Assessing the levels of microsatellite allelic dropout in museum specimens using high-throughput sequencing and genotyping by synthesis.

Stella Yuan[1], Eric Malekos[1], and Melissa Hawkins[2]

[1]Humboldt State University
[2]Smithsonian Institution

May 27, 2020

Abstract

The use of museum specimens held in natural history repositories for population and conservation genetic research is increasing in tandem with the use of next generation sequencing technologies. Short Tandem Repeats (STRs), or microsatellite loci, are commonly used genetic markers in population and conservation genetic studies. However, they traditionally suffered from a host of issues: fragment size homology, high costs, and low throughput as a result of capillary electrophoresis genotyping and difficulty in reproducibility across laboratories. Next generation sequencing technologies can address these problems, but the incorporation of DNA derived from museum specimens suffers from significant fragmentation and contamination with exogenous DNA. Combatting these issues requires extra measures of stringency in the lab and during data analysis, yet there have not been any studies evaluating microsatellite allelic dropout from museum specimen extracted DNA. In this study, we explore a high throughput sequencing method to evaluate the amount of variation found within museum specimen DNA extracts for previously characterized microsatellites across PCR replicates. We found it useful to classify samples based on quality after replicated PCRs, which determined the rate by which genotypes were accurately recovered. We also found that longer microsatellites performed worse in all museum specimens, so when designing a study invoking museum specimens, short markers (under 250 bp) should be preferentially selected. Allelic dropout rates across loci were dependent on sample quality. The high quality museum specimens performed as well, and recovered nearly as high quality metrics as our tissue sample. Mitochondrial DNA sequences were not predictive of nuclear DNA presence, as all samples recovered cytochrome b fragments yet many lacked microsatellite genotypes, particularly in samples deemed low quality. Based on our results, we have provided a set of best practices for screening, quality assurance, and incorporation of reliable genotypes from museum specimens.

Introduction:

Natural history repositories represent invaluable collections of specimens for scientific use across diverse fields (Blagoderov, Kitching, Livermore, Simonsen, & Smith, 2012; Lane, 1996; Lister & Group, 2011; S. L. Williams, 1999). Many of these specimens represent populations of plants and animals that no longer exist due to land use change and human alterations of landscapes over the past century (Smith et al., 2013). Additionally, museum specimens often represent the few or only representatives of endangered or rare species, and provide important vouchers for comparison with modern samples, as well as genetic resources for species which may be logistically difficult or impossible to sample in wild habitats (W. Miller et al., 2009; White, Mitchell, & Austin, 2018). As such, usage of museum specimens for modern research incorporating DNA analysis is increasing. In turn, destructive sampling requests are increasing, many of which propose molecular sequencing from specimens as a justification for consumption of source material.

The degraded DNA associated with museum specimens is known to require extra measures of stringency in order to combat issues with exogenous DNA sequences (Paabo et al., 2004; Rizzi, Lari, Gigli, De Bellis, &

Caramelli, 2012), and the use of PCR based methods have identified issues with nuclear copies of mitochondrial DNA that confound degraded or ancient DNA mitochondrial sequence results (den Tex, Maldonado, Thorington, & Leonard, 2010). The extracted DNA in each sample is often contaminated by exogenous sources (humans, bacteria, pests) and the endogenous DNA can be highly fragmented (Campana et al., 2012; Hawkins, Hofman, et al., 2016; McDonough, Parker, Rotzel McInerney, Campana, & Maldonado, 2018). Studies which reliably sequence DNA from museum specimens undergo stringent protocols to combat the low quantity and highly fragmented nature of museum specimen extracts. As such, these studies must process the specimens with additional precautions in order to prevent cross contamination of samples, and should be processed in appropriate lab spaces dependent on the material. Downstream from wet lab procedures additional bioinformatic steps should be taken to ensure that the resulting genetic sequence data represents the target taxa. Truly ancient samples (derived from archaeological samples, permafrost specimens, coprolites, sediments, mummies, and others) have been shown to offer patterns of degradation associated with misincorporation of various nucleotides – namely cytosine to uracil deamination – from which characteristic patterns can be tested for to provide authenticity to the recovered sequences (Hofreiter, Serre, Poinar, Kuch, & Paabo, 2001; Jónsson, Ginolhac, Schubert, Johnson, & Orlando, 2013). Patterns of degradation are only starting to be understood, and vary depending on the type of samples being processed (Shapiro, 2012; Weiß et al., 2016), with museum specimens lacking the characteristic cytosine to uracil deamination (McDonough et al., 2018).

A study of mitochondrial genome enrichment from museum specimens (Hawkins, Hofman, et al., 2016) found that sample type was more predictive of amplification success rather than age. Another study also concluded that success rates as well as endogenous DNA content varied widely depending on the type of consumed sample (McDonough et al., 2018), and Campana et al., (2012) found that recovery of longer mitochondrial (D-loop) PCR products did not correlate with the success of nuclear DNA amplification. When granted destructive sampling permissions, institutions often set individual policies on what types of samples are provided to approved research projects. As such, the most desired sample types may not be approved for consumption in DNA extraction.

Short Tandem Repeats (STRs), also commonly referred to as microsatellite loci, have been useful markers for numerous applications, such as forensics, cancer diagnosis, and widely implemented in the fields of conservation genetics to evaluate genetic diversity and population structure in organisms ranging from bacteria, to plants and animals (e.g. Bilska & Szczecińska, 2016; Thatte, Joshi, Vaidyanathan, Landguth, & Ramakrishnan, 2018). Historically, microsatellites were isolated from a specific species of interest for use on population level analyses, a process which took time and funding to develop prior to any analysis on the taxa of interest (Fisher, Gardner, & Richardson, 1996; Glenn & Schable, 2005; Lian, Wadud, Geng, Shimatani, & Hogetsu, 2006). Cross species amplification has been shown to work in some taxa, but comparisons across different species must be done cautiously due to issues with homoplasy and ascertainment bias (Bailey et al., 2015; Crawford et al., 1998; Estoup, Jarne, & Cornuet, 2002; Grimaldi & Crouau-Roy, 1997; Li & Kimmel, 2013).

Next generation sequencing technology has allowed for a much more rapid identification of microsatellite loci in non-model organisms (Duan, Li, Sun, Wang, & Zhu, 2014; Griffiths et al., 2016; M. P. Miller, Knaus, Mullins, & Haig, 2013; Silva, Martins, Gouvea, Pessoa-Filho, & Ferreira, 2013) by allowing tandem repeat regions to be identified at a genomic scale, and allowing the simultaneous sequencing of thousands of putative microsatellite loci as compared to traditional cloning based methods (Glenn & Schable, 2005). In addition to the cost reduction of microsatellite isolation, some of the issues known to occur when genotyping microsatellites via capillary electrophoresis (CE hereafter) can be alleviated using next generation sequencing technologies (Vartia et al., 2016). For example, fragment size analysis via CE has been known to provide (albeit sometimes predictably) shifted sizes when samples are run on different machines (Morin, Manaster, Mesnick, & Holland, 2009). Access to the raw sequences from next generation sequencing would allow precise sizing of alleles (Darby, Erickson, Hervey, & Ellis-Felege, 2016).

A number of studies have evaluated how to transform these sequence based microsatellite reads into genotypes recovered from capillary sequencing (Barbian et al., 2018; Darby et al., 2016; De Barba et al., 2017; Jónsson

et al., 2013; Pimentel et al., 2018; Šarhanová, Pfanzelt, Brandt, Himmelbach, & Blattner, 2018; Vartia et al., 2016; Zhan et al., 2017). Each of these genotyping by synthesis (GBS hereafter) studies has evaluated some aspects of the biases induced when comparing sequences from high-throughput sequencing platforms as opposed to fragment size analysis genotyping. For instance, GBS studies have resulted in recovery of additional alleles due to the reconstruction from DNA sequences as opposed to fragment size analysis from CE. Some of the other most commonly addressed issues included evaluation of stutter, PCR artifacts and size homoplasy (Barbian et al., 2018; De Barba et al., 2017). Although challenges exist for direct comparison of high-throughput sequencing based microsatellite genotypes with those from capillary sequencers via fragment size analysis, the ability to generate comparable datasets is paramount in order to build off previous research, and inform larger, potentially landscape based conservation plans.

Despite the wide range of studies already published on genotyping using high throughput sequencing, there are no studies which have specifically evaluated the degree of variation which occurs from museum specimen sourced DNA. GBS studies have estimated the amount of allelic dropout from chimpanzees (Barbian et al., 2018) and bears (De Barba et al., 2017) from fecal samples, as well as tissue (Vartia et al., 2016). Here we explore a high throughput sequencing method to evaluate the amount of variation found within DNA extracts from museum specimens for previously characterized microsatellites across various PCR replicates. We analyzed three types of datasets: a dataset containing individual PCR replicates, a pooled dataset where the individual replicates were mixed together prior to library preparation, and a bioinformatically pooled dataset where the replicates were combined via bash scripting. The rates of allelic dropout generated here will serve as the first for high throughput sequencing of museum specimens and provide best practices for subsequent studies on museum derived specimens.

Methods:

Samples

In order to establish baseline data for quantifying allelic dropout and sequencing errors/false genotypes in degraded source material, total genomic DNA from 147 samples were extracted for a population genetic study on the Humboldt's flying squirrel (*Glaucomys oregonensis,*Yuan, 2020). From these samples, a subset was evaluated to test rates of dropout across replicated microsatellite PCRs. From the screened samples, three museum specimens which reliably amplified in PCR were deemed 'high quality' museum specimens, HQMS hereafter (as assessed from agarose gel visualization of PCR products). Additionally, three samples which repeatedly failed in PCR were deemed 'low quality' samples (LQMS hereafter). This differential selection was intentional to determine if allelic dropout rates varied based on reliability of PCR results. Additionally, one frozen tissue sourced specimen was included to observe if similar rates of allelic dropout would be detected from that sample. All samples' details are provided in Table 1.

Mitochondrial DNA Sequencing

A short fragment of mitochondrial cytochrome *b* was amplified for all individuals in order to provide a maternal signature of inheritance, as is common practice in population genetic studies. Cytochrome *b*was selected as it is informative and widely used in mammalian population and phylogenetics. Due to the degraded nature of museum specimens only the first ~300 bp of the cytochrome*b* gene were selected for amplification in this study. Mitochondrial DNA is present in hundreds to thousands of copies per cell, and historically the recovery of mtDNA has been routinely published in ancient DNA, with nuclear DNA (only one copy per cell) being more difficult to recover. As such, mtDNA sequences serve as a maternal signal as well as a control to ensure endogenous DNA was present in the tested samples.

Microsatellite Selection

We tested three sample types: tissue, 'high', and 'low' quality museum specimens to evaluate differential amplification across microsatellites of varying lengths. Degradation of samples is thought to occur quickly, but age alone is a poor predictor of the overall quality (Campana et al., 2012; Hawkins, Hofman, et al., 2016; McDonough et al., 2018). Previously published microsatellites from the northern flying squirrel (*G.*

sabrinus) - the sister taxa of the Humboldt's flying squirrel - were used in this study (Loci names: GS-2, GS-4; Zittlau, Davis, & Strobeck, 2000, and GLSA-12, GLSA-22, and GLSA-52; Kiesow, Wallace, & Britten, 2011, Table 2).

