Model-based genotype and ancestry estimation for potential hybrids with mixed-ploidy

Vivaswat Shastry[1], Paula E. Adams[2], Dorothea Lindtke[3], Elizabeth G. Mandeville[4], Thomas L. Parchman[5], Zachariah Gompert[6], and C. Alex Buerkle[1]

[1] Department of Botany, University of Wyoming, Laramie, Wyoming 82071, USA
[2] Department of Biological Sciences, University of Alabama, Tuscaloosa, Alabama 35401, USA
[3] Institute of Plant Sciences, University of Bern, 3013 Bern, Switzerland
[4] Department of Integrative Biology, University of Guelph, Guelph, Ontario N1G 2W1, Canada
[5] Department of Biology, University of Nevada–Reno, Reno, Nevada 89557, USA
[6] Department of Biology, Utah State University, Logan, Utah 84322, USA

Corresponding author: C. Alex Buerkle
1000 E. University Ave.
Department of Botany, 3165
University of Wyoming
Laramie, WY 82071, USA
buerkle@uwyo.edu

Keywords: *admixture, ancestry, polyploids, genotype likelihoods, hybridization, introgression*

Running title: Genotype and ancestry estimation

Abstract

Non-random mating among individuals can lead to spatial clustering of genetically similar individuals and population stratification. This deviation from panmixia is commonly observed in natural populations. Consequently, individuals can have parentage in single populations or involving hybridization between differentiated populations. Accounting for this mixture and structure is important when mapping the genetics of traits and learning about the formative evolutionary processes that shape genetic variation among individuals and populations. Stratified genetic relatedness among individuals is commonly quantified using estimates of ancestry that are derived from a statistical model. Development of these models for polyploid and mixed-ploidy individuals and populations has lagged behind those for diploids. Here, we extend and test a hierarchical Bayesian model, called `entropy`, which can utilize low-depth sequence data to estimate genotype and ancestry parameters in autopolyploid and mixed-ploidy individuals (including sex chromosomes and autosomes within individuals). Our analysis of simulated data illustrated the trade-off between sequencing depth and genome coverage and found lower error associated with low depth sequencing across a larger fraction of the genome than with high depth sequencing across a smaller fraction of the genome. The model has high accuracy and sensitivity as verified with simulated data and through analysis of admixture among populations of diploid and tetraploid *Arabidopsis arenosa*.

Introduction

Species are distributed across geographic ranges and potentially heterogeneous environments, and experience barriers to dispersal. Thus, a species rarely corresponds to a single, genetically homogeneous, panmictic population. This differentiation across the geographic range can consist of clinal variation, genetic subdivisions into local populations or 'demes', or some combination of both (Endler, 1977; Bradburd *et al.*, 2013; Gompert & Buerkle, 2016). Even

species with high rates of dispersal can have geographic ranges that are large relative to dispersal distances (e.g., Novembre *et al.*, 2008; Phifer-Rixey *et al.*, 2018), such that the distribution of traits and alleles is commonly heterogeneous and stratified among geographic locations.

Quantifying population heterogeneity and stratification is a fundamental component of empirical population genetics, both to provide a context for the study of evolutionary dynamics and as a component of learning about trait genetics in natural populations. Information about population structure and mixtures can reveal aspects of the underlying evolutionary processes and has played a significant role in shaping our understanding of the nature of hybridization, speciation, and adaptation. This includes knowledge of the prevalence of gene flow and introgression, as well as variability in introgression among geographic sites and genomic regions (e.g., Nadeau *et al.*, 2012; Abbott *et al.*, 2013; Gompert *et al.*, 2014b; Mandeville *et al.*, 2017; Meier *et al.*, 2017). For example, `structure`-like models are commonly used to quantify the proportion of an individual's genome inherited from each of K hypothetical source populations, which corresponds to their ancestry or admixture composition (Pritchard *et al.*, 2000; Falush *et al.*, 2003; Gompert *et al.*, 2014b). Comparisons of parameter estimates from models with different numbers (K) of source populations can guide an understanding of hierarchy and spatial genetic structure and admixture among the sampled individuals. Beyond `structure`-like models, there is considerable interest in estimates of locus-specific ancestry and introgression, with a corresponding wealth of existing and continuously developing methods in computational statistics (e.g., Sankararaman *et al.*, 2008; Gompert & Buerkle, 2013; Gompert, 2016; Rosenzweig *et al.*, 2016; Ottenburghs *et al.*, 2016; Schumer *et al.*, 2019, for a review, see Gompert *et al.* 2017). These include parametric methods for detecting loci with ancestry that is concordant with the remainder of the genome (e.g., Szymura & Barton, 1986; Gompert & Buerkle, 2011a), or for detecting breakpoints and tracts of ancestry among chromosomal blocks or haplotypes (e.g., Wegmann *et al.*, 2011; Lawson *et al.*, 2012; Sohn *et al.*, 2012; Gompert, 2016). Similarly, researchers have contrasted

ancestry and introgression of sex chromosomes relative to ancestry of autosomes in hybrid zones (Harrison & Larson, 2016; Chaturvedi *et al.*, 2020).

Accounting for population stratification and mixtures is typically a critical component of trait mapping in natural populations. Accounting for population stratification can reduce the number of false positive associations between loci and trait variation (e.g., Pritchard & Donnelly, 2001; Haworth *et al.*, 2019). Admixture coefficients or genetic kinship matrices can quantify diffuse genetic effects that are attributable to the genetic background of individuals (overall ancestry), rather than the effects of individual genetic loci (Zhou *et al.*, 2013; Hellwege *et al.*, 2017).

Despite the abundance of non-parametric statistical methods (e.g., `EIGENSTRAT`, Price *et al.* 2006 and `DAPC`, Jombart *et al.* 2010) and parametric models for population structure, methods for quantifying admixture in autopolyploid or mixed-ploidy individuals (combination of autosomes and sex chromosomes within individuals, or a mixture of ploidal levels among individuals in a population) are not fully developed. This is true even though 16% of all plant species contain some ploidal variation (Rice *et al.*, 2015). The dynamics of mixed-ploidy species can reveal processes governing polyploid evolution and the role of ploidal variation in adaptation and speciation (Kolář *et al.*, 2017). Autopolyploids harbour multiple complete haploid subgenomes with sets of homologous chromosomes that share recent common ancestry and that aggregate and then segregate randomly in meiosis, leading to polysomic inheritance. Hence, methods for autopolyploid genetics should contain the ability to treat each allele copy at a locus as being independent. In contrast, in allotetraploids with disomic inheritance, loci can be modeled as having diploid genotype values (and use the methods previously developed for diploids), instead of modeling complete tetraploid genotypes as with autotetraploids (even with minimal information on the origin of reads from the two different subgenomes using the model presented in Blischak *et al.*, 2017). `structure` can be used with autopolyploid and mixed-ploidy individuals, but lacks the ability to utilize genotype likelihoods as input data and thereby account for uncertainty in genotype calls,

and requires a model misspecification to accommodate variable ploidy (i.e., by assuming a single ploidal level for input genotype data across all individuals, Meirmans *et al.*, 2018; Stift *et al.*, 2019). Differences in genotyping errors could occur across ploidal levels and cause potential artefacts if `structure` were applied to a mixed-ploidy data set, though the magnitude of such effects in estimation have not been well studied (Ferretti *et al.*, 2018). As a result, `structure` cannot make full use of low-depth sequencing, or be used as a population model for estimating genotypes (including imputation of missing genotypes). Other methods that utilize genotype likelihoods and low depth sequences have not been extended to polyploids (Skotte *et al.*, 2013; Meisner & Albrechtsen, 2018). The use of the full distribution of genotype likelihoods (from `GATK`, McKenna *et al.* 2010, `SAMtools`, Li 2011, or `FreeBayes`, Garrison & Marth 2012), rather than point estimates of genotypes, is particularly appropriate for polyploids in which a heterozygous genotype can arise from multiple dosages of alternative alleles (e.g., 1:3, 2:2, and 3:1 in a tetraploid) that will be difficult to distinguish, particularly with low sequencing depth. More generally, methods that utilize genotype likelihoods from all appropriately filtered loci will make more complete and better use of the available genomic data to estimate ancestry and genotypes (Gompert & Buerkle, 2011b; Nielsen *et al.*, 2012; Buerkle & Gompert, 2013; Vieira *et al.*, 2013), including for estimating genotypes to map phenotypes to the genomes of polyploids (Grandke *et al.*, 2016). In addition to the class of `structure`-like models for population allele frequencies and individual ancestry, methods have been developed to estimate genotypes from polyploid sequence data, without considering population structure and admixture (`EBG`, Blischak *et al.* 2017; `updog`, Gerard *et al.* 2018; `polyRAD`, Clark *et al.* 2019).

A recent simulation study (Stift *et al.*, 2019) showed that model-based approaches like `structure` outperform other ancestry-estimation methods for the analysis of mixed-ploidy populations. Likewise, using an evolutionary model for allele frequencies in populations, including a `structure`-like model, improves estimates of genotypes from sequence data relative to methods that do not use population models (Gompert *et al.*, 2014b; Clark *et al.*,

2019). At the same time, the assumption of `structure`-like models of admixture among ancestral demes should be tested, so as to avoid model misspecification and being misled for some populations and instances of gene flow (e.g., when there is additional substructure within the assumed ancestral populations, or inference is based on discrete samples along a continuous, isolation by distance gradient, etc., see Lawson *et al.*, 2018). Whereas the model can apply well to cases of contemporary hybridization and population mixtures, model misspecification can lead to incorrect inferences. Hence, the importance of model choice and fit has spurred further development of methods to gauge the appropriateness of the model for individual studies (Gompert *et al.*, 2014b; Garcia-Erill & Albrechtsen, 2019; Chaturvedi *et al.*, 2020).

With these motivations, we extend and thoroughly test the performance of a model similar to the admixture model implemented in `structure` (a version for diploids was presented previously as part of analyses in Gompert *et al.*, 2014b) to detect and quantify contemporary population structure in mixed-ploidy populations. In our `entropy` software, we specifically model mixed-ploidy by allowing for variable ploidal level across individuals (ranging from haploid to hexaploid). We have implemented methods for autopolyploids, since allopolyploids can be modelled as a lower ploidy, given sufficient knowledge of genome organization and chromosome pairing, including which loci occur on the pairs of homoeologous chromosomes (Bourke *et al.*, 2018). Herein we also restate a novel ancestry-estimation method (*ancestry complement* model for diploids, previously presented in the Supplementary Material of Gompert *et al.*, 2014b) that considers the ancestry of allele combinations in diploid genotypes, rather than allele copies independently, which provides additional information about the composition of early generation hybrids. We quantify the ability of the `entropy` model to recover true parameters from polyploid and mixed-ploidy sequencing data in simulations (with varying sequence depth, population differentiation, and percent of missingness) and through reanalysis of previously published data for population structure and admixture of mixed-ploidy *Arabidopsis arenosa* in Monnahan *et al.* (2019). From our testing, we

conclude that the `entropy` model has high accuracy rate in recovering true genotype and ancestry estimates from a variety of simulations, and further resolves population mixtures in empirical data from a diploid-tetraploid hybrid zone.

Methods

Model specification

Our hierarchical Bayesian model describes the probability of parameters of interest (genotype, population allele frequency, admixture proportion, etc.) given the data (genotype likelihoods for individual SNPs), and is similar to the admixture model implemented in the software `structure` (Pritchard *et al.*, 2000; Falush *et al.*, 2003). This model has multiple hierarchical levels, such that the joint product across the hierarchy does not have a closed form, analytical solution. Instead, we rely on Markov chain Monte-Carlo (MCMC) methods to obtain samples from the posterior probability distributions of these parameters. Several related models have been implemented over the years that use a similar idea to obtain parameter estimates through various computational techniques, the most commonly used being Bayesian MCMC (e.g., Pritchard *et al.*, 2000) and Expectation-Maximization (EM) of a likelihood (e.g., Tang *et al.*, 2005), and more recently, variational inference of a posterior (e.g., Raj *et al.*, 2014). We chose to use Bayesian MCMC so as to obtain measures of uncertainty associated with the estimates of our parameters, especially since we wanted the model to be usable with uneven and low depth DNA sequence data. The measures of uncertainty are useful in interpretation of point estimates and can be carried forward into subsequent analyses.

We deviate from several previous models by using genotype likelihoods as input instead of fixed genotypes, as a way of propagating this uncertainty from the data to the inference of parameters. One can think of `entropy` as a data generative model that tries to match the

genotype likelihoods (or genotypes) that are observed to an evolutionary process (parameterized by the allele frequency p, ancestry z and so on) that could have generated the data (Figure 1).

The evolutionary process we assume here starts with an ancestral population (characterized by allele frequency, π) that evolves through drift (parameterized by F using the Balding-Nichols model, Balding & Nichols, 1995) to give rise to the K 'parental' populations (each characterized by allele frequency, p) from which potentially admixed individuals are drawn, with admixture quantified by proportion q. We then use the observed genotype likelihoods (obtained from sequencing individuals) to match this evolutionary process and estimate our parameters through a hierarchical Bayesian model. In the following subsections, we explain what each parameter is and how it fits into the assumed evolutionary model.

The process of sampling and estimating these parameters of interest in code follows common methods for MCMC. After initialization, we begin by updating parameters at the lowest level of the model hierarchy (parameter γ and α), followed by updates of parameters in conditional probability functions in the next higher level of the hierarchy (here, $\mathbf{q}$ or $\mathbf{Q}$, π and F). We continue this type of sampling at each level, updating parameters individually by either Gibbs or Metropolis updates (depending on whether we have a conjugate prior for our conditional likelihood), until we reach the top level of the hierarchy, the probability of the data conditional on the model parameters. At this step, the estimates for the parameter are informed by the data (in this case, genotype likelihoods $\mathbf{X}$) and the parameter's prior probability given the current values of other parameters in the model. This type of one-at-a-time sampling and updating of parameters takes place at each step in a run of the model, and steps are iterated sufficiently in an MCMC run (a chain) to converge to stationary distributions for all the parameters in the hierarchy. In the sections that follow, we describe each of the conditional probabilities, moving from the base of the model hierarchy to the likelihood (Figure 1). A more detailed description of conditional distributions of parameters and MCMC sampling techniques is provided in the Supplementary Material.

Admixture proportion (q and Q)

In this model, admixture proportion or ancestry in an individual is the proportion of an individual's genome that is derived from one of K source populations. The admixture proportions are estimates of the average genome-wide or global ancestry for an individual and, with information on the individuals descended solely from parental populations, can be used to describe hybridization among the demes represented in the sample (as shown in Gompert *et al.*, 2017). As a result, this quantity is a vector of length K that sums to one for each individual. By modeling potential admixture in individuals, the model applies to both individuals coming entirely from a single deme, and also to individuals that are the progeny of crosses between demes (such as F1, F2, F3, and backcrosses). Conditional on a certain K number of demes and their allele frequencies, the model accounts for the genotypes that may be present in an individual and the individual's fractional ancestry in each deme (Pritchard *et al.*, 2000).

The `entropy` model includes two different models for the estimation of admixture proportion of individuals. The first is the q-model, is similar to the `structure` admixture model with correlated allele frequencies (Falush *et al.*, 2003). Here, we specify a vector of admixture proportions, denoted $\mathbf{q} = [q_1, q_2, \ldots, q_k]$ to indicate the proportion of an individual's genome that was inherited from each source population. These parameters are the probability of sampling a particular ancestry for an individual allele copy at the locus, independent of other alleles, which is equivalent to Hardy-Weinberg expectations that arise from random mating.

The second model in `entropy` is the *ancestry complement* model and considers the combination of ancestry for pairs of alleles across all loci in diploid individuals (the model is specified for diploids only). In early generation hybrids, interspecific (or inter-demic) combinations of alleles are expected to be common. Parameterization of the ancestry combination of the pair of allele copies in the *ancestry complement* model allows for deviations from

192 independence, which is assumed among allele copies in the simpler q-model. The ancestry
193 combinations are represented in a $k \times k$ dimension matrix, $\mathbf{Q}$. For example, with $K = 2$
194 demes the ancestry complement matrix $\mathbf{Q}$ is 2×2 in dimension. Q_{11} denotes the proportion
195 of the individual's genome in which both allele copies are descended from source population
196 1. Similarly, Q_{22} denotes the proportion of the individual's genome in which both copies are
197 descended from source population 2, and $Q_{12} = Q_{21}$ denotes the proportion of the genome
198 in which one allele copy is from source population 1 and the other allele copy is descended
199 from source population 2 (since the order of the allele copy does not matter, Q_{12} is equal
200 to Q_{21}). In the *ancestry complement* model, the admixture proportion vector $\mathbf{q}$ is a de-
201 rived quantity from the admixture complement matrix $\mathbf{Q}$. For instance, with $K = 2$ demes,
202 $q_i = Q_{ii} + \sum_{\substack{k'=1 \\ k' \neq i}}^{2} \frac{Q_{ik'}}{2}$ for $i = \{1, 2\}$.

