Past and present: what museum specimens and detection dogs bring to pangolin conservation

Hyeon Jeong Kim

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Samuel K. Wasser, Chair

Jennifer Ruesink

Scott Kennedy

Program Authorized to Offer Degree:
Biology Department

University of Washington

Abstract

Past and present: what museum specimens and detection dogs bring to pangolin conservation

Hyeon Jeong Kim

Chair of the Supervisory Committee:
Dr. Samuel K. Wasser
Biology Department

Pangolins, a group of scaly anteaters, have recently garnered the attention of both the public and scientific community by the magnitude of trade in their meat and scales. Despite the large volume of legal and illegal trade in pangolins, there is limited information about the impact of this trade on the group of species, especially from a genetics perspective. My dissertation focuses on harnessing the genomic information in museum specimens and trialing detection dog methodology for pangolin sample collection to provide information for pangolin conservation. In Chapter 1, using whole genome sequence data generated from museum samples, historic and modern, as well as publicly available pangolin data, I examine the feasibility of using museum specimens as genetic resource to glean information for conservation genetics. I estimate two metrics of genetic diversity to assess if the demographic decline of pangolins is followed by genetic decline in seven pangolin species. The comparison of runs of homozygosity and heterozygosity within each species show a general decline in genetic diversity over time. In Chapter 2, I identify a putative set of single nucleotide polymorphisms (SNPs) for wildlife forensic applications. By using geo-referenced museum specimens across seven species and applying stringent filtering criteria, a set of 1094 SNPs are identified. The set of SNPs provide decent accuracy in genetic assignment tests, but the results are limited by the small sample size. I outline future steps to validate and select a set of SNPs for forensic application. In Chapter 3, a pilot study on the training and use of detection dogs for the collection of Chinese and

Indian pangolins fecal samples is described. The proof-of-concept study show that detection dogs are able to detect pangolin fecal samples across a range of landscapes and elevation. The chapter details the detection dog methods for pangolin sample collection and identifies solutions to difficulties in acquiring training samples and lack of prior ecological knowledge of pangolin populations.

Table of Contents

List of Figures

List of Tables

1

Genome-wide analysis of pangolins using historic and modern museum specimens

1.1 Abstract

A wide array of species are currently undergoing population decline. Population decline alone is concerning as it poses a risk of extirpation and extinction of species, but it can also reduce genetic diversity and the adaptive potential of species. Using temporal sampling of species pre and post-decline provides an opportunity to study the changes in genetic diversity. Natural history museum collection hold the key to identifying temporal and spatial changes in genetic variation. In this chapter, 62 historic and modern pangolin samples from seven species collected between 1905 and 2017 were sequenced and combined with six publicly available data sets to assess the feasibility of using museum specimens as genetic resources to glean information for conservation genetics and assess levels of genetic diversity. Using genome wide single nucleotide polymorhism (SNP) data, species and population structure was analyzed and genetic diversity metrics were calculated to identify, if any, erosion of genetic diversity. After correcting for potential DNA damage, a general trend of declining genetic diversity was observed over several decades of collection dates. Further investigation using larger sample size with overlapping geographic location is recommended to confirm the trend and timing of genetic diversity loss with levels of harvest The results of this study establish baseline data on genomic diversity of samples collected over a period of 100 years to inform future studies.

1.2 Introduction

Around one million animal and plant species are threatened with extinction due to human activities including changes in land and sea use, direct exploitation of organisms, climate change, pollution and invasive alien species (IPBES, 2019). Trade in wildlife, both legal and illegal, is a global phenomenon impacting species across the phylogenetic spectrum (Lenzen et al., 2012; Scheffers, Oliveira, Lamb, & Edwards, 2019). Unregulated harvest of species can lead to population decline over time, which may result in erosion of genetic diversity through genetic drift and inbreeding (Hedrick & Kalinowski, 2000; Morton, Scheffers, Haugaasen, & Edwards, 2021). As genetic diversity provides the foundation for evolutionary change, loss of diversity can lead to a potential decline in the adaptive potential of the population. Quantifying and monitoring levels of genetic diversity to determine the extent to which declining species are losing genetic variation is a fundamental part of conservation priorities.

The extent of decline in genetic diversity varies among taxa experiencing population loss. Some species exhibit limited or no observable decline in genetic diversity with a decrease in population size while others show reduced diversity and associated deleterious effects (Hedrick & Kalinowski, 2000). This difference in the impact of population decline on genetic diversity is hypothesized to be influenced by life history of the species, with highly mobile and migratory species being better able to maintain genetic diversity compared to species with lower dispersal capabilities (Shi, Chen, Su, Zhang, & Wasser, 2019; Willoughby et al., 2017). The wide variation in genetic diversity among mammals renders it difficult to compare different taxonomic groups to infer an healthy levels of genetic diversity (Robinson et al., 2016).

Temporal sampling, which entails sampling from a population before and after population decline, provides the best opportunity to understand the level of inherent genetic diversity and possible genetic erosion (Díez-del-Molino, Sánchez-Barreiro, Barnes, Gilbert, & Dalén, 2018). Historic specimens in natural history museum collections offer a unique opportunity to obtain a baseline for genetic diversity, especially if the collections include specimens collected prior to the species' population decline (Besnard et al., 2016; Díez-del-Molino et al., 2018). A large portion of historic samples were collected prior to the molecular era and therefore not stored with genetic analysis as a goal, and have hampered genomic studies in the

past. The recent advances in high-throughput sequencing has opened avenues of harnessing the genomic information in these samples. Whole genome shotgun sequencing can sequence the surviving fragments of DNA in the samples and make it possible to examine individual genomes of a species.

Pangolins, a group of scaly anteaters in the family Manidae (Order Pholidota), have recently garnered world-wide scientific attention as the most trafficked mammal in the world. The eight species of pangolins, four species in Africa and four species in Asia, are hypothesized to have undergone significant population decline in the last century due to harvesting for their use for luxury goods, traditional medicine and consumption (Heinrich et al., 2016). Trade records show that tens of thousands of animals were traded annually between 1980 to 2000 (Heinrich et al., 2016). While the exact timing of the decline in pangolin populations is difficult to assess due to their cryptic nature and wide distribution, there are documented concerns for the Chinese pangolin dating back to 1938 stating that the demand of pangolin carcass makes it necessary to protect the populations in south China and that scales from 4000 or 5000 individuals were imported from Java annually (Anonymous, 1938). Analysis of the CITES Trade database, which includes data from 1977 onwards shows that trade in pangolins species is not recent and tens of thousands of animals were traded annually until 2000 (Heinrich et al., 2016). Population decline as a result of this trade has been documented through local ecological knowledge (Nash, Wong, & Turvey, 2016; Newton, Van Thai, Roberton, & Bell, 2008), shifts in trade numbers and routes (UNODC, 2020), and anecdotal evidence, there is a lack of information on which populations are genetically distinct and have undergone the most significant population decline.

In this chapter, a distribution-wide sampling of historic pangolins specimens ranging from 1905 to 1995 and modern samples ranging from 2000 to 2017 from genome banks and of seven species are sequenced and combined with publicly available data to assess the levels of genetic diversity in each species of pangolins. These data are used to establish a baseline for genetic variation of each sampled species, understand the impact of population decline on pangolins, and identify conservation priorities.

1.3 Methods

1.3.1 Sample information

Historic museum samples in the form of dried skin tissue from geo-referenced pangolin specimens were sampled from National Museum Of Natural History (NMNH), Smithsonian Institution, American Museum of Natural History (AMNH), and Natural History Museum (NHM), London, and frozen tissue samples were acquired from the Burke Museum of Natural History and Culture's Genetic Resources Collection and the NMNH, Smithsonian Institution. Pangolin samples from seven species across its distribution were sequencing. This included 10 cryopreserved frozen tissue samples and 46 dried skin samples from historic museum specimens (Figure 1.1 and 1.2). This was combined with six samples from publicly available data to form a dataset that included 17 *Manis. javanica* samples, 17 *M. pentadactyla* samples, four *M. crassicaudata* samples, four *M. gigantea* samples, nine *M. tricuspis* samples, six *M. tetradactyla* samples, and five *M. temminckii* samples collected between 1905 to 2017. Historic samples were collected between 1905 and 1995 and modern samples were sampled between 2000 and 2017. The range of collection date for each species is shown in Figure 1.2.

1.3.2 DNA extraction and sequencing

The historic museum specimen DNA extraction was carried out in two batches: the first batch of samples was extracted in a low template pre-PCR room and the second batch of samples was extracted in a dedicated cleanroom with positive pressure where no pangolin DNA had previously been introduced. Approximately 10 mg of dried tissue was used for DNA extraction using QIAamp DNA Micro Kit (Qiagen) following the manufacturer's protocol for tissue with modifications to increase tissue lysis. Prior to extraction, samples were rinsed twice in ethanol and once in PBS and soaked in PBS for up to 5 min to re-hydrate the tissue to aid in cutting and lysis. The duration of lysis was increased to a minimum of 24 hours and proteinase K was added at regular intervals until the sample was digested. Incubation was set to 56C to minimize risk of denaturing the DNA fragments. For samples with low DNA concentrations, DNA was extracted multiple times and the elutions were pooled and concentrated using Amicon ultra-

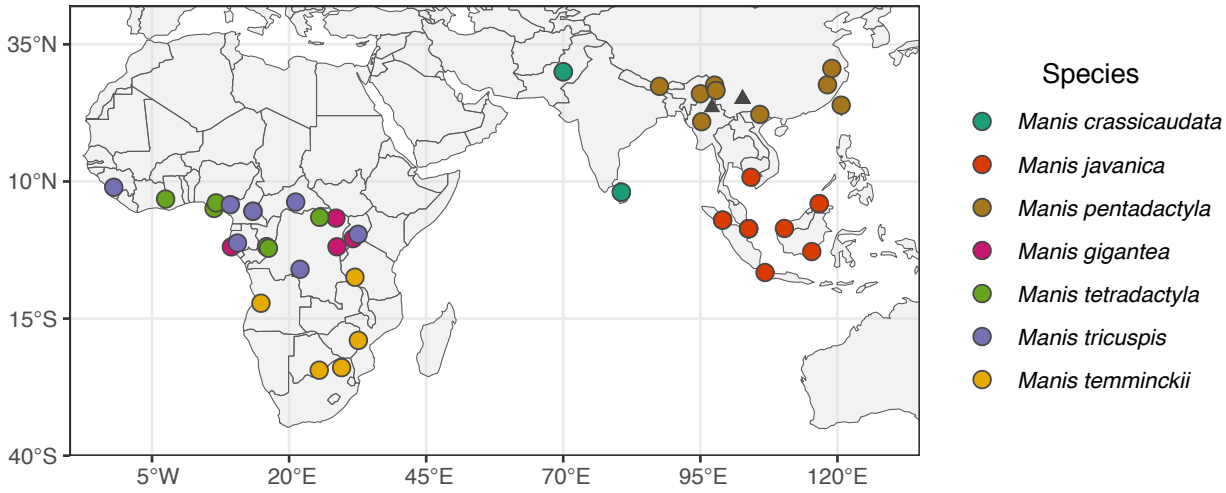

Figure 1.1: Map showing the approximate location of sample collection for specimens with geographic provenance information. Samples sequenced in this study are shown in circles and samples downloaded from public repositories are shown in triangles.

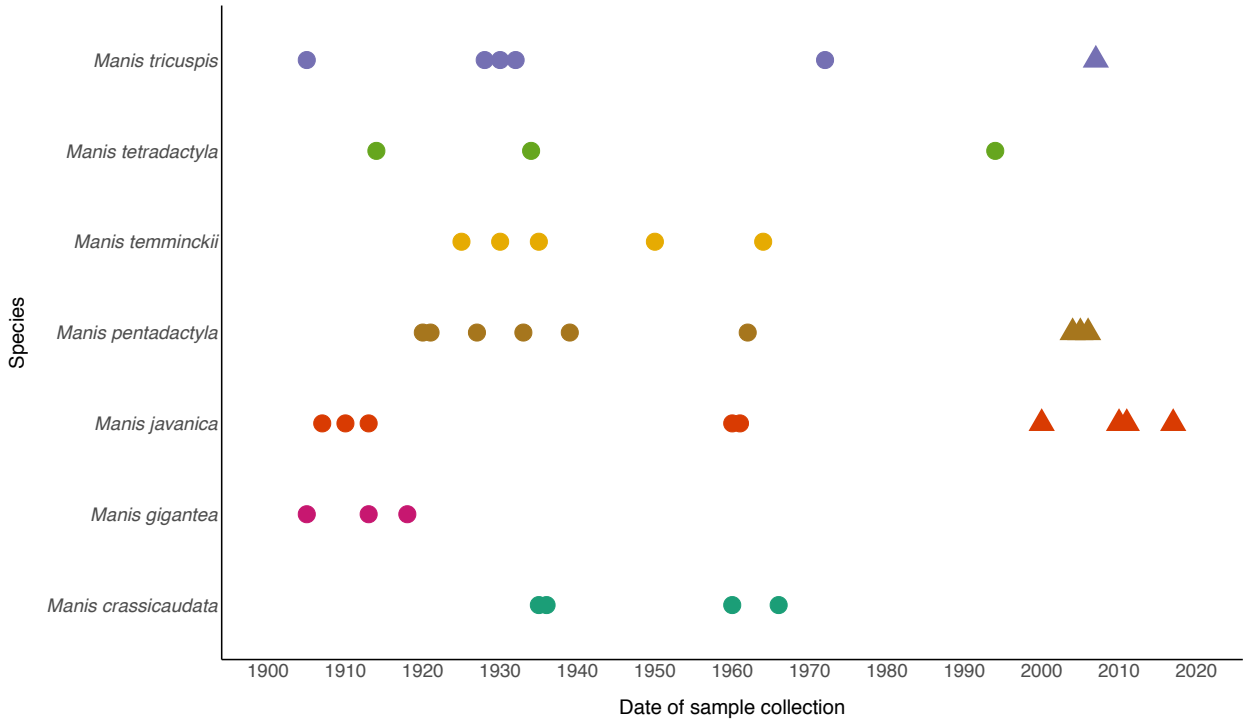

Figure 1.2: The year of sample collection of each species. Samples sequenced in this study are shown in circles and samples downloaded from public repositories are shown in triangles.

0.5 30kDA (Millipore Sigma). DNA quality and quantity were checked using dsDNA High Sensitivity Assay Kit on Qubit Fluorimeter (Thermo Fisher Scientific) and TapeStation Genomic or High Sensitivity DNA ScreenTape analysis (Agilent). A total of 110 samples were extracted and 40 samples with a minimum average fragment size of 80 bp and 600 ng were submitted to Northwest Genomics Center, University of Washington, for whole genome shot-gun sequencing libraries preparation and sequencing. The samples were sequenced in two batches using 100 bp paired-end sequencing on HiSeq 4000 and NovaSeq 6000 platforms. Modern samples (frozen tissue and blood samples) were acquired from the Burke Museum of Natural History and Culture's Genetic Resources Collection and NMNH, Smithsonian Institution. DNA from modern samples was extracted using Qiagen DNeasy Blood and Tissue kit following manufacturer's protocol for each sample type. DNA extracts were sent to the University of Washington's Center for Precision Diagnostics for library preparation and 150 bp paired-end sequencing on two separate HiSeq 4000 platforms.

1.3.3 Published pangolin sequence data

To complete our reference baseline, raw sequencing data (fastq files) of four *M. javanica* samples and one *M. pentadactyla* sample from Hu et al. (2020), one *M. tricuspis* sample from an assembly, and one *M. javanica* sample and one *M. pentadactyla* sample from DNAZoo (Dudchenko et al., 2017) were downloaded from SRA and included in all analyses. These sequencing data were classified as modern samples.

1.3.4 Data analysis

All sequence data, historic and modern, were checked for quality using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and processed using the Paleomix v1.3.2 bam pipeline (Schubert et al., 2014). For sequence data generated in this study, AdapterRemoval v2.3.1 (Schubert, Lindgreen, & Orlando, 2016) was used to trim adapters, filter low quality reads, and remove reads less than 25 bp. For data downloaded from public sources, adapters were trimmed if adapters could be identified. Reads were not collapsed to retain pair-end information and paired-end reads were aligned to a *M. javanica* genome (ManJav 2.0; Hu et al. (2020)) using bwa-mem of BWA version 0.7.17 (Li, 2013).

Post alignment, Picard Tools v2.23 ("Picard toolkit," 2019) was used to mark duplicates and FGBio's Clipbam (https://github.com/fulcrumgenomics/fgbio) was used to eliminate overlap between each paired end read. Where read groups overlapped between samples due to identical index sequences from different sequencing batches, they were renamed using Picard Tools. To ensure that the results of genomic analysis are not biased due to higher coverage in modern sample, modern samples with coverage higher than 15X were downsampled to an average coverage of 15X with DownsampleSam function in PicardTools. The quality of the bam files were assess using Qualimap v2.2.2 (Okonechnikov, Conesa, & García-Alcalde, 2016). Two samples with low coverage and high missing data were excluded from downstream analysis.

Mitochondrial genomes were assembled in two steps. The raw reads were aligned to a whole mitochondrial genome of the same species using BWA version 0.7.17 (Li, 2013) with a mismatch score of 4 to retain all possible reads that may map to the mitochondrial genome. Then, the mapped reads were assembled and annotated using MitoFinder 1.4 (Allio et al., 2020) with the assembler Megahit (Li et al., 2016) and annotator MiTfi (Jühling et al., 2012). The list of mitochondrial genomes used as reference genomes for alignment are shown in Supplementary Table 1.5.

1.3.5 Variant calling and filtering

Variant calling was performed using genotype likelihoods as implemented in ANGSD (Korneliussen, Albrechtsen, & Nielsen, 2014) to accommodate the uncertainties arising from low coverage data from historic specimens. Within ANGSD, SAMtools model (-GL 1) was used to estimate genotype likelihoods with SAMtools default BAQ computation (-baq 1; Li (2011)) and the following filters and parameters were applied: uniquely mapping reads (-uniqueOnly 1), remove anomalous reads (-remove_bads 1), minimum mapping quality (-minMapQ 20), minimum base quality (-minQ 20), and a minimum number of individuals (-minInd 42). SNPs were called using a uniform prior to calculate the posterior genotype probability (-doPost 2), major and minor alleles were inferred from genotype likelihoods (-doMajorMinor 1) and sites that had a minimum minor allele frequency of 0.05 (-minMaf 0.05). The following criteria were used to filter SNPs: -snp_pval 1e-6), -sb_pval 1e5 and -hetbias_pval 1e05. To calculate individual specific

metrics of diversity, SNPs were called on samples from each identified species separately to reduce bias in the SNP calls. Variants were called using the same filters, but with a minimum number of individuals set to one less than the total number of samples.

1.3.6 Genomic analysis

To identify the number of clusters of pangolins in the seven known pangolin species, a principal component analysis (PCA) was conducted using PCAngsd and clustering analysis was performed using ADMIXTURE using the full SNP dataset with all 62 samples.

Genetic diversity change were estimated by calculating two metrics of genetic diversity. Heterozygosity was calculated from SNP data using VCFtools for each species separately. Runs of homozygosity (ROH) was identified using PLINK. ROH was calculated following parameters described previously (Ceballos, Hazelhurst, & Ramsay, 2018), with `--homozyg-kb` of 50, `--homozyg-kb` of 300, `--homozyg-window-snp` of 20, and `--homozyg-window-het` of 5. Historic samples are prone to sequencing errors from DNA damage resulting in higher rates of heterozygosity and influencing genetic diversity calculations (Axelsson, Willerslev, Gilbert, & Nielsen, 2008). To be conservative all analyses were conducted with and without transitions in the SNP dataset.

1.4 Results

DNA was extracted from 110 museum samples of eight species of pangolins. A total of 46 samples from seven species were selected for sequencing based on DNA quantity by total ng of DNA and quality as assessed by fragment size (Supplementary table 1.4). There is variation in average DNA concentration among species but also within each species (Supplementary Figure 1.10).

Mitochondrial genomes were assembled for 44 museum specimens and nine modern samples that were sequenced. The assembled mitochondrial COI gene were blasted separately to confirm the species of the samples. The genomic clustering using identity by state matrix of the nuclear data also resulted in congruent results to the mitochondrial DNA species identification results (Figure 1.3). Two samples were

genetically identified with both mitochondrial and nuclear DNA to be a different species than the museum labels. These samples were re-assigned to the genetically confirmed species: *M. longicaudata* (*M. crassicaudata*) to *M pentadactyla*, and *M. tetradactyla* to *M tricuspis*. Two samples of unknown species were identified as *M pentadactyla*.

The 46 historic specimens were sequenced at an average coverage of 10X (0.591X to 14.9X) with an average insert size of 114 bp (Supplementary Table 1.3).The samples were mapped to the *M. javanica* pangolin genome (ManJav2.0) and resulted in an average mapping rate of 93.5% (Asian pangolin average mapping rate = 95.9%; African pangolin average mapping rate = 91.2%). The 10 modern samples were sequenced at an average coverage of 50X with an average insert size of 349 bp. These samples were downsampled to an average coverage of 14X (11.4-14.8X). Two historic samples, PK001 (*M. crassicaudata*) and MZ001 (*M. temminckii*), with a coverage of 0.59X and 2.2X were excluded from further analysis due to low coverage. The Palawan pangolin, *M. culionensis*, sample did not yield sufficient DNA for sequencing, but due to their limited geographic distribution of the Palawan island, its absence did not pose significant gaps in analysis. Variant calling resulted in a total of 107,501,585 SNPs across the seven species and 15,842,342, 13,860,151, 19,399,314, 39,111,061, 23,115,202, 28,138,458, 23,942,596 SNPs were identified for *M. javanica*, *M. pentadactyla*, *M. crassicaudata*, *M. tricuspis*, *M. tetradactyla*, *M. gigantea*, and *M. temminckii*, respectively. As multiple species were mapped to a single species genome, a large number of SNPs were identified and in general an increasing number of SNPs were identified with increasing divergence from *M. javanica*.

