2020

Genetic Analysis of Snow Leopard Population Employing Next Generation Sequencing For Its Improved Conservation And Management

Safia Janjua
Wright State University

GENETIC ANALYSIS OF SNOW LEOPARD POPULATION EMPLOYING NEXT GENERATION SEQUENCING FOR ITS IMPROVED CONSERVATION AND MANAGEMENT

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

By

SAFIA JANJUA
MPhil, Biochemistry, PMAS-Arid Agriculture University Rawalpindi, Pakistan, 2010
B.S., University of the Punjab, Pakistan, 2005

2020
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

July 29th, 2020

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Safia Janjua ENTITLED Genetic Analysis of Snow Leopard Population Employing Next Generation Sequencing For Its Improved Conservation And Management BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Thomas P. Rooney, Ph.D.
Dissertation Director

Don Cipollini, Ph.D.
Director, Environmental Sciences
Ph.D. Program

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

Committee on Final Examination:

Jeffrey L. Peters, Ph.D.

Fakhar-i-Abbas, Ph.D.

Byron Weckworth, Ph.D.

Volker Bahn, Ph.D.

Abstract

Janjua, Safia. Ph.D., Environmental Sciences Ph.D. Program, Wright State University, 2020. Genetic Analysis of Snow Leopard Population Employing Next Generation Sequencing For Its Improved Conservation And Management.

Snow leopards (*Panthera uncia*) are an enigmatic, high-altitude species whose challenging habitat, low population densities and patchy distribution have presented challenges for scientists studying its biology, population structure, and genetics. To address these important ecological, conservation, and evolutionary questions, scientists are tailoring laboratory and computational methods to better extract the information from non-invasive samples, only available source of DNA for this species. These samples with very low quantity and quality of DNA, present unique methodological challenges. ddRAD-seq, one of next generation sequencing method is used here to develop reference sequence library for snow leopard using five blood samples from Mongolian snow leopards. 697 SNPs are identified through this method. This genetic data from ddRAD-seq will be invaluable for conducting population and landscape scale studies that can inform snow leopard conservation strategies. Then probes are designed for target DNA capture, a widely used method for studying low quality and quantity of DNA from ancient DNA samples, eDNA, and forensics, using developed ddRAD-seq reference sequence library. Non-invasive fecal scats of snow leopards from seven different countries are used for target DNA capture. In addition to target DNA, high number of non-targeted mtDNA of snow leopard and prey species are also obtained. This non-targeted DNA is used to identify prey species in snow leopard scats that are collected from different regions/locations. 3369 bp of snow leopard mtDNA are used to identify 22 parsimony informative sites that can be useful for future mitochondrial gene-based population genetics and structure studies of snow leopards.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2

Chapter 3

LIST OF TABLES

Chapter 1

Chapter 3

Chapter 4

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisors Dr. Tom Rooney and Dr Jeffrey L Peters for their support and guidance throughout this project. I would also like to thank my committee members, Dr. Fakhar-i-Abbas, Dr Byron Weckworth and Dr. Volker Bahn for feedbacks and encouragement. I must also thank Dr. Philip Lavretsky, Dr. Michael L. Raymer, SoonJye Kho and Clayton P. Allex-Buckner for helping me in data analysis. I am thankful to Panthera for funding this project through Sabin Snow Leopard grant and Dr Byron for helping me get the samples from different countries. I would like to thank Centre of Bioresource Research (CBR) team for providing samples from Pakistan. Thanks to Dr Afsar Mian for continuous feedback on my writing and encouragement throughout my PhD.

My special thanks to Kendra Millam for her unconditional support, I found a good friend and beautiful human being in her. Thanks to Sara Seibert and other lab fellows for friendly support. Finally, my sincere gratitude to my Mother, my husband Sharjeel, my siblings, other family members and friends for support during this time. Last but not the least big thanks and warm hugs to my kids, Abiha and Ashas, their cheeky smiles and love is source of refreshment during this time.

OVERVIEW

The enigmatic snow leopard (*Panthera uncia*) occupies particularly inaccessible mountainous habitat in central Asia, covering an area of more than 1.6 million km^2 across 12 countries including Afghanistan, Bhutan, China, India, Kyrgyzstan, Kazakhstan, Nepal, Mongolia, Pakistan, Russia, Tajikistan and Uzbekistan (Nowell and Jackson 1996). Despite its high profile as a charismatic carnivore, snow leopard information is scarce and difficult to obtain due to its cryptic nature and remote habitat. As a result, reliable information is lacking on the number of snow leopards remaining, locations of peripheral and core populations, and areas where they are in decline. However, such knowledge is critical for the conservation of this flagship species. Considerable effort is underway to determine the status of the species across its extensive and highly fragmented range (Ale *et al*. 2007; Jackson *et al*. 2006; Janecka *et al*. 2008; Lovari *et al*. 2009; McCarthy *et al*. 2008; Xu *et al*. 2008).

A handful of studies are available on snow leopard genetics. The first reported genetic study was of Zhang *et al*. (2007), which involved development of snow leopard-specific primers to amplify a section of the control region (CR) in the mitochondrial genome. This study used the CR sequence to demonstrate a phylogenetic relationship of snow leopard and other *Panthera* species. Later the full mitochondrial genome sequence was reported by Wei *et al*. (2009), which motivated scientists to study the phylogeography and evolutionary pattern of snow leopard using different mitochondrial regions. One such study (Davis *et al.,* 2010) involves a combination of both nuclear and mitochondrial genes of snow leopards to develop the phylogenetic relationship of genus *Panthera*. Furthermore, Caragiulo *et al*. (2014) used complete mtDNA of snow leopard to resolve the phylogenetic position of snow leopard among Pantherines. The primary focus of these mitochondrial genome-based studies was to determine the phylogenetic relationship of snow leopards among Pantherine species. These efforts resulted in a consensus among the scientific community that snow leopard is a member of genus *Panthera* hence named *Panthera uncia* instead of *Uncia uncia*, considering it as a separate genus. On one hand, mtDNA provided useful information on snow leopard taxonomy and phylogenetics. However, on the other hand, this 15-18 kbp circular maternally inherited haploid DNA has some disadvantages over the nuclear genome, especially for studying population structure (i.e. the composition of population in terms of sex, age, spatial arrangements, movement of individuals between or within populations, etc) and genetics of species (Ballard and Whitlock, 2004; Choleva *et al*. 2014; Peters *et al*. 2014). Additionally, this rapidly evolving DNA accumulates enough differences to differentiate between closely related species but is less suitable to give information for phylogeography (Toews and Brelsford, 2012; Janecka *et al.,* 2017; Lavretsky *et al*. 2015).

Given the limitations of mtDNA, Waits *et al*. (2007) used polymorphic nuclear microsatellite markers, originally developed for domestic cats, for identifying snow leopard individuals. This was followed by some studies that involved modification of these domestic cat microsatellites to make them snow leopard specific. Seven such modified and specific markers were reported by Janecka *et al*. (2008). This improved the efficiency of microsatellites for non-invasive samples that were available for snow leopard genetic analysis at the time. Janecka *et al*. (2008) were one of the first to examine wild snow leopards in different portions of their range (northwest India, central China, and southern Mongolia) using noninvasive genetic techniques. Due to small sample sizes, the study was merely a descriptive one, but provided the first baseline genetic information on the species. Janecka *et al*. (2008) was followed by other noninvasive genetic studies on snow leopard populations in Nepal (Karmacharya *et al.*, 2011; Aryal *et al.*, 2014), but again these studies were either for optimization of Janecka *et al*. (2008) microsatellites in specific lab conditions or with low quantity and quality of available DNA from a small number of samples. Reproducibility in different labs and low quality and quantity of DNA is a major problem in such studies (Andrews *et al.*, 2018).

To date, the most comprehensive effort done to understand snow leopard genetics is Janecka *et al*. (2017) which involved 33 microsatellites, designed specifically for snow leopards, tested on a comparatively comprehensive sampling regime. This study raised a new debate in the snow leopard research/interest group by reporting three sub-species of snow leopard across its range (Senn *et al.*, 2018 and Janecka *et al.*, 2018).

There is no doubt about the high power of microsatellites, which is the reason for their wide acceptance and utilization in studying wild populations. However, a major disadvantage of microsatellites is that genotyping involves subjective visual interpretation of images, which can lead to relatively high genotyping error rates and the inability to directly compare data across laboratories (Bonin *et al.,* 2004; Pompanon *et al.,* 2005). In addition, the discovery and genotyping of microsatellite loci can be expensive and time-consuming (Andrews *et al.,* 2018).

Because of the problems associated with microsatellites, single nucleotide polymorphisms (SNPs) have long been considered advantageous over microsatellites for addressing many ecological and evolutionary questions (Brumfield *et al.*, 2003; Morin *et* *al.,* 2004; Lopez-Herraez *et al.,* 2005; Anderson and Garza, 2006). Unlike microsatellites, SNPs are less subjective and are reproducible across labs (Andrew *et al.,* 2018). Furthermore, genotyping of SNPs is typically less time-consuming than microsatellites, and SNP loci are more abundant in the genome than microsatellites (Andrews *et al.,* 2018). The problem of lower polymorphism of SNPs compared to microsatellites is overcome by a significantly higher number of SNP loci throughout the genome that contribute to strong analytical power (Glaubitz *et al.*, 2003; Hauser *et al.,* 2011; Tokarska *et al.,* 2009).

In the past few years, next-generation sequencing (NGS) has provided an unprecedented amount of sequence data that is useful for discovering SNPs and studying population structure, genetics, evolutionary patterns, genetic drift, and other forces and processes affecting genetic variation of species. This information is valuable for conservation efforts. Furthermore, NGS outputs short stretches of DNA, making it well-suited for analyzing non-invasive samples, where collected DNA tends to be fragmented (Waits and Paetkau, 2005). A number of NGS methods are available for discovering SNPs, including whole-genome sequencing (Ekblom and Wolf, 2014), RNA-seq (De Wit, *et al.,* 2015), DNA capture (Jones and Good, 2016) and restriction site-associated DNA sequencing (RAD-seq) (Andrews *et al.,* 2016).

