Version dated: September 13, 2017

Minor allele frequency thresholds strongly affect population structure inference with genomic datasets

ETHAN B. LINCK[1] AND C.J. BATTEY[1]

[1] *Department of Biology & Burke Museum of Natural History & Culture, University of Washington, Seattle, WA, 918195, USA*

Corresponding author: Ethan B. Linck, Department of Biology & Burke Museum of Natural History & Culture, University of Washington, Seattle, WA, 918195, USA; E-mail: elinck@uw.edu.

Abstract.—

Across the genome, the effects of different evolutionary processes and historical events can result in different classes of genetic variants (or alleles) characterized by their relative frequency in a given population. As a result, population genetic inference can be strongly affected by biases in laboratory and bioinformatics treatments that affect the site frequence spectrum, or SFS. Yet despite the widespread use of reduced-representation genomic datasets with nonmodel organisms, the potential consequences of these biases for downstream analyses remain poorly examined. Here, we assess the influence of minor allele frequency (MAF) thresholds implemented during variant detection on inference of population structure. We use simulated and empirical datasets to evaluate the effect of

MAF thresholds on the ability to discriminate among populations and quantify admixture with both model-based and non-model-based clustering methods. We find model-based inference of population structure is highly sensitive to choice of MAF, and may be confounded by either including singletons or excluding all rare alleles. In contrast, non-model-based clustering is largely robust to MAF choice. Our results suggest that model-based inference of population structure can fail due to either natural demographic processes or assembly artifacts, with broad consequences for phylogeographic and population genetic studies. We propose a simple hypothesis to explain this behavior and recommend a set of best practices for researchers seeking to describe population structure using reduced-representation libraries.
(Keywords: MAF; population structure; population genetic inference; *structure*; clustering; genomics; bioinformatics)

The distribution of genetic variation within and among individuals is the crucial to understanding the organization of biological diversity and its underlying causes. Across the genome, the effects of different evolutionary processes and historical events can result in different classes of genetic variants (or alleles) characterized by their relative frequency in a given population. An excess of common alleles may reflect the signature of population bottlenecks (Marth et al. 2004), purifying selection (Fay et al. 2001), or the absence of population subdivision (Pritchard et al. 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 and Slatkin 1986; Gompert et al. 2014). Because the distribution of allele frequencies (also known as the site frequency spectrum, or SFS) therefore reflects the unique combination of these varied factors, downstream analyses are sensitive to the influence of sampling methodologies. Yet despite the explosive recent growth of population genetics provided by advent of affordable reduced-representation genome sequencing for nonmodel organisms, there remain significant gaps in our knowledge of how data collection biases population genetic inference.

These biases may originate either in wet lab and bioinformatic treatments. Prior to sequencing, the SFS may be shaped by ascertainment bias in library preparation: RADseq-style methods introduce genealogical biases (Arnold et al. 2013) and nonrandom patterns of missing data (Gautier et al. 2013) due to reliance on the presence of restriction cut sites, while hybridization capture with ultraconserved element (UCE) probesets necessarily involves targeting sites highly-conserved across evolutionarily distant taxa (Faircloth et al. 2012). During sequencing itself, relatively high error rates are accepted in individual reads, under the assumption they will be corrected during bioinformatic processing steps (Nielsen et al. 2012). However, the absence of standard bioinformatic pipelines in ecology and evolutionary biology is itself a source of uncertainty (Shafer et al. 2016) 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. 2016). 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 single nucleotide polymorphisms or 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. But despite its potential importance, the two most popular comprehensive bioinformatic pipelines for RADseq data alternatively include (Catchen et al. 2013) or exclude (Eaton 2014) the option to set minor allele frequency thresholds during variant calling, with the result that among empirical studies, MAF choices are only sometimes reported (e.g., Winger 2017; Blanco-Bercial and 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. Broadly speaking, methods to detect population structure fall into two categories: model-based (or parametric) approaches, and nonparamatric approaches. Model-based methods, exemplified by the influential program *structure* (Pritchard et al. 2000), typically assume a hypothetical K populations characterized by P allele frequencies at a set of loci L, and seek to probabilistically assign individuals to each of these populations given their genotypes. When allowing for admixture, an additional parameter Q models proportion of each individual's genome that originated from a given population. While other programs differ from *structure* in using variational inference (*fastSTRUCTURE*; Raj et al. 2014) or a maximum likelihood framework (e.g. *ADMIXTURE* or *FRAPPE*; Alexander et al. 2009; Tang et al. 2006), they are united in proposing an explicit generative model for input data, assuming linkage equilibrium between loci and Hardy Weinberg equilibrium between alleles. In contrast, nonparametric methods such as principal components analysis and k-means clustering (Jombart 2008; 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's possible these sensitivities also reflect the influence of the type datasets 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 initially applied on microsatellite data screened for appropriate frequency distributions (Pritchard et al. 2000; Li et al. 2002), the characteristics of unfiltered modern SNP datasets may present unanticipated challenges to accurate population genetic inference. Yet to the best of our knowledge, no studies have directly addressed this potential source of error in population genetic and phylogeographic studies. Here, we assess the influence of minor allele frequency (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 datasets and find that model-based approaches are highly sensitive to the choice of MAF cutoff. We propose a simple hypothesis to explain this behavior and recommend a set of best practices for researchers seeking to describe population structure using reduced-representation libraries.