Table 1.1: Information on historic and modern samples sequences in this study.

Sample ID	Sequencing ID	Sequencing year	Platform	Museum	Museum ID	Species	Sample collection date
PK001	336747	2019	HiSeq 4000	AMNH	185599	*Manis crassicaudata*	1960
NP001	336742	2019	HiSeq 4000	NHM London	63.36	*Manis pentadactyla*	1962
NG004	336743	2019	HiSeq 4000	NHM London	1995.25	*Manis tetradactyla*	1994
KH001	336744	2019	HiSeq 4000	Smithsonian	321551	*Manis javanica*	1961
ZW001	336745	2019	HiSeq 4000	Smithsonian	368617	*Manis temminckii*	1964
GA004	336746	2019	HiSeq 4000	AMNH	119964	*Manis tetradactyla*	NA
AO002	442962	2020	NovaSeq S4	AMNH	80817	*Manis temminckii*	1925
BW001	442963	2020	NovaSeq S4	AMNH	168954	*Manis temminckii*	1950
CD001	442964	2020	NovaSeq S4	Smithsonian	101448	*Manis tricuspis*	NA
CD004	442965	2020	NovaSeq S4	NHM London	12.12.3.2	*Manis tetradactyla*	NA
CD005	442966	2020	NovaSeq S4	AMNH	53845	*Manis gigantea*	1913
CD019	442967	2020	NovaSeq S4	AMNH	53862	*Manis tetradactyla*	1914
CF001	442968	2020	NovaSeq S4	AMNH	88437	*Manis tricuspis*	1930
CG001	442969	2020	NovaSeq S4	NHM London	61.1096	*Manis gigantea*	1913
CM002	442970	2020	NovaSeq S4	NHM London	48.1325	*Manis tricuspis*	1932
CM004	442971	2020	NovaSeq S4	AMNH	241130	*Manis tricuspis*	1972
CM005	442972	2020	NovaSeq S4	AMNH	241131	*Manis tricuspis*	1972
CN012	442973	2020	NovaSeq S4	AMNH	47852	*Manis pentadactyla*	1920
CN015	442974	2020	NovaSeq S4	AMNH	57059	*Manis pentadactyla*	1921
GA003	442975	2020	NovaSeq S4	Smithsonian	220404	*Manis gigantea*	1918
GA005	442976	2020	NovaSeq S4	AMNH	119963	*Manis tricuspis*	NA
GH006	442977	2020	NovaSeq S4	NHM London	35.10.22.157	*Manis tetradactyla*	1934
ID010	442978	2020	NovaSeq S4	Smithsonian	198319	*Manis javanica*	1913

ID013	442979	2020	NovaSeq S4	Smithsonian	267388	*Manis javanica*	NA
ID014	442980	2020	NovaSeq S4	Smithsonian	267389	*Manis javanica*	NA
ID017	442981	2020	NovaSeq S4	NHM London	9.1.5.857	*Manis javanica*	1907
IN001	442982	2020	NovaSeq S4	NHM London	21.8.2.27	*Manis pentadactyla*	1920
LK001	442983	2020	NovaSeq S4	NHM London	76.138	*Manis crassicaudata*	1935
LK002	442984	2020	NovaSeq S4	NHM London	1950.29	*Manis crassicaudata*	1936
MM002	442985	2020	NovaSeq S4	NHM London	34.3.4.1	*Manis pentadactyla*	1933
MM003	442986	2020	NovaSeq S4	NHM London	50.752	*Manis pentadactyla*	1939
MM005	442987	2020	NovaSeq S4	NHM London	14.7.19.238	*Manis pentadactyla*	NA
MY002	442988	2020	NovaSeq S4	Smithsonian	317198	*Manis javanica*	1960
MY003	442989	2020	NovaSeq S4	Smithsonian	317199	*Manis javanica*	1960
MY007	442990	2020	NovaSeq S4	NHM London	32637	*Manis javanica*	1910
MZ001	442991	2020	NovaSeq S4	AMNH	42349	*Manis temminckii*	NA
NG005	442992	2020	NovaSeq S4	NHM London	1998.302	*Manis tetradactyla*	NA
PK002	442993	2020	NovaSeq S4	AMNH	244406	*Manis crassicaudata*	1966
PK003	442994	2020	NovaSeq S4	AMNH	244407	*Manis crassicaudata*	1966
SL001	442995	2020	NovaSeq S4	NHM London	46.888	*Manis tricuspis*	1905
TW007	442996	2020	NovaSeq S4	AMNH	183148	*Manis pentadactyla*	NA
TZ001	442997	2020	NovaSeq S4	AMNH	83772	*Manis temminckii*	1935
UG001	442998	2020	NovaSeq S4	NHM London	61.1095	*Manis gigantea*	1905
UG002	442999	2020	NovaSeq S4	NHM London	28.9.8.37	*Manis tricuspis*	1928
VN006	443000	2020	NovaSeq S4	NHM London	27.12.1.203	*Manis pentadactyla*	1927
ZA001	443001	2020	NovaSeq S4	AMNH	83609	*Manis temminckii*	1930
EEM957	10594	2017	HiSeq 4000	Burke	EEM957	*Manis javanica*	2010
CSW8292	10595	2017	HiSeq 4000	Burke	CSW8292	*Manis javanica*	2010
CSW8308	10596	2017	HiSeq 4000	Burke	CSW8308	*Manis javanica*	NA

CSW8288	10600	2017	HiSeq 4000	Burke	CSW8288	*Manis javanica*	2011
599832	10597	2017	HiSeq 4000	Smithsonian	599832	*Manis pentadactyla*	2006
599833	10598	2017	HiSeq 4000	Smithsonian	599833	*Manis pentadactyla*	2006
599835	10599	2017	HiSeq 4000	Smithsonian	599835	*Manis pentadactyla*	2005
587996	10601	2017	HiSeq 4000	Smithsonian	587996	*Manis pentadactyla*	2005
587997	10602	2017	HiSeq 4000	Smithsonian	587997	*Manis pentadactyla*	2004
599829	10603	2017	HiSeq 4000	Smithsonian	599829	*Manis pentadactyla*	2006

Species and population structure

The samples from each of the seven sequenced pangolin species clustered by continent and by species in the PCA and admixture analysis (Figure 1.4 and Figure 1.5). Each species showed distinct clusters, with the two ground pangolins species (*M. temminckii* and *M. gigantea*) clustering more closely and the two tree pangolins (*M. tricuspis* and *M. tetradactyla*) clustering closely together (Figure 1.4).

Admixture analysis of the four African pangolin species and the three Asian pangolin species separately revealed within species structure Figure 1.6 and Figure 1.7. When the samples were analyzed by species, the Chinese pangolin, *M. pentadactyla*, showed two clusters with one cluster consisting of samples from China, Viet Nam, and Taiwan, and the other cluster consisting of samples from Nepal and Myanmar. The Sunda pangolin, *M. javanica*, also grouped into two separate clusters with samples in the northern and southern range clustering separately. The northern cluster of Sunda pangolins consist entirely of samples included from Hu et al. (2020), but are also the only samples in the dataset from that geographic area. White-bellied pangolins, *M. tricuspis* also showed two distinct clusters with one clade consisting of samples from Sierra Leone and Cameroon, West Africa, and the other cluster including samples from Gabon, Democratic Republic of Congo, Central African Republic, and Uganda, Central Africa (Figure 1.6).

Genetic diversity estimates between historic and modern samples.

Heterozygosity (He) was calculated for each species using a dataset with and without transition SNPs. There were minimal difference between the two dataset despite a reduced number of SNPs. The results shown are those calculated from the SNP data without transitions. Average heterozygosity of each species ranged from 0.158 in *M. pentadactyla* to 0.435 in *M. gigantea*; the average He was 0.123 for modern and 0.272 for historic samples (Figure 1.8). Historic and modern *M javanica* samples have an average He of 0.203 and 0.133, respectively and historic and modern *M. pentadactyla* samples show an average He of 0.199 to 0.112. It is not possible to draw conclusions from these values because a large portion of the modern *M. pentadactyla* samples may be from Taiwan and show lower He values due to being an island

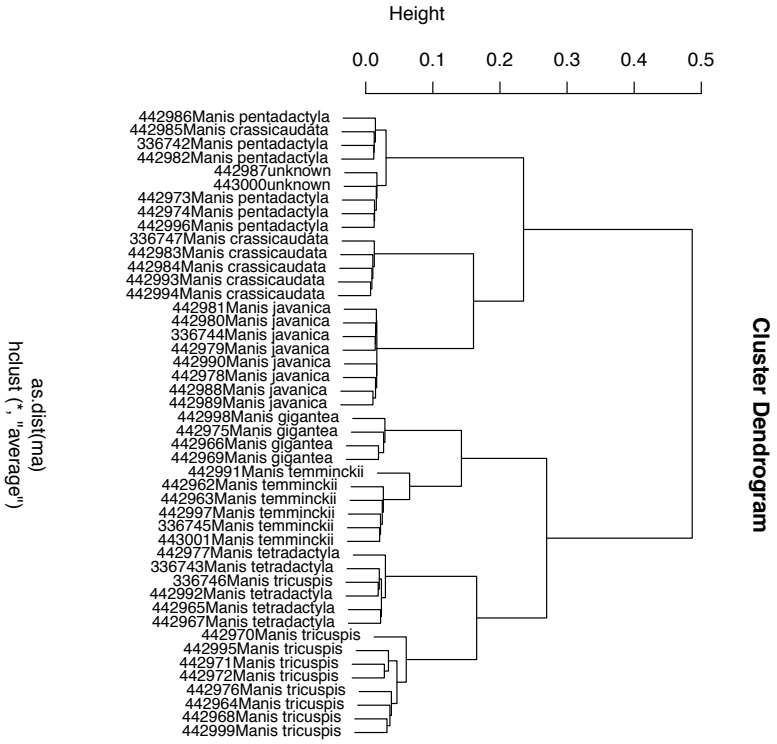

Figure 1.3: Dendrogram based on IBS matrix of SNP data.

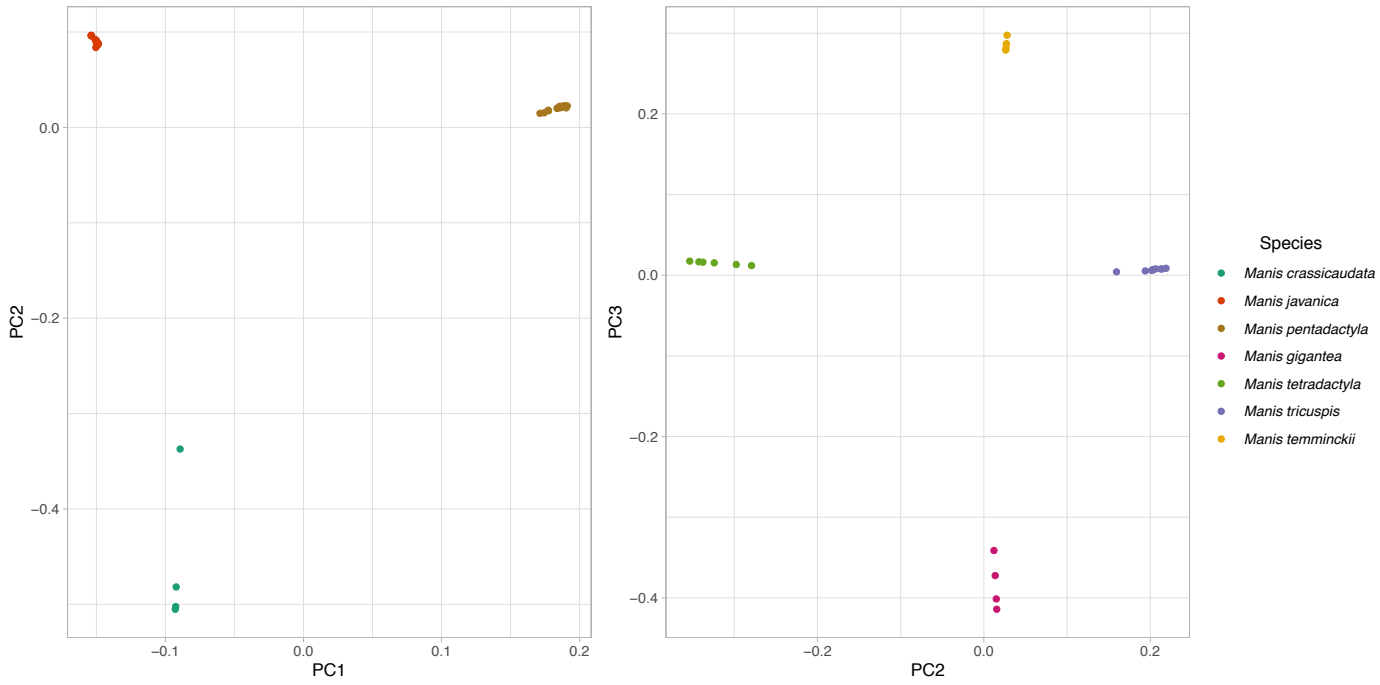

Figure 1.4: PCA plot of 62 samples.

population. The one historic *M. pentadactyla* sample in the dataset from Taiwan does not have a collection date, but has a He of 0.12 while other historic samples from China and Viet Nam have values of around 0.2. A total of 12,114 runs of homozygosity (ROH) was identified in 60 samples. The average total length of ROH was 956,692 in historic and 1,915,940 in modern samples. When compared within species, historic *M. javanica* sample had an average total length of 1,201,423 kb while modern samples had an average total length of 1,663,395 kb. Similar trends in increase in ROH total length was seen in *M pentadactyla* and *M. tricuspis* while there were no modern samples to compare with for the remaining species (Table 1.2). There is a correlation between the number of ROH in a genome and the total length of ROH, however, a number of sample have a longer total length of ROH given the number of ROH (Figure 1.9). There is also an increase in the average length of ROH over time with the most recently collected samples showing the longer ROH (Figure 1.9).

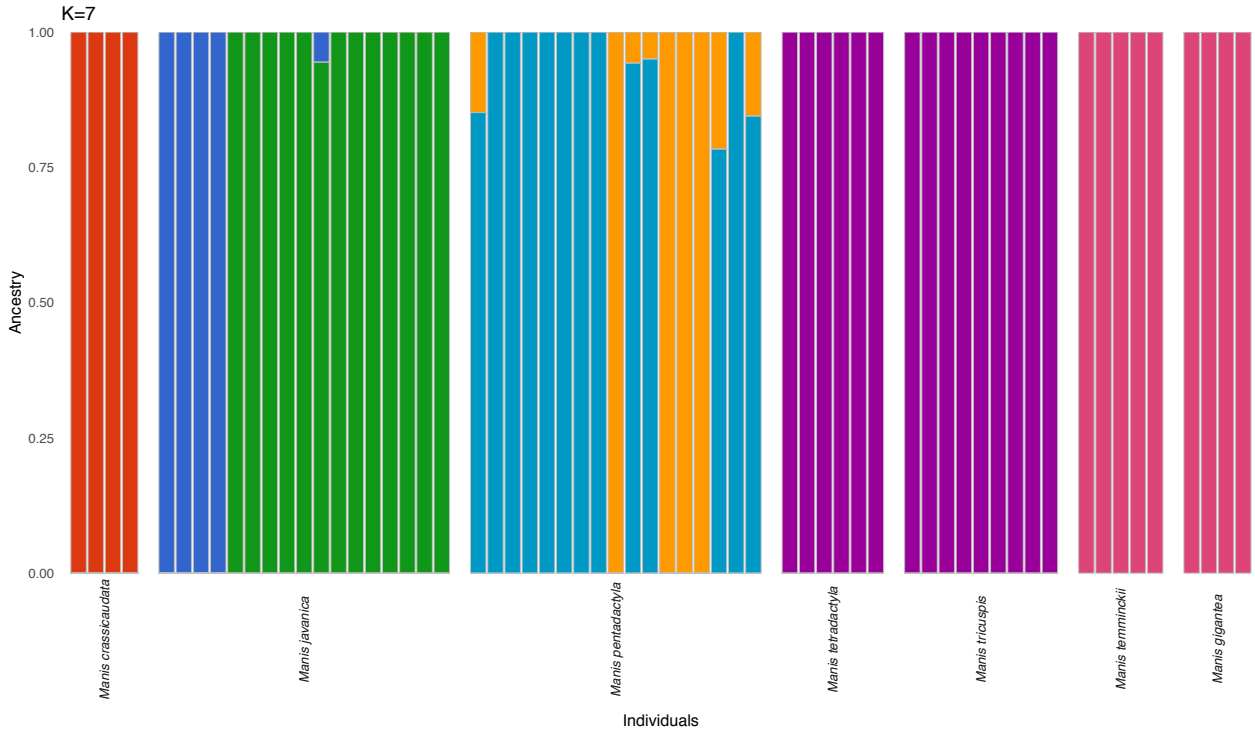

Figure 1.5: Admixture plot.

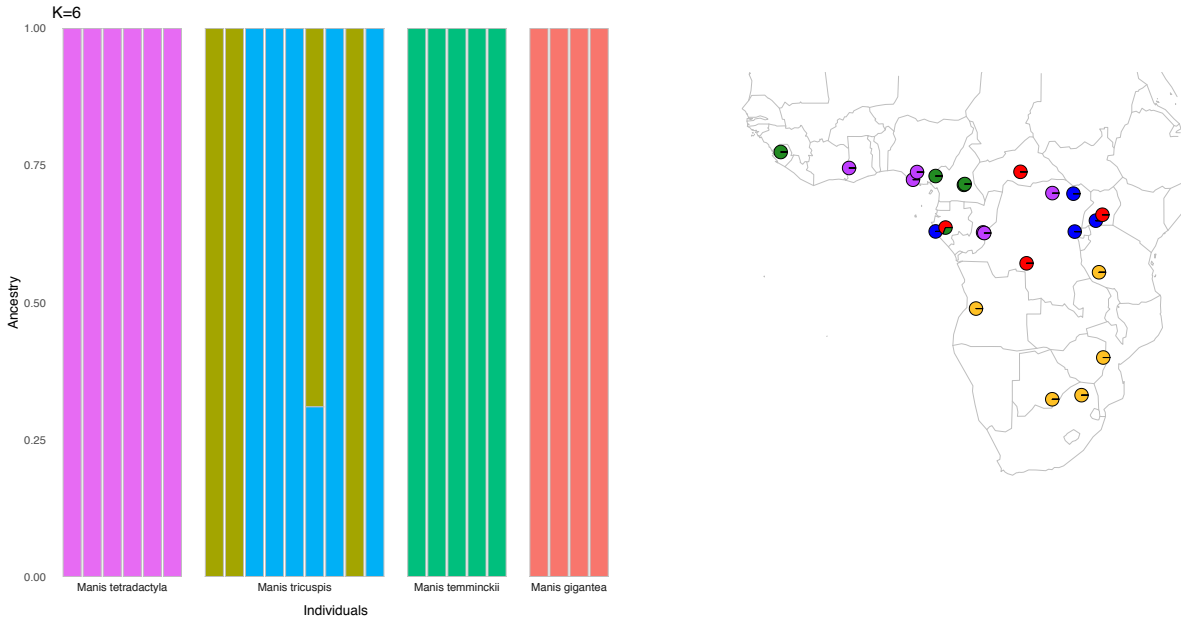

Figure 1.6: Admixture plot and ancestry proportion of the African pangolin samples.

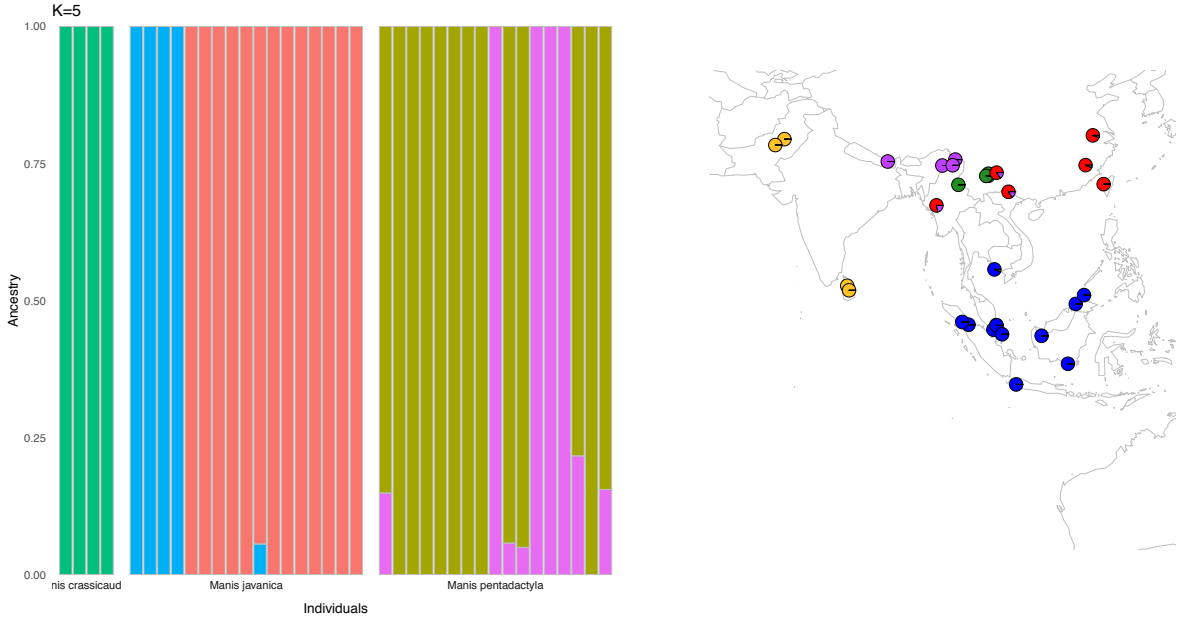

Figure 1.7: Admixture plot and ancestry proportion of the Asian pangolin samples..

