



Consideration of sample source for establishing reliable genetic microsatellite data from mammalian carnivore specimens held in natural history collections

ROBERT C. LONSINGER,* DAVID DANIEL, JENNIFER R. ADAMS, AND LISETTE P. WAITS

Department of Natural Resource Management, South Dakota State University, Brookings, SD 57007, USA (RCL)

Economics, Applied Statistics and International Business Department, New Mexico State University, Las Cruces, NM 88003, USA (DD)

Department of Fish and Wildlife Sciences, University of Idaho, Moscow, ID 83844, USA (JRA, LPW)

* Correspondent: Robert.Lonsinger@sdstate.edu

Specimens from natural history collections (NHCs) are increasingly being used for genetic studies and can provide information on extinct populations, facilitate comparisons of historical and contemporary populations, produce baseline data before environmental changes, and elucidate patterns of change. Destructive sampling for DNA may be in disagreement with NHC goals of long-term care and maintenance. Differentiating quality among sample sources can direct destructive sampling to the source predicted to yield the highest quality DNA and most reliable data, potentially reducing damage to specimens, laboratory costs, and genotyping errors. We used the kit fox (*Vulpes macrotis*) as a model species and evaluated the quality and reliability of genetic data obtained from carnivoran specimens via three different sample sources: cranial bones, nasal bones, and toepads. We quantified variation in microsatellite amplification success and genotyping error rates and assessed the reliability of source-specific genic data. Toepads had the highest amplification success rates and lowest genotyping error rates. Shorter loci had higher amplification success and lower allelic dropout rates than longer loci. There were substantial differences in the reliability of resulting multilocus genotypes. Toepads produced the most reliable data, required the fewest replicates, and therefore, had the lowest costs to achieve reliable data. Our results demonstrate that the quality of DNA obtained from specimens varies by sample source and can inform NHCs when evaluating requests for destructive sampling. Our results suggest that prior to large-scale specimen sampling, researchers should conduct pilot studies to differentiate among source-specific data reliability, identify high performing loci, reduce costs of analyses, and minimize destructive sampling.

Key words: consensus genotypes, DNA degradation, genotyping error, historical DNA, microsatellite loci, natural history collections, reliability, *Vulpes macrotis*

Natural history collections (NHCs) provide an extensive and retrospective resource for researchers (e.g., Wandeler et al. 2007; Lister et al. 2011; Holmes et al. 2016; McLean et al. 2016). NHCs offer ecologists an opportunity to investigate populations and biodiversity through time and examine changes following accelerated anthropogenic disturbances (e.g., habitat loss and modification, climate change, invasive species—Shaffer et al. 1998; Gardner et al. 2014). With the advancement of molecular techniques, NHCs are increasingly being used to investigate macroevolutionary and microevolutionary (e.g., changes in genetic diversity) processes (Austin and Melville 2006; Wandeler et al. 2007; Holmes et al. 2016). This is especially relevant for imperiled species as conservation geneticists can use NHCs to evaluate the genetic and demographic consequences of

declining population size by comparing historical and contemporary samples (Wisely et al. 2002; Miller and Waits 2003; Culver et al. 2008; Lonsinger et al. 2018a).

Requesting genetic material from NHCs requires sampling of a finite resource, a process that is restricted and requires careful consideration (Wandeler et al. 2007; Holmes et al. 2016). For historical carnivoran specimens, common sources of genetic samples have included skin tissues, dried muscles, soft tissues remaining on skeletal specimens, toepads, claws, bones, and teeth (e.g., Wisely et al. 2004; Schwartz et al. 2007; Casas-Marce et al. 2010; Holbrook et al. 2012; Jordan et al. 2012; Jansson et al. 2014; McDonough et al. 2018). Similar to noninvasive genetic samples (i.e., DNA collected from the environment such as from feces or hair), DNA from

historical specimens is often degraded (Wandeler et al. 2007). Consequently, destructive sampling does not guarantee acquisition of usable or reliable DNA, and researchers should select the sample source(s) predicted to yield the highest quality DNA and that minimizes destruction to the specimen (Casas-Marce et al. 2010).

The quality of DNA from historical specimens is generally diminished, as cellular repair processes that counteract DNA degradation in living cells cease at death. Natural processes driving DNA degradation in historical (and ancient) DNA include nucleases, damage by microorganisms (e.g., bacteria), oxidative and hydrolytic lesions, and other chemical processes (reviewed in detail by Pääbo et al. 2004). These processes fragment DNA, reducing the size of sequences available for amplification. Beyond natural DNA degradation, specimen preparation techniques such as tanning, formalin, and bleaching may further fragment DNA (Wandeler et al. 2007). Thus, differences in preservation techniques within and among sample sources (e.g., bones, skins) may drive variation in DNA quality, even for specimens that are the same age (Pääbo et al. 2004).

The low quality of DNA in historical samples can negatively influence polymerase chain reaction (PCR) amplification success and may increase the probability of genotyping errors (i.e., allelic dropout and false alleles—Wandeler et al. 2007). Allelic dropout (i.e., the failure to detect one allele in a heterozygote genotype during a successful PCR amplification) can result from low quantities of DNA, whereas false alleles (i.e., an allele identified during a PCR amplification that is not a true allele in the respective genotype) tend to be PCR-generated artifacts from polymerase slippage during PCR (Broquet and Petit 2004; Wandeler et al. 2007). Heterogeneity in DNA degradation processes, preparation techniques, and preservation conditions among different sample sources are expected to drive variation in DNA quality and the resulting PCR success and genotyping error rates (Greenwood et al. 1999).

Accordingly, studies using microsatellite loci have investigated DNA degradation among sample sources for historical carnivoran specimens by evaluating variation in amplification success rates and, to a lesser extent, genotyping error rates. Conflicting results have been reported across taxa. For example, bone samples produced higher DNA amplification success rates than skin samples for specimens of arctic fox (*Vulpes lagopus*—Nyström et al. 2006) and black-footed ferret (*Mustela nigripes*—Wisely et al. 2004). In contrast, tissues from footpads had higher amplification success rates than bone samples for Eurasian lynx (*Lynx lynx*—Polanc et al. 2012). No difference was observed in genotyping error rates for bone and skin samples of black-footed ferret specimens (Wisely et al. 2004), while footpad samples produced lower error rates than bone samples for Eurasian lynx (Polanc et al. 2012).

Low amplification success can limit usefulness of samples and error-prone samples (i.e., samples amplifying with the presence of genotyping errors) may be particularly problematic if they produce results with questionable reliability. Genotyping errors can bias the results of analyses (e.g., genetic diversity, population genetic structure, effective population

size, parentage) if erroneous multilocus genotypes are assigned to individuals (Miller et al. 2002; Wandeler et al. 2007). For example, allelic dropout has been suggested as a potential cause of heterozygote deficiency observed for a microsatellite locus used to study historical arctic fox specimens (Nyström et al. 2006). In noninvasive genetic sampling studies, it has become common practice to quantify the reliability of genetic data, particularly for multilocus genotypes observed in only a single sample (e.g., Kitchen et al. 2006; Stenglein et al. 2010; Lonsinger et al. 2018b). Though not as common, researchers have begun to incorporate reliability metrics into studies employing historical specimens (e.g., Polanc et al. 2012). In addition to influencing data reliability, error-prone samples may require a greater number of replicates to resolve genotypic differences observed among individual replicates, potentially increasing laboratory costs (Pompanon et al. 2005; Lonsinger et al. 2015).

We used the kit fox (*Vulpes macrotis*) as a model carnivoran species and evaluated the quality of data from nuclear DNA (nDNA) microsatellite loci extracted from historical specimens to identify the optimal DNA source and assess multilocus genotype reliability. We quantified variation in nDNA amplification success rates and genotyping error rates among three sample sources: cranial bones, nasal bones, and toepads. Additionally, we used two metrics to compare the relative reliability of genotypes derived from each sample source: mean reliability across samples and mean proportion of deviations among loci (between the source-specific and specimen-level consensus genotypes). It has been suggested that bone may tend to preserve DNA better than soft tissues (Greenwood et al. 1999). Thus, we predicted that bone samples would have higher amplification success rates, lower genotyping error rates, and higher data reliability than toepads (Casas-Marce et al. 2010). Based on the processes degrading DNA in historical specimens, we predicted that amplification success would be negatively related to sample age (Wandeler et al. 2003, 2007) and locus length (Wandeler et al. 2003; Buchan et al. 2005, Lonsinger et al. 2015). With respect to genotyping errors, we predicted that allelic dropout would be positively related to sample age and locus length, and that false alleles would be positively related to sample age (Wandeler et al. 2003). Based on the processes generating false alleles (Broquet and Petit 2004; Wandeler et al. 2007) and previous research (Wandeler et al. 2003), we did not expect locus length to influence the rate of false alleles in historical samples.

