Received: 23 October 2018 | Revised: 14 December 2018 | Accepted: 4 January 2019

DOI: 10.1111/1755-0998.12995

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Minor allele frequency thresholds strongly affect population structure inference with genomic data sets

Ethan Linck[1] | C. J. Battey[2]

[1]Department of Biology and Burke Museum of Natural History and Culture, University of Washington, Seattle, Washington

[2]Department of Biology and Institute of Ecology and Evolution, University of Oregon, Eugene, Oregon

Correspondence
Ethan Linck, University of Washington Department of Biology, Seattle, WA.
Email: elinck@uw.edu

Funding information
U.S. Army, Grant/Award Number: 32 CFR 168a

Abstract

A common method of minimizing errors in large DNA sequence data sets is to drop variable sites with a minor allele frequency (MAF) below some specified threshold. Although widespread, this procedure has the potential to alter downstream population genetic inferences and has received relatively little rigorous analysis. Here we use simulations and an empirical single nucleotide polymorphism data set to demonstrate the impacts of MAF thresholds on inference of population structure—often the first step in analysis of population genomic data. We find that model-based inference of population structure is confounded when singletons are included in the alignment, and that both model-based and multivariate analyses infer less distinct clusters when more stringent MAF cutoffs are applied. We propose that this behaviour is caused by the combination of a drop in the total size of the data matrix and by correlations between allele frequencies and mutational age. We recommend a set of best practices for applying MAF filters in studies seeking to describe population structure with genomic data.

KEYWORDS
minor allele frequency, population genetic structure, principal components analysis, STRUCTURE

1 | INTRODUCTION

The distribution of genetic variation within and among individuals is crucial to understanding the organization of biological diversity and its underlying causes. Across the genome, variation in mutational age, the effects of different evolutionary processes and the influence of historical events can result in different classes of genetic variants characterized by their relative frequency in a given population (Griffiths & Tavaré, 1999; Kimura & Ohta, 1973; Mathieson & McVean, 2014; Slatkin, 1985). An excess of common alleles may reflect the signature of population bottlenecks (Marth, Czabarka, Murvai, & Sherry, 2004), purifying selection (Fay, Wyckoff, & Wu, 2001) or the absence of population subdivision (Pritchard, Stephens, & Donnelly, 2000). Alternatively, high frequencies of rare alleles can provide evidence of population expansion (Marth et al., 2004), detailed information on mutation rates and gene flow (Slatkin, 1985), and reveal geographically localized population subdivision (Barton & Slatkin, 1986; Gompert et al., 2014).

Because the distribution of allele frequencies across sites (also known as the site frequency spectrum, or SFS) reflects the unique combination of these varied factors, downstream analyses are sensitive to the influence of sampling methodologies that alter the SFS. Yet despite the explosive recent growth in population genetics provided by the advent of affordable reduced-representation genome sequencing for nonmodel organisms, there remain significant gaps in our knowledge of how population genetic inference is affected by data collection biases and filtering steps that preferentially shape the SFS.

These biases may originate either in wet lab or bioinformatic treatments. Prior to sequencing, the SFS may be shaped by ascertainment bias in library preparation: restriction site-associated DNA sequencing (RADseq)-style methods introduce genealogical biases (Arnold, Corbett-Detig, Hartl, & Bomblies, 2013) and nonrandom

patterns of missing data (Gautier et al., 2012) due to reliance on the presence of restriction cut sites; hybridization capture with ultraconserved element (UCE) probe sets necessarily involves targeting sites that are highly conserved across evolutionarily distant taxa (Faircloth et al., 2012); and single nucleotide polymorphism (SNP) arrays (or "chips") explicitly screen for variation at a particular frequency cutoff. During sequencing itself, relatively high error rates are accepted in individual reads, under the assumption they will be corrected during bioinformatic processing steps (Nielsen, Korneliussen, Albrechtsen, Li, & Wang, 2012). However, the absence of standard bioinformatic pipelines in ecology and evolutionary biology is itself a source of uncertainty (Shafer et al., 2017) because specific methodologies and parameter choices may dramatically affect the composition of data matrices.

For organisms lacking a suitable reference genome, de novo sequence assemblies may introduce substantial errors that affect both the SFS and inference of population genetic structure (Shafer et al., 2017). During read-mapping, SNP variation can result in higher rates of successful alignments in reads sharing the reference allele (Degner et al., 2009). Parameters used during variant detection can also play a significant role in determining the number and distribution of SNPs (Nielsen et al., 2012), the most frequently used marker type in modern population genetics. In particular, minor allele frequency (MAF) thresholds directly influence the SFS by imposing a cutoff on the minimum allele frequency allowed to incorporate a specific genetic variant. However, despite its potential importance, the two most popular comprehensive bioinformatic pipelines for RADseq data alternatively include (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013) or exclude (Eaton, 2014) the option to set MAF thresholds during variant calling, with the result that among empirical studies MAF thresholds are only sometimes reported (e.g., Winger, 2017, Blanco-Bercial & Bucklin, 2016).