203 As noted above, the benefit to the admixture complement parameterization is that it
204 explicitly models the the combination of ancestry states at a locus, which is particularly
205 beneficial in distinguishing among early generations of hybrid individuals (i.e., F1, F2, F3,
206 and BC1). For first generation hybrids between parental taxa (F1) and between hybrids
207 that have no parentage involving backcrossing (F2, F3, etc.) the expected value for q_1 is 0.5,
208 with some variance in observed individuals. This means that with the $\mathbf{q}$ vector alone, we
209 can distinguish recent hybrids from the parentals and maybe backcrosses but not distinguish
210 F1s from later generation hybrids. Likewise, distinguishing backcrosses from the parentals
211 for later generations of hybrids is difficult with the admixture proportion vector $\mathbf{q}$ alone,
212 given chance deviations from the expected values (Lindtke *et al.*, 2014). Particularly for
213 early generations of hybridization between a pair of taxa, the combination of information
214 in the admixture complement matrix $\mathbf{Q}$ (particularly Q_{12}) and $\mathbf{q}$ can support assignment
215 of individuals to hybrid generations (Figure 2). Use of the admixture complement model
216 will typically be restricted to low levels of K, because interpretation becomes increasingly
217 complex for $K > 2$, requiring multi-dimensional plots for combinations of higher K values
218 in the Q matrix (see Figure 2). In empirical study of systems for which $K = 2$ was well

supported, the *ancestry complement* model has been used to learn about patterns of hybrid matings among *Lycaeides* butterflies (Gompert *et al.*, 2014b; Chaturvedi *et al.*, 2020), and *Catostomus* fish (Mandeville *et al.*, 2017, 2019), and in a related model used to study assortative mating among *Populus* species and their hybrids (Lindtke *et al.*, 2014). The *ancestry complement* model for diploids is not found in `structure`, or other `structure`-like models. As noted previously, the implementation of the *ancestry complement* model only was made in software for diploids, because the number of dimensions required to represent this matrix for higher ploidal levels was unwieldy and difficult to summarize into interpretable statistics.

Locus-specific, local ancestry (z)

The local ancestry parameter (z) is a marker for the population of origin of each allele copy at a given locus (for the **q** model; see below for the *ancestry complement* model) for an individual in a data set. It follows that the ancestry at a locus in an individual is informed by the genome-wide admixture proportions of that individual, reflecting different source populations, with z indicating the appropriate source population. So the prior probability for local ancestry of an individual i at locus j is given by the admixture proportion for that individual, q_i: $P(z_{ija} = k) = q_{ik}$ for allele copy $a \in \{1, 2, \ldots, n\}$ in autopolyploid individuals (since we model each allele copy to be independently derived). This allows for each allele copy at a locus to be derived from a different source population. This number is a single draw from a multinomial distribution conditioned on the admixture proportions in **q**.

The z vector for diploid individuals in the *ancestry complement* model functions similarly, in that we assume the conditional probability for local ancestry to be $P(z_{ij} = kk'|\mathbf{Q}) = Q_{kk'}$. This means that the probability that both allele copies at a locus were inherited from source population k is equal to the proportion of the individual's genome in which both allele copies are inherited from population k, and so on. This allows for the combinations of interspecific ancestry to be modeled explicitly as it considers the possibility of separate ancestry states at a locus.

Allele frequency (p)

The allele frequency in inferred demes is an important parameter that allows sharing of information among loci by quantifying their shared evolutionary divergence from allele frequencies in an idealized ancestral population (parameterized by F and π). The allele frequency in `entropy` for a locus j in population k, p_{jk} is modeled with an F-model prior as in Nicholson *et al.* (2002); Falush *et al.* (2003); Gaggiotti & Foll (2010)

$$P(p_{jk}|\pi_j, F_k) \sim \text{beta}(\pi_j \frac{1-F_k}{F_k}, (1-\pi_j)\frac{1-F_k}{F_k})$$

where π_j denotes the allele frequency at locus j in the hypothetical population that was ancestral to the K source populations. F_k denotes the extent to which the k^{th} source population has diverged from the ancestral population. This is analogous to Wright's F_{ST} under some conditions, and can be thought of as being directly proportional to the amount of genetic divergence between the ancestral and the derived populations. The prior on π_j is beta(α, α) and the prior on F_k is uniform(0,1), where α is inversely proportional to genetic variation in the ancestor and is estimated from the data. This formulation does not change for polyploid populations as is shown in the Implementation section of Nicholson *et al.* (2002).

Since the allele frequency in the ancestral population π is drawn from a beta(α, α), we obtain a symmetric distribution that could take various shapes for different values of α, but the distribution is constrained to a mean ancestral allele frequency of 0.5.

Genotype (g)

In the `entropy` model the genotypes are treated as parameters and are estimated from the element-wise product of the genotype likelihood (the input data) and the prior probability for the genotypes, $\mathbf{GL} \times P(g_{ij}|p_j, z_{ij})$. With contemporary DNA sequencers, genotypes are

not observed directly, but instead information about genotype likelihoods (**GL**) is obtained through bioinformatic steps and a model for the observed sequences. The genotype likelihood is calculated based on the observed sequence data (incorporating read counts, base quality scores, mapping quality scores, etc.) for each of the possible genotypes at a locus. Because these likelihoods are for discrete genotypes, they can be readily rescaled so that they sum to one and can be used as a discrete probability distribution. Often, during the analysis of DNA sequencing data, software is used to call a genotype, for each locus and individual, to be the most likely genotype given the sequence data at the locus (i.e., the mode of the genotype likelihood). The use of genotype likelihoods rather than point estimates of genotype allows uncertainty stemming from sequencing depth and mapping quality to be incorporated into a probability distribution, while maximizing the use of information in sequence data. Genotype likelihoods can be obtained from most variant-calling softwares (e.g., `GATK` McKenna *et al.* 2010, `FreeBayes` Garrison & Marth 2012, or `SAMtools` Li 2011), which can take into account the base and mapping qualities, haplotypic information, along with read counts to estimate a likelihood for the genotype.

The prior probability of each genotype is calculated from the allele frequencies in the corresponding source population, as determined by the ancestry of the allele copy or the ancestry combination of a pair of alleles in the *ancestry complement* model. This assumes genotypes arise from random draws of alleles. The genotype prior probabilities for a n-ploid individual i at locus j is given as

$$P(g_{ij}|\mathbf{p_j}, \mathbf{z_{ij}}) = \prod_{k} \prod_{a=1}^{n} \begin{cases} p_{jk}^{g_{ija}}(1-p_{jk})^{n-g_{ija}} & \text{when } k = z_{ija} \\ 1 & \text{otherwise} \end{cases}$$

Here, $\mathbf{z_{ij}} = [k_1, k_2, \ldots, k_n]$ denotes the local ancestry of the n allele copies for individual i, and z_{ija} denotes the local ancestry of the specific allele copy, a in the individual. The term p_{jk} denotes the corresponding allele frequency in the k^{th} source population. The above

expression yields a discrete posterior probability distribution of length $n+1$ for each genotype (g) in a n-ploid individual ($\{0, 1, \ldots, n\}$) i.e., number of possible alternate alleles at a locus, of the same size as the vector **GL**.

Model initialization and comparison

Given the potentially large number of loci and individuals in a contemporary study, the model will include large numbers of parameters, including loci $\times$ individuals genotypes, loci $\times$ ploidy(n) $\times$ individuals $\times$ populations locus specific ancestries (z), loci $\times$ populations allele frequencies (p), and individuals $\times$ populations admixture proportions (q). Given the large number of parameters and Bayesian MCMC estimation, the efficiency of the estimation (faster convergence in this highly dimensional space) benefits from starting the chains as close to the stationary distributions as possible. Also due to the arbitrary nature of the model's indexing of population or demes, estimation could include label switching among MCMC chains (i.e., the possibility of having ancestry or deme categories have different label indexes across chains, because the arbitrary indexing does not result in a change in the likelihood of the parameters given the data; see Stephens 2000). To speed convergence and avoid label switching, in practice one can initialize values based on a statistical procedure or taxonomic categories. We have used K-means clustering on the output of a linear discriminant analysis of the first five principal components (as specified in Jombart *et al.* 2010) to obtain estimates of the assignment probabilities to the K clusters for all the individuals. This analysis is run on point estimates of the genotypes from the genotype likelihoods. This statistical approach yields a probability of assignment of individuals to demes (the K-means clusters), without admixture. We have used the estimated assignment probabilities as mean initialization values (with some variance) for $\mathbf{q}$ in the `entropy` model and software (e.g., Gompert *et al.*, 2014b; Mandeville *et al.*, 2015; Haselhorst *et al.*, 2019). Additionally, starting values for the admixture proportions could come from taxonomic labels or justified strata in the sampling. The software implementation uses the initial q values to compute the initial population allele

frequency in each of the K populations. This is calculated by finding the number of alleles with ancestry in a certain population (given by the initial $\mathbf{q}$) and then dividing this number by the total number of allele copies in the population. This step initializes the population allele frequencies consistently among chains and limits the possibility of a label switch among chains.

The fit of an `entropy` model for a given K (i.e., the set of parameters) to the observed sequence data can be quantified by using a measure of 'deviance' or the likelihood of the data given the parameters. The `entropy` model provides values of deviance (i.e., the negative log probability of the data given the parameters) and this can be used to calculate the Watanabe-Akaike Information Criterion (Watanabe, 2010). This WAIC value is a combination of the log predictive pointwise density (*lppd*), similar to the model likelihood output by `structure`, with a penalization term for the number of parameters in the model (since models with more parameters fit the data better). Consequently, for the WAIC and the negative log-likelihood, a lower value signifies a better fit. This WAIC value differs from the Deviance Information Criterion (DIC, Spiegelhalter *et al.* 2002) in that the log-likelihood value is averaged across all posterior samples instead of being calculated on a single average value of the posterior samples. Similarly, the effective number of parameters, which is the penalization term, is also computed using the variance of the log-likelihood (i.e. 'deviance') across all samples. This measure is suggested to work well with a hierarchical model in which the parameters increase in number with the dimensions of the data (Gelman *et al.*, 2014), which is the case in our model.

The summary of model fit from WAIC, in combination with graphical analyses of $\mathbf{q}$ estimates, can contribute to an understanding of the number of potential demes (K) involved in admixture, particularly for taxa with contemporary hybridization and in the context of other information about the evolutionary history of the groups. However, this measure of model fit only allows contrasts among models for different choices of the number of demes (K). As such, `structure`-like models cannot themselves provide evidence for demic popula-

tion structure (as noted in Pritchard *et al.* 2000), but instead must rely on complementary analyses and knowledge of the system. If the true population histories differ significantly from the underlying demic model, contrasts of the WAIC for different K can indicate which model best approximates the system. However, all of the demic models could fit poorly if genetic differences among individuals include substantial isolation by distance, additional substructure within the ancestral populations, rather than, or in addition to differences in the actual number of demes (K). Additionally, inference of the number of demes using `structure`-like models (or ordinations) can be misled by uneven sampling of individuals from putative demes, since very uneven sampling can introduce spurious substructure and lead to an underestimation of 'true' number of subpopulations (Puechmaille, 2016). Finally, aside from the difficulty of inferring the number of demes that are consistent with the data, the choice of K does not affect the estimation of genotypes. Instead, genotype estimates can be averaged over the posterior distributions of genotypes across all K runs to obtain a point estimate at a given locus for an individual (e.g., Gompert *et al.*, 2014b).

Model performance

Our measures of model performance build on previous testing of the model and software for diploids (Gompert *et al.*, 2014b) and emphasize tests of the model extensions to data from polyploid and mixed-ploidy samples. As noted above, the diploid portion of the model has been used previously for several empirical analyses (e.g., Gompert *et al.*, 2014b; Mandeville *et al.*, 2015; Chaturvedi *et al.*, 2020). We do not explicitly test the performance of the *ancestry complement* model as it has been done previously in Gompert *et al.* (2014b) and has been used subsequently in several studies (e.g., Mandeville *et al.*, 2017; Chaturvedi *et al.*, 2020), to better distinguish among different classes of early generation hybrids and to distinguish recent from more advanced generation hybrids in diploid individuals (Figure 2).

Simulated data

We used simulations to quantify the performance of the model using three different metrics: accuracy in genotype and ancestry estimates under various simulation parameters (for each 2n, 3n, 4n, 6n, and 2n-4n data set), ability to impute missing data under varying missingness percentages (for each 2n, 3n, 4n, and 6n data set), and accuracy in ancestry estimates for a trade-off between coverage and sequence depth (4n data set).

The genotypic data for 2,000 loci and 100 individuals were simulated using the following evolutionary history. Individuals were assumed to be descended (either completely or partially) from one of $K = 3$ demes. The demes were a result of evolution with drift relative to an ancestral population, with the ancestral allele frequency at each locus drawn from a beta$(0.5, 0.75)$ distribution to simulate the allele frequency spectrum expected in a real population with a skew towards low-frequency alleles. Separate simulations considered different amounts of evolution relative to the ancestral population, using an F-model for derived allele frequencies and $F \in \{0.05, 0.1, 0.2, 0.4\}$ (ranging from low to high evolutionary divergence), and the differentiation this induced among demes. Based on the allele frequencies in demes, genotypes were simulated from a binomial distribution given the individual's ploidy and their local ancestry across different populations. For instance, in a tetraploid individual, the genotype at a locus was drawn from binomial distribution with four draws (number of allele copies) and the success probability being the frequency of the alternate allele, weighted by its proportional ancestry in the source population. An individual could either be a parental, F1, back-cross between an F1 and a parental (BC1), F2, or F3. The genotypes were then converted to genotype likelihoods based on a range of sequence depths (drawn from a Poisson distribution with means: $1\times, 2\times, 4\times, 6\times,$ and $12\times$) following the `GATK HaplotypeCaller` model, assuming a constant sequencing and mapping quality so as to isolate any bias in these numbers on estimating our parameters.

Similarly, to explicitly validate the mixed-ploidy portion of our model, we simulated

genotypic data for a hundred individuals, with fifty tetraploids and fifty diploids. Here, the genotypic data for the loci were drawn from the same evolutionary process as stated above with a change in the binomial sampling to yield the correct number of allele copies. The simulations were run for $F \in \{0.05, 0.1\}$ and for an average sequence depth of $2\times$ for diploid individuals and $4\times$ for tetraploid individuals. As input, we also provided the ploidy of each individual along with the genotype likelihoods to `entropy`. The goal here was to primarily test the ability of our software to handle mixed-ploidy input and secondarily, to test the minimal ability of our model to recover the simulated parameters. From these simulations, model performance was quantified by calculating the accuracy in estimation of genotype and ancestry across all individuals and loci.

One measure of model performance would be the extent to which the model could correctly impute missing (i.e., left-out) data from a simulation. To quantify the ability of the model to impute left-out data, subsets of the genotypic data from above were randomly excluded from a complete data set to achieve varying proportions of missingness $(10\%, 20\%, 30\%, 40\%)$ over loci and individuals. This metric was important to test the performance of the model, not only for assessing the accuracy in imputing missing values, but also to mimic real empirical sequencing in which regions of the genome are not sampled at all for a number of individuals. This form of testing is akin to conducting a posterior predictive check in a Bayesian modeling framework (first introduced in Rubin, 1984) by quantifying the ability of our model to recover simulated parameters, especially from held-out (in our case, missing) data. Secondarily, this test of performance gives an indication of how data missingness would affect inferences of genotype and admixture proportions with empirical data, and we considered a range of missingness that one might encounter in empirical studies. The missing data in our simulations refers to a case when there is no sequence information (i.e., no reads) at a certain locus in an individual (for instance in a `vcf` file for diploids, `./.` would indicate that we have no information to make a genotype call at that locus, equivalent to having missing data). For instance, in the simulated data set where we have an average

10% missingness in the data, we only have sequence data or genotype likelihood information for 1,800 of our total 2,000 loci. The rest of the genotypes have no likelihood information, and the genotype will be directly estimated from the population prior that has been built up in the hierarchy. To simulate this missing data, we randomly selected loci to have equally probable genotype likelihoods (i.e., every dosage/genotype is equally likely given no other information) to mimic the absence of sequence information at this locus. This setup is similar to encountering a 0× read depth from a Poisson distribution, however more useful, since this framework allows us to systematically test the idea of having a fixed proportion of missing data, instead of it being an artefact of the sampling process for which we have no control on the proportion of missing data that could be obtained.

To test the hypothesis that it is better to estimate average genome-wide ancestry by capturing more of the genome (i.e., loci, via greater genome coverage, defined here to mean extent of the genome covered by the sequence data) at a lower sequencing depth than it is to sequence a smaller region of the genome (i.e., lower coverage) at a higher depth (i.e., more reads), we ran a simulation for 100 tetraploid individuals across three pairs of values for coverage and corresponding sequence depth. The assumed evolutionary process was the same as the one used before, in which we simulated the tetraploid individuals as being descended from one of 3 possible demes (differentiated by $F = 0.05$) sequenced at: 4× and 1,000 loci ('low' coverage), 2× and 2,000 loci ('medium' coverage), and 1× and 4,000 loci ('high' coverage). Our testing here focuses on the "middle" tetraploid case, since we expect a similar mechanism to be operating at lower and higher ploidal levels (Buerkle & Gompert, 2013).