MATERIALS AND METHODS

Sample Collection and Laboratory Methods

Museum sampling.—We collected biological samples from historical kit fox specimens (i.e., individual kit foxes) maintained by the Natural History Museum of Utah (UMNH). Kit fox specimens were originally collected between 1936 and 1970 in the Great Basin Desert (although five specimens did not have information on the date of collection; [Supplementary Data SD1](#)). We collected samples from each specimen from up

to two different specimen parts (i.e., skulls, skins, or both) and three different sample sources when possible (i.e., two from the skull and one from the skin; Fig. 1, Supplementary Data SD1). From skulls, we sampled 1) maxilloturbينات (hereafter nasal bones; sensu Wisely et al. 2004), and 2) inner cranial cavity bones (i.e., tentorium and internal occipital protuberance; hereafter cranial bones; sensu Miller and Waits 2003). We carefully dislodged nasal and cranial bone samples with sterilized tweezers or forceps. From skins, we removed a portion of the toepads with a sterile razor blade. We collected each sample onto sterile foil prior to weighing and transferring to a coin envelope. All samples were stored in a sealed plastic bag with silica desiccant until extraction. All sampling and procedures were approved by the Natural History Museum of Utah and collection techniques attempted to minimize damage to specimens.

DNA extraction and PCR amplification.—To minimize the possibility of contamination, we extracted DNA and set up PCR reactions for historical samples in an isolated clean laboratory that had not previously been used to store or process high-quality (e.g., blood, tissue) or contemporary (e.g., scats) sources of vertebrate DNA (Greenwood et al. 1999; Wisely et al. 2004; Lonsinger et al. 2018a). To compare PCR success rates and genotyping error rates (i.e., allelic dropout and false

alleles) across sample sources, we attempted to standardize the amount of material extracted from each sample. We extracted DNA from ~ 0.06 g of each sample (see Results); when < 0.06 g was available, we used the entire sample. We ground cranial and nasal bone samples into a powder using liquid nitrogen and a sterilized mortar and pestle. For toepads, we used a sterile razor blade to cut each toepad to the smallest pieces possible. We then extracted DNA from each sample using the “silica” method (Boom et al. 1990; Höss and Pääbo 1993) to a final elution of 180 µL; negative controls were included with each extraction event to monitor for contamination.

We amplified all historical samples with a multiplex of nine nDNA microsatellite loci (Ostrander et al. 1993, 1995; Fredholm and Wintero 1995; Holmes et al. 1995; Francisco et al. 1996; Cullingham et al. 2006; Table 1), which have been used to characterize the genetic diversity of historical and contemporary kit fox populations (Lonsinger et al. 2018a). The PCR fragment length for each microsatellite locus ranged in size from 70 to 251 base pairs (bp; Table 1). In addition to the microsatellite loci, two sex identification primers developed for red fox (*V. vulpes*; Berry et al. 2007) were included in the multiplex as part of a concurrent study that used the same samples (Lonsinger et al. 2018a). The X-linked *CF-hprt* primer (Forward: 5′-AGT CAA CGG GGG ACA TAA AAG-3′; Reverse: 5′-ACC ATT TTT

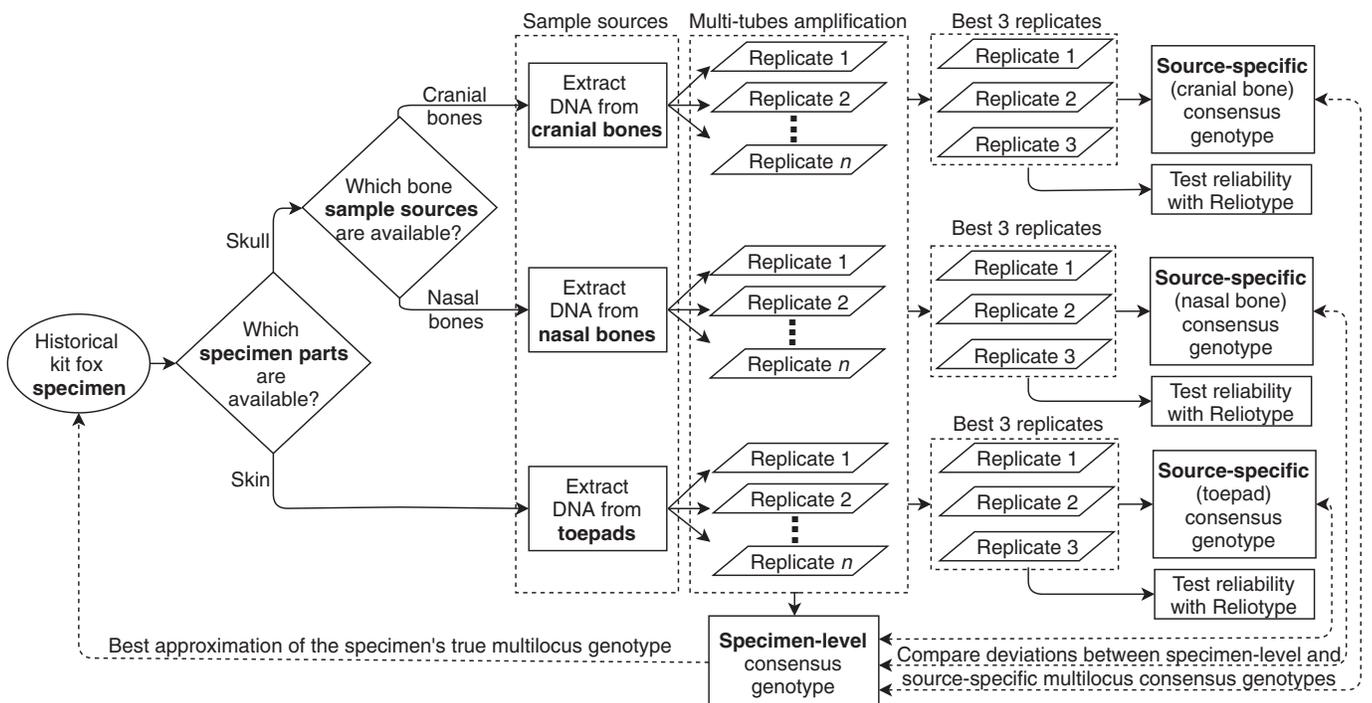


Fig. 1.—Diagram illustrating the workflow used to evaluate the reliability of genetic data derived from three sources (cranial bones, nasal bones, and toepads) from 76 historical kit fox (*Vulpes macrotis*) specimens originally collected between 1936 and 1970 in the Great Basin Desert of Utah (United States), maintained by the Natural History Museum of Utah (UMNH), and sampled for DNA in 2013 (see Table 1). For each specimen (i.e., individual kit fox), multiple independent amplifications across all available sample sources were combined to establish a specimen-level consensus genotype. For each available sample source, a source-specific consensus genotype was derived from the best three replicates from that source. The best replicates were defined as those with the highest amplification success rates (i.e., number of successful amplifications/total number attempted). Mean reliability across replicates and the proportion of replicates with a reliability ≥ 95% were estimated with the program RELIOTYPE (Miller et al. 2002). Source-specific consensus genotypes were compared to the specimen-level consensus genotype to calculate the mean proportion of deviations among loci.

Table 1.—The repeat motif (Repeat), range of allele sizes (Length) in base pairs (bp), number of alleles (N_A), allelic richness (A_r), observed (H_o) and unbiased expected heterozygosity (H_e), and amplification (PCR) success and genotyping error (i.e., allelic dropout [ADO] and false alleles [FA]) rates for nine microsatellite loci amplified for 76 historical kit fox (*Vulpes macrotis*) specimens originally collected between 1936 and 1970 in the Great Basin Desert of Utah (United States), maintained by the Natural History Museum of Utah (UMNH), and sampled for DNA in 2013. PCR success was the number of successful amplifications/total number attempted across replicates. ADO represented the proportion of replicates in which only one allele in a heterozygous genotype amplified, whereas FA represented the proportion of replicates in which an allele amplified that was not present in the consensus genotype.