Restriction site-associated DNA sequencing (RAD-seq) is extensively used in molecular genetic studies (Davey and Blaxter 2010; Etter *et al*. 2012; Puritz *et al*. 2014), including genome wide association studies (Davey *et al*. 2011) and phylogeography (Andrews *et al*. 2016). This pseudorandom sequencing method involves de-novo genome sequencing and subsequent analysis, involving fragmentation of DNA using restriction enzymes, followed by tagging of digested fragments and high throughput sequencing (Willing *et al.*, 2011; Peterson *et al.*, 2012; Puritz *et al.*, 2014). The insufficient sequence data available for snow leopards makes RAD-seq a promising method for studying snow leopard populations.

The major problem with RAD-seq, however, is that it needs relatively high quality and quantity of DNA, which is almost impossible under snow leopard research circumstances. Another NGS method that can address this problem is DNA capture. It is suitable for non-invasive samples (Perry *et al.,* 2010). This method has already exhibited its potential for very low quality and contaminated DNA in different fields of science including medicine, forensics, paleontology, evolutionary genetics, etc. (Avila-Arcos *et al.*, 2011; Fujii *et al.,* 2019). It involves designing probes that can hybridize with specific DNA sequences of target organisms, leaving behind contaminant DNA that is usually present in non-invasive (scat) samples.

Double Digest Restriction site associated DNA sequencing (ddRAD-seq)

As mentioned earlier, RAD-seq is one method which provides high-resolution population genomic data at low cost and has become an important component in ecological and evolutionary studies. This method involves restriction enzymes to generate DNA fragments from which thousands of SNPs can be identified using high throughput sequencing. RAD-seq does not require a fully sequenced reference genome as loci can be reconstructed de novo from sequencing reads, greatly widening the types of organisms that can be studied beyond traditional model species (Baird *et al.,* 2008; Davey and Blaxter, 2010; Miller *et al.,* 2007). In order to make it more feasible for organisms of interest and cutting the cost and time, some modifications are introduced in RAD-seq; one such modification is double digest restriction site associated DNA sequencing (ddRAD-seq) that uses two digestion enzymes instead of one (Peterson *et al.,* 2012). The digestion with double enzyme allows the selection of a smaller fraction of the genome compared to RAD-seq, improving the targeting of SNPs (though smaller in number) in a greater number of samples (Peterson *et al*. 2012; Puritz *et al.,* 2014; Kess *et al.* 2016). Because of its flexibility, ddRAD-seq is one of the most widely used tools in modern genetics, including medical genetics, diagnostics, ancestry, population genetics and population structure.

ddRAD-seq is a widely accepted method not just for discovering SNPs (Peters *et al.*, 2016; Ba *et al.,* 2017) but also for helping conservation geneticists answer important questions related to wild (Peters *et al.*, 2016; Lavretsky *et al.*, 2016) or captive populations (Ba *et al.,* 2017, Hosoya *et al.,* 2018). It has been used successfully in vertebrates (Ba *et al.,* 2017; Lavretsky *et al.,* 2019), invertebrates (Souza *et al.*, 2017; Kotsakiozi *et al.*, 2017; Lam *et al.,* 2018; Choquet *et al.*, 2019), and plants (Zhou *et al.*, 2014; Yang *et al.*, 2016; Westergaard *et al.,* 2019).

Inter- and intraspecific genetic diversity is the result of past and present geographic, ecological, and behavioral events that either prohibit or favor gene flow, which influences the structure of populations. Such information is important for conservation of species. Deep coverage of genotypes using ddRAD-seq loci enables scientists to identify and study evolutionary processes responsible for contemporary population structure (Koizumi *et al.,* 2012; Lavretsky *et al.,* 2019). This method helped to identify the conservation units in different taxa (Peters *et al.,* 2016) thus helping to improve the conservation efforts for proper management of wild populations. For example, Lah *et al*. (2016) used both microsatellites and ddRAD-seq loci to study population structure in harbor porpoise (*Phocoena phocoena*). The study concluded that ddRAD-seq loci outperform microsatellite markers for identifying population structure in these mobile marine mammals.

Moreover, species where interspecies hybridization is common and is not considered good for genetic integrity of species, ddRAD-seq provided estimates of the frequency and timing of hybridization (Lavretsky *et al.,* 2016; Tezuka *et al.,* 2018; Lavretsky *et al.,* 2019). The data obtained from ddRAD-seq also helped to predict the direction of gene flow among sub-populations (Peters *et al.,* 2016; Saenz-Agudelo *et al.*, 2015).

Phylogeographic analysis, which involves the study of the historical distributions and environmental events that are responsible for current genetic diversity and structure of populations (Koizumi *et al.,* 2012; Mitsui and Setoguchi, 2012), is an area of interest for most conservation geneticist and wildlife managers. ddRAD-seq have enabled low-cost discovery and genotyping of thousands of genetic markers for phylogeographic studies and conservation genetics of different species (Andrews *et al.,* 2016; Valencia *et al.,* 2018), helping to identify local adaptations in populations (Ikeda and Setoguchi, 2010).

Reintroduction programs aim to restore self-sustaining populations of threatened species to their historic range. However, demographic restoration may not reflect genetic restoration, which is necessary for the long-term persistence of populations. To assess the genetic success of reintroduction programs of four small mammals in Australia, White *et* *al*. (2018) used ddRAD-seq to study the genetic diversity of the reintroduced population and founder population. This method helped to identify genetic diversity and highlighted the power of admixture as a tool for conservation management. In conclusion, ddRAD-seq is a method of choice, not only because of its low cost and feasibility, but because of the vast variety of information it can provide about population/species of interest.

DNA Capture Method

Despite all the application and depth of information obtained from ddRAD-seq analysis one of the major problems with this method is its requirement of good quality and quantity of DNA. For species like snow leopard where opportunistic samples (non-invasive - mostly scats) are the only available genetic samples, this method is not very useful because fecal DNA (fDNA) has a high amount of contamination, including DNA from prey items, gut fauna, decomposers, and the environment. fDNA also is highly fragmented. In such scenarios advance target enrichment methods are often necessary to obtain quality data (Wall *et al.,* 2016; Snyder-Mackler *et al.,* 2016; Hernandez-Rodriguez *et al.,* 2018). These methods involve hybridization of target specific DNA/RNA probes (aka baits) that are tagged with biotin, which are subsequently isolated and sequenced for further analysis (Gnirke *et al.,* 2009; Albert *et al.,* 2007). Because of the high sensitivity and specificity, DNA capture is widely used in different fields of research (Avila-Arcos *et al.,* 2011; Fujii *et al.,* 2019).

One of the applications of this method is in environmental DNA (eDNA) sampling, which involves species detection and studying community structure using environmental samples (soil or water). Environmental samples have DNA shed by the species living in the community, but the quantity and quality of this DNA is highly compromised. However, DNA capture is proving to have strong potential for helping scientist to study communities using eDNA (Wilcox *et al.,* 2018; Aylward *et al.,* 2018; Xia *et al.,* 2018; Fujii *et al.,* 2019). In eDNA studies the research question is limited to presence or absence of species and is typically using mtDNA.

DNA capture is also used in studying ancient DNA (aDNA). Samples available in these types of studies are of low quality and quantity DNA. Targeting different genes scientist use this method to study fossils, mummies, and ancient genetic samples (Cruz-Dávalos *et al.,* 2018; Wasef *et al.,* 2018; Richards *et al.,* 2019). Using this approach Paleo-geneticists have examined phylogenetics and evolutionary patterns by comparing aDNA with contemporary samples (Pickrell and Reich, 2014; Lavretsky *et al.,* 2019). Modification and progress of this method is also going on in this field to improve the depth of information obtained.

Its wide acceptance in eDNA and aDNA analysis, where DNA is contaminated and highly fragmented/degraded, which is also common in fecal DNA (fDNA), makes this method a suitable candidate for analyzing fecal DNA. The method is in a phase of discovery and optimization for fDNA; to date, few studies have used this method with fDNA to study wild populations, mostly primates (Perry *et al.*, 2010; De Manuel *et al.*, 2016; Hernandez-Rodriguez *et al.,* 2018).

Considering the advantages of different methods and available genetic sample resources for snow leopard, we used a two-step strategy for studying snow leopard population genetics. We first developed a reference sequence library for SNP discovery using a few invasive samples (blood), which are suitable for the ddRAD-seq protocol. We then used this developed library for designing probes for the DNA capture method. Unfortunately, the DNA capture run was not successful (detailed in chapter 2), but it gave us good quantity of non-targeted (or by-product) data. These by-product data included mtDNA from both snow leopard and prey items. We utilized the mtDNA of prey species for studying snow leopard diet in samples collected from different countries. Snow leopard's mtDNA was used to study genetic diversity in samples based on mtDNA. Both aspects are important for understanding snow leopard ecology.

