Ascertainment Bias in Spatially Structured Populations: A Case Study in the Eastern Fence Lizard

ERICA BREE ROSENBLUM AND JOHN NOVEMBRE

From the Museum of Vertebrate Zoology, University of California, Berkeley CA 94720 (Rosenblum); and the Department of Integrative Biology, University of California, Berkeley CA 94720 (Novembre). Erica Bree Rosenblum is now at the Department of Molecular and Cell Biology, University of California, Berkeley CA 94720; and John Novembre is now at the Department of Human Genetics, University of Chicago, Chicago, IL 60637.

Address correspondence to J. Novembre at the address above, or e-mail: jnovembre@uchicago.edu.

Abstract

Despite increased interest in applying single nucleotide polymorphism (SNP) data to questions in natural systems, one unresolved issue is to what extent the ascertainment bias induced during the SNP discovery phase will impact available analysis methods. Although most studies addressing ascertainment bias have focused on human populations, it is not clear whether existing methods will work when applied to other species with more complex demographic histories and more significant levels of population structure. Here we present findings from an empirical approach to exploring the effect of population structure on issues of ascertainment bias in the Eastern Fence Lizard, *Sceloporus undulatus*. We find that frequency spectra and summary statistics were highly sensitive to SNP discovery strategy, necessitating careful selection of the initial ascertainment panel. Randomly selected ascertainment panels performed equally well as ascertainment panels chosen to jointly sample geographic, phenotypic, and genetic diversity. Geographically restricted panels resulted in larger biases. Additionally, we found existing ascertainment bias correction methods, which were not developed for geographically structured data sets, were largely effective at reducing the impact of ascertainment bias. Because bias correction methods performed well even when underlying assumptions were violated, our results suggest tools are currently available to analyze SNP data in structured populations.

A number of recent studies have drawn attention to the utility of single nucleotide polymorphisms (SNPs) in ecology and evolution (e.g., Brumfield et al. 2003; Morin et al. 2004). Multiple unlinked markers provide increased power to estimate demographic parameters, identify loci under selection, and reconstruct lineage histories. SNPs are particularly effective as multilocus markers because they can provide extensive genomic coverage, simpler mutational models than microsatellites, and more reliable characterization than fingerprinting methods (Kuhner et al. 2000; Brumfield et al. 2003; Morin et al. 2004). For species with no preexisting nuclear sequence data, SNPs can be found by constructing genomic libraries and screening anonymous regions of the genome for nucleotide polymorphisms (e.g., Karl and Avise 1993; Bagley et al. 1997; Brumfield et al. 2003). This method allows for the efficient identification of large numbers of markers for population study. With declining costs of sequencing, de novo SNP development is increasingly feasible to implement on a large-scale and SNP data sets are likely to become increasingly popular in studies of nonmodel species.

Although biologists now have the technical ability to generate large numbers of multilocus markers, a number of challenges remain for SNP data analysis in nonmodel species. For one, species in the wild often have complex evolutionary histories and spatially structured populations. Understanding the impact of population structure on SNP analysis methods is particularly important for issues of ascertainment bias. SNPs are generally identified using an "ascertainment panel," a small sample of individuals from a population or region of interest. Because the ascertainment panel contains a subset of individuals, only a fraction of all variable sites sorting in the population will be discovered. When SNPs detected within the ascertainment panel are then genotyped on a larger sample of individuals, an "ascertainment bias" is introduced that depends on the number and source of individuals in the ascertainment panel (Eberle and Kruglyak 2000; Kuhner et al. 2000; Wakeley et al. 2001). In particular, the observed SNP frequency spectrum will deviate from the true SNP frequency spectrum because existing SNPs, particularly those at low frequencies, have been under-sampled. Ascertainment bias

therefore affects the accuracy of any demographic parameter estimated based on the allele frequency spectrum. This bias is problematic because many commonly used summary statistics are functions of the allele frequency spectrum (e.g., θ_π, Tajima's *D*, Fay and Wu's H). Ultimately, ascertainment bias can lead to mistaken inferences about demographic history. For instance, ascertainment bias can lead to an excess of intermediate frequency alleles and a deficiency of low-frequency alleles that can obscure signatures of population expansion and generate the appearance of population structure where there is none.