Locus	Repeat	Length (bp)	N_A	A_r	H_o	H_e	PCR	ADO	FA
CXX103	(TG) ₁₇	70–86	6	5.3	0.58	0.57	91.7%	5.9%	3.7%
FH2088	(TTTA) ₁₀ (TTCA) ₄	117–149	8	7.9	0.69	0.68	86.6%	9.5%	1.7%
CXX250	(AC) ₁₈ A ₂ (TC) ₄	135–145	6	6.0	0.44	0.46	32.9%	14.6%	3.1%
FH2001	(GATA) ₈	145–161	6	5.3	0.54	0.49	84.7%	9.1%	0.8%
CPH3	(GA) ₂ TA(GA) ₁₇	154–160	4	3.5	0.39	0.42	78.8%	6.7%	0.9%
FH2054	(GATA) ₁₆	163–187	7	6.5	0.66	0.70	76.9%	10.6%	1.8%
CXX377	(AC) ₁₂	175–195	12	11.4	0.71	0.82	80.5%	12.0%	0.9%
FH2010	(ATGA) ₁₀	216–232	5	4.7	0.75	0.72	75.3%	7.6%	1.0%
VVE-M19	(TTTC) ₂₅	214–251	9	8.6	0.73	0.73	75.7%	6.8%	0.5%

GGA TTA TAC TGC–3′) produced a 148 bp product, whereas the *VV-sry* primer (Forward: 5′–GAA TCC CCA AAT GCA AAA CTC G–3′; Reverse: 5′–CCA TTT TTC GGC TTC TGT AAG C–3′) linked to testis development in foxes produced a 75 bp product. The PCR conditions for each primer pair in the multiplex and the thermal profile followed methods detailed in Lonsinger et al. (2018a). We set up PCRs in the same facility that was used for extraction of historical samples to minimize the risk of contamination. We added DNA from a contemporary kit fox scat sample that had previously amplified across loci (Lonsinger et al. 2015) as a positive control. Each positive control was added to the PCR plate in a separate clean lab used to extract noninvasive genetic samples by uncapping only the PCR positive well. All PCR procedures were conducted on a BioRad Tetrad thermocycler (Bio-Rad, Hercules, California) with negative and positive controls to monitor contamination and amplification success, respectively. We used a 3130xl DNA Analyzer (Applied Biosystems, Foster City, California) and Genemapper 3.7 (Applied Biosystems) to visualize results and score allele sizes, respectively.

Data Analysis

PCR amplification success and genotyping error rates.—We employed a multi-tubes approach to generate multilocus genotypes, assess DNA quality, and limit genotyping error rates (Taberlet et al. 1996). The multi-tubes approach is commonly used when working with degraded DNA and is one of the most accepted methods for quantifying DNA quality and limiting genotyping errors (Zhan et al. 2010). The multi-tubes approach employs multiple PCR amplifications (i.e., replicates) from each DNA extract (Taberlet et al. 1996; Pompanon et al. 2005; Fig. 1); comparisons of multilocus genotypes across replicates are then used to establish consensus genotypes (i.e., the best approximation of the true multilocus genotype) and quantify genotyping error rates. For each specimen, we performed three replicates per available sample source (i.e., cranial bones, nasal bones, or toepads; Fig. 1). We determined a specimen-level consensus genotype at each locus by combining replicates for each specimen across sample sources (Taberlet et al. 1996;

Pompanon et al. 2005), and requiring that heterozygous and homozygous alleles be observed at least two and three times, respectively (Broquet and Petit 2004; Lonsinger and Waits 2015). After running the initial three replicates per sample source, we increased the number of replicates until we achieved specimen-level consensus genotypes (Fig. 1) across at least seven loci for the specimen, or until we reached a maximum of six replicates for each available sample source with successful amplifications. Additional replicates beyond the initial three replicates per sample source were randomly selected among available sample sources (excluding sample sources that failed to produce a positive amplification over the initial three replicates). The specimen-level consensus genotype was used as the best approximation of the specimen's true multilocus genotype (Fig. 1).

We quantified DNA quality based on measures of amplification success and genotyping errors across replicates. We calculated PCR amplification success rates as the number of successful amplifications/total number attempted, where the total number attempted was the number of replicates performed multiplied by the number of microsatellite loci (i.e., nine). We calculated PCR sample-level success rates as the proportion of replicates that amplified at $\geq 50\%$ of the microsatellite loci. We calculated PCR amplification and sample-level success rates by sample sources and overall (i.e., combined across sample sources). To estimate genotyping error rates, we compared each replicate to its respective specimen-level consensus genotype and classified the observation of an allele not present in the specimen-level consensus genotype as a false allele, and the amplification of only one allele in a heterozygous specimen-level consensus genotype as allelic dropout (Broquet and Petit 2004; Lonsinger and Waits 2015). We did not detect any evidence of contamination (see “Results”) and therefore all PCR replicates were included in the PCR success and genotyping error rate calculations. We compared replicates to establish specimen-level consensus genotypes and calculated amplification success and genotyping error rates with the R programming language using ConGenR (Lonsinger and Waits 2015; R Core Team 2018).

Genetic diversity.—We calculated number of alleles, observed heterozygosity (H_o), and Nei's unbiased expected heterozygosity (H_e) across each of the nine microsatellite loci with GenAlEx 6.5 (Peakall and Smouse 2006), and allelic richness with FStat 2.9.3.2 (Goudet 1995). For each locus, we quantified locus-specific rates of PCR success, false alleles, and allelic dropout across all sample sources as before with ConGenR (Lonsinger and Waits 2015; R Core Team 2018).

Mixed-effects logistic regression.—We used mixed effects logistic regression analyses to evaluate the impact of locus length, sample weight, sample age, and sample source on each of PCR success, allelic dropout, and false alleles. We evaluated PCR success, allelic dropout, and false alleles as binary response variables with mixed-effect logistic regression models. For models of PCR amplification success, a successful amplification was coded as a one, whereas a failure to amplify was coded as a zero. For models of each form of genotyping error (i.e., allelic dropout or false allele), the presence of an error was coded as a one, whereas the lack of an error (for a sample with a successful amplification) was coded as zero. In an effort to account for a possible influence of sample source on the proportion of cells available in the sample weight, we also attempted to examine the interaction effect between sample weight and sample source. The mixed model was needed to account for the random effect of specimen. Independent variables not significant at $\alpha = 0.05$ were removed from the model. Prior to model evaluation, multicollinearities were investigated using generalized variance inflation factors (GVIFs—Fox and Monette 1992). A variance inflation factor (VIF) is a metric that indicates how much the variance of an independent variable's coefficient estimate will be inflated as a consequence of having the other independent variables in the model. A VIF will generally be larger when there is greater overlap in information contained in that variable and the information contained in the other independent variables (i.e., when there is a greater multicollinearity). Furthermore, if independent variables involved in a strong multicollinearity are included in a model together, P -values will often be inflated and coefficient estimates can vary substantially, including changing sign. VIFs are generally useful for assessing strictly numeric independent variables, whereas GVIFs are an extension to both numeric and categorical variables. While GVIFs were quite low (< 1.3) for all independent variables (i.e., locus length, sample age, sample weight, and sample sources), low GVIFs are not necessarily indicative that there are no problems due to multicollinearity. Because of this, we additionally tested each independent variable alone in the model, and also tested the addition of previously removed variables back into a model again later to help ensure that multicollinearities were not contributing to their initial removal. This process indicated that sample source and specimen weight were highly significant in modeling PCR success when each was in the model by itself ($P < 0.0001$), but specimen weight was far from being significant ($P = 0.65$) when both variables were included, suggesting that these two variables were collinear. Examining an interactively rotatable, three-dimensional plot (with jittering of points) of specimen

weight (referred to as specimen amount in [Supplementary Data SD2](#)) against the two dummy variables used to represent sample source, we saw that there were very few low values of specimen weight for the sample source of nasal bone (this was more difficult to properly visualize in two-dimensional plots). Hence, a collinearity existed in our data between specimen weight and one of the dummy variables used to represent sample source. The implication is that we were unable to reliably include both variables in the model together. This also made it impossible to investigate an interaction effect between specimen weight and sample source.

After final models were attained for each of the three response variables (i.e., probabilities of PCR success, false alleles, and allelic dropout), a log-odds estimating function was obtained, and from that a function estimating the probability of success ($\hat{\pi}$) was obtained. Sample age had 306 missing values (from the five specimens; [Supplementary Data SD1](#)), representing 5.9% of the data. These observations were removed from analyses when sample age was considered, but ultimately this variable was not significant in any of the final models and, hence, the entire data set was used for all the final models. All mixed-effects logistic regression analyses were conducted using program R ([R Core Team 2018](#)).