Microsatellites were selected by length and type of repeat motif. Two microsatellites were selected in our 'short' size range (GS-2 and GS-4), which was considered any marker under 150 bp in length according to published allele sizes, two 'medium' (GLSA-12 and GLSA-22, 150-200 bp in length) and one 'long' microsatellite (GLSA-52, >200 bp in length). The repeat motif was also considered, and of these microsatellites, one was considered a 'complex' motif (GLSA-12) while all others were considered 'simple'. The specific motif composition can be found in Table 2. The rationale for testing microsatellites with both types of motifs was to examine homoplasy. Direct sequencing can reveal if microsatellites with complex motifs accrew variants in different regions of the repeat complex that result in the same genotype in a different individual with polymorphisms in another region.

DNA extraction

Total genomic DNA was isolated using Qiagen QIAamp DNA Mini Kits (Qiagen, Valencia, CA) in a lab designated for degraded and ancient DNA, while DNA from the single tissue sample was processed in a designated, high quality DNA facility using Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA), following the standard protocol for animal tissue samples. DNA was eluted in buffer AE in a volume of 100 ul for all museum specimens and 200 ul for the tissue sample. The degraded DNA laboratory included additional steps of placing a blank extraction control for minimally every 12 samples, bleaching, exposing both plastic consumables and pipettes to ultraviolet radiation, routine changing of gloves, and processing of samples under a hood in order to prevent contamination during DNA extraction.

Polymerase Chain Reaction

Mitochondrial DNA- Cytochrome b

The entire (˜1140 bp) cytochrome *b* gene was amplified in PCR using universal primers (L14724, Irwin & Kocher, 1991; H15910, Oshida, Lin, Masuda, & Yoshida, 2000) for the tissue sample (HSU 8180). In the museum specimens, an approximately 300 bp region of the mitochondrial cytochrome *b* gene was amplified with the primer L14724 (Irwin & Kocher, 1991) and a newly designed reverse primer, GOR_R1, from a *G. o. californicus* GenBank sequence (Accession #: AF063060) using Primer3 (Untergasser et al., 2012). Fragments were mapped to published sequences and visualized in Geneious Prime v 2020.0.4 (Kearse et al., 2012). Primers were ordered from IDT at 35.4 nmol concentration, solubilized, and diluted to 100 µM prior to PCR. Singleplexed PCR was performed in 16 µl reactions containing 2.0 µl of DNA template, 4.5 µl of ddH_2O, 0.5 µl of each primer, and 8.5 µl of DreamTaq Green PCR Master Mix (Thermo Fisher Scientific Inc., Waltham, MA) with the following thermocycler profile: 95°C for 1 min; 30 cycles (tissues) or 35 cycles (museum samples) of 95°C for 30 sec, 45°C for 30 sec, 72°C for 30 sec; 72°C for 5 mins. All amplifications were performed on either an Applied Biosystems MiniAmp or BioRad T-100 cycler. Successful PCR amplifications were replicated twice in each sample to ensure validity, and a 1-1.5% agarose gel was run to visualize PCR products.

Nuclear DNA- Microsatellites

Singleplexed PCR was performed in 16 µl reactions containing 2.0 µl of DNA template, 4.5 µl of ddH_2O, 0.5 µl of each primer, and 8.5 µl of DreamTaq Green PCR Master Mix. When amplifying microsatellites with the GS-2 and GS-4 primers, 0.3 µl of ddH_2O was replaced with bovine serum albumin (BSA, New England Biolabs, Inc, 12 mg) to promote reaction specificity. For all samples a touchdown PCR profile was used with the following conditions: 95°C for 1 min; 2 cycles of 95°C for 15 sec, 60°C for 30 sec, 72°C for 45 sec; 2 cycles changing 60°C to 58°C; 2 cycles changing 58°C to 54°C; 2 cycles changing 54°C to 52°C; 35 cycles of 95°C for 15 sec, 50°C for 30 sec, 72°C for 45 sec; 72°C for 5 mins. Successful PCR amplifications were replicated twice in tissue samples and three times in museum samples. Following amplification, 1.5% agarose gels were run with a 100 bp size standard (Invitrogen) and stained with GelRed (Biotium) or SYBR Green (Invitrogen).

To remove residual primers, dNTPs and nontarget molecules, solid phase reversible immobilization (SPRI hereafter) cleaning via magnetic beads was performed (following Rohland & Reich, 2012) in a ratio of 1 part PCR product to 1.5X magnetic beads (KAPA beads, Roche). Cleaned PCR products were eluted in 20 ul ddH_2O and quantified on a NanoDrop Lite Spectrophotometer (ThermoScientific).

Library Preparation

In order to evaluate the variation of resulting genotypes, individual PCR replicates required library preparation for Illumina sequencing. Individual dual iTru indices (Glenn et al., 2019) were ligated to each PCR replicate using KAPA Illumina Library Preparation Kits (Roche, # KK8232) following the reduced reactions previously published (Hawkins, Leonard, et al., 2016). Additionally, all PCR replicates across all microsatellites were pooled together for another separate library preparation in order to determine if pooling across replicates influenced resulting genotypes (hereafter referred to as the pooled dataset). Two µl of each PCR replicate (mitochondrial DNA and microsatellites) was pooled from museum specimens prior to library preparation, and 2 or 3 µl from tissue replicates of mitochondrial DNA and microsatellites respectively. A library preparation control (consisting of ddH_20) was included in all steps. Libraries, as well as a negative control, were amplified in 25 µl reactions consisting of 1.25 µl of each iTru adapter, 2.5 µl ddH_2O, 7.5 µl of adapter-ligated DNA, and 12.5 µl of KAPA HiFi HotStart ReadyMix. The thermocycler conditions for library amplification were: 98°C for 45 sec; 10 cycles (tissue sample) or 14 cycles (museum samples) of 98°C for 15 sec, 60°C for 30 sec, 72°C for 1 min; 72°C for 5 mins. After completing library prep, an agarose gel was run on all replicates to ensure successful ligation of Illumina and iTru adapters. Products were SPRI cleaned as detailed above. Each individually prepared replicate was then pooled together by placing 2 ul of cleaned product into a 0.5 mL tube, which was quantified for Illumina sequencing on an ABI QuantStudio 3 using the KAPA Biosystems Library Quantification Kit (Roche, # KK4824), standard Illumina Primers and following the protocol therein.

Sequencing

The quantified pool was sequenced on an Illumina MiSeq using a 2 x 300 PE version 3 kit at the Center for Conservation Genomics, Smithsonian Conservation Biology Institute, Washington DC or using a 2 x 250 PE version 2 Nano kit at the Laboratory of Analytical Biology, National Museum of Natural History, Smithsonian Institution, Washington DC. Reads were demultiplexed and downloaded from the BaseSpace Server.

Quality Filtering

Samples were run through FastQC v 0.11.9 (Andrews 2010) and CutAdapt v 1.18 (Martin, 2011) for quality filtering and additional checks for adapter removal. Phred scores were required to be [?] 20 averaged across each read, and the command for cutadapt was: cutadapt –report=minimal -q 20 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT -o {R1_output.fastq} -p {R2_output.fastq} {R1_input.fastq.gz} {R2_input.fastq.gz}.

In order to assess the overall read quality, prinseq v 0.20.4 (Schmieder & Edwards, 2011) was run on each individual library preparation to determine the proportion of low quality reads. 'perl prinseq-lite.pl -fastq {R1.fastq} -out_format 3 -derep_min 4 -log {NAME}.fastq.log -min_qual_mean 20 -verbose'. This data was run for evaluative purposes and not passed through the CHIIMP genotyping pipeline (Barbian et al., 2018) as the only requirements for that pipeline is for the adapters to be removed via CutAdapt. All quality data is summarized in Table 3.

Mitochondrial DNA reconstruction

Data from only the pooled run contained the two cytochrome *b* PCR replicates. Raw data was uploaded to Geneious Prime (Biomatters Inc.) where a published reference (GenBank Accession # AF030390, Demboski, Jacobsen, & Cook, 1998) was used to map reads. Mapping was done in Geneious via the bowtie2 plugin v 2.3.0 (Langmead & Salzberg, 2012), and consensus sequences were extracted. All sequences were aligned with MAFFT v1.4.0 (Katoh & Standley, 2013) via the Geneious plugin, and primers were trimmed off. Data

were combined from 147 total individuals (Yuan, 2020, data not shown) to reconstruct haplotypes using PopArt (*http://popart.otago.ac.nz*) and a median spanning network (Bandelt, Forster, & Rohl, 1999).

Genotyping

The CHIIMP v 0.3.1 (Barbian et al., 2018) pipeline was used on reads filtered with CutAdapt to generate genotypes across all replicates using identical parameters. Each genotype was called if there were minimally 5 reads (counts.min =5), alleles were identified only if the number of recovered reads constituted minimally 5% of the filtered reads (fraction.min = 0.05), and all loci were given a length buffer of 20 bp. The output from CHIIMP not only identified the most likely allele calls, but also provided data on whether or not the sample had PCR associated stutter, PCR artifacts removed and more than two prominent sequences. Data for the individually library prepared samples, the pooled data and a bioinformatically pooled dataset, where the 2-3 replicates of each individually prepared reads were concatenated together, were put through quality filtering in order to evaluate the possibility of low coverage of the library preparations leading to erroneous genotypes. All of the recovered genotypes were manually compared in order to evaluate any patterns across the tested samples.

Statistical Analyses

Descriptive statistics, a single factor analysis of variance (ANOVA), and a linear regression were calculated in Microsoft Excel v16.16.20 using the data analysis add in. The ANOVA was performed on the quality data (represented by the percentage of reads passing quality from prinseq) compared against the sample types.

The program MicroDrop v1.01 (Wang & Rosenberg, 2012) was run to evaluate the rates of allelic dropout within samples and across loci. The program was run twice, once on the genotypes called directly by CHIIMP based on the pooled data, and once on the genotypes determined by our best practices, shown in the 'Manually Processed Genotypes' column of Table 4. The program was run using the default parameters, and we did not enforce Hardy-Weinberg Equilibrium on our data due to the low number of alleles and samples. Individual replicates were not run on MicroDrop, as the program is designed to work on non-replicated datasets, although the pooled data had multiple replicates of each individual pooled prior to library preparation.

Results:

Mitochondrial Cytochrome b

Following amplification of either the entire cytochrome *b* gene (tissue) or the first 300 base pairs (museum specimens), all tested samples recovered sequences which mapped to the *G. sabrinus*reference and ranged from 0-100% amplification success as assessed from gel electrophoresis (details in Appendix 1). Coverage of the museum specimens ranged from an average of 22.7- 1933x. Interestingly, the highest coverage was from one of the poor quality museum specimens (MVZ 5211), as was the lowest (MVZ 2088). Regardless, all samples recovered reliable cytochrome *b* sequences, and only the lowest coverage (MVZ 2088) had an error rate over 0.0001% (0.064%, Appendix 2). Despite poor gel visualization confirmation, all samples recovered mitochondrial sequences, even when many PCRs had been deemed as failed (see all poor quality sample results in Appendix 1). Average coverage across all sample types was 624x, 612x across the high quality museum specimens, and 760x for the poor quality museum specimens (which was largely biased by the very high coverage recovered in MVZ 5211, without this sample the average was 174x). Additional quality metrics for each sample are detailed in Appendix 2. The quality scores across the samples did not vary substantially. The sample recovering the lowest Q20 was HSU 1836 with 95.8%, and the highest was MVZ 2088 with 98.6%. At a quality of Q30 the lowest was again HSU 1836 and the highest was UMMZ 79755 with 98%. The tissue sample had 98.2% score at Q20, and 94% at Q30, which is likely artificially low since this sample had the entire cytochrome*b* amplified, then fragmented prior to library preparation (Yuan, 2020). Regardless of sample type, the reads appear to be high quality as determined by the quality metrics, expected rates of errors and Q scores. It is also noteworthy that the cytochrome *b* data was extracted solely from the pooled data where all microsatellites and cytochrome *b* fragments were pooled prior to library prep.