Diet Ecology

It is very important to have a clear understanding of resources utilized by species for its effective management and conservation planning. Snow leopard, being an apex predator, has a key role in maintaining biodiversity of the ecosystem through population dynamics and trophic cascades (Sergio *et al.*, 2008; Baum and Worm, 2009). Despite the difficulty in gathering information, researchers are trying hard to collect information on snow leopard diets in different regions of its habitat (Shehzad *et al.*, 2012, Janecka *et al.*, 2020). The snow leopard is an opportunistic predator with diverse prey species in different regions (Lyngdoh *et al.*, 2014). A major portion of its diet constitutes wild ungulates that range in size from 36-76 kg (Janecka *et al.*, 2020, Lyngdoh *et al.*, 2014, Shehzad *et al.*, 2012). In addition, it occasionally feeds on medium and small size mammals such as marmots, martens, rodents etc. (Janecka *et al.*, 2007; Jumbay-Uulu *et al.*, 2014, Shehzad *et al.*, 2012). Reports of livestock predation are also common across its range (Oli *et al*. 1994; Shehzad et., 2012). Livestock predation often results in retaliatory killing of snow leopards (Jackson and Wangchuk, 2004) and thus is a major threat to the snow leopard population (Li *et al.,* 2016 and Jackson *et al.,* 2010). Current information on snow leopard diet is not enough to estimate the exact extent of livestock depredation, reason for killing livestock and regional scenario of livestock killings. This is because of a lack of a robust method that can identify prey species from snow leopard scats.

Based on available information and challenges associated with DNA capture method this document has the following sections/chapter:

1. Development of a reference library for snow leopard using double-digest restriction-associated DNA sequencing (ddRAD-seq) for designing probes for DNA capture.
2. DNA capture: challenges and utility for non-invasive samples
3. Diet analysis of snow leopards using non-invasive genetic sampling.
4. SNP identification in mitochondrial genes of snow leopards from samples collected from different regions across its range.

LITERATURE CITED

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J. and Weinstock, G.M. (2007). Direct selection of human genomic loci by microarray hybridization. Nature methods, 4(11), 903-905.

Ale, S. B., Yonzon, P., and Thapa, K. (2007). Recovery of snow leopard Uncia uncia in Sagarmatha (Mount Everest) National Park, Nepal. Oryx, 41(01), 89-92.

Anderson, E. C., and Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. Genetics, 172(4), 2567-2582.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., and Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 17(2), 81.

Andrews, K.R., Adams, J.R., Cassirer, E.F., Plowright, R.K., Gardner, C., Dwire, M., Hohenlohe, P.A. and Waits, L.P., (2018). A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RAD seq data. Molecular ecology resources, 18(6), 1263-1281.

Aryal, A., Brunton, D., Ji, W., Karmacharya, D., McCarthy, T., Bencini, R., and Raubenheimer, D. (2014). Multipronged strategy including genetic analysis for assessing conservation options for the snow leopard in the central Himalaya. Journal of Mammalogy, 95(4), 871-881.

Ávila-Arcos, M.C., Cappellini, E., Romero-Navarro, J.A., Wales, N., Moreno-Mayar, J.V., Rasmussen, M., Fordyce, S.L., Montiel, R., Vielle-Calzada, J.P., Willerslev, E. and Gilbert, M.T.P. (2011). Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. Scientific reports, 1, 74.

Aylward, M. L., Johnson, S. E., Sullivan, A. P., Perry, G. H., and Louis, E. E. (2018). A novel environmental DNA (eDNA) sampling method for aye-ayes from their feeding traces. bioRxiv, 272153.

Ba, H., Jia, B., Wang, G., Yang, Y., Kedem, G., and Li, C. (2017). Genome-Wide SNP Discovery and Analysis of Genetic Diversity in Farmed Sika Deer (Cervus nippon) in Northeast China Using Double-Digest Restriction Site-Associated DNA Sequencing. G3: Genes, Genomes, Genetics, 7(9), 3169-3176.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one, 3(10), e3376.

Ballard, J. W. O., and Whitlock, M. C. (2004). The incomplete natural history of mitochondria. Molecular ecology, 13(4), 729-744.

Baum, J. K., and Worm, B. (2009). Cascading top-down effects of changing oceanic predator abundances. Journal of Animal Ecology, 78(4), 699-714.

Bonin, A., Bellemain, E., Bronken Eidesen, P., Pompanon, F., Brochmann, C., and Taberlet, P. (2004). How to track and assess genotyping errors in population genetics studies. Molecular ecology, 13(11), 3261-3273.

Brumfield, R. T., Beerli, P., Nickerson, D. A., and Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology and Evolution, 18(5), 249-256.

Caragiulo, A., Dias-Freedman, I., Clark, J.A., Rabinowitz, S., Amato, G., 2014. Mitochondrial DNA sequence variation and phylogeography of Neotropic pumas (Puma concolor). Mitochondrial DNA 25, 304–312.

Choleva, L., Musilova, Z., Kohoutova-Sediva, A., Paces, J., Rab, P., and Janko, K. (2014). Distinguishing between incomplete lineage sorting and genomic introgressions: complete fixation of allospecific mitochondrial DNA in a sexually reproducing fish (Cobitis; Teleostei), despite clonal reproduction of hybrids. PLoS One, 9(6), e80641.

Choquet, M., Smolina, I., Dhanasiri, A. K., Blanco-Bercial, L., Kopp, M., Jueterbock, A., ... and Hoarau, G. (2019). Towards population genomics in non-model species with large genomes: a case study of the marine zooplankton Calanus finmarchicus. Royal Society Open Science, 6(2), 180608.

Cruz-Dávalos, D.I., Nieves-Colón, M.A., Sockell, A., Poznik, G.D., Schroeder, H., Stone, A.C., Bustamante, C.D., Malaspinas, A.S. and Ávila-Arcos, M.C., (2018). In-solution Y-chromosome capture-enrichment on ancient DNA libraries. BMC genomics, 19(1), 608.

Davey, J. W., and Blaxter, M. L. (2010). RADSeq: next-generation population genetics. Briefings in functional genomics, 9(5-6), 416-423.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, 12(7), 499.

Davis, B.W., Li, G., Murphy, W.J., 2010. Supermatrix and species tree methods resolve phylogenetic relationships within the big cats, Panthera (Carnivora: Felidae). Mol. Phylogenet. Evol. 56 (1), 64–76.

De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V.C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P. and Schmidt, J.M. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science, 354(6311), 477-481.

De Wit, P., Pespeni, M. H., and Palumbi, S. R. (2015). SNP genotyping and population genomics from expressed sequences–current advances and future possibilities. Molecular ecology, 24(10), 2310-2323.

Ekblom, R., and Wolf, J. B. (2014). A field guide to whole-genome sequencing, assembly and annotation. Evolutionary applications, 7(9), 1026-1042.

Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A., and Cresko, W. A. (2012). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. In Molecular methods for evolutionary genetics (pp. 157-178). Humana Press.

Fujii, K., Doi, H., Matsuoka, S., Nagano, M., Sato, H., and Yamanaka, H. (2019). Environmental DNA metabarcoding for fish community analysis in backwater lakes: A comparison of capture methods. PloS one, 14(1), e0210357.

Glaubitz, J. C., Rhodes Jr, O. E., and DeWoody, J. A. (2003). Prospects for inferring pairwise relationships with single nucleotide polymorphisms. Molecular Ecology, 12(4), 1039-1047.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. and Gabriel, S. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature biotechnology, 27(2), 182-189.

Hauser, L., Baird, M., Hilborn, R. A. Y., Seeb, L. W., and Seeb, J. E. (2011). An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (Oncorhynchus nerka) population. Molecular ecology resources, 11, 150-161.

Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., Angedakin, S., Casals, F., Navarro, A., Vigilant, L. and Kühl, H.S. (2018). The impact of endogenous content, replicates, and pooling on genome capture from faecal samples. Molecular ecology resources, 18(2), 319-333.

Hosoya, S., Kikuchi, K., Nagashima, H., Onodera, J., Sugimoto, K., Satoh, K., ... and Ueda, K. (2018). Assessment of genetic diversity in Coho salmon (Oncorhynchus kisutch) populations with no family records using ddRAD-seq. BMC research notes, 11(1), 548.

Ikeda, H., and Setoguchi, H. (2010). Natural selection on PHYE by latitude in the Japanese archipelago: insight from locus specific phylogeographic structure in Arcterica nana (Ericaceae). Molecular ecology, 19(13), 2779-2791.

Jackson, R. M., and Wangchuk, R. (2004). A community-based approach to mitigating livestock depredation by snow leopards. Human dimensions of wildlife, 9(4), 1-16.

Jackson, R. M., Mishra, C., McCarthy, T. M., and Ale, S. B. (2010). Snow leopards: conflict and conservation. The Biology and Conservation of Wild Felids, 417-430.

Jackson, R. M., Roe, J. D., Wangchuk, R., and Hunter, D. O. (2006). Estimating snow leopard population abundance using photography and capture-recapture techniques. Wildlife Society Bulletin, 34(3), 772-781.

Janecka, J. E., Jackson, R., Yuquang, Z., Diqiang, L., Munkhtsog, B., Buckley-Beason, V., and Murphy, W. J. (2008). Population monitoring of snow leopards using noninvasive collection of scat samples: a pilot study. Animal Conservation, 11(5), 401-411.

Janecka, J. E., Janecka, M. J., Helgen, K. M., and Murphy, W. J. (2018). The validity of three snow leopard subspecies: response to Senn *et al*. Heredity, 120(6), 586.

Janecka, J. E., Zhang, Y., Li, D., Munkhtsog, B., Bayaraa, M., Galsandorj, N., Wangchuk T. R., Karmacharya, D., Li, J., Lu, Z., Uulu, K. Z. Gaur, A., Kumar, S., Kumar, K., Hussain, S., Muhammad, G., Jevit, M., Hacker, C., Burger, P., Wultsch, C., Janecka, M. J. Helgen, K., Murphy, W. J., Jackson, R. (2017). Range-wide snow leopard phylogeography supports three subspecies. Journal of Heredity, 108(6), 597-607.

Janecka, J.E., Hacker, C., Broderick, J., Pulugulla, S., Auron, P., Ringling, M., Nelson, B., Munkhtsog, B., Hussain, S., Davis, B. and Jackson, R. (2020). Noninvasive Genetics and Genomics Shed Light on the Status, Phylogeography, and Evolution of the Elusive Snow Leopard. In Conservation Genetics in Mammals (pp. 83-120). Springer, Cham.