Reliability.—To evaluate the relative quality of sample sources, we compared the specimen-level consensus genotype to source-specific consensus genotypes determined based on the best three replicates for each source-specific sample, following Hedmark and Ellegren (2005) (Fig. 1). We identified the best replicates as those replicates that had the highest amplification success rates across loci (i.e., the highest number of successful amplifications/total number attempted). Based on this definition, we had at least three replicates for each sample source available across all specimens. While considering the best replicates may provide an optimistic estimation, it reduces the potential of including replicates that failed due to stochastic processes (e.g., pipetting errors) and because it was consistent across sample sources, should still provide a measure of relative performance. For each available sample source within each specimen, we used ConGenR to determine a source-specific consensus genotype by combining the best three replicates, and requiring that heterozygous and homozygous alleles be observed at least two and three times, respectively (Broquet and Petit 2004; Lonsinger and Waits 2015; R Core Team 2018). We then calculated the mean proportion of loci from each source-specific consensus genotype that deviated from the specimen-level consensus genotype (Fig. 1), and interpreted greater departures as indicating lower reliability. To further evaluate the relative reliability of each sample source, we used the program RELIOTYPE (Miller et al. 2002) to assess the reliability of the three best replicates (Fig. 1). RELIOTYPE uses a maximum likelihood-based approach that considers allelic dropout genotyping errors and population allele frequencies to quantify the reliability of a multilocus genotype. For each sample source, we then summarized the results by the mean reliability across replicates, proportion of replicates that were reliable (i.e., reliability $\geq 95\%$), and the mean predicted number of additional replicates per locus required to achieve $\geq 95\%$ reliability

(for those samples with reliability < 95%). We excluded source-specific samples for specimens characterized by only a single sample source in reliability tests, as their inclusion would have artificially increased reliability since the replicate set used to define both the specimen-level consensus genotype and source-specific consensus genotype would be identical.

RESULTS

Museum sampling.—We collected samples from 76 unique kit fox specimens originally collected between 1936 and 1970 from Utah (70), Nevada (2), and Colorado (2; [Supplementary Data SD1](#)). The location and date of collection were unknown for two and five specimens, respectively. Nineteen specimens included both a skull and skin, yielding samples from all three source materials (cranial bone, nasal bone, and toepad). Only a skull was available for 39 specimens, providing cranial and nasal bone samples. Only partial skulls were available for three specimens; we collected nasal bone samples from two and a cranial bone sample from one. The remaining 15 specimens included only skins, from which we collected toepad samples. The mean mass of nasal bone samples ($\bar{X} \pm SD$; 0.16 ± 0.08 g) was greater than cranial bone samples (0.08 ± 0.04 g). Toepads yielded in the lowest mean sample mass (0.07 ± 0.02 g). Specimen details are available in [Supplementary Data SD1](#).

DNA extraction, PCR success, and genotyping errors.—Approximately 0.057 ± 0.01 g of material was used in each extraction. We did not detect any evidence of contamination based on extraction negatives and PCR negatives, nor did we detect any evidence of cross-contamination among specimens (as might as might possibly occur in NHCs due to samples being stored together). Cross-contamination among specimens would be identified, had it occurred, through the confirmation of more than two alleles per locus across multiple loci. Mean number of replicates performed was comparable among sample sources: cranial bones = 3.2 ± 0.6 , nasal bones = 4.2 ± 1.3 , and toepads = 4.1 ± 1.1 . Overall amplification success rates (number of successful amplifications/total number attempted) were highest for toepads and lowest for cranial bones ([Table 2](#)). Sample-level amplification success rates (i.e., proportion of replicates amplifying at $\geq 50\%$ of the loci) followed a similar pattern and were only slightly higher than overall success rates ([Table 2](#)). Cranial bones had the highest genotyping error rates, while toepads produced the fewest errors ([Table 2](#)). Overall and among sample sources, allelic dropout rates were approximately 5–6 times higher than false allele rates ([Table 2](#)). When considering all three sample sources, overall amplification success rates ranged from 32.9% to 91.7% across loci, whereas allelic dropout and false allele rates ranged from 5.9% to 14.6% and 0.5% to 3.7%, respectively ([Table 1](#)). We achieved specimen-level consensus genotypes at ≥ 7 nDNA microsatellite loci for 70 of 76 (92.1%) kit fox specimens; these multilocus consensus genotypes were achieved with a mean of 7.44 ± 2.4 replicates across all source-specific samples.

Mixed-effects logistic regression.—As described in the methods, sample weight was not considered in the final models due to its collinearity with sample source. Sample age did not

Table 2.—Polymerase chain reaction (PCR) amplification success rates and genotyping error rates for samples collected from historical kit fox (*Vulpes macrotis*) specimens originally collected between 1936 and 1970 in the Great Basin Desert of Utah (United States), maintained by the Natural History Museum of Utah (UMNH), and sampled for DNA in 2013. Overall PCR success was the number of successful amplifications/total number attempted (where the total number attempted is the number of replicates performed multiplied by nine microsatellite loci), whereas sample-level PCR success was the proportion of replicates amplifying at $\geq 50\%$ of the loci. Allelic dropout represented the proportion of replicates in which only one allele in a heterozygous genotype amplified. False alleles represented the proportion of replicates in which an allele amplified that was not present in the consensus genotype. All proportions are based on all replicates from the respective sample sources and across all sampled specimens.

Sample source	PCR success		Genotyping errors	
	Overall	Sample-level	Allelic dropout	False alleles
Cranial bones	58.8%	60.4%	14.3%	2.4%
Nasal bones	78.1%	84.3%	8.8%	1.5%
Toepads	94.9%	99.3%	4.1%	0.9%
All samples	75.9%	80.2%	8.8%	1.5%

significantly influence the probabilities of PCR success, false alleles, or allelic dropout. Consequently, the final model for all three response variables (i.e., PCR success, false alleles, or allelic dropout) was the response variable \sim locus length + sample source + (1|Specimen). Detailed analysis considerations are available in [Supplementary Data SD2](#).

Analysis for PCR success led to an estimating model of

$$\hat{\pi} = \frac{1}{1 + e^{-1.075 + 0.003867 L - 1.734 I_{NB} - 2.959 I_{TP}}},$$

where $\hat{\pi}$ is the probability of PCR success, L is locus length, I_{NB} is 1 if the sample source is nasal bone and zero otherwise, and I_{TP} is 1 if the sample source is toepad and zero otherwise. Sample age was dropped from the final model due to its non-significance ($P = 0.97$). The implication of this model estimate is that the probability of PCR success decreases as a function of increasing locus length, and that sample sources of cranial bone, nasal bone, and toepad resulted in progressively higher probabilities of PCR success, respectively ([Fig. 2A](#)).

The analysis of false alleles led to the estimating model of

$$\hat{\pi} = \frac{1}{1 + e^{2.2583 + 0.01115 L + 0.5092 I_{NB} + 1.0366 I_{TP}}},$$

where $\hat{\pi}$ is the probability of a false allele. Sample age was dropped from the model due to its non-significance ($P = 0.44$). This equation implies that the probability of a false allele decreases as a function of increasing locus length, and sample sources of cranial bone, nasal bone, and toepad resulted in progressively lower probabilities of false allele error, respectively ([Fig. 2B](#)).

Analysis of allelic dropout resulted in an estimating model of

$$\hat{\pi} = \frac{1}{1 + e^{2.5527 - 0.003836 L + 1.336 I_{NB} + 1.934 I_{TP}}},$$

where $\hat{\pi}$ is the probability of an allelic dropout. Sample age was dropped from the model due to its non-significance ($P = 0.33$).