Methods

Simulated data.

We simulated genome-wide SNP datasets under a custom demographic model in *fastsimcoal2* (Excoffier et al. 2013) in order to assess the impacts of MAF filtering on population structure inference in the absence of sequencing or assembly error. The underlying demographic model was designed 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 increase in size exponentially, reaching a final N_e of 50,000 for the "outgroup" lineage and 500,000 for the remaining populations. Migration is allowed between all populations after the final divergence event. To avoid genealogical bias in the SFS of simulated SNP data, we included a mutation rate parameter of $2x10^{-6}$, equivalent to selecting a single SNP from a 200bp region in an organism with an average genome-wide mutation rate of $1x10^{-8}$ (see *fastsimcoal2* user manual). Missing data – a common feature of reduced-representation library SNP datasets – was simulated by randomly dropping 25% of the alleles at each simulated locus. We generated 100 independent simulations using the same starting parameter values. Each simulation was initialized with 5,000 loci across 10 individuals in each of the 3 populations. After converting *fastsimcoal2* output to *structure's* input file format, we generated MAF-filtered datasets at each of the following cutoffs: 1/60, 2/60, 3/60, 4/60, 5/60, 8/60, and 20/60.

Empirical data

We collected genome-wide SNP data from 40 individuals of the widespread North American passerine *Regulus satrapa,* the Golden-crowned Kinglet. Our geographic aimed to represent three areas of the species' breeding range a previous study with mitochondrial DNA suggested were distinct populations (Klicka 2017): subspecies *satrapa* in the Eastern US / Canada; subspecies *olivaceous* / *apache* in the coastal and Rocky Mountain US / Canada, respectively; and subspecies *azteca* in the Sierra Madre del Sur and Transvolcanic Belt of Mexico. We extracted whole genomic DNA using Qiagen DNEasy extraction kits and prepared reduced-representation libraries via the ddRADseq protocol (Peterson et al. 2012) using the digestion enzymes Sbf1 and Msp1 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. We assembled reads into sequence alignments de novo using the program ipyrad v. 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. To exclude paralogs from the final dataset, we

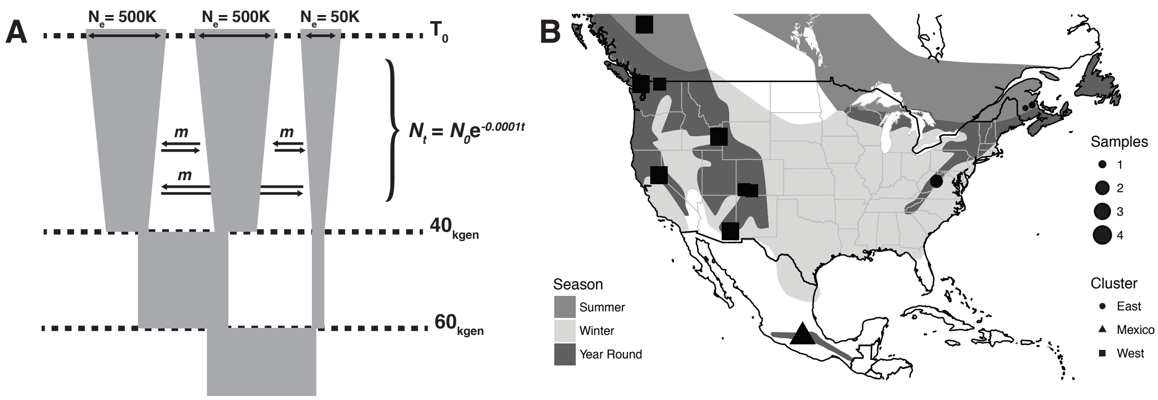

Figure 1: (A) The demographic model used in simulating SNP datasets. (B) Sampling localities and sizes for *Regulus satrapa*, with *a priori* population assignments.

filtered out loci sharing a heterozygous site in 50% of samples. We define "locus" throughout this manuscript as a cluster of sequence reads putatively representing the same 50-bp region downstream of an Sbf1 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 behavior, we removed 7 individuals from our dataset 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.