In total, 1,210 simulations were run to quantify accuracy in estimation, ability to impute missing data, and a trade-off between coverage and sequence depth across different levels of ploidy (2n, 3n, 4n, 6n, and 2n-4n), range of missingness percentages, varying levels of sequence depth, and admixture from three ancestral populations (at varying levels of evolutionary divergence). For simulations that contained missing data, we used the correlation

441 metric between the point estimate of our parameter (the average of posterior distributions
442 across chains) and the simulated truth to measure how well we could recapture this simu-
443 lated parameter given a certain percentage of missing data. For the rest of the simulations,
444 we calculated the root mean squared error (RMSE) between the inferred values for the geno-
445 type and admixture proportions and the known true values that were simulated, as a way
446 of measuring our ability to recover the truth.

447 Empirical data

448 As a further test of the performance of the model and software, we reanalyzed an empir-
449 ical mixed-ploidy data set that includes DNA sequences of individuals from diploid and
450 autotetraploid populations of *Arabidopsis arenosa* across Europe (Monnahan *et al.*, 2019).
451 We compared estimates of admixture proportions from this mixed-ploidy sample from both
452 `entropy` and `structure` softwares. We used, as input, sequence data obtained from the `vcf`
453 files for eight scaffolds, which were shared by the authors of Monnahan *et al.* (2019). From
454 these, we sampled single variable loci randomly in 50,000 base pair windows (within each
455 scaffold) to retain loci that were more likely to vary independently due to recombination
456 and independent evolution. This left us with a set of 5655 loci across 287 individuals (105
457 diploids in 15 populations and 182 tetraploids in 24 populations) with 22.4% missing data.
458 The previous analysis in Monnahan *et al.* (2019) used 9543 loci across 287 individuals with
459 2.4% missing data. The different number of loci and missingness for the two analyses is
460 because we randomly thinned variants over windows of 50 kb to reduce the effect of linkage.
461 So, the percentage of missing data (no call in `vcf` file) in our thinned set of loci reflected
462 the average among variants in the original `vcf` file. We note that this process of random
463 thinning only affects the credible intervals, and not the point estimates of the admixture
464 proportions from the model. By including more loci in our analysis, we would get a more
465 accurate point estimate for our genome-wide admixture proportion with a tighter credible
466 interval. However, we also note that there is a diminishing return to including more loci in

the analysis if the goal is to simply obtain point estimates for admixture proportion. The input to `structure` (version 2.3.4) was a file with the called values of genotypes (GT field in the `vcf` file) of selected loci and individuals, called using `GATK HaplotypeCaller` (version 3.5, McKenna *et al.*, 2010). Given that the maximum ploidy included four allele copies, the loci for the diploid individuals were encoded with four allele copies and the extra two allele copies as missing data (since the `structure` manual indicates that all individuals in the sample should have a single ploidal level, Meirmans *et al.*, 2018). The input to `entropy` was a file with the genotype likelihoods (PL field in the `vcf` file) of the selected loci, rather that the point estimates of the genotypes. The `entropy` model was initialized using the discriminant function method described previously, to reduce the chance of label switching among chains and speed MCMC convergence. We compared admixture proportion estimates from `structure` and `entropy` primarily for $K = 6$, which was regarded by Monnahan *et al.* (2019) as the most likely model given other knowledge of the evolutionary history of *A. arenosa*.

The `structure` admixture model was run three times for 600,000 iterations and 100,000 burn-in, which took approximately 102 hours each. This number of iterations was chosen based on multiple runs of different lengths and picking the shortest run that arrived at approximately the same estimates as the longer runs. Since `structure` stores every sample after burn-in as a draw from the posterior distribution, the admixture proportions were estimated based on 500,000 draws. On the other hand, the `entropy` model was run with three chains simultaneously for 30,000 total iterations with 10,000 burn-in each, which took approximately 24 hours in total. The number of steps was chosen based on the convergence of previous data sets of similar size. The quicker convergence times were likely a result of starting our chains with plausible initial admixture proportions (as mentioned in the Model initialization and comparison section). Researchers could typically use `fastStructure` (Raj *et al.*, 2014) in this case, but this software only allows for diploid samples, which requires a downsampling at each tetraploid locus to fit the requirements for the input data (Monnahan *et al.*, 2019). The samples collected were thinned to retain every 10^{th} step to remove

autocorrelation within the chain. We were finally left with 6,000 (2,000 $\times$3) samples from the posterior distribution for admixture proportion. The chains were tested for convergence by looking at the trace plots (to check for sufficient exploration of parameter space) and the average $\hat{R}$ statistic ($\approx$ 1.01) across parameters (Gelman & Rubin, 1992). To validate the number of clusters statistically, we ran the entire *A. arenosa* data set for values of K ranging from 4 through 10 to obtain WAIC estimates, and compared them to estimates from Monnahan *et al.* (2019).

Results

Simulated data

We present the effect of sequence depth, F, number of ancestral demes, and ploidy on the ability of our model to accurately predict estimates of genotype and ancestry from simulated data (Figures 3 and 4). Based on the different axes of variation in our simulation parameters, we found that sequence depth had the strongest effect on our ability to estimate both genotypes and admixture proportions accurately, followed by the degree of differentiation F between our simulated demes. From our simulations containing missing data, as expected, the model performed better at recapturing the missing genotypes when we had lower percentages of missing data (Figures S2 and S3). Holding all else constant, we also found that we did better at accurately estimating admixture proportions of tetraploid individuals when we had higher coverage (number of loci across the genome) over higher sequencing depth (read depth at a locus; Figure 5).

With regard to the estimation of genotypes, we better distinguished discrete genotype classes at higher sequence depth and higher F values. The reason we obtained larger values of RMSE for higher ploidy is a consequence of a wider range of genotypes that are possible. So as a consequence of the RMSE statistic, we are bound to get higher error values for higher

number of genotype classes (i.e., higher ploidy) in our data. But, based on the approximately constant correlation between simulated and estimated genotypes in our missing data sets, we show that higher ploidal levels do not translate to a higher error rate in estimation (Figure S2). However, the spread around the RMSE across each ploidal level suggests that the degree of differentiation (F) and number of ancestral demes did not play a major role in our accuracy of prediction, as shown by the inset plot in Figure 3 and Figure S5.

Similarly, in the estimation of admixture proportions we found that the sequence depth had the biggest effect on our ability to recover the truth, followed by the degree of differentiation between the simulated demes (Figures 4(a), S1 and S4). With a sequence depth of $1\times$, we only observed one allele in a tetraploid and, as a result, our genome-wide ancestry estimates were solely guided by a single allele at each locus. However, the model performed much better once we observed, on average, two alleles ($2\times$) at a given site and hit diminishing returns in sequencing beyond $4\times$ (as seen in the estimates from Figure 4). Based on a comparison study between the teraploid and hexaploid data sets, we found that we did better at recovering the true number of clusters K from the simulated data with a higher ploidy level (Tables S3 and S4). However, the differences in WAIC values between the two tables indicate the erratic nature of using information criteria in making a choice about the K value in empirical studies. For the mixed-ploidy portion of the simulations, we found that we did equally well in recovering genotype estimates as the fully diploid and fully tetraploid simulations, regardless of F and the number of ancestral demes (Table S1). However, we found that WAIC recovered the correct number of simulated clusters K given the simulated mixed-ploidy data set in most cases, except for when $F = 0.05$ (which is equivalent to a highly differentiated cluster in a $K = 1$ model).

For the simulations involving missing genotype likelihood data, we found that the correlation (r) between the estimated genotypes at the missing sites and the simulated truth was between 0.76 and 0.88 across ploidal levels, indicating an increasing correlation with a decrease in missing percentage (Figure S6). This correlation translates to the fact that, on

average across all simulation parameters, we could predict approximately 70% of our missing genotypes accurately (coefficient of determination, expressed as a percentage, is equal to r^2). We also found that we have a higher correlation of estimated and true genotypes for higher ploidy levels, given the same missingness percentage and degree of differentiation (Figure S2). We believe the higher correlation between estimated and true genotypes for higher ploidal levels is due to the ability of our model to better recapture the number of true clusters K from a higher ploidy data set in our simulations (as shown with WAIC values in Tables S3 and S4). Similarly, for admixture proportion estimation, we found a correlation between 0.96 and 1 for 296 out of 320 simulations. The set of outlier simulations were for low $F = 0.05$ and a high missingness (40%) value across all ploidal levels, for which the correlation was only 0.83 (Figure S3). For more than 80 percent of simulations, we predicted approximately 95% of our missing admixture proportions accurately (Figure S7).

Based on the RMSE metric, we found that the model estimated genome-wide ancestry for tetraploid individuals more accurately with higher coverage across the genome and a lower sequencing depth (4,000 loci at 1×) than with lower coverage and a higher sequencing depth (1,000 loci at 4×) (Figure 5).

Empirical data

Overall, the admixture estimates from `entropy` closely match the estimates from `structure` (Figure 6). Similarly, the `entropy` and `structure` admixture estimates largely match those presented in Monnahan *et al.* (2019), which used a combination of analyses from `fastStructure` (Raj *et al.*, 2014) and a non-parametric K-means clustering technique (with a confirmatory analysis in `structure` for $K = 6$). Below we note some of the differences that were found between the ancestry estimates from `entropy` and `structure` (and shown in Figure 6).

Firstly, the admixture proportion estimates for the diploid individuals in populations

from the Pannonian region were calculated to be different by `entropy` and `structure`. The `entropy` model assigned these individuals to a separate cluster, but the `structure` model found these same individuals to be genetically intermediate between the Dinaric (orange) and E. Alps (yellow) regional ancestries. Based on the evolutionary history of the plant, the Pannonian populations are the most divergent and should separate out as their own cluster (as shown in Figure 1 of Monnahan *et al.*, 2019). This hypothesis was further supported when both `entropy` and `structure` placed the Pannonian population into a distinct cluster when run for a $K = 5$ model on only the diploid individuals and a $K = 7$ model on all the *A. arenosa* individuals (as shown in Figures S8 and S10 for `entropy` and Figure S5 of Monnahan *et al.* 2019 for `structure`), as expected when running the analysis with a higher K in `structure`-like models. Secondly, the tetraploid individuals in the S. Carpathians are estimated to share some ancestry (between $\sim$ 20% and 50%) with their W. Carpathian (green) counterparts in `entropy` but this was not found to be the case with the estimates from `structure`. This shared, intermediate or hybrid ancestry in the S. Carpathian populations is to be expected from the single origin of tetraploidy in the populations of the Carpathian mountain range, as confirmed through coalescent simulations of this mixed-ploidy hybrid zone (as presented in Figure 4 and Figure S9 of Monnahan *et al.*, 2019).

From our range of runs for $K = 4$ to 10, we found the lowest WAIC value for $K = 9$, as opposed to $K = 6$ that was found by Monnahan *et al.* (2019). The authors of the original study used a combination of the Bayesian Information Criterion (BIC, Schwarz *et al.* 1978), and the similarity index proposed by Nordborg *et al.* (2005) to inform their choice of K. However, the range of BIC values for $K = 5$ to 10 were found to be within five points of each other, indicating similar support for the given cases of K, highlighting the potential challenge of choosing K in empirical studies.

Discussion

In the context of recent hybridization or admixture among divergent lineages, estimates of admixture proportions are a fundamental component of analyses of evolutionary processes or of learning the effects of population stratification in genome-wide association studies (Gompert & Buerkle, 2013; Harrison & Larson, 2016; Gompert *et al.*, 2017). Hybrids commonly occur between taxa that have sex chromosomes (and mixed-ploidy within genomes of the heterogametic sex) and sex chromosomes may contribute disproportionately to their reproductive isolation (Payseur *et al.*, 2004; Sæther *et al.*, 2007; Presgraves, 2008; Macholán *et al.*, 2011; Chaturvedi *et al.*, 2020). Additionally, many species complexes involve interactions and potential hybridization between individuals of different ploidy, including autopolyploids (e.g., Otto & Whitton, 2000; Kolář *et al.*, 2017; Van de Peer *et al.*, 2017). Population genetic analyses of polyploids would benefit from models that correctly specify the number of allele copies at a locus, rather than misspecified models that do not fully use the available data (e.g., encoding diploids as tetraploids with missing data so that `structure` can be used to analyze mixed ploidy individuals). Additionally, given genotype uncertainty in contemporary low-depth sequencing data from populations, we make better use of the data with models that formally incorporate uncertainty through the use of genotype likelihoods as input (Buerkle & Gompert, 2013; Fumagalli *et al.*, 2013). Here, we address these needs and present additional benefits, with improvements in running time, and ability to assess convergence of chains using appropriate metrics, in the form of a population model for allele frequencies and admixture of individuals that follows the precedent of the `structure` model (Pritchard *et al.*, 2000; Falush *et al.*, 2003) and its several derivatives. We present and analyze the performance of `entropy`, a hierarchical Bayesian model that can use genotype likelihoods to estimate genotype and ancestry for polyploid and mixed-ploidy individuals.

We found that the `entropy` model performed well to capture the truth from simulated mixed-ploidy data sets. We used estimates from the model for simulated autopolyploid data

and quantified similarity of our estimates to the known values using RMSE and correlation statistics. The `entropy` software implements a population model and information sharing among individuals and loci that provides stronger evidence for low-depth, or missing genotypes (especially with polyploids and low-depth sequencing) than methods that do not model populations (consistent with Clark *et al.*, 2019). With the extension of the model and software to mixed-ploidy, we can also model haploid loci or hemizygous regions of the genome. Thus, given knowledge of the genomic position of loci, the model will support contrasts of ancestry between sex chromosomes and autosomes (e.g., Hamilton *et al.*, 2013; Parchman *et al.*, 2013; Harrison & Larson, 2016). For the analysis of diploid hybrids, as shown previously in Gompert *et al.* (2014b), the *ancestry complement* model considers the combination of ancestry in diploid genotypes and allows genotypic data to more readily distinguish among different classes of early generation hybrids (Figure 2) and to distinguish recent from more advanced generation hybrids (e.g., Gompert *et al.*, 2014b; Mandeville *et al.*, 2017; Chaturvedi *et al.*, 2020).

In our simulations, we found that sequence depth had the largest effect on accurately estimating the genotype and ancestry of an individual, similar to findings in Gerard *et al.* (2018). The degree of differentiation among demes (driven by F divergence from the ancestral population) had the second largest effect on the accuracy of our estimates. For admixture proportion q, we found no difference in our ability to estimate ancestry across the different ancestry classes (F1, F2, BC, etc.), but do markedly better with increasing sequence depth, as seen in Figure 4. Across our simulations, we also found that the percent of missingness did not affect how well we could estimate true parameters. For example, when going from 40% to 10% missingness in sequence data there was only a 2% gain in accuracy of prediction for tetraploid genotypes. In summary, from the different combinations of the simulation parameters, it was the hardest to recover parameters accurately when we had low sequence depth (for higher ploidy) and minimal differentiation between populations ($F < 0.05$), as was expected. Based on the ability to recover the truth in various simulations, for analyses

of admixture proportions with this model we recommend choosing a median sequence depth of $\frac{n}{2}\times$ (i.e., $2\times$ for tetraploids) and sampling more individuals and populations rather than sequencing deeply (consistent with findings from our simulations in Figures 4(a) and (c) and Buerkle & Gompert, 2013; Fumagalli *et al.*, 2013). The structure of the hierarchical model is such that enough information is shared across loci to accurately estimate admixture proportions even without full information about genotypes. However, if the goal of an analysis is highly accurate genotype estimates, as expected, sequencing to $6\times$ or greater depth might be warranted (Figure S4).

Our direct comparison of `entropy` estimates to estimates from `structure` for an empirical mixed-ploidy data set (diploid and tetraploid) of *Arabidopsis arenosa* (Monnahan *et al.*, 2019) validated the software implementation of the model and revealed some differences of admixture proportions and inferred ancestry for a few populations. The data used with `entropy` and `structure` contained fewer loci (5655 loci versus 9543 loci) and a higher percentage of missing data that is typical in a RADseq or similar dataset (22.4% versus 2.4%), compared to the original analysis in Monnahan *et al.* (2019). Nevertheless, the two models were still able to capture the previously inferred population structure. The estimates from the `entropy` model for a cluster of admixed tetraploid individuals indicated a portion of their ancestry belonged to a previously undetected neighboring cluster, a finding that was supported by coalescent simulations in Monnahan *et al.* (2019), but was not captured by the `structure` model. Additionally, the `entropy` model distinguished and assigned some exceptional individuals to a distinct cluster instead of classifying them as belonging to an admixed group, as done by `structure`.