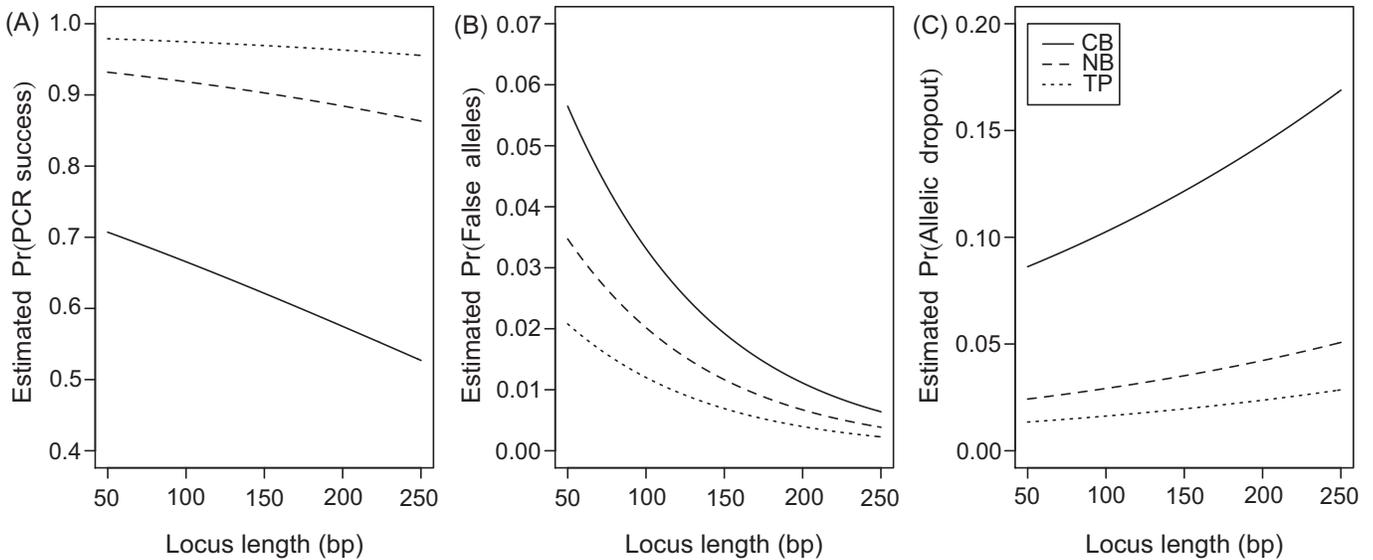


Fig. 2.—Mixed-effects logistic regression model results for the probabilities of (A) PCR success, (B) false alleles, and (C) allelic dropout as a function of locus length in base pairs (bp) for three sample sources—cranial bone (CB), nasal bone (NB), and toepad (TP)—from historical kit fox (*Vulpes macrotis*) specimens originally collected between 1936 and 1970 in the Great Basin Desert of Utah (United States), maintained by the Natural History Museum of Utah (UMNH), and sampled for DNA in 2013.

Table 3.—Reliability in multilocus consensus genotypes from cranial bone, nasal bone, and toepad samples from historical kit fox (*Vulpes macrotis*) specimens originally collected between 1936 and 1970 in the Great Basin Desert of Utah (United States), maintained by the Natural History Museum of Utah (UMNH), and sampled for DNA in 2013. Proportion (%) of deviations indicates the mean proportion of loci in which the source-specific consensus genotype differed from the specimen-level consensus genotypes. Proportion (%) reliable indicates the proportion of source-specific samples that achieved $\geq 95\%$ reliability with three replicates, whereas mean reliability indicates the mean reliability across source-specific samples. For samples with $< 95\%$ reliability, additional replicates (Add reps) indicate the average predicted number of additional replicates (over the three considered) per locus required to achieve $\geq 95\%$ reliability.

Sample source	% Deviations		% Reliable	Reliability		Add reps	
	Mean	SD		Mean	SD	Mean	SD
Cranial bones	53.4%	39.2%	51.7%	66.5%	41.1%	1.6	2.7
Nasal bones	23.6%	31.2%	72.4%	84.7%	30.1%	0.5	1.5
Toepads	13.5%	10.2%	94.7%	97.8%	7.5%	0.02	0.2

The implication of this model estimate is that the probability of allelic dropout increases as a function of locus length, and that sample sources of cranial bone, nasal bone, and toepad resulted in progressively lower probabilities of allelic dropout, respectively (Fig. 2C).

Source-specific sample reliability.—Source-specific consensus genotypes resulting from cranial bones deviated the most from the specimen-level consensus genotypes (i.e., that were based on all available sample sources and expected to best represent the true multilocus genotype), whereas toepads produced source-specific consensus genotypes that were most consistent with the specimen-level consensus genotypes (Table 3). Among sample sources, toepads were reliable most frequently and had the highest mean reliability based on maximum likelihood estimates (Table 3). Nasal bones were not as reliable as toepads, but were more reliable than cranial bones, which were reliable only about one-half of the time (Table 3). Consequently, cranial bones were predicted to require the greatest number of additional replicates per locus to achieve reliable source-specific consensus genotypes (Table 3).

DISCUSSION

Natural history collections maintain millions of specimens that can offer insights into evolutionary processes and population response to altered ecosystem dynamics (Wandeler et al. 2007). As molecular techniques have improved, and costs have declined, the number of conservation genetic studies employing museum specimens has increased (Casas-Marce et al. 2010), and this trend is predicted to continue (McDonough et al. 2018). With increasing demands on finite resources, researchers and managers of NHCs require information on the reliability of commonly used sample sources to inform requests and decisions related to destructive sampling of museum specimens. Differentiating quality among sample sources can direct destructive sampling to the sample source predicted to yield the most reliable data, potentially reducing both damage to museum specimens and laboratory costs (via fewer replicates required to generate reliable data). We examined the quality of genetic data derived from three common sample sources used for carnivoran specimens and demonstrated that sample source

and locus length influenced the quality and reliability of resulting nDNA multilocus genotypes. We used these genetic data to demonstrate how deviations between source-specific and specimen-level consensus genotypes can be used to identify the optimal sample source; this provides researchers with a technique for making comparisons among samples that can be used in future pilot studies to optimize museum specimen sampling. We demonstrated that deviation-based measures of data reliability were consistent with measures of reliability based on maximum-likelihood. Finally, our study provided a quantitative assessment of how differences in sample-source quality for historical specimens translated to variation in laboratory efforts (i.e., number of replicates required to generate reliable data), and therefore, costs.

Factors influencing DNA degradation.—Soft tissue samples (e.g., skin, toepad) and hard tissue samples (e.g., bones, teeth) are among the most commonly sampled sources for historical carnivore DNA. Previous research has suggested that hard tissue samples from historical or ancient specimens tended to produce higher DNA amplification rates and fewer genotyping errors than soft tissue samples (e.g., Cooper et al. 1992; Wisely et al. 2004; Nyström et al. 2006; Casas-Marce et al. 2010; Jansson et al. 2014; McDonough et al. 2018). Consequently, researchers interested in securing the highest quality DNA may be inclined to sample hard tissues, whereas museum curators may prefer the less-damaging collection of soft tissue when both sample sources are available (Wandeler et al. 2007). In contrast to this previous work and our predictions, we found that soft tissue toepads had significantly higher PCR amplification success rates and fewer genotyping errors than bone samples. Similar patterns were observed in historical Eurasian lynx specimens, with footpads yielding higher microsatellite amplification success rates and lower error rates than bones (Polanc et al. 2012). Variation in the relative performance of sample sources among studies may relate to differences in preparation and preservation of historical specimens, but these details are rarely documented or reported for comparison among studies.

The DNA within historical specimens is likely degraded and, therefore, microsatellite loci with shorter PCR amplicons are expected to perform better (i.e., have greater PCR success rates and lower genotyping error rates) than longer loci (Wandeler et al. 2003; Buchan et al. 2005). Previous researchers have recommended using microsatellite loci with allele lengths smaller than approximately 250 bp when working with historical DNA (Nielsen et al. 1999). Our loci conformed to this recommendation, and we had relatively high PCR success rates (> 75%) across eight of the nine nDNA loci. Within the range of locus lengths analyzed, we still found higher success rates for shorter loci and higher allelic dropout rates for longer loci, as expected. Within a locus, a wide range of allele sizes could also result in higher rates of allelic dropout for longer alleles than shorter alleles (i.e., “large allele dropout”—Nielsen et al. 1999), although we did not detect this pattern with our data.

In contrast to our initial prediction, false allele rates appeared to decrease with increasing locus length. Closer examination of locus-specific false allele rates suggest that this pattern may

have been driven by two loci (i.e., CXX103 and CX250) that were among the three shortest loci, and had relatively high false allele rates. Some loci can perform poorly (i.e., have high failure rates and error rates) even if their allele lengths are short. For example, CXX250 was among the shortest loci used in this study but had markedly lower PCR success relative to other loci, the highest allelic dropout rate, and the second highest rate of false alleles. This locus performed poorly in an earlier study of historical arctic fox specimens (Nyström et al. 2006) and was removed from analyses of effective population size for kit foxes due to poor performance (Lonsinger et al. 2018a). In contrast, CXX103 was the locus with the highest PCR success rate and lowest rate of allelic dropout, but the highest rate of false alleles. One potential cause of a false allele is an unambiguous peak resulting from spectral overlap with another marker in the multiplex. In our multiplex, the sizes of CXX103 and CXX250 overlapped with that of the *VV-sry* (75 bp) and *CF-hprt* (148 bp) sex identification markers (Berry et al. 2007), respectively, and spectral overlap may have generated a higher frequency of false alleles than would be typically expected for these shorter loci. Although CXX103 and CXX250 are dinucleotide repeat motifs, which may be more prone to slippage than tetranucleotides (Broquet et al. 2007), we did not see high false allele rates in the other dinucleotide loci (i.e., CPH3 and CXX377). These patterns suggest researchers may be able to improve the quality of genetic data from historical samples by selecting shorter loci (perhaps with narrower ranges of allele lengths), as noted by others (e.g., Nielsen et al. 1999; Wandeler et al. 2003), but that loci performance should be evaluated to identify and cull out those that may be error prone or have low amplification success rates.