Although the issue of ascertainment bias is important to consider in all SNP studies, it may be exacerbated in studies of species with complex (or unknown) population structures. When populations are highly structured, it is likely difficult to sample existing variation in all subpopulations during the SNP detection phase, and erroneous inferences become more likely. For example in humans, ascertainment of polymorphisms in European populations has lead to spurious inferences of low heterozygosities outside of Europe and of significant differences in ancestral allele frequencies among human populations (Mountain and Cavalli-Sforza 1994).

Methods do exist to help account for ascertainment strategy in analyses of SNP data (e.g., Wakeley et al. 2001; Polanski and Kimmel 2003; Marth et al. 2004; Nielsen et al. 2004). However, these methods are based either on 1) models that do not incorporate population structure (e.g., Polanski and Kimmel 2003; Marth et al. 2004), 2) models of population structure and growth that represent human demographic history (e.g., Wakeley et al. 2001), or 3) nonparametric models that assume the ascertainment panel is a random sample from a population (e.g., Nielsen et al. 2004). The last of these approaches is the most generally applicable to studies of natural populations. However, in organisms where population structure is not known, a priori and/or sampling is challenging or geographically limited, it may be logistically difficult to obtain ascertainment panels that represent a random sample of the population. Studies of natural systems with varying demographic histories are therefore necessary to understand how different ascertainment strategies affect inferences of demographic parameters and how adequately existing ascertainment bias correction methods perform for spatially structured populations.

Here we evaluate the performance of the Nielsen et al. (2004) method for correcting ascertainment bias. Although this correction method has been previously evaluated using human genome-wide polymorphism data (Clark et al. 2005), we investigate how well it can be applied in practice to spatially structured populations of a nonhuman, nonmodel organism. We focus on how individuals are chosen for the ascertainment panel (whether chosen from multiple or single demes) and we analyze data from the Eastern Fence Lizard, *Sceloporus undulatus*, as the basis of our case study.

Sceloporus undulatus is a broadly distributed taxon with extensive geographic variation and a complex evolutionary history (e.g., Leaché and Reeder 2002). Although there are many scales at which the effect of spatial structure can be evaluated in *S. undulatus,* we chose to focus on a portion of the range

that presents particularly compelling evolutionary questions. The focal region of this study is the Tularosa Basin of southern New Mexico, which was colonized by *S. undulatus* following the last glacial maximum. Subsequently, populations adapted to multiple substrate environments (i.e., dark soil, white sand, and black lava habitats), and there is evidence for ongoing gene flow between divergent morphs (Rosenblum 2006). The complex history of population expansion, local adaptation, and ongoing gene flow has left a strong signature of population structure. Overall F_{ST} in this region based on mitochondrial data is 0.54 (Rosenblum 2006). Although our study focuses on only a portion of the *S. undulatus* range, our multilocus nuclear estimate of nucleotide diversity in this region ($\pi = 0.007$) is nearly an order of magnitude higher than that observed in ethnically diverse samples of humans ($\pi = 0.0008$, International SNP Map Working Group 2001).

Methods

We conducted our analyses using a data set containing 19 anonymous nuclear loci sequenced for a set of 91 individuals chosen to sample known geographic variation in *S. undulatus* populations in southern New Mexico. Thirty-seven "wild-type" animals from 4 dark soil localities, 29 blanched animals from 3 white sand localities, and 25 melanic animals from 3 lava flow localities were sampled; whenever possible, 10 individuals were sampled per population. The total data set consisted of 4949 bp of sequence data from across 19 loci with a total of 198 variable sites. Details on primer design, experimental conditions, and data handling are presented in Rosenblum et al. (2007). Raw sequence data have been deposited in GenBank (accession numbers EF411269–EF412962). The 19 loci varied in observed levels of sequence variation; 1–12 SNPs were found per 100 bp, with an average of 4.4 SNPs/100 bp. The frequency spectrum of the 198 SNPs exhibited, as expected, a long tail of high-frequency SNPs. When categorized into bins, there were many low-frequency SNPs and increasingly fewer high-frequency SNPs (see Figure 1). Because these data are based on full sequences from all 91 individuals, there is no ascertainment bias in this sample, and hereafter, we will refer to it as the sample with no ascertainment bias.