Limitations and further directions

Whereas the model specified in `entropy` will be useful in many contexts, we recognize some of its limitations, including ones that pertain generally to inferring ancestry using a `structure`-like model and other forms of model misspecification. Population genetic variation among

673 natural populations arises due to clinal isolation by distance, more abrupt barriers to dis-
674 persal that result in actual 'demic' substructure within species, or some combination of both
675 (Bradburd *et al.*, 2013; Gompert & Buerkle, 2016). Analyses with `structure`-like models do
676 not incorporate clinal isolation by distance variation. Thus, inferences regarding the number
677 of demes (K) or admixture proportions could be based on a false inference of subdivision
678 and be very misleading (Lawson *et al.*, 2018; Garcia-Erill & Albrechtsen, 2019). In particu-
679 lar, discrete geographic sampling of widely spaced populations along a spatial gradient can
680 give the misimpression of discrete population differences (Witherspoon *et al.*, 2006; Gom-
681 pert & Buerkle, 2016). The model in `entropy` and other `structure`-like models do not
682 address these directly. Linear models for population genetic differences can test for evidence
683 of demic structure beyond what could be predicted from geographic distance alone (e.g.,
684 Gompert *et al.*, 2014a; Parchman *et al.*, 2016; Crow *et al.*, 2020). Additionally, alterna-
685 tive models can explicitly parameterize continuous clinal variation and guide understanding
686 of the contribution of demic and clinal variation to population structure (Bradburd *et al.*,
687 2013, 2016; Battey *et al.*, 2020). When feasible, structured and planned geographic sampling
688 can assist in quantifying the contributions of isolation by distance and demes to population
689 variation. Deviation from the assumed evolutionary model (model misspecification) can be
690 quantified through the correlated differences in a population between predicted and observed
691 genotypes (i.e., correlated residual error), which can guide model choice and interpretation
692 (Garcia-Erill & Albrechtsen, 2019). This limitation of inferring population structure along
693 a cline may be reduced for mixed-ploidy systems, where we expect some level of genetic
694 differentiation across ploidal levels (even with cross-ploidy gene flow), and `structure`-like
695 models would likely correctly partition individuals of different ploidal levels into different
696 demes.

697 The q model in `entropy` also does not formally include deviations from Hardy-Weinberg
698 equilibrium due to inbreeding (or due to potential double reduction in autopolyploids, see
699 Luo *et al.* 2006 and Bourke *et al.* 2015) and the resulting excess homozygosity of individuals

(F_{IS}) in the prior probabilities for genotype. With sufficient sequencing depth, genotype estimates will strongly reflect the data rather than the prior probabilities and inbreeding (F_{IS}) could be estimated in a separate model. Alternatively, the `entropy` model could readily be extended to formally model excess homozygosity.

Even though the `entropy` model does not explicitly account for allopolyploids, we note that the model can still provide estimates of admixture proportion for loci within separate subgenomes or chromosomes in a higher ploidy individual by treating them as coming from a lower ploidal level. For instance, in allotetraploid individuals with disomic inheritance, we can run a diploid `entropy` analysis on a set of loci coming from one pair of homoeologous chromosomes and a similar analysis on the set of loci coming from the other pair of homoeologous chromosomes, and compare the admixture estimates of the individuals from these two separate analyses as independent realizations of their shared evolutionary history. The pipeline presented in Blischak *et al.* (2017) can be used to obtain `vcf` files with appropriate genotype likelihoods for SNPs in allopolyploid individuals that can then be used as input to `entropy`, by specifying the appropriate ploidal level. The model should probably not be applied to polyploids for which the mode of inheritance is not known, given the potential for spurious clustering due to model misspecification.

The genetic composition of individuals could be the result of a combination of ancient and more recent (i.e., contemporary) hybridization (Gompert *et al.*, 2017; Chaturvedi *et al.*, 2020). Analysis of recent hybridization can benefit from the study of population structure through ancestry-estimation methods such as `entropy`. However, recent hybridization can obfuscate signals of more ancient gene flow (Eriksson & Manica, 2012). Regardless of the extent of contemporary hybridization, alternative models are beneficial to evaluate evidence for more ancient introgression (e.g., Sankararaman *et al.*, 2014; Gompert, 2016; Schumer *et al.*, 2016).

The software for the model is written in C++ using the GNU Scientific Library (Galassi *et al.*, 2009) and the output being written to a Hierarchical Data Format (The HDF5 Group,

2010) file. However, even though the program is written in a low-level language with optimized libraries, given the large size of the estimation problem with typical data sets, the process of converging to a stationary distribution using the Gibbs and Metropolis sampling scheme for MCMC can be time intensive. In future versions of the software, this runtime could be shortened by using techniques like variational inference (as in Raj *et al.*, 2014; Gopalan *et al.*, 2016) and non-negative matrix factorization (as in Engelhardt & Stephens, 2010; Meisner & Albrechtsen, 2018) to arrive at the posterior parameter estimates without using MCMC sampling. However, dealing with the heterogeneity in parameter dimensions that comes with a mixed-ploidy data set will be an algorithmic challenge. For now, in practice we reduce the dimensions of a model run by treating different chromosomes (or other large genome scaffolds) as independent sampling units. This allows one to run separate, parallel analyses of loci on different chromosomes (or scaffolds) that can be distributed across multiple computing cores or nodes.

With these limitations and potential extensions in mind, we find that the `entropy` model can contribute to our understanding of contemporary hybridization and population structure. In particular, the `entropy` model provides a rigorous and beneficial framework for genotype and ancestry estimation from economical, low-depth sequencing data. The model also supports analysis of a wide range of ploidy (from haploid to hexaploid) and mixed-ploidy individuals within a single analysis, which will facilitate a diversity of studies.

Data Accessibility

All simulation and analysis code is available as part of the Bitbucket repository that hosts the source code. The program can be installed via the bioconda channel (`https://anaconda.org/bioconda/popgen-entropy`) or from source by cloning the Bitbucket repository (`https://bitbucket.org/buerklelab/mixedploidy-entropy/`), which also houses the on-going developmental code base. A software vignette is part of the Supplementary Material and

is also found in the Bitbucket repository. Raw sequence data for *Arabidopsis arenosa* are available at `https://www.ncbi.nlm.nih.gov/bioproject/484107`.

Author Contributions

CAB, ZG, EM, DL and TP wrote the diploid model specification and developed the initial software for diploids. The software was tested and improved by DL, PA, EM, TP, and VS. VS extended the model and software to incorporate variable and mixed ploidy, and performed all analyses. VS and CAB wrote the manuscript with input from the co-authors.

Acknowledgments

We thank colleagues who have contributed to development of the model and its use in various empirical contexts (incl. C. Nice, J. Fordyce, M. Forister, M. Haselhorst, and S. Lebeis). We thank C. Wagner and K. Hufford for helpful comments on drafts of this manuscript. We also wish to thank F. Kolář for sharing the mixed-ploidy *A. arenosa* data from Monnahan *et al.* (2019), and for helpful discussion regarding our reanalysis of their data. This interaction was initiated at the ForBio course "Population Genetics in Polyploids" in 2018, which VS attended with financial support from an NIH INBRE grant to the University of Wyoming. This work was funded in part by the National Science Foundation (DEB-1638602 to CAB). Computing was performed in the Teton Computing Environment at the Advanced Research Computing Center (University of Wyoming, `https://doi.org/10.15786/M2FY47`).

References

Abbott R, Albach D, Ansell S, *et al.* (2013) Hybridization and speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

773 Balding DJ, Nichols RA (1995) A method for quantifying differentiation between populations
774 at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*,
775 **96**, 3–12.

776 Battey C, Ralph PL, Kern AD (2020) Space is the place: Effects of continuous spatial
777 structure on analysis of population genetic data. *Genetics*, **215**, 193–214.

778 Blischak PD, Kubatko LS, Wolfe AD (2017) SNP genotyping and parameter estimation in
779 polyploids using low-coverage sequencing data. *Bioinformatics*, **34**, 407–415.

780 Bourke PM, Voorrips RE, Visser RG, Maliepaard C (2015) The double-reduction landscape
781 in tetraploid potato as revealed by a high-density linkage map. *Genetics*, **201**, 853–863.

782 Bourke PM, Voorrips RE, Visser RG, Maliepaard C (2018) Tools for genetic studies in
783 experimental populations of polyploids. *Frontiers in plant science*, **9**, 513.

784 Bradburd GS, Ralph PL, Coop GM (2013) Disentangling the effects of geographic and eco-
785 logical isolation on genetic differentiation. *Evolution*, **67**, 3258–3273.

786 Bradburd GS, Ralph PL, Coop GM (2016) A spatial framework for understanding population
787 structure and admixture. *PLoS genetics*, **12**.

788 Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how
789 low should we go? *Molecular Ecology*, **22**, 3028–3035.

790 Chaturvedi S, Lucas LK, Buerkle CA, *et al.* (2020) Recent hybrids recapitulate ancient
791 hybrid outcomes. *Nature Communications*, **11**, 1–15.

792 Clark LV, Lipka AE, Sacks EJ (2019) polyRAD: Genotype calling with uncertainty from
793 sequencing data in polyploids and diploids. *G3: Genes, Genomes, Genetics*, **9**, 663–673.

794 Crow TM, Runcie DE, Hufford K (2020) Implications of genetic heterogeneity for plant
795 translocation during ecological restoration. *bioRxiv*.

796 Endler JA (1977) *Geographic Variation, Speciation, and Clines*. Princeton University Press,
797 Princeton, NJ.

798 Engelhardt BE, Stephens M (2010) Analysis of population structure: A unifying framework
799 and novel methods based on sparse factor analysis. *PLoS genetics*, **6**, e1001117.

800 Eriksson A, Manica A (2012) Effect of ancient population structure on the degree of poly-
801 morphism shared between modern human populations and ancient hominins. *PNAS*, **109**,
802 13956–13960.

803 Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilo-
804 cus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

805 Ferretti L, Ribeca P, Ramos-Onsins SE (2018) The site frequency/dosage spectrum of au-
806 topolyploid populations. *Frontiers in genetics*, **9**, 480.

807 Fumagalli M, Vieira FG, Korneliussen TS, *et al.* (2013) Quantifying population genetic
808 differentiation from Next-Generation Sequencing data. *Genetics*, **195**, 979–992.

809 Gaggiotti OE, Foll M (2010) Quantifying population structure using the F-model. *Molecular*
810 *Ecology Resources*, **10**, 821–830.

811 Galassi M, Davies J, Theiler J, *et al.* (2009) *GNU Scientific Library: Reference Manual*.
812 Network Theory Ltd.

813 Garcia-Erill G, Albrechtsen A (2019) Evaluation of model fit of inferred admixture propor-
814 tions. *bioRxiv*.

815 Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing.
816 *arXiv preprint arXiv:1207.3907*.

817 Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for
818 Bayesian models. *Statistics and Computing*, **24**, 997–1016.

819 Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences.
820 *Statistical Science*, **7**, 457–511.

821 Gerard D, Ferrão LFV, Garcia AAF, Stephens M (2018) Genotyping polyploids from messy
822 sequencing data. *Genetics*, **210**, 789–807.

823 Gompert Z (2016) A continuous correlated beta process model for genetic ancestry in ad-
824 mixed populations. *PLoS One*, **11**, e0151047.

825 Gompert Z, Buerkle CA (2011a) Bayesian estimation of genomic clines. *Molecular Ecology*,
826 **20**, 2111–2127.

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

829 Gompert Z, Buerkle CA (2013) Analyses of genetic ancestry enable key insights for molecular
830 ecology. *Molecular Ecology*, **22**, 5278–5294.

831 Gompert Z, Buerkle CA (2016) What, if anything, are hybrids: enduring truths and chal-
832 lenges associated with population structure and gene flow. *Evolutionary Applications*, **9**,
833 909–923.

834 Gompert Z, Comeault AA, Farkas TE, *et al.* (2014a) Experimental evidence for ecological
835 selection on genome variation in the wild. *Ecology Letters*, **17**, 369–379.

836 Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, Nice CC (2014b) Admixture
837 and the organization of genetic diversity in a butterfly species complex revealed through
838 common and rare genetic variants. *Molecular Ecology*, **23**, 4555–4573.

839 Gompert Z, Mandeville EG, Buerkle CA (2017) Analysis of population genomic data from
840 hybrid zones. *Annual Review of Ecology, Evolution, and Systematics*, **48**, 207–229.

841 Gopalan P, Hao W, Blei DM, Storey JD (2016) Scaling probabilistic models of genetic
842 variation to millions of humans. *Nature genetics*, **48**, 1587.

843 Grandke F, Singh P, Heuven HC, De Haan JR, Metzler D (2016) Advantages of continuous
844 genotype values over genotype classes for GWAS in higher polyploids: a comparative study
845 in hexaploid chrysanthemum. *BMC genomics*, **17**, 672.

846 Hamilton JA, Lexer C, Aitken SN (2013) Genomic and phenotypic architecture of a spruce
847 hybrid zone (*Picea sitchensis* x *P. glauca*). *Molecular Ecology*, **22**, 827–841.

848 Harrison RG, Larson EL (2016) Heterogeneous genome divergence, differential introgression,
849 and the origin and structure of hybrid zones. *Molecular Ecology*, **25**, 2454–2466.

850 Haselhorst MSH, Parchman TL, Buerkle CA (2019) Genetic evidence for species cohesion,
851 substructure and hybrids in spruce. *Molecular Ecology*, **28**, 2029–2045.

852 Haworth S, Mitchell R, Corbin L, *et al.* (2019) Apparent latent structure within the UK
853 Biobank sample has implications for epidemiological analysis. *Nature Communications*,
854 **10**, 1–9.

855 Hellwege JN, Keaton JM, Giri A, Gao X, Velez Edwards DR, Edwards TL (2017) Population
856 stratification in genetic association studies. *Current protocols in human genetics*, **95**, 1–22.

857 Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a
858 new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

859 Kolář F, Čertner M, Suda J, Schönswetter P, Husband BC (2017) Mixed-ploidy species:
860 progress and opportunities in polyploid research. *Trends in Plant Science*, **22**, 1041–1055.

861 Lawson DJ, van Dorp L, Falush D (2018) A tutorial on how not to over-interpret STRUC-
862 TURE and ADMIXTURE bar plots. *Nature Communications*, **9**, 3258.

863 Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using
864 dense haplotype data. *PLoS Genetics*, **8**, e1002453.

865 Li H (2011) Improving SNP discovery by base alignment quality. *Bioinformatics*, **27**, 1157–
866 1158.

867 Lindtke D, Gompert Z, Lexer C, Buerkle CA (2014) Unexpected ancestry of *Populus*
868 seedlings from a hybrid zone implies a large role for postzygotic selection in the main-
869 tenance of species. *Molecular Ecology*, **23**, 4316–4330.

870 Luo Z, Zhang Z, Zhang R, *et al.* (2006) Modeling population genetic data in autotetraploid
871 species. *Genetics*, **172**, 639–646.

872 Macholán M, Baird SJE, Dufková P, Munclinger P, Bimová BV, Piálek J (2011) Assessing
873 multilocus introgression patterns: a case study on the mouse X chromosome in Central
874 Europe. *Evolution*, **65**, 1428–1446.

875 Mandeville EG, Parchman TL, McDonald DB, Buerkle CA (2015) Highly variable reproduc-
876 tive isolation among pairs of *Catostomus* species. *Molecular Ecology*, **24**, 1856–1872.

877 Mandeville EG, Parchman TL, Thompson KG, *et al.* (2017) Inconsistent reproductive isola-
878 tion revealed by interactions between catostomus fish species. *Evolution letters*, **1**, 255–268.

879 Mandeville EG, Walters AW, Nordberg BJ, Higgins KH, Burckhardt JC, Wagner CE (2019)
880 Variable hybridization outcomes in trout are predicted by historical fish stocking and
881 environmental context. *Molecular Ecology*, **28**, 3738–3755.

882 McKenna A, Hanna M, Banks E, *et al.* (2010) The Genome Analysis Toolkit: A MapReduce
883 framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**,
884 1297–1303.

885 Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O (2017) Ancient hy-
886 bridization fuels rapid cichlid fish adaptive radiations. *Nature Communications*, **8**, 14363.

887 Meirmans PG, Liu S, van Tienderen PH (2018) The analysis of polyploid genetic data.
888 *Journal of Heredity*, **109**, 283–296.

889 Meisner J, Albrechtsen A (2018) Inferring population structure and admixture proportions
890 in low-depth ngs data. *Genetics*, **210**, 719–731.

Monnahan P, Kolář F, Baduel P, *et al.* (2019) Pervasive population genomic consequences of genome duplication in *Arabidopsis arenosa*. *Nature ecology & evolution*, **3**, 457.

Nadeau NJ, Whibley A, Jones RT, *et al.* (2012) Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 343–353.

Nicholson G, Smith AV, Jonsson F, Gustafsson O, Stefansson K, Donnelly P (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society Series B-Methodological*, **64**, 695–715.

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

Nordborg M, Hu TT, Ishino Y, *et al.* (2005) The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biology*, **3**.

Novembre J, Johnson T, Bryc K, *et al.* (2008) Genes mirror geography within Europe. *Nature*, **456**, 98–101.