Table 1.2: Mean total length of ROH by species and type of sample.

Species	Sample type	Mean length (bp)
Manis crassicaudata	historic	965835.4
Manis gigantea	historic	113738.3
Manis javanica	historic	1201422.6
Manis javanica	modern	1663394.9
Manis pentadactyla	historic	1769085.2
Manis pentadactyla	modern	2161708.6
Manis temminckii	historic	148414.1
Manis tetradactyla	historic	386085.8
Manis tricuspis	historic	1043824.1
Manis tricuspis	modern	2222695.4

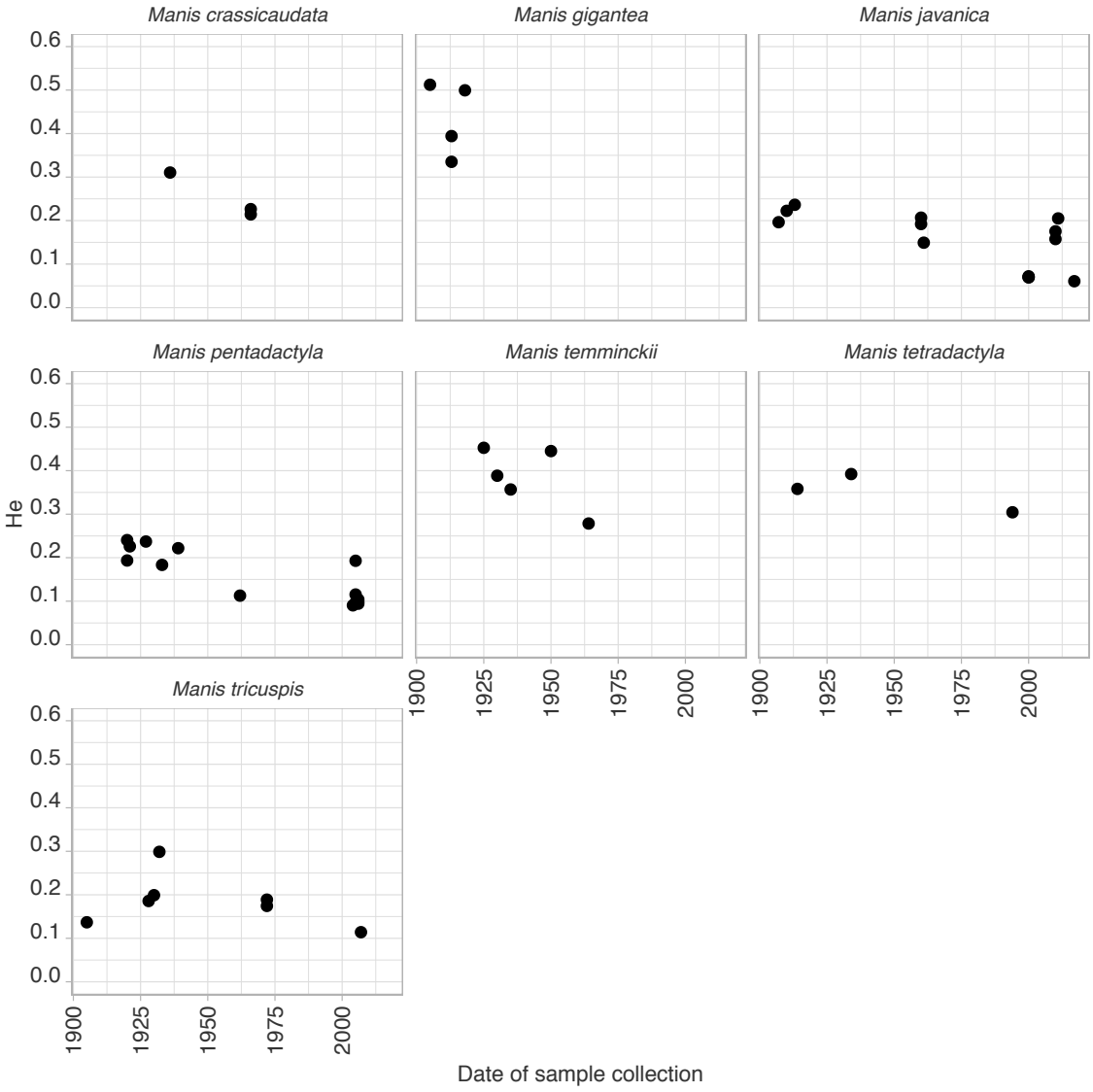

Figure 1.8: Heterozygosity over time by species.

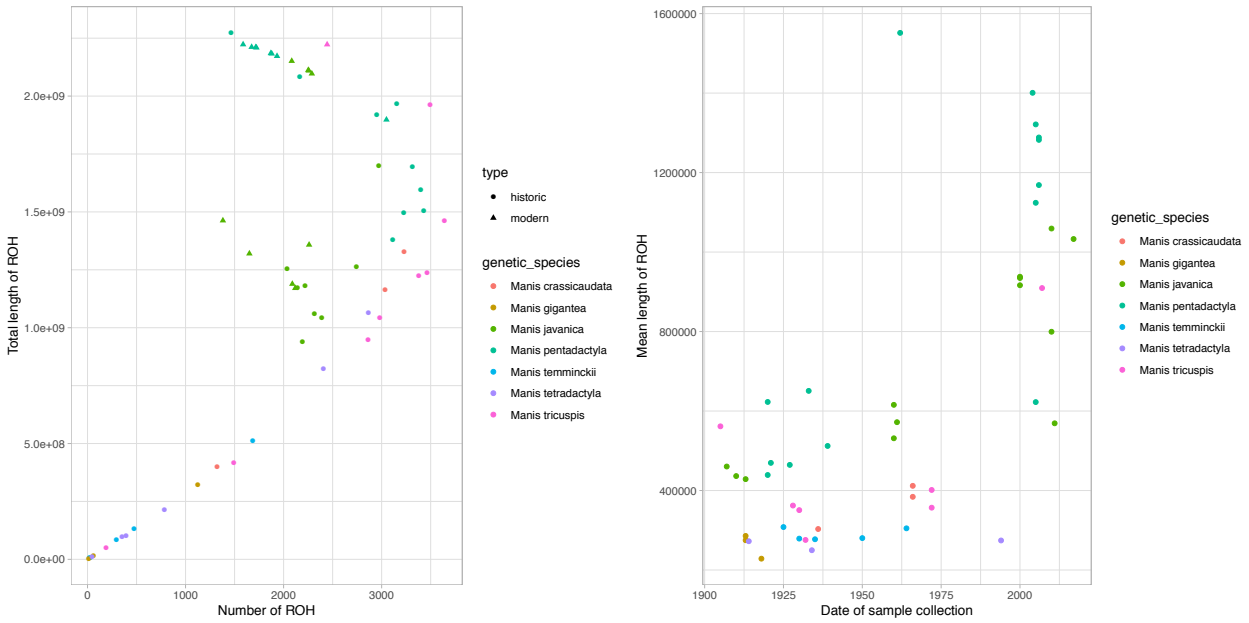

Figure 1.9: Sum and total number of ROH by species, type and sample collection date.

1.5 Discussion

Genetic diversity of Manidae species over temporal and spatial scales are assessed in this chapter. Whole genome data of seven species across two continents and over a century reveal varying levels of genetic diversity in time, space, and species. Eight species of pangolins are currently documented with four species in Asia and four in Africa. This chapter sequenced and analyzed seven of the eight species and revealed clustering of the seven known species.

The population structure within Sunda and Chinese pangolins identified in this chapter was also observed in a population genomic study of *M. javanica* and *M. pentadactyla* (Hu et al., 2020). Four Sunda pangolin samples from the Hu et al. (2020) study were included in this chapter. These samples were labeled as seized in either Myanmar or Yunnan, China, and marked as known origin samples. There is debate over whether Sunda pangolins are or have been distributed in China, but there is anecdotal evidence of Sunda pangolin occurrence in southern China. Complete mitochondrial genome from an animal caught in GuangXi province and identified to be a Chinese pangolin by the authors was later re-identified to

be a Sunda pangolin based on sequence comparison (Qin et al. 2012; Gaubert & Agostinho 2015; Hassanin et al. 2015; Luczon et al. 2016). Two other individuals found in Yunnan and Guizhou province were identified as Chinese pangolins (Feng & Shi 1983, Quan & Wang 1984), but reported to have the same number of chromosomes as *M. javanica*, 2n = 38 and possibly misidentified. It is possible that the range of Sunda pangolins extend into southern China but are at very low densities due to population declines. This warrants further investigation. The three genetic lineages of Sunda pangolins within the islands of Southeast Asia among Borneo, Java, and Sumatra/Singapore observed in Nash et al. 2018 were not seen in this data This is most likely due to combination of the small sample size for each island and the larger genetic difference between the island populations and the continental population masking any subdivisions.

Based on mitochondrial DNA analysis, *M. tricuspis* has previously been reported to show cryptic structuring within the species with 6 separate lineages spread out across western and central Africa (Gaubert et al., 2016). Only two clusters were identified for this species, but again there is a small sample size with insufficient number to form populations. Additional sampling of *M. tricuspis* is necessary to elucidate whether nuclear data reveals the same population structure.

The genetic diversity metrics were calculated without C to T and G to A transitions to avoid inflating genetic diversity in historic samples from DNA damage error. There is a decline in heterozygosity in species for which both historic and modern samples are available. However, lack of geographic overlap among the two types of samples limits interpretation of the general decline. Choo et al. (2016) reported high levels of heterozygosity in Sunda pangolins in comparison to Chinese pangolins, most likely due to the Chinese pangolin sample originating from Taiwan. Conversely, Hu et al. (2020) found higher levels of He in Chinese pangolins in comparison to Sunda pangolins. (Sun et al., 2020) report low levels of genetic diversity. There is consensus that the population of *M. pentadactyla* in Taiwan exhibit lowest levels of He (Hu et al, 2020, Sun et al., 2020). Similarly patterns were found in the length of ROH in each individual genome with modern samples showing higher than average total length of ROH given the total number of ROH. Several other studies have reported numbers of ROH in pangolin genomes, but the different parameters used to identify ROH renders comparison difficult (Brüniche-Olsen, Kellner, Anderson, &

DeWoody, 2018). The small sample size makes it difficult to draw conclusions from the changes in genetic diversity over space and time, a general trend of lower genetic diversity is seen in the three species where modern samples are available.

1.6 Acknowledgements

This work was supported by the USAID Wildlife Crime Tech Challenge. Additionally, the work was supported by University of Washington Biology Department's WRF-Hall International Fellowship and the Hall International Fellowship. The historic and modern pangolin samples used in this study were kindly provided by the American Museum of Natural History, Burke Museum of Natural History and Culture, Natural History Museum, London, and National Museum Of Natural History, Smithsonian Institution. We thank George Amato and Eleanor Hoeger at AMNH, Nicole Edmison, Piper Mullins and Darrin Lunde at NMNH, Louise Tomsett at NHM London, and Sharlene Santana, Sharon Birks and Jeffrey Bradley at Burke Museum for their assistance in acquiring the samples. A special thank you to the Eisenberg Lab for the use of the cleanroom and to the Leache and Kennedy Labs for generously sharing their lab space and equipment for genomic analysis. Unpublished genome assemblies and sequencing data for *M. javanica* and *M. pentadactyla* are used with permission from the DNA Zoo Consortium (dnazoo.org). This work was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system and funded by the STF at the University of Washington.

1.7 Data availability

Code used to conduct the analysis and information to access the sequence data are available on Github (https://github.com/kimhii/dissertation).

1.8 Supplementary materials

Table 1.3: Details of the sequencing results.

Sample ID	Sequencing ID	Sequence year	Platform	Species	Date	Total read	Average GC	Insert size	Median coverage	Percent aligned
PK001	336747	2019	HiSeq 4000	*Manis crassicaudata*	1960	281283572	46.14	60	90.50	0.59
NP001	336742	2019	HiSeq 4000	*Manis pentadactyla*	1962	211511418	43.36	157	96.78	5.41
NG004	336743	2019	HiSeq 4000	*Manis tetradactyla*	1994	534420220	38.45	91	93.64	7.77
KH001	336744	2019	HiSeq 4000	*Manis javanica*	1961	524401354	39.38	163	93.48	14.07
ZW001	336745	2019	HiSeq 4000	*Manis temminckii*	1964	614752992	41.77	159	84.49	12.44
GA004	336746	2019	HiSeq 4000	*Manis tetradactyla*	NA	609927758	40.93	164	89.22	13.86
AO002	442962	2020	NovaSeq S4	*Manis temminckii*	1925	704517646	41.19	82	89.44	8.23
BW001	442963	2020	NovaSeq S4	*Manis temminckii*	1950	756782004	42.09	88	87.47	9.19
CD001	442964	2020	NovaSeq S4	*Manis tricuspis*	NA	685216350	41.91	79	91.15	8.79
CD004	442965	2020	NovaSeq S4	*Manis tetradactyla*	NA	614555660	38.90	111	93.52	11.35
CD005	442966	2020	NovaSeq S4	*Manis gigantea*	1913	642791340	41.43	95	93.30	9.65
CD019	442967	2020	NovaSeq S4	*Manis tetradactyla*	1914	549977618	40.42	111	92.51	9.98
CF001	442968	2020	NovaSeq S4	*Manis tricuspis*	1930	591912342	41.74	94	89.38	9.02
CG001	442969	2020	NovaSeq S4	*Manis gigantea*	1913	639231878	40.42	127	92.57	12.88
CM002	442970	2020	NovaSeq S4	*Manis tricuspis*	1932	760598556	42.53	47	88.51	4.52
CM004	442971	2020	NovaSeq S4	*Manis tricuspis*	1972	355441456	41.21	117	94.26	6.58
CM005	442972	2020	NovaSeq S4	*Manis tricuspis*	1972	475655694	41.29	105	95.00	8.02
CN012	442973	2020	NovaSeq S4	*Manis pentadactyla*	1920	557763800	41.73	108	98.07	10.81
CN015	442974	2020	NovaSeq S4	*Manis pentadactyla*	1921	6488889384	41.37	125	97.79	14.07
GA003	442975	2020	NovaSeq S4	*Manis gigantea*	1918	615164830	44.44	69	94.03	6.01
GA005	442976	2020	NovaSeq S4	*Manis tricuspis*	NA	637088096	41.11	134	92.71	13.67
GH006	442977	2020	NovaSeq S4	*Manis tetradactyla*	1934	637436626	41.42	69	89.22	6.85
ID010	442978	2020	NovaSeq S4	*Manis javanica*	1913	603216030	42.20	118	97.52	13.19
ID013	442979	2020	NovaSeq S4	*Manis javanica*	NA	606277688	41.55	133	98.35	14.87
ID014	442980	2020	NovaSeq S4	*Manis javanica*	NA	483263524	41.36	134	98.06	12.04
ID017	442981	2020	NovaSeq S4	*Manis javanica*	1907	569348614	40.20	117	98.90	12.72
IN001	442982	2020	NovaSeq S4	*Manis pentadactyla*	1920	654038674	42.62	124	97.85	13.64
LK001	442983	2020	NovaSeq S4	*Manis crassicaudata*	1935	485476834	41.34	64	60.37	4.27
LK002	442984	2020	NovaSeq S4	*Manis crassicaudata*	1936	551140260	39.68	111	94.02	10.75

MM002	442985	2020	NovaSeq S4	*Manis pentadactyla*	1933	437924982	40.57	122	97.56	9.21
MM003	442986	2020	NovaSeq S4	*Manis pentadactyla*	1939	622539156	38.10	85	97.96	8.20
MM005	442987	2020	NovaSeq S4	*Manis pentadactyla*	NA	543940918	41.21	111	97.88	10.71
MY002	442988	2020	NovaSeq S4	*Manis javanica*	1960	364326188	42.39	196	97.39	11.05
MY003	442989	2020	NovaSeq S4	*Manis javanica*	1960	392562676	42.88	229	96.94	12.38
MY007	442990	2020	NovaSeq S4	*Manis javanica*	1910	457585712	41.88	112	98.92	9.95
MZ001	442991	2020	NovaSeq S4	*Manis temminckii*	NA	635347604	46.06	27	59.81	2.16
NG005	442992	2020	NovaSeq S4	*Manis tetradactyla*	NA	521873340	40.74	147	88.36	11.07
PK002	442993	2020	NovaSeq S4	*Manis crassicaudata*	1966	394304688	41.64	180	97.69	11.53
PK003	442994	2020	NovaSeq S4	*Manis crassicaudata*	1966	545633448	40.98	133	98.36	13.30
SL001	442995	2020	NovaSeq S4	*Manis tricuspis*	1905	628582438	39.94	95	95.78	9.28
TW007	442996	2020	NovaSeq S4	*Manis pentadactyla*	NA	574220724	42.40	156	97.23	14.70
TZ001	442997	2020	NovaSeq S4	*Manis temminckii*	1935	408792188	40.65	118	88.55	7.24
UG001	442998	2020	NovaSeq S4	*Manis gigantea*	1905	726169652	41.94	71	88.60	7.57
UG002	442999	2020	NovaSeq S4	*Manis tricuspis*	1928	539085512	42.36	94	91.75	8.45
VN006	443000	2020	NovaSeq S4	*Manis pentadactyla*	1927	564031648	41.18	108	97.64	10.56
ZA001	443001	2020	NovaSeq S4	*Manis temminckii*	1930	586531122	41.61	113	89.94	10.03
EEM957	10594	2017	HiSeq 4000	*Manis javanica*	2010	10931239O6	43.75	366	95.89	14.63
CSW8292	10595	2017	HiSeq 4000	*Manis javanica*	2010	218648140 8	41.59	397	96.38	14.70
CSW8308	10596	2017	HiSeq 4000	*Manis javanica*	NA	13322555 36	41.50	383	96.20	14.43
CSW8288	10600	2017	HiSeq 4000	*Manis javanica*	2011	267181496	41.75	391	95.51	11.44
599832	10597	2017	HiSeq 4000	*Manis pentadactyla*	2006	20114111982	41.50	389	95.68	14.47
599833	10598	2017	HiSeq 4000	*Manis pentadactyla*	2006	336605440	41.06	382	95.84	14.00
599835	10599	2017	HiSeq 4000	*Manis pentadactyla*	2005	539014570	41.00	345	96.26	14.32
587996	10601	2017	HiSeq 4000	*Manis pentadactyla*	2005	10292181 82	41.00	353	96.02	14.24
587997	10602	2017	HiSeq 4000	*Manis pentadactyla*	2004	20467275 26	41.47	361	96.32	14.83
599829	10603	2017	HiSeq 4000	*Manis pentadactyla*	2006	819562860	41.00	342	96.32	14.46

1.8.1 Supplementary materials and methods

A spin-column-based DNA extraction method was combined with filter concentration method to reduce the possibility of denaturing the DNA fragments through ethanol precipitation (Svaren, Inagami, Lovegren, & Chalkley, 1987) or high temperature (Schaeffer, Kolb, & Buc, 1982). Phenol-chloroform method has been shown to increase the yield of DNA, but was not used due to the toxicity of the reagents and the possibility of DNA oxidation (Finnegan, Herbert, Evans, Griffiths, & Lunec, 1996).

About 10 mg of dried skin tissue was excised from the 110 collected dried skin sample using sterile blades. The sample was washed twice using 70-80% ethanol and then air dried before washing in 1X phosphate buffered saline (PBS). The samples were then cut into small pieces using a sterile scalpel blade and placed in 1.5 mL microcentrifuge tubes. The manufacturer's protocol were followed for the QIAamp Micro Kit with lysis conducted at 56 C for a minimum of 24 hrs. During this period, the samples were vortexed regularly and 10 µL of Proteinase K was added at regular intervals until the sample was digested. The samples were eluted twice using 100 µL of buffer AE. For samples with low concentration, each sample was extracted multiple times following the same protocol and then concentrated using Amicon ultra-0.5 30kDA to a final volume of ~ 100 µL. The sample concentration was measured with a Qubit HS kit and fragment size was analyzed using TapeStation Genomics tape or for highly fragmented sample D1000 or D5000 tapes. A total of 40 samples with >600 ng of DNA and >80 bp average DNA fragment size were selected for whole genome sequencing.

Table 1.4: Details of DNA extraction results.