DNA is expected to degrade over time, and both genetic (Wandeler et al. 2003; Casas-Marce et al. 2010) and genomic (McDonough et al. 2018) studies of historical DNA have reported decreasing DNA quality with increasing sample age. Studies of DNA degradation using microsatellite loci to evaluate noninvasive genetic samples (e.g., fecal DNA) have commonly reported nonlinear patterns of degradation, with PCR success rates declining precipitously at first and the rate of decline slowing over time (DeMay et al. 2013; Lonsinger et al. 2015; Woodruff et al. 2015). If historical DNA degradation follows similar nonlinear patterns, the effect of sample age may be more significant when sampling over longer historical periods, or when comparing historical samples to contemporary samples. For example, Nielsen et al. (1999) reported that some microsatellite loci used to analyze historical salmonid scales had markedly reduced amplification success when they were more than 10 years old. Wandeler et al. (2003) reported increasingly nonlinear patterns of degradation with increasing locus length for four microsatellite loci used to analyze historical red fox teeth samples, with the most precipitous declines in amplification success occurring over the first ~ 25 years of storage. At the time of extraction, our samples spanned 35 years and ranged from 43 to 77 years old, yet we did not detect an effect of sample age on amplification success or genotyping error rates. Thus, we suggest that if the rate of DNA degradation for

our specimens was nonlinear, the most precipitous rate of DNA degradation may have occurred prior to 43 years (the age our youngest specimen), and consequently the time span covered by our samples may not have been sufficiently long to detect an effect of age. In contrast, [McDonough et al. \(2018\)](#) did find an effect of age on nDNA of historical samples (i.e., bones, claws, osteocrusts, and skin) using a next-generation sequencing approach, and this could relate to potentially greater resolution offered by sequence data, or because they sampled specimens that covered a period approximately twice as long (1898–1968) as our range. [Casas-Marce et al. \(2010\)](#) and [Wandeler et al. \(2003\)](#) also detected an effect of age on nDNA microsatellite loci for historical samples that spanned 53 years (1954–2006) and 32 years (1969–2000), respectively, but these studies both compared historical samples to more contemporary samples that were less than 10 years old. Although degradation of historical DNA is likely variable, further research is needed to characterize broadly how DNA degradation rates within museum specimens and among sample sources vary over time.

Sample weight may also influence amplification success and genotyping errors. Sample weight significantly influenced amplification success of historical red fox teeth samples ([Wandeler et al. 2003](#)). Similarly, sample weight significantly influenced nDNA microsatellite locus amplification success for historical Iberian lynx (*L. pardinus*) sampled via hard (e.g., claw, nasal bones, and teeth) and soft (e.g., skin, toepads) tissues ([Casas-Marce et al. 2010](#)). In both of these studies, increasing sample weight resulted in increased probability of PCR amplification success. We attempted to standardize sample weights used for extraction, but there was still variation in weight both within and among sample sources. Among the sample sources we evaluated, the toepads had the lowest mean sample weight but still produced the highest amplification success rates, lowest genotyping error rates, and highest reliability. Thus, despite the lower sample weights, toepads outperformed both bone sample sources. Between the bone sample sources, nasal bones had higher mean sample weights than cranial bones and this may have led to the higher amplification success rates, lower genotyping error rates, and higher reliability attributed to nasal bones over cranial bones. Still, this suggests that achieving reliable results may be more challenging with cranial bones than nasal bones for some species, as the sampling of kit fox cranial bones (i.e., the tentorium and internal occipital protuberance) consumed all of the available material, whereas the sampling of nasal bones (i.e., maxilloturbinates) only consumed a portion of the available material. Due to patterns of collinearity between sample weights and sample sources, sample weight was excluded from consideration in our final models, and we were unable to formally evaluate the influence of sample weight. Sample weight may be confounded to some degree by differences in cell densities among different tissue types (e.g., hard versus soft tissues) or by variation in cell density among species or individuals, but we were unable to assess this with our study design.

The quality and reliability of genetic data from historical specimens is likely influenced by specimen preparation and

preservation techniques ([Wandeler et al. 2007](#)). Preparation of soft tissue, such as tanning, arsenic or salting preservation, and the use of formalin, can result in DNA degradation and inhibition that reduces the quality of resulting DNA data ([Hedmark and Ellegren 2005](#); [Casas-Marce et al. 2010](#); [Polanc et al. 2012](#)). Preparation of hard tissue (e.g., bleaching and boiling) may also degrade DNA ([Wandeler et al. 2003](#); [McDonough et al. 2018](#)). Still, hard tissue is believed to provide better protection of DNA and potentially minimize degradation of DNA from preparation techniques ([Cooper et al. 1992](#); [Greenwood et al. 1999](#)). As is common with historical samples, the preparation and preservation histories of our samples were unknown, and we were therefore unable to evaluate the influence of specimen preservation on DNA quality and reliability. More than one-half of our samples were collected by a single researcher (H. Egoscue; [Supplementary Data SD1](#)), and there was not a discernible pattern between amplification success rates and researchers that might imply different preparation methods among researchers impacted the results.

Influence of DNA degradation on data reliability and cost.—Previous studies have investigated amplification success rates and, to a lesser extent, genotyping error rates of various sample sources for historical specimens, but have not evaluated the reliability of the resulting genetic data and formally quantified its implications for laboratory effort (and therefore, costs). Detection of a genotyping error (i.e., allelic dropout or false allele) is conditional upon a successful PCR amplification. Thus, despite the common use of amplification success as a metric for DNA quality (e.g., [Wisely et al. 2004](#); [Casas-Marce et al. 2010](#); [Polanc et al. 2012](#)), genotyping error rates may be a better measure of DNA quality when error rates vary among samples sources. Our results indicate that some sample sources are more prone than others to genotyping errors. As the probability of genotyping errors increases, additional PCR replicates are required to reconcile genotypic differences observed among replicates, increasing laboratory costs for generating reliable data.

Our results indicate that allelic dropout occurs more frequently than false alleles, similar to the findings of others (e.g., [Miller et al. 2003](#); [Wandeler et al. 2007](#)). Allelic dropout results from low quantities of DNA, whereas false alleles tend to be artifacts of polymerase slippage during PCR ([Broquet and Petit 2004](#)). Genotyping errors are expected to have low reproducibility ([Waits and Paetkau 2005](#)), but the relative frequencies of allelic dropouts and false alleles, combined with the processes generating genotyping errors, make allelic dropout more likely to be reproduced than false alleles in multiple replicates of a sample. Thus, it is more likely to incorrectly characterize a heterozygous genotype as homozygous than vice versa, and this could artificially result in underestimation of genetic diversity from historical specimens with low-quality DNA ([Taberlet et al. 1999](#); [Wandeler et al. 2007](#)). For example, [Nyström et al. \(2006\)](#) could not rule out allelic dropout in a microsatellite locus as the cause of heterozygote deficiency in a study of arctic foxes based on historical teeth, bone, and skin samples.

We used two different approaches to evaluate reliability of genotypic data by sample source, and both methods resulted

in similar conclusions. Data reliability was strongly influenced by sample source, with the mean reliability of cranial bones (66.5%) and nasal bones (84.7%) being well below critical thresholds (e.g., 95–99% reliability) often used to cull low-quality samples. On average, source-specific consensus genotypes from cranial bones and nasal bones deviated from the specimen-level consensus genotypes at more than one-half and nearly a quarter of loci, respectively. In contrast, mean reliability of toepads was > 95% and source-specific consensus genotypes deviated from specimen-level consensus genotypes at only a single locus on average. Consequently, for our historical specimens, reliance on cranial bones or nasal bones alone would be expected to produce data with lower reliability (and potentially artificially suppressed heterozygosity) relative to data generated from only toepads.