The large number of individuals sequenced in this data set provided a unique opportunity to investigate the effect of different ascertainment strategies on SNP detection and summary statistics describing genetic diversity. Specifically, we were able to empirically investigate the effect various ascertainment procedures would have had if they had been applied on this data set. To mimic the ascertainment process, we constructed mock ascertainment panels using a subset of the 91 individuals we sequenced. For each mock panel, we scored polymorphic sites as discovered SNPs and then we constructed a data set as if we had genotyped only the discovered SNPs in the full sample of 91 individuals. By comparing the patterns of variation observed using different ascertainment panels, we assessed how panel design affected the extent of ascertainment bias.
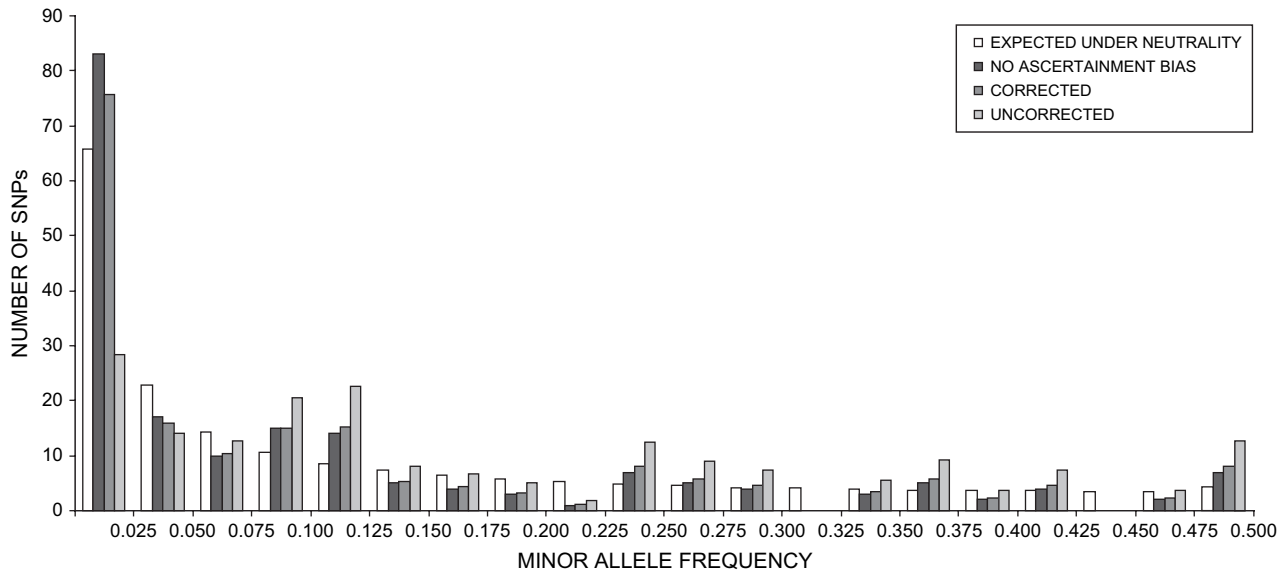
Figure 1. SNP frequency spectra. The spectrum labeled "no ascertainment bias" is derived from all SNPs observed in the *Sceloporus undulatus* data set of 91 individuals. The "uncorrected" frequency spectrum represents frequencies for only those SNPs that were observed in one example of applying the random ascertainment panel strategy. The "corrected" spectrum results from the application of ascertainment bias correction of Nielsen et al. (2004) to the "uncorrected" spectrum.

We focused our investigation specifically on the effects of changing the geographic and genetic composition of individuals in the ascertainment panel. Three different categories of mock ascertainment panel designs were tested. First, as might be typical for species that will be studied using SNPs, prior data may exist that might guide the selection of an "intentionally diverse" panel. Here, prior data (Rosenblum 2006) were used to ensure representation of different geographic, genetic, and phenotypic groups in the "intentionally diverse" panels. These panels contained 3 individuals from each of 2 geologically recent habitat types (white sand and lava flow) and 4 individuals from geographically disparate potentially ancestral dark soil populations; where possible individuals within habitats were chosen to be from divergent mitochondrial DNA clades. Second, "geographically restricted" panels were comprised of individuals from a single geographical population. This sample design may arise in practice because of logistical limitations to sampling or because a set of SNPs discovered in a localized study is later used for a survey of genetic variation in a more cosmopolitan study. Third, "random" panels consisted of randomly chosen individuals. Each panel contained 10 individuals and had a unique set of individuals, although some individuals are present in multiple panels.