The haplotypes recovered from the different samples included here consisted of three closely related haplotypes, separated by three to six substitutions. All of the *G. o. californicus* recovered the same haplotype, the single *G. o. lascivus* (HSU 8180) recovered a second, the *G. o. stephensi* (HSU 1836) sample recovered the third haplotype (Appendix 3). The haplotypes were common across a wider study of *G. oregonensis* (Yuan, 2020), and the fact that none of the LQMS recovered a unique haplotype provides support for the authenticity of the data. Importantly, samples included in this study represent three subspecies from four geographic locations, so numerous haplotypes were expected.

Microsatellites- effects of sample quality

When prinseq was run on all reads from individually library prepared replicates a general pattern emerged where PCR success was predictive of quality. The mean quality across all sample types was 85.6%, with a range from 57.43% - 99.48% (MVZ 5211 GS-2 replicate two and HSU 1836 GLSA-52 replicate one represent the lowest and highest). The median was 92.6% and mode was 96.82%. If sample types were separated the mean quality was as follows: 95.99% (SE +-0.99%), 95.06% (SE +-0.68%), and 73.84% (SE +-2.10%) for tissue, HQMS and LQMS respectively (Table 5). All aforementioned and additional descriptive metrics are shown in Appendix 2. The ANOVA was significant (P= <0.001) and the regression resulted in an R^2 of 0.44 which was also highly significant (P=<0.001), indicating our assessed quality from gel electrophoresis was predictive of genotyping success.

The CHIIMP genotypes were accurate for the single tissue sample, especially if the PCR replicates were pooled and genotyped together. The reads per replicate and percentage of good reads (as determined by standard prinseq quality filtration) are reported in Table 3. All recovered genotypes are summarized in Table 4. In the GLSA-52 locus the pooled run recovered a second allele 251 from HSU 8180 despite none of the other genotypes recovering that allele. In the pooled run HSU 8180 recovered a total of 79,417 reads, across all microsatellite loci and complete cytochrome *b* . The 251 allele was recovered in the pooled dataset with a frequency of 5.3% whereas the 257 allele was recovered at 17.6%, more than three times the frequency of the 251 allele. When all other replicates of this sample were evaluated only the 257 allele was recovered, and with rates ranging from 12.9-17.6% (see details in Appendix 4).

Mismatched alleles were recovered most frequently in low quality samples which routinely appeared to fail PCR across numerous replicates. Mismatches were often associated with one or more of the following: PCR stutter sequences, PCR artifacts and more than two prominent sequences as identified by the CHIIMP pipeline (Table 4). Individual samples did not appear to recover specific CHIIMP flags across all replicates, neither did specific microsatellites, however the locus GS-2 recovered frequent flags for all three metrics (Table 4).

The HQMS samples had routinely high quality sequences as determined by prinseq metrics. Only a single PCR replica from UMMZ 79755 had less than 85% of sequences pass quality metrics. All other replicates were over 85%, and most recovered over 95% of sequences passing quality filtration. Interestingly, the LQMS samples had high variation between PCR replicates performed here (Figure 1). The average for each of the LQMS samples were 81.3, 69.6, and 70.6% with a high degree of variation between replicates. For example, LACM 95619 ranged from 67.7-92.05% passing quality filters for GS-4. In this instance, across the three replicates three completely different sets of alleles were recovered providing no confidence in those genotypes despite one replicate recovering 92.05% high quality sequences.

Microsatellites- effects of microsatellite length

As expected the tissue sample had fairly consistent genotypes across all amplifications, except in one instance where a 2 bp difference was detected between the two PCR replicates. Interestingly, in this case (HSU 8180 marker GS-4) both the bioinformatically combined and pooled runs recovered a homozygous genotype. Based on read depth it is possible the minor alleles (92 and 94 in replicates 1 and 2 respectively) were sequencing errors or PCR stutter related to the high depth of coverage.

We recovered more frequent, and arguably more reliable genotypes across all samples for shorter microsatel-

lites. In one sample, LACM 95619 at the GS-2 locus, it appeared that both the bioinformatically combined and pooled run genotypes consisted of alleles from replicate 1 (82/82) and replicate 3 (84/84). Unfortunately the second replicate failed to produce an allele, so it is unknown from this study if both recovered alleles across the two PCRs are accurate, or if a dropout event is depicted in this instance. These genotypes were called from 4,235 raw reads for GS-2 replicate 1 and 4,449 reads from replicate 3. The second replicate recovered only 1,034 reads, and was very poor quality (only 659 passed standard prinseq quality filters, a mere 63.73% of the reads), which explains the lack of resulting genotypes from that replicate (Table 3).

The high quality museum specimens performed as well, and occasionally better, than the tissue sample. Despite reliable performance and genotyping success for the HQMS, there were still a handful of missing allele calls ranging from GS-4 in sample UMMZ 79755, to the longest marker GLSA-52, where both UMMZ 79755 and UMMZ 79760 lacked calls in one replicate each. Overall, the high quality specimens worked remarkably well across all loci, but often had more than two prominent sequences as flagged by CHIIMP (see Table 4 for details). The resulting genotypes were highly reliable, and only appeared to lack confirmation in GLSA-52, the longest microsatellite evaluated here. Two of the three HQMS had different calls between the individual replicates and the bioinformatically combined and pooled runs. Interestingly, UMMZ 79760 recovered 1 bp separated genotypes, 250/251, which does not make evolutionary sense for a dinucleotide microsatellite. Upon further investigation, the two alleles in the pooled run were recovered due to the following: allele 1 (251 bp) had an additional 'CA' repeat, but only 16 bp of the reverse primer, and allele 2 (250 bp) had one fewer 'CA' repeat and 17 bp of the reverse primer, resulting in a difference of 1 bp. The 255 bp alleles called had the same number of repeats as allele 1 but included the entire reverse primer sequence (20 bp), and the 253 bp allele called had the same number of repeats as allele 2 but included the entire reverse primer sequence.

The low quality samples could recover accurate genotypes, however much more variation occurred in the quality of the data (see Figure 1) and as such the reliability of the resulting genotypes requires stringent evaluation. The length appeared to make a difference, even on the shortest marker GS-2 four out of nine replicates did not recover a genotype, and two of the remaining five did not match between replicates. As the length of the microsatellites increased, the generation and reliability of the microsatellite decreased. For the low quality museum specimens, GS-2 had four missing genotypes, GS-4 had one missing and five mismatches, GLSA-12 had eight missing genotypes and one mismatch, GLSA-22 had seven missing and GLSA-52 was missing all nine genotypes.

Microsatellites- effects of repeat motif

In this study we evaluated four 'simple' repeat motif markers and a single 'complex' repeat motif. The GLSA-12 marker was the only 'complex' motif, characterized by a repeat region that was interrupted multiple times, and contained three regions of dinucleotide elements, see Table 2. All other microsatellites included dinucleotide repeats which were not interrupted. Based on the PCR replicate results (summarized in Appendix 1) the GLSA-12 locus performed the same as the other medium length microsatellite locus for all museum specimens, and failed a single PCR replicate of the HSU 8180 tissue sample (when repeated a third time this sample produced a matching genotype). Across all samples GLSA-12 amplified successfully 36.3% of the time as compared with 34% for the other medium length marker (GLSA-22). When evaluating the resulting genotypes, GLSA-12 recovered four out of seven genotypes versus five out of seven genotypes for GLSA-22 (see Table 4). The difference between GLSA-12 and GLSA-22 was one LQMS (MVZ 2088) which was amplified in GLSA-22. The complex motif, however, was not recovered as previously published in *G. sabrinus* (Kiesow et al., 2011).The published motif, (GT)5A(TG)3TTT(GT)5 varied from what we recovered here: (GT)13TT(GT)5 (HSU 1836) and (GT)11TT(GT)5 (HSU 8180). Overall, it does not appear that the complexity of the repeat motif influenced the rate of genotyping success in these samples, however we have limited data to conclude if the complexity affected genotyping rates.

Rates of Allelic Dropout

MicroDrop was run twice, first on the pooled run output from CHIIMP, and a second time on the manually

quality filtered dataset. Each MicroDrop analysis outputs individual and locus rates of allelic dropout. In the first run of the pooled genotypes, the locus specific dropout rates ranged from 0 (GLSA-22) to 48% (GLSA-52). Following manual processing of genotypes the rates of dropout ranged from 0 (GLSA-22 and GS-4) to 36.3% (GLSA-52).

Dropout rates were also calculated for individual samples, and ranged from 0 (HSU 1836, UMMZ 79760 and LACM 95619) to 40.58% (MVZ 2088), as calculated from the pooled dataset (see Table 6 for a complete list). HSU 8180 recovered a dropout rate of 14.73% and we found individual replicates recovering shorter alleles where primer sequences were clipped differentially, see Appendix 3 for specific details. Following a manual evaluation of genotypes recovered across replicates, we reevaluated the allelic dropout rates, which ranged from 0 (HSU 1836, UMMZ 79760) to 100% in MVZ 5211 where all genotypes were missing. When MVZ 5211 was removed the highest dropout recovered was in LACM 95619 with 77.3%. The LQMS had higher rates of dropout after manually processing genotypes due to the removal of low confidence genotypes.

Discussion:

Mitochondrial results

Unsurprisingly, the sensitivity of Illumina sequencing recovered mitochondrial PCR products which were not easily deciphered from agarose gel electrophoresis. PCR success, as determined solely from gel electrophoresis, was a poor indicator of template existence in low quality samples for mitochondrial DNA. All museum samples recovered the 300 bp fragment of cytochrome *b* , despite only an average of 11% PCR success in the LQMS. The haplotypes recovered here matched with others in a rangewide study of *G. oregonensis* (Yuan, 2020).

Three haplotypes were recovered, and all of the LQMS recovered high quality, reliable sequences. The lowest coverage was in MVZ 2088 which still had minimally 98.6% reads above Q20 and 95.2% reads Q30 and above. From the limited coverage in this sample (22.7x average) there were only about 7 expected sequencing errors. The low error rate combined with a shared haplotype recovered provide evidence that this data was reliable.