Jones, M. R., and Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. Molecular ecology, 25(1), 185-202.

Jumaby-Uulu K, Wegge P, Mishra C, Sharma K (2014) Large carnivores and low diversity of optimal prey: a comparison of the diets of snow leopards *Panthera uncia* and wolves *Canis lupus* in Sarychat-Ertash Reserve in Kyrgyzstan. Oryx 48:529–535.

Karmacharya DB, Thapa K, Shrestha R, Dhakal M, Janecka JE. 2011. Noninvasive genetic population survey of snow leopards (*Panthera uncia*) in Kangchenjunga conservation area, Shey Phoksundo National Park and surrounding buffer zones of Nepal. BMC Res Notes. 4:516.

Kess, T., Gross, J., Harper, F., and Boulding, E. G. (2016). Low-cost ddRAD method of SNP discovery and genotyping applied to the periwinkle Littorina saxatilis. Journal of Molluscan Studies, 82(1), 104-109.

Koizumi, I., Usio, N., Kawai, T., Azuma, N., and Masuda, R. (2012). Loss of genetic diversity means loss of geological information: the endangered Japanese crayfish exhibits remarkable historical footprints. PloS one, 7(3), e33986.

Kotsakiozi, P., Richardson, J. B., Pichler, V., Favia, G., Martins, A. J., Urbanelli, S., ... and Caccone, A. (2017). Population genomics of the Asian tiger mosquito, Aedes albopictus: insights into the recent worldwide invasion. Ecology and evolution, 7(23), 10143-10157.

Lah, L., Trense, D., Benke, H., Berggren, P., Gunnlaugsson, Þ., Lockyer, C., Öztürk, A., Öztürk, B., Pawliczka, I., Roos, A. and Siebert, U. (2016). Spatially explicit analysis of genome-wide SNPs detects subtle population structure in a mobile marine mammal, the harbor porpoise. PloS one, 11(10), e0162792.

Lam, A.W., Gueuning, M., Kindler, C., Van Dam, M., Alvarez, N., Panjaitan, R., Shaverdo, H., White, L.T., Roderick, G.K. and Balke, M. (2018). Phylogeography and population genomics of a lotic water beetle across a complex tropical landscape. Molecular ecology, 27(16), 3346-3356.

Lavretsky, P., Dacosta, J. M., Hernández-Baños, B. E., Engilis, A., Sorenson, M. D., and Peters, J. L. (2015). Speciation genomics and a role for the Z chromosome in the early stages of divergence between Mexican ducks and mallards. Molecular ecology, 24(21), 5364-5378.

Lavretsky, P., Janzen, T., and McCracken, K. G. (2019). Identifying hybrids and the genomics of hybridization: Mallards and American black ducks of Eastern North America. Ecology and Evolution.

Lavretsky, P., Peters, J.L., Winker, K., Bahn, V., Kulikova, I., Zhuravlev, Y.N., Wilson, R.E., Barger, C., Gurney, K. and McCracken, K.G., (2016). Becoming pure: identifying generational classes of admixed individuals within lesser and greater scaup populations. Molecular Ecology, 25(3), 661-674.

Li, J., Xiao, L., and Lu, Z. (2016). Challenges of snow leopard conservation in China. Science China. Life Sciences, 59(6), 637.

Lopez-Herraez, D., Schäfer, H., Mosner, J., Fries, H. R., and Wink, M. (2005). Comparison of microsatellite and single nucleotide polymorphism markers for the genetic analysis of a Galloway cattle population. Zeitschrift für Naturforschung C, 60(7-8), 637-643.

Lovari, S., Boesi, R., Minder, I., Mucci, N., Randi, E., Dematteis, A., and Ale, S. B. (2009). Restoring a keystone predator may endanger a prey species in a human-altered ecosystem: the return of the snow leopard to Sagarmatha National Park. Animal Conservation, 12(6), 559-570.

Lyngdoh, S., Shrotriya, S., Goyal, S. P., Clements, H., Hayward, M. W., and Habib, B. (2014). Prey preferences of the snow leopard (*Panthera uncia*): regional diet specificity holds global significance for conservation. PloS one, 9(2).

McCarthy, K. P., Fuller, T. K., Ming, M., McCarthy, T. M., Waits, L., and Jumabaev, K. (2008). Assessing estimators of snow leopard abundance. The Journal of Wildlife Management, 72(8), 1826-1833.

Miller,M.R., Dunham, J.P., Amores, A., Cresko, W.A., Johnson, E.A., 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248.

Mitsui, Y., and Setoguchi, H. (2012). Recent origin and adaptive diversification of Ainsliaea (Asteraceae) in the Ryukyu Islands: molecular phylogenetic inference using nuclear microsatellite markers. Plant systematics and evolution, 298(5), 985-996.

Morin, P. A., Luikart, G., and Wayne, R. K. (2004). SNPs in ecology, evolution and conservation. Trends in ecology and evolution, 19(4), 208-216.

Nowell, K., and Jackson, P. (Eds.). (1996). Wild cats: status survey and conservation action plan (pp. 1-382). Gland: IUCN.

Oli, M. K., Taylor, I. R., and Rogers, M. E. (1994). Snow leopard *Panthera uncia* predation of livestock: an assessment of local perceptions in the Annapurna Conservation Area, Nepal. Biological Conservation, 68(1), 63-68.

Perry, G. H., Marioni, J. C., Melsted, P., and Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. Molecular Ecology, 19(24), 5332-5344.

Peters, J. L., Lavretsky, P., DaCosta, J. M., Bielefeld, R. R., Feddersen, J. C., and Sorenson, M. D. (2016). Population genomic data delineate conservation units in mottled ducks (Anas fulvigula). Biological Conservation, 203, 272-281.

Peters, J. L., Sonsthagen, S. A., Lavretsky, P., Rezsutek, M., Johnson, W. P., and McCracken, K. G. (2014). Interspecific hybridization contributes to high genetic diversity and apparent effective population size in an endemic population of mottled ducks (Anas fulvigula maculosa). Conservation Genetics, 15(3), 509-520.

Peterson B. K., J. N. Weber, E. H. Kay, H. S. Fisher H. E. Hoekstra (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7(5): e37135. doi:10.1371/journal.pone.0037135.

Pickrell, J. K., and Reich, D. (2014). Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics, 30(9), 377-389.

Pompanon, F., Bonin, A., Bellemain, E., and Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. Nature Reviews Genetics, 6(11), 847.

Puritz, J. B., Hollenbeck, C. M., and Gold, J. R. (2014). dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ, 2, e431.

Richards, S.M., Hovhannisyan, N., Gilliham, M., Ingram, J., Skadhauge, B., Heiniger, H., Llamas, B., Mitchell, K.J., Meachen, J., Fincher, G.B. and Austin, J.J. (2019). Low-cost cross-taxon enrichment of mitochondrial DNA using in-house synthesised RNA probes. PloS one, 14(2), e0209499.

Saenz-Agudelo, P., Dibattista, J. D., Piatek, M. J., Gaither, M. R., Harrison, H. B., Nanninga, G. B., and Berumen, M. L. (2015). Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. Molecular ecology, 24(24), 6241-6255.

Senn, H., Murray-Dickson, G., Kitchener, A. C., Riordan, P., and Mallon, D. (2018). Response to Janecka *et al*. 2017. Heredity, 120(6), 581.

Sergio, F., Caro, T., Brown, D., Clucas, B., Hunter, J., Ketchum, J., McHugh, K. and Hiraldo, F. (2008). Top predators as conservation tools: ecological rationale, assumptions, and efficacy. Annual review of ecology, evolution, and systematics, 39, 1-19.

Shehzad, W., McCarthy, T. M., Pompanon, F., Purevjav, L., Coissac, E., Riaz, T., and Taberlet, P. (2012). Prey preference of snow leopard (*Panthera uncia*) in South Gobi, Mongolia. PloS one, 7(2).

Snyder-Mackler, N., Majoros, W.H., Yuan, M.L., Shaver, A.O., Gordon, J.B., Kopp, G.H., Schlebusch, S.A., Wall, J.D., Alberts, S.C., Mukherjee, S. and Zhou, X. (2016). Efficient genome-wide sequencing and low-coverage pedigree analysis from noninvasively collected samples. Genetics, 203(2), 699–714.

Souza, C.A., Murphy, N., Villacorta-Rath, C., Woodings, L.N., Ilyushkina, I., Hernandez, C.E., Green, B.S., Bell, J.J. and Strugnell, J.M. (2017). Efficiency of ddRAD target enriched sequencing across spiny rock lobster species (Palinuridae: Jasus). Scientific reports, 7(1), 6781.

Tezuka, A., Takasu, M., Tozaki, T., and Nagano, A. J. (2018). The ability of ddRAD-Seq to estimate genetic diversity and genetic introgression in endangered native livestock. bioRxiv, 454108.

Toews, D. P., and Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology, 21(16), 3907-3930.

Tokarska, M., Marshall, T., Kowalczyk, R., Wójcik, J.M., Pertoldi, C., Kristensen, T.N., Loeschcke, V., Gregersen, V.R. and Bendixen, C., (2009). Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison. Heredity, 103(4), 326-332.

Valencia, L. M., Martins, A., Ortiz, E. M., and Di Fiore, A. (2018). A RAD-sequencing approach to genome-wide marker discovery, genotyping, and phylogenetic inference in a diverse radiation of primates. PloS one, 13(8), e0201254

Waits, L. P., and Paetkau, D. (2005). Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection. Journal of Wildlife Management, 69(4), 1419-1433.

