integrated forwards at a constant size for time T_1 before experiencing an instantaneous bottleneck lasting for time T_2. Frequency spectra from these populations were sampled and combined in the same way and compared to the frequency spectra from SLiM.

Segmental allopolyploids: For segmental allotetraploids, the simulation setup was similar to the one used for allotetraploids. We included the same initial period of divergence (T_1), as well as the secondary period of time after polyploid formation (T_2). During this secondary period, we added two levels of allelic exchange between subgenomes. The levels we chose were $e_{i\leftrightarrow j} = 5 \times 10^{-5}$ and $e_{i\leftrightarrow j} = 5 \times 10^{-7}$. These levels correspond to one exchange event every 10 or every 1,000 generations, respectively. Parameters for the initial period of divergence, T_1, were kept the same. We also did simulations with bottlenecks, using bottlenecks of the same sizes, ν, during the time period T_2 for the corresponding set of models. Setting up the model in dadi for segmental allotetraploids was identical to allotetraploids except with the addition of the exchange parameter, $E_{i\leftrightarrow j} = 2Ne_{i\leftrightarrow j}$, during the integration for T_2.

Parameter inference and identifiability

In a separate set of SLiM simulations, we also sought to understand how well demographic parameters could be inferred from a combined polyploid SFS in dadi. For these simulations, we used a subset of the parameterizations listed above for allotetraploids and segmental allotetraploids, simulating 10 independent frequency spectra for each parameter combination. We also included an additional layer of complexity in these simulations by incorporating two different types of data generation meant to mimic data collected using genotyping by sequencing (GBS) and whole-genome resequencing data (WGS). For each GBS simulation, we generated 100 bp segments across 5000 independent SLiM runs and combined them into a single SFS. For each WGS simulation, we generated 5 Mb regions across 10 independent SLiM runs and combined them into a single SFS.

Models in dadi were specified as described in the previous section and were used to maximize the composite likelihood for the simulated input data. Parameters were initialized at random starting points and were estimated using the `nlopt`-enabled optimizer in dadi (BOBYQA algorithm; Powell 2009; Johnson 2014). We conducted 50 independent optimization replicates for each simulated data set for all models and simulation parameters to assess convergence on the same set of maximum likelihood parameter estimates from different starting points. After this, we used R v3.6 to sort the results by likelihood value, keeping the parameters producing the highest likelihood from the 50 optimizations across the twenty replicates for each simulation, and compared these estimates to the true value used to simulate the data.

Empirical example: *Capsella bursa-pastoris*

To model the demographic history of shepherd's purse, we first obtained the multidimensional SFS with the *C. bursa-pastoris* subgenomes separated from Douglas *et al.* (2015). We then combined the subgenomes into a one-dimensional frequency spectrum with Eq. 4 and used this SFS as input for parameter inference in dadi. Within dadi, we specified two models for comparison: the allotetraploid bottleneck model and the segmental allotetraploid bottleneck model. These two models are identical apart from the inclusion of allelic exchange between subgenomes in the segmental allotetraploid version. Parameters for each model were estimated with the BOBYQA algorithm in dadi using 100 independent optimization runs from different random starting points. We then used the parameter estimates with the best likelihood to compare the models with the observed data, as well as conducting a likelihood ratio test. Parameters were then converted from dadi units to real units using a mutation rate of 7×10^{-9} and total sequence length (L) of 773,748 bp, the same values used by Douglas *et al.* (2015).

Data availability

Supplemental files, including scripts for performing simulations, analyzing results, and generating plots, can be found on GitHub (https://github.com/pblischak/polyploid-demography.git) and are archived on figshare (https://doi.org/10.6084/m9.figshare.20635635.v3). All simulated data files, optimization results, and data files for analyses of *C. bursa-pastoris*, are also available on both GitHub and figshare.

Results and Discussion

We developed a generalized diffusion approximation for polyploids by extending previous work on the multi-population diffusion equation. Within this framework, we are able to accommodate the full continuum of polyploid formation types by explicitly parameterizing homoeologous exchanges between subgenomes. We then used simulations to validate the diffusion approximation by comparing it to results from forward simulations using collapsed frequency spectra to emulate the difficulties of separating parental subgenomes. We also investigated the accuracy of parameter inference under the diffusion framework for a subset of parameter values for allotetraploids and segmental allotetraploids. Finally, we used the polyploid diffusion model to infer the demographic history of *Capsella bursa-pastoris*, a widely distributed allotetraploid, finding evidence for allelic exchange between subgenomes.