One potential consequence of ambiguous MAF choice is variation in the ability to detect population subdivision (or structure), a fundamental goal of many population genetic studies. Previous empirical work suggests analyses of population structure are sensitive to filtering by allele frequency class. For example, estimates of Wright's fixation index F_{ST}—commonly employed to quantify population subdivision—are strongly restricted by the site frequency spectrum (Jakobsson, Edge, & Rosenberg, 2013). Similarly, studies using more geographically explicit test statistics (Mathieson & McVean, 2012) and/or clustering methods (Gompert et al., 2014; De La Cruz & Raska, 2014) inferred significantly different patterns and levels of population genetic structure when alternately using only common and rare variants. These results highlight the need for a detailed investigation of the behaviour of these methods using commonly applied MAF filters.

Clustering methods are particularly widespread in population genetic studies of nonmodel organisms where researchers generally lack a priori knowledge of population structure. They generally fall into one of two categories: model-based (or parametric) approaches and nonparamatric approaches. Model-based methods, exemplified by the influential program STRUCTURE (Pritchard et al., 2000; Falush, Stephens, & Pritchard, 2003), typically assume hypothetical *K* populations characterized by a set of alleles with frequency *p* at locus *l*, and seek to probabilistically assign individuals to each of these populations given their genotypes. When allowing for admixture, an additional parameter *Q* models the proportion of each individual's genome that originated from a given population. While other programs differ from STRUCTURE in using variational inference (FASTSTRUCTURE: Raj, Stephens, & Pritchard, 2014) or a maximum-likelihood framework (ADMIXTURE: Alexander, Novembre, & Lange, 2009; FRAPPE: Tang, Coram, Wang, Zhu, & Risch, 2006), they are united in proposing an explicit causal model for input data, assuming linkage equilibrium between loci and Hardy–Weinberg equilibrium between alleles. In contrast, nonparametric methods such as principal components (PCA) analysis and *K*-means clustering (Jombart, Devillard, & Balloux, 2010; Novembre et al., 2008) first reduce the dimensionality of an allele frequency matrix and then seek to identify groups of individuals that minimize an objective function without explicitly modelling the attributes of genetic data.

Because of these differences, parametric and nonparametric approaches may show different sensitivities to SFS generated through biased data collection methods. It is possible that these sensitivities also reflect the influence of the type of data sets available during each program's initial development: for example, as STRUCTURE's underlying algorithm was tested prior to widespread adoption of high-throughput sequencing methods and was initially applied to microsatellite data screened for appropriate frequency distributions (Li, Korol, Fahima, Beiles, & Nevo, 2002; Pritchard et al., 2000), the characteristics of unfiltered modern SNP data sets may present unanticipated challenges to accurate population genetic inference.

Here, we build on previous studies of the relationship between population subdivision and allele frequencies (Gompert et al., 2014; Jakobsson et al., 2013; Mathieseon & McVean, 2012; Mathieson & McVean, 2014) to systematically assess the influence of MAF thresholds on inference of population structure. We evaluate the ability of model-based and nonparametric clustering methods to describe population structure in both simulated and empirical genomic data sets using diallelic SNPs and find that STRUCTURE is confounded by singletons and that both approaches are sensitive to variation in MAF thresholds. We propose a simple hypothesis to explain this behaviour and recommend a set of best practices for researchers seeking to describe population structure using multilocus data sets.

2 | METHODS

2.1 | Simulated data

We simulated genome-wide SNP data sets under a custom demographic model in FASTSIMCOAL2 version 2.5.2.21 (Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013) in order to assess the impacts of MAF filtering on population structure inference in the absence of sequencing or assembly error. Model parameters were chosen to reflect a plausible demographic history for our empirical case (see below), with one population experiencing successive splits 60,000 and 40,000 generations in the past after which all populations

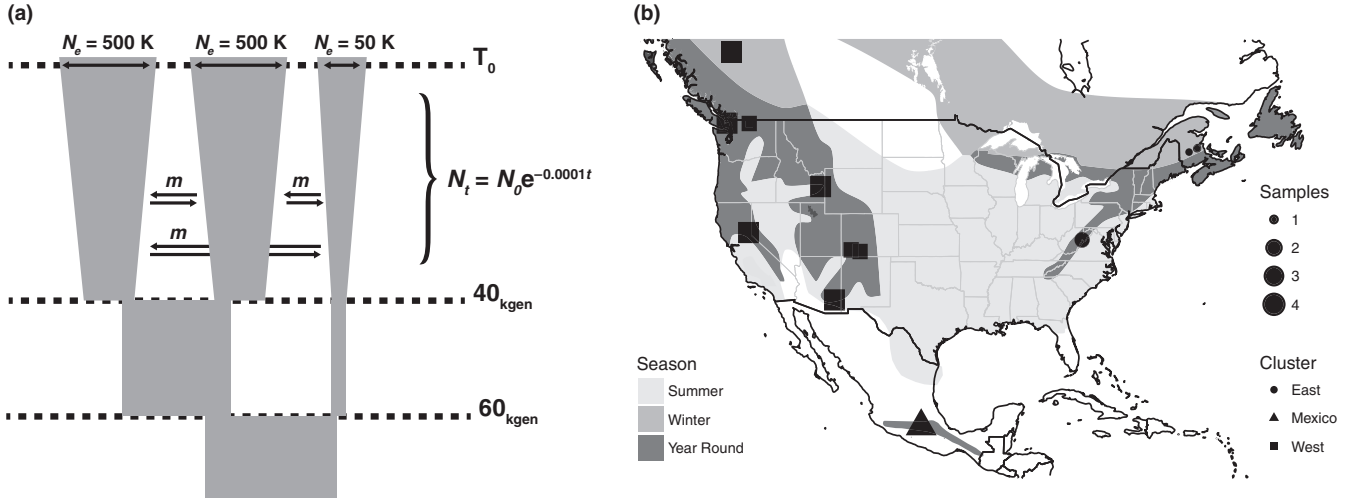

FIGURE 1 (a) The demographic model used in simulating SNP data sets. (b) Sampling localities and sizes for *Regulus satrapa*, with a priori population assignments