Population structure analyses

We ran alignments for all MAF filters of simulated (n=700) and empirical data (n=7) in structure using the correlated allele frequency model with admixture for 250,000 generations each, with 10,000 generations of burn-in. *Structure* was run once for each simulated dataset and 10 times for each empirical dataset, with all runs initialized using a random seed value drawn from a uniform distribution with range (0 - 10,000). To test whether variation in population assignments were due to issues related to *structure's* implementation and not representative of model-based methods more broadly, we also ran *fastSTRUCTURE* once for each simulated dataset once and 10 times for each empirical dataset. (While *fastSTRUCTURE* shares parameters with structure, it employs variational Bayesian inference for rapid computation rather than a Markov chain Monte Carlo approach to sample posterior distributions of global ancestry parameters.) We performed k-means clustering and discriminant analysis of principal components (DAPC) using the R package *adegenet* Jombart (2008) and the same datasets as input. For both nonparametric analyses, we performed 100 replicate cluster assignments on each alignment and calculated the percentage of correct population identifications, given K=3. For DAPC, we cross-validated population assignments by randomly selecting half the samples in each k-means cluster, conducting a DAPC on these samples, and predicting the group assignments of remaining individuals with the "trained" DAPC model. To summarize the sensitivity of each clustering method to MAF threshold, we performed an analysis of variance and a Tukey honest significant difference test in R Team (2017), and averaged differences across all MAF thresholds within each clustering method. We then scaled summary statistics describing population assignment accuracy to a range of 0-1 and asked if the distribution of differences in accuracy across MAF cutoffs differed between methods using a second Tukey test. These analyses were conducted separately for empirical and simulated datasets.

Results

Simulations and sequence assembly

Following MAF filtering, our simulated datasets retained an average range of 3942 (for MAF=1) to 242 (for MAF=20) loci. 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 22x. After clustering across individuals and applying paralog and depth-of-coverage filters, we retained an average of 4286 loci per sample. Prior to applying MAF filters and removing individuals for excess missing data, our alignment included 3898 unlinked diallelic SNPs from sequenced in at least 30 of the original 40 samples. Our final MAF-filered datasets ranged from 3419 (MAF=1) to 431 (MAF=20) loci. Site frequency spectra of simulated and empirical datasets were similar in decreasing in an approximately exponential manner as MAF increased, but with two notable differences. First, the SFS of our unfiltered simulated datasets featured greater proportion of singletons than our ddRAD data – 44.6% to 36.7% in a representative example. However, when excluding singletons, simulated datasets generally had fewer rare alleles overall, e.g. 12.6% to 16.9% and 6.8% to 8%, respectively (Figure S1).

Parametric clustering

The ability to detect population subdivision in both simulated and empirical datasets varied widely across MAF thresholds using the model-based method *structure* (Figure 2). Across 100 replicates of each of 7 MAF thresholds tested for simulated datasets, population discrimination (defined as the Euclidean distance between populations in Q-matrix space, indicating the proportion of an individual's ancestry from a given ancestral population) and the natural logarithm of the mixing parameter alpha (ln(alpha), indicating the relative level of admixture present in an individual's genome) were significantly different for all but one pairwise comparisons in a Tukey HSD test (adjusted p=0-0.06). Population discrimination and ln(alpha) were significantly different for 11/21 pairwise comparisons with empirical *Regulus satrapa* data (adjusted p=0-0.99). We observed broadly similar