The diffusion approximation in polyploids

We compared the expected SFS from the polyploid diffusion approximation as implemented in dadi with frequency spectra generated by forward simulations in SLiM. Figure 3 shows these results for a sample of parameter combinations for models including bottlenecks across autotetraploids, allotetraploids, and segmental allotetraploids. For these examples, as well as for the other parameters used for the simulations, we find good qualitative agreement between the frequency spectra from dadi and SLiM.

As might be expected for autopolyploids, the frequency spectra appear similar to what we would obtain for a diploid but with double the sample size (Fig. 3A&B). This is the case even though we simulated the autotetraploid in SLiM as two populations forming the subgenomes of a single polyploid population. For allotetraploids and segmental allotetraploids, combining allele frequencies between divergent parental lineages results in a characteristic spike in sites with allele frequencies of 50% (Fig. 3C-F). Much of this pattern is driven by opposite alleles drifting toward fixation in the two subgenomes, leading to fixed heterozygosity. This is more pronounced in the frequency spectra for populations experiencing a stronger bottleneck ($\nu = 0.1$ versus $\nu = 0.5$), where the spike at 50% frequency is higher and the drop off in the prevalence of sites with allele frequencies over 0.5 is greater. Segmental allotetraploids also differ in the appearance of their 50% frequency spike, having small shoulders of increased counts of sites with frequencies around 50% (Fig. 3C&D). This is caused by the allelic exchange between subgenomes generating allelic combinations that are not possible in allotetraploids due to the complete separation of subgenomes. As the exchange rate in segmental allotetraploids increases, the spike continues to level

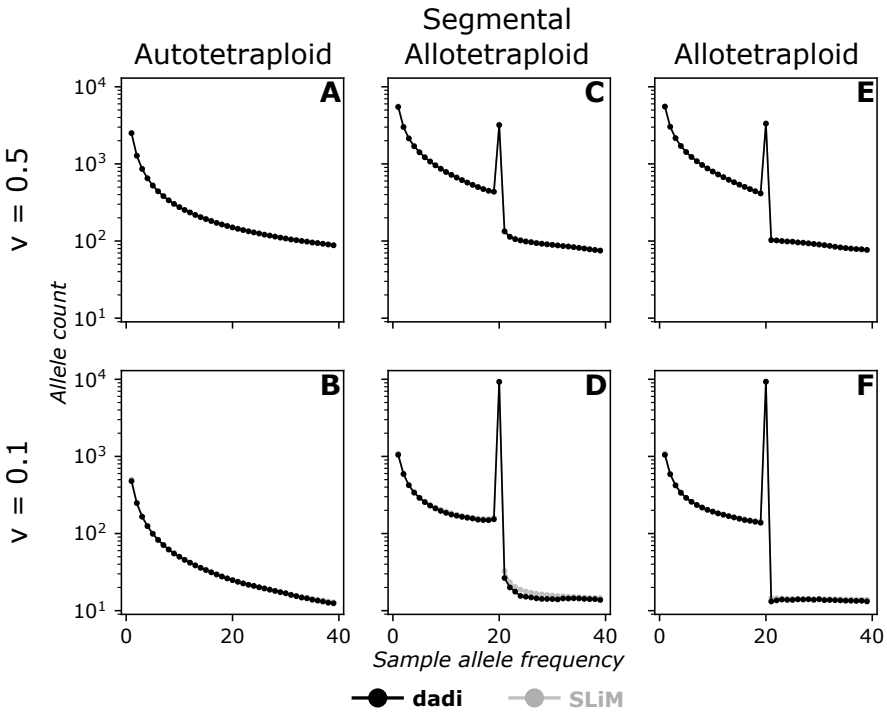

Figure 3 Comparison of collapsed frequency spectra between SLiM (black) and dadi (grey) for two different bottleneck sizes ($\nu = 0.5$ [top row] and $\nu = 0.1$ [bottom row]) for autotetraploids (A & B), segmental allotetraploids (C & D), and allotetraploids (E & F). The rate of homoeologous exchange for segmental allotetraploids was set to $e_{i \leftrightarrow j} = 5 \times 10^{-6}$.

out and eventually becomes visually indistinguishable from an autotetraploid (Fig. 4).