Sample ID	Museum ID	Species	Museum	Concentration (ng/uL)	Peak fragment size
AO001	80816	*Manis temminckii*	AMNH	0.099	NA
AO001	80816	*Manis temminckii*	AMNH	0.258	NA
AO002	80817	*Manis temminckii*	AMNH	5.22	162
AO002	80817	*Manis temminckii*	AMNH	6.91	NA
AO002	80817	*Manis temminckii*	AMNH	10.7	173
AO003	90095	*Manis temminckii*	AMNH	0.052	NA
AO003	90095	*Manis temminckii*	AMNH	0.075	NA
BW001	168954	*Manis temminckii*	AMNH	0.22	NA
BW001	168954	*Manis temminckii*	AMNH	0.314	103
BW001	168954	*Manis temminckii*	AMNH	1.15	123
BW001	168954	*Manis temminckii*	AMNH	3.54	158
BW001	168954	*Manis temminckii*	AMNH	0.12	105
CD001	101448	*Manis tricuspis*	SI	3.64	169
CD001	101448	*Manis tricuspis*	SI	3.67	149
CD001	101448	*Manis tricuspis*	SI	3.94	186
CD002	101449	*Manis gigantea*	Smithsonian	0.607	NA
CD002	101449	*Manis gigantea*	Smithsonian	0.931	NA
CD002	101449	*Manis gigantea*	Smithsonian	1.06	NA
CD002	101449	*Manis gigantea*	Smithsonian	1.89	NA
CD004	12.12.3.2	*Manis tetradactyla*	london	1.38	151
CD004	12.12.3.2	*Manis tetradactyla*	london	17	239
CD005	53845	*Manis gigantea*	AMNH	1.33	NA
CD005	53845	*Manis gigantea*	AMNH	2.61	NA
CD005	53845	*Manis gigantea*	AMNH	10.1	170
CD006	53846	*Manis gigantea*	AMNH	0.14	NA
CD007	53847	*Manis gigantea*	AMNH	0.066	NA
CD008	53848	*Manis gigantea*	AMNH	0.109	NA
CD009	53849	*Manis gigantea*	AMNH	too low	NA
CD010	53850	*Manis gigantea*	AMNH	0.135	NA
CD010	53850	*Manis gigantea*	AMNH	too low	NA
CD011	53851	*Manis gigantea*	AMNH	0.155	NA
CD012	53853	*Manis gigantea*	AMNH	too low	NA
CD013	53854	*Manis gigantea*	AMNH	0.077	NA
CD014	53855	*Manis gigantea*	AMNH	0.061	NA

CD014	53855	*Manis gigantea*	AMNH	0.117	NA
CD015	53857	*Manis gigantea*	AMNH	0.13	NA
CD015	53857	*Manis gigantea*	AMNH	0.259	NA
CD016	53858	*Manis gigantea*	AMNH	0.155	NA
CD017	53859	*Manis gigantea*	AMNH	0.097	NA
CD018	53860	*Manis gigantea*	AMNH	0.515	76
CD018	53860	*Manis gigantea*	AMNH	0.516	NA
CD018	53860	*Manis gigantea*	AMNH	0.734	84
CD019	53862	*Manis tetradactyla*	AMNH	2.64	90
CD019	53862	*Manis tetradactyla*	AMNH	4.13	223
CD019	53862	*Manis tetradactyla*	AMNH	18.2	254
CD020	53863	*Manis tetradactyla*	AMNH	0.094	NA
CD020	53863	*Manis tetradactyla*	AMNH	too low	NA
CD021	53864	*Manis tetradactyla*	AMNH	0.054	NA
CD021	53864	*Manis tetradactyla*	AMNH	too low	NA
CD022	53866	*Manis tetradactyla*	AMNH	0.492	128
CD022	53866	*Manis tetradactyla*	AMNH	1.11	116
CD022	53866	*Manis tetradactyla*	AMNH	1.14	146
CD022	53866	*Manis tetradactyla*	AMNH	1.86	119
CD023	53867	*Manis tetradactyla*	AMNH	0.067	NA
CD023	53867	*Manis tetradactyla*	AMNH	0.083	NA
CD037	53887	*Manis tricuspis*	AMNH	0.069	NA
CD037	53887	*Manis tricuspis*	AMNH	too low	NA
CD040	53890	*Manis tricuspis*	AMNH	0.104	NA
CD041	53891	*Manis tricuspis*	AMNH	0.085	NA
CD074	53930	*Manis tricuspis*	AMNH	0.088	NA
CD075	53931	*Manis tricuspis*	AMNH	0.083	NA
CD076	53932	*Manis tricuspis*	AMNH	0.126	NA
CD079	53936	*Manis tricuspis*	AMNH	0.457	NA
CD079	53936	*Manis tricuspis*	AMNH	0.468	NA
CD081	55066	*Manis tricuspis*	AMNH	too low	NA
CD082	55067	*Manis tricuspis*	AMNH	0.461	NA
CD082	55067	*Manis tricuspis*	AMNH	0.831	NA
CD082	55067	*Manis tricuspis*	AMNH	1.53	104
CD083	86840	*Manis tetradactyla*	AMNH	0.629	86
CF001	88437	*Manis tricuspis*	AMNH	21.1	193
CF001	88437	*Manis tricuspis*	AMNH	22.7	212

CG001	61.1096	*Manis gigantea*	london	14	489
CI001	89399	*Manis tetradactyla*	AMNH	too low	NA
CI001	89399	*Manis tetradactyla*	AMNH	too low	NA
CM002	48.1325	*Manis tricuspis*	london	7.03	NA
CM002	48.1325..	*Manis tricuspis*	london	21	986
CM002	48.1325.	*Manis tricuspis*	london	28.5	1092
CM004	241130	*Manis tricuspis*	AMNH	6.63	233
CM004	241130	*Manis tricuspis*	AMNH	19.5	209
CM005	241131	*Manis tricuspis*	AMNH	6.92	172
CM005	241131	*Manis tricuspis*	AMNH	10.2	223
CN004	1938.9.7.58	*Manis pentadactyla*	london	1.18	NA
CN005	1938.9.7.57	*Manis pentadactyla*	london	1.07	NA
CN006	26636	*Manis pentadactyla*	AMNH	0.177	NA
CN009	43162	*Manis pentadactyla*	AMNH	0.211	186
CN012	47852	*Manis pentadactyla*	AMNH	8.18	232
CN012	47852	*Manis pentadactyla*	AMNH	9.55	180
CN015	57059	*Manis pentadactyla*	AMNH	11.9	281
CN015	57059	*Manis pentadactyla*	AMNH	32.5	341
GA001	218835	*Manis tetradactyla*	Smithsonian	0.399	78
GA001	218835	*Manis tetradactyla*	Smithsonian	1.98	85
GA001	218835	*Manis tetradactyla*	Smithsonian	2.16	NA
GA001	218835	*Manis tetradactyla*	Smithsonian	too low	NA
GA002	220402	*Manis tetradactyla*	Smithsonian	0.438	91
GA003	220404	*Manis gigantea*	Smithsonian	0.68	92
GA003	220404	*Manis gigantea*	Smithsonian	0.799	66
GA003	220404	*Manis gigantea*	Smithsonian	0.916	NA
GA003	220404	*Manis gigantea*	Smithsonian	6.55	159
GA005	119963	*Manis tricuspis*	AMNH	12.7	387
GH006	35.10.22.157	*Manis tetradactyla*	london	0.44	152
GH006	35.10.22.157	*Manis tetradactyla*	london	1.08	152
GH006	35.10.22.157	*Manis tetradactyla*	london	1.64	203
GH006	35.10.22.157	*Manis tetradactyla*	london	1.67	NA
GH006	35.10.22.157	*Manis tetradactyla*	london	1.92	171
GH006	35.10.22.157	*Manis tetradactyla*	london	2.74	169
GQ001	598576	*Manis gigantea*	Smithsonian	0.052	NA
GQ002	598577	*Manis gigantea*	Smithsonian	0.347	109
GQ002	598577	*Manis gigantea*	Smithsonian	0.518	77

GQ002	598577	*Manis gigantea*	Smithsonian	0.589	NA
GQ002	598577	*Manis gigantea*	Smithsonian	3.04	NA
GQ002	598577	*Manis gigantea*	Smithsonian	5.21	212
ID007	142461	*Manis javanica*	Smithsonian	0.559	NA
ID010	198319	*Manis javanica*	Smithsonian	6.36	209
ID010	198319	*Manis javanica*	Smithsonian	13.4	192
ID013	267388	*Manis javanica*	Smithsonian	9.1	212
ID013	267388	*Manis javanica*	Smithsonian	9.84	329
ID014	267389	*Manis javanica*	Smithsonian	5.9	226
ID014	267389	*Manis javanica*	Smithsonian	14	243
ID015	9.1.5.858	*Manis javanica*	london	1.46	175
ID017	9.1.5.857	*Manis javanica*	london	14.1	200
ID019	101562	*Manis javanica*	AMNH	too low	NA
ID023	101566	*Manis javanica*	AMNH	0.142	NA
ID035	102096	*Manis javanica*	AMNH	0.327	NA
ID037	102098	*Manis javanica*	AMNH	too low	NA
ID039	102149	*Manis javanica*	AMNH	0.144	NA
ID040	102180	*Manis javanica*	AMNH	0.1	NA
ID045	17600	*Manis javanica*	AMNH	0.363	NA
ID045	17600	*Manis javanica*	AMNH	0.423	NA
ID045	17600	*Manis javanica*	AMNH	0.454	NA
ID045	17600	*Manis javanica*	AMNH	0.467	NA
ID045	17600	*Manis javanica*	AMNH	0.636	NA
ID045	17600	*Manis javanica*	AMNH	0.729	NA
ID045	17600	*Manis javanica*	AMNH	0.9	NA
IN001	21.8.2.27	*Manis pentadactyla*	london	6.34	NA
IN001	21.8.2.27	*Manis pentadactyla*	london	8.93	217
KH001	321551	*Manis javanica*	Smithsonian	25	258
LA002	87622	*Manis javanica*	AMNH	0.064	NA
LA003	87624	*Manis javanica*	AMNH	too low	187
LK001	76.138	*Manis crassicaudata*	london	10.6	182
LK001	76.138	*Manis crassicaudata*	london	14.3	277
LK002	1950.29	*Manis crassicaudata*	london	19	294
MM001	124242	*Manis javanica*	Smithsonian	0.05	NA
MM002	34.3.4.1	*Manis crassicaudata/pentadactyla*	london	9.11	254
MM002	34.3.4.1	*Manis crassicaudata/pentadactyla*	london	30.9	399
MM003	50.752..	*Manis pentadactyla*	london	1.63	NA

MM003	50.752..	*Manis pentadactyla*	london	4.36	NA
MM003	50.752..	*Manis pentadactyla*	london	6.88	NA
MM004	50.751	*Manis pentadactyla*	london	1.2	NA
MM005	14.7.19.238	*Manis pentadactyla*	london	14.7	NA
MM005	14.7.19.238	*Manis pentadactyla*	london	19.4	270
MY002	317198	*Manis javanica*	Smithsonian	too high	1483
MY003	317199	*Manis javanica*	Smithsonian	too high	2121
MY004	317200	*Manis javanica*	Smithsonian	5.56	NA
MY005	488088	*Manis javanica*	Smithsonian	1.3	NA
MY006	98.8.3.6	*Manis javanica*	london	0.251	NA
MY007	32637	*Manis javanica*	AMNH	12.7	172
MZ001	42349	*Manis temminckii*	AMNH	0.77	NA
MZ001	42349	*Manis temminckii*	AMNH	1.48	254
MZ001	42349	*Manis temminckii*	AMNH	2.39	281
MZ001	42349	*Manis temminckii*	AMNH	3.18	304
MZ002	216259	*Manis temminckii*	AMNH	0.055	NA
MZ002	216259	*Manis temminckii*	AMNH	0.115	NA
NG005	1998.302	*Manis tetradactyla*	london	51	784
PH001	203298	*Manis culionensis*	AMNH	0.636	NA
PH001	203298	*Manis culionensis*	AMNH	1.68	NA
PH001	203298	*Manis culionensis*	AMNH	3.51	NA
PH002	242095/24795	*Manis culionensis*	AMNH	0.112	NA
PH002	242095	*Manis culionensis*	AMNH	0.715	NA
PK002	244406	*Manis crassicaudata*	AMNH	too high	569
PK003	244407	*Manis crassicaudata*	AMNH	45.4	567
SL001	46.888..	*Manis tricuspis*	london	8.17	227
SL001	46.888..	*Manis tricuspis*	london	19	269
TH005	257682	*Manis javanica*	Smithsonian	0.152	NA
TH007	167959	*Manis pentadactyla*	AMNH	0.192	NA
TW007	183148	*Manis pentadactyla*	AMNH	11.2	463
TW009	184958	*Manis pentadactyla*	AMNH	0.707	NA
TZ001	83772	*Manis temminckii*	AMNH	0.773	NA
TZ001	83772	*Manis temminckii*	AMNH	8.88	208
TZ001	83772	*Manis temminckii*	AMNH	0.451	141
UG001	61.1095..	*Manis gigantea*	london	0.774	165
UG001	61.1095..	*Manis gigantea*	london	2.18	115
UG001	61.1095	*Manis gigantea*	london	2.62	151

UG001	61.1095	*Manis gigantea*	london	6.2	170
UG002	28.9.8.37	*Manis tricuspis*	london	1.64	208
UG002	28.9.8.37	*Manis tricuspis*	london	4.5	254
UG002	28.9.8.37	*Manis tricuspis*	london	7.16	191
VN006	27.12.1.203	*Manis pentadactyla*	london	0.066	276
VN006	27.12.1.203	*Manis pentadactyla*	london	2.06	241
VN006	27.12.1.203	*Manis pentadactyla*	london	3.66	66
VN006	27.12.1.203	*Manis pentadactyla*	london	9.73	201
ZA001	83609	*Manis temminckii*	AMNH	4.76	225
ZA001	83609	*Manis temminckii*	AMNH	8.93	205
ZA001	83609	*Manis temminckii*	AMNH	9.43	207
ZW001	368617	*Manis temminckii*	Smithsonian	2.32	NA
ZW001	368617	*Manis temminckii*	Smithsonian	3.08	NA
TH004	236633	*Manis javanica*	Smithsonian	NA	NA
TH001	83250	*Manis javanica*	Smithsonian	NA	221
KH001	321551	*Manis javanica*	Smithsonian	NA	263
MM001	124242	*Manis javanica*	Smithsonian	NA	872
TH003	86789	*Manis javanica*	Smithsonian	NA	NA
NP001	1963.36	*Manis pentadactyla*	london	NA	433
NG004	1995.25	*Manis tetradactyla*	london	NA	589
TH004	236633	*Manis javanica*	Smithsonian	NA	NA
ZW001	368617	*Manis temminckii*	Smithsonian	NA	NA
GA004	119964	*Manis tricuspis*	AMNH	NA	721
PK001	185599	*Manis crassicaudata*	AMNH	NA	804

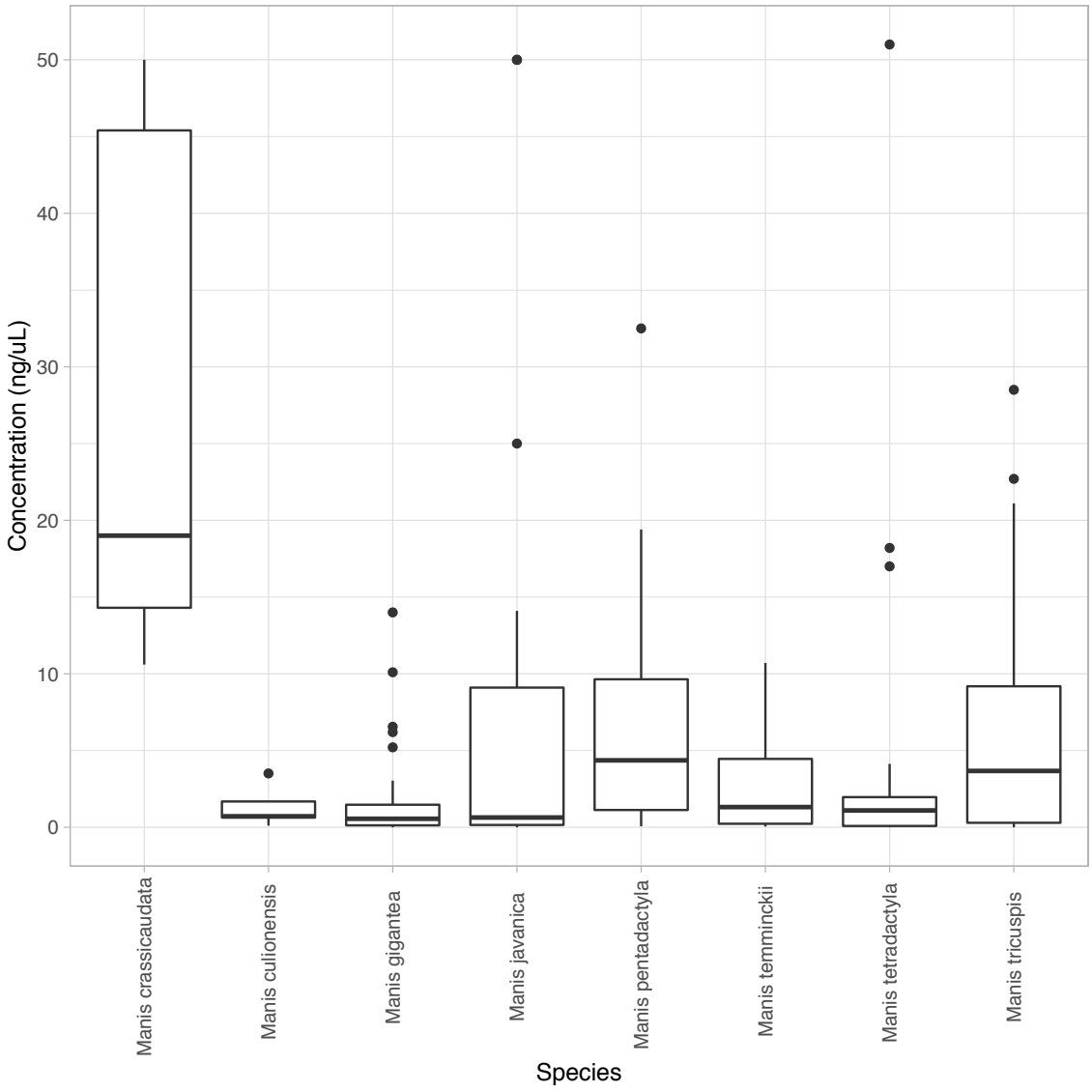

Figure 1.10: DNA concentration by species.

Table 1.5: Mitochondrial genomes used as referene genomes for mitochondrial sequence alignment and assembly.

Species	Accession Number
Manis pentadactyla	MG196307
Manis crassicaudata	MG196304
Manis javanica	MG196302
Manis gigantea	MG196301
Manis temminckii	MG196300
Manis tetradactyla	MG196299
Manis tricuspis	MG196296

2

Forensic genetic marker development to track the illegal pangolin trade

2.1 Abstract

Illegal wildlife trade has burgeoned over the last few decades and impacts a diverse array of taxonomic groups. One group of mammals, pangolins (Manidae species), has been subject to trade for their meat and scale over many decades, but has recently been intensively traded to fulfill rising demands. This has led to concerted enforcement efforts to understand and curb the large transnational illegal trade of pangolins. The geographic provenance of the large scale seizures have been a persistent enigma due to the lack of accurate tools to assign origin to pangolin specimens and parts in trade. In this chapter, pangolin specimens from natural history museum collections are used to identify genetic markers for wildlife forensic and establish a genetic database for genetic assignment. Whole genome shot-gun sequencing was conducted on 48 geo-referenced pangolin samples from seven species from both historic and modern samples generating genome-wide data. Single nucleotide polymorphisms (SNPs) were identified using two variant callers resulting in 60,082,067 SNPs. Stringent filtering criteria were applied and 1094 putative SNPs were retained as a putative set of markers that could be validated for use in wildlife forensics to identify geographic origin of pangolin seizures. The SNP panel was tested for species and population differentiation and geographic assignment capabilities using a software that implements the 'Smoothed Continuous Assignment Technique (SCAT). The limited number of samples in the discovery phase calls for the need to genotype a larger sample pool to identify the most informative SNPs, however, this results of this chapter provides the foundation for a pangolin forensics marker set for future studies.

2.2 Introduction

Wildlife trafficking has become a global phenomenon with links to transnational organized crime impacting a wide range of taxonomic groups (Anagnostou, 2021; Wyatt, Uhm, & Nurse, 2020). Pangolins are caught in this illicit trade driven largely by demand for their meat eaten as a delicacy, scales used for traditional medicine and use in luxury products. High levels of trade in pangolins and estimated decrease in the number of pangolins in the wild (Nash et al., 2016; Newton et al., 2008) have granted pangolins national and international protection from harvest and trade. At the international level, all pangolins species, four in Africa and four in Asia, are listed in Appendix I of the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) and no commercial international trade is allowed. Their risk for extinction was recognized and have been placed on the IUCN Red List of Threatened Species. The four Asian pangolins are either Critically endangered (*Manis pentadactyla*, *M. javanica*, *M. culionensis*) or Endangered (*M crassicaudata*) while the four African pangolins are either Endangered (*M. gigantea*, *M. tricuspis*) or Vulnerable (*M. temminckii*, *M. tetradactyla*).

The high levels of protection afforded to pangolins have not curbed the demand for their products. In conjunction with the protection and low density of remaining Asian pangolins, there has been a shifts in source countries from Asia to Africa (Heinrich et al., 2016) and recently from legal to illegal trade (UNODC, 2020). The source and trade routes of this large scale illegal trade have been difficult to identify because of the wide distribution of pangolins and their elusive nature. Source countries reported in shipments are unreliable because contraband can be shipped from a country that is different from where the animals were originally poached (Wasser et al., 2015). Even when the source country and species can be identified, pangolins could have been poached in any of dozens of different protected areas in a single country.