Microsatellite Recovery

Microsatellite genotypes were recovered at a higher rate than expected based on failed results from low quality samples in PCR. The LQMS recovered an overall PCR success rate of 21.4% (calculated from the average success for all recorded PCRs of the LQMS from Appendix 1) yet recovered 42% genotyping success (calculated from a total of 30 recovered genotypes from the possible 72 for the LQMS, Table 4). This percentage did not include removal of problematic genotypes. When only genotypes that were agreeable were included, this reduced the 30 genotypes to 6 or 8.3%. Despite this very low rate of confirmation among the LQMS this study quantifies rates for genotyping success via GBS on poorly amplifying museum specimens for the first time. Alternatively, for HQMS the rate of recovered genotypes was 91.7% (66 of 72), and 86.1% (62 of 72) of agreeable genotypes. For the tissue sample 100% (16 of 16) of the replicates resulted in a genotype, with 87.5% (14 of 16) confirmed by the second genotype recovered. This provides robust support that the rate of disagreement shown in the HQMS is negligible and only 1.4% less than the tissue sample.

Samples which routinely amplified produced reliable genotypes. Low quality samples had variable genotypes which seemed stochastic and unreliable across various PCR replicates. Additionally, it appeared that these variable genotypes may have resulted from differences between replicated PCR. This was apparent when sequencing reads were bioinformatically combined as well as from the pooled run. Short microsatellites amplified more frequently and certain microsatellites had more reliable calls than others. This study has shown that samples which do not reliably amplify as assessed by gel electrophoresis may be prone to poor performance in PCR and result in inaccurate genotypes.

We were not able to compare the resulting genotypes recovered here to other population genetics studies on *G. oregonensis* , as this is the first time such a study has occurred in this species. The study performed by Barbian et al., (2018) used GBS to re-genotype chimpanzees with known life-history data and previous

capillary electrophoresis (CE) genotypes, and noted a shift of 1-3 bp in genotype results between the CE and GBS data. They also noted the recovery of additional allelic diversity, which was traditionally lost in homoplasy (and confirmed by pedigree analysis). We also recovered homoplastic events (detailed in Appendix 5) in our data. Darby et al., (2016) noted a 44% increase in alleles due to homoplasy in their dataset as well, and increased alleles from 164 to 294. In other words, 130 novel alleles were recovered from GBS over CE.

Allelic Dropout Rates

Here we recover high rates of detected allelic dropout in samples classified as low quality museum specimens from many rounds of PCR amplification of nuclear and mitochondrial DNA markers. The rates of allelic dropout averaged 12.6% across the five evaluated microsatellite loci for the pooled dataset. When we applied our 'best practices' method from this study the average rate of allelic dropout was reduced to 9.8% across all loci and including the LQMS. When data from the pooled run was evaluated, our rates of allelic dropout in the LQMS (average of 19.6%) conformed to rates reported for various studies of avian fecal allelic dropout (mean of 21%, Regnaut, Lucas, & Fumagalli, 2006) but was much higher than rates reported from chimpanzee fecal samples via GBS (7% rate of allelic dropout; Barbian et al., 2018). This may be due to the fact that Barbian et al., (2018) only genotyped samples at loci that had over 500 reads (counts.min = 500). After manually processing genotypes from all the a priori information (the replicates, pooled run, and bioinformatically pooled data) the rates of dropout in the LQMS increased dramatically following processing to 79%. However, that was due to most genotypes lacking verification, and resulted in only four total genotypes called across all LQMS samples, with MVZ 2088 recovering three genotypes and MVZ 5211 recovering one. It is also worth mentioning that the LQMS samples had very high rates of individual allelic dropout due to unreliable genotypes which were ultimately removed in the manually processed (best practices) genotypes.

The HQMS samples performed better than expected, all recovering very low rates of allelic dropout (2.4% from the pooled run and <0.001% following our best practices). Two samples were collected from 1926 and one from 1975, and provide robust evidence for the utility of museum specimens for the recovery of microsatellite genotypes. These three samples performed on par with GBS studies derived from tissue samples (Darby et al., 2016).

The rate of allelic dropout in our tissue sample was higher upon raw results from CHIIMP, at 14.7% compared to 0.4% in Darby et al., (2016). However, this was partially due to one instance of primer region trimming (GS-4) for this sample. Following implementation of our best practices the rate was reduced to 8.5%, which is still higher than other studies of tissue samples (Darby et al., 2016). Additionally, the aberrant call of 251 from HSU 8180 (GSLA-52, pooled run) was removed during our best practices.

The pooled run CHIIMP results for dropout when separated by length were: 7.5% for short, 0.35% for medium and 48% for long loci. The higher rates for the short loci is likely attributed to non-specific amplification or higher amounts of stutter and PCR artifacts as identified from CHIIMP. Following our manually processed genotypes we recovered 0.0025%, 6.3%, and 36% occurrence of allelic dropout for short, medium and long loci.

Number of Alleles

Here we found that the number of alleles recovered by CHIIMP on the pooled dataset was higher than our manually processed genotypes. The number of alleles per locus stayed the same at two loci (GS-2 and GLSA-12, although a different allele was present in the processed genotypes for both loci; allele '96' in GS-2 and '162' in GLSA-12), and decreased in the other three loci. Due to the nature of the microsatellite loci previously described for *G. sabrinus* , and implemented here, we may have recovered a higher proportion of stutter alleles due to the dinucleotide repeat motif in all loci (even the complex motif microsatellite GLSA-12 had motifs including dinucleotide repeats; Kiesow et al., 2011; Zittlau et al., 2000).

When comparing our allele counts to those from *G. sabrinus* we recovered fewer alleles for four loci, GS-2, GS-4, GLSA-22 and GLSA-52, and recovered more alleles in GLSA-12. However, previous research in the sister species *G. sabrinus* included far more individuals and a wider geographic distribution. Therefore, we

do not expect the alleles recovered here to be exhaustive for *G. oregonensis* , and our allele counts seem reasonable for a sample size of seven individuals.

Best Practices:

Museum collections are increasingly being used for molecular sequencing, yet comparative studies on the retrieval and reliability of microsatellite genotypes from these data sources are not readily available. Here we show that while museum specimens can recover reliable, and important genotypes for rare, endangered and elusive species, additional precautions must be made prior to acceptance of genotypes.

From our data we recommend a minimum of three successful amplifications for each marker of interest. The samples which routinely amplified (HQMS) recovered genotypes with very similar rates of genotype confirmation between replicates, and when compared to the tissue sample. Poor performing samples may require additional replication compared to better performing samples. We noticed in the LQMS that the longer the microsatellite locus, the worse the marker amplified. This was apparent when none of the LQMS recovered genotypes for the ˜250 bp microsatellite locus GLSA-52. All of the HQMS samples recovered reliable genotypes across replicates tested here, and for all marker lengths and types of repeat motif.

Our data separated sample types into three categories, tissue, HQMS and LQMS, the latter two designations were only applied after many rounds of PCR and agarose gel visualization. During project design samples should be evaluated so that adequate replicates of PCR can be performed to ensure accurate genotypes. Additionally, calibration/confirmation of the genotypes generated by GBS can be done via CE or other fragment visualizing instruments (Fragment Analyzer, Advanced Analytical Ankeney, IA). It is notable that genotypes may be predictably shifted from comparison of GBS and CE methods as detailed in Barbian et al., (2018).

In order to reduce the inaccurate genotypes, optimization of PCRs should be performed prior to GBS. The addition of various reagents has been shown to increase specificity and reduce non-specific amplification, as has been widely published over the past 30 years (Boleda, Briones, Farres, Tyfield, & Pi, 1996; Robertson & Walsh-Weller, 1998; J. F. Williams, 1989). The PCRs performed here incorporated a touchdown protocol, which starts at a high annealing temperature (60degC) for 2 cycles of PCR before reducing to the lowest annealing temperature of 50degC for 35 cycles. Touchdown PCR was used on the museum specimens and across loci as it was shown to effectively amplify all microsatellite markers. Two microsatellites (GS-2 and GS-4) had Bovine Serum Albumin (BSA) added since, during initial PCR testing, BSA improved amplification success. GS-4 however, recovered numerous unconfirmed genotypes, which could be related to input DNA quality, or from poor performance in PCR. This locus in particular would benefit from additional optimization in order to determine if non-specific amplification could be reduced. GS-2 may have also benefited from additional optimization as that locus recovered numerous flags from the CHIIMP pipeline including PCR artifacts, PCR stutter and more than two prominent sequences.

The CHIIMP pipeline worked well on our samples after modification of published protocols (Barbian et al., 2018). We found it useful to combine the CHIIMP genotypes with the quality data as determined by the proportion of reads passing prinseq filtering to evaluate which samples may be more prone to false/inaccurate genotypes. The combination of multiple rounds of PCR, prinseq quality filtering and manual evaluation of CHIIMP results allowed increased confidence in the genotypes recovered by museum specimens in this study. This process is illustrated in Figure 2, and summarized here. First, we would assign our samples as high or low quality following multiple attempts of PCR with agarose gel visualization. Second, based on these findings, we suggest recovering minimally 3 successful PCR replicates prior to genotyping. If PCRs continue to fail, optimization of each locus may be helpful, as well as evaluation of DNA extracts for the presence of PCR inhibitors, which has been shown to affect recovery of ancient DNA and environmental DNA samples (Matheson et al., 2009; McKee, Spear, & Pierson, 2015; Pontiroli et al., 2011). Once successful amplification has occurred across all samples and markers, perform library preparation on successfully amplified PCR products and sequence on an Illumina platform with adequate insert length for the included microsatellites. Sequence to a minimum depth of 1000 reads per sample per microsatellite marker. For our data that would

entail 5000 sequences per sample.

Demultiplexed data should have CutAdapt and FastQC performed in order to run CHIIMP v 0.3.1. Simultaneously, run prinseq as a parallel analysis to determine the overall quality of the samples. Samples with higher proportions of low quality reads should be noted as they may be more prone to erroneous genotypes. When alleles are recovered only in low quality samples, it is imperative to look at the output from CHIIMP, and determine if the differences are associated with primer sequences or repeat elements. If primer sequence varies, manually correct the length when the entire primer sequence would be included, and ignore primer site size mismatches in allele calls as this is likely an artifact of sequencing or amplification errors. Traditionally, fragment size analysis via CE would ignore peaks outside of the expected size range via programs like GeneMapper (Applied Biosystems). If an allele does not have a priming site error, it is important to evaluate if the size shift follows microsatellite evolutionary patterns, for example, if it is two base pairs shifted in a dinucleotide sequence that makes evolutionary sense. However, stutter sequences are often frame shifted by the size of the repeat motif. While CHIIMP evaluates for stutter, and allows filter manipulation, we found it was possibly including false alleles due to the nature of our markers. Dinucleotide sourced microsatellites are more challenging regarding stutter evaluation due to the short difference in size between true alleles and stutter peaks. (Barbian et al., 2018; O'reilly, Canino, Bailey, & Bentzen, 2000). In order to further scrutinize stutter sequences, we calculated the proportion of reads associated with the various alleles. If one call makes up a very small percentage (30% the number of reads as the other allele) it is likely a stutter peak. Further visualization via electropherograms could illuminate this process if rampant in a locus. Additionally, the number of reads represented by the allele can also provide insight into whether or not the polymorphism is due to sequencing error. By following these practices we reduced our allelic dropout by about 3% across loci. We also had to remove many of the LQMS genotypes as we could not be certain they were authentic. The CHIIMP pipeline also allows for optimization and customization of commands for recovering more strict versus lenient genotypes. Here we modified the published parameters based on the depth of coverage of our samples.