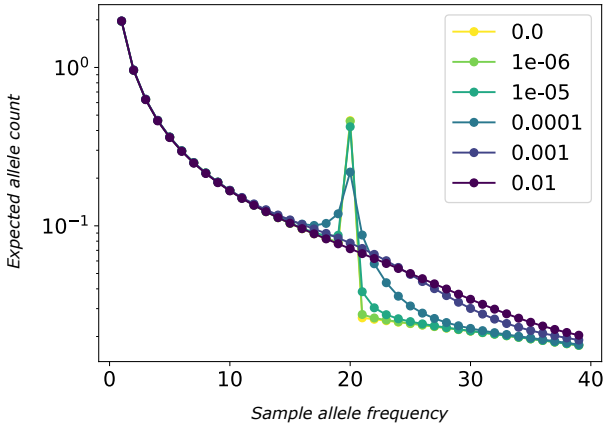

Figure 4 Illustration of the effect of the homoeologous exchange rate on the collapsed polyploid site frequency spectrum for a tetraploid. Here exchange rates vary from $e_{i \leftrightarrow j} = 0$ (allotetraploid) to $e_{i \leftrightarrow j} = 0.01$. At the high end of this range, the SFS no longer has a distinctive peak at 50% frequency due to the exchanges mixing alleles, making the SFS appear more similar to that of an autopolyploid.

Inferring demographic parameters in polyploids

As a follow-up to our validating simulations, we also sought to understand how well we could infer demographic parameters using the numerical approaches implemented in dadi for collapsed allotetraploid and segmental allotetraploid frequency spectra. For the parameters that describe the formation of the polyploid population itself, we are typically able to obtain precise parameter estimates across our simulated scenarios (Fig. 5). As expected, parameter estimates for the GBS data simulations generally show more variation than the WGS simulations.

Estimates for parameters that describe the demographic model before the formation of the polyploid are generally less precise and often showed unstable behaviors when searching for optimal values (see Supplemental Files). For example, estimates of the combined parental population sizes (denoted N_0 in our models) were consistently unstable/unbounded. This suggests that parameters estimated for the parental lineages of polyploid populations from a one-dimensional collapsed frequency spectrum should be interpreted with caution. For diploid populations and piecewise constant population histories, such instabilities have been well characterized by the geometry of the SFS and the effect of sampling error in the SFS on estimating parameter values on the boundary of the search space (Rosen *et al.* 2018).

Demographic history of *Capsella bursa-pastoris*

Shepherd's purse (*Capsella bursa-pastoris*; Brassicaceae) is a well-studied allotetraploid species resulting from hybridization between an outcrossing species, *C. grandiflora*, and a selfing species, *C. orientalis*. Previous work on the demographic history of *C. bursa-pastoris* found that it formed roughly 100 kya, with a current distribution spreading across Europe, the Middle East, and Asia (Douglas *et al.* 2015; Kryvokhyzha *et al.* 2019). These three regions also correspond to three major genetic groups within the species, all of which have experienced varied evolutionary trajec-