Ottenburghs J, van Hooft P, van Wieren SE, Ydenberg RC, Prins HH (2016) Birds in a bush: Toward an avian phylogenetic network. *The Auk: Ornithological Advances*, **133**, 577–582.

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

Parchman TL, Buerkle CA, Soria-Carrasco V, Benkman CW (2016) Genome divergence and diversification within a geographic mosaic of coevolution. *Molecular Ecology*, **25**, 5705–5718.

Parchman TL, Gompert Z, Braun MJ, *et al.* (2013) The genomic consequences of adaptive

914 divergence and reproductive isolation between species of manakins. *Molecular Ecology*, **22**,
915 3304–3317.

916 Payseur BA, Krenz JG, Nachman MW (2004) Differential patterns of introgression across
917 the X chromosome in a hybrid zone between two species of house mice. *Evolution*, **58**,
918 2064–2078.

919 Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy.
920 *Nature Reviews Genetics*, **18**, 411.

921 Phifer-Rixey M, Bi K, Ferris KG, *et al.* (2018) The genomic basis of environmental adaptation
922 in house mice. *PLoS Genetics*, **14**, e1007672.

923 Presgraves DC (2008) Sex chromosomes and speciation in Drosophila. *Trends in Genetics*,
924 **24**, 336–343.

925 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal
926 components analysis corrects for stratification in genome-wide association studies. *Nature*
927 *Genetics*, **38**, 904–909.

928 Pritchard JK, Donnelly P (2001) Case–control studies of association in structured or admixed
929 populations. *Theoretical population biology*, **60**, 227–237.

930 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using mul-
931 tilocus genotype data. *Genetics*, **155**, 945–959.

932 Puechmaille SJ (2016) The program structure does not reliably recover the correct population
933 structure when sampling is uneven: subsampling and new estimators alleviate the problem.
934 *Molecular Ecology Resources*, **16**, 608–627.

935 Raj A, Stephens M, Pritchard JK (2014) faststructure: Variational inference of population
936 structure in large snp data sets. *Genetics*, **197**, 573–589.

937 Rice A, Glick L, Abadi S, *et al.* (2015) The chromosome counts database (ccdb)–a community
938 resource of plant chromosome numbers. *New Phytologist*, **206**, 19–26.

939 Rosenzweig BK, Pease JB, Besansky NJ, Hahn MW (2016) Powerful methods for detecting
940 introgressed regions from population genomic data. *Molecular Ecology*, **25**, 2387–2397.

941 Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applies
942 statistician. *The Annals of Statistics*, pp. 1151–1172.

943 Sæther SA, Sætre GP, Borge T, *et al.* (2007) Sex chromosome-linked species recognition and
944 evolution of reproductive isolation in flycatchers. *Science*, **318**, 95–97.

945 Sankararaman S, Mallick S, Dannemann M, *et al.* (2014) The genomic landscape of Nean-
946 derthal ancestry in present-day humans. *Nature*, **507**, 354–357.

947 Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in
948 admixed populations. *American Journal of Human Genetics*, **82**, 290–303.

949 Schumer M, Cui R, Powell DL, Rosenthal GG, Andolfatto P (2016) Ancient hybridization
950 and genomic stabilization in a swordtail fish. *Molecular Ecology*, **25**, 2661–2679.

951 Schumer M, Powell DL, Corbett-Detig R (2019) Versatile simulations of admixture and
952 accurate local ancestry inference with mixnmatch and ancestryinfer. *bioRxiv*, p. 860924.

953 Schwarz G, *et al.* (1978) Estimating the dimension of a model. *The annals of statistics*, **6**,
954 461–464.

955 Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture propor-
956 tions from next generation sequencing data. *Genetics*, **195**, 693–702.

957 Sohn KA, Ghahramani Z, Xing EP (2012) Robust estimation of local genetic ancestry in
958 admixed populations using a non-parametric Bayesian approach. *Genetics*.

959 Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model
960 complexity and fit. *Journal of the Royal Statistical Society: Series B*, **64**, 583–639.

961 Stephens M (2000) Dealing with label switching in mixture models. *Journal of the Royal*
962 *Statistical Society: Series B (Statistical Methodology)*, **62**, 795–809.

963 Stift M, Kolář F, Meirmans PG (2019) Structure is more robust than other clustering meth-
964 ods in simulated mixed-ploidy populations. *Heredity*, **123**, 429–441.

965 Szymura JM, Barton NH (1986) Genetic analysis of a hybrid zone between the fire-bellied
966 toads, *Bombina bombina* and *B. variegata*, near Cracow in southern Poland. *Evolution*,
967 **40**, 1141–1159.

968 Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: Analytical
969 and study design considerations. *Genetic Epidemiology*, **28**, 289–301.

970 The HDF5 Group (2010) *Hierarchical data format version 5, 2000-2010*.
971 http://www.hdfgroup.org/HDF5.

972 Vieira FG, Fumagalli M, Albrechtsen A, Nielsen R (2013) Estimating inbreeding coefficients
973 from ngs data: impact on genotype calling and allele frequency estimation. *Genome re-*
974 *search*, pp. gr–157388.

975 Watanabe S (2010) Asymptotic equivalence of bayes cross validation and widely applicable
976 information criterion in singular learning theory. *Journal of Machine Learning Research*,
977 **11**, 3571–3594.

978 Wegmann D, Kessner DE, Veeramah KR, *et al.* (2011) Recombination rates in admixed
979 individuals identified by ancestry-based inference. *Nature Genetics*, **43**, 847–853.

980 Witherspoon D, Marchani E, Watkins W, *et al.* (2006) Human population genetic structure
981 and diversity inferred from polymorphic L1 (LINE-1) and Alu insertions. *Human heredity*,
982 **62**, 30–46.

983 Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with Bayesian sparse linear
984 mixed models. *PLoS Genetics*, **9**, e1003264.

Figures

X	Sequence data
g	Genotype
z	Locus-specific, local ancestry
p	Allele frequency
q	Admixture proportion
π	Ancestral allele frequency
F	Genetic differentiation
γ	Prior membership in cluster K
α	Genetic variation in ancestor

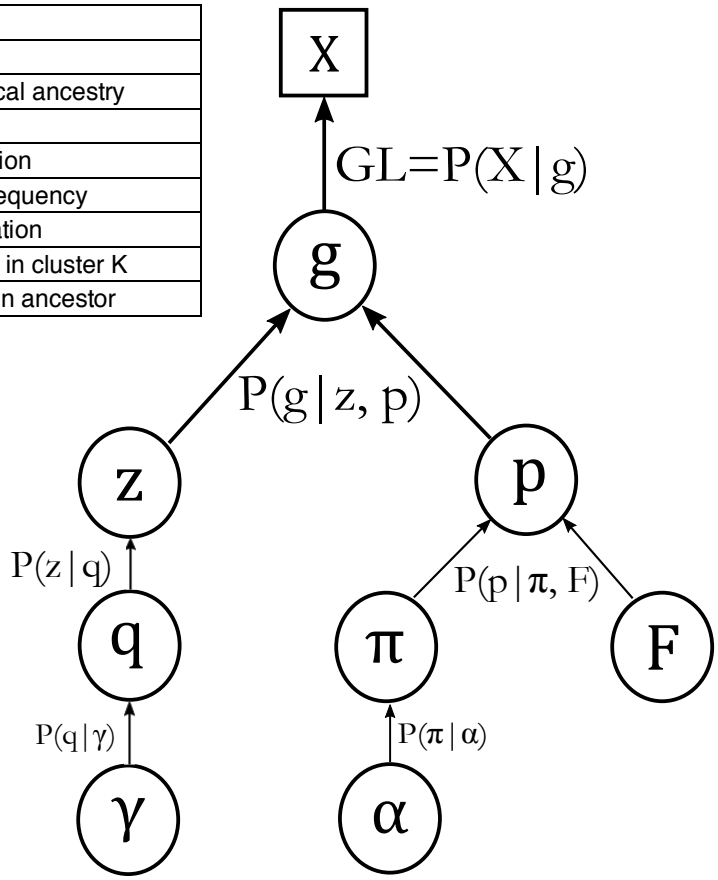

Figure 1: The graphical representation of the `entropy` model illustrates the information sharing in the model. Parameters that are being estimated are represented inside circles and the input sequence data are represented inside the square. The probability functions that generate these quantities are presented below each parameter. Typically, in a hierarchical framework we start at the bottom with new estimates of random values following a prior probability distribution. Then, conditional on these parameters, in the next highest level in the hierarchy we obtain estimates, and so forth, until the top level (the likelihood, here $GL = P(X|g)$) where estimates are constrained and estimated according to information in the data and prior probabilities. The parameter $\mathbf{q}$ is replaced by $\mathbf{Q}$ when using the *ancestry complement* model.

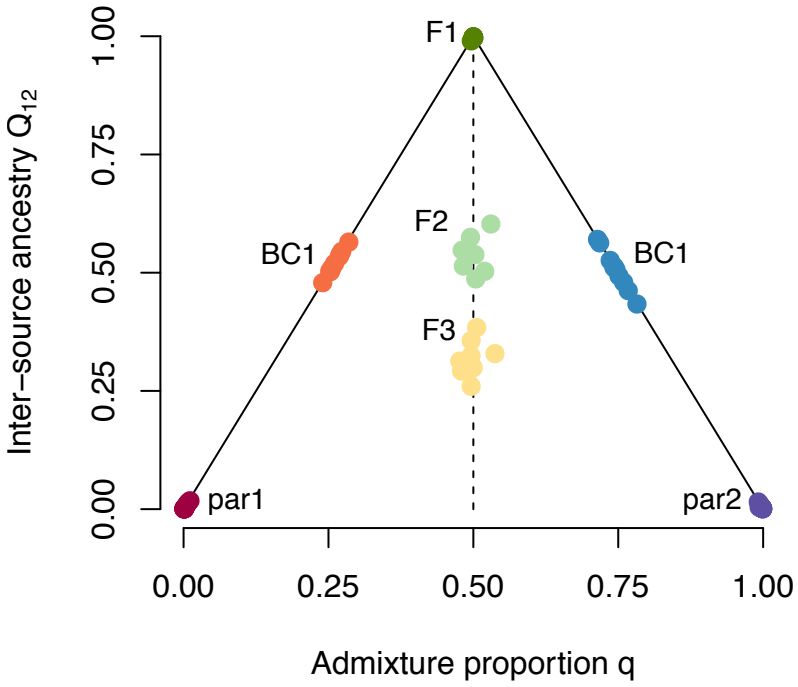

Figure 2: The diploid *ancestry complement* model in `entropy` with F1 hybrids (along the $q = 0.5$ line) between two parental populations ($K = 2$, at $q = 0.0$ and 1.0) reveals parameter differences between different early hybrid generations. The combination of admixture proportion q and ancestry complement $\mathbf{Q}$ distinguishes among F1, F2, and F3 hybrids. The admixture proportion (q) values for these three classes of ancestry are all 0.5 with some variance, but Q_{12} declines from 1 in the F1 with each generation of hybridization. Additionally, BC hybrids have maximal Q_{12} for a given q. The solid lines for the triangle indicate individuals with maximal possible Q_{12} values, corresponding to having at least one non-admixed parent.

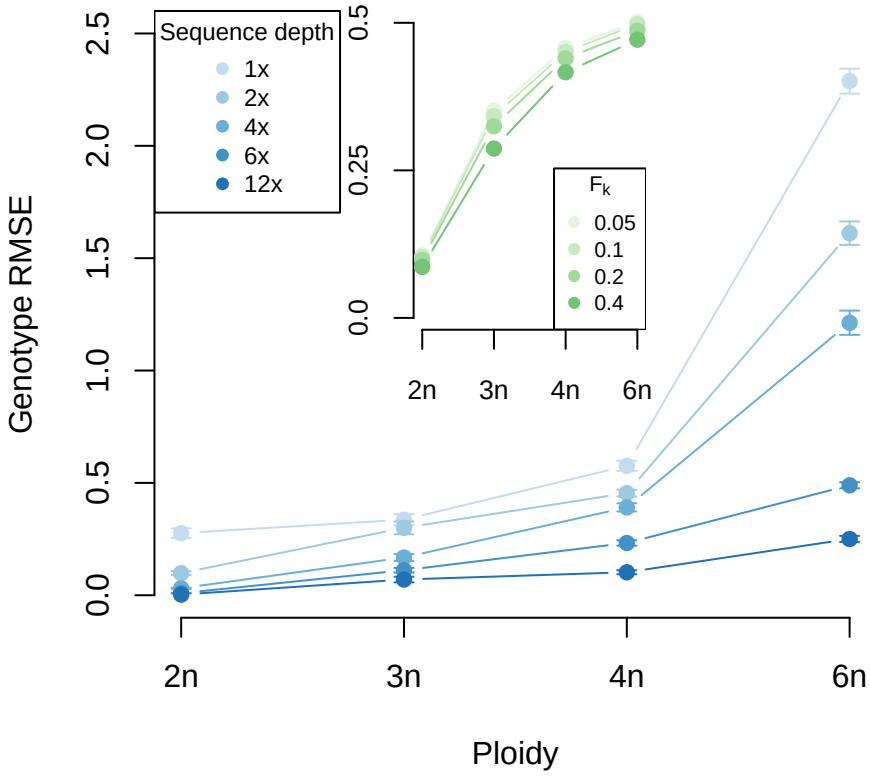

Figure 3: Error in genotype estimation across a range of ploidy decreased with greater sequence depth. The outer plot depicts change in RMSE for different ploidal levels versus sequence depth, across $F = \{0.05, 0.1, 0.2, 0.4\}$ and number of source populations $K = \{1, 2, 3\}$. Error increased with the number of allele copies in polyploids, because the range of variation and possible error in genotype is greater for polyploids than diploids. We found consistently higher error for lower sequence depth (across all ploidal levels). The inner plot depicts the change in RMSE for different ploidy and population differentiation (F) across a sequence depth of $n\times$ (with n ploidy and $K = 2$). Error in genotype estimation increased with ploidal level, but was affected very little by the extent of population differentiation.

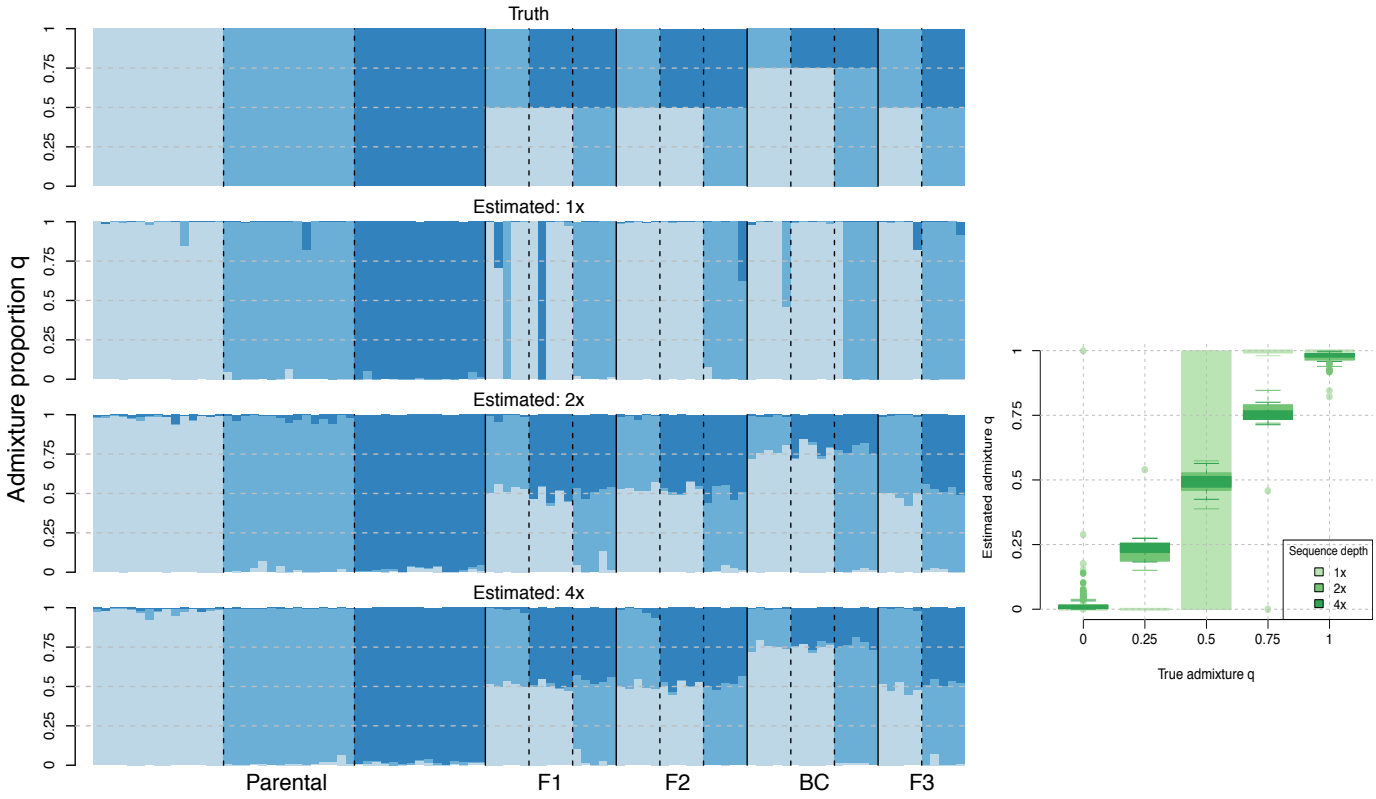

Figure 4: With increasing sequencing depth (rows 2–4) the model more accurately estimates true admixture proportions (row 1), particularly among hybrids. With 1× average sequence depth, `entropy` accurately estimated ancestry of parentals but did much less well with hybrids. With an average of 2× sequence depth at a locus `entropy` more accurately estimated admixture proportion, and at 4× average sequence depth estimates are very close to the truth. The subset plot contains a visual summary of the diminishing returns with higher depth, averaged across all ancestry groups. This plot contains comparisons of estimated and true admixture proportions for varying sequence depths across $K = 3$ source populations and 100 tetraploid individuals.