Conclusions

Genotyping by synthesis is an effective way to generate affordable genotype results for degraded specimens when stringent protocols and deep sequencing is performed. Our costs were under ~$15 per sample, details provided in Appendix 6. This was very comparable to other GBS studies (Darby et al., 2016), and notably does not require the initial investment in fluorescently labelled primers, but does require sequencing adapters, as well as the ability to fill a sequencing run. We also only performed singleplex PCR, and if time was spent on designing multiplex PCRs the cost of taq could be significantly reduced. If for example, two microsatellites were multiplexed the cost per sample would be reduced to $13.30/sample, and if three microsatellites were pooled the overall cost would be reduced to $12.81/sample.

Several bioinformatic pipelines have already been developed to generate microsatellite genotypes from HTS data (Barbian et al., 2018; De Barba et al., 2017; Pimentel et al., 2018; Tibihika et al., 2019), and have screened a variety of starting template types including tissue samples, hair and fecal samples. This is the first time GBS methods (employing an existing pipeline developed for fecal samples) has been applied to evaluate the error rates from museum specimen derived DNA samples. Our results show that when reliable amplification occurs, robust genotyping can be recovered from museum specimens, especially samples deemed HQMS. The rates of agreement between genotypes were nearly identical between the HQMS and our tissue sample. For low quality samples repeated PCR is necessary, and does not completely eliminate the opportunity for a false genotype to be included in a dataset. This, however, is also known from CE fragment size analysis, and many studies have reported shifted alleles of the same PCR products on different runs of an automated capillary sequencer, or with a different size standard (Ellis et al., 2011; Haberl & Tautz, 1999). We believe that our allele calls for the HQMS are robust and contribute valuable data points to studies where historical data is not available. This study provides best practices for the genotyping of degraded source samples.

Previous studies have shown that the type of museum specimen sample obtained (bone, skin, hair, cartilage,

nail) may have more of an effect than age on recovery of DNA (Hawkins, Hofman, et al., 2016; McDonough et al., 2018), yet here, based on our limited sample size the worst performing samples for microsatellites were in fact the oldest (1905-1919 Table 1). Hawkins, Hofman, et al., (2016) only evaluated mitochondrial DNA recovery and from in-solution hybridization and McDonough et al., (2018) recovered variable concentrations of mtDNA versus nDNA, with mtDNA unexpectedly recovering approximately an order of magnitude more sequencing depth than nDNA. It is worth noting however, that many samples from our expanded dataset (Yuan, 2020) were as old as the LQMS here, yet reliably amplified for the same microsatellite loci. Due to these factors we refrain from further speculation on the patterns of degradation associated with age for nuclear DNA content in museum specimens.

The LQMS genotypes recovered require fine scale evaluation to ensure accuracy and repeatability for downstream analyses, as inaccurate allele calls can affect population genetic inferences. Variable genotypes were much more prevalent in the low quality samples (16 instances in the LQMS versus only two in the HQMS). These variable genotypes may not be specifically due to allelic dropout which is commonly seen in fecal samples (Piggott, Bellemain, Taberlet, & Taylor, 2004; Regnaut et al., 2006), since alleles which appear to be outside the expected bin sizes were recovered (see GS-4 for the LQMS), and only rarely did potential allelic dropout appear (see GS-2 for LACM 95619). Further optimization of the CHIIMP pipeline may allow for elimination of those genotypes with the size buffer setting. Additionally, all samples recovered reliable mtDNA signatures, where many (particularly the LQMS) lacked nDNA at many loci. One sample (MVZ 5211) had incredibly high cytochrome *b*coverage, yet no reliable nDNA genotypes.

The integration of using microsatellite markers on degraded samples and using the improved resolution from GBS will allow further comparison to the plethora of published studies on microsatellites. Museum specimens are very important to utilize as they give both temporal perspective and representation of rare species. But, appropriate QC measures need to be undertaken to ensure accuracy of recovered genotypes. We believe that this data illuminates the possibility of reliably incorporating microsatellite genotypes from specimens from the early 20th century museum collection in combination with modern surveys to evaluate genetic shifts and population genomics through space and time.

Data Accessibility

Cytochrome *b* sequences can be found on GenBank accessions MT498442-MT498448. Microsatellite output files from the CHIIMP pipeline can be found at Dryad DOI:*https://doi.org/10.5061/dryad.t1g1jwt07*

Competing Interests Statement

The authors declare no competing interests

Author Contributions

MTRH and SCY conceived of the study and performed laboratory work. SCY and EM performed bioinformatics. MTRH wrote the manuscript and performed statistical analyses. All authors analyzed the data, and approved of the final manuscript.

Acknowledgements

The authors wish to thank a number of individuals who contributed to this study. Laboratory assistance was provided by Clare O'Connell, Michael Kiso, Jack Lemke and Evan Miller. Nancy Rotzel and Katie Murphy are thanked for performing the sequencing runs at the Center for Conservation Genomics, Smithsonian Conservation Biology Institute, and the Laboratory of Analytical Biology, National Museum of Natural History, Smithsonian Institution, respectively. Several museums allowed for destructive sampling of specimens; the Museum of Vertebrate Zoology, University of California Berkeley, Chris Conroy, Jim Patton, Eileen Lacey and Michael Nachman, Los Angeles County Museum, Jim Dines and Kayce Bell, and the Humboldt State University Vertebrate Museum, Alyssa Semerdjian, Nick Kerhoulas and Allison Bronson, and the University of Michigan Museum of Zoology, Cody Thompson. We also express gratitude to Beatrice Hahn and Jesse Connell for their assistance implementing the CHIIMP pipeline. This study was funded by MTRH's dis-

cretionary funds, as well as a Grants-in-Aid award from the American Society of Mammalogists, Sigma Xi Grants-in-Aid of Research award (G201903158734905) and the Humboldt State University Department of Biology Master's Student Grant.

References:

Bailey, C. A., McLain, A. T., Paquette, S. R., McGuire, S. M., Shore, G. D., & Lei, R. (2015). Evaluating the genetic diversity of three endangered lemur species (Genus: Propithecus) from northern Madagascar. *Journal of Primatology* , *5* , 132.

Bandelt, H. J., Forster, P., & Rohl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* , *16* (1), 37–48. doi: 10.1093/oxfordjournals.molbev.a026036

Barbian, H. J., Connell, A. J., Avitto, A. N., Russell, R. M., Smith, A. G., Gundlapally, M. S., ... Wroblewski, E. E. (2018). CHIIMP: An automated high-throughput microsatellite genotyping platform reveals greater allelic diversity in wild chimpanzees. *Ecology and Evolution* , *8* (16), 7946–7963.

Bilska, K., & Szczecińska, M. (2016). Comparison of the effectiveness of ISJ and SSR markers and detection of outlier loci in conservation genetics of Pulsatilla patens populations. *PeerJ* , *4* , e2504.

Blagoderov, V., Kitching, I. J., Livermore, L., Simonsen, T. J., & Smith, V. S. (2012). No specimen left behind: industrial scale digitization of natural history collections. *ZooKeys* , (209), 133.

Boleda, M. D., Briones, P., Farres, J., Tyfield, L., & Pi, R. (1996). Experimental design: a useful tool for PCR optimization. *BioTechniques* , *21* (7), 134–140.

Campana, M. G., Lister, D. L., Whitten, C. M., Edwards, C. J., Stock, F., Barker, G., & Bower, M. A. (2012). Complex relationships between mitochondrial and nuclear DNA preservation in historical DNA extracts. *Archaeometry* , *54* (1), 193–202.

Crawford, A. M., Kappes, S. M., Paterson, K. A., deGotari, M. J., Dodds, K. G., Freking, B. A., ... Beattie, C. W. (1998). Microsatellite evolution: testing the ascertainment bias hypothesis. *Journal of Molecular Evolution* , *46* (2), 256–260.

Darby, B. J., Erickson, S. F., Hervey, S. D., & Ellis-Felege, S. N. (2016). Digital fragment analysis of short tandem repeats by high-throughput amplicon sequencing. *Ecology and Evolution* ,*6* (13), 4502–4512.

De Barba, M., Miquel, C., Lobréaux, S., Quenette, P. Y., Swenson, J. E., & Taberlet, P. (2017). High-throughput microsatellite genotyping in ecology: improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. *Molecular Ecology Resources* ,*17* (3), 492–507.

Demboski, J. R., Jacobsen, B. K., & Cook, J. A. (1998). Implications of cytochrome b sequence variation for biogeography and conservation of the northern flying squirrels (Glaucomys sabrinus) of the Alexander Archipelago, Alaska. *Canadian Journal of Zoology* , *76* (9), 1771–1777.

den Tex, R.-J., Maldonado, J. E., Thorington, R., & Leonard, J. A. (2010). Nuclear copies of mitochondrial genes: another problem for ancient DNA. *Genetica* , *138* (9–10), 979–984. doi: 10.1007/s10709-010-9481-9

Duan, C., Li, D., Sun, S., Wang, X., & Zhu, Z. (2014). Rapid development of microsatellite markers for Callosobruchus chinensis using Illumina paired-end sequencing. *PloS One* , *9* (5).

Ellis, J. S., Gilbey, J., Armstrong, A., Balstad, T., Cauwelier, E., Cherbonnel, C., ... Crozier, W. (2011). Microsatellite standardization and evaluation of genotyping error in a large multi-partner research programme for conservation of Atlantic salmon (Salmo salar L.). *Genetica* , *139* (3), 353–367.

Estoup, A., Jarne, P., & Cornuet, J.-M. (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology* , *11* (9), 1591–1604.

Fisher, P. J., Gardner, R. C., & Richardson, T. E. (1996). Single locus microsatellites isolated using 5' anchored PCR. *Nucleic Acids Research* , *24* (21), 4369–4371.

Glenn, T. C., Nilsen, R. A., Kieran, T. J., Sanders, J. G., Bayona-Vasquez, N. J., Finger, J. W., ... Louha, S. (2019). Adapterama I: universal stubs and primers for 384 unique dual-indexed or 147,456 combinatorially-indexed Illumina libraries (iTru & iNext).*PeerJ* , *7* , e7755.

Glenn, T. C., & Schable, N. A. (2005). Isolating microsatellite DNA loci. In *Methods in enzymology* (Vol. 395, pp. 202–222). Elsevier.

Griffiths, S. M., Fox, G., Briggs, P. J., Donaldson, I. J., Hood, S., Richardson, P., ... Preziosi, R. F. (2016). A Galaxy-based bioinformatics pipeline for optimised, streamlined microsatellite development from Illumina next-generation sequencing data.*Conservation Genetics Resources* , *8* (4), 481–486.

Grimaldi, M.-C., & Crouau-Roy, B. (1997). Microsatellite allelic homoplasy due to variable flanking sequences. *Journal of Molecular Evolution* , *44* (3), 336–340.

Haberl, M., & Tautz, D. (1999). Comparative allele sizing can produce inaccurate allele size differences for microsatellites. *Molecular Ecology* , *8* (8), 1347–1349.

Hawkins, M. T., Hofman, C. A., Callicrate, T., McDonough, M. M., Tsuchiya, M. T., Gutierrez, E. E., ... Maldonado, J. E. (2016). In-solution hybridization for mammalian mitogenome enrichment: Pros, cons and challenges associated with multiplexing degraded DNA.*Molecular Ecology Resources* , *16* (5), 1173–1188.