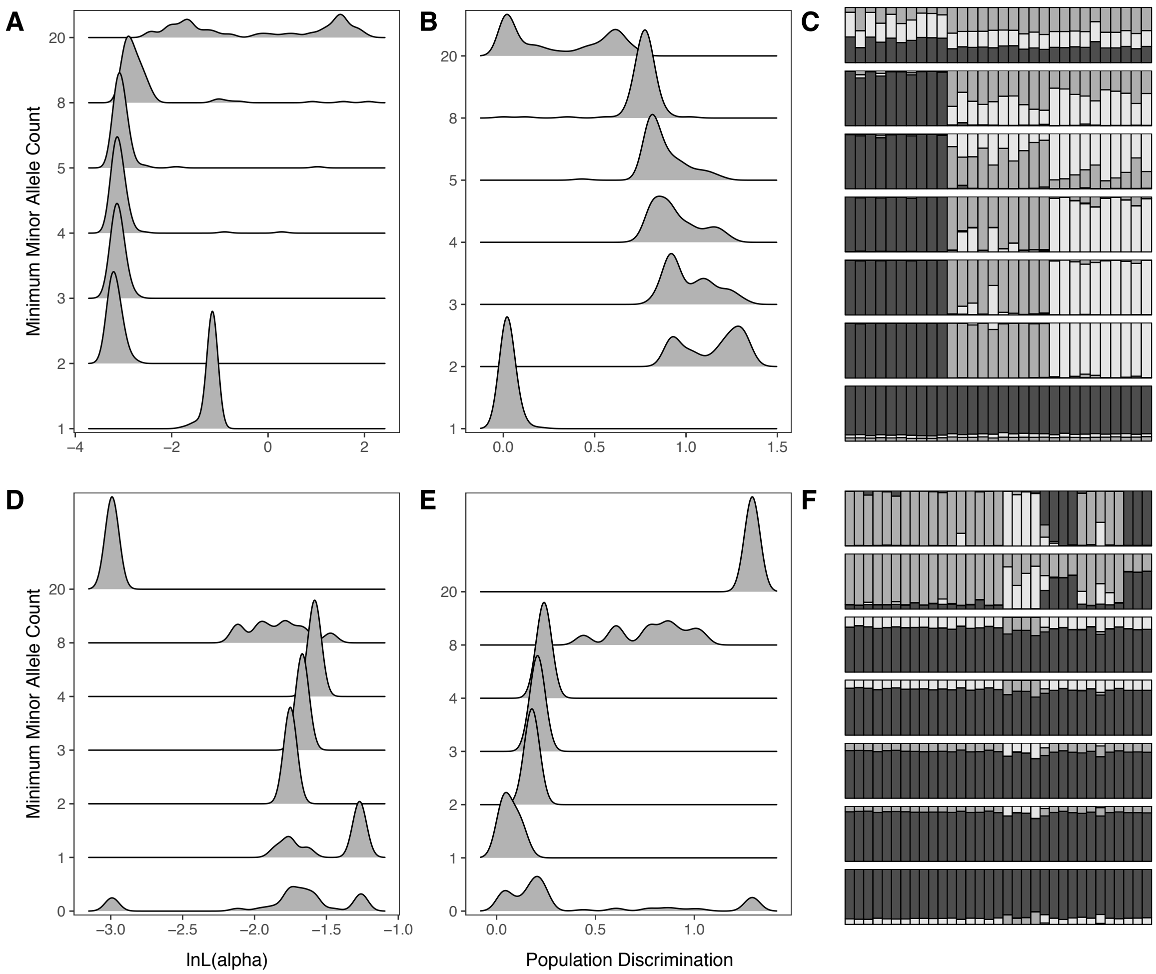

Figure 2: The influence of minor allele count on the natural log of *structure's* admixture parameter (A), population discrimination (B), and highest likelihood clustering plots (C) for simulated SNP datasets; identical results for empirical *Regulus satrapa* data shown below (D,E,F).

sensitivity in *fastSTRUCTURE* results, but with considerably less accuracy and ability to detect admixture (Figures S2 and S3).

Nonparametric clustering

In contrast to structure, both k-means and DAPC with cross-validation showed relatively little sensitivity to MAF threshold (Figure 3). In simulated alignments, all pairwise differences in accuracy across MAF cutoffs were statistically significant – likely due to large sample sizes – but were much smaller than found in structure (mean differences of 0.066 for k-means clustering versus 0.33 for structure population discrimination). For empirical data, both methods achieved near-perfect accuracy under all MAF cutoffs, resulting in no significant differences in accuracy. After scaling structure population discrimination to a range of (0,1), we found that the differences across MAF levels were significantly greater for *structure* (mean difference in population discrimination 0.33 for simulated data and 0.34 for empirical data) than in either k-means (0.066 simulated, 0 for empirical) or DAPC (0.078 simulated, 0 empirical; Figure 4, $p < 0.01$ for all comparisons).

Discussion

Parametric inference of population structure is sensitive to MAF

Our results demonstrate model-based inference of population structure can be strongly influenced by choice of MAF threshold. However, differences in patterns of detected subdivision across MAF values between our two datasets suggest that programs like *structure* may fail when input a variety of SFS distributions resulting from genome-wide sampling. Our simulations of a three-deme model with migration and exponential growth