increase in size exponentially, reaching a final effectice population size N_e of 50,000 for the "outgroup" lineage and 500,000 for the remaining populations (Figure 1a). Migration is allowed among all populations after the final divergence event with a rate of $m = 5 \times 10^{-5}$. We included a mutation rate parameter of 2×10^{-6} in simulated data—equivalent to selecting a single SNP from a 200-bp region in an organism with an average genome-wide mutation rate of 1×10^{-8} (see FASTSIMCOAL2 user manual). Missing data—a common feature of reduced-representation library SNP datasets—were simulated by randomly dropping 25% of the alleles at each simulated locus.

We generated 10 independent simulations using the same starting parameter values and replicated analyses 10 times for each data set. Each simulation was initialized with 5,000 loci across 10 individuals in each of the three populations. After converting FASTSIMCOAL2 output to STRUCTURE's input file format, we used a custom R script to apply the following MAF cutoffs to all populations in our simulated data: 1/60, 2/60, 3/60, 4/60, 5/60, 8/60, 11/60 and 15/60.

To test whether variation in inferred admixture levels was caused by MAF thresholds specifically rather than a drop in the total size of the data matrix after filtering, we reran the above simulations but initialized with 40,000 loci and then randomly down-sampled all alignments to 1,000 sites after applying MAF cutoffs.

2.2 | Empirical data

We collected genome-wide SNP data from 40 individuals of the widespread North American passerine *Regulus satrapa*, the golden crowned kinglet. Our geographical sampling aimed to represent three areas of the species' breeding range that a previous study with mitochondrial DNA (mtDNA) suggested were distinct populations (J. Klicka, unpublished data): subspecies *satrapa* in the Eastern US/Canada; subspecies *olivaceous/apache* in the coastal and Rocky Mountains US/Canada, respectively; and subspecies *azteca* in the Sierra Madre del Sur and Transvolcanic Belt of Mexico (Figure 1b). We extracted whole genomic DNA using Qiagen DNEasy extraction kits and prepared reduced-representation libraries via the double digest (dd)RADseq protocol (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012) using the digestion enzymes *Sbf*1 and *Msp*1 and a size-selection window of 415–515 bp. We sequenced the resulting libraries for 50-bp single-end reads on an Illumina HiSeq 2500 sequencer.

We assembled reads into sequence alignments de novo using the program IPYRAD version 0.7.11 (https://github.com/dereneaton/ipyrad). We set a similarity threshold of 0.88 for clustering reads within and between individuals, a minimum coverage depth of 6 per individual, and a maximum depth of 10,000. We filtered out loci sharing a heterozygous site in 50% of samples as probable clusters of paralogues with a fixed difference. (We define "locus" in the context of ddRADseq data as a cluster of sequence reads putatively representing the same 50-bp region downstream of an *Sbf*1 cut site.) Because missing data can have a strong influence on population genetic inference (Arnold et al., 2013, Gautier et al., 2013) and preliminary exploration suggested anomalous clustering behaviour, we removed seven individuals from our data set prior to all downstream analysis. Of these final 33 samples, we required each locus to be sequenced in at least half of samples and randomly selected one SNP per locus. We then applied the same set of MAF cutoffs as used in our simulation study to all populations (1/60 to 15/60).

2.3 | Population structure analyses

We ran 10 replicate clustering analyses using STRUCTURE version 2.3.4 for all MAF filters of simulated ($n = 80$) and empirical data ($n = 8$) using the correlated allele frequency model with admixture for 250,000 generations each, setting $K = 3$ and discarding the initial 10,000 generations as burn-in. All runs were initialized using a random seed value drawn from a uniform distribution with range 0–10,000. No prior population assignment information was included in the model. All other settings were left at default values.

PCA, *K*-means clustering and discriminant analysis of principal components (DAPC) were conducted using the R package ADEGENET

version 2.1.0 (Jombart et al., 2010) and the MAF-filtered STRUCTURE files as input. Missing data were replaced with the mean values across the full sample before running PCAs. All PCs were retained for *K*-means clustering, as recommended by the ADEGENET documentation (https://github.com/thibautjombart/adegenet/wiki/Tutorials), which we performed with a fixed value of $K = 3$ on the basis of the apparent level of subdivision in preliminary mtDNA data (J. Klicka et al., unpublished data). DAPCs were initialized using the *K*-means clustering solution and tested by training the model on half the individuals in each population, then predicting the population assignment of the remaining individuals. We retained three PCs and two discriminant axes after manually examining several runs with both simulated and empirical data. PCA and *K*-means analyses were repeated 10 times per input data set, and DAPC cross validations were repeated 10 times per *K*-means replicate.