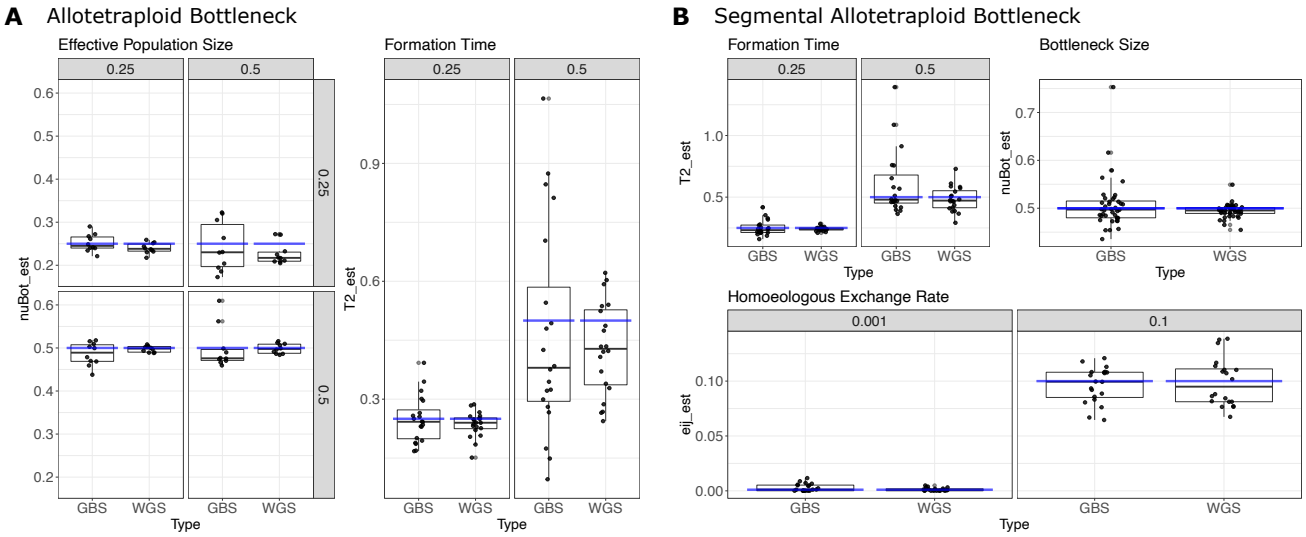

Figure 5 (A) Parameters estimates from dadi for bottleneck size (left panel) and formation time (right panel) for the allotetraploid bottleneck model simulated with SLiM across two different data types: Genotyping-by-sequencing (GBS) and whole-genome resequencing (WGS). For estimates of the bottleneck size, the secondary divisions in the plots show the true formation time (T_2 = 0.25, 0.5) in the rows and the true bottleneck size (ν_{Bot} = 0.25, 0.5) in the columns. (B) Parameters estimates from dadi for formation time (top-left panel), bottleneck size (top-right panel), and homoeologous exchange rate (bottom panel) for the segmental allotetraploid bottleneck model simulated with SLiM across GBS and WGS data types.

tories including bottlenecks and changes in life history, though the Middle Eastern region is where the species is inferred to have originated (Cornille *et al.* 2016). Here we use data for shepherd's purse from Douglas *et al.* (2015) as an empirical example to compare the results of conducting demographic inference with a one-dimensional collapsed SFS to those obtained by the original study.

Using a collapsed representation of the SFS from Douglas *et al.* (2015), we modeled the demographic history of *Capsella bursa-pastoris* using the allotetraploid bottleneck and segmental allotetraploid bottleneck models (Table 1). Both models resulted in similar estimates for the parameters regarding the formation of the polyploid lineage (ν_{bot} and T_2; see Fig. 2), finding that *C. bursa-pastoris* was formed around ∼270–285 kya with an initial effective population size of ∼50,000–55,000 individuals. These estimates differ somewhat from the estimates reported in (Douglas *et al.* 2015), who found that *C. bursa-pastoris* formed more recently, between ∼22–177 kya. However, Douglas *et al.* (2015) estimated separate effective populations sizes for each subgenome, finding similar values to our combined estimate: ∼6,000–55,000 individuals for the *C. grandiflora* subgenome (*C. bursa-pastoris* A) and ∼12,000–101,000 for the *C. orientalis* subgenome (*C. bursa-pastoris* B). For most other parameters, the allotetraploid and segmental allotetraploid models had drastically different estimates of the ancestral and combined parental population sizes (N_A and N_0, respectively) and a roughly 0.75X difference in the estimate of parental divergence (T_1). This is consistent with the instablility observed in our simulation studies for pre-polyploid formation parameters. The estimated exchange rate ($e_{i\leftrightarrow j}$) for the segmental allotetraploid model was 6×10^{-8}, suggesting that rare bouts of allelic exchange may have occurred between the subgenomes of *C. bursa-pastoris*. This is corroborated by the fact that the composite log-likelihood for the segmental allotetraploid model was 138 units higher than for the allotetraploid model, resulting in a likelihood ratio test statistic of 277, corresponding to a vanishingly small p-value.

Our estimates of the timing and population size impacts of polyploid formation for *C. bursa-pastoris* are similar to those of Douglas *et al.* (2015), but our estimates of pre-polyploid parameters differ substantially. This is not unexpected, given that our simulations experiments exhibited unbounded behavior in likelihood optimization for those parameters. Furthermore, the parental taxa forming *C. bursa-pastoris* have different life history strategies, with *C. orientalis* being highly inbred. This could be leading to the differences we see in our estimates of divergence between the parental lineages and their effective population sizes. Incorporating this inbreeding into the model should be possible using sampling schemes that build inbreeding in their derivation of the expected SFS (Blischak *et al.* 2020).