Hawkins, M. T., Leonard, J. A., Helgen, K. M., McDonough, M. M., Rockwood, L. L., & Maldonado, J. E. (2016). Evolutionary history of endemic Sulawesi squirrels constructed from UCEs and mitogenomes sequenced from museum specimens. *BMC Evolutionary Biology* ,*16* (1), 80.

Hofreiter, M., Serre, D., Poinar, H., Kuch, M., & Paabo, S. (2001). Ancient DNA. *Nature Reviews Genetics* , *2* , 353–359.

Irwin, D. M., & Kocher, T. D. (1991). Evolution of the cytochromeb gene of mammals. *Journal of Molecular Evolution* , *32* (2), 128–144.

Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P. L., & Orlando, L. (2013). mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* , *29* (13), 1682–1684.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.*Molecular Biology and Evolution* , *30* (4), 772–780. doi: 10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., ... Thierer, T. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* , *28* (12), 1647–1649.

Kiesow, A. M., Wallace, L. E., & Britten, H. B. (2011). Characterization and isolation of five microsatellite loci in northern flying squirrels, Glaucomys sabrinus (Sciuridae, Rodentia). *Western North American Naturalist* , *71* (4), 553–556.

Lane, M. A. (1996). Roles of Natural History Collections. *Annals of the Missouri Botanical Garden* , *83* (4), 536–545. doi: 10.2307/2399994

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* , *9* (4), 357–359. doi: 10.1038/nmeth.1923

Li, B., & Kimmel, M. (2013). Factors influencing ascertainment bias of microsatellite allele sizes: impact on estimates of mutation rates.*Genetics* , *195* (2), 563–572.

Lian, C. L., Wadud, M. A., Geng, Q., Shimatani, K., & Hogetsu, T. (2006). An improved technique for isolating codominant compound microsatellite markers. *Journal of Plant Research* , *119* (4), 415–417.

Lister, A. M., & Group, C. C. R. (2011). Natural history collections as sources of long-term datasets. *Trends in Ecology & Evolution* ,*26* (4), 153–154.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal* , *17* (1), 10. doi: 10.14806/ej.17.1.200

Matheson, C. D., Marion, T. E., Hayter, S., Esau, N., Fratpietro, R., & Vernon, K. K. (2009). Removal of metal ion inhibition encountered during DNA extraction and amplification of copper-preserved archaeological bone using size exclusion chromatography. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* , *140* (2), 384–391.

McDonough, M. M., Parker, L. D., Rotzel McInerney, N., Campana, M. G., & Maldonado, J. E. (2018). Performance of commonly requested destructive museum samples for mammalian genomic studies. *Journal of Mammalogy* , *99* (4), 789–802.

McKee, A. M., Spear, S. F., & Pierson, T. W. (2015). The effect of dilution and the use of a post-extraction nucleic acid purification column on the accuracy, precision, and inhibition of environmental DNA samples. *Biological Conservation* , *183* , 70–76.

Miller, M. P., Knaus, B. J., Mullins, T. D., & Haig, S. M. (2013). SSR_pipeline: A bioinformatic infrastructure for identifying microsatellites from paired-end Illumina high-throughput DNA sequencing data. *Journal of Heredity* , *104* (6), 881–885.

Miller, W., Drautz, D. I., Janecka, J. E., Lesk, A. M., Ratan, A., Tomsho, L. P., ... Qi, J. (2009). The mitochondrial genome sequence of the Tasmanian tiger (Thylacinus cynocephalus). *Genome Research* , *19* (2), 213–220.

Morin, P. A., Manaster, C., Mesnick, S. L., & Holland, R. (2009). Normalization and binning of historical and multi-source microsatellite data: overcoming the problems of allele size shift with allelogram.*Molecular Ecology Resources* , *9* (6), 1451–1455.

O'reilly, P. T., Canino, M. F., Bailey, K. M., & Bentzen, P. (2000). Isolation of twenty low stutter di-and tetranucleotide microsatellites for population analyses of walleye pollock and other gadoids.*Journal of Fish Biology* , *56* (5), 1074–1086.

Oshida, T., Lin, L. K., Masuda, R., & Yoshida, M. C. (2000). Phylogenetic relationships among Asian species of Petaurista (Rodentia, Sciuridae), inferred from mitochondrial cytochrome b gene sequences.*Zoologica Science* , *17* (1), 123–128.

Paabo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., ... Hofreiter, M. (2004). Genetic Analyses from Ancient DNA. *Annual Review of Genetics* , *38* (645–679).

Piggott, M. P., Bellemain, E., Taberlet, P., & Taylor, A. C. (2004). A multiplex pre-amplification method that significantly improves microsatellite amplification and error rates for faecal DNA in limiting conditions. *Conservation Genetics* , *5* (3), 417–420.

Pimentel, J. S., Carmo, A. O., Rosse, I. C., Martins, A. P., Ludwig, S., Facchin, S., ... Kalapothakis, E. (2018). High-throughput sequencing strategy for microsatellite genotyping using neotropical fish as a model. *Frontiers in Genetics* , *9* , 73.

Pontiroli, A., Travis, E. R., Sweeney, F. P., Porter, D., Gaze, W. H., Mason, S., ... Wellington, E. M. H. (2011). Pathogen quantitation in complex matrices: a multi-operator comparison of DNA extraction methods with a novel assessment of PCR inhibition. *PloS One* ,*6* (3).

Regnaut, S., Lucas, F. S., & Fumagalli, L. (2006). DNA degradation in avian faecal samples and feasibility of non-invasive genetic studies of threatened capercaillie populations. *Conservation Genetics* ,*7* (3), 449–453.

Rizzi, E., Lari, M., Gigli, E., De Bellis, G., & Caramelli, D. (2012). Ancient DNA studies: new perspectives on old samples. *Genetics Selection Evolution* , *44* (1), 21.

Robertson, J. M., & Walsh-Weller, J. (1998). An introduction to PCR primer design and optimization of amplification reactions. In*Forensic DNA profiling protocols* (pp. 121–154). Springer.

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research* , gr.128124.111. doi: 10.1101/gr.128124.111

Šarhanová, P., Pfanzelt, S., Brandt, R., Himmelbach, A., & Blattner, F. R. (2018). SSR-seq: Genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring. *Ecology and Evolution* ,*8* (22), 10817–10833.

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)* ,*27* (6), 863–864. doi: 10.1093/bioinformatics/btr026

Shapiro, B. (2012). *Ancient DNA: Methods and Protocols* . Retrieved from http://library.wur.nl/WebQuery/clc/1989945

Silva, P. I., Martins, A. M., Gouvea, E. G., Pessoa-Filho, M., & Ferreira, M. E. (2013). Development and validation of microsatellite markers for Brachiaria ruziziensis obtained by partial genome assembly of Illumina single-end reads. *Bmc Genomics* , *14* (1), 17.

Smith, A. B., Santos, M. J., Koo, M. S., Rowe, K. M., Rowe, K. C., Patton, J. L., ... Moritz, C. (2013). Evaluation of species distribution models by resampling of sites surveyed a century ago by Joseph Grinnell. *Ecography* , *36* (9), 1017–1031.

Thatte, P., Joshi, A., Vaidyanathan, S., Landguth, E., & Ramakrishnan, U. (2018). Maintaining tiger connectivity and minimizing extinction into the next century: Insights from landscape genetics and spatially-explicit simulations. *Biological Conservation* ,*218* , 181–191.

Tibihika, P. D., Curto, M., Dornstauder-Schrammel, E., Winter, S., Alemayehu, E., Waidbacher, H., & Meimberg, H. (2019). Application of microsatellite genotyping by sequencing (SSR-GBS) to measure genetic diversity of the East African Oreochromis niloticus. *Conservation Genetics* , *20* (2), 357–372.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Untergasser, Andreas, et al. "Primer3—new capabilities and interfaces. *Nucleic Acids Research* , *40* (15), e115.

Vartia, S., Villanueva-Cañas, J. L., Finarelli, J., Farrell, E. D., Collins, P. C., Hughes, G. M., ... Cross, T. F. (2016). A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. *Royal Society Open Science* , *3* (1), 150565.

Wang, C., & Rosenberg, N. A. (2012). MicroDrop: a program for estimating and correcting for allelic dropout in nonreplicated microsatellite genotypes version 1.01. *See Https://Web. Stanford. Edu/Group/Rosen/Berglab/Microdrop. Html* .

Weiß, C. L., Schuenemann, V. J., Devos, J., Shirsekar, G., Reiter, E., Gould, B. A., ... Burbano, H. A. (2016). Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *Royal Society Open Science* , *3* (6), 160239.

White, L. C., Mitchell, K. J., & Austin, J. J. (2018). Ancient mitochondrial genomes reveal the demographic history and phylogeography of the extinct, enigmatic thylacine (Thylacinus cynocephalus).*Journal of Biogeography* , *45* (1), 1–13.

Williams, J. F. (1989). Optimization strategies for the polymerase chain reaction. *Biotechniques* , *7* (7), 762–769.

Williams, S. L. (1999). *Destructive preservation: A review of the effect of standard preservation practices on the future use of natural history collections* .

Yuan, S. (2020). *PHYLOGENETIC AND POPULATION GENETIC ANALYSIS OF THE HUMBOLDT'S FLYING SQUIRREL USING HIGH-THROUGHPUT SEQUENCING DATA* . Humboldt State University, Arcata, CA.

Zhan, L., Paterson, I. G., Fraser, B. A., Watson, B., Bradbury, I. R., Nadukkalam Ravindran, P., ... Bentzen, P. (2017). MEGASAT: automated inference of microsatellite genotypes from sequence data. *Molecular Ecology Resources* , *17* (2), 247–256.

Zittlau, K. A., Davis, C. S., & Strobeck, C. (2000). Characterization of microsatellite loci in northern flying squirrels (Glaucomys sabrinus). *Molecular Ecology* , *9* (6), 826–827.

Figures:

Figure 1: Scatter plot of average quality of PCR replicates following QC from prinseq-lite. The replicates for each specimen are shown across the x axis, and the % of 'good' reads are shown on the y-axis. Samples are sorted by type: tissue, high quality museum specimen and low quality museum specimens. Individuals of the same type are separated by a dashed line.

Figure 2: A flow chart of best practices for sample analysis performed here, with particular emphasis on how to assess low quality museum specimens. Note: many steps detailed under 'CHIIMP Output' can be manipulated when running the pipeline.

Tables:

Table 1: Summary of samples used in this study. Color coded throughout as green= tissue sample, blue = high quality museum specimen (also referred to as HQMS hereafter), and red = low quality museum specimens (LQMS). Samples were acquired from approved destructive sampling requests from several national museum collections as follows: HSU= Humboldt State University Vertebrate Museum, UMMZ= University of Michigan Museum of Zoology, MVZ= Museum of Vertebrate Zoology, University of California Berkeley, and LACM= Los Angeles County Museum.