For each mock ascertainment panel, we mimicked the SNP discovery phase by recording the variable sites present in the 10 individuals of that particular panel (a subsample of all existing SNPs discovered when all 91 individuals were sequenced). We then observed the frequency spectrum (described below) and computed summary statistics (described below) using all 91 individuals but using only the particular SNPs detected from each mock ascertainment panel. All results for the mock ascertainment trials were therefore cal-

culated based on data from 91 individuals but using a different subsample of variable sites depending on which individuals had been utilized for SNP ascertainment.

Typically out-group sequences may be used to classify each allele at a SNP locus as ancestral or derived, and then the frequency spectrum is constructed by making a histogram of the frequencies of the derived allele at each site. These are referred to as unfolded frequency spectra. Here, because of our interest in nonmodel organisms, we assume the out-group sequence is unavailable and so we calculated "folded" frequency spectra by constructing a histogram of the frequencies of the minor (less-frequent) allele at each site (see e.g., Figure 1). The 2 forms of frequency spectra are closely related and both are more detailed depictions of the patterns of polymorphism in the data than can be found using summary statistics.

Typically frequency spectra are constructed under the assumption that all sites have the same sample size, but in reality missing data resulted in a variable number of chromosomes observed per site (122–192 chromosomes). As a solution, we used the relative frequencies of each allele at each site to construct a frequency spectrum standardized to the minimum observed sample size of 122 chromosomes. This approach is conservative in that it results in a sample size that is low relative to the actual number of observations across sites; however, it can generate modest distortions to the frequency spectrum (Marth et al. 2004).

We also calculated 3 commonly used summary statistics (θ_π, θ_W, and Tajima's D). θ_π and θ_W are each estimates of the genetic diversity in a population, and Tajima's D is a normalized version of the difference between these 2 estimates that is sensitive to departures from a standard neutral model. θ_π and θ_W were calculated using the folded frequency spectrum

as input to Equations 5 and 6 of Fu (1994), and Tajima's D was calculated based on these 2 estimates. To assess the effects of ascertainment bias, we compared the results from 10 "intentionally diverse" panels, 6 "geographically restricted" panels, and 100 random panels with results from the full sample that lacks ascertainment bias.

For each panel, we also evaluated the performance of the ascertainment bias correction method by Nielsen et al. (2004). The method of Nielsen et al. (2004) is a general method to correct a frequency spectrum given a particular ascertainment-sampling scheme. Their approach does not make any assumptions regarding population structure, but does assume that the ascertainment panel is obtained as a random sample from the whole population. We used a straightforward extension of their results to folded frequency spectra and to renormalize the spectra.

To describe the extension to folded spectra, let us define the observed allele frequency spectrum in terms of n_i^*, the observed number of sites with minor allele frequency i in a sample of size n chromosomes. The frequency i of the minor allele can only take on integer values from 1 to $[n/2]$ where brackets denote the operation of rounding the number down to the nearest integer. The vector $(n_1^*, n_2^*, \ldots, n_{[n/2]}^*)$ represents the observed allele frequency spectra. We presume that the observed spectrum is affected by ascertainment bias. Specifically, we suppose it has been obtained by genotyping S_0 SNP loci that are initially discovered on the basis of being variable within an ascertainment sample size of d chromosomes. Let p_i^* be the proportion of nucleotide sites with minor alleles of frequency i in a sample without ascertainment bias. As in Nielsen et al. (2004), we define the reconstituted frequency spectra by the vector $P^* = (p_1^*, p_2^*, \ldots, p_{[n/2]}^*)$. The goal is to estimate the reconstituted allele frequency spectra based on the observed allele frequency spectra. A straightforward extension of the derivations for an unfolded frequency spectra in Nielsen et al. (2004) can show the maximum likelihood estimate P^* is given by

$$P_k^* = \frac{n_k^*}{\Pr(\mathrm{Asc}\,|\,Y=k)} \left[\sum_{j=1}^{[n/2]} \frac{n_j^*}{\Pr(\mathrm{Asc}\,|\,Y=j)} \right]^{-1}$$
$$k = 1, 2, \ldots, [n/2],$$

where

$$\Pr(\mathrm{Asc}\,|\,Y=k) = 1 - \frac{\binom{k}{d} + \binom{n-k}{d}}{\binom{n}{d}},$$

and Y is a random variable equal to the observed number of copies of the minor allele at a site.