Another important distinction between our analyses and those of Douglas *et al.* (2015) is their use of a four-population model for their demographic inferences, including both parents, *C. grandiflora* and *C. orientalis*, and separating the corresponding subgenomes of *C. bursa-pastoris*. In our combined analysis, we are deliberately excluding this additional information to better understand the limits of demographic inference within a single polyploid population. The result is that we are not able to reliably estimate all model parameters. However, we do recover strong evidence for the occurrence of rare homoeologous exchanges between the two subgenomes of shepherd's purse. Future work using our modeling framework with the inclusion of parental lineages, inbreeding, and other demographic factors will help us to better disentangle the evolutionary forces affecting *C. bursa-pastoris* as well as other polyploid species.

General considerations for modeling the demographic history of polyploids

Researchers studying polyploid taxa are faced with numerous challenges when analyzing genomic data to understand more about the evolutionary history of their study organism(s). Here

Table 1 Parameter estimates for *Capsella bursa-pastoris* demographic history. Population sizes (N_*) are reported as the number of individuals and time intervals (T_*) are reported as the number of years.

Parameter	Allotetraploid Bottleneck	Segmental Allotetraploid Bottleneck
N_A	248970	7723
N_0	24897004	772310
N_{bot}	50371	55346
T_1	1027900	1544621
T_2	268515	283783
e_{ij}	–	6.0×10^{-8}
% misidentified	0.017	0.031

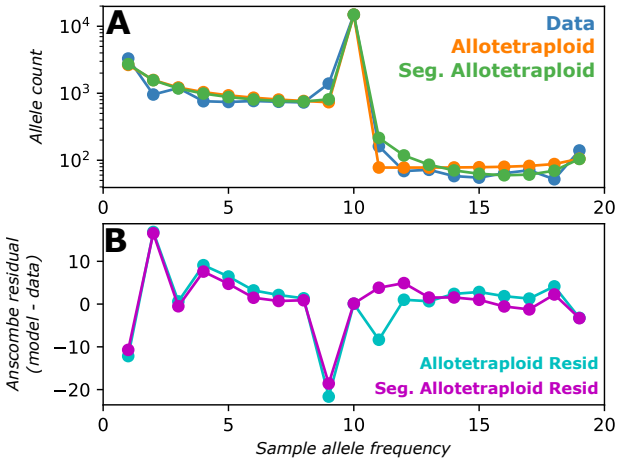

Figure 6 (A) Site frequency spectra resulting from the maximum likelihood parameters estimated for the allotetraploid bottleneck (orange) and segmental allotetraploid bottleneck (green) models for *C. bursa-pastoris*. The observed data are also shown in blue. (B) Residual plots comparing each entry in the SFS between the allotetraploid bottleneck (cyan) and segmental allotetraploid bottleneck (magenta) models with the observed data.

we have proposed a model to help alleviate some of these issues by explicitly parameterizing the continuum of polyploid formation types based on the expectations for a collapsed polyploid SFS. Given the results of our simulation and empirical analyses, there are clear advantages and disadvantages to analyzing data in this way. One advantage is that there is no longer a need to separate variation occurring within potential subgenomes, because the model accommodates auto-, allo-, and segmental allopolyploids. Assuming complete separation of subgenomes in an allopolyploid or completely free recombination in an autopolyploid can lead to unintentionally ignoring signals for intermediate patterns that may have important consequences. Furthermore, there may be only certain parts of the genome that experience exchanges. For example, allotetraploid *Arabidopsis suecica*, while primarily having bivalent pairing of chromosomes within each of its subgenomes, was found to have variation in patterns of gene expression owing to a relatively small number of homoeologous exchanges present in some samples (Burns *et al.* 2021). In this case, the authors were able to leverage the robust genomic resources available in *Arabisopsis* to recognize a signal for homoeologous exchange. For most species, however, identifying the mode of formation will have to be done in the absence of a reference genome or genomes. With our model, the shape of the collapsed SFS can be a good initial indicator of whether the species is an auto- or allopolyploid, with shoulders around any peaks at 50% frequency providing additional evidence for the presence of homoeologous exchanges. Further investigation using the likelihood-based framework in dadi to perform model comparisons for determining the mode of formation could also provide a robust means of identifying even small amounts of homoeologous exchange in polyploid lineages.