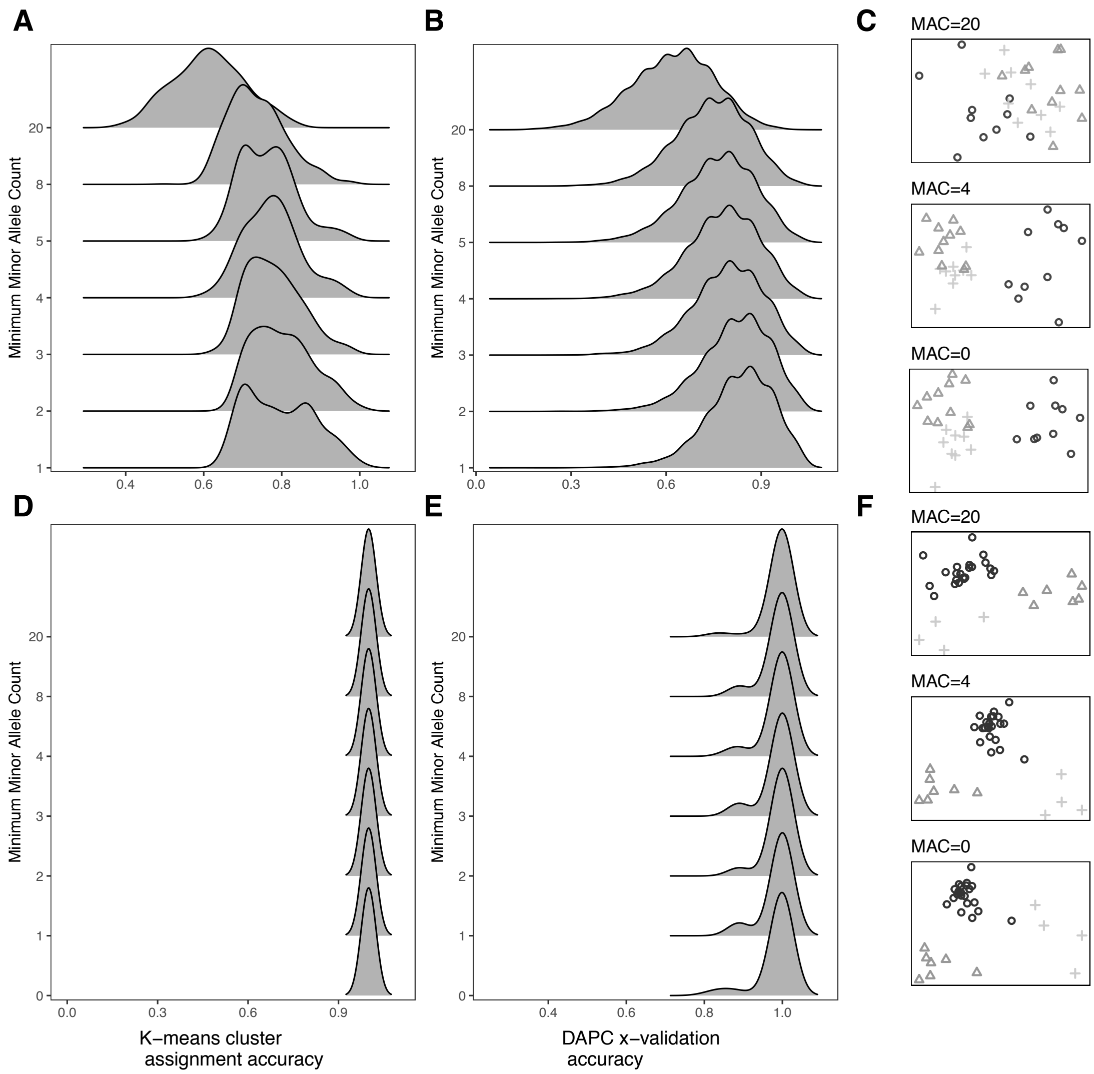

Figure 3: The influence of minor allele count on k-means assignment accuracy (A), DAPC cross-validation accuracy (B), and principal component analysis (C) for simulated SNP datasets; identical results for empirical *Regulus satrapa* data shown below (D,E,F).

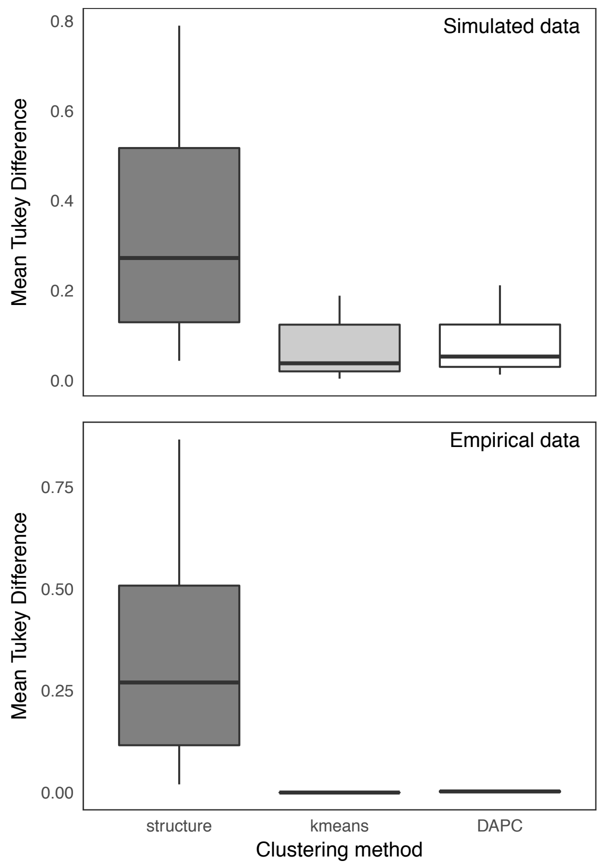

Figure 4: Mean difference in effect size from a Tukey's honest significant difference test, compared across three clustering programs.

following divergence resulted in a high proportion of singletons, which can demonstrate fine-scale patterns of geographic structure (Barton and Slatkin 1986; Gompert et al. 2014). Yet across replicate datasets and *structure* runs, most of analyses failed to detect population limits when singletons were included. In contrast, including rare alleles (MAF=2 to MAF=8) allowed for highly accurate population inference, followed by a decay in accuracy when only common alleles were included.

Our empirical ddRAD *Regulus satrapa* data, featuring fewer singletons, exhibited different behavior in *structure*. While the full, unfiltered dataset occasionally resulted in the correct identification of our three *a priori* population assignments, most runs failed to detect any structure at all. This result repeated for increasingly stringent MAF filters until all rare alleles were excluded, and was replicated in ln(alpha) values.

We believe both simulated and empirical cases reflect behavior of *structure's* likelihood function, with overfitting as a result of either a high frequency of uninformative rare alleles or a high frequency of uninformative common alleles (discussed in the context of maximum likeihood methods in Alexander and Lange 2011). In both scenarios, population k_1 receives an allele frequency distribution averaging 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 receive high frequencies of singletons or otherwise uninformative rare alleles, resulting in the additional broad bands of minority ancestry. With our simulated data, rare but non-singleton alleles reflect fine population structure and thus harm inference when excluded; with our empirical data, rare alleles are uninformative and serve only as noise to common allele frequency distributions reflecting true population history.