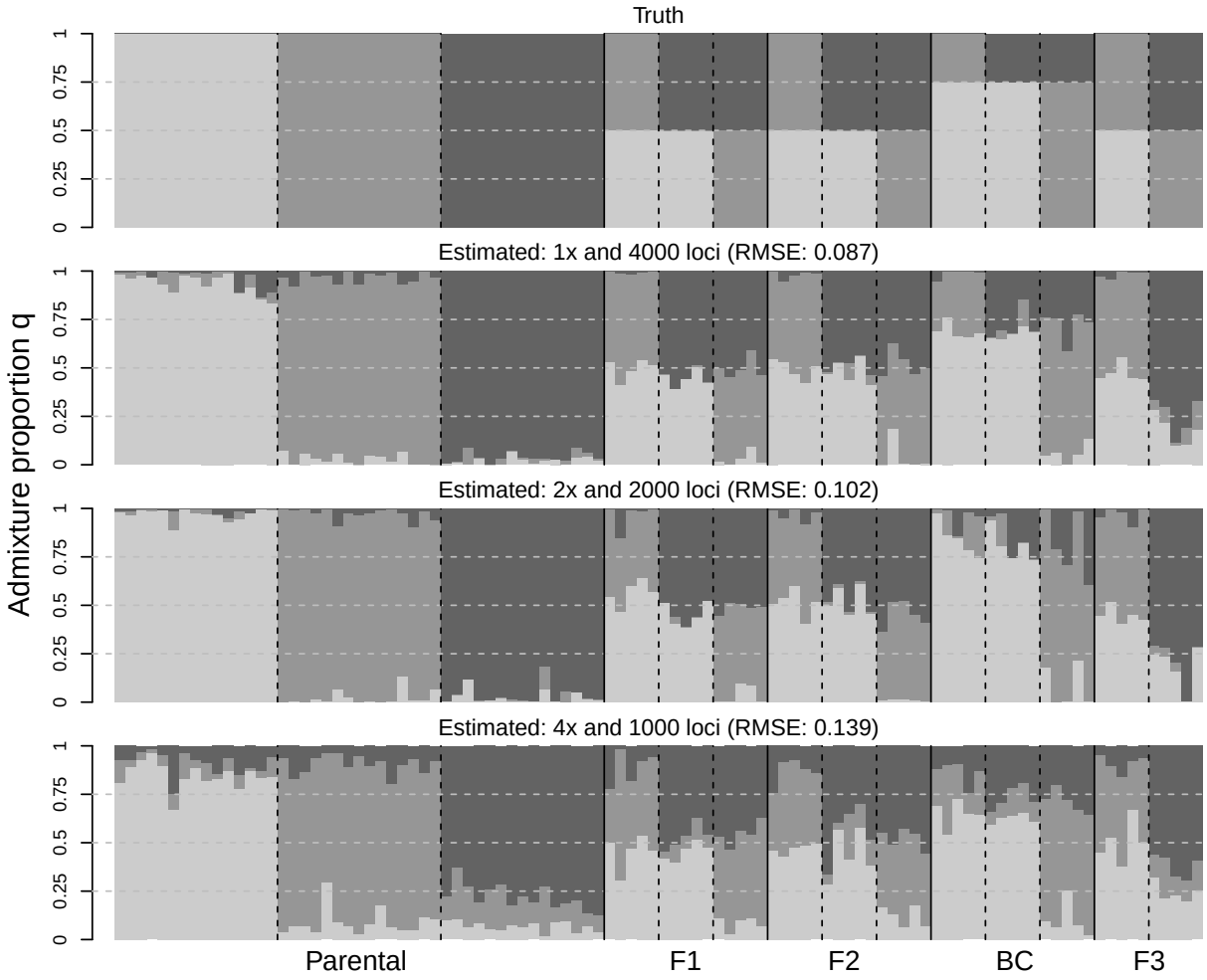

Figure 5: Admixture proportion is more accurately estimated with higher coverage and lower sequence depth. With higher coverage (i.e., more loci across the genome, 4000 loci) and a lower average sequence depth (1×), the estimates are closer to the simulated truth for global ancestry over the same data set subsampled to lower coverage (1000 loci) and a correspondingly higher sequence depth (4×). Admixture proportion estimates and RMSE values shown here for a continuum between 1× and 4× average sequence depth, with a corresponding coverage between 1000 loci and 4000 loci.

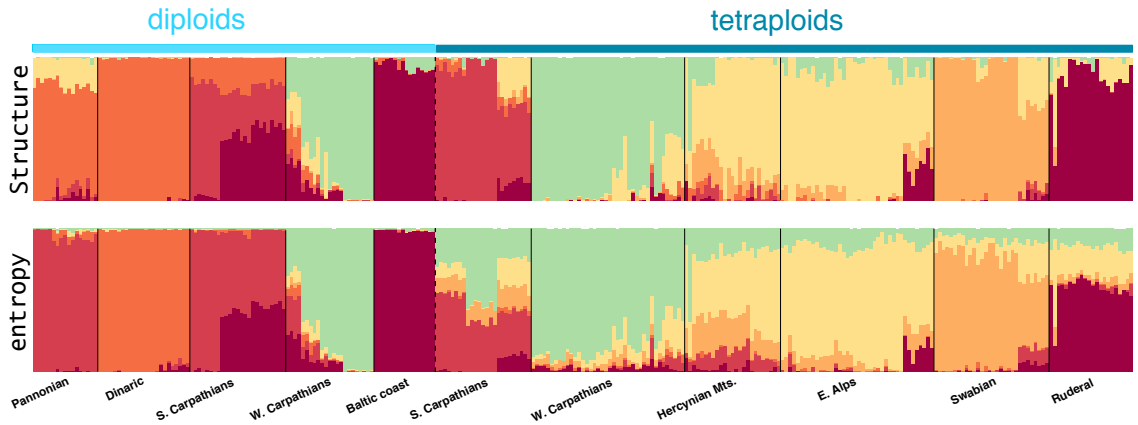

Figure 6: Admixture proportion estimates from `structure` and `entropy` agree very well for most the 287 *Arabidopsis arenosa* individuals from a mixed diploid and autotetraploid sample of populations across Europe for a $K = 6$ model with a median sequence depth of 10× (data from Monnahan *et al.*, 2019). The $K = 6$ model was the preferred model in the analysis by Monnahan *et al.* (2019). The two most notable differences between the `structure` and `entropy` estimates were the labeling of the Pannonian individuals (far left) as red ancestry by `entropy` versus a mixture of orange & yellow by `structure` and the different contributions to the composition of diploid and tetraploid S. Carpathians in the `entropy` and `structure` analyses.

Supplementary Material

Model description

We present a hierarchical Bayesian model to jointly infer genotypes and admixture proportions from sequence data of polyploid and mixed-ploidy populations. This model is implemented in software called `entropy`. This model is similar to the admixture model presented in `structure` (Pritchard *et al.*, 2000; Falush *et al.*, 2003), but with the exception that our model uses genotype likelihood data to incorporate uncertainty arising from sequence data, in the estimation of our downstream parameters. A graphical description of the model is presented in Figure 1 of the main text. The software to sample from the posterior distribution of the parameters (using MCMC) was written in C++, using the GNU Scientific Library (Galassi *et al.*, 2009) and HDF5 (The HDF5 Group, 2010). The program can be installed via the bioconda channel (`https://anaconda.org/bioconda/popgen-entropy`) or from source by cloning the Bitbucket repository (`https://bitbucket.org/buerklelab/mixedploidy-entropy/`), which also houses the on-going developmental code base.

Below we provide a more detailed description of the model, with information on the sampling distributions and how it differs from the diploid version in Gompert *et al.* (2014b). We present the two models: the admixture proportion and ancestry complement models, as presented in the main text.

Model 1 (admixture proportion model) As described in the main text, the probability of observing the genotype $\mathbf{g}$ is conditional on the unknown population of origin $\mathbf{z}$ of each allele that forms the genotype, and the unknown allele frequencies $\mathbf{p}$ in the source populations, $P(\mathbf{g}|\mathbf{z},\mathbf{p})$. We use genotype likelihoods, $L(\mathbf{g}|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{g})$ rather than raw sequence data $\mathbf{X}$ as model input. The genotype likelihoods are pre-calculated, taking into account the number of reads, number of genotypes, read specific error rate, haplotypic information, etc., given by the sequence data $\mathbf{X}$ (using softwares such as from `GATK`, McKenna *et al.* 2010, `SAMtools`, Li 2011, or `FreeBayes`, Garrison & Marth 2012). The genotype likelihoods were

normalized to sum to 1.

As we restrict our model to work with bi-allelic loci, for an n-ploid individual, we can expect to see $n+1$ genotypic states or dosage values at each locus j. Each allele at a locus is encoded as either 0 for the reference or 1 for the alternate. The sum at each locus, across all allele copies (e.g., four allele copies in a tetraploid), denotes the genotype at that locus. We can then calculate the probability of each genotype as the product over the probabilities of each allelic state across all alleles, conditional on $\mathbf{z}$ and $\mathbf{p}$. This discrete probability distribution is given as a Bernoulli distribution with a single draw at each allele, repeated n times for each locus in an n-ploid individual with probability equal to the allele frequency of the alternative allele in the population of origin k.

$$P(g_{ij}|\mathbf{p_j}, \mathbf{z_{ij}}) = \prod_{k}\prod_{a=1}^{n} \begin{cases} p_{jk}^{g_{ija}}(1-p_{jk})^{n-g_{ija}} & \text{when } k = z_{ija} \\ 1 & \text{otherwise} \end{cases}$$

Here, $\mathbf{z_{ij}} = [k_1, k_2, \ldots, k_n]$ denotes the local ancestry of the n allele copies for individual i, and z_{ija} denotes the local ancestry of the specific allele copy, a in the individual. The term p_{jk} denotes the corresponding allele frequency in the k^{th} source population.

The remainder of our model deviates little from the `structure` *admixture* model with correlated allele frequencies presented in Falush *et al.* (2003). We specify a set of admixture proportions, denoted by $q_1, q_2, \ldots, q_k$ to indicate the proportion of the individual's genome inherited from each of k source populations. These admixture proportions give the prior for the local ancestry $\mathbf{z}$ in a simple fashion, i.e., $P(z_{ija} = k) = q_{ik}$ for $a \in \{0, 1, \ldots, n\}$. We then place a Dirichlet prior on the admixture proportions for each individual with a scale parameter, λ, estimated from the data.

The probability of the unobserved allele frequency p_{jk} of locu j in source population k is calculated assuming an F-model (as presented in Balding & Nichols (1995)), where the population allele frequency is the result of divergence F_k from an ancestral population,

characterized by allele frequency π_j. We draw p_{jk} from a Beta distribution with shape parameters π_j and $(1-\pi_j)$, both multiplied by $(1/F_k - 1)$. F_k can be seen as a measure of genetic divergence from the ancestral population, analogous to F_{ST}:

$$P(p_{jk}|\pi_j, F_k) \sim \text{beta}(\pi_j \frac{1-F_k}{F_k}, (1-\pi_j)\frac{1-F_k}{F_k})$$

The allele frequencies π_j are obtained from a symmetrical beta distribution, $P(\pi_j|\alpha) \sim beta(\alpha, \alpha)$. The hyperparameter α can be seen as a measure of genetic diversity in the ancestral population, and is drawn from a Uniform distribution, $\alpha \sim \text{Uniform}(0, 10, 0000]$. F_k is assigned an uninformative prior $beta(1, 1)$ to indicate equal support for any value of $F_k \in [0, 1]$.

We specify the prior probability for the genome-wide admixture proportion vector $\mathbf{q}$ with a Dirichlet distribution with parameter vector $\gamma = (\gamma_1, \ldots, \gamma_K)$. To assign the same prior probability for each ancestral deme, we specify identical values for all γ_k. This is appropriate when assuming that neither individuals with ancestry from a single population, nor any particular class of hybrids dominate the hybrid zone. The hyperparameter γ_k is drawn from a Uniform distribution between 0 and 10.

Thus, the posterior probability distribution for the `entropy` model is given by

$$P(\mathbf{g}, \mathbf{z}, \mathbf{p}, \mathbf{q}, \pi, \mathbf{F}, \alpha, \gamma|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{g})P(\mathbf{g}|\mathbf{p}, \mathbf{z})P(\mathbf{z}|\mathbf{q})P(\mathbf{q}|\gamma)P(\mathbf{p}|\pi, \mathbf{F})P(\pi|\alpha)P(\phi)$$

where $P(\phi)$ is the joint probability of the terminal parameters in the hierarchy.

Model 2 (ancestry complement model) This model is almost identical to the previous model, except in the formulation of admixture proportion. Here, we make use of a matrix $\mathbf{Q}$ instead of the vector $\mathbf{q}$ that is used in `structure`, called the ancestry complement matrix to specify interspecific (or inter-demic) ancestry at a locus, as we shall see below. This model is only available for diploid individuals.

We can obtain additional information on genome-wide admixture by considering a combination of ancestry states at each locus, instead of treating each allele copy as being derived independently from a source population. Therefore, we calculated the probability for locus-specific ancestry jointly for both allele copies (in a diploid) by working with ancestry $\mathbf{z_{ij}}$ as a whole, instead of ancestry for each allele copy z_{ija} separately. The ancestral parameter $\mathbf{z_{ij}}$ can be seen as a $K \times K$ matrix with all its elements set to zero except the element at row $k = z_{ij1}$ and column $k' = z_{ij2}$ set to one. This means that $\mathbf{z_{ij}}$ is represented as a "one-hot" vector with the index for the corresponding source population denoted by a one, with the remaining entries being zero. This indexing of allele copies to a source population lets us select the corresponding allele frequency for the individual when calculating downstream parameters. The probability of the locus-specific ancestry is then calculated conditional on the genome-wide ancestry complement matrix $\mathbf{Q_i}$ for individual i. $\mathbf{Q_i}$ is another $K \times K$ matrix that gives the prior probabilities for genome-wide admixture, or genome composition, for each of the possible states of $\mathbf{z_{ij}}$, with all elements in $\mathbf{Q_i}$ summing to one. The elements on and off the main diagonal give the probabilities for intra-source and inter-source ancestry, respectively. The probability for locus-specific ancestry conditional on genome-wide admixture follows a categorical distribution (or a multinomial distribution with one draw) and is given by

$$P(z_{ij_{kk'}} = 1|\mathbf{Q_i}) = Q_{i_{kk'}}$$

with k and k' giving the rown and column of the $\mathbf{z_i j}$ and the $\mathbf{Q_i}$ matrix. The genome-wide admixture proportion $\mathbf{q}$ is not included as a model parameter but can be calculated marginally from $\mathbf{Q}$ within each interation as

$$q_{i_k} = \frac{1}{2}(\sum_{s=1}^{K} Q_{i_{ks}} + \sum_{t=1}^{K} Q_{i_{tk}})$$

Similar to the previous model, the $\gamma = (\gamma_{11}, \ldots, \gamma_{KK})$ prior is now a matrix instead of a vector, drawn from a Dirichlet distribution with an equal weighting on each ancestral deme.

Thus, the posterior probability distribution for all parameters in this hierarchical Bayesian model is given by

$$P(\mathbf{g}, \mathbf{z}, \mathbf{p}, \mathbf{Q}, \pi, \mathbf{F}, \alpha, \gamma | \mathbf{X}) \propto P(\mathbf{X}|\mathbf{g})P(\mathbf{g}|\mathbf{p}, \mathbf{z})P(\mathbf{z}|\mathbf{Q})P(\mathbf{Q}|\gamma)P(\mathbf{p}|\pi, \mathbf{F})P(\pi|\alpha)P(\phi)$$

where $P(\phi)$ is the joint probability of the terminal parameters in the hierarchy, with the only replacement being $\mathbf{Q}$ for $\mathbf{q}$.

MCMC updates

Below we describe the process for sampling from the posterior for each of our parameters (using various techniques) given the conditional distributions mentioned above. This process also acts as a proxy for the formulation in code with each update step written into a separate function. We will move downward from the graph presented in Figure 1 of the main text. The following text was adapted from the Supplement of Lindtke *et al.* (2014), with minor changes for dealing with higher ploidal levels.