In practice most clustering solutions are assessed visually by comparing bar plots of STRUCTURE output or scatter plots of PCs 1 and 2. To quantitatively compare clustering results across methods and MAF cutoffs, we estimated two summary statistics: the proportion of correct population assignments, and the ratio of distances between individuals within populations to those between all individuals (we refer to this as "PC_{ST}" in analogy to F_{ST} and ϕ_{ST}). The proportion of correct population assignments was estimated by assigning each individual to a single cluster (for STRUCTURE results individuals were assigned to the cluster with the highest *q* value), swapping cluster labels to account for stochastic label switching during inference, and comparing inferred and true population assignments. Within-to-total population distance ratios were calculated as:

$$PC_{ST} = 1 - \frac{1}{k}\sum_{k}\frac{\bar{d}_{k_i k_j}}{\bar{d}_{ij}}.$$

where k is the population index, i and j are the indices of individuals, and $\bar{d}$ is the mean Euclidean distance between individuals in a k-dimensional space described by the first k principal components or the columns of the q matrix returned by STRUCTURE. More simply, this ratio is the average distance between individuals in the same population over the average distance between all individuals. High values indicate that inferred clusters are discrete, while low values indicate that clusters overlap—reflecting either uncertainty in individual assignments or admixture among populations. We fitted linear mixed models in R version 3.5.1 (R Core Team, 2018) to evaluate the relationship between MAF cutoffs and the values of assignment accuracy and PC_{ST}, using simulation number as a random effect to account for the nonindependence of replicated analyses on the same data set. We visualized results using ggplot2 version 3.1.0 (Wickham, 2016) and ggridges version 0.5.1 (Wilke, 2018).

3 | RESULTS

3.1 | Simulations and sequence assembly

Following MAF filtering, our simulated data sets retained an average range of 3,942 (for MAF = 1) to 242 (for MAF = 15) loci. Constant-length data sets were always subsampled to 1,000 bp. For our *Regulus satrapa* ddRAD libraries, Illumina sequencing returned an average of 781,011 quality-filtered reads per sample. Clustering within individuals identified 35,722 putative loci per sample, with an average depth of coverage of 22×. After clustering across individuals and applying paralogue and depth-of-coverage filters, we retained an average of 4,286 loci per sample. Prior to applying MAF filters and removing individuals for excess missing

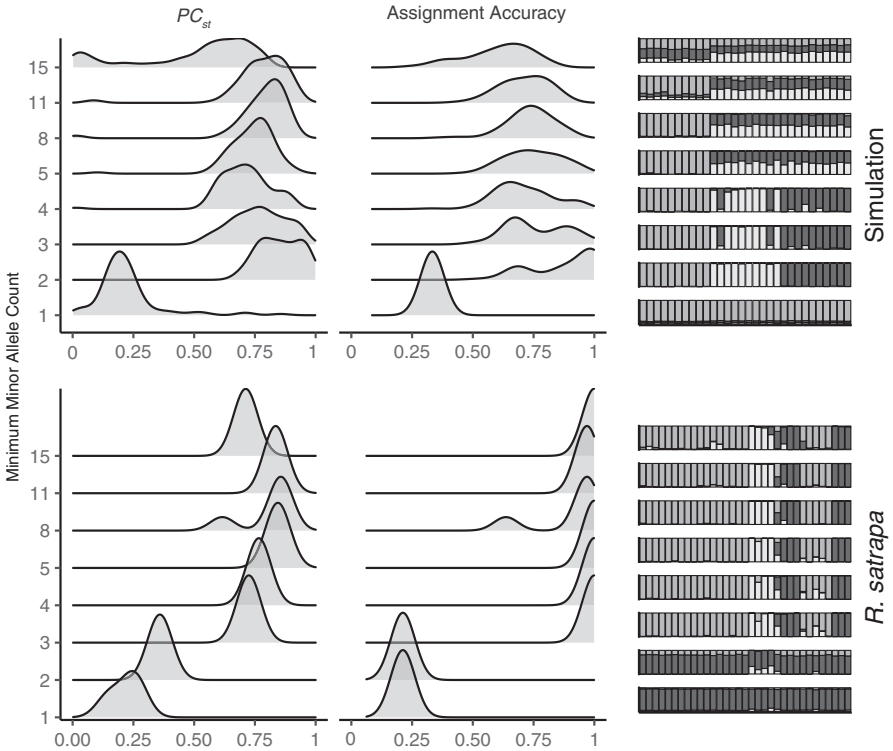

FIGURE 2 The influence of minor allele count on STRUCTURE's assignment accuracy under the admixture model, and PC_{ST} for simulated and empirical data sets