The primary disadvantage of analyzing data using a collapsed spectrum is the inability to reliably infer ancestral dynamics prior to polyploid formation. This pattern was observed across our simulations and the analysis of *C. bursa-pastoris*. As we mentioned above, one way to potentially deal with these issues would be to include the parental lineages in the demographic model. This does require additional knowledge about the study system but should be feasible within dadi, if the data are available, by adding up to four or five populations using the newly implemented graphics processing unit acceleration (Gutenkunst 2021). Theoretical investigations into the limitations of using a collapsed polyploid frequency spectrum could also further illuminate parameter and model identifiability and could be used to guide additional innovation in the construction

of more informative demographic models for polyploids.

Conclusions

Disentangling the roles of demography and selection in polyploids will be a major step toward better understanding the role they play in the generation and maintenance of biodiversity. Additionally, an appreciation for the continuum-like nature of polyploids will remain essential as more methods are developed to model their evolutionary histories. The method we have developed here provides a foundation for further exploration of diffusion-based demographic models for polyploids and reveal important pain points and considerations for how to approach demographic modeling with a collapsed polyploid SFS. Including parental lineages and gaining a better theoretical understanding of the effect of fixed differences between subgenomes on demographic inference will be key advancements in future iterations of our modeling approach. Combining more robust demographic models for polyploids with existing frameworks for studying selection (e.g., Huang *et al.* 2021) will then provide a powerful framework for revealing their evolutionary importance in the short term.

Acknowledgements

The authors thank Justin Conover for comments that helped to improve the manuscript and Stephen Wright for sharing the data for *Capsella bursa-pastoris*. We also would like to thank J. Wendel, T. Steussey, and attendees of the Polyploid Webinar (https://www.barkerlab.net/polyweb) for their thoughtful input regarding effective population sizes in polyploids, the nature of segmental allopolyploidy, and for exhibiting a general enthusiasm for all things whole-genome duplication.

Funding

This work was supported by a National Science Foundation Postdoctoral Research Fellowship in Biology (IOS-1811784 to P.D.B.) and by the National Institute of General Medical Sciences of the National Institutes of Health (R01GM127348 to R.N.G.).

Conflicts of Interest

None declared.

Literature cited

Arnold, B. J., B. Lahner, J. M. DaCosta, C. M. Weisman, J. D. Hollister, *et al.*, 2016 Borrowed alleles and convergence in serpentine adaptation. Proceedings of the National Academy of Sciences USA **113**: 8320–8325.

Baduel, P., S. Bray, M. Vallejo-Marin, F. Kolář, and L. Yant, 2018 The "polyploid hop": Shifting challenges and opportunities over the evolutionary lifespan of genome duplications. Frontiers in Ecology and Evolution **6**.

Baniaga, A. E., H. E. Marx, N. Arrigo, and M. S. Barker, 2020 Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. Ecology Letters **23**: 68–78.

Blischak, P. D., M. S. Barker, and R. N. Gutenkunst, 2020 Inferring the demographic history of inbred species from genome-wide snp frequency data. Molecular Biology and Evolution **37**: 2124–2136.

Blischak, P. D., L. S. Kubatko, and A. D. Wolfe, 2018a SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. Bioinformatics **34**: 407–415.

Blischak, P. D., M. E. Mabry, G. C. Conant, and J. C. Pires, 2018b Integrating networks, phylogenomics, and population genomics for the study of polyploidy. Annual Review of Ecology, Evolution, and Systematics **49**: 253–278.

Burns, R., T. Mandáková, J. Gunis, L. M. Soto-Jiménez, C. Liu, *et al.*, 2021 Gradual evolution of allopolyploidy in *Arabidopsis suecica*. Nature Ecology & Evolution **5**: 1367–1381.

Clark, L. V., A. E. Lipka, and E. J. Sacks, 2019 polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids. G3 Genes | Genomes | Genetics **9**: 663–673.

Clevenger, J. P., W. Korani, P. Ozias-Akins, and S. Jackson, 2018 Haplotype-based genotyping in polyploids. Frontiers in Plant Science **9**: 564.

Clevenger, J. P. and P. Ozias-Akins, 2015 SWEEP: A Tool for Filtering High-Quality SNPs in Polyploid Crops. G3 Genes | Genomes | Genetics **5**: 1797–1803.

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

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

Cui, L., P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, *et al.*, 2006 Widespread genome duplications throughout the history of flowering plants. Genome Research **16**: 738–749.

Douglas, G. M., G. Gos, K. A. Steige, A. Salcedo, K. Holm, *et al.*, 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 **112**: 2806–2811.