Table 2: Primers used for microsatellite and mitochondrial cytochrome b amplification. Microsatellites were characterized by the length (S=short, under 150 bp, M=medium, 150-200 bp, and L= long, over 200 bp) and repeat motif as simple (standard dinucleotide repeat motif in all included microsatellites) and complex (where repeat motifs were interrupted by variable repeat units- only seen in GLSA12 here). The reverse primer for mitochondrial cytochrome *b* was newly designed for this study as previously published internal cytochrome *b*primers did not amplify in *G. oregonensis* .

Table 3. Quality of individually prepared libraries as assessed by prinseq-lite. The sample name and replicate number are indicated in the Sample ID field, and the 'all' row includes the replicate counts when bioinformatically summed together. The 'combined' ID (also shown in bold) represents the second library prep where all microsatellite replicates + cytochrome *b* were pooled and run through the CHIIMP pipeline. Only quality data from the forward read (R1) is shown here. Samples retain the same color coding for tissue= green, HQMS= blue and LQMS= red. Average quality is shown per replicate, read counts and the percentage of reads passing prinseq-lite quality filtering are shown across all microsatellite replicates as well. Range was calculated from each individual replicate and excluded bioinformatically summed and combined library prep data.

Table 4: Summary of recovered genotypes from the CHIIMP pipeline (Barbian et al. 2018). All recovered genotypes are provided, any areas where a '-' is found indicates no recovered genotype from that replicate. The accuracy across each amplification was calculated as well as the average accuracy per microsatellite and across sample type (tissue, HQMS and LQMS). The bioinformatically combined dataset as well as the pooled dataset genotypes are also provided. CHIIMP output provides various metrics on quality and as such a * represents where possible PCR stutter was removed, a $^{r}epresentswhere PCR artifacts were removed and represents where more than two prominent sequences were found. Genotypes$

Table 5: Descriptive Statistics of samples sorted by type, either 'tissue', 'HQMS' or 'LQMS'.

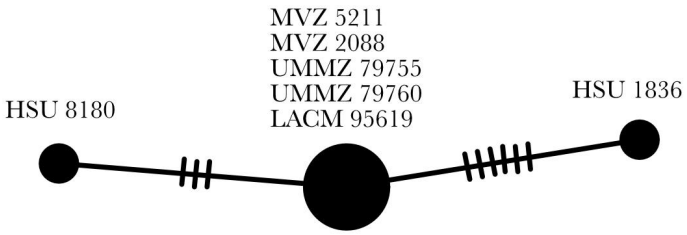
Table 6: MicroDrop results for bioinformatically combined datasets with a comparison of the raw, initial CHIMP output to the final, processed data. Both locus specific and individual rates of estimated allelic dropout are provided.

Appendices

Appendix 1: PCR Amplification success of the five included microsatellites and mitochondrial cytochrome b gene. Success was determined by the presence of a band in the expected size range from an agarose gel. Note: HSU 8180 represents a tissue sample so only two PCR replicates of each marker were performed.

Appendix 2: Coverage of cytochrome *b* across all samples, bowtie 2 v 2.3.0 was used to map reads and Geneious Prime calculated the included quality metrics. Quality metrics provided from the 300 bp fragment amplified in all samples except HSU 8180 for which the entire CDS (1,140 bp) was amplified and analyzed.

Appendix 3: Mitochondrial minimum spanning network.



Appendix 4: Details of variable genotypes recovered here, separated by locus.

GS 2

Forward primer: AACATTCTCGCCACATCTAA

Repeat motif: GT

Reverse primer: CTACACCCCCAGCCCTACAA

Reverse complement: TTGTAGGGCTGGGGGTGTAG

Nucleotide differences within the sample

>IndividualName_Replicate **BOLDbp** = allele call allele/locus/total (# of reads of specific allele, total reads for specific locus, total reads for the sample)

HSU_1836

Pooled:

>HSU1836-GS2_1 **98bp** allele/locus/total=6/33/6548

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTCTTTTTAC

GT : 16

Without primers = 58 bp

Allele Call= 98

>HSU1836-GS2_2 **104bp** allele/locus/total=4/33/6548

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTC

GT : 19

Without primers = 64 bp

Allele Call= 104

Individual Replicates:

Replicate 1:

>HSU1836GS21-1-GS2_1 **98bp**allele/locus/total=207/1611/12066

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTC

GT : 16

Without primers = 58 bp

>HSU1836GS21-1-GS2_2 **104bp**allele/locus/total=174/1611/12066

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTC

GT : 19

Without primers = 64 bp

Replicate 2:

>HSU1836GS22-2-GS2_1 **98bp**allele/locus/total=319/1925/10206

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTC

GT : 16

Without primers = 58 bp

>HSU1836GS22-2-GS2_2 **96bp**allele/locus/total=197/1925/10206

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTCTTTTTAGTC

GT : 15

Without primers = 56 bp

Replicate 3:

>HSU1836GS23-3-GS2_1 **98bp**allele/locus/total=269/959/9780

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTC

GT : 16

Without primers = 58 bp

>HSU1836GS23-3-GS2_2 **96bp**allele/locus/total=152/959/9780

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTCTTTTTAGTC

GT : 15

Without primers = 56 bp

Bioinformatically Combined:

>HSU1836GS2-GS2_1 **98bp**allele/locus/total=801/4487/32051

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTCTTTTTAC

GT : 16

Without primers = 58 bp

>HSU1836GS2-GS2_2 **96bp**allele/locus/total=435/4487/32051

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTTTCTTTTTAGTC

GT : 15

Without primers = 56 bp

LACM_95619

Pooled:

>LACM95619-GS2_1 **84bp**allele/locus/total=318/933/61226

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTTTCTTTTTAGTCTATTGTAGGGCT

GT : 9

Without primers = 44 bp

>LACM95619-GS2_2 **82bp**allele/locus/total=285/933/61226

AACATTCTCGCCACATCTAAATCTGCTCTGGGGTGTGTGTGTGTGTGTTTCTTTTTAGTCTATTGTAGGGCTGG

GT : 8

Without primers = 42 bp

Individual Replicates:

Replicate 1:

>LACM95619GS21-1-GS2_1 **82bp**allele/locus/total=2372/3463/8470

AACATTCTCGCCACATCTAAATCTGCTCTGGGGTGTGTGTGTGTGTGTTTCTTTTTAGTCTATTGTAGGGCTGG

GT : 8

Without primers = 42 bp

Replicate 3:

>LACM95619GS23-3-GS2_1 **84bp**allele/locus/total=1700/2952/8898

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTGTTTCTTTTTAGTCTATTGTAGGGCT

GT : 9

Without primers = 44 bp

Bioinformatically Combined:

>LACM95619GS2-GS2_1 **82bp**allele/locus/total=2373/6414/19436

AACATTCTCGCCACATCTAAATCTGCTCTGGGGTGTGTGTGTGTGTGTTTCTTTTTAGTCTATTGTAGGGCTGG

GT : 8

Without primers = 42 bp

>LACM95619GS2-GS2_2 **84bp**allele/locus/total=1700/6414/19436

AACATTCTCGCCACATCTAAGTCTGCTCTGGGGTGTGTGTGTGTGTGTTTCTTTTTAGTCTATTGTAGGGCT

GT : 9

Without primers = 44 bp

GS 4

Forward primer: CTTCTTGAGTTGCTGGGGTGAC

Repeat motif: GT

Reverse primer: ATCTAAACCATAACACACACACA

Reverse complement: TGTGTGTGTGTTATGGTTTAGAT

>IndividualName_Replicate **BOLDbp** = allele call allele/locus/total (# of reads of specific allele, total reads for specific locus, total reads for the sample)

HSU_8180

Pooled:

>HSU8180VM2551-GS4_1 **96bp**allele/locus/total=352/1411/158834

CTTCTTGAGTTGCTGGGGTGACAGGTGTGTGCCACCATGTGGTGAGCCTCATATTCTTTTTAGTGTGTGTGTGT

GT : 3

Without primers = 51 bp

Individual Replicates:

Replicate 1:

>HSU8180GS41-1-GS4_1 **96bp**allele/locus/total=232/1236/8132

CTTCTTGAGTTGCTGGGGTGACAGGTGTGTGCCACCATGTGGTGAGCCTCATATTCTTTTTAGTGTGTGTGTGT

GT : 3

Without primers = 51 bp

>HSU8180GS41-1-GS4_2 **92bp**allele/locus/total=209/1236/8132

CTTCTTGAGTTGCTGGGGTGACAGGTGTGTGCCACCATGTGGTGAGCCTCATATTCTTTTTAGTGTGTGTGTGT

GT : 3

Without primers = 51 bp

Replicate 2:

>HSU8180Gs42-2-GS4_1 **96bp**allele/locus/total=236/909/6772

CTTCTTGAGTTGCTGGGGTGACAGGTGTGTGCCACCATGTGGTGAGCCTCATATTCTTTTTAGTGTGTGTGTGT

GT : 3

Without primers = 51 bp

>HSU8180Gs42-2-GS4_2 **94bp**allele/locus/total=89/909/6772

CTTCTTGAGTTGCTGGGGTGACAGGTGTGTGCCACCATGTGGTGAGCCTCATATTCTTTTTAGTGTGTGTGTGT

GT : 3

Without primers = 49 bp

Bioinformatically Combine):

>HSU8180GS4-GS4_1 **96bp**allele/locus/total=680/2146/14904

CTTCTTGAGTTGCTGGGGTGACAGGTGTGTGCCACCATGTGGTGAGCCTCATATTCTTTTTAGTGTGTGTGTGT

GT : 3

Without primers = 51 bp

MVZ_5211

Pooled:

No genotype

Individual Replicates:

Replicate 1:

>MVZ5211GS41-1-GS4_1 **104bp**allele/locus/total=505/1033/3152

CTTCTTGAGTTGCTGGGGTGACACTGTTCAGCGTGTGTTGCGGGTGTGTGTTTGTGTGCGTGCTGCGGGTGTC

GT : ?

Without primers = 59 bp

Replicate 2:

>MVZ5211GS42-2-GS4_1 **97bp**allele/locus/total=32/203/1948

CTTCTTGAGTTGCTGGGGTGACATCTAAAGCGGAATTATAATAATTGTGATGATGATGATGTTGATGATAGTGTT

GT : ?

Without primers = 52 bp

>MVZ5211GS42-2-GS4_2 **105bp**allele/locus/total=27/203/1948

CTTCTTGAGTTGCTGGGGTGACTGTCTGTGTGTCTGTGTGTGTGTCTGTCTATGTGTGTCTGTCTATGTGTGTC

GT : ?

Without primers = 60 bp

Bioinformatically Combined:

>MVZ5211GS4-GS4_1 **104bp**allele/locus/total=260/2304/20202

CTTCTTGAGTTGCTGGGGTGACACTGTTCAGCGTGTGTTGCGGGTGTGTGTTTGTGTGCGTGCTGCGGGTGTC

GT : ?

Without primers = 59 bp

>MVZ5211GS4-GS4_2 **100bp**allele/locus/total=249/2304/20202

CTTCTTGAGTTGCTGGGGTGACACTGTTCAGCGTGTGTTGCGGGTGTGTGTTTGTGTGCGTGCTGCGGGTGTC

GT : ?