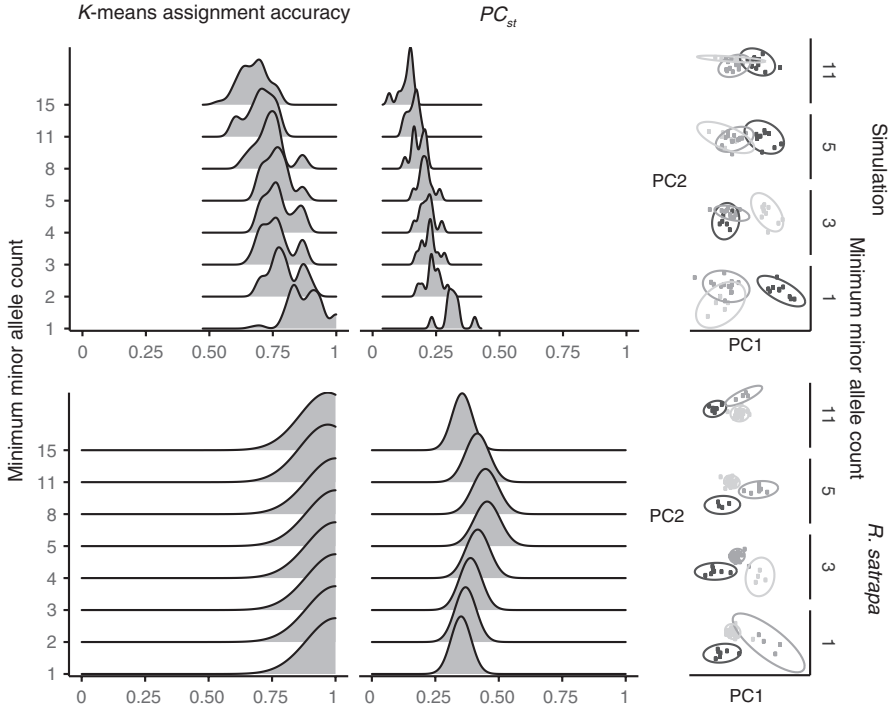

FIGURE 3 The influence of minor allele count *K*-means assignment and PC_{ST} for simulated and empirical data sets. On PCA plots, *x*-axis values are PC1 and *y*-axis values are PC2

data, our alignment included 3,898 unlinked diallelic SNPs that were sequenced in at least 30 of the original 40 samples. Our final MAF-filtered data sets ranged from 3,419 (MAF = 1) to 431 (MAF = 20) loci.

3.2 | Parametric clustering

The ability to detect population subdivision in both simulated and empirical data sets varied widely across MAF thresholds using the model-based method STRUCTURE (Figure 2). In both constant- and variable-length data sets, including singletons caused STRUCTURE to assign all individuals to the same majority ancestry cluster. For variable-length simulated data sets, after excluding alignments with singletons, higher MAF thresholds are also associated with lower population discrimination (PC_{ST} ~ minor_allele_count * sim_num; $p < 2e{-}16$, $R^2 = 0.207$, $df = 696$; Supporting Information Figure S1) and assignment accuracy (accuracy ~minor_allele_count * sim_num; $p < 2e{-}16$, $R^2 = 0.195$, $df = 696$; Supporting Information Figure S1). The association between high MAF cutoffs and population discrimination is reversed in constant-length data sets—more stringently filtered data sets infer more discrete clusters—although the effect is much weaker (PC_{ST} ~ minor_allele_count * sim_num: $p < 2e{-}16$, $R^2 = 0.098$, $df = 696$; accuracy ~minor_allele_count * sim_num: $p < 0.005$, $R^2 = 0.01$, $df = 696$; Supporting Information Figures S1 and S2).

3.3 | Nonparametric clustering

In contrast to STRUCTURE, both *K*-means clustering accuracy and PC_{ST} were robust to inclusion of singletons. However, both measures were highly sensitive to MAF thresholds in simulated data (Figure 3). Both PC_{ST} and *K*-means assignment accuracy decline as the MAF threshold is increased (PC_{ST} ~ minor_allele_count * sim_num; $p < 2e{-}16$, $R^2 = 0.642$, $df = 796$; kmeans_accuracy ~minor_allele_count * sim_num; $p < 2e{-}16$, $R^2 = 0.409$, $df = 796$; Supporting Information Figure S3). As with STRUCTURE these relationships are reversed but weaker when alignment length is held constant (PC_{ST} ~ minor_allele_count * sim_num; $p < 2e{-}16$, $R^2 = 0.246$, $df = 796$; kmeans_accuracy ~minor_allele_count * sim_num; $p < 2e{-}16$, $R^2 = 0.116$, $df = 796$; Supporting Information Figure S3), although the relationship remains negative across MAF cutoffs in the range of 1/60–3/60 (Supporting Information Figure S4). For empirical data, both methods achieved near-perfect assignment accuracy under all MAF cutoffs (Figure 3).

4 | DISCUSSION

4.1 | Inference of population structure is sensitive to MAF

Our results demonstrate that inference of population structure can be strongly influenced by choice of MAF threshold with both model-based and multivariate approaches. STRUCTURE fails to detect even moderate population subdivision ($F_{ST} \cong 0.05$) when singletons are included in the alignment, and both methods generally infer increasing levels of admixture as the minimum MAF of sites included in the alignment is increased. These trends do not occur when alignment length is held constant, suggesting that most of the effect is driven by a drop in the total size of the data matrix after filtering by MAF. In practice this will occur in most empirical data sets when genotypes are estimated from sequencing data. For chip-based approaches in which SNPs are first screened for variation at some cutoff, our

analysis suggests that clustering results should be relatively robust to implicit MAF cutoffs applied during chip design. This is a particularly important concern for ancient DNA studies, which frequently collect data with these methods (e.g., Rasmussen et al., 2011).