1. Update $\mathbf{g}$ (sampled from the full distribution)
2. Update $\mathbf{z}$ (sampled from the full distribution)
3. Update $\mathbf{p}$ (Gibbs sampling)
4. Update $\boldsymbol{\pi}$ (Metropolis sampling)
5. Update $\mathbf{F}$ (Metropolis sampling)
6. Update α (Metropolis sampling)
7. Update $\mathbf{q}/\mathbf{Q}$ (Gibbs sampling)
8. Update γ (Metropolis sampling)

To implement this sampling procedure, we cycle through each parameter in the model and run the update step for this parameter by holding all other parameters constant at their current value. Once, we are through all the parameters in the model, we will start back up at the 'top' of the hierarchy with the likelihood of the sequence data and run through the same sampling process again. The update steps for each parameter are specified in more detail below:

1. Update $\mathbf{g}$:

$$P(g_{ij}|L(g_{ij}|x_{ij}), \mathbf{z_{ij}}, \mathbf{p_j}) = \frac{L(g_{ij}|x_{ij})P(p_{jk}|\mathbf{z_{ij}}, g_{ij})}{\sum_{g_{ij1}=0}^{1} \dots \sum_{g_{ijn}=0}^{1} L(g_{ij}|x_{ij})P(p_{jk}|\mathbf{z_{ij}}, g_{ij})}$$

Here, $g_{ij} = \{g_{ij1}, \dots, g_{ijn}\}$ and $L(g_{ij}|x_{ij})$ gives the pre-calculated likelihood of each genotype (the input data). For example, in a triploid ($n = 3$),
$g_{ij} \in \{000, 001, 010, 011, 100, 101, 110, 111\}$ for each allele copy.
$P(p_{jk}|\mathbf{z_{ij}}, g_{ij}) = p_{jk^1}^{g_{ij1}}(1 - p_{jk})^{1-g_{ij1}} \dots p_{jk^n}^{g_{ijn}}(1 - p_{jk^n})^{1-g_{ijn}}$ is the product of the allele frequencies for the first to the n^{th} allele copy in genotype g_{ij} in population $k^1 = z_{ij1}, \dots, k^n = z_{ijn}$, respectively. This update step essentially combines the likelihood of observing a certain genotype given the read data, scaled by the expected frequency of that genotype at that locus (given by the $P(\mathbf{p}|\mathbf{z}, \mathbf{g})$ term).

2. Update $\mathbf{z}$: For the admixture proportion model, we have

$$P(z_{ijk} = 1|g_{ij}, \mathbf{p_j}, \mathbf{q_i}) = \frac{q_i P(p_{jk}|g_{ij})}{\sum_{k=1}^{K} q_i P(p_{jk}|g_{ij})}$$

where $P(p_{jk}|g_{ij})$ is given in the previous update step (for a certan $z_{ijk} = 1$). Here, we are multiplying two 1-dimensional vectors of length K and dividing each element by the average to obtain normalized values between 0 and 1. This update follows a similar pattern to the update for the genotypes. Here, we sample from a full distribution because we obtain a value for each cluster k in the discrete probability distribution

from 1 through K. From these vector of values, we perform a single multinomial draw (i.e., $n = 1$) to obtain an index for the putative ancestral cluster.

Similarly, for the ancestry complement model, we have

$$P(z_{ijk} = 1|g_{ij}, \mathbf{p_j}, \mathbf{Q_i}) = \frac{q_i P(p_{jkk'}|g_{ij})}{\sum_{k=1}^{K}\sum_{k'=1}^{K} Q_i P(p_{jkk'}|g_{ij})}$$

where $P(p_{jkk'}|g_{ij})$ is given in the previous update step with only two k values since we are dealing with diploid loci and two allele copies, meaning two ancestral populations in k and k'.

3. Update $\mathbf{p}$:

$$P(p_{jk}|\mathbf{z_j}, \mathbf{g_j}, F_k, \pi_j) \sim \text{beta}(\pi_j(\frac{1}{F_k} - 1) + r_{ijk1}, (1 - \pi_j)(\frac{1}{F_k} - 1) + r_{ijk0})$$

where

$$r_{ijk1} = \sum_i \sum_n \begin{cases} g_{ijn} & \text{when } k = z_{ijn}, \\ 0 & \text{when } k \neq z_{ijn} \end{cases}$$

and

$$r_{ijk0} = \sum_i \sum_n \begin{cases} (1 - g_{ijn}) & \text{when } k = z_{ijn}, \\ 0 & \text{when } k \neq z_{ijn} \end{cases}$$

give the counts for the alternate and reference allele copies assigned to an ancestral population k, respectively.

4. Update $\boldsymbol{\pi}$: Propose a new π'_j from

$$\pi'_j|\pi_j \sim \text{Uniform}(\pi_j - 0.1, \pi_j + 1)$$

and accept the proposed value as the new update for π_j with probability $\min(1, r)$ if

$0 < \pi'_j < 1$, with

$$r = \frac{P(\pi'_j|\alpha)}{P(\pi_j|\alpha)}$$

Using Bayes' rule,

$$r = \frac{P(\alpha|\pi'_j)}{P(\alpha|\pi_j)} \prod_k \frac{P(\pi'_j, \theta_k|p_{jk})}{P(\pi_j, \theta_k|p_{jk})}$$

where

$$P(\pi_j, \theta_k|p_{jk}) = \frac{p_{jk}^{pi_j\theta_k - 1}(1 - p_{jk})^{(1-\pi_j)\theta_k - 1}}{\text{beta}(\pi_j\theta_k, (1 - \pi_j)\theta_k)}$$

and

$$P(\alpha|\pi_j) = \frac{\pi_i^{\alpha-1}(1 - \pi_j)^{\alpha-1}}{\text{beta}(\alpha, \alpha)}$$

with $\theta_k = \frac{1}{F_k} - 1$, and the probabilities for π'_j are computed in a similar manner.

5. Update $\mathbf{F}$: Proposal for a new F'_k from

$$F'_k|F_k \sim \text{Uniform}(F_k - 0.01, F_k + 0.01)$$

and accept F'_k as new update for F_k (represented here as θ'_k and θ_k) with probability $\min(1, r)$ if $0 < F'_k < 1$, with

$$r = \prod_j \frac{P(\pi_j, \theta'_k|p_{jk})}{P(\pi_j, \theta_k|p_{jk})}$$

where $P(\pi_j, \theta_k|p_{jk})$ is given from the previous update step, with $\theta'_k = \frac{1}{F'_k} - 1$.

6. Update α: Proposal for a new α' from

$$\alpha'|\alpha \sim \text{Uniform}(\alpha - 20, \alpha + 20)$$

and accept α' as new update for α with probability $\min(1, r)$ if $0 < \alpha' \leq 10000$, with

$$r = \prod_i \frac{P(\alpha'|\pi_j)}{P(\alpha|\pi_j)}$$

where $P(\alpha|\pi_j)$ is given from the previous update step.

7. Update $\mathbf{q}/\mathbf{Q}$: This step involves a simple counting procedure and a single Gibbs update by multiplying a mulitnomial likelihood with a Dirichlet prior, which gives us a Dirichlet distribution with updated parameters.

$$P(\mathbf{q_i}|\mathbf{z_i}, \gamma) \sim \text{Dirichlet}(\gamma_1 + \sum_n \sum_j z_{ijn1}, \ldots, \gamma_K + \sum_n \sum_j z_{ijnK})$$

where z_{ijn1} denotes the local ancestry values (either 0 or 1) for each locus j in an n-ploid individual i descended from source population $k = 1$. Similarly, for the ancestry complement model, we have

$$P(\mathbf{Q_i}|\mathbf{z_i}, \gamma) \sim \text{Dirichlet}(\gamma_{11} + \sum_j z_{ij11}, \ldots, \gamma_{KK} + \sum_j z_{ijKK})$$

where we sum over the local ancestry values across all loci j in the genome to obtain an estimate for genome-wide admixture proportion.

8. Update γ: In the model, all elements of γ_k are identical (for both matrix and vector form). We, therefore, propose new γ' by proposing a single element γ'_k from:

$$\gamma'_k|\gamma_k \sim \text{Uniform}(\gamma_k - 0.05, \gamma_k + 0.05)$$

and accept the new update with probability $\min(1, r)$ if $0 < \gamma'_k \leq 10$, with

$$r = \frac{P(\gamma'_k|\mathbf{q})}{P(\gamma_k|\mathbf{q})}$$

Using Bayes' rule,

$$r = \frac{P(\mathbf{q}|\gamma'_k)P(\gamma'_k|\gamma_k)}{P(\mathbf{q}|\gamma_k)P(\gamma_k|\gamma'_k)}$$

$$r = \prod_i \frac{P(q_i|\gamma'_k)}{P(q_i|\gamma_k)}$$

with the probabilities given in the previous update step. Since we adopt a Metropolis sampling scheme (i.e., symmetric proposal distributions with $P(\gamma'_k|\gamma_k) = P(\gamma_k|\gamma'_k)$), the second component to our update is equal to 1. This allows us to calculate the probability of acceptance, r, without considering this second term.

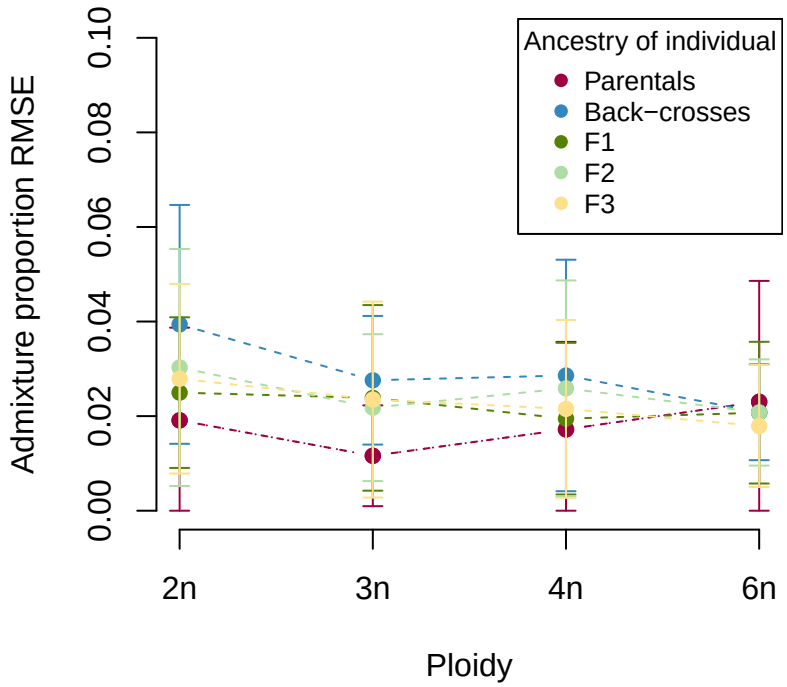

Figure S1: Change in admixture proportion RMSE across ploidal levels. Different ancestry classes in our simulation did not systematically affect the ability to recover true admixture proportions. This is shown for the case of sequence depth equal to $n\times$ and across various F values and number of source populations.

Supplementary Figures and Tables

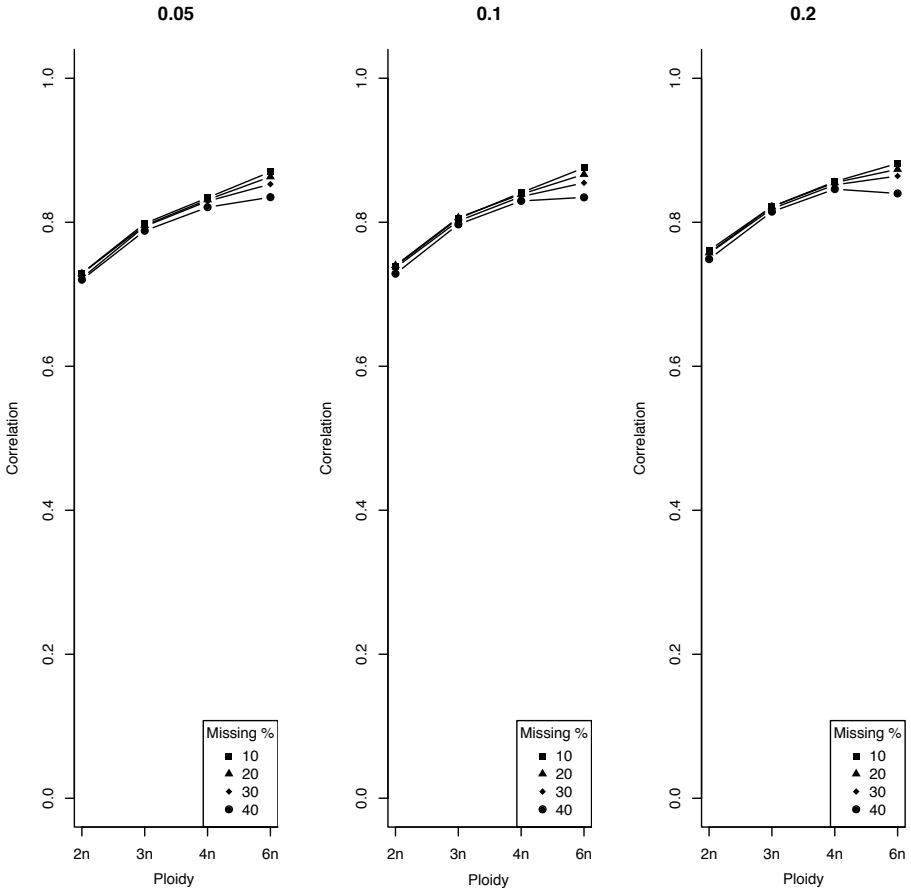

Figure S2: Correlation of estimated genotypes at missing sites to true genotypes over varying levels of genetic differentiation F (0.05, 0.1, and 0.2; panes of the plot). There was a slight gain in estimation accuracy when going from 40% missingness to 10% missingness. There was little or no effect of genetic differentiation on estimation accuracy across ploidy levels.

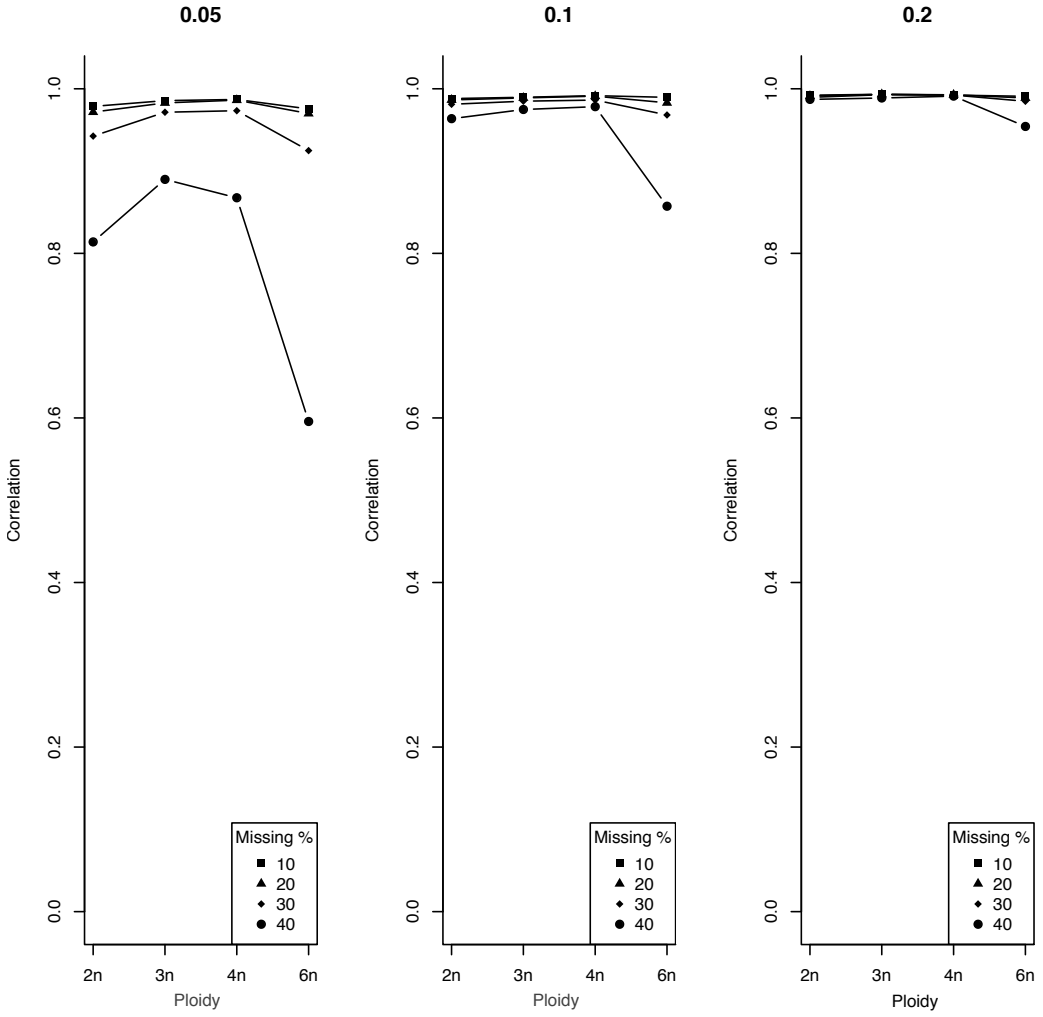

Figure S3: Correlation of estimated admixture proportion in individuals over varying levels of missingness and genetic differentiation F (0.05, 0.1, and 0.2; panes of the plot). The percentage of missingness was very important for simulations of demes that were genetically similar ($F = 0.05$) and hexaploid individuals, but even with high levls of missingness, more differentiated parental populations supported highly accurate admixture proportion estimates. The correlation between estimated parameters and the true was very high across ploidy levels (average ≈ 0.97).