Waits, L. P., Buckley-Beason, V. A., Johnson, W. E., Onorato, D., and McCarthy, T. O. M. (2007). A select panel of polymorphic microsatellite loci for individual identification of snow leopards (*Panthera uncia*). Molecular Ecology Notes, 7(2), 311-314.

Wall, J.D., Schlebusch, S.A., Alberts, S.C., Cox, L.A., Snyder-Mackler, N., Nevonen, K.A., Carbone, L. and Tung, J., (2016). Genome-wide ancestry and divergence patterns from low-coverage sequencing data reveal a complex history of admixture in wild baboons. Molecular ecology, 25(14), 3469-3483.

Wasef, S., Huynen, L., Millar, C.D., Subramanian, S., Ikram, S., Holland, B., Willerslev, E. and Lambert, D.M., (2018). Fishing for Mitochondrial DNA in The Egyptian Sacred Ibis Mummies. bioRxiv, 473454.

Wei, L., Wu, X., Jiang, Z., 2009. The complete mitochondrial genome structure of snow leopard *Panthera uncia*. Mol. Biol. Rep. 36, 871–878.

Westergaard, K. B., Zemp, N., Bruederle, L. P., Stenøien, H. K., Widmer, A., and Fior, S. (2019). Population genomic evidence for plant glacial survival in Scandinavia. Molecular ecology, 28(4), 818-832.

White, L. C., Moseby, K. E., Thomson, V. A., Donnellan, S. C., and Austin, J. J. (2018). Long-term genetic consequences of mammal reintroductions into an Australian conservation reserve. Biological Conservation, 219, 1-11.

Wilcox, T. M., Zarn, K. E., Piggott, M. P., Young, M. K., McKelvey, K. S., and Schwartz, M. K. (2018). Capture enrichment of aquatic environmental DNA: A first proof of concept. Molecular ecology resources, 18(6), 1392-1401.

Willing, E. M., Hoffmann, M., Klein, J. D., Weigel, D., and Dreyer, C. (2011). Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics, 27(16), 2187-2193.

Xia, Z., Zhan, A., Gao, Y., Zhang, L., Haffner, G. D., and MacIsaac, H. J. (2018). Early detection of a highly invasive bivalve based on environmental DNA (eDNA). Biological invasions, 20(2), 437-447.

Xu, A., Jiang, Z., Li, C., Guo, J., Da, S., Cui, Q., Yu, S., and Wu, G. (2008). Status and conservation of the snow leopard *Panthera uncia* in the Gouli Region, Kunlun Mountains, China. Oryx, 42(3), 460-463.

Yang, G.Q., Chen, Y.M., Wang, J.P., Guo, C., Zhao, L., Wang, X.Y., Guo, Y., Li, L., Li, D.Z. and Guo, Z.H. (2016). Development of a universal and simplified ddRAD library preparation approach for SNP discovery and genotyping in angiosperm plants. Plant methods, 12(1), 39.

Zhang F., Jiang, Z., Zeng, Y., McCarthy, T., 2007. Development of primers to characterize the mitochondrial control region of the snow leopard (Uncia uncia). Mol. Ecol. Notes 7, 1196–1198.

Zhou, X., Xia, Y., Ren, X., Chen, Y., Huang, L., Huang, S., Liao, B., Lei, Y., Yan, L. and Jiang, H. (2014). Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). BMC genomics, 15(1), 351.

Chapter 1

IMPROVING OUR CONSERVATION GENETIC TOOLKIT: DDRADSEQ FOR SNPS IN SNOW LEOPARDS

1.1. ABSTRACT

Snow leopards (*Panthera uncia*) are an enigmatic, high-altitude species whose challenging habitat, low population densities and patchy distribution have presented challenges for scientists studying its biology, population structure, and genetics. Molecular scatology brings a new hope for conservation efforts by providing valuable insights about snow leopards, including their distribution, population densities, connectivity, habitat use, and population structure for assigning conservation units. However, traditional amplification of microsatellites from non-invasive sources of DNA are accompanied by significant genotyping errors due to low DNA yield and poor quality. These errors can lead to incorrect inferences in the number of individuals and estimates of genetic diversity. Next generation technologies have revolutionized the depth of information we can get from a species' genome. Here we used ddRAD-seq, a well-established technique for studying non-model organisms, to develop a reference sequence library for snow leopards using blood samples from five Mongolian individuals. Our final data set reveals 4504 loci with a median size range of 221 bp. We identified 697 SNPs and low nucleotide diversity (0.00032) within these loci. However, the probability that two random individuals will share identical genotypes is about 10^{-168}. We developed probes for DNA capture using this sequence library which can now be used for genotyping individuals from scat samples. Genetic data from ddRAD-seq will be invaluable for conducting population and landscape scale studies that can inform snow leopard conservation strategies.

1.2. INTRODUCTION

Despite being a high profile and charismatic carnivore, information on snow leopard ecology, population structure, and genetics lags behind most other large carnivores (Caragiulo *et al*. 2016 and Fox and Chundawat, 2016). This lack of information is largely because of the animal's cryptic nature and remote habitat. Therefore, it is difficult to obtain sufficient data on numbers, locations of peripheral and core populations, and areas where the snow leopard populations are in decline. Such information is critical for the conservation of this apex predator.

Efforts are underway to determine the status of the species across its extensive (approximately 1.6 million km^2) but highly fragmented range (McCarthy *et al*. 2016). One such effort is the genomic analysis of DNA, extracted from non-invasively collected samples. This can benefit conservation efforts by providing information about population densities, connectivity, source/sink dynamics, and habitat use, among others, as well as how to define distinct conservation units upon which to focus specific conservation efforts.

Several studies have genotyped non-invasively collected snow leopard DNA samples (Waits *et al*. 2007; Janecka *et al.,* 2008; Karmacharya *et al*. 2011; Aryal *et al*. 2014). However, a common problem associated with these studies is that the genetic markers used were either not snow leopard specific or had to be modified to improve their specificity. The methods used are also challenged by genotyping errors (McKelvey and Schwartz, 2004) due to the inherent low yield and poor quality of DNA from such samples. These errors can lead to incorrect inferences, including the misidentification of individuals (Waits and Paetkau, 2005), which in turn can lead to incorrect estimates of population size and patterns of genetic diversity. Janecka *et al*. (2016) attempted to overcome these limitations by designing 33 snow leopard-specific microsatellite markers and using them to evaluate snow leopard populations across its range. This is the most detailed genetic study of this cat to date. However, it still has conventional limitations associated with microsatellites, such as amplification failure, allele dropouts, and the appearance of false alleles. Hence, the issues remained unresolved.

Today, cutting-edge technologies, like next generation sequencing (NGS), have revolutionized the depth of information we can get from a species' genome by providing sequence information from thousands of loci. Thus, NGS enables scientists to supplement and further refine existing research by thoroughly analyzing the genomes of organisms to better evaluate evolutionary patterns and signatures that can be beneficial for conservation efforts. Additionally, NGS amplifies short stretches of DNA, making it well-suited for analyzing non-invasive samples, where collected DNA tends to already be fragmented (Waits and Paetkau, 2005).

Double digest restriction-site associated DNA sequencing (ddRAD-seq) is a well-established NGS technique used to study non-model organisms (Peterson *et al*. 2012). This method involves pseudorandom sampling of whole genomes of organisms (Miller *et al.*, 2007; Baird *et al.*, 2008) and subsequent discovery of SNPs from sequenced genomes (Peterson *et al*. 2012). We developed ddRAD-seq libraries for snow leopard and identified a SNP panel for studying their population genetics. Given the need for high quality snow leopard genetic data and the limitations associated with existing methods, the primary objectives of this study were:

i- Genome-wide SNP discovery in snow leopards;

ii- Determine the utility of this SNP panel for identifying individuals; and

iii- Compare the resolution between SNPs and microsatellites

1.3. MATERIALS AND METHODS

1.3.1. Sampling and DNA Extraction

We used five blood samples of wild Mongolian snow leopards, archived at the American Museum of Natural History, which were collected as part of another project (Johansson *et al.*, 2013). DNA from blood samples was extracted by using a DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA). The DNA extractions were quantified using a NanoDrop (Thermo Scientific).

1.3.2. ddRAD-seq Library

The extracted DNA was used to generate ddRAD-seq libraries following DaCosta and Sorenson (2014). Briefly, 0.5–1 μg of genomic DNA was digested using 20 U of the restriction enzymes *EcoRI* and *SbfI*, producing fragments with sticky ends. Uniquely barcoded adapters were ligated to the digestion products. DNA in the size range of 300 to 450 bp was excised from a 2% low melt agarose gel and purified using a MinElute Gel Extraction Kit (Qiagen, Valencia, CA). This size-selected DNA was then amplified using standard PCR with Phusion high fidelity DNA polymerase (Thermo Scientific, Pittsburgh, PA), and the products were purified using magnetic SPRI beads (Company). We used real-time PCR to quantify PCR products using an Illumina library quantification kit (KAPA Biosystems, Wilmington, MA) and an ABI 7900HT SDS (Applied Biosystems, Foster City, CA). Equimolar concentrations of libraries with unique barcodes were pooled and sequenced (single-end, 150 base pair reads) on an Illumina HiSeq 2000 at Tufts University.

1.3.3. Data processing

DaCosta and Sorenson's (2014) computational pipeline was used to process the raw data obtained from Illumina sequence reads (Python scripts available at http://github.com/BU-RAD-seq/ddRAD-seqPipeline). For each individual, identical reads were collapsed while retaining read counts and the highest quality score for each position. Individual reads that were >10% divergent from all others (using the UCLUST function in USEARCH v. 5; Edgar, 2010) and/or those with an average Phred score < 20 were removed. We then clustered the filtered reads from all individuals into putative loci using UCLUST (–id setting of 0.85). The highest quality read from each cluster (i.e., putative locus) was localized in the tiger (*Panthera tigris*) genome (GCA_000464555.1 PanTig1.0), using BLASTN v. 2 (Altschul *et al.*, 1990). Clusters that did not generate a BLAST hit were carried through the pipeline as anonymous loci. After combining clusters that had the same BLAST hits, we aligned reads from each cluster using MUSCLE v. 3 (Edgar, 2004).