Degraded genetic samples can influence laboratory efforts (and costs) at various stages. It has become common practice to screen degraded samples (e.g., noninvasive genetic samples [Lonsinger et al. 2018b] and historical genetic samples [Polanc et al. 2012]) and exclude samples failing at > 50% of loci (Paetkau 2003). Therefore, consideration of DNA quality and reliability among sample sources can direct researchers to sample sources most likely to produce high amplification success rates, reducing the proportion of samples that are excluded following repeated amplification attempts. For degraded genetic samples that successfully amplify at $\geq 50\%$ of loci, researchers commonly then use a multi-tubes approach that compares multiple PCR amplifications (i.e., replicates) to resolve differences among replicates and establish consensus genotypes (Taberlet et al. 1996). For our historical specimens, our data suggested that significant differences in probability of PCR success among sample sources would result in higher laboratory costs for screening cranial bones than nasal bones. Still, when considering the mean number of replicates performed and additional replicates required to achieve reliable data, our data suggested that despite significant differences in probability of genotyping errors, the total number of replicates (approximately six), and therefore costs, required to generate reliable data were similar for those cranial and nasal bones that were retained following screening. Toepads had the highest PCR success rates and achieved acceptable levels of reliability ($\geq 95\%$) with two fewer replicates (approximately four) on average than bone samples. For our historical specimens, then, toepad samples would have the lowest laboratory costs and result in the least amount of unnecessary destructive sampling (i.e., fewer samples collected via destructive sampling would fail to produce reliable data).

Implications for future studies.—This study demonstrates how genetic data obtained from historical mammalian specimens varied based on the source of the sample, and presents methods for evaluating data reliability and laboratory costs that can be applied to future studies using historical specimens. Although our results are based on a single species and set of samples, the findings inform studies of mammalian species in which multiple potential sample sources are available. We

concur with Wandeler et al. (2007) that pilot studies are invaluable for investigations using historical specimens. This study highlights the importance of considering sample source and demonstrates how a pilot study can 1) differentiate among data reliability by sample source, 2) identify loci that are expected to have high performance or be problematic, 3) decrease laboratory costs, and 4) minimize destructive sampling.

Our results contrast with many earlier studies that found hard tissue sample sources produced higher quality genetic (microsatellite) data (e.g., Cooper et al. 1992; Nyström et al. 2006; Casas-Marce et al. 2010). Collection of hard tissue samples can be more destructive than sampling soft tissues (Wandeler et al. 2007), and it would therefore be prudent for future studies relying on museum specimens to evaluate the sample source quality for a subset of specimens prior to sampling many (or all) specimens. When working with historical specimens, researchers may have little control over the original environmental conditions in which a specimen was collected, the storage conditions prior to museum preparation, or the conditions in which the specimen was preserved, all of which may influence heterogeneity in DNA degradation. The specimens we analyzed were collected in a cold-desert environment (i.e., the Great Basin Desert, which experiences hot and dry summers and cold winters), and these conditions may have minimized DNA degradation of the soft tissues relative to other studies using specimens from more humid regions.

A common application of genetic data from historical mammalian specimens is to evaluate population changes (e.g., genetic diversity, effective population size, gene flow) by comparing data from historical and contemporary samples (Wandeler et al. 2007; Holbrook et al. 2012; Jordan et al. 2012; Polanc et al. 2012; Lonsinger et al. 2018a). Thus, the use of historical sample sources producing genetic data of low quality and reliability could artificially decrease historical genetic diversity and lead to erroneous conclusions about population change (Taberlet et al. 1999). Finally, directing sampling of historical specimens to the sample source most likely to produce reliable results reduces unnecessary destructive sampling and contributes to traditional goals of NHCs of long-term care and maintenance of specimens (Wisely et al. 2004), while facilitating genetic research and retrospective analyses essential for understanding and improving the conservation of contemporary populations (e.g., Miller and Waits 2003) and biodiversity (Shaffer et al. 1998).

ACKNOWLEDGMENTS

We thank the Natural History Museum of Utah for granting access to their collections, E. Rickart for his oversight and assistance during our time at the museum, N. Bosque-Perez for providing laboratory space, and E. Gese for thoughtful insights and discussions. We are indebted to the many biologists and researchers before us, who diligently collected and cataloged the specimens that fill natural history collections. Funding was provided by T&E Incorporated's Grants for Conservation Biology,

the U. S. Army Research Laboratory and the U. S. Army Research Office (Grant: RC-201205), and the U.S. Department of Defense Environmental Security Technology Certification (Grant: 12 EB-RC5-006) and Legacy Resource Management (Grant: W9132T-12-2-0050) programs. We thank two anonymous reviewers and J. Light for comments and suggestions that greatly improved this manuscript.

DATA ACCESSIBILITY

Data for models of PCR success, allelic dropout, and false alleles are available from Open PRAIRIE at https://openprairie.sdstate.edu/nrm_data_sets/1/.

SUPPLEMENTARY DATA

Supplementary data are available at *Journal of Mammalogy* online.

Supplementary Data SD1.—Description of kit fox (*Vulpes macrotis*) museum specimens sampled from the Natural History Museum of Utah (UMNH).

Supplementary Data SD2.—Analysis of PCR data for kit fox (*Vulpes macrotis*) museum specimens sampled from the Natural History Museum of Utah (UMNH).

LITERATURE CITED

- AUSTIN, J. J., AND J. MELVILLE. 2006. Incorporating historical museum specimens into molecular systematic and conservation genetics research. *Molecular Ecology Notes* 6:1089–1092.
- BERRY, O., S. D. SARRE, L. FARRINGTON, AND N. AITKEN. 2007. Faecal DNA detection of invasive species: the case of feral foxes in Tasmania. *Wildlife Research* 34:1–7.
- BOOM, R., C. J. SOL, M. M. SALIMANS, C. L. JANSEN, P. M. WERTHEIM-VAN DILLEN, AND J. VAN DER NOORDAA. 1990. Rapid and simple method for purification of nucleic acids. *Journal of Clinical Microbiology* 28:495–503.
- BROQUET, T., N. MÉNARD, AND E. PETIT. 2007. Noninvasive population genetics: a review of sample source, diet, fragment length and microsatellite motif effects on amplification success and genotyping error rates. *Conservation Genetics* 8:249–260.
- BROQUET, T., AND E. PETIT. 2004. Quantifying genotyping errors in noninvasive population genetics. *Molecular Ecology* 13:3601–3608.
- BUCHAN, J. C., E. A. ARCHIE, R. C. VAN HORN, C. J. MOSS, AND S. C. ALBERTS. 2005. Locus effects and sources of error in noninvasive genotyping. *Molecular Ecology Notes* 5:680–683.
- CASAS-MARCE, M., E. REVILLA, AND J. A. GODOY. 2010. Searching for DNA in museum specimens: a comparison of sources in a mammal species. *Molecular Ecology Resources* 10:502–507.
- COOPER, A., C. MOURER-CHAUVIRÉ, G. K. CHAMBERS, A. VON HAESLER, A. C. WILSON, AND S. PÄÄBO. 1992. Independent origins of New Zealand moas and kiwis. *Proceedings of the National Academy of Sciences* 89:8741–8744.
- CULLINGHAM, C. I., C. SMEETON, AND B. N. WHITE. 2006. Isolation and characterization of swift fox tetranucleotide microsatellite loci. *Molecular Ecology Notes* 7:160–162.
- CULVER, M., P. W. HEDRICK, K. MURPHY, S. O'BRIEN, AND M. G. HORNOCKER. 2008. Estimation of the bottleneck size in Florida panthers. *Animal Conservation* 11:104–110.
- DEMAY, S. M., P. A. BECKER, C. A. EIDSON, J. L. RACHLOW, T. R. JOHNSON, AND L. P. WAITS. 2013. Evaluating DNA degradation rates in faecal pellets of the endangered pygmy rabbit. *Molecular Ecology Resources* 13:654–662.
- FOX, J., AND G. MONETTE. 1992. Generalized collinearity diagnostics. *Journal of the American Statistical Association* 87:178–183.
- FRANCISCO, L., A. LANGSTON, C. MELLERSH, C. NEAL, AND E. OSTRANDER. 1996. A class of highly polymorphic tetranucleotide repeats for canine genetic mapping. *Mammalian Genome* 7:359–362.
- FREDHOLM, M., AND A. WINTERO. 1995. Variation of short tandem repeats within and between species belonging to the Canidae family. *Mammalian Genome* 6:11–18.
- GARDNER, J. L., T. AMANO, W. J. SUTHERLAND, L. JOSEPH, AND A. PETERS. 2014. Are natural history collections coming to an end as time-series? *Frontiers in Ecology and the Environment* 12:434–436.
- GOUDET, J. 1995. FSTAT: a computer program to calculate F-statistics. *Journal of Heredity* 86:485–486.
- GREENWOOD, A. D., C. CAPELLI, G. POSSNERT, AND S. PÄÄBO. 1999. Nuclear DNA sequences from late pleistocene megafauna. *Molecular Biology and Evolution* 16:1466–1473.
- HEDMARK, E., AND H. ELLEGREN. 2005. Microsatellite genotyping of DNA isolated from claws left on tanned carnivore hides. *International Journal of Legal Medicine* 119:370–373.
- HOLBROOK, J. D., R. W. DEYOUNG, M. E. TEWES, AND J. H. YOUNG. 2012. Demographic history of an elusive carnivore: using museums to inform management. *Evolutionary Applications* 5:619–628.
- HOLMES, M. W., ET AL. 2016. Natural history collections as windows on evolutionary processes. *Molecular Ecology* 25:864–881.
- HOLMES, N. G., H. F. DICKENS, H. L. PARKER, M. M. BINNS, C. S. MELLERSH, AND J. SAMPSON. 1995. Eighteen canine microsatellites. *Animal Genetics* 26:132–133.
- HÖSS, M., AND S. PÄÄBO. 1993. DNA extraction from pleistocene bones by a silica-based purification method. *Nucleic Acids Research* 21:3913–3914.
- JANSSON, E., J. HARMOINEN, M. RUOKONEN, AND J. ASPI. 2014. Living on the edge: reconstructing the genetic history of the finnish wolf population. *BMC Evolutionary Biology* 14:64.
- JORDAN, N. R., J. MESSENGER, P. TURNER, E. CROOSE, J. BIRKS, AND C. O. REILLY. 2012. Molecular comparison of historical and contemporary pine marten (*Martes martes*) populations in the British Isles: evidence of differing origins and fates, and implications for conservation management. *Conservation Genetics* 13:1195–1212.
- KITCHEN, A. M., E. M. GESE, L. P. WAITS, S. M. KARKI, AND E. R. SCHAUSTER. 2006. Multiple breeding strategies in the swift fox, *Vulpes velox*. *Animal Behaviour* 71:1029–1038.
- LISTER, A. M.; CLIMATE CHANGE RESEARCH GROUP. 2011. Natural history collections as sources of long-term datasets. *Trends in Ecology & Evolution* 26:153–154.
- LONSINGER, R. C., J. R. ADAMS, AND L. P. WAITS. 2018a. Evaluating effective population size and genetic diversity of a declining kit fox population using contemporary and historical specimens. *Ecology and Evolution* 8:12011–12021.
- LONSINGER, R. C., E. M. GESE, S. J. DEMPSEY, B. M. KLUEVER, T. R. JOHNSON, AND L. P. WAITS. 2015. Balancing sample accumulation and DNA degradation rates to optimize noninvasive genetic sampling of sympatric carnivores. *Molecular Ecology Resources* 15:831–842.