To calculate corrected summary statistics from P^* (such as θ_π, θ_W, and Tajima's D), the spectrum should be renormalized to sum to S, the number of segregating sites in a sample without ascertainment bias. However, S is unobserved and must be estimated. Assuming sites are independent, then the number of ascertained sites, S_0, is binomial random variable resulting from S trials each with success probability $\Pr(\mathrm{Asc})$, where $\Pr(\mathrm{Asc})$ is the probability a site is ascertained:

$$\Pr(\mathrm{Asc}) = \sum_{i=1}^{[n/2]} \Pr(\mathrm{Asc}\,|\,Y=i)\Pr(Y=i).$$

Here we approximate $\Pr(Y=\mathrm{i})$ with its maximum likelihood estimate p_i^*. Our resulting method-of-moments estimator of S is

$$\hat{S} = S \left[\sum_{i=1}^{[n/2]} \Pr(\mathrm{Asc}\,|\,Y=i)p_i^* \right]^{-1}.$$

We then use S_0 to renormalize the reconstituted allele frequency spectrum.

As a simple means to handle missing data within an ascertainment panel, which leads to variable "depths" (i.e., sample sizes among sites in the ascertainment panel), we chose to use correction of Nielsen et al. (2004) with the depth, d, fixed at its maximum value across sites of 20 chromosomes. By assuming d to be fixed at a value larger than or equal to the actual value at any site, our approach undercorrects for the effects of ascertainment bias, thereby providing a conservative test of the robustness of this method to violations of underlying assumptions.

Using the corrected frequency spectra, we again calculated the values of θ_π, θ_W, and Tajima's D. The values of each of the 3 summary statistics were compared among panel designs. Again, to assess how well the different ascertainment methods performed, we compared the values obtained from each mock panel design to the values obtained using the full sequence data from 91 individuals.

Results

All ascertainment panel strategies recovered fewer SNPs than were found in the full data set; however, some ascertainment strategies performed better than others. The random panels and the diverse panels detected on average nearly identical numbers of SNPs (Table 1). Although these 2 strategies only recovered just more than half the total existing SNPs, they recovered 25% more than the geographically restricted panels (Table 1).

For the summary statistics, SNPs ascertained from the random and diverse panels provided estimates that were more similar to the values calculated with no ascertainment bias than those ascertained from the geographically restricted panel (Table 1). However, raw data from all panel designs differed substantially from the results obtained with the full sample with no ascertainment bias. Although θ_π was fairly similar to the result with no ascertainment bias, θ_W and Tajima's D were not. As expected when under-sampling rare alleles, Tajima's D from all subsampling schemes was large and positive when it should have been small and negative.

The ascertainment bias correction method of Nielsen et al. (2004) was effective at reconstituting the allele frequency spectra (Figure 1). Taking the random panel design as an example, rare SNPs were under-sampled in the raw data (Figure 1). However, once the correction was applied, the shape of the SNP frequency spectrum was more closely

Table 1. Performance of 3 different ascertainment strategies. Results from the full sample (no ascertainment bias) are compared with average results from 100 random, 10 intentionally diverse, and 6 geographically restricted ascertainment panels

Sampling strategy	No ascertainment bias	Random panel	Intentionally diverse panel	Geographically restricted panel
Average no. SNPs ascertained	191	105 (SD = 5.54)	104 (SD = 4.03)	85 (SD = 8.35)
Average θ_π (uncorrected)	30.85	26.91 (SD = 0.78)	26.81 (SD = 0.44)	23.27 (SD = 1.59)
Average θ_π (corrected)		29.33 (SD = 0.91)	29.18 (SD = 0.56)	25.09 (SD = 1.72)
Average θ_W (uncorrected)	35.52	19.47 (SD = 1.03)	19.27 (SD = 0.75)	15.78 (SD = 1.55)
Average θ_W (corrected)		31.11 (SD = 3.73)	30.10 (SD = 2.53)	25.33 (SD = 4.39)

matched to neutral expectations and to the spectrum derived from the full data set (Figure 1). Summary statistics calculated from the corrected spectra also provided estimates that were closer to those obtained with no ascertainment bias, particularly for θ_W and Tajima's D, which were poorly estimated with the raw data (Table 1). For example, the value of Tajima's D based on the corrected spectra decreased in all cases. For SNPs ascertained using the random and diverse strategies, the corrected Tajima's D was quite close in direction and magnitude to that calculated from the full sample with no ascertainment bias (Figure 2). However, the correction was less effective for SNPs ascertained from the geographically restricted subsample, a pattern observed across all summary statistics (Table 1).