The final alignments were used to genotype individuals at each locus using the RADGenotypes.py script. Individuals were scored as homozygous at a locus if ≥93% of reads were consistent with a single haplotype across polymorphic sites, and heterozygous if a second haplotype was represented by at least 29% of reads (DaCosta and Sorenson, 2014). We also scored individuals as heterozygous if a second allele was represented by as few as 20% of reads, but only if the second allele was known from other individuals in the sample. Individual genotypes that did not meet either of these criteria, or had evidence of more than two haplotypes, were flagged (0.001% of all genotypes in the final data set); for these samples, we retained only the allele represented by the majority of reads and scored the second allele as missing data. Similarly, the second allele was scored as missing for apparently homozygous genotypes based on 1 to 5 reads, which were considered “low depth” (1.1% of all genotypes). We retained for analysis all loci that had complete genotypes for at least four of our five individuals.

1.3.4. Estimates of genetic diversity

Standard estimates of genetic diversity were calculated using the R package – PopGenome (Pfeifer *et al.*, 2014) and Structure 2.2.3 (Pritchard *et al.*, 2000). Observed heterozygosity (H_{obs}), expected heterozygosity (H_e), nucleotide diversity, and inbreeding coefficient (F_{IS}) were calculated. Probability of Identity (PID), which is the probability that two random individuals will have identical genotypes, was calculated using Paetkau and Strobeck (1994) equation: $PID = \sum P_i^4 + \sum (2P_iP_j)^2$. Where *Pi* is the frequency of the i^{th} allele and *Pj* is the frequency of j^{th} allele.

1.3.5. Comparison with Microsatellites

The individual snow leopards used in this study were also genotyped by Caragiulo *et al*. (unpublished) using 12 microsatellite loci (Caragiulo *et al*. 2015). We calculated H_{obs}, H_e, F_{IS}, and PID values for the microsatellite data as we did for ddRAD-seq loci for comparisons.

1.4. RESULTS AND DISCUSSION

Given significant knowledge gaps on the genetics of snow leopards, and the importance of this information for conservation, we used ddRAD-seq for a high resolution and low-cost development of a reference sequence library. This method is widely becoming an important component of ecological and evolutionary studies (Andrews *et al.*, 2018; Ba *et al.*, 2017; Peters *et al.*, 2016), especially for organisms like snow leopards where little is known about their genome. We obtained an average of ~ 4,000,000 high quality sequence reads per individual. After assembling these reads, we retained 4504 loci that were recovered from a minimum of 80% of individuals and contained no flagged genotypes. Among these reads median fragment size was 221 bp. The final data set comprised 511 loci with one or more polymorphic sites plus 3993 constant loci and a total of 7,428,063 aligned nucleotides and 697 SNPs. Overall nucleotide diversity in the five Mongolian samples was 0.00032 with nucleotide diversity ranging from 0.00081 to 0.07935 among the variable loci.

To evaluate the utility of these SNPs, we estimated different descriptive statistics important in population genetics, comparing them between traditional microsatellites and our SNP panel. The average expected heterozygosity (He) was 0.042, which was slightly lower than the observed heterozygosity (Ho) of 0.047 for ddRAD-seq. However, He and Ho did not differ significantly (p=0.07). In contrast, He was slightly higher than Ho for microsatellites but the difference is statically insignificant (p = 0.341) (Table 1.1). Fixation index (F) for both microsatellites and ddRAD-seq was negative. However, these values were calculated for only five samples from Mongolia and are not necessarily representative of the total genetic diversity within the population.

To evaluate the power of ddRAD-seq loci for individual identification, we calculated PID. This is the probability that two individuals drawn at random from a population will have the same genotype at multiple loci (Waits *et al*. 2001, Valiere 2002). The PID is widely used to assess the statistical confidence for individual identification, for non-invasive sampling (Reed *et al*. 1997; Kohn *et al*. 1999; Mills *et al*. 2000; Waits and Leberg 2000). It is therefore useful for estimating the number of individuals with higher confidence (Ernest *et al*. 2000). The value for PID for ddRAD-seq is very low (i.e. 1.55 x 10^{-168}) compared to the value for microsatellite loci (i.e. 2.35 x 10 $^{-7}$), Therefore the power of ddRAD-seq loci to identify individuals is over 100 orders of magnitude higher than that of microsatellites.

As the role of genomic methods (e.g. SNPs) has evolved in conservation practice, the transition from, and juxtaposition with, traditional conservation genetics methods (Ouburg *et al*. 2010) has brought to light advantages and challenges (Allendorf *et al*. 2010). For snow leopards, the advantages are particularly critical. Microsatellites are notoriously difficult to standardize across different labs. In the case of non-invasive samples, there is the added challenge of genotyping error due to allelic drop-out and false alleles. Specifically, in snow leopards, microsatellites appear to have generally low variability (Caragiulo *et al*. 2016). These challenges limit the utility of microsatellites as a universal maker to use across labs and research groups. In addition, an often overlooked obstacle is that given the species' legal status and the political sensitivities in many of its range countries, it is sometimes impossible to transport DNA samples across borders, further precluding work being done in a single, standardized lab. Instead, lab work must occur within the country where the samples were collected. A hypervariable SNP panel for snow leopards, such as done here, circumvents all the challenges of microsatellites. The next step is to design probes from our ddRAD-seq libraries for a targeted DNA capture method that is well suited for non-invasive samples (Perry *et al.*, 2010). This work is ongoing and will provide a crucial new tool for conducting conservation genetic research on this imperiled species.

Table 1.1: The mean genetic diversity estimated in 5 snow leopards genotyped using 12 microsatellites and ddRAD-seq loci.

	Microsatellites	ddRAD-seq loci
N	4.333	4.973
Na	2.917	1.118
Ho	0.536	0.047
He	0.544	0.042
F	-0.023	-0.113

N = average sample size, Na = average number of different alleles, Ho = observed heterozygosity, He = expected heterozygosity, F = fixation index.

1.5. LITERATURE CITED

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. Nature Reviews Genetics 11, 697-709.

Andrews, K. R., Adams, J. R., Cassirer, E. F., Plowright, R. K., Gardner, C., Dwire, M., ... and Waits, L. P. (2018). A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RAD seq data. Molecular ecology resources.

Ba, H., Jia, B., Wang, G., Yang, Y., Kedem, G., and Li, C. (2017). Genome-Wide SNP Discovery and Analysis of Genetic Diversity in Farmed Sika Deer (Cervus nippon) in Northeast China Using Double-Digest Restriction Site-Associated DNA Sequencing. G3: Genes, Genomes, Genetics, 7(9), 3169-3176.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PloS one, 3(10), e3376.

Caragiulo A., G. Amato, B. Weckworth (2016) Conservation genetics of snow leopards. In: Snow Leopards – Biodiversity of the world: conservation from genes to landscapes (Eds. P. J. Nyhus, T. McCarthy and D. Mallon). Elsevier Inc., London, UK.

Caragiulo A., Y. Kang, S. Rabinowitz, I. Dias-Freedman, S. Loss, X.W. Zhou, W.D. Bao, G. Amato (2015) Presence of the endangered Amur tiger Panthera tigris altaica in Jilin Province, China, detected using non-invasive genetic techniques. Oryx 49(4) 632-635.

DaCosta, J. M., and Sorenson, M. D. (2014). Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. PloS one, 9(9), e106713.

Janečka, J. E., Jackson, R., Yuquang, Z., Diqiang, L., Munkhtsog, B., Buckley-Beason, V., and Murphy, W. J. (2008). Population monitoring of snow leopards using noninvasive collection of scat samples: a pilot study. Animal Conservation, 11(5), 401-411.

Johansson, Ö., Malmsten, J., Mishra, C., Lkhagvajav, P., and McCarthy, T. (2013). Reversible immobilization of free-ranging Snow Leopards (Panthera uncia) with a combination of medetomidine and tiletamine-zolazepam. Journal of Wildlife Diseases, 49(2), 338-346.

Lavretsky, P., Dacosta, J. M., Hernández-Baños, B. E., Engilis, A., Sorenson, M. D., and Peters, J. L. (2015). Speciation genomics and a role for the Z chromosome in the early stages of divergence between Mexican ducks and mallards. Molecular ecology, 24(21), 5364-5378.

McKelvey, K. S., and Schwartz, M. K. (2004). Genetic errors associated with population estimation using non-invasive molecular tagging: problems and new solutions. Journal of Wildlife Management, 68(3), 439-448.

McKelvey, K. S., and Schwartz, M. K. (2004). Genetic errors associated with population estimation using non-invasive molecular tagging: problems and new solutions. Journal of Wildlife Management, 68(3), 439-448.

Mills, L. S., Citta, J. J., Lair, K. P., Schwartz, M. K., and Tallmon, D. A. (2000). Estimating animal abundance using noninvasive DNA sampling: promise and pitfalls. Ecological applications, 10(1), 283-294.

Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma RK, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. Trends in Genetics 26, 177-187

Perry, G. H., Marioni, J. C., Melsted, P., and Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. Molecular Ecology, 19(24), 5332-5344.

Peters, J. L., Lavretsky, P., DaCosta, J. M., Bielefeld, R. R., Feddersen, J. C., and Sorenson, M. D. (2016). Population genomic data delineate conservation units in mottled ducks (Anas fulvigula). Biological Conservation, 203, 272-281.