Due to the morphological similarity among species of each continent, identification of the geographic origin of the specimen is difficult, if not, impossible, based on morphology alone. Genetic methods based on mitochondrial, microsatellite, and single nucleotide polymorphisms (SNPs) have been applied to identify the species in trade and geographic origin of the specimens. While all of the studies have revealed novel

information on the species composition of seizures (Hsieh et al., 2011; Mwale et al., 2017), trade routes, and large scale geographic origin assessment (Nash et al., 2018; Zhang et al., 2020), the lack of a genetic database, or reference map, for comparison to perform genetic assignment has limited the precision of geographic origin identification. Previous work on the illegal trade in African elephant ivory has shown that with a large number of samples in a reference map, major elephant poaching hotspots can be identified (Wasser et al., 2015, 2007; Wasser, Shedlock, et al., 2004).

The power of genetic assignment testing relies on an extensive database of genetic profiles of each known population that can be used for comparison, but such samples are difficult to acquire for endangered species that are also elusive. Natural history museums provide valuable resources for building a reference map as a large number of samples held in museum collections are geo-referenced. The recent advances in genetic sequencing has lowered the barrier for using historical museum specimens, most of which have been collected and stored prior to the molecular era. The highly fragmented DNA in historic samples and low DNA yield can be overcome by sequencing all of the DNA fragments in the sample using shotgun sequencing. Using this genome-wide data, SNP loci that are informative for population differentiation can be identified for use in geographic origin assignment (Benestan et al., 2015; Puckett & Eggert, 2015).

In this chapter, high throughput sequencing is employed to harness the genomic information in geo-referenced pangolin samples in natural history museums with the aim of establishing a foundation for a wildlife forensic tool for pangolins. This is conducted by identifying high quality SNPs across the genome in seven pangolin species across their distribution and establishing a reference map for geographic origin assignment of pangolin samples.

2.3 Methods

2.3.1 Sample information

For identification of SNPs for wildlife forensic applications, samples with geographic provenance information were selected from the previous chapter's dataset. The sampling locality provided by the museum was used to identify geographic coordinates. Coordinates were assigned by either mapping the sampling

location on Google Maps, searching articles that reference the location, or identifying the location on maps from the time of the sample collection. For samples with ambiguous locations, a coordinate was picked based on best available information. For species with limited number of samples, samples that only had information on the country of collection were used and the country's centroid coordinates were used in place. Details of the source used to assign coordinates are in Supplementary table 2.2.

2.3.2 Analysis

Variant calling and filtering

Variant calling was conducted as a full sample set of 48 samples and separately for each species using Freebayes v1.3.5 (Garrison & Marth, 2012) and ANGSD v0.935 (Korneliussen et al., 2014). Criteria of minimum mapping quality of 20, minimum base quality of 20, minor allele frequency of 0.02 and a minimum individual missing of 42 was applied. Concordant SNPs, those that have the same reference and alternate alleles, between the two variant callers were identified using `bcftools isec` then a set of site quality filters were applied to retain loci with a minimum depth of 3 and a maximum depth of 25, a quality score of 20, missing rate of 0.05, and a minor allele frequency of 0.03 using `vcftools`. The set of SNPs for each species were compared to each other to identify loci that were present in at least two of the seven species SNP set. Genotype of these loci were extracted for all 48 samples from the full sample set from the ANGSD and Freebayes variant call results. The loci were thinned to retain only those that were 20,000 bp from each other to avoid selecting loci that are linked due to close proximity.

Selection of SNPs

The set of SNPs was used to estimate the informativeness of the SNPs were for population differentiation in each species, where possible. To identify a panel of forensic SNPs, the informativeness of each SNP for population assignment was estimated using TRES v1.0 (Kavakiotis et al., 2015) and informativeness for assignment (In; Ding et al. (2011) and Rosenberg, Li, Ward, & Pritchard (2003)) and delta (Shriver et al., 1997) measures were calculated to select the 100 loci ranked highest for each measure of marker

informativeness. Populations used to calculated the measures were designated by geographic distance of the samples for each species and using previously identified population clusters (Gaubert et al., 2016; Hu et al., 2020; Nash et al., 2018). A principal component analysis was performed on the forensic SNP panel in 48 samples to assess the informativeness of the SNP panel.

Assignment test analysis

Using the genotype data of the 48 individuals, a leave-one-out analysis was conducted using the program Smooth Continuous Assignment Technique (SCAT version 3.0; Wasser, Shedlock, et al. (2004); Wasser et al. (2007)) to assess accuracy of assignment. The program SCAT uses three input files, a genotype file of known origin samples, a location file of each sample in the known origin file, and a genotype file of unknown individuals. The location file was created using the same population divisions used to identify informative SNPs in the selection process. Assignment was tested by running SCAT with three MCMC chains using the default settings and the first 100 values were discarded as burn-in.

2.4 Results

The 48 geo-referenced samples analyzed together resulted in 156,181,707 and 105,420,052 SNPs by Freebayes and ANGSD, respectively, of which 60,082,067 SNPs were identified by both variant callers. The samples from seven species of pangolins were mapped to a reference genome of a single species (*M. javanica*) resulting in a high number of SNPs, especially due to the high divergence between the reference genome and the African pangolin species. SNPs were also called by species and the number of concordant SNPs between the two variant callers for each species is shown in Table 2.1. On average, around 29% of SNPs passed site based quality filters, but ranged from 17.5% for *M. pentadactyla* retaining 1,195,116 SNPs to 60% for *M. javanica* leaving 6,109,074 SNPs. The highest number of high quality SNPs were trained for *M. javanica*, which had the largest sample size of 12 and also was the species of the reference genome the raw reads were aligned to, while the fewest number of SNPs were retained for *M. crassicaudata* and *M. gigantea*, which also had the smallest sample size of four each.

Table 2.1: Information on the SNPs identified in this study by species and variant caller.

Species	Number of samples	Number of SNPs Freebayes	Number of SNPs ANGSD	Number of concordant SNPs	Number of SNPs post-filtering
Full sample set	48	156181707	105420052	60082067	NA
Manis javanica	12	40805548	13883662	10260918	6109074
Manis pentadactyla	9	62208969	13069196	6836037	1195116
Manis crassicaudata	4	43742273	6391597	2968613	635024
Manis tetradactyla	6	69493764	10355450	3588540	1101413
Manis tricuspis	8	75864885	21873882	8589451	1518091
Manis temminckii	5	68474740	10294902	3362259	1098778
Manis gigantea	4	67909455	11186914	3218305	838127

Seven species-specific SNP sets were compared to each other to identify concordant loci. Setting the number of species that loci needed to be present to be retained from 2 to 5 resulted in 215,575, 17,567, 1,502, and 57 loci. To ensure sufficient number of SNPs for filtering and retain species-specific informativeness, loci present in a minimum of two species, 215575 loci, were used for further analysis. Out of the 215,575 loci, 175,646 were identified in the full data set variant calls by either variant callers. After thinning the loci to retain those at a minimum 20,000 bp apart, 57,064 SNPs remained.

The 57,064 SNPs were analyzed for informativeness for population assignment within each species. The SNPs were ranked based on the In and delta metric scores and the top 100 SNPs for each metric were retained for each species. The distribution of In and delta metrics across the 57,064 SNPs are shown in Supplementary figures 2.3, 2.4, 2.5, and 2.6. Due to some overlap in SNPs, the 1400 SNPs resulted in 1094 unique SNPs with the highest level of informativeness across the species and measures for validation for forensic application. Detailed statistics on the allele and genotype frequencies are in Supplementary table 2.3. In brief, the alternative allele frequency ranged from 0.02 to 0.84 and heterozygote frequency ranged from 0 to 0.563. PCA of the seven species using the forensic SNP panel showed that the selected SNPs were able to delineate species and within species structure (Figure 2.1). The clustering of each species was not as distinct as when the entire SNP set is used, but within species differentiation can be seen in *M tricuspis*, *M.*

javanica, and *M. pentadactyla*. A leave-one-out analysis removed each sample from the reference dataset and used it as a sample of unknown locality to perform SCAT and assess the accuracy of the results. Due to the small sample size, samples that came from an existing population in the reference dataset led to high accuracy with minimal variation around the actual coordinates of the sample. For samples that lacked other samples from the same population in the reference dataset (i.e. were the sole sample of the population), results were more varied. The SCAT results of two such samples, one *M. pentadactyla* sample from Nanping, Fujian Province, China, and one *M. temminckii* sample from Mutare, Zimbabwe, are shown in Figure 2.2.

2.5 Discussion

The natural history museum collections provided a valuable resource for establishing baseline genetic information on a rare and endangered group of mammals. Combining the use of high-throughput sequencing for large scale SNP development and the georeference information of the museum specimens allowed for the identification of a putative SNP set that can be validated for wildlife forensic use.

Developing a set of forensic markers from a small number of museum specimens results in ascertainment bias due to the small sample size used to identify the genetic markers in comparison to the large distribution of pangolin . There were not sufficient samples for each species to delineate populations to assess the power of the SNP panel for population differentiation. Although data from seven out of the eight known species of pangolins were included and covered a majority of the distribution range, there were insufficient sample size for each species to define population structure based on genetic data. The population defined in this chapter were based on geographic distance between samples and based on previously published population structure information.

The small sample size and population numbers render drawing conclusion on the accuracy of the assignment tests difficult. The use of smooth continuous assignment technique is superior to traditional assignment methods as it allows for the assignment of unknown samples to locations without a reference sample. By inferring a continuous allele frequency map across the pangolin distribution based on the

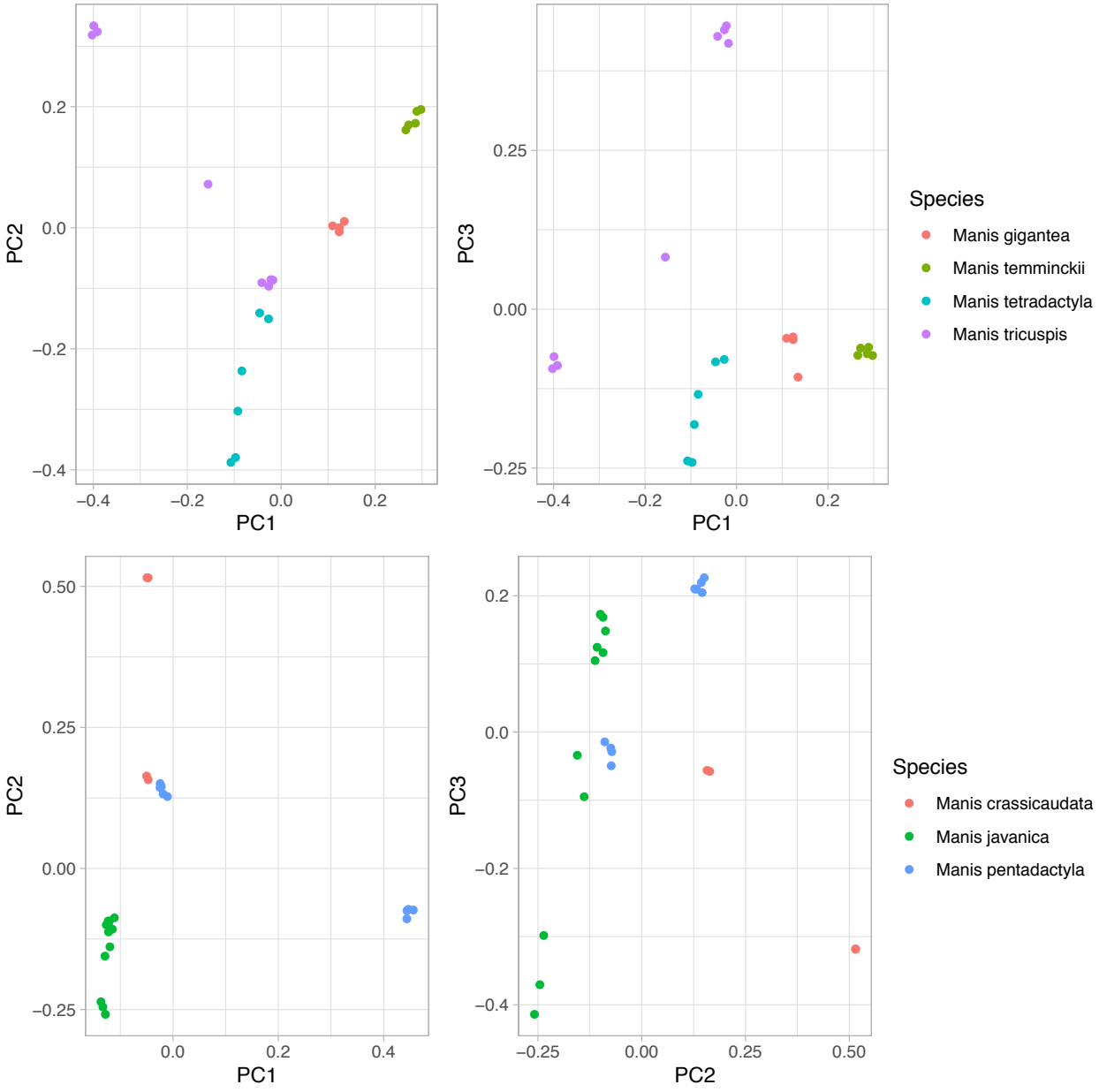

Figure 2.1: PCA plot of 46 samples from genotype data of 1904 SNPs.

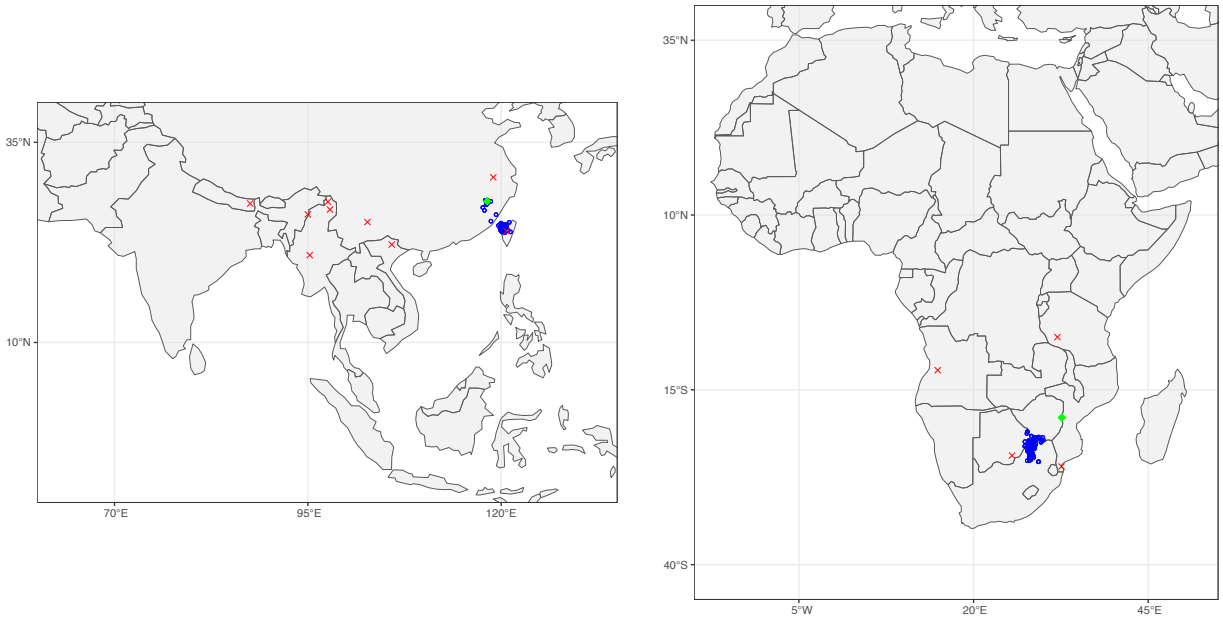

Figure 2.2: Genetic assignment results using SCAT. The known location of sample collection is shown as a green diamond and the inferred location using SCAT is shown in blue dots. The cross marks are locations of samples of the same species included in the reference database.

input data, the SCAT software is able to assign samples to locations without reference samples. The ability to assign individuals to anywhere in the species' distribution range makes SCAT especially useful for pangolins due to their wide distribution and the difficulty of sampling pangolin populations across their wide distribution. When a larger sample size is acquired and loci under selection can be positively identified, the addition of those loci may increase the power of assignment while reducing the total number of SNPs required (Freamo, O'Reilly, Berg, Lien, & Boulding, 2011; Moore et al., 2014).

The number of samples and SNPs required for accurate assignment is related to the power for detecing population differentiation. Statistical power to identify evolutionary significant units is dependent on the level of differentiation between populations. A simulation study showed that to reach a power of 0.8 with 75 SNPs, 100 samples are required to reach a power of 0.8 when the FST is low (0.0025), but only 30 samples when the FST is higher (0.01; Morin, Martien, & Taylor (2009)). However, when the number of SNPs are increased substantially, it is possible to detect genetic differentiation even with a small sample size with a simulation showing that with 3000 SNPs, it is possible to detect FST of 0.01 with four individuals (Willing, Dreyer, & Van Oosterhout, 2012).

Museum specimens that have been preserved for several decades contain sufficient amount of DNA for genetic analysis, but the DNA that remains in the sample is fragmented and may have been subjected to post-mortem damage (Keighley et al., 2021; Nakahama, 2021). In addition, the preservation method used to store the specimens over multiple decades can also have an impact on the integrity of the DNA structure (Zimmermann et al., 2008). The genetic markers that have been identified to be polymorphic may be an artifact or DNA damage error and not a real polymorphism in the population. This requires laboratory validation by genotyping a larger number of samples across multiple populations to assess information content of the SNP and validate that the SNP is real and not an artifact.

The best approach to validate the SNPs on a large pool of museum specimens is through the use of an ultra accurate sequencing technology called Duplex Sequencing (Kennedy et al., 2014). Post-mortem damage on DNA frequently results in deamination of cytosine to uracil, which will result in incorporation of an adenine during PCR and leads to a G to A and C to T substitution on a single strand. Duplex sequencing is based on the ligation of a a double-stranded complementary unique molecular identifier (UMI) and

then enriched for the SNP loci using two rounds of hybridization capture. As each strand of each DNA molecule is tagged with unique barcode that is complementary, it is possible to identify the sequences resulting from each DNA strand of each molecule to build a single and then a double-stranded consensus sequence of the molecule. This allows one to for the identification of mutations that are found only on one strand due to DNA damage, PCR replication, and sequencing errors. True mutations should be present on both strands.

The set of SNPs identified in this study show promise for differentiation between and within species of pangolins. The SNP loci are amplifiable across multiple species and based on a small sample size of 48 show high levels of informativeness for population assignment. The SNPs have also been vetted for quality to increase the possibility that the loci are real variants and not artifacts of DNA damage and error. Further validation of the SNP loci are need to identify those that are polymorphic and informative across a large population sample to finalize a set of SNP markers for forensic use in pangolins.

2.6 Acknowledgements

This work was supported by the USAID Wildlife Crime Tech Challenge. This work was facilitated through the use of advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system and funded by the STF at the University of Washington.

2.7 Data availability

Code used to perform the analysis and the set of putative forensic SNPs are available of GitHub (https://github.com/kimhii/dissertation).

2.8 Supplementary materials

Table 2.2: Locality information on the samples included in the identification of SNPs.