Discussion

Choice of ascertainment panels for spatially structure populations has profound implications for inferences derived from SNP data. Any ascertainment strategy using a reduced pool of individuals will find only a subset of existing SNPs, and rare alleles will be particularly under-sampled. However, certain ascertainment strategies perform better than others, especially for subdivided populations. In this study, we found that ascertainment panels containing geographically

Figure 2. Comparison of Tajima's D calculated with all available SNPs with that calculated with those SNPs ascertained with 3 different strategies. Raw results are compared with results after the Nielsen et al. (2004) ascertainment bias correction. Bars represent 2 standard errors.

restricted samples (i.e., all individuals from the same local population) were the least effective in capturing the SNP frequency distribution that would be obtained in a sample with no ascertainment bias. Perhaps surprisingly, panels comprised of randomly sampled individuals and those containing individuals known a priori to be geographically, genetically, and phenotypically diverse performed similarly well. Researchers studying species with high levels of population structure should therefore choose ascertainment panels to sample known geographic variation at the spatial scale of interest or should use a random panel design.

Although many methods used to correct for ascertainment bias assume a randomly chosen ascertainment panel, the correction method employed here (Nielsen et al. 2004) was robust to some deviation from this assumption. Corrected frequency spectra and corrected values of summary statistics (e.g., Tajima's D) were much closer to the values obtained without ascertainment than uncorrected values, and after correction, both the random panel strategy and intentionally diverse panel strategy produced summary statistics close to those with no ascertainment bias. In agreement with the results of Clark et al. (2005), we found that while the correction method greatly improved the fit of the data to expectations, some deviation from the results from the full data set was still observed (Figure 1). The conservative approach we took for dealing with missing data in the ascertainment panel should lead to an under-correction of the frequency spectra. Indeed, the frequency spectra and corrected Tajima's D values are consistent with this prediction. Future applications of this correction can improve on the results found here by explicitly handling the variable amounts of missing data per site. Nielsen et al. (2004) suggest an alternative correction for when the panel size, d, varies among loci, and this alternative correction may be useful for incorporating information regarding missing data.

Given that random and intentionally diverse ascertainment panels recovered nearly identical numbers of SNPs and bias correction methods were fairly successful using both strategies, are there practical reasons to choose one panel strategy over the other? Because the randomly chosen panels performed as well as the intentionally diverse panels in this study, there was no "cost" to using a randomly selected ascertainment panel. Additionally, available correction methods assume randomly chosen ascertainment panels. Although the correction method employed here was robust to deviations from the assumption of random sampling during SNP discovery, we have not tested the extent to which
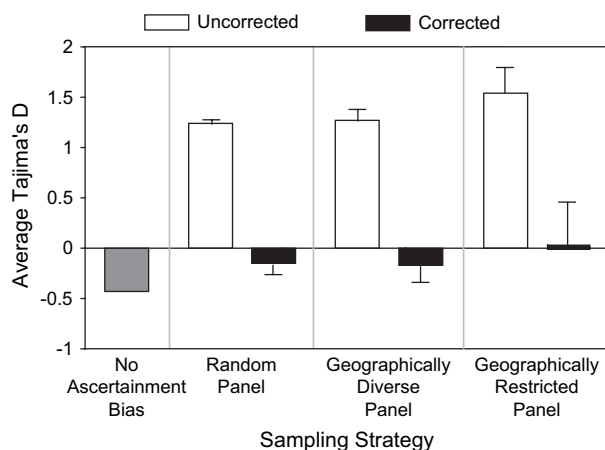
this assumption can be violated. In some systems, nonrandom choice of ascertainment panels may be more difficult to model and may lead to poor performance of available correction methods. Therefore, a randomly chosen ascertainment panel is likely the method of choice for many researchers. Using random panels will also simplify the first step of SNP discovery because a priori genetic data on population structure is unnecessary; individuals for the panel can simply be chosen at random over the spatial scale of interest.