Peterson B. K., J. N. Weber, E. H. Kay, H. S. Fisher H. E. Hoekstra (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7(5): e37135. doi:10.1371/journal.pone.0037135

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

Waits, L. P., and Paetkau, D. (2005). Noninvasive genetic sampling tools for wildlife biologists: a review of applications and recommendations for accurate data collection. Journal of Wildlife Management, 69(4), 1419-1433.

Waits, L. P., Luikart, G., and Taberlet, P. (2001). Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. Molecular ecology, 10(1), 249-256.

Chapter 2

DNA CAPTURE: CHALLENGES AND UTILITY FOR NON-INVASIVE SAMPLES

2.1. ABSTRACT

Our capacity to address important ecological, conservation, and evolutionary questions increased rapidly with the advances in DNA sequencing and genotyping technologies. Non-invasive samples are often the only available source of DNA for many wild species, but these samples yield low quantity and quality of DNA. Such samples present unique methodological challenges, which is the prime reason advanced genetic techniques are not used in wildlife and conservation studies. Researchers are tailoring laboratory and computational methods to better extract the information contained in non-invasive DNA. We obtained non-invasive fecal scats of snow leopards from seven different countries, and applied DNA capture, a widely used method for studying human populations, ancient DNA samples, eDNA, and forensics. We successfully sequenced endogenous DNA from samples. Additionally, we sequenced a significant amount of non-target DNA, mostly mitochondrial (mt) DNA. This non-target mtDNA originated from snow leopard and prey species. Here, we present how we can use this non-target data to harvest important genetic information about snow leopard populations, such as mtDNA-based analysis of population structure and diet.

2.2. INTRODUCTION

The next generation sequencing (NGS) technologies have drastically changed the paradigm of sequencing and information scientists can obtain from the DNA of organisms.

Both the amount of information and the ability to harvest information in a time- and cost-effective manner make NGS methods successful. Based on the requirements for specific research questions, there are a number of options for NGS based methods, including whole-genome sequencing (Ekblom and Wolf, 2014), RNA-seq (De Wit, *et al.*, 2015), DNA capture (Jones and Good, 2016) and restriction site-associated DNA sequencing (RAD-seq) (Andrews *et al.*, 2016), etc. DNA capture is used especially when genetic information source is of low quality and quantity of DNA, e.g. ancient DNA (aDNA), fecal DNA (fDNA), forensic DNA, environmental DNA (eDNA) etc. DNA capture involves hybridization of target specific DNA/RNA probes (aka baits) that are tagged with biotin, which are subsequently isolated and sequenced for further analysis (Gnirke *et al.*, 2009; Albert *et al.*, 2007). Because of the high sensitivity and specificity, DNA capture is widely used in different fields of research (Avila-Arcos *et al.*, 2011; Fujii *et al.*, 2019).

Environmental DNA (eDNA) sampling involves species detection and studying community structure using environmental samples (soil or water). Environmental samples have DNA shed by the species living in the community, but the quantity and quality of this DNA is highly compromised. However, DNA capture is proving to have a strong potential for helping scientists study communities using eDNA (Wilcox *et al.*, 2018; Aylward *et al.*, 2018; Xia *et al.*, 2018; Fujii *et al.*, 2019). In eDNA studies the research question is limited to presence or absence of species and is typically using mtDNA.

DNA capture is also used in studying ancient DNA (aDNA). Samples available in these types of studies are of low quality and quantity DNA. Targeting different genes scientist use this method to study fossils, mummies, and ancient genetic samples (Cruz-

Dávalos *et al.,* 2018; Wasef *et al.,* 2018; Richards *et al.,* 2019). Using this approach Paleo-geneticists tried to find out phylogenetics and evolutionary patterns by comparing aDNA with contemporary samples (Pickrell and Reich, 2014; Lavretsky *et al.,* 2019). Modification and progress of this method is also going on in this field to improve the depth of information obtained.

Wide acceptance in eDNA and aDNA analysis, where DNA is contaminated and highly fragmented/degraded, which is also common character of fecal DNA (fDNA), makes this method a suitable candidate and conservation geneticists are experimenting with the method to study species of interest (Wilcox *et al.,* 2018; Aylward *et al.,* 2018; Xia *et al.,* 2018; Fujii *et al.,* 2019). Applying this method to fecal samples is still in the discovery phase and requires additional optimization; to date, few studies have used this method to study wild populations, mostly primates (Perry *et al.,* 2010; De Manuel *et al.,* 2016; Hernandez-Rodriguez *et al.,* 2018). In this chapter, we report on a DNA capture trial using fDNA extracted from snow leopard scats, and we discuss problems identified and some recommendations for improving the method.

2.3. MATERIALS AND METHODS

2.3.1. Sampling and DNA extraction

For this study, 96 scat samples were used (Appendix 1: List of samples). Fifty-three (53) samples were collected from different locations of Gilgit-Baltistan (Pakistan) between December 2015 and April 2016. Forty-three (43) were obtained from the archives at the American Museum of Natural History, which were collected by WCS-China, WCS-Afghanistan, Panthera, and Snow Leopard Trust. Eleven samples were from Tajikistan and Mongolia each, 10 were from China, 9 from Kyrgyzstan, and 1 from Nepal and

Afghanistan each. QIAamp Fast DNA Stool Mini Kit (Cat No./ID: 19593) was used to extract DNA from fecal samples. The quantity and quality of extracted DNA was checked on a NanoDrop (Thermo Scientific).

2.3.2. DNA Capture Method

Arbor Biosciences designed probes for DNA capture (myBaits®) using the ddRAD-seq library developed from snow leopard samples (Janjua *et al.,* 2020). These myBaits probes are biotinylated oligonucleotides that are complementary to the target sequences and are used to separate the DNA of the target taxa from nontarget DNA. The two main steps of DNA capture are library preparation and capture. They are detailed below.

Step 1: Library Preparation

DNA extracted from scats was sheared using a Covaris ultra-sonicator (Covaris, MA, USA) to generate double-stranded DNA (dsDNA) fragments of ~250 bp. We used 52.5 µL of DNA in Covaris specific tubes, microTUBE-50 AFA Fiber Screwcap, for this shearing step with a total cycle of 130s. This sheared dsDNA was used for library preparation using Illumina® TruSeq® Nano DNA Library Prep kits by following instructions provided in the user manual. Briefly, the sheared DNA was cleaned using sample purification beads, and the ends of fragments were repaired with End Repair Mix (ERP). This step clips off or add nucleotides to the end of sheared fragments to produce blunt ends. Then a single adenine (A) nucleotide was attached on the 3' ends to inhibit ligation of fragments to each other. The addition of an adenine makes a sticky end DNA fragment for ligation of adaptors that have complementary thiamine (T) on their 3' end.

Finally, PCR enrichment provided us with multiple copies of fragments that were quantified using qPCR, and these fragments were used for the subsequent step of DNA capture.

Step 2: Capture or Hybridization of probes

The enriched library was used for hybridization of probes by following protocols provided in the user manual (myBaits® Manual Version: v4 -4.01). Briefly, the enriched library was allowed to hybridize with probes at 55^{o}C overnight. These captured fragments were single-stranded DNA; therefore, amplification of the second strand was completed using Dynabeads™ MyOne™ Streptavidin C1 beads with KAPA HiFi DNA Polymerase (Kapa Biosystems, USA) and the following conditions; 98 °C for 30 s, followed by 5 cycles of 98 °C for 20 s, 68 °C for 30 s and 72 °C for 45 s, and a final extension of 5 min at 72 °C, using primers specific to the attached adaptors (i.e., i5 and i7 at the ends of the Illumina library (Glenn *et al.,* 2016)).

2.3.3. Data analysis:

All libraries were dual barcoded and de-multiplexed based on perfect barcode matches only, limiting any potential of barcode sequence jumping. Across samples, Geneious version 10 (http://www.geneious.com; Kearse *et al.,* 2012) was used to merge paired-end files, filter for quality (phred score ≥ 30) per base and total sequence (average phred score ≥ 30), as well as remove PCR duplicates. All samples were then analyzed for per base sequence quality, per sequence GC content, sequence duplication levels, overrepresented sequences, and adapter content in FastQC version 0.11.5 (Andrews 2010). Any samples not passing FastQC quality checks were further filtered or discarded. All pass-filter, merged and unmerged, sequences were retained for downstream processing. Burrows Wheeler Aligner v. 07.12 (bwa; Li and Durbin, 2009) was used to index and align per sample sequences to our reference – comprised of initial ddRAD sequences used to build the capture array (Janjua *et al.,* 2020) – and using optimized parameters as described in Schubert *et al*. (2012: -T 25 -I 1024). Samples were then sorted and indexed in Samtools v. 1.6 (Li *et al.,* 2009) and then combined using the "mpileup" function with following parameters "-c –A -Q 30 -q 30." Final variant calling was done using program bcftool s v. 1.6 (Li *et al.,* 2009). VCF files for each marker, as well as concatenated autosomal and Z-chromosome markers were converted to fasta files using the program PGDspyder v. 2.1.1.2 (Lischer and Excoffier, 2012). In short, base pairs were retained based on sequence coverage of a minimum allele depth of 5x (10x per genotype) and quality Phred scores of ≥30. Using custom scripts, each FASTA file was further filtered for samples having < 50% sequence coverage, and base positions having < 80% of sample representation.

2.4. RESULTS

2.4.1. DNA Extraction: Quantity and Quality

Quantities of DNA extracted from all samples ranges from a very low concentration of 0.01 ng/µL to a reasonably high concentration of 433 ng/µL. Samples collected in recent years from Pakistan in 2015-16, had significantly higher mean concentrations (17.36 ± 32.09) compared to samples collected from other locations in 2008-12 (2.25 ± 5.78; t-test, t = -3.30, df = 54, *P* = 0.001; Fig. 2.1). However, the quality ratio (A260/A280) of the two groups was not significantly different (Fig. 2.1; 1.34 for 2015-16, and 1.46 for 2008-12; t-test, t = 1.23, df = 71, P = 0.22).