Two factors may explain the pattern of increased admixture in more stringently filtered data sets: variation in the total size of the data matrix, and the distribution of mutations on a coalescent tree. In simulated data sets of varying size (as in nearly all empirical cases), increasing the MAF cutoff decreases the total size of the data matrix and leads to much higher estimates of individual admixture. This is in part an interpretive issue, as the strong effect of the size of the data matrix suggests that the high q-matrix values reflect uncertainty in individual assignments rather than higher admixture levels. However, because parametric approaches are typically interpreted in light of their generative model, many users are likely to see this pattern as evidence of higher gene flow.

A secondary cause of increased admixture in more stringently filtered data sets is the time distribution of mutations in a coalescent tree. Under a standard coalescent model the expected number of sites with a derived allele present in i samples (s_i) is the total length of branches subtending i descendants (τ_i), multiplied by the expected number of mutations per unit time $\left(\frac{\theta}{2}\right)$:

$$E[s_i] = \frac{\theta}{2} E[\tau_i] \quad \text{(Wakeley, 2009, equation 4.15)}$$

Low-frequency alleles represent mutations that occurred on branches with few descendants, and these branches are typically found close to the present (Figure 4; see Appendix S1 for simulation details). They therefore contain a disproportionate amount of information about recent events. Removing them is similar to drawing a horizontal line across a coalescent tree and dropping mutations that occur beyond that line. In the absence of recent pulses of gene flow (where ancient alleles from a donor are rare in the recipient and thus confound the relationship between frequency and mutation age), this "pruning" process causes populations to appear less differentiated as the MAF threshold increases, seen in PCA output as reduced distance between clusters and in STRUCTURE output as increased admixture within individuals (although some would argue that this is simply a misinterpretation of STRUCTURE's output, e.g., Lawson, van Dorp, & Falush, 2018). In the presence of recent pulses of gene flow, the true signal of admixture is instead replaced with inferred admixture (or reduced distance between clusters) as a function of a loss in information content, to uncertain effect.

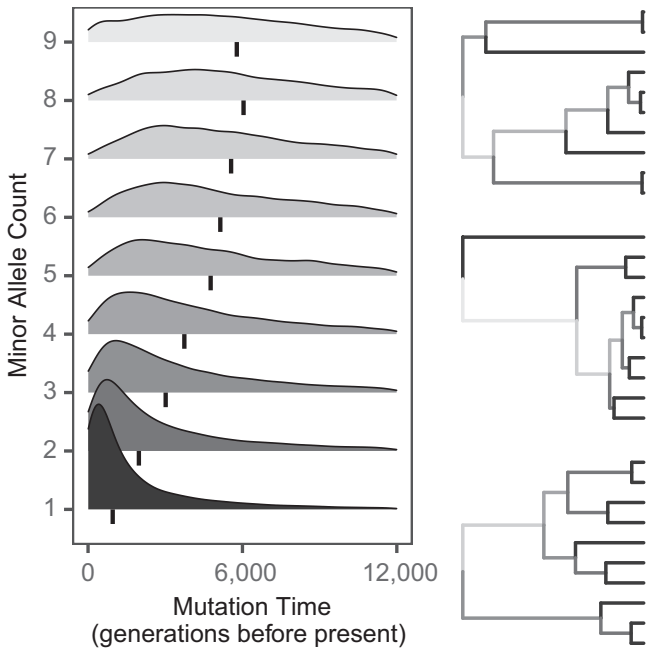

FIGURE 4 Time distribution of mutations with varying derived allele counts

The failure of model-based analyses to recover a clear signal of population subdivision when singletons are included in the alignment is more difficult to explain. The issue appears to be related to overfitting as a result of either a high frequency of uninformative singletons or a high frequency of uninformative common alleles (Alexander & Lange, 2011). As a verbal model, this is intuitive: an allele found at a frequency of $1/2N$ lacks information on broader patterns of population structure because it only serves to distinguish a single individual from all others, while a common allele found may be uninformative because of the absence of differences in its frequency across populations. We hypothesize that under STRUCTURE's algorithm, a population k_1 is assigned a site frequency spectrum that averages out true population specific-frequencies of common alleles, resulting in the broad band of majority ancestry visible in Figure 2. Subsequently, populations k_2, ..., k_n are assigned site frequency spectra characterized by high frequencies of singletons or other rare alleles, resulting in the additional bands of minority ancestry shared across all individuals. With our simulated data, rare but nonsingleton alleles reflect fine population structure and thus harm inference when excluded; with our empirical data, these rare alleles are uninformative and serve only as noise to obscure the common allele frequency distributions reflecting true population history.

This hypothesis is consistent with a pathology related to STRUCTURE's inability to model mutation of modern alleles, previously identified as a potential obstacle to accurate inference of population structure under certain histories (Shringapure & Xing, 2009). Because STRUCTURE assumes each unique allele in the input data set has a distinct frequency in its parent population, recent mutations (e.g., derived alleles) are erroneously treated as representative of a separate population-specific allele frequency profile rather than as descendants of ancestral copies. If a sufficient number of singletons are present in the data set, the noise from these false allele frequency profiles may mask the signal from alleles indicative of "true" populations. Although most multivariate analyses were robust to inclusion of singletons, a similar pattern of low accuracy and population discrimination was observed in PCA when alignment length was held constant—probably because low-frequency alleles hold less information about intergroup differences than moderate-frequency alleles, and low-frequency alleles will be a larger proportion of the total data matrix in this case.