It is important to note that there are no ubiquitous characteristics of nonmodel species, and therefore, the generality of our results will require further exploration in additional taxa. For example, levels of population structure and nucleotide diversity in *S. undulatus* may be higher than those observed in many species (e.g., Brumfield et al. 2003; Ryynänen and Primmer 2006). However, many species in the wild do exhibit high densities of SNPs (e.g., Hughes and Mouchiroud 2001; Primmer et al. 2002). Because issues of ascertainment bias have been evaluated to date in species with comparatively low levels of population structure and nucleotide diversity, developing a case study from the opposite end of the continuum—as we have done—is particularly informative. As large multilocus nuclear data sets become available for more nonmodel species, the sensitivity of our conclusions to different levels of population structure can be better evaluated.

Although the ascertainment bias correction method evaluated in this study (Nielsen et al. 2004) allowed us to calculate basic summary statistics that are useful for studying natural populations, assessing the confidence intervals and significance of these statistics requires an additional step. Coalescent simulations should be used to accurately take into account the variances that arise due to the linkage relationships among SNPs and the variance induced by using a corrected frequency spectrum rather than a directly observed one. Future developments in ascertainment bias corrections will likely lead to direct corrections for estimators that will preclude the need for simulations to obtain confidence intervals (Polanski and Kimmel 2003; Nielsen et al. 2004). However, our results demonstrate that practical solutions are currently available for handling SNP ascertainment bias in structured populations.

Funding

Acknowledgments

References

Bagley MJ, Medrano JF, Gall GAE. 1997. Polymorphic molecular markers from anonymous nuclear DNA for genetic analysis of populations. Mol Ecol. 6:309–320.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV. 2003. The utility of single nucleotide polymorphisms in inferences of population history. Trends Ecol Evol. 18:249–256.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 15:1496–1502.

Eberle MA, Kruglyak L. 2000. An analysis of strategies for discovery of single-nucleotide polymorphisms. Genet Epidemiol. 19:S29–S35.

Fu YX. 1994. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics. 138:1375–1386.

Hughes S, Mouchiroud D. 2001. High evolutionary rates in nuclear genes of squamates. J Mol Evol. 53:70–76.

International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 409:928–933.

Karl SA, Avise JC. 1993. PCR-based assay of Mendelian polymorphisms from anonymous single-copy nuclear DNA: techniques and applications for population genetics. Mol Biol Evol. 10:342–361.

Kuhner MK, Beerli P, Yamato J, Felsenstein J. 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics. 156:439–447.

Leaché AD, Reeder TW. 2002. Molecular systematics of the Eastern Fence Lizard (Sceloporus undulatus): a comparison of parsimony, likelihood, and Bayesian approaches. Syst Biol. 51:44–68.

Marth GT, Czabarka E, Murvai J, Sherry ST. 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.

Morin PA, Luikart G, Wayne RK, SNP Workshop Group. 2004. Applications of single nucleotide polymorphism (SNPs) in ecology, evolution, and conservation. Trends Ecol Evol. 19:208–216.

Mountain JL, Cavalli-Sforza LL. 1994. Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc Natl Acad Sci USA. 91:6515–6519.

Nielsen R, Hubisz MJ, Clark AG. 2004. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics. 168:2373–2382.

Polanski A, Kimmel M. 2003. New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics. 165:427–436.

Primmer CR, Borge T, Lindell J, Sætre G-P. 2002. Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome. Mol Ecol. 11:603–612.

Rosenblum EB. 2006. Convergent evolution and divergent selection: lizards at the White Sands ecotone. Am Nat. 167:1–15.

Rosenblum EB, Belfiore N, Moritz C. 2007. Anonymous nuclear marker discovery in the Eastern Fence Lizard, Sceloporus undulatus. Mol Ecol Notes. 7:113–116.

Ryynänen HJ, Primmer CR. 2006. Single nucleotide polymorphism (SNP) discovery in duplicated genomes: intron-primed exon-crossing (IPEC) as a strategy for avoiding amplification of duplicated loci in Atlantic salmon (Salmo salar) and other salmonid fishes. BMC Genomics. 7:192.

Wakeley J, Nielsen R, Liu-Cordero SN, Ardlie K. 2001. The discovery of single-nucleotide polymorphisms-and inferences about human demographic history. Am J Hum Genet. 69:1332–1347.