All samples were categorized into three main categories based on initial DNA concentrations (<10 ng/µL, 10 ng/ µL -100 ng/ µL, and >100 ng/µL). The frequency of samples falling in each category were 78.13%, 18.75% and 3.12%, respectively. Furthermore, 64.15% of Pakistani samples fall within the <10 ng/µL category, 30.18% in the 10 ng/µL - 100 ng/µL category, and 5.66% in the >100 ng/µL category. In contrast, 95.34% of samples collected from other regions have DNA <10 ng/µL, 4.65% 10 ng/µL - 100 ng/µL and none of the sample falls in >100 ng/µL category. These results suggest that older samples have less DNA compared to samples collected recently (Figure 2.1).

2.4.2. DNA Capture: Quantity

DNA extracted from each sample was processed for DNA capture and quantified before sending sample for final sequencing. Figure 2.2 is a whisker plot showing comparison of initial DNA concentration and quantities after DNA capture. Samples with initial concentrations <10 ng/µL DNA have low post-capture DNA concentration, whereas those with initial concentrations of >100 ng/µL have high DNA concentration after capture. The middle group has wider variation but most of samples have higher post-capture concentrations compared to the first group and less than the third group. However, the correlation coefficient (R^2 = 0.195), indicates that initial DNA concentration is not strongly correlated with after capture concentration. Four out of nine outliers with DNA concentration of <10 ng/µL were samples collected in 2008-12 and the remaining five were recently collected (2015-16) samples.

2.4.3. Comparison of DNA Concentration and Number of Reads

The number of reads obtained after initial filtration was compared with the initial DNA concentration. DNA concentration in the 10-100 ng/µL range had the highest number of reads (Figure 2.3). The R^2 value from a linear regression was 0.0107, indicating that the number of reads was not related to the initial DNA quantity.

2.4.4. Comparison of DNA Concentration and Percent Target Sequences

To extract target sequence reads from the total number of reads for each sample, we aligned sequenced reads with the ddRAD-seq loci from Janjua *et al*. (2020) (see Chapter 1). The target sequence reads were very low, ranging from 0.002% to 47.34%. Percentage target is not correlated with DNA quantity (R^2 = 0.0083; Figure 2.4). Unexpectedly the percentage of target reads is lowest in DNA >10ng/µL. However, ANOVA indicated that the percentage of target reads did not differ significantly among the three categories of initial concentration (p = 0.607). Overall, 64.1% of samples have target DNA reads less than 1% while 34.8% have target reads ranging from 1-16% (Figure 2.5).

2.5. DISCUSSION

There is no doubt about the wide application potential of NGS in large-scale genomic studies. However, the fields of conservation, genetics, ecology, evolution, etc. of nonhuman, non-model organisms, especially research on natural populations of endangered mammals, have yet to benefit extensively from the recent availability of next generation sequencing technologies. One of the major impediments to such studies is low quality DNA. Based on challenges and risks associated with collection of invasive samples (like blood and tissue), researchers often opt for non-invasive sampling. In principle, DNA for genetic analyses can also be isolated from non-invasive samples such as feces and shed hair. Such samples can be collected readily without harm (sometimes even without direct observation of the animal) and are thus ideal in many respects for genetic studies of natural populations. However, this non-invasive source of DNA is not easy to work with because it is often fragmented and degraded due to oxidative hydrolytic damage (i.e. cytosine deamination) (Paabo *et al.,* 1989; Shapiro and Hofreiter, 2012). Moreover, it is often low in quantity, and contaminated by exogenous DNA sources and chemicals that are potential inhibitors. These limitations make PCR based analyses more challenging (Taberlet *et al.*, 1999).

Based on limitations of non-invasive samples, traditional techniques to study population genetics have largely been restricted to mitochondrial DNA (mtDNA) sequencing and microsatellite-based genotyping. However, these traditional methods must deal with misidentification and allelic dropout-related challenges (Arandjelovic *et al.*, 2009; Buchan *et al*. 2005; McKelvey and Schwartz 2004). While such work has provided important insights into taxonomy, population structure, and the relationship between relatedness and behavior in natural populations in a number of species (e.g., Kohn and Wayne, 1997; Piggott and Taylor, 2003; DeSalle and Amato, 2004; Vigilant and Guschanski 2009), a new approach is required for genomic-level analyses of non-invasive DNA that will facilitate large-scale genetic studies in natural populations.

Research and development in NGS-based techniques is trying to overcome several traditional limitations associated with DNA from non-invasive samples. To deal with low quantity, scientists have developed methods that are sensitive to low quality and quantity of DNA (Miotto *et al.,* 2012, Wultsch *et al.,* 2014, Russello *et al.,* 2015). DNA extracted from such samples is thought to be overwhelmed by DNA from exogenous sources, like gut bacteria (Stephen and Cummings, 1980), prey species, plants, and other environmental contaminants. Target enrichment is one approach to separate endogenous DNA from the (often) large amounts of exogenous DNA. The ability to select and target the appropriate markers would represent a powerful new tool for molecular ecology, phylogenetics, archaeology and biomedical studies.

These NGS technologies require only a small quantity (often less than 10 ng) of DNA and they require fragmented DNA, therefore obviating the problems associated with fecal DNA based genetic analyses. DNA capture is one of the NGS based method which use target specific probes to enrich the endogenous DNA against the exogenous DNA. The ability to carry out such studies would represent a powerful new tool for conservation and evolutionary ecology studies (Kohn *et al*. 2006; Ouborg *et al*. 2010).

Despite the potential application and implications of DNA capture, its use in wild population is still in discovery and optimization phase. Limited data are available on the method for studying wild populations and especially large carnivores. Our probes designed using a snow leopard library developed by ddRAD-seq (Janjua *et al.,* 2020) failed to give the desired information. We have seen that initial DNA quantities of less than 10 ng/µL were unlikely to yield a good quantity of DNA after capture (Figure 2.2) and so it affected the number of reads obtained from those samples (Figure 2.2). The problem of low quantity of DNA can be addressed by performing multiple DNA extractions per sample if possible (Hernandez-Rodriguez *et al.,* 2018). Unfortunately, we were short of samples and relied on a low quantity of DNA for this optimization run. It is also important to mention here that samples with DNA quantity greater than 100 ng/µL also had low performance in the number of reads compared to samples that were between 10 – 100 ng/µL. This is supported by previous studies using different molecular techniques to study non-invasive samples

(Arandjelovic and Vigilant, 2018; Carroll *et al.*, 2018). One of the problems with our data is also due to pooling of samples before sequencing run. Hernandez-Rodriguez *et al*. (2018) suggested that in order to minimize drowning of low quantity of DNA in pool of high-quality DNA, one should pool similar quantity samples thus doing multiple runs of sequencing by sub-pooling the samples based on initial DNA concentrations.

Furthermore, the number of probes selected for targeting nuclear regions was too high compared to numbers reported in a handful of studies (O'Leary *et al.*, 2018; Bose *et al.,* 2018). One of the biggest reasons for selecting higher number of probes was that we were using reference sequence library developed for snow leopard developed from five samples from Mongolian population (Janjua *et al.*, 2020). We assumed that the regions that are variable in these samples might not be variable in other samples from other populations or even from the same population. This idea led us to include all conserved ~3900 ddRAD-seq loci for probes designing, but this approach might have contributed to the low success of our protocol. Higher numbers of probes likely decreased the coverage and depth of results. Regardless, the number of reads from 87.5% of samples fall in the range of 2,047,91 and 14,885,356 which is a good number for successful studies using NGS in wild population studies (Cosart *et al.,* 2011; Andrews *et al.,* 2018).

Another problem with the designed probes was recorded while processing data through the bioinformatics pipeline with 5X coverage. We retained only 18 loci that met this threshold, and when those loci were aligned with the tiger genome, we found multiple hits indicating that multiple copies were present for these loci. There are different kinds of repetitive sequences present in eukaryotic cells including transposable elements, satellite

DNAs, simple sequences, and tandem repeats (Kidwell, 2002). In addition, the phenomenon responsible for this gene duplication can be result of ectopic recombination, retro transposition event, aneuploidy, polyploidy, and replication slippage (Conant and Wolfe, 2008; Flagel and Wendel, 2009), each of which is an interesting phenomenon to study evolutionary history and can be later targeted for study of snow leopards. However, our data were not adequate for studying such perspectives at this point. These repetitive sequences in our capture data are making it difficult to differentiate true homologues and heterologues.

We have redesigned our capture array to target fewer loci and to remove all loci that had multiple hits in the tiger genome. Doing so should increase the number of reads per target locus and by-pass the problem associated with heterologues. This capture array still needs to be tested, however.

Percentage of target reads ranged from 0 to 16%, which is very common is non-invasive samples, especially fDNA (Carroll *et al.*, 2018; Arandjelovic and Vigilant, 2018; Hernandez-Rodrigues *et al.,* 2018). There are some methods reported to improve the target yield, such as target the methylation based (Chiou *et al.,* 2018), where differences in CpG-methylation sites between eukaryotes and prokaryotic bacterial genomes is targeted to preferentially bind with host DNA and leaving behind bacterial DNA, and additional or prolonged hybridization step (Hernandez-Rodrigues *et al.,* 2018).

In conclusion, DNA capture method needs optimization for studying population genetics of snow leopards. After this first trial, we know we need to re-design probes to improve coverage and depth. To enhance endogenous DNA reads we can manipulate the library preparation and DNA capture step like increasing the capture hybridization step timing or adding additional cycles. These modifications can help us to get more useful data from fDNA that can help us studying snow leopard populations.