4.2 | Recommendations for setting MAF thresholds in population genetic studies

Our results suggest that SFS distributions that can cause STRUCTURE and other model-based programs to erroneously fail to detect structure that may be generated by either normal demographic processes (e.g., exponential population growth with relatively recent divergence, as in our simulated example) or assembly errors (potentially present in our empirical example, and well documented in other de novo RADseq data sets, e.g., Shafer et al., 2017). As a consequence, a broad set of empirical studies may be affected. We recommend researchers using model-based programs to describe population structure observe the following best practices: (a) duplicate analyses with nonparametric methods such as PCA and DAPC with cross validation; (b) exclude singletons; and (c) compare alignments with multiple assembly parameters. When seeking to exclude only singletons in alignments with missing data (a ubiquitous problem for reduced-representation library preparation methods), it is preferable to filter by the count (rather than frequency) of the minor allele, because variation in the amount of missing data across an alignment will cause a static frequency cutoff to remove different SFS classes at different sites. The scripts used to filter STRUCTURE input files for this manuscript are available at https://github.com/cjbattey/LinckBattey2017_MAF_clustering.

4.3 | Population genetics of *Regulus satrapa*

Although describing population structure and phylogeographical patterns of the golden-crowned kinglet was not the primarily goal of our study and will be elaborated on elsewhere, our data provide novel evidence for deep splits across the range of the species, corroborating previous mtDNA evidence (J. Klicka, unpublished data). Curiously, the results of our model-based population structure inference suggest not only singletons but all rare alleles (MAF ≤ 8/80) have a high noise to signal ratio, while common alleles (MAF ≥ 10/80) accurately reflect expected relationships. This pattern may be driven by either purifying selection eliminating geographically localized variants (Jackson, Campos, & Zeng, 2015; Nelson et al., 2012), a population bottleneck (Gattepaille, Jakobsson, & Blum, 2013; Nei, Maruyama, & Chakraborty, 1975), a burst of recent migration following exponential population growth (Slatkin, 1985), or assembly artefacts resulting in a high proportion of uninformative/erroneous sites (Shafer et al., 2017). While all scenarios are probably contributing to some extent, studies of genetic variation in similar taxa provide support for post-Pleistocene expansion and gene flow among populations separated by ice sheets (Spellman & Klicka, 2006), processes that may result in similar SFS distributions to our example.

4.4 | Future directions

With simulated and empirical cases reflecting similar (if nonidentical) site frequency spectra, our focus was on a necessarily narrow range of demographic scenarios and a relatively narrow range of SFS distributions. Future examinations of the sensitivity of population genetic inference to MAF thresholds with data sets simulated under a diversity of evolutionary histories may shed light on the biological processes generating problematic SFS, and lead to the development of more robust model-based programs. While other parametric population structure inference programs share STRUCTURE's underlying model and we believe the broad patterns reported here will be similarly reflected, differences in implementation (e.g., Markov chain Monte Carlo mixing) may shape specific sensitivities. A broader survey of model-based population structure inference methods will help to clarify which approaches are best suited to next-generation sequencing data, and lead to the development of more robust software for describing the fundamental units of biological organization.

ACKNOWLEDGEMENTS

This research was supported by a National Defense Science and Engineering Graduate (NDSEG) Fellowship to E.B.L. Comments from Alex Buerkle, an anonymous reviewer, and numerous readers of a preprint version of this article significantly improved its contents. We especially thank Dave Slager for his early contributions to our thinking on this problem.

AUTHOR CONTRIBUTIONS

EBL and CJB jointly conceived and designed the study, collected and analyzed data, and drafted the manuscript

DATA ACCESSIBILITY

Simulation results are available from the Dryad Digital Repository, https://doi.org/10.5061/dryad.hr1hh75. Raw sequence data are available from the NCBI SRA, accession PRJNA514868. Code used in the study is available via GitHub: https://github.com/cjbattey/LinckBattey2017_MAF_clustering.

ORCID

Ethan Linck https://orcid.org/0000-0002-9055-6664
C. J. Battey https://orcid.org/0000-0002-9958-4282

REFERENCES

Alexander, D. H., & Lange, K. (2011). Enhancements to the admixture algorithm for individual ancestry estimation. *BMC Bioinformatics*, *12*(246), 1–6. https://doi.org/10.1186/1471-2105-12-246

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*, 1655–1664. https://doi.org/10.1101/gr.094052.109

Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to

nonrandom haplotype sampling. *Molecular Ecology*, *22*, 3179–3190. https://doi.org/10.1111/mec.12276

Barton, N. H., & Slatkin, M. (1986). A Quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. *Heredity*, *56*, 409–415. https://doi.org/10.1038/hdy.1986.63

Blanco-Bercial, L., & Bucklin, A. (2016). New view of population genetics of zooplankton: RAD-seq analysis reveals population structure of the North Atlantic planktonic copepod Centropages typicus. *Molecular Ecology*, *25*, 1566–1580.

Catchen, J. C., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*, 3124–3140. https://doi.org/10.1111/mec.12354

Cruz, O., & Raska, P. (2014). Population structure at different minor allele frequency levels. *BMC Proceedings*, *8*, S55.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., & Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, *25*, 3207–3212. https://doi.org/10.1093/bioinformatics/btp579

Eaton, D. A. R. (2014). PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, *30*, 1844–1849. https://doi.org/10.1093/bioinformatics/btu121

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLoS Genetics*, *9*, e1003905. https://doi.org/10.1371/journal.pgen.1003905

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*, 717–726. https://doi.org/10.1093/sysbio/sys004

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*, 1567–1587.