This hypothesis is consistent with a pathology related *structure's* inability to model mutation of modern alleles, previously identified as a potential obstacle to accurate inference of population structure under certain histories (Shringarpure and Xing 2008).

Because *structure* assumes each unique allele in the input dataset 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 dataset, the noise from these false allele frequency profiles may mask the signal from alleles indicative of "true" populations.

Nonparametric methods show little sensitivity to MAF choice

Relative to model-based population structure inference, the results of nonparametric methods of k-means clustering and DAPC with cross validation were little influenced by different MAF threshold choices. For simulated datasets, accuracy only slightly decreased as MAF threshold was increased, though large sample sizes contributed to both the breadth of distributions and statistical significance across nearly all MAF values. Across empirical datasets, MAF choice showed no effect on the ability to detect predefined population clusters with either k-means or DAPC with cross validation. Visualization of the first two principal components similarly showed little pattern of increased or decreased separation across groups. These results are unsurprising given both k-means clustering and DAPC heavily weight genetic variants with the ability to distinguish among groups, while retaining properties capable of reconstructing an accurate approximation of the source dataset (Jombart et al. 2010). Indeed, PCA-based methods are widely used in human and *Drosophila* genomics, subfields with a longer history of access to large, computationally intensive SNP datasets (Paschou et al. 2007; Novembre et al. 2008; Pool et al. 2012). Though nonparametric approaches lack the ability to directly model important population genetic statistics such as levels of admixture, their flexibility (particularly in combination with demographic modeling using predefined population assignments) makes them an important tool for systematists and other biologists looking to identify population subdivision in nonmodel organisms.

Sources of error and best practices

Importantly, our results suggest the specific SFS distributions that can cause *structure* and other model-based programs to erroneously fail to detect structure may be generated by either normal demographic processes (e.g., exponential population growth, as in our simulated example) or by assembly errors (potentially present in our empirical example, and well documented in other de novo RADseq datasets (e.g. Shafer et al. 2016). 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: 1) duplicate analyses with nonparametric methods such as PCA and DAPC with cross validation; 2) exclude singletons or compare unfiltered and filtered datasets; 3) compare alignments with multiple assembly parameters.

Population genetics of Regulus satrapa

Though describing population structure and phylogeographic 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 (Klicka 2017). Curiously, the results of our model-based population structure inference suggest not only singletons but all rare alleles ($MAF \leq 8/80$) have a high noise to signal ratio, while common alleles ($MAF \geq 10/80$) accurately reflect expected relationships. This pattern may be driven by either purifying selection eliminating geographically localized variants (Nelson et al. 2012; Jackson et al. 2015), a population bottleneck (Nei et al. 1975; Gattepaille et al. 2013), a burst of recent migration following exponential population growth (Slatkin 1985), or assembly artifacts resulting in a high proportion of uninformative / erroneous sites (Shafer et al. 2016). While all scenarios are likely 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 and Klicka 2006), processes that may result in similar SFS distributions to our example.

Future directions

With simulated and empirical cases reflecting similar (if non-identical) 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 datasets 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., MCMC versus maximum likelihood, Gelman et al. 1996) may shape specific sensitivities. A broader survey of model-based population structure inference methods will help clarify which approaches are best suited to genomic datasets, and lead to the development of more robust software for describing the fundamental units of biological organization.

Conclusions

Our study demonstrates model-based methods to infer population structure are highly affected by minimum MAF choice, while non-model-based methods show relatively little sensitivity. Model-based methods lacking parameters to account for recent mutation may fail due to overfitting as a result of uninformative singletons, or due to limited variation in common alleles given stringent MAF thresholds. As problematic site frequency distributions can be generated by either assembly artifacts or natural demographic processes prior to dataset filtering, we suggest researchers observe our reccomended best practices while conducting exploratory data analysis with these programs.

Acknowledgements

We thank Andy Mack for being better at statistics than us.

*

References

Alexander, D. H. and K. Lange. 2011. Enhancements to the admixture algorithm for individual ancestry estimation. BMC Bioinformatics 12:246.

Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19:1655–1664.

Arnold, B., R. B. Corbett-Detig, D. Hartl, and K. Bomblies. 2013. Radseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Molecular Ecology 22:3179–3190.

Barton, N. and M. Slatkin. 1986. A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. Heredity 56 (Pt 3):409–15.

Blanco-Bercial, L. and A. Bucklin. 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., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124–3140.

Degner, J. F., J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard. 2009. Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. Bioinformatics 25:3207–3212.

Eaton, D. A. 2014. Pyrad: assembly of de novo radseq loci for phylogenetic analyses. Bioinformatics 30:1844–1849.

Excoffier, L., I. Dupanloup, E. Huerta-Sánchez, V. C. Sousa, and M. Foll. 2013. Robust demographic inference from genomic and snp data. PLoS Genetics 9:e1003905.

Faircloth, B. C., J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, and T. C. Glenn. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Systematic Biology 61:717–726.

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

Gattepaille, L. M., M. Jakobsson, and M. G. Blum. 2013. Inferring population size changes with sequence and snp data: lessons from human bottlenecks. Heredity 110:409.