- LONSSINGER, R. C., P. M. LUKACS, E. M. GESE, R. N. KNIGHT, AND L. P. WAITS. 2018b. Estimating densities for sympatric kit foxes (*Vulpes macrotis*) and coyotes (*Canis latrans*) using noninvasive genetic sampling. *Canadian Journal of Zoology* 96:1080–1089.
- LONSSINGER, R. C., AND L. P. WAITS. 2015. ConGenR: rapid determination of consensus genotypes and estimates of genotyping errors from replicated genetic samples. *Conservation Genetics Resources* 7:841–843.
- MCDONOUGH, M. M., L. D. PARKER, N. R. MCINERNEY, M. G. CAMPANA, AND J. E. MALDONADO. 2018. Performance of commonly requested destructive museum samples for mammalian genomic studies. *Journal of Mammalogy* 99:789–802.
- MCLEAN, B. S., ET AL. 2016. Natural history collections-based research: progress, promise, and best practices. *Journal of Mammalogy* 97:287–297.
- MILLER, C. R., P. JOYCE, AND L. P. WAITS. 2002. Assessing allelic dropout and genotype reliability using maximum likelihood. *Genetics* 160:357–366.
- MILLER, C. R., AND L. P. WAITS. 2003. The history of effective population size and genetic diversity in the Yellowstone grizzly (*Ursus arctos*): implications for conservation. *Proceedings of the National Academy of Sciences* 100:4334–4339.
- NIELSEN, E. E., M. M. HANSEN, AND V. LOESCHCKE. 1999. Analysis of DNA from old scale samples: technical aspects, applications and perspectives for conservation. *Hereditas* 130:265–276.
- NYSTRÖM, V., A. ANGERBJO, AND L. DALEN. 2006. Genetic consequences of a demographic bottleneck in the Scandinavian arctic fox. *Oikos* 114:84–94.
- OSTRANDER, E. A., F. A. MAPA, M. YEE, AND J. RINE. 1995. One hundred and one new simple sequence repeat-based markers for the canine genome. *Mammalian Genome* 6:192–195.
- OSTRANDER, E. A., G. F. SPRAGUE, JR, AND J. RINE. 1993. Identification and characterization of dinucleotide repeat (CA)_n markers for genetic mapping in dogs. *Genomics* 16:207–213.
- PÄÄBO, S., ET AL. 2004. Genetic analyses from ancient DNA. *Annual Review of Genetics* 38:645–679.
- PAETKAU, D. 2003. An empirical exploration of data quality in DNA-based population inventories. *Molecular Ecology* 12:1375–1387.
- PEAKALL, R., AND P. E. SMOUSE. 2006. genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6:288–295.
- POLANC, P., M. SINDIČIĆ, M. JELENCIĆ, T. GOMERČIĆ, I. KOS, AND D. HUBER. 2012. Genotyping success of historical Eurasian lynx (*Lynx lynx* L.) Samples. *Molecular Ecology Resources* 12:293–298.
- POMPANON, F., A. BONIN, E. BELLEMAIN, AND P. TABERLET. 2005. Genotyping errors: causes, consequences and solutions. *Nature Reviews. Genetics* 6:847–859.
- R CORE TEAM. 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- SCHWARTZ, M. K., ET AL. 2007. Inferring geographic isolation of wolverines in California using historical DNA. *Journal of Wildlife Management* 71:2170–2179.
- SHAFFER, H. B., R. N. FISHER, AND C. DAVIDSON. 1998. The role of natural history collections in documenting species declines. *Trends in Ecology & Evolution* 13:27–30.
- STENGLEIN, J. L., L. P. WAITS, D. E. AUSBAND, P. ZAGER, AND C. M. MACK. 2010. Efficient, noninvasive genetic sampling for monitoring reintroduced wolves. *Journal of Wildlife Management* 74:1050–1058.
- TABERLET, P., ET AL. 1996. Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Research* 24:3189–3194.
- TABERLET, P., L. P. WAITS, AND G. LUIKART. 1999. Noninvasive genetic sampling: look before you leap. *Trends in Ecology & Evolution* 14:323–327.
- WAITS, L. P., AND D. PAETKAU. 2005. Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection. *Journal of Wildlife Management* 69:1419–1433.
- WANDELER, P., P. E. HOECK, AND L. F. KELLER. 2007. Back to the future: museum specimens in population genetics. *Trends in Ecology & Evolution* 22:634–642.
- WANDELER, P., S. SMITH, P. A. MORIN, R. A. PETTIFOR, AND S. M. FUNK. 2003. Patterns of nuclear DNA degeneration over time—a case study in historic teeth samples. *Molecular Ecology* 12:1087–1093.
- WISELY, S. M., S. W. BUSKIRK, M. A. FLEMING, D. B. McDONALD, AND E. A. OSTRANDER. 2002. Genetic diversity and fitness in black-footed ferrets before and during a bottleneck. *The Journal of Heredity* 93:231–237.
- WISELY, S. M., J. E. MALDONADO, AND R. C. FLEISCHER. 2004. A technique for sampling ancient DNA that minimizes damage to museum specimens. *Conservation Genetics* 5:105–107.
- WOODRUFF, S. P., T. R. JOHNSON, AND L. P. WAITS. 2015. Evaluating the interaction of faecal pellet deposition rates and DNA degradation rates to optimize sampling design for DNA-based mark-recapture analysis of sonoran pronghorn. *Molecular Ecology Resources* 15:843–854.
- ZHAN, X., X. ZHENG, M. W. BRUFORD, F. WEI, AND Y. TAO. 2010. A new method for quantifying genotyping errors for noninvasive genetic studies. *Conservation Genetics* 11:1567–1571.

Submitted 13 December 2018. Accepted 19 June 2019.

Associate Editor was Jessica Light.