Furlong, R. F. and P. W. H. Holland, 2001 Were vertebrates octoploid? Philosophical Transactions of the Royal Society B: Biological Sciences **357**: 531–544.

Gerard, D., L. F. V. Ferrão, A. A. F. Garcia, and M. Stephens, 2018 Genotyping polyploids from messy sequencing data. Genetics **210**: 789–807.

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

Gutenkunst, R. N., 2021 dadi.CUDA: Accelerating Population Genetics Inference with Graphics Processing Units. Molecular Biology and Evolution **38**: 2177–2178.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics **5**: e1000695.

Haller, B. C. and P. W. Messer, 2019 SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. Molecular Biology and Evolution **36**: 632–637.

Huang, X., A. L. Fortier, A. J. Coffman, T. J. Struck, M. N. Irby, *et al.*, 2021 Inferring Genome-Wide Correlations of Mutation Fitness Effects between Populations. Molecular Biology and Evolution **38**: 4588–4602.

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

Johnson, S. G., 2014 *The NLopt nonlinear-optimization package, http://github.com/stevengj/nlopt*.

Kimura, M., 1964 Diffusion models in population genetics. Jour-

nal of Applied Probability **1**: 177–232.

Korani, W., J. P. Clevenger, Y. Chu, and P. Ozias-Akins, 2019 Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants. Plant Genome **12**: 180023.

Kryvokhyzha, D., A. Salcedo, M. C. Eriksson, T. Duan, N. Tawari, *et al.*, 2019 Parental legacy, demography, and admixture influenced the evolution of the two subgenomes of the tetraploid *Capsella bursa-pastoris* (Brassicaceae). PLOS Genetics **15**: 1–34.

Leebens-Mack, J. H., M. S. Barker, E. J. Carpenter, M. K. Deyholos, M. A. Gitzendanner, *et al.*, 2019 One thousand plant transcriptomes and the phylogenomics of green plants. Nature **574**: 679–685.

Li, Z. and M. S. Barker, 2020 Inferring putative ancient whole-genome duplications in the 1000 Plants (1KP) initiative: access to gene family phylogenies and age distributions. GigaScience **9**: giaa004.

Li, Z., G. P. Tiley, S. R. Galuska, C. R. Reardon, T. I. Kidder, *et al.*, 2018 Multiple large-scale gene and genome duplications during the evolution of hexapods. Proceedings of the National Academy of Sciences **115**: 4713–4718.

Mason, A. S. and J. F. Wendel, 2020 Homoeologous exchanges, segmental allopolyploidy, and polyploid genome evolution. Frontiers in Genetics **11**: 1014.

McIntyre, P. J. and S. Strauss, 2017 An experimental test of local adaptation among cytotypes within a polyploid complex. Evolution **71**: 1960–1969.

Meirmans, P. G. and P. H. van Tienderen, 2013 The effects of inheritance in tetraploids on genetic diversity and population divergence. Heredity **110**: 131–137.

Monnahan, P., F. Kolář, P. Baduel, C. Sailer, J. Koch, *et al.*, 2019 Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. Nature Ecology & Evolution **3**: 457–468.

Ohno, S., 1970 *Evolution by gene duplication*. Springer.

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

Powell, M. J. D., 2009 The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report 2009/NA06, Department of Applied Mathematics and Theoretical Physics, Cambridge University.

Rosen, Z., A. Bhaskar, S. Roch, and Y. S. Song, 2018 Geometry of the sample frequency spectrum and the perils of demographic inference. Genetics **210**: 665–682.

Roux, C. and J. R. Pannell, 2015 Inferring the mode of origin of polyploid species from next-generation sequence data. Molecular Ecology **24**: 1047–1059.

Roux, C., X. Vekemans, and J. Pannell, 2021 Inferring the demographic history of tetraploid species from genomic data. bioRxiv p. 10.1101/2021.07.10.451876.

Sawyer, S. A. and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics **132**: 1161–1176.

Selmecki, A. M., Y. E. Maruvka, P. A. Richmond, M. Guillet, N. Shoresh, *et al.*, 2015 Polyploidy can drive rapid adaptation in yeast. Nature **519**: 349–352.

Serang, O., M. Mollinari, and A. A. F. Garcia, 2012 Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. PLOS ONE **7**: 1–13.

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

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, *et al.*, 2009 The frequency of polyploid speciation in vascular plants. Proceedings of the National Academy of Sciences **106**: 13875–13879.