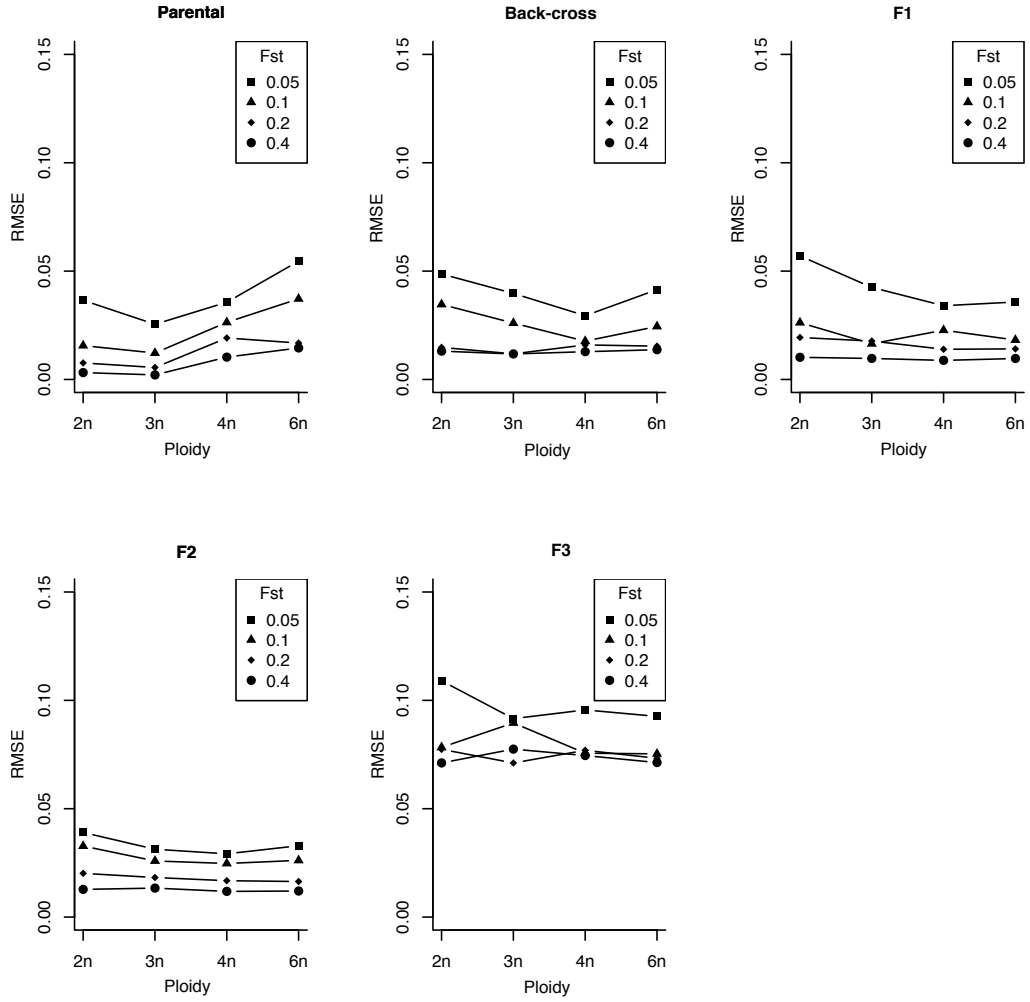

Figure S4: Mean squared error of admixture proportion across five early generation hybrid categories and ploidy levels for varying levels of F. Error in estimates decrease with increasing genetic differentiation for all categories of hybrids. The higher overall error with the F3 individuals was because it is harder to estimate the accurate q value given the high realized variance of individual genetic composition around the expectation.

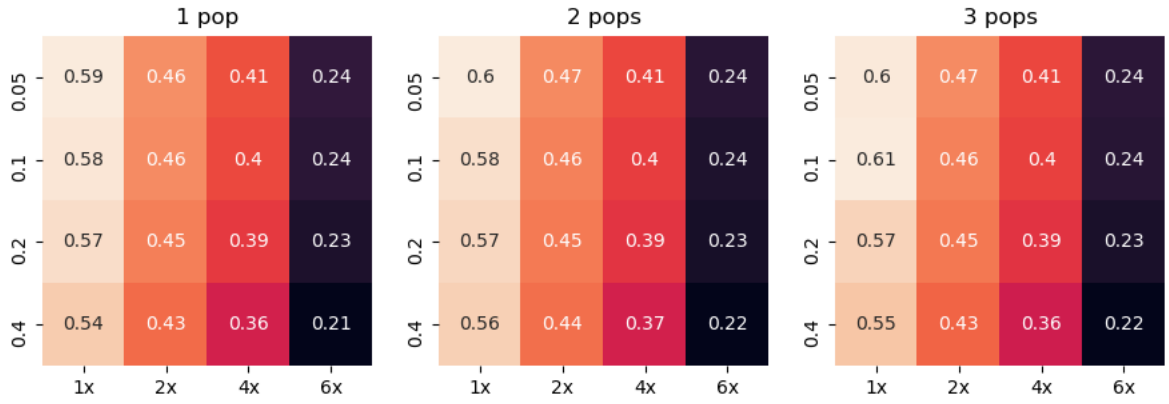

Figure S5: Mean squared error for estimation of genotypes in tetraploid individuals across sequence depths and population differentiation F and number of source populations. The steepest gradient in error was across the sequence depths (i.e., better estimation with greater sequence depth and slight improvement with higher values of genetic differentiation). The lowest error occurred with $F = 0.4$ and $6\times$ sequence depth.

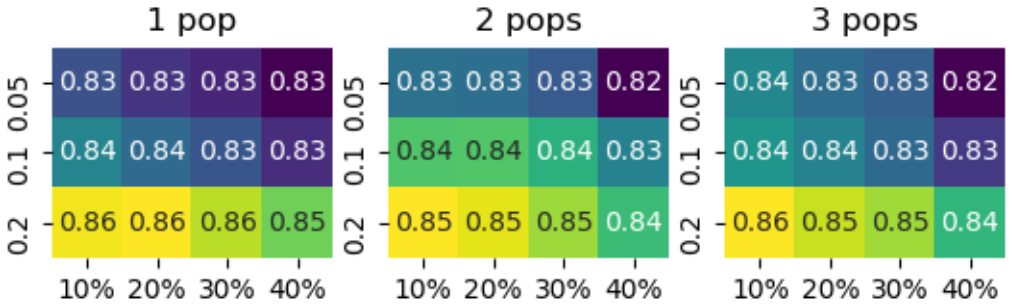

Figure S6: Consistently high correlation of the estimated tetraploid genotypes across degrees of data missingness, three levels of genetic differentiation, and number of source populations. The `entropy` model estimates had a $\approx 83\%$ correlation with the true genotypes at missing sites. Correlations were unaffected or increased slightly with higher differentiation and lower missingness percentage.

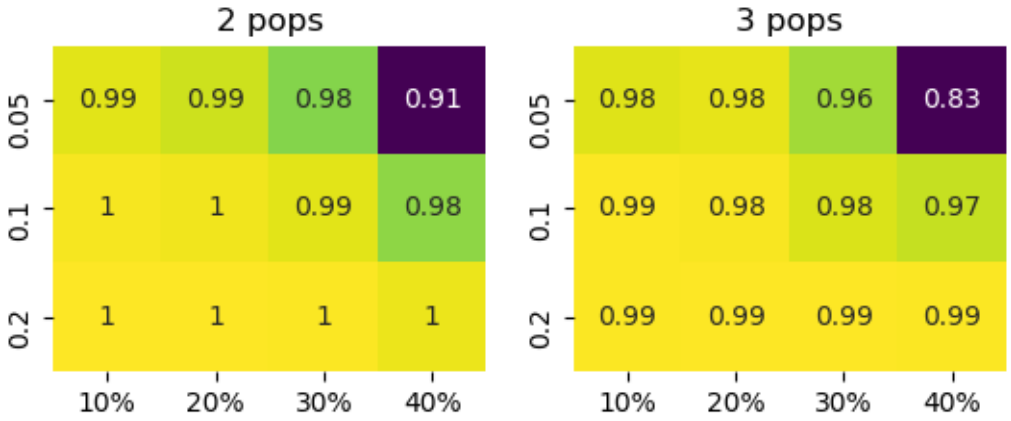

Figure S7: High correlation between estimated and simulated admixture proportion in tetraploid individuals across missingness percentage and genetic differentiation. The model estimated ancestry of individuals with high accuracy, across different levels of missingness in the loci. For example, in a $K = 2$ simulation with $F = 0.05$, the correlation between the true and estimated admixture proportions for individuals with 30% of their sites missing was 0.98. Correlations were lower in simulations with minimal genetic differentiation and high missingness in the data.

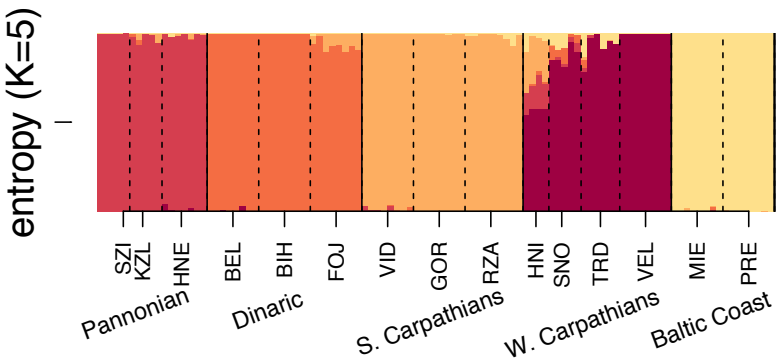

Figure S8: Admixture proportions of only the 105 diploid *A. arenosa* individuals for a K=5 model in `entropy`. The populations in the Pannonian region at the far left (labeled by *SZI, KZL, HNE*) fall into a distinct cluster compared to the rest of the individuals. The Pannonian cluster (red) is genetically the most district from the remaining ancestry groups and was expected to form a distinct group based on the analysis in Monnahan *et al.* (2019). However, this distinction was not found with the `structure` model applied to the whole data set with a K=6 model (as seen in Figure S9).

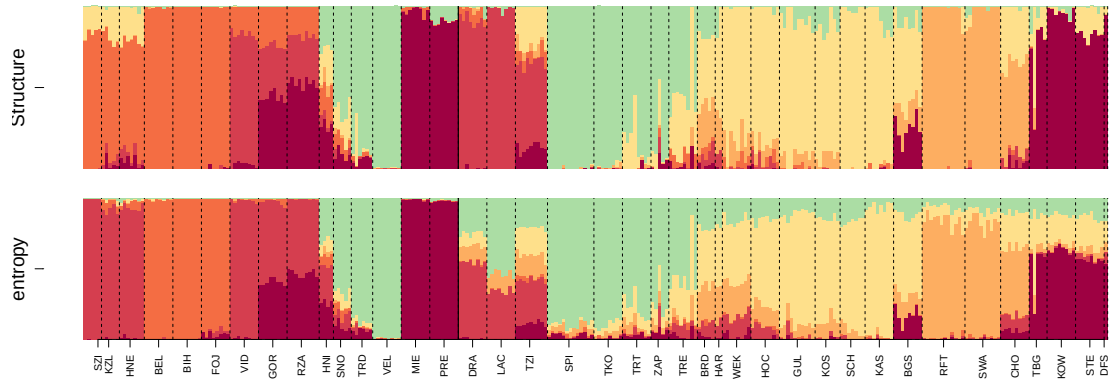

Figure S9: Admixture proportions of all 287 *A. arenosa* individuals for a K=6 model run in `entropy` plotted with population codes instead of regional codes.

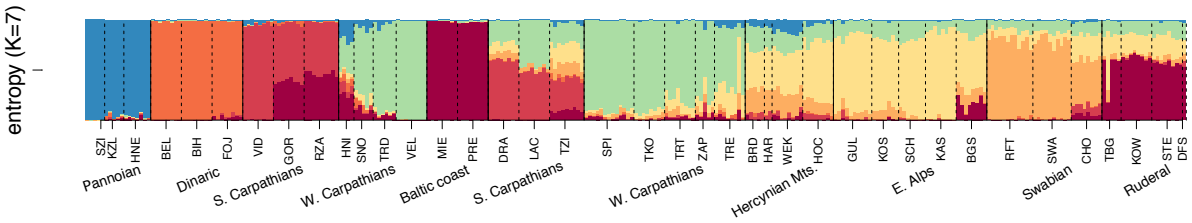

Figure S10: Admixture proportions of all 287 *A. arenosa* individuals for a K=7 model run through `entropy`. The populations in the Pannonian region to the far left (categorized by *SZI, KZL, HNE*) fell into a distinct cluster (blue) compared to the rest of the individuals, further confirming it as genetically differentiated relative to the other populations in the data set.

	K=2 & F=0.05	K=3 & F=0.05	K=2 & F=0.1	K=3 & F=0.1
2n	0.106	0.106	0.101	0.100
2n-4n	0.256	0.250	0.248	0.244
4n	0.409	0.407	0.399	0.400

Table S1: Error rates for genotype estimates in mixed diploid-tetraploid populations are in between the fully diploid (2n) and fully tetraploid (4n) population. This table contains RMSE values for genotypes in diploid (2n), diploid-tetraploid (2n-4n) and tetraploid (4n) populations for different numbers of ancestral demes K and levels of evolutionary divergence F.

Model	WAIC	lppd	neff
K=4	3079.1	-421.5	1118.0
K=5	3056.1	-415.9	1112.1
K=6	3021.5	-411.8	1098.9
K=7	2981.9	-404.4	1086.4
K=8	2963.0	-401.0	1080.5
K=9	2947.3	-399.4	1074.2
K=10	2958.5	-398.1	1081.1

Table S2: Table of WAIC values for various K models for the *Arabidopsis arenosa* data set with the best-fit model being $K = 9$. However, Monnahan *et al.* (2019) found $K = 6$ to be the best-fit, informed by a combination of the Bayesian Information Criterion (BIC, Schwarz *et al.* 1978) and a similarity index. Similar to other information criteria, the *WAIC* value provides the support for a certain value of K and is a combination of the log-predictive posterior density (*lppd*, similar to deviance in DIC) and the penalization term for the total number of effective parameters in the model (*neff*).

			Assumed		
F=0.05	K=1	K=2	K=3	K=4	K=5
K=2	61891.57	61321.97	61373.16	61372.37	61430.01
K=3	61598.17	61179.33	60842.83	60929.4	60931.71
F=0.2	K=1	K=2	K=3	K=4	K=5
K=2	61395.55	59922.69	59892.87	59957.28	59956.59
K=3	62235.11	61187.8	60205.53	60183.08	60203.56
F=0.4	K=1	K=2	K=3	K=4	K=5
K=2	60957.32	53878.83	53888.24	53867.42	53889.24
K=3	63773.64	58131.46	53072.45	53083.38	53052.75

Table S3: The assumed K is found to be equal to the simulated K only 33% of the time for a tetraploid data set. This table contains WAIC values to infer best-fit K from `entropy` for different simulation parameters. There is no apparent effect of F on the ability of our model to estimate number of demes K. These results differ drastically from the simulated data for hexaploid individuals presented in Table S4.

			Assumed		
F=0.05	K=1	K=2	K=3	K=4	K=5
K=2	240331.5	240496	244859.4	244908.7	243719.5
K=3	241100.1	241432.6	241246.5	242149.9	244027.2

F=0.2	K=1	K=2	K=3	K=4	K=5
K=2	244342	244109.4	244231.7	247922.2	244270.4
K=3	243375.4	243195.5	242924.9	243012	246588.3

F=0.4	K=1	K=2	K=3	K=4	K=5
K=2	264797.5	261440	261647.6	261706.7	261645.4
K=3	267000.9	264881.6	262363.8	262476.6	262560.2

Table S4: The assumed K is found to be equal to the simulated K more than 80% of the time for a hexaploid data set. This table contains WAIC values to infer best-fit K (for a range) from `entropy` for different simulation parameters, and we see that with higher F we capture 'true' K, which differs from the results for the simulated tetraploid data set presented in Table S3.