Without primers = 59 bp

LACM_95619

Pooled:

>LACM95619-GS4_1 **124bp**allele/locus/total=968/4317/61226

CTTCTTGAGTTGCTGGGGTGACCGGTTTGTTGCGTGTACTTACTTGTGACAGTTGATGCGCGGATGTCTAGGTC

GT : 9 ? [SCY3]

Without primers = 79 bp

>LACM95619-GS4_2 **109bp**allele/locus/total=360/4317/61226

CTTCTTGAGTTGCTGGGGTGACTGCACACAGACGTAGTAGGCTGCTATCAGTGTGAGACAGATCCAGGACAAGA

GT : 10 ?

Without primers = 64 bp

Individual Replicates:

Replicate 1:

>LACM95619GS41-1-GS4_1 **120bp**allele/locus/total=399/2924/9152

CTTCTTGAGTTGCTGGGGTGACCGGTTTGTTGCGTGTACTTACTTGTGACAGTTGATGCGCGGATGTCTAGGTC

GT : 9 ?

Without primers = 79 bp

>LACM95619GS41-1-GS4_2 **124bp**allele/locus/total=382/2924/9152

CTTCTTGAGTTGCTGGGGTGACCGGTTTGTTGCGTGTACTTACTTGTGACAGTTGATGCGCGGATGTCTAGGTC

GT : 9 ?

Without primers = 79 bp

Replicate 2:

>LACM95619GS42-2-GS4_1 **109bp**allele/locus/total=689/2833/10532

CTTCTTGAGTTGCTGGGGTGACTGCACACAGACGTAGTAGGCTGCTATCAGTGTGAGACAGATCCAGGACAAGA

GT : 10 ?

Without primers = 64 bp

>LACM95619GS42-2-GS4_2 **105bp**allele/locus/total=614/2833/10532

CTTCTTGAGTTGCTGGGGTGACTGCACACAGACGTAGTAGGCTGCTATCAGTGTGAGACAGATCCAGGACAAGA

GT : 10 ?

Without primers = 64 bp

Replicate 3:

>LACM95619GS43-3-GS4_1 **98bp**allele/locus/total=13/136/6352

CTTCTTGAGTTGCTGGGGTGACAAGCATGCACTATGGCACCTGGTTTTGTTTGTTTATTTGTGTGTGTGTGTGT

GT : 13 ?

Without primers = 53 bp

>LACM95619GS43-3-GS4_2 **87bp**allele/locus/total=9/136/6352

CTTCTTGAGTTGCTGGGGTGACAGGGGCTTCTTTGAAAAGGAGTGATGTGTGTGTGTGTGTGTGTGTGTGTGT

GT : 14 ?

Without primers = 42 bp

Bioinformatically Combined:

>LACM95619GS4-GS4_1 **109bp**allele/locus/total=689/5888/26036

CTTCTTGAGTTGCTGGGGTGACTGCACACAGACGTAGTAGGCTGCTATCAGTGTGAGACAGATCCAGGACAAGA

GT : 10 ?

Without primers = 64 bp

>LACM95619GS4-GS4_2 **105bp**allele/locus/total=614/5888/26036

CTTCTTGAGTTGCTGGGGTGACTGCACACAGACGTAGTAGGCTGCTATCAGTGTGAGACAGATCCAGGACAAGA

GT : 10 ?

Without primers = 64 bp

GLSA 22

Forward primer: CCTGAAAATGATGCATGTGG

Repeat motif: CA

Reverse primer: AGAGTAGGCTGTTCCTTTGAGG

Reverse complement: CCTCAAAGGAACAGCCTACTCT

Nucleotide differences within the sample

>IndividualName_Replicate **BOLDbp** = allele call allele/locus/total (# of reads of specific allele, total reads for specific locus, total reads for the sample)

MVZ_2088

Pooled:

>MVZ2088-GLSA22_1 **179bp**allele/locus/total=1054/2103/118054

CCTGAAAATGATGCATGTGGCTATACTTCCAAAGTCTTACACTTCCTTAACAGAATATGATGAGAGACTGTAGAT

CA : 7 ?

Without primers = 137 bp

Individual Replicates:

Replicate 2:

>MVZ2088GLSA222-2-GLSA22_1 **179bp**allele/locus/total=10/40/24822

CCTGAAAATGATGCATGTGGCTATACTTCCAAAGTCTTACACTTCCTTAACAGAATATGATGAGAGACTGTAGAT

CA : 18

Without primers = 137 bp

>MVZ2088GLSA222-2-GLSA22_2 **177bp**allele/locus/total=6/40/24822

CCTGAAAATGATGCATGTGGCTATACTTCCAAAGTCTTACACTTCCTTAACAGAATATGATGAGAGACTGTAGAT

CA : 17

Without primers = 135 bp

Replicate 3:

>MVZ2088GLSA223-3-GLSA22_1 **179bp**allele/locus/total=1644/3249/7686

CCTGAAAATGATGCATGTGGCTATACTTCCAAAGTCTTACACTTCCTTAACAGAATATGATGAGAGACTGTAGAT

CA : 7 ?

Without primers = 137 bp

Bioinformatically Combined:

>MVZ2088GLSA22-GLSA22_1 **179bp**allele/locus/total=1644/3287/33506

CCTGAAAATGATGCATGTGGCTATACTTCCAAAGTCTTACACTTCCTTAACAGAATATGATGAGAGACTGTAGAT

CA : 7 ?

Without primers = 137 bp

GLSA 52

Forward primer: TCCATCCACAGTGTGTGAGC

Repeat motif: CA

Reverse primer: CCTGGAGTCCACTCAAGCAT

Reverse complement: ATGCTTGAGTGGACTCCAGG

Nucleotide differences within the sample

>IndividualName_Replicate **BOLDbp** = allele call allele/locus/total (# of reads of specific allele, total reads for specific locus, total reads for the sample)

HSU_8180

Pooled:

>HSU8180VM2551-GLSA52_1 **257bp**allele/locus/total=1815/10305/158834

TCCATCCACAGTGTGTGAGCCTGTGCGAGCATGCACACACACACACACACACACACATACACACACACAGAGGACAC

CA : 11 ?

Without primers = 217 bp

>HSU8180VM2551-GLSA52_2 **251bp**allele/locus/total=548/10305/158834

TCCATCCACAGTGTGTGAGCCTGTGCGCGCGTGCACACACACACACACACACACACACACACAGAGGACAGATGTT

CA : 14

Without primers = 211 bp

Individual Replicates:

Replicate 1:

>HSU8180GLSA521-1-GLSA52_1 **257bp**allele/locus/total=332/2572/5956

TCCATCCACAGTGTGTGAGCCTGTGCGAGCATGCACACACACACACACACACACACATACACACACACAGAGGACAC

CA : 11 ?

Without primers = 217 bp

Replicate 2:

>HSU8180GLSA522-2-GLSA52_1 **257bp**allele/locus/total=522/3024/7106

TCCATCCACAGTGTGTGAGCCTGTGCGAGCATGCACACACACACACACACACACACATACACACACACAGAGGACAC

CA : 11 ?

Without primers = 217 bp

Bioinformatically Combined:

>HSU8180GLSA52-GLSA52_1 **257bp**allele/locus/total=872/5594/13062

TCCATCCACAGTGTGTGAGCCTGTGCGAGCATGCACACACACACACACACACACACATACACACACACAGAGGACAC

CA : 11 ?

Without primers = 217 bp

UMMZ_79760

Pooled:

UMMZ-79760-GLSA52_1 **251bp** allele/locus/total=75/760/26736

TCCATCCACAGTGTGTGAGCCTGTGCGCGCGTGCACACACACACACACACACACACACACACACACACAGAGGACAGA

CA : 16

Without primers = 215 bp

UMMZ-79760-GLSA52_2 **250bp** allele/locus/total=53/760/26736

TCCATCCACAGTGTGTGAGCCTGTGCGCGCGTGCACACACACACACACACACACACACACACACACACAGAGGACAGATG

CA : 15

Without primers = 213 bp

Individual Replicates:

Replicate 1:

>UMMZ79760GLSA52_1-GLSA52_1 **255bp**allele/locus/total=84/1596/9730

TCCATCCACAGTGTGTGAGCCTGTGCGCGCGTGCACACACACACACACACACACACACACACACACACAGAGGACAGA

CA : 16

Without primers = 215 bp

Replicate 3:

>UMMZ79760GLSA52_3-GLSA52_1 **255bp**allele/locus/total=289/2346/5632

TCCATCCACAGTGTGTGAGCCTGTGCGCGCGTGCACACACACACACACACACACACACACACACACACAGAGGACAGA

CA : 16

Without primers = 215 bp

>UMMZ79760GLSA52_3-GLSA52_2 **253bp**allele/locus/total=175/2346/5632

TCCATCCACAGTGTGTGAGCCTGTGCGCGCGTGCACACACACACACACACACACACACACACACACAGAGGACAGATG

CA : 15

Without primers = 213 bp

Bioinformatically Combined:

>UMMZ79760GLSA52-GLSA52_1 **255bp**allele/locus/total=473/6292/23088

TCCATCCACAGTGTGTGAGCCTGTGCGCGCGTGCACACACACACACACACACACACACACACACACAGAGGACAGA

CA : 16

Without primers = 215 bp

Appendix 5: Identification of size homoplasy in samples UMMZ 79760 and HSU 1836 at GLSA-52 from the bioinformatically pooled dataset. UMMZ 79760 is truly homozygous, while HSU 1836 is not. There also seems to be different polymorphisms (highlighted in yellow) in all three alleles even though they are all 255 bp fragments.

From the Methods Combined data:

- Samples: UMMZ 79760 and HSU 1836
- Locus: GLSA-52
- Primers

> **UMMZ79760** GLSA52-GLSA52_1 255bp allele/locus/total=473/6292/23088

TCCATCCACAGTGTGTGAGCCTGTGCGCGCGTGCACACACACACACACACACACACACACACACACAGAGGACAGA

>**HSU1836** GLSA52-GLSA52_1 255bp allele/locus/total=948/10535/26164

TCCATCCACAGTGTGTGAGCCTGTGCACGCGTGCACACACACACACACACACACATACACACACACAGAGGACAGAT

>**HSU1836** GLSA52-GLSA52_2 255bp allele/locus/total=571/10535/26164

TCCATCCACAGTGTGTGAGCCTGTGCGCGCGTGCACACACACACACACACACACACACACACACACAGAGGACAGA

Appendix 6: Cost Calculations

Qiagen QiaAmp Mini Kit (Qiagen.com #51304): $170 for 50 samples = $3.40/sample

KAPA Biosystems Illumina Library Prep Kit (Roche.com #KK8301): $971 for 48*4 samples = $5.05/sample

KAPA Beads (#KK8001) $450.00 for 30 mL, ˜200 ul per sample = $3.00/sample

DreamTaq ReadyMix (ThermoFisher Item #FERK1081) $91.93 per 5000 ul = $0.17 per reaction = $2.89 per sample across all PCR replicates

Illumina MiSeq v3 2x300 PE kit (Illumina.com #MS-102-3003) =$1674.00 only used 7% lane here = $117.18 for all samples N=300 = $0.40/sample

Total: $14.74 per sample

NOTE: These costs exclude Illumina individual TruSeq style adapters, all standard lab equipment, qPCR reagents and PCR primers- which were all unlabeled and ordered from IDT (IDTdna.com)

The rest of the Illumina MiSeq lane contained samples for other projects.