Gautier, M., K. Gharbi, T. Cezard, J. Foucaud, C. Kerdelhu, P. Pudlo, J.-M. Cornuet, and A. Estoup. 2013. The effect of rad allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology 22:3165–3178.

Gelman, A., G. O. Roberts, W. R. Gilks, et al. 1996. Efficient metropolis jumping rules. Bayesian Statistics 5:42.

Gompert, Z., L. K. Lucas, C. A. Buerkle, M. L. Forister, J. A. Fordyce, and C. C. Nice. 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.

Jackson, B. C., J. L. Campos, and K. Zeng. 2015. The effects of purifying selection on patterns of genetic differentiation between drosophila melanogaster populations. Heredity 114:163.

Jombart, T. 2008. adegenet: a r package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.

Klicka, J. 2017. mtdna variation in golden-crowned and ruby-crowned kinglets.

Li, Y.-C., A. B. Korol, T. Fahima, A. Beiles, and E. Nevo. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Molecular Ecology 11:2453–2465.

Marth, G. T., E. Czabarka, J. Murvai, and S. T. Sherry. 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.

Nei, M., T. Maruyama, and R. Chakraborty. 1975. The bottleneck effect and genetic variability in populations. Evolution 29:1–10.

Nelson, M. R., D. Wegmann, M. G. Ehm, D. Kessner, P. S. Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337:100–104.

Nielsen, R., T. Korneliussen, A. Albrechtsen, Y. Li, and J. Wang. 2012. Snp calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PloS ONE 7:e37558.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, et al. 2008. Genes mirror geography within europe. Nature 456:98.

Paschou, P., E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. 2007. Pca-correlated snps for structure identification in worldwide human populations. PLoS Genetics 3:e160.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra. 2012. Double digest radseq: an inexpensive method for de novo snp discovery and genotyping in model and non-model species. PloS ONE 7:e37135.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchen, J. Emerson, P. Saelao, D. J. Begun, et al. 2012. Population genomics of sub-saharan drosophila melanogaster: African diversity and non-african admixture. PLoS Genetics 8:e1003080.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Raj, A., M. Stephens, and J. K. Pritchard. 2014. faststructure: variational inference of population structure in large snp data sets. Genetics 197:573–589.

Shafer, A., C. R. Peart, S. Tusso, I. Maayan, A. Brelsford, C. W. Wheat, and J. B. Wolf. 2016. Bioinformatic processing of rad-seq data dramatically impacts downstream population genetic inference. Methods in Ecology and Evolution .

Shringarpure, S. and E. P. Xing. 2008. mstruct: a new admixture model for inference of population structure in light of both genetic admixing and allele mutations. Pages 952–959 *in* Proceedings of the 25th International Conference on Machine Learning ACM.

Slatkin, M. 1985. Rare alleles as indicators of gene flow. Evolution 39:53–65.

Spellman, G. M. and J. Klicka. 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., M. Coram, P. Wang, X. Zhu, and N. Risch. 2006. Reconstructing genetic ancestry blocks in admixed individuals. The American Journal of Human Genetics 79:1–12.

Team, R. C. 2017. R: A language and environment for statistical computing.

Winger, B. M. 2017. Consequences of divergence and introgression for speciation in andean cloud forest birds. Evolution 71:1815–1831.

Supplemental Material

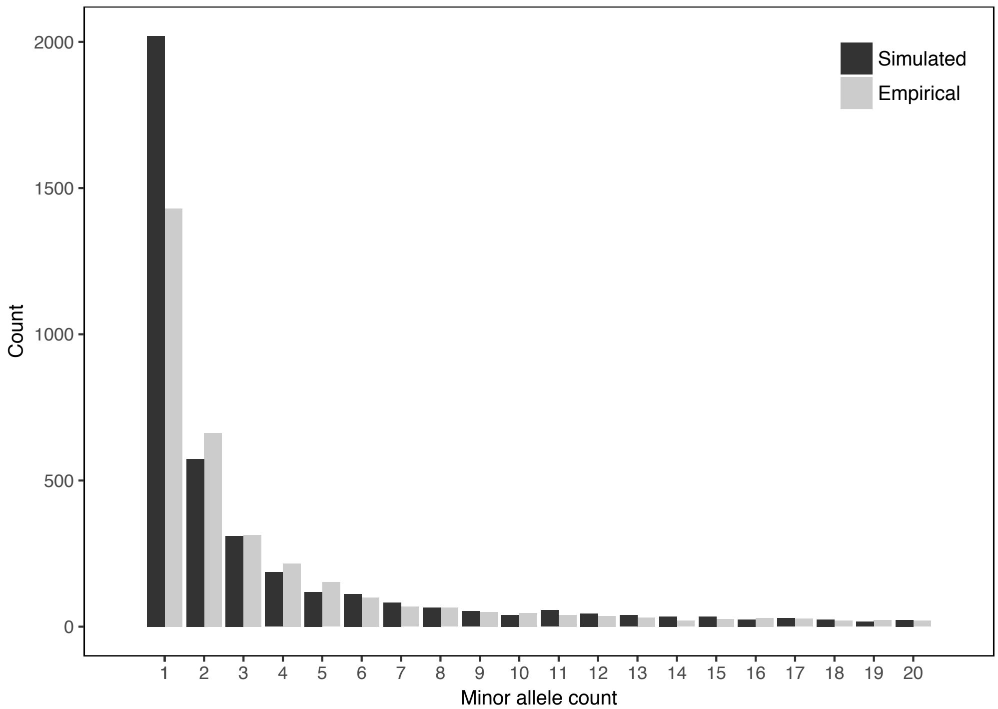

Figure S1: Site frequency spectra for the simulated and empirical datasets.