Fay, J. C., Wyckoff, G. J., & Wu, C.-I. (2001). Positive and negative selection on the human genome. *Genetics*, *158*, 1227–1234.

Gattepaille, L. M., Jakobsson, M., & Blum, M. G. (2013). Inferring population size changes with sequence and SNP data: Lessons from human bottlenecks. *Heredity*, *110*, 409–419. https://doi.org/10.1038/hdy.2012.120

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., … Estoup, A. (2012). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, *22*, 3165–3178. https://doi.org/10.1111/mec.12089

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., & Pudlo,P. … Estoup,A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, *22*, 3165–3178.

Gompert, Z., Lucas, L., Buerkle, C. A., Forister, M. L., Fordyce, J. A., & Nice, C. C. (2014). Admixture and the organization of genetic diversity in a butterfly species complex revealed through common and rare genetic variants. *Molecular Ecology*, *23*, 4555–4573. https://doi.org/10.1111/mec.12811

Griffiths, R. C., & Tavaré, S. (1999). The ages of mutations in gene trees. *The Annals of Applied Probability*, *9*, 567–590. https://doi.org/10.1214/aoap/1029962804

Jackson, B. C., Campos, J. L., & Zeng, K. (2015). The effects of purifying selection on patterns of genetic differentiation between *Drosophila melanogaster* populations. *Heredity*, *114*, 163–174. https://doi.org/10.1038/hdy.2014.80

Jakobsson, M., Edge, M. D., & Rosenberg, N. A. (2013). The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics*, *193*, 515–528.

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*, 1–15. https://doi.org/10.1186/1471-2156-11-94

Kimura, M., & Ohta, T. (1973). The age of a neutral mutant persisting in a finite population. *Genetics*, *75*, 199–212.

Lawson, D., van Dorp, L., & Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, *9*, 3258–4000. https://doi.org/10.1038/s41467-018-05257-7

Li, Y. C., Korol, A. B., Fahima, T., Beiles, A., & Nevo, E. (2002). Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. *Molecular Ecology*, *11*, 2453–2465. https://doi.org/10.1046/j.1365-294X.2002.01643.x

Marth, G. T., Czabarka, E., Murvai, J., & Sherry, S. T. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, *166*, 351–372. https://doi.org/10.1534/genetics.166.1.351

Mathieson, I., & McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, *44*, 243–246. https://doi.org/10.1038/ng.1074

Mathieson, I., & McVean, G. (2014). Demography and the age of rare variants. *PLoS Genetics*, *10*, e1004528. https://doi.org/10.1371/journal.pgen.1004528

Nei, M., Maruyama, T., & Chakraborty, R. (1975). The bottleneck effect and genetic variability in populations. *Evolution*, *29*, 1–10. https://doi.org/10.1111/j.1558-5646.1975.tb00807.x

Nelson, M., Wegmann, D., Ehm, M., Kessner, D., St. Jean, P., Verzilli, C., … Mooser, V. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, *337*, 100–104. https://doi.org/10.1126/science.1217876

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. https://doi.org/10.1371/journal.pone.0037558

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., … Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*, 98–101. https://doi.org/10.1038/nature07331

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, *7*, e37135. https://doi.org/10.1371/journal.pone.0037135

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *19*, 573–589. https://doi.org/10.1534/genetics.114.164350

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K., Rasmussen, S., Albrechtsen, A., … Willerslev, E. (2011). An aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, *334*(6052), 94–98. https://doi.org/10.1126/science.1211177

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Molecular Ecology and Evolution*, *8*, 907–917. https://doi.org/10.1111/2041-210X.12700

Shringapure, S., & Xing, E. P. (2009). mStruct: Inference of population structure in light of both genetic admixing and allele mutations. *Genetics*, *182*, 575–593. https://doi.org/10.1534/genetics.108.100222

Slatkin, M. (1985). Rare alleles as indicators of gene flow. *Evolution*, *39*, 53–65. https://doi.org/10.1111/j.1558-5646.1985.tb04079.x

Spellman, G. M., & Klicka, J. (2006). Testing hypotheses of Pleistocene population history using coalescent simulations: Phylogeography of the pygmy nuthatch (*Sitta pygmaea*). *Proceedings of the Royal Society of London B: Biological Sciences*, *273*, 3057–3063.

Tang, H., Coram, M., Wang, P., Zhu, X., & Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, *79*, 1–12. https://doi.org/10.1086/504302

Wakeley, J. (2009). *Coalescent theory: An introduction*. Greenwood Village, CO: Roberts & Company Publishers.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.

Wilke, C. O. (2018). ggridges: Ridgeline Plots in 'ggplot2'. R package version 0.5.0.

Winger, B. M. (2017). Consequences of divergence and introgression for speciation in Andean cloud forest birds. *Evolution*, *71*, 1815–1831. https://doi.org/10.1111/evo.13251

SUPPORTING INFORMATION

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

How to cite this article: Linck E, Battey CJ. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Mol Ecol Resour*. 2019;19:639–647. https://doi.org/10.1111/1755-0998.12995