Sample ID	Species	Museum name	Museum ID	Location	Latitude	Longitude	Notes
442962	*Manis temminckii*	AMNH	80817	Mombolo Cuansa Sul, Angola	-12.198333	14.868333	Branch et al. 2017
442963	*Manis temminckii*	AMNH	168954	Molepolole Kweneng, Botswana	-24.406667	25.495000	Wikipedia
442964	*Manis tricuspis*	Smithsonian	101448	Kassai, Congo Democratic Rep. Of	-6.000000	22.000000	NA
442965	*Manis tetradactyla*	NHM London	12.12.3.2	Uele, DR Congo	3.470046	25.579216	NA
442966	*Manis gigantea*	AMNH	53845	Bafuka Haut Zaire, Zaire	3.320000	28.550000	Google maps
442967	*Manis tetradactyla*	AMNH	53862	Bolobo, Zaire	-2.166667	16.233333	Google maps
442968	*Manis tricuspis*	AMNH	88437	Ippy, Central African Republic	6.250000	21.200000	Google maps
442969	*Manis gigantea*	NHM London	61.1096	Zambo, Kiru E Congo	-1.900815	28.730915	Google maps
442970	*Manis tricuspis*	NHM London	48.1325	Mamfe, British Cameroons	5.766667	9.283333	Wikipedia for Mamfe, Cameroon
442971	*Manis tricuspis*	AMNH	241130	30 km West of Bertoua, Est Region Cameroon	4.575539	13.406743	Picked a location 30 km west of Bertoua, east region
442972	*Manis tricuspis*	AMNH	241131	30 km West of Bertoua, Est Region, Cameroon	4.575539	13.406743	Picked a location 30 km west of Bertoua, east region
442973	*Manis pentadactyla*	AMNH	47852	Nanping, Fujian Province, China	27.642000	118.179000	Wikipedia
442974	*Manis pentadactyla*	AMNH	57059	Ningkuofu an bwei, China	30.623617	118.981841	Google maps location for ningguo city in an hui (help from http://bdcconline.net/en/stories/mather-percy-cunningham)
442975	*Manis gigantea*	Smithsonian	220404	Anguanamo, Ngori	-1.927517	9.469287	Picked an island in Iguela Lagoon based on Rich (2012), which stated 'on island of Anguanamo on Iguela Llagoon"
336746	*Manis tricuspis*	AMNH	119964	Gamboma, Gabon	-1.874424	15.881490	Based on the map in Malbrant and Maclatchy (1949) - Picked the Gamboma by the river (google maps shows two gambomas this is the one to the east)
442976	*Manis tricuspis*	AMNH	119963	Fougamou, Gabon	-1.219304	10.588160	Google maps
442977	*Manis tetradactyla*	NHM London	35.10.22.157	[Goaso], Ashanti [Chana]	6.801538	-2.519406	Google maps
442978	*Manis javanica*	Smithsonian	198319	Madang Borneo, Indonesia	-2.780875	115.329724	Google maps

442979	*Manis javanica*	Smithsonian	267388	Siantar Sumatra, Indonesia	2.960000	99.060000	Wikipedia for Pematangsiantar
442980	*Manis javanica*	Smithsonian	267389	Siantar Sumatra, Indonesia	2.960000	99.060000	Wikipedia for Pematangsiantar
442981	*Manis javanica*	NHM London	9.1.5.857	Butenzorg Java	-6.597104	106.802663	Google maps for Bogor
442982	*Manis pentadactyla*	NHM London	21.8.2.27	Kheusa, Naga Hills Assam	25.994838	95.003150	Google maps for Naga hills (can not find Kheusa)
336744	*Manis javanica*	Smithsonian	321551	Damrey Phong, 2 Km SE, Cambodia	10.771149	104.231188	Google maps - picked a location 2 km SE of Damrey Phong
442983	*Manis crassicaudata*	NHM London	76.138000000000005	Kekkirawa [SSri Lanka]	8.048180	80.596218	Google maps for Kekkirawa Sri Lanka
442984	*Manis crassicaudata*	NHM London	1950.29	Kekirawa, N.C. Prov. Ceylon	8.047416	80.595789	Google maps for Kekkirawa Sri Lanka
442985	*Manis crassicaudata*	NHM London	34.3.4.1	30km NE Fort Herty Upper Burma	27.566773	97.595407	Google maps - location 30 km NE of Wikipedia's Fort Herty location
442986	*Manis pentadactyla*	NHM London	50.752000000000002	Htingnan Upper Burma	26.600000	97.866667	Museum location - 26 deg 36 N, 97deg 52'E
442987	*unknown*	NHM London	14.7.19.238	Mt. Poppa Upper Burma	20.905471	95.248499	Google maps - location near Mt Poppa
442988	*Manis javanica*	Smithsonian	317198	Ranau Sabah Borneo, Malaysia	5.966667	116.683333	Wikipedia for Ranau District
442989	*Manis javanica*	Smithsonian	317199	Ranau Sabah Borneo, Malaysia	5.966667	116.683333	Wikipedia for Ranau district
442990	*Manis javanica*	AMNH	32637	Sarawak, 10 mi S of Kuching, Malaysia	1.437321	110.333409	Google maps - picked a location 10 mi south of Kuching
442991	*Manis temminckii*	AMNH	42349	Lourenco Marques, Mozambique	-25.966667	32.583333	Google maps for Maputa (Lourenco Marques until 1976)
336743	*Manis tetradactyla*	NHM London	1995.25	Aseingbene, West bank, Lower Taylor Creek, Niger Delta, Nigeria	5.083333	6.316667	Coordinates from Luiselli et al. 2000
442992	*Manis tetradactyla*	NHM London	1998.30199999999999	Obiofu	6.098378	6.699914	Google maps - Obiofu, Nigeria
336742	*Manis pentadactyla*	NHM London	63.36	Sanghu, Maewa Khola E Nepal	27.350000	87.550000	Museum location - 27deg 21'N, 87 deg 33'E
336747	*Manis crassicaudata*	AMNH	185599	Hab River Valley, Pakistan	26.244099	67.207068	Google maps - Hub river
442993	*Manis crassicaudata*	AMNH	244406	Pakistan	30.000000	70.000000	NA
442994	*Manis crassicaudata*	AMNH	244407	Pakistan	30.000000	70.000000	NA
442995	*Manis tricuspis*	NHM London	46.887999999999998	Binkold, N of Makeni Sierra Leone	8.947433	-11.979302	Binkolo ne of makeni sierra leone (possible a typo)
442996	*Manis pentadactyla*	AMNH	183148	Nan-Tou Hsien, Wu-shen, Taiwan	23.915356	120.684533	Google maps - Nantou city
442997	*Manis temminckii*	AMNH	83772	Kawue, Tabora Region, Tanzania	-7.452765	32.001884	Google maps - Picked a point next to Kawue lake, but the lake is not in Tabora region.
442998	*Manis gigantea*	NHM London	61.1094999999999997	Buddu S Uganda	-0.416670	31.666670	Wikipedia for Buddu of Uganda

442999	*Manis tricuspis*	NHM London	28.9.8.37	near Kampala Uganda	0.377558	32.582245	Google maps - location near Kampala
443000	*Manis pentadactyla*	NHM London	27.12.1.203	Bac-kam, Tonkin [Vietnam]	22.229716	105.855133	Google maps - location in Bac-Kan province, can not find Bac-Kam, possibly an error.
443001	*Manis temminckii*	AMNH	83609	Transvaal, South Africa	-23.946169	29.580526	Google map - central point in the province
336745	*Manis temminckii*	Smithsonian	368617	Umtali, Zimbabwe	-18.966667	32.633333	Wikipedia for Mutare (name change in 1983)
10594	*Manis javanica*	Burke	EEM957	Singapore	1.386234	103.804540	NA
10595	*Manis javanica*	Burke	CSW8292	Singapore	1.386234	103.804540	NA
10596	*Manis javanica*	Burke	CSW8308	Singapore	1.386234	103.804540	NA
10600	*Manis javanica*	Burke	CSW8288	Singapore	1.386234	103.804540	NA

Table 2.3: Allele frequency information for the top 30 most common frequencies.

n	ref (n)	ref (freq)	het (n)	het (freq)	alt (n)	alt (freq)	mis (n)	mis (freq)
34	43	0.896	2	0.042	3	0.063	0	0.000
33	42	0.875	2	0.042	4	0.083	0	0.000
31	41	0.854	3	0.063	4	0.083	0	0.000
27	44	0.917	0	0.000	4	0.083	0	0.000
24	43	0.896	3	0.063	2	0.042	0	0.000
21	44	0.917	2	0.042	2	0.042	0	0.000
21	45	0.938	0	0.000	3	0.063	0	0.000
21	43	0.896	0	0.000	4	0.083	1	0.021
21	42	0.875	3	0.063	3	0.063	0	0.000
20	43	0.896	1	0.021	4	0.083	0	0.000
19	45	0.938	1	0.021	2	0.042	0	0.000
19	44	0.917	1	0.021	3	0.063	0	0.000
19	42	0.875	1	0.021	5	0.104	0	0.000
18	40	0.833	5	0.104	3	0.063	0	0.000
17	44	0.917	0	0.000	3	0.063	1	0.021
17	41	0.854	4	0.083	3	0.063	0	0.000
16	42	0.875	4	0.083	2	0.042	0	0.000
14	39	0.813	4	0.083	5	0.104	0	0.000
13	40	0.833	3	0.063	5	0.104	0	0.000
13	40	0.833	4	0.083	4	0.083	0	0.000
12	43	0.896	0	0.000	5	0.104	0	0.000
12	46	0.958	0	0.000	2	0.042	0	0.000
11	44	0.917	3	0.063	1	0.021	0	0.000
11	38	0.792	5	0.104	5	0.104	0	0.000
11	39	0.813	3	0.063	6	0.125	0	0.000
10	41	0.854	1	0.021	6	0.125	0	0.000
10	45	0.938	0	0.000	2	0.042	1	0.021
9	39	0.813	5	0.104	4	0.083	0	0.000
9	41	0.854	2	0.042	5	0.104	0	0.000
9	42	0.875	1	0.021	4	0.083	1	0.021

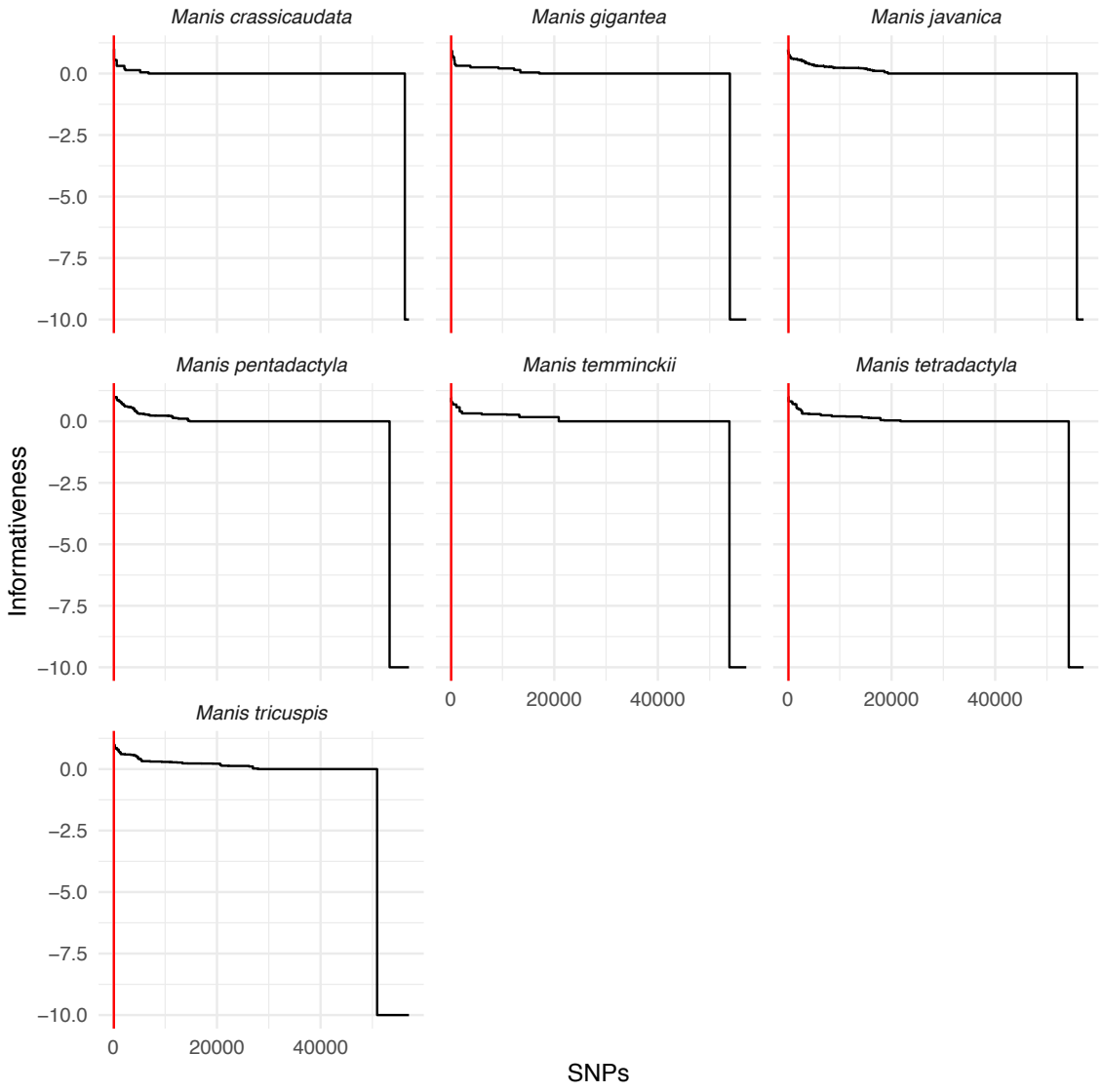

Figure 2.3: Informativeness scores of the SNPs for each species. The red vertical line denotes the cut off of the top 100 SNPs that were retained for the forensic SNP panel.

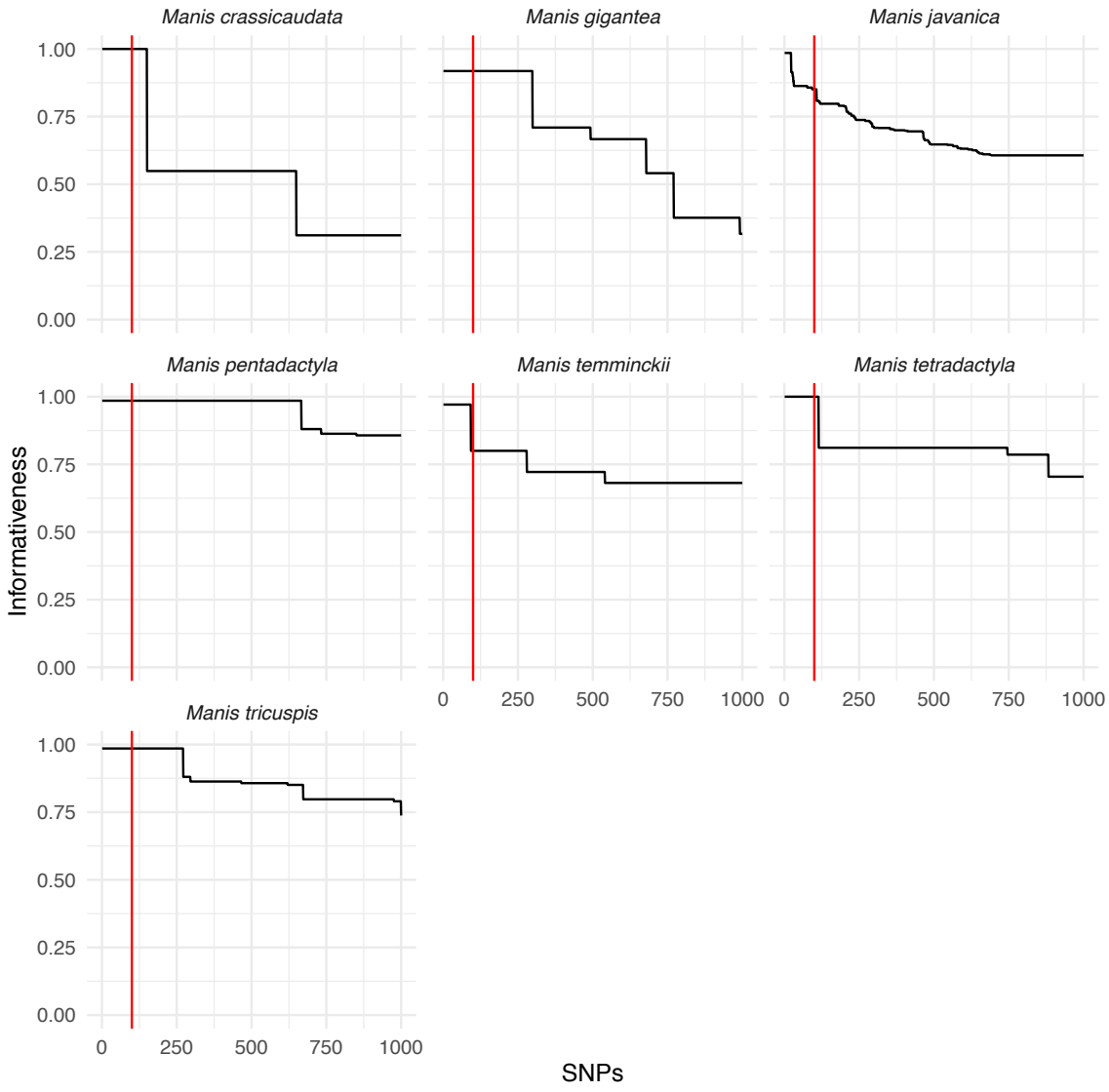

Figure 2.4: Informativeness scores of the top 1000 SNPs for each species. The red vertical line denotes the cut off of the top 100 SNPs that were retained for the forensic SNP panel.

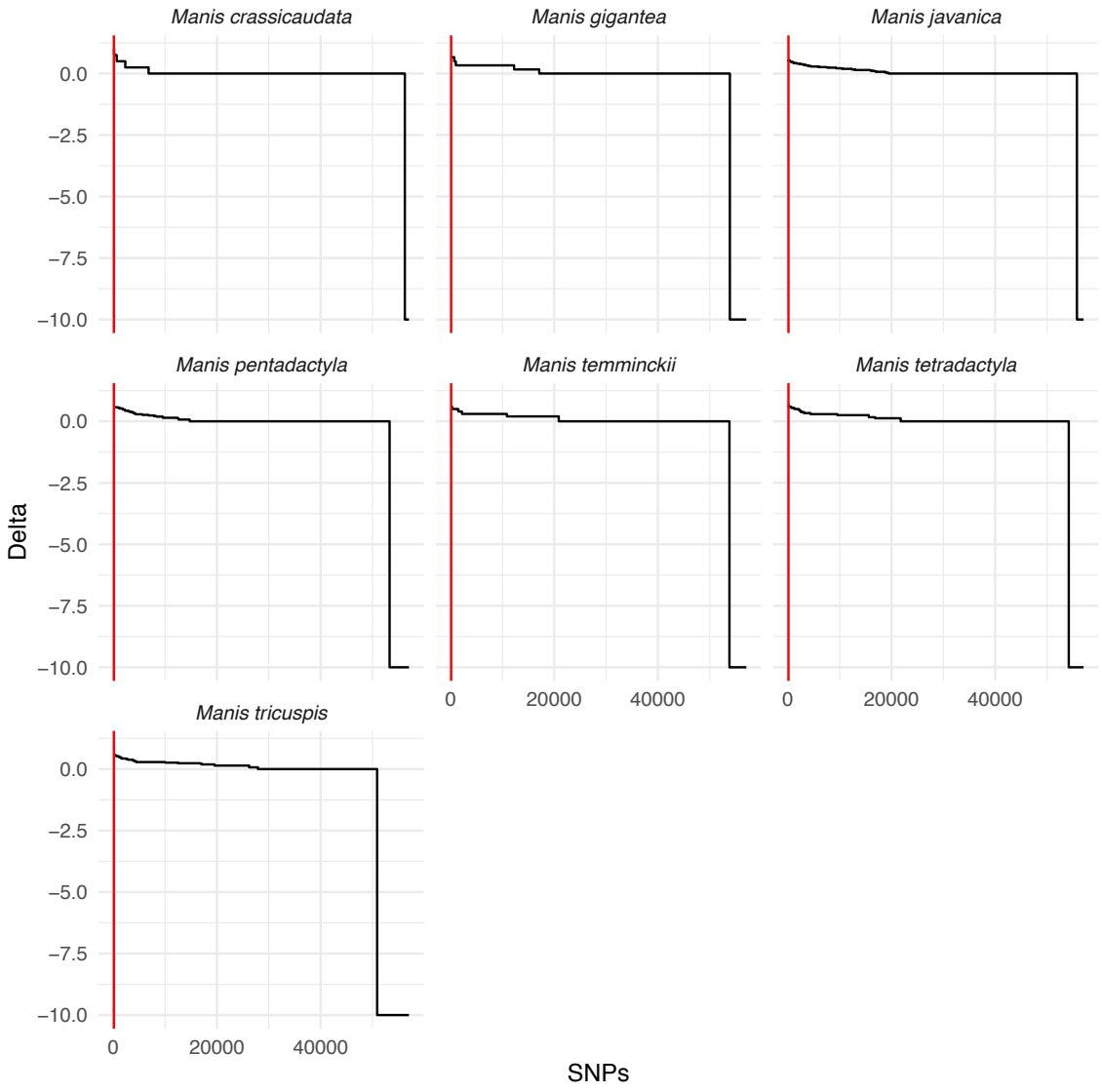

Figure 2.5: Delta scores of the SNPs for each species. The red vertical line denotes the cut off of the top 100 SNPs that were retained for the forensic SNP panel.

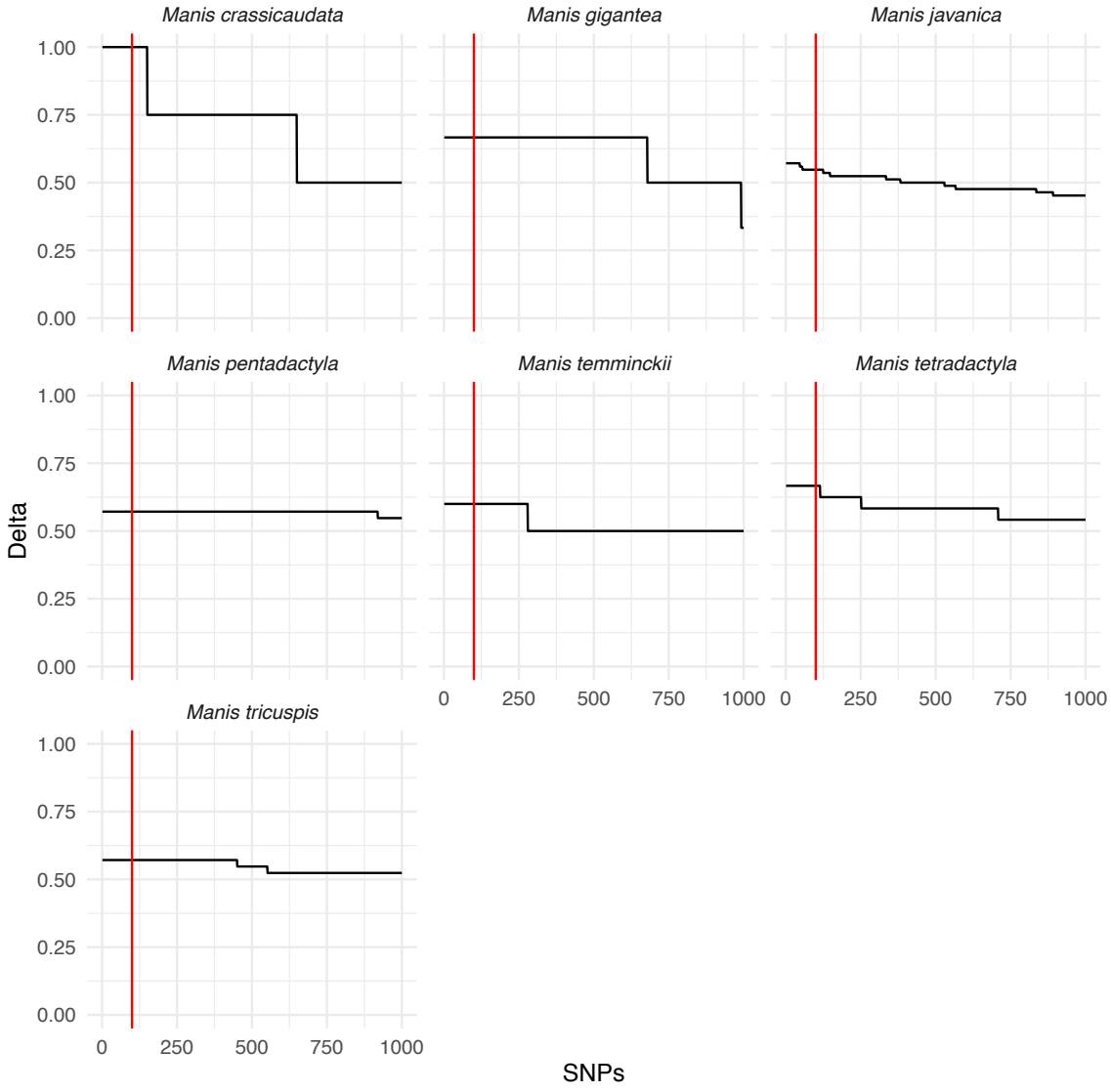

Figure 2.6: Delta scores of the top 1000 SNPs for each species. The red vertical line denotes the cut off of the top 100 SNPs that were retained for the forensic SNP panel.