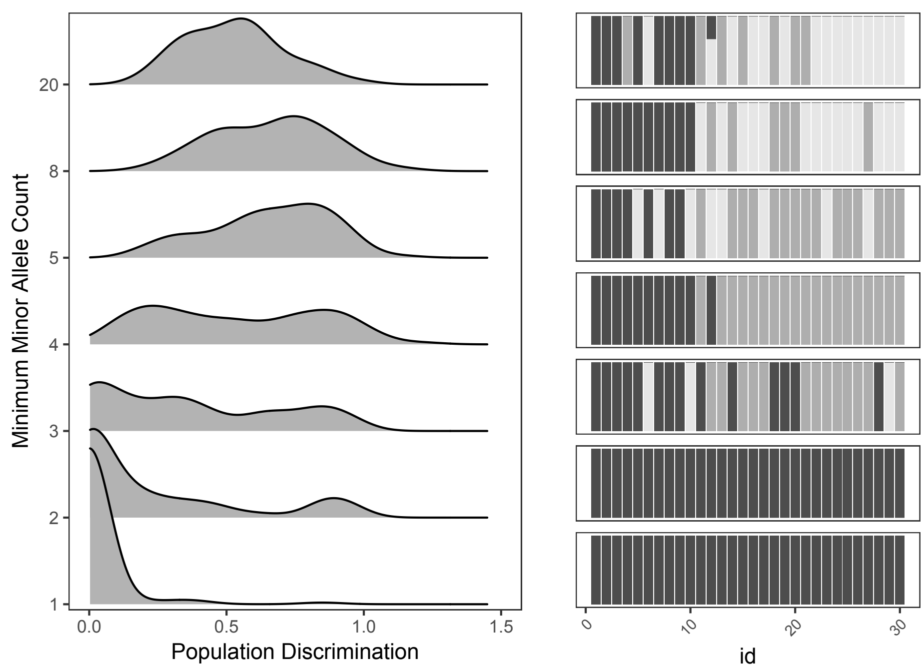

Figure S2: Population discrimination and representative clustering plots for simulated SNP data.

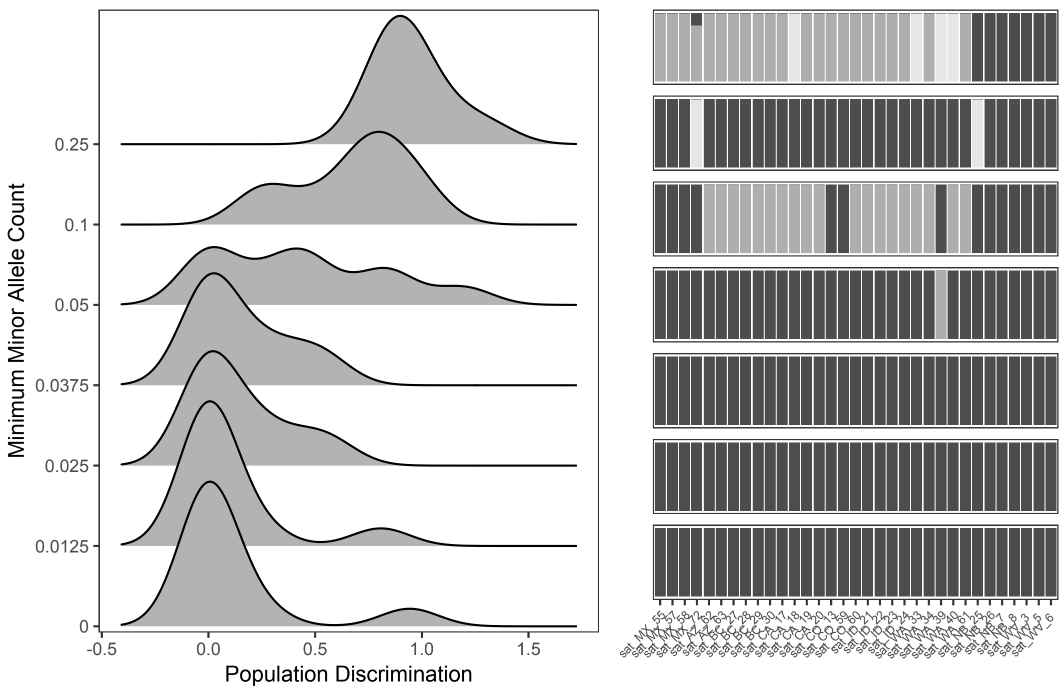

Figure S3: Population discrimination and representative clustering plots for empirical *Regulus satrapa* data.