3

Detection dog methodology to identify pangolin fecal samples

3.1 Abstract

The elusive nature of pangolins coupled with their rapidly declining populations have made detection and monitoring of pangolins in the wild challenging. Traditional methods of surveying mammals including camera trapping, transect surveys, and sign surveys have low detection rates or accuracy in pangolin species. Scat detection by dogs, specially trained to locate fecal samples of targeted endangered species, have been successfully used to collect samples of rare and wide-ranging species. In this chapter, scat detection dogs were trialed in a proof-of-concept study on Chinese and Indian pangolins (*Manis pentadactyla* and *M. crassicaudata*) to assess the feasibility of this survey method for genetic sample collection. Factors were identified that will aid future scat detection dog monitoring studies. Two detection dog teams trained on wild Chinese pangolin scat and surveyed five different regions of Nepal in 17 surveys covering 215 km over 175 sampling hours. Dogs alerted to pangolin scat in only three of these regions, for a total of 29 scats found in surveys. Preliminary genetic analysis was conducted to ensure specificity of the detection dog collected samples and to test extraction and amplification of genetic material from collected pangolin fecal samples. The two samples located by detection dogs and one training sample yielded amplifiable mitochondrial DNA and were confirmed as Chinese pangolin. The chapter details the detection dog methods for pangolin sample collection and identifies solutions to difficulties in acquiring training samples and lack of prior ecological knowledge of the pangolin populations.

3.2 Introduction

Pangolins, a group of scaly anteaters, have recently garnered the attention of both the public and scientific community by the combination of the magnitude of trade and the paucity of information on the group of species. More than a million individuals are estimated to have been poached in just one decade (Challender, Waterman, & Baillie, 2014). Despite the large volume of trade in pangolin meat and scales (Heinrich et al., 2016), there are no known estimates of population sizes, densities or habitat preferences. They have not been widely studied due to their elusive nature and low-density as a result of their population declines. This has resulted in gaps of knowledge on the ecology, habitat and behavior of all eight pangolins species, despite their endangered status.

Traditional methods of monitoring species such as camera traps, sign surveys, and transect surveys have yielded low detection rates (Willcox et al., 2019). This has resulted in gaps of knowledge on their ecology, habitat and behavior. The eight species of pangolins have been afforded protection at national and international levels, but have continued to be traded over the last few years. In addition, despite their status as endangered species, there is still a lack of information on pangolin biology, ecology, and behavior. The high detection probability of well-trained dogs can address these unique challenges by non-invasively locating scats to be genotyped, thereby providing needed data on genetic variation, habitat preferences, and other ecological parameters.

Since their inception as a tool for ecological and conservation studies in the late 1990's, scat detection dogs have been used for a wide range of species across various habitats (Mengüllüoğlu, Fickel, Hofer, & Förster, 2019; Rolland et al., 2007; Wasser et al., 2017). They have been successfully incorporated into study designs including mark capture recapture (Akenson, Henjum, Wertz, & Craddock, 2001; Arandjelovic et al., 2015; Cablk & Heaton, 2006), finding presence of species (Vynne et al., 2011), and studying wide ranging species (Becker et al., 2017).

This chapter describes the piloted use of detection dogs to locate fecal samples of Chinese pangolin (*Manis pentadactyla*) and Indian pangolin (*M. crassicaudata*) in Nepal. The aim was to assess the feasibility and value of using detection dogs for pangolins for genetic sample collection and outline lessons learned to

Table 3.1: Details of the detection dog teams.

Dog	Breed	Origin	Sex	Age	Weight (kg)	Dog experience (years)	Handler experience (years)
Athena	Cattle dog/Border Collie mix	Owner surrender	Female - neutered	2.5	18	1	13
Skye	Australian Kelpie mix	Rescue center	Female - neutered	3.0	17	1	5

inform future detection dog studies on pangolins.

3.3 Methods

3.3.1 Ethics statement

Sample collection methods were approved by the University of Washington's Institutional Animal Care and Use Committee (IACUC) under permit numbers 2850-08 and 2850-11.

3.3.2 Scat detectrion dog training

Detection dog teams from the University of Washington Center for Conservation Biology's Conservation Canines Facility in Eatonville, Washington, were used for this study. In depth protocol on detection dog selection and training is described in Wasser et al (2004).For this study, the handlers were selected based on the years of experience in detection dog work, while the detection dogs were selected to meet the 32 kg size limit on airlines carriers (Table 3.1). A short acclimation period (around two weeks) was followed by odor-specific training and hotspot training. The odor-specific training was conducted at the National Trust for Nature Conservation - Biodiversity Conservation Center, (NTNC-BCC) in Sauraha, Nepal, using one desiccated Chinese pangolin (*M. pentadactyla*) fecal sample. The fecal sample was placed in an open area with no other scats nearby and the handlers brought the dogs directly into the sample for a quick reward. Once the dogs were trained on the odor, hotspot training of the two detection dogs was conducted in Rani Community Forest, where there was a confirmed population of pangolins. Hotspot

training is conducted in areas where there is a relatively high concentration of wild samples from the target species and serves to focus the dog's alerts onto wild samples under natural conditions. For hotspot training, one team was sent to search a "hotspot" - an area with high potential for containing the target odor. The samples found by the first team were left on site, not disturbed or collected, and the second team followed directly after in order to have the second dog locate the same samples.

3.3.3 Survey sites

Permission to conduct detection dog work in Nepal was granted outside of national parks and buffer zones due to national legislation banning domesticated animals inside national parks. The survey sites were chosen in Community Forestry User Group (CFUG) and agricultural lands based on expert knowledge and available scientific literature as well as the area accessibility. The target species in Taplejung, Chitwan, Makwanpur, and Bhaktapur districts was Chinese pangolin and in Kanchanpur district was Indian pangolin (Figure 3.1). The distribution range of each species were derived from the digital Distribution Maps on The IUCN Red List of Threatened Species (IUCN, 2016). Prior to conducting surveys, permission was gained from the CFUG or residents to conduct field work. During this time, an informal interview was conducted to ask residents about pangolin sightings and burrows in the forest with a request to see the burrows for confirmation, where possible.

3.3.4 Detection dog surveys

During a four week period between March 3 and April 7, 2017, 31 detection dog surveys were conducted in five districts. The survey teams, at a minimum, consisted of a detection dog team (one dog and one handler) and a local guide, a wildlife field technician. The survey teams searched freely within a designated area with the dogs searching off-leash, but within sight of the handler. Each handler guided the dog to keep its searches downwind, when necessary, and to help focus the dog's attention on relevant signs (e.g. burrows, termite mounds). Dogs were also instructed to search each burrow or hole encountered during the survey.

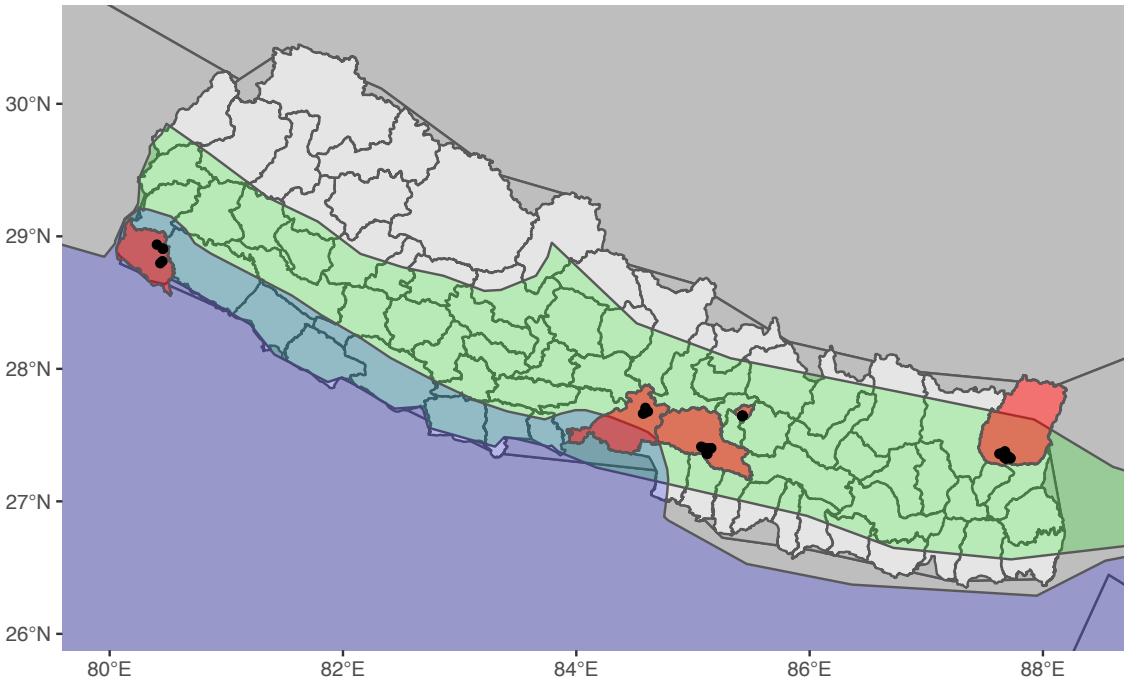

Figure 3.1: Detection dog survey locations in relation to published Chinese and Indian pangolin distribution range. Detection dog survets were conducted in five districts shown in red. The estimated Chinese pangolin distribution is shown in green and the Indian pangolin distribution is shown in blue.

3.3.5 Sample collection

When the dog located a fecal sample, the handler rewarded the dog with around 90 seconds of play with a toy and recorded data on the sample before collection. Survey routes were continuously recorded, in duplicate, along with sample location using Columbus V-900 Bluetooth Global Positioning System (GPS) Data Loggers. One location was captured by the continuously recording GPS data tracker worn by the dog and the other by the GPS data tracker held by the dog handler. Data pertaining to each scat collected were recorded on Android cell phones using the application Open Data Kit (ODK; full list of variables collected can be found in Supplementary table 3.3).

Depending on the field conditions, samples were either collected in Ziploc bags or swabbed using sterile cotton swabs in duplicate in the field and collected in Ziploc bags as backup. Sample collection via swabbing was conducted by placing a sterile cotton swab in phosphate buffered saline for five seconds then gently rolling the swab over the external surface of the fecal sample. The cotton swab was place in a 1.5 mL microcentrifuge tube with about 150 µL of ZR DNA Stabilizer solution (ZymoResearch). This was repeated to collect samples in duplicate. For collecting the entire fecal sample, the fecal sample was placed in a Ziploc bag. Samples were kept cold during transfer and stored at -20 C once transferred to the laboratory.

3.3.6 Genetic analysis

The genetic analysis was conducted in the molecular biology laboratory at NTNC-BCC. Three fecal samples, one training sample collected by a person and two samples located by detection dogs, were extracted in duplicate using a modified ZR Fecal DNA MiniPrep Kit protocol (ZymoResearch). In lieu of fecal matter, swabs were used as the starting material for extraction. The sample was vortexed for a minimum of 20 min on a desktop vortex at the ZR BashingBead Lysis step to ensure complete lysis. A 402 bp fragment of the mitochondrial DNA cytochrome b gene was amplified following protocol from Hsieh et al. (2001) and sequenced at Nepal Academy of Science and Technology (NAST), Kathmandu, Nepal. The forward and reverse sequences were assembled and aligned using Geneious and the resulting consen-

sus sequence was identified to species using NCBI Nucleotide BLAST.

3.3.7 Survey analysis

The tracklogs from the dog's GPS trackers were analyzed with the R package trackeR v1.5.2 using R version 3.6.3 (2020-02-29). In the two instances where the dog's tracklogs did not collect data, the handler's tracklog was used in its place.

3.4 Results

The two detection dogs were trained on the odor of Chinese pangolin scats using one wild fecal sample. Although three putative Chinese pangolin fecal samples were found for training, only one sample gave high enough confidence as a pangolin sample to be used due for training in the absence of timely genetic confirmation. This sample was found adjacent to a fresh pangolin burrow in Rani Community Forest by a forest guard and accepted as a pangolin scat based on the experience and knowledge of the forest guard in addition to the confirmed presence of pangolins in the forest.

The dogs' ability to locate Chinese pangolin fecal samples was subsequently confirmed during hotspot training where the detection dogs alerted to two separate fecal samples in the forest. In addition, the training samples and two fecal samples collected by the detection dog teams were genetically confirmed by a 402 bp fragment of the cytochrome b gene. The sequences varied in length due to quality and an alignment of 315 bp was recovered, which formed one haplotype. The haplotype had a >95% identify match with *M. pentadactyla* when compared to other COI sequences using Nucleotide BLAST.

Detection dog surveys were conducted over 17 survey days covering 215 km over 175 sampling hours. The tracklogs of each survey and the location of the fecal sample in relationship to the survey route is shown in Figure 3.2. The surveys included 16 surveys by detection dog, Athena, totaling 98 km, and 15 surveys by detection dog, Skye, totaling 117 km.

Thirty-two high confidence pangolin fecal samples were collected across three districts. Three surveys were conducted in Kanchanpur district in Nepal that historically harbored Indian pangolins, but did not

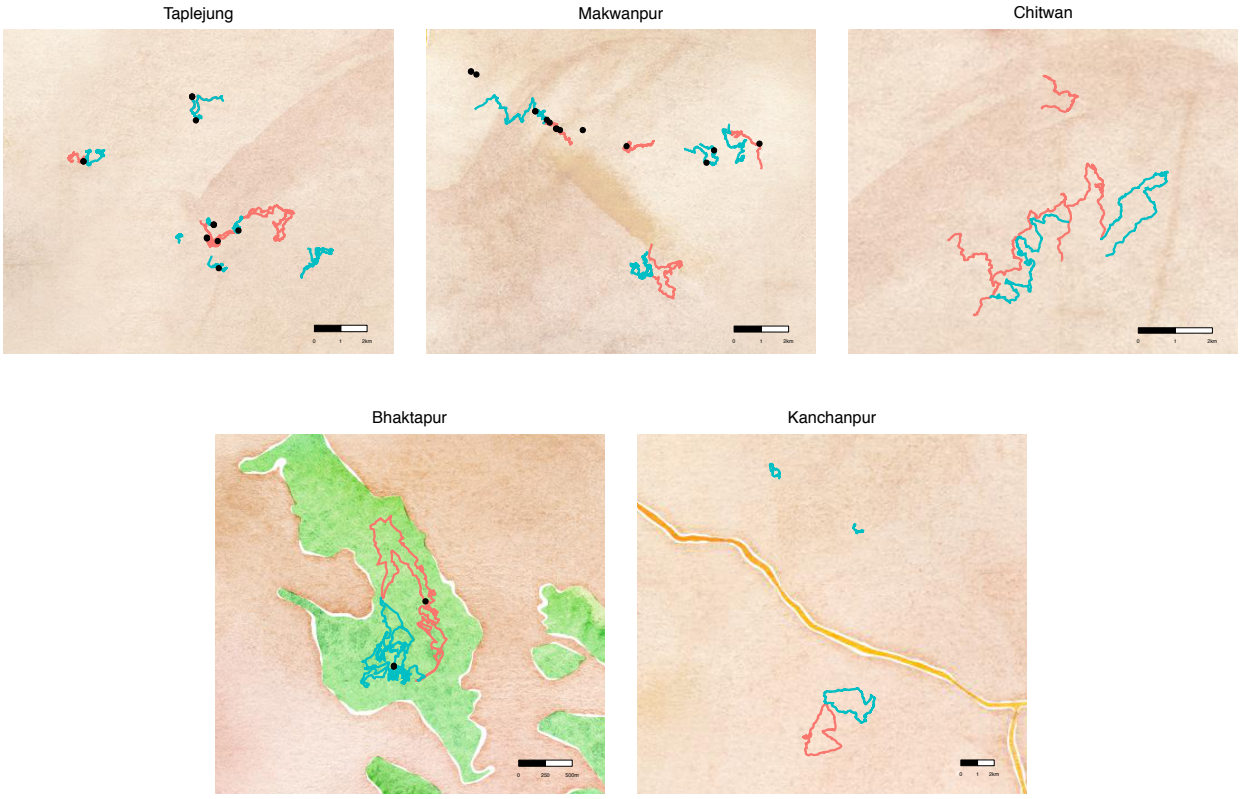

Figure 3.2: Location of collected fecal samples are shown as black dots and the detection dog team track logs are shown in red for detection dog Athena and blue for detection dog Skye. Scale of each map is different and shown as a bar in each figure. Maps are rendered in low resolution without coordinates to not disclose location of pangolin samples.

Table 3.2: Total number of fecal samples found during surveys by type and detection dog team for Chinese pangolin in Nepal

Type	Located by	Dog	Count	Buried	Near burrow
training	forest guard	-	1	0% (0/1)	100% (1/1)
hotspot	detection dog team	-	2	0% (0/2)	0% (0/2)
survey	detection dog team	Skye	13	38% (5/13)	62% (8/13)
survey	detetction dog team	Athena	13	62% (8/13)	77% (10/13)
survey	survey team	-	3	0% (0/3)	100% (3/3)

Figure 3.3: Each pangolin fecal samples was genetically confirmed as *M. pentadactyla* and were found close to a burrows. First column: Sample NS04 located by detection dog Skye buried in old termite mound near a burrow, second column: Sample NA06 located by detection dog Athena near a burrow, third column; Sample NA02 located by a person next to a burrow.

result in finding fecal samples. Three of the collected samples were located by people as putative training samples and two samples were collected by detection dogs as part of the hotspot training. The remaining 28 samples were collected during surveys. Twenty-six of the 28 samples were found by detection dogs (0.12 scats/km for detection dogs) and two were located by people on the survey teams (Table 3.2). The pangolin fecal samples consisted largely of soil with almost no visible ants and/or termite carapaces to the naked eye (Figure 3.3).

Thirteen of the 26 fecal samples located by detection dogs during surveys were buried under a layer of soil (5+ cm of soil) and not visible to humans. Eighteen of the 26 samples were found next to or inside the opening of a visible pangolin burrow. While each pangolin burrow was checked by dogs, they were not explored further to minimize potential disturbance to live pangolins and the lack of a feasible method to search the burrows for fecal samples. The fecal samples were found at elevations ranging from 427 m in Makwanpur district up to 1705 m in Taplejung district.

3.5 Discussion

Detecting and monitoring pangolins has become increasingly difficult due to their elusive nature and a low density across a wide distribution. For such species, detection dogs can provide a viable solution for locating individuals or fecal samples for non-invasive sampling of the species. Detection dogs have been employed in species that are difficult to detect using traditional methods including spotted owls, Javan rhinos, bats, and giant armadillos (Brook et al., 2012; Chambers, Vojta, Mering, & Davenport, 2015; Silveira et al., 2009; Wasser et al., 2012), but have not yet been trialed for pangolins species. Here, we used scat detection dogs in Nepal to present a proof-of-concept study on the feasibility of using detection dogs for Chinese and Indian pangolin fecal sample collection.

When training detection dogs for a scat collection project, samples from multiple individuals from different areas are used to train the dogs to the species (Long, Donovan, MacKay, Zielinski, & Buzas, 2007; Orkin, 2016; Wasser, Davenport, et al., 2004). The variety of samples ensures that the dogs can generalize the odor to the target species and not to a particular scent of a single individual. This poses difficulties for

endangered species where fecal samples are sparse and especially for pangolins which are not common in zoos or in captivity. Having no genetically confirmed training samples, confirmation by local residents was used as a substitute to identify pangolin scats and the training was limited to a single scat. The inclusion of hotspot training, training in a location with high potential for pangolin fecal samples, allowed the handlers to confirm the detection dogs' ability to detect the target species, and lock in the scent of wild pangolins, helping the detection dogs to generalize the scent to other individuals of Chinese pangolins.

Detection dog studies report varying detection distances depending on environmental conditions. In a terrestrial study, dogs were able to detect >75% of scats located within 10 m and in a marine study up to 1.8 km (Reed, Bidlack, Hurt, & Getz, 2011; Wasser et al., 2017). Chinese pangolin fecal sample detection distance is likely to be smaller than reported distances due to the target being buried about 50% of the time. The buried fecal samples lead to a target that less odiferous than fecal samples of other species. Short and confined survey tracks focused to areas with burrows maximized possibility of detection and refined searches were necessary to help dogs work into smaller scent cones. Chinese pangolin burrows were helpful in narrowing down the search area and guide the dogs into a tighter and thorough search.

Targeted search are possible when prior information on the presence of a species and their population densities are known or where reliable signs of the animal is known to gather information. For many species of pangolins, this information is not available. Only a few fossorial species leave distinct signs such as burrows and tracks for easy identification to anchor detection dog surveys (Prakash, Nepali, Khatiwada, Shambhu, & others, 2013; Willcox et al., 2019). Physical signs of arboreal pangolins such as claw marks, and tree holes are difficult to identify and pangolin habitat preferences are lacking (Willcox et al., 2019). The residents of the study area were able to provide information on the pangolin's presence to guide the targeted searches. Local ecological knowledge has been used to glean information on pangolin presence to target camera trapping and assess spatiotemporal dynamics (Simo et al., 2020; Zanvo et al., 2020). Similarly, this knowledge can fill the gap in information on pangolin populations and may provide a solution to targeting survey sites for pangolin detection dog studies.

This pilot study shows promise of using detection dogs for studying sample location of Chinese pangolins for presence and population level monitoring. It highlights possible ways to overcome factors that can

limit the efficacy of detection dog surveys. The lack of a wide variety of training samples can be mitigated by inclusion of a "hotspot" with a high probability of locating samples of the target species. This chapter also demonstrates how alternative methods can be substituted to overcome the lack of ecological knowledge for studying highly reclusive, endangered species.

3.6 Acknowledgements

This work was supported by the USAID Wildlife Crime Tech Challenge and the Mohamed bin Zayed Species Conservation Fund. Field work in Nepal was conducted in collaboration with the National Trust for Nature Conservation and the support of wildlife technicians at the NTNC-Biodiversity Conservation Center. A special thank you goes to Raju and Bishnu. The work would not have been possible without the help and support of the local residents and CFUG users that helped us conduct detection in their farms and forests.

3.7 Supplementary materials

Table 3.3: Variables collected by the detection dog teams for each fecal sample and their description.

Collected variable	Details
Sample ID	Sample ID numbered sequentially for each fecal sample
Start	Time and date of sample collection
Latitude	Collected by GPS
Longitude	Collected by GPS
Altitude	Collected by GPS
Accuracy	Collected by GPS
Image_Scat	Picture of fecal sample
Dog Rewarded	Yes/no on whether dog was rewarded for finding sample
Handler Gestalt	Low/medium/high on the handler's certainty of target species
Dog_Gestalt	Low/medium/high on the dog's
Scat_Condition	Old/fresh/dry/moist to describe the state of the fecal sample
Fungus	Yes/no on whether fungus was found on sample
Species	Host species of the collected fecal sample
Habitat	Mixed forest/deciduous forest/agricultural land/grassland/other to describe immediate surrounding of fecal sample
Burrow	Yes/no to indicate presence of a nearby burrow
Image_Burrow	If burrow was found, picture of the burrow
Water_Source	Yes/no to indicate presence of a nearby water source
Water_Feature	If a nearby water source was found, description of the water souce
Other_Scat_Present	Yes/No and species to indicate if other species' fecal samples were found
Notes	Additional notes regarding the fecal sample or dog's reaction to sample

References

Akenson, J. J., Henjum, M. G., Wertz, T. L., & Craddock, T. J. (2001). Use of dogs and mark-recapture techniques to estimate American black bear density in northeastern Oregon. *Ursus*, 203–209.

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources*, *20*(4), 892–905.

Anagnostou, M. (2021). Synthesizing knowledge on crime convergence and the illegal wildlife trade. *Environmental Challenges*, 100222.

Anonymous. (1938). Chinese medicine and the pangolin. *Nature*, 1–1.

Arandjelovic, M., Bergl, R. A., Ikfuingei, R., Jameson, C., Parker, M., & Vigilant, L. (2015). Detection dog efficacy for collecting faecal samples from the critically endangered Cross River gorilla (*Gorilla gorilla diehli*) for genetic censusing. *Royal Society Open Science*, *2*(2), 140423–14.

Axelsson, E., Willerslev, E., Gilbert, M. T. P., & Nielsen, R. (2008). The effect of ancient DNA damage on inferences of demographic histories. *Molecular Biology and Evolution*, *25*(10), 2181–2187.

Becker, M. S., Durant, S. M., Watson, F. G., Parker, M., Gottelli, D., M'soka, J., ... others. (2017). Using dogs to find cats: Detection dogs as a survey method for wide-ranging cheetah. *Journal of Zoology*, *302*(3), 184–192.

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (*Homarus americanus*). *Molecular Ecology*, *24*(13), 3299–3315.

Besnard, G., Bertrand, J. A., Delahaie, B., Bourgeois, Y. X., Lhuillier, E., & Thébaud, C. (2016). Valuing museum specimens: High-throughput DNA sequencing on historical collections of New Guinea crowned pigeons (Goura). *Biological Journal of the Linnean Society*, *117*(1), 71–82.

Branch, W. R., Haacke, W., Pinto, P. V., Conradie, W., Verburgt, L., & Baptista, N. (2017). Loveridge's angolan geckos, afroedura karroica bogerti and pachydactylus scutatus angolensis (sauria, gekkonidae): New distribution records, comments on type localities and taxonomic status.

Brook, S. M., Coeverden de Groot, P. van, Scott, C., Boag, P., Long, B., Ley, R. E., ... Hai, B. T. (2012). Integrated and novel survey methods for rhinoceros populations confirm the extinction of *Rhinoceros sondaicus annamiticus* from Vietnam. *Biological Conservation*, *155*(C), 59–67.

Brüniche-Olsen, A., Kellner, K. F., Anderson, C. J., & DeWoody, J. A. (2018). Runs of homozygosity have utility in mammalian conservation and evolutionary studies. *Conservation Genetics*, *19*(6), 1295–1307.

Cablk, M. E., & Heaton, J. S. (2006). Accuracy and reliability of dogs in surveying for desert tortoise (*Gopherus agassizii*). *Ecological Applications : A Publication of the Ecological Society of America*, *16*(5), 1926–1935.

Ceballos, F. C., Hazelhurst, S., & Ramsay, M. (2018). Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data, 1–12.

Challender, D. W., Waterman, C., & Baillie, J. E. (2014). Scaling up pangolin conservation. *IUCN SSC Pangolin Specialist Group Conservation Action Plan. Zoological Society of London, London*.

Chambers, C. L., Vojta, C. D., Mering, E. D., & Davenport, B. (2015). Efficacy of scent-detection dogs for locating bat roosts in trees and snags. *Wildlife Society Bulletin*, *39*(4), 780–787.

Ding, L., Wiener, H., Abebe, T., Altaye, M., Go, R. C., Kercsmar, C., ... Chakraborty, R. (2011). Comparison of measures of marker informativeness for ancestry and admixture mapping. *BMC Genomics*, *12*(1), 622.

Díez-del-Molino, D., Sánchez-Barreiro, F., Barnes, I., Gilbert, M. T. P., & Dalén, L. (2018). Quantifying temporal genomic erosion in endangered species. *Trends in Ecology & Evolution*, *33*(3), 176–185.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., ... others. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95.

Finnegan, M. T. V., Herbert, K. E., Evans, M. D., Griffiths, H. R., & Lunec, J. (1996). Evidence for sensitisation of dna to oxidative damage during isolation. *Free Radical Biology and Medicine*, *20*(1), 93–98. http://doi.org/https://doi.org/10.1016/0891-5849(95)02003-9

Freamo, H., O'Reilly, P., Berg, P. R., Lien, S., & Boulding, E. G. (2011). Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources*, *11*, 254–267.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv Preprint*.

Gaubert, P., Njiokou, F., Ngua, G., Afiademanyo, K., Dufour, S., Malekani, J., ... others. (2016). Phylogeography of the heavily poached African common pangolin (Pholidota, *Manis tricuspis*) reveals six cryptic lineages as traceable signatures of Pleistocene diversification. *Molecular Ecology*, *25*(23), 5975–5993.

Hedrick, P. W., & Kalinowski, S. T. (2000). Inbreeding depression in conservation biology. *Annual Review of Ecology and Systematics*, *31*(1), 139–162.

Heinrich, S., Wittmann, T. A., Prowse, T. A., Ross, J. V., Delean, S., Shepherd, C. R., & Cassey, P. (2016). Where did all the pangolins go? International CITES trade in pangolin species. *Global Ecology and Conservation*, *8*, 241–253.

Hsieh, H.-M., Lee, J. C.-I., Wu, J.-H., Chen, C.-A., Chen, Y.-J., Wang, G.-B., ... Tsai, L.-C. (2011). Establishing the pangolin mitochondrial D-loop sequences from the confiscated scales. *Forensic Science International: Genetics*, *5*(4), 303–307.

Hu, J.-Y., Hao, Z.-Q., Frantz, L., Wu, S.-F., Chen, W., Jiang, Y.-F., ... others. (2020). Genomic consequences of population decline in critically endangered pangolins and their demographic histories. *National Science Review*, *7*(4), 798–814.

IPBES. (2019). *Global assessment report of the intergovernmental science-policy platform on biodiversity and ecosystem services* (p. 1753). Book, Bonn, Germany: IPBES Secretariat. http://doi.org/10.5281/zenodo.3831673

Jühling, F., Pütz, J., Bernt, M., Donath, A., Middendorf, M., Florentz, C., & Stadler, P. F. (2012). Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids Research*, *40*(7), 2833–2845.

Kavakiotis, I., Triantafyllidis, A., Ntelidou, D., Alexandri, P., Megens, H.-J., Crooijmans, R. P. M. A., ... Vlahavas, I. (2015). TRES: identification of discriminatory and informative SNPs from population genomic data. *Journal of Heredity*, *106*(5), 672–676.

Keighley, X., Bro-Jørgensen, M. H., Ahlgren, H., Szpak, P., Ciucani, M. M., Sánchez Barreiro, F., ... Olsen, M. T. (2021). Predicting sample success for large-scale ancient DNA studies on marine mammals. *Molecular Ecology Resources*, *21*(4), 1149–1166.

Kennedy, S. R., Schmitt, M. W., Fox, E. J., Kohrn, B. F., Salk, J. J., Ahn, E. H., ... Loeb, L. A. (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. *Nature Protocols*, *9*(11), 2586–2606.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*(1), 1–13.

Lenzen, M., Moran, D., Kanemoto, K., Foran, B., Lobefaro, L., & Geschke, A. (2012). International trade drives biodiversity threats in developing nations. *Nature*, *486*(7401), 109–112.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., ... Lam, T.-W. (2016). MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, *102*, 3–11.

Li, H. (2011). Improving SNP discovery by base alignment quality. *Bioinformatics*, *27*(8), 1157–1158.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Preprint arXiv:1303.3997*.

Long, R. A., Donovan, T. M., MacKay, P., Zielinski, W. J., & Buzas, J. S. (2007). Effectiveness of scat detection dogs for detecting forest carnivores. *Journal of Wildlife Management*, *71*(6), 2007–2017.

Luiselli, L., Politano, E., & Angelici, F. M. (2000). Ecological correlates of the distribution of terrestrial and freshwater chelonians in the niger delta, nigeria: A biodiversity assessment with conservation implications. *Revue d'écologie*.

Malbrant, R., Maclatchy, A., & others. (1949). Faune de l'équateur africain français.

Mengüllüoğlu, D., Fickel, J., Hofer, H., & Förster, D. W. (2019). Non-invasive faecal sampling reveals spatial organization and improves measures of genetic diversity for the conservation assessment of territorial species: Caucasian lynx as a case species. *PLoS ONE*, *14*(5), e0216549–20.

Moore, J.-S., Bourret, V., Dionne, M., Bradbury, I., O'Reilly, P., Kent, M., ... Bernatchez, L. (2014). Conservation genomics of anadromous Atlantic salmon across its North American range: Outlier loci identify the same patterns of population structure as neutral loci. *Molecular Ecology*, *23*(23), 5680–5697.

Morin, P. A., Martien, K. K., & Taylor, B. L. (2009). Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, *9*(1), 66–73.

Morton, O., Scheffers, B. R., Haugaasen, T., & Edwards, D. P. (2021). Impacts of wildlife trade on terrestrial biodiversity. *Nature Ecology & Evolution*, *5*(4), 540–548.

Mwale, M., Dalton, D. L., Jansen, R., De Bruyn, M., Pietersen, D., Mokgokong, P. S., & Kotzé, A. (2017).

Forensic application of DNA barcoding for identification of illegally traded African pangolin scales. *Genome*, *60*(3), 272–284.

Nakahama, N. (2021). Museum specimens: An overlooked and valuable material for conservation genetics. *Ecological Research*, *36*(1), 13–23.

Nash, H. C., Low, G. W., Choo, S. W., Chong, J. L., Semiadi, G., Hari, R., ... others. (2018). Conservation genomics reveals possible illegal trade routes and admixture across pangolin lineages in Southeast Asia. *Conservation Genetics*, *19*(5), 1083–1095.

Nash, H. C., Wong, M. H., & Turvey, S. T. (2016). Using local ecological knowledge to determine status and threats of the Critically Endangered Chinese pangolin em(Manis pentadactyla) in Hainan, China. *Biological Conservation*, *196*, 189–195.

Newton, P., Van Thai, N., Roberton, S., & Bell, D. (2008). Pangolins in peril: Using local hunters' knowledge to conserve elusive species in Vietnam. *Endangered Species Research*, *6*(1), 41–53.

Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*, *32*(2), 292–294.

Orkin, J. D. (2016). Cost-effective scat-detection dogs: unleashing a powerful new tool for international mammalian conservation biology. *Nature Publishing Group*, 1–10.

Picard toolkit. (2019). *Broad Institute, GitHub repository*. http://broadinstitute.github.io/picard/; Broad Institute.

Prakash, T., Nepali, S., Khatiwada, A., Shambhu, P., & others. (2013). Distribution and conservation status of Chinese pangolin in Nangkholang VDC, Taplejung, Eastern Nepal. *Tigerpaper*, *40*(3), 22–27.

Puckett, E. E., & Eggert, L. S. (2015). Comparison of SNP and microsatellite genotyping panels for spatial assignment of individuals to natal range: A case study using the American black bear (*Ursus americanus*). *Biological Conservation*, *193*(C), 86–93.

Reed, S. E., Bidlack, A. L., Hurt, A., & Getz, W. M. (2011). Detection distance and environmental factors in conservation detection dog surveys. *The Journal of Wildlife Management*, *75*(1), 243–251.

Rich, J. (2012). *Missing links: The african and american worlds of rl garner, primate collector*. University of Georgia Press.

Robinson, J. A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., Marsden, C. D., Lohmueller, K. E., ... others. (2016). Genomic flatlining in the endangered island fox. *Current Biology*, *26*(9), 1183–1189.

Rolland, R. M., Hamilton, P. K., Kraus, S. D., Davenport, B., Gillett, R. M., & Wasser, S. K. (2007). Faecal sampling using detection dogs to study reproduction and health in North Atlantic right whales (*Eubalaena glacialis*). *Journal of Cetacean Research and Management*, *8*(2), 121.

Rosenberg, N. A., Li, L. M., Ward, R., & Pritchard, J. K. (2003). Informativeness of Genetic Markers for Inference of Ancestry. *The American Journal of Human Genetics*, *73*(6), 1402–1422.

Schaeffer, F., Kolb, A., & Buc, H. (1982). Point mutations change the thermal denaturation profile of a short DNA. *The EMBO Journal*, *1*(1), 99–105.

Scheffers, B. R., Oliveira, B. F., Lamb, I., & Edwards, D. P. (2019). Global wildlife trade across the tree of life. *Science*, *366*(6461), 71–76.

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., ... others. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols*, *9*(5), 1056–1082.

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*(1), 1–7.

Shi, Y., Chen, J., Su, J., Zhang, T., & Wasser, S. K. (2019). Genetic resilience of a once endangered species, Tibetan antelope *(Pantholops hodgsonii). bioRxiv*, 628727.

Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R., & Ferrell, R. E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Ge-*

netics, *60*(4), 957.

Silveira, L., Almeida Jácomo, A. T. de, Furtado, M. M., Torres, N. M., Sollmann, R., & Vynne, C. (2009). Ecology of the Giant Armadillo (*Priodontes maximus*) in the Grasslands of Central Brazil. *Edentata*, *8-10*(10), 25–34.

Simo, F., Difouo Fopa, G., Kekeunou, S., Ichu, I. G., Esong Ebong, L., Olson, D., & Ingram, D. J. (2020). Using local ecological knowledge to improve the effectiveness of detecting white-bellied pangolins (*Phataginus tricuspis*) using camera traps: A case study from Deng-Deng national park, cameroon. *African Journal of Ecology*, *58*(4), 879–884.

Sun, N. C.-M., Chang, S.-P., Lin, J.-S., Tseng, Y.-W., Pei, K. J.-C., & Hung, K.-H. (2020). The genetic structure and mating system of a recovered Chinese pangolin population (*Manis pentadactyla* Linnaeus, 1758) as inferred by microsatellite markers. *Global Ecology and Conservation*, *23*, e01195.

Svaren, J., Inagami, S., Lovegren, E., & Chalkley, R. (1987). DNA denatures upon drying after ethanol precipitation. *Nucleic Acids Research*, *15*(21), 8739–8754.

The iucn red list of threatened species. Https://www.iucnredlist.org. Downloaded on june 2016. (n.d.).

UNODC. (2020). World wildlife crime report 2020: Trafficking in protected species.

Vynne, C., Skalski, J. R., Machado, R. B., Groom, M. J., Jácomo, A. T. A., Marinho-Filho, J., ... Wasser, S. K. (2011). Effectiveness of scat-detection dogs in determining species presence in a tropical savanna landscape. *Conservation Biology : The Journal of the Society for Conservation Biology*, *25*(1), 154–162.

Wasser, S. K., Brown, L., Mailand, C., Mondol, S., Clark, W., Laurie, C., & Weir, B. (2015). Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots. *Science*, *349*(6243), 84–87.

Wasser, S. K., Davenport, B., Ramage, E. R., Hunt, K. E., Parker, M., Clarke, C., & Stenhouse, G. (2004). Scat detection dogs in wildlife research and management: application to grizzly and black bears in the Yellowhead Ecosystem, Alberta, Canada. *Canadian Journal of Zoology*, *82*(3), 475–492.

Wasser, S. K., Hayward, L. S., Hartman, J., Booth, R. K., Broms, K., Berg, J., ... Smith, H. (2012). Using Detection Dogs to Conduct Simultaneous Surveys of Northern Spotted (*Strix occidentalis caurina*) and Barred Owls (*Strix varia*). *PLoS ONE*, *7*(8), e42892–8.

Wasser, S. K., Lundin, J. I., Ayres, K., Seely, E., Giles, D., Balcomb, K., ... Booth, R. (2017). Population growth is limited by nutritional impacts on pregnancy success in endangered Southern Resident killer whales (*Orcinus orca*). *PLoS One*, *12*(6), e0179824.

Wasser, S. K., Mailand, C., Booth, R., Mutayoba, B., Kisamo, E., Clark, B., & Stephens, M. (2007). Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proceedings of the National Academy of Sciences*, *104*(10), 4228–4233.

Wasser, S. K., Shedlock, A. M., Comstock, K., Ostrander, E. A., Mutayoba, B., & Stephens, M. (2004). Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. *Proceedings of the National Academy of Sciences*, *101*(41), 14847–14852.

Willcox, D., Nash, H. C., Trageser, S., Kim, H. J., Hywood, L., Connelly, E., ... others. (2019). Evaluating methods for detecting and monitoring pangolin (Pholidata: Manidae) populations. *Global Ecology and Conservation*, *17*, e00539.

Willing, E.-M., Dreyer, C., & Van Oosterhout, C. (2012). Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers.

Willoughby, J. R., Sundaram, M., Wijayawardena, B. K., Lamb, M. C., Kimble, S. J., Ji, Y., ... Dewoody, J. A. (2017). Biome and migratory behaviour significantly influence vertebrate genetic diversity. *Biological Journal of the Linnean Society*, *121*(2), 446–457.

Wyatt, T., Uhm, D. van, & Nurse, A. (2020). Differentiating criminal networks in the illegal wildlife trade: Organized, corporate and disorganized crime. *Trends in Organized Crime*, *23*, 350–366.

Zanvo, S., Gaubert, P., Djagoun, C. A., Azihou, A. F., Djossa, B., & Sinsin, B. (2020). Assessing the spatiotemporal dynamics of endangered mammals through local ecological knowledge combined with direct evidence: The case of pangolins in Benin (West Africa). *Global Ecology and Conservation*, *23*,

e01085.

Zhang, H., Ades, G., Miller, M. P., Yang, F., Lai, K.-w., & Fischer, G. A. (2020). Genetic identification of African pangolins and their origin in illegal trade. *Global Ecology and Conservation*, *23*, e01119.

Zimmermann, J., Hajibabaei, M., Blackburn, D. C., Hanken, J., Cantin, E., Posfai, J., & Evans, T. C. (2008). DNA damage in preserved specimens and tissue samples: A molecular assessment. *Frontiers in Zoology*, *5*(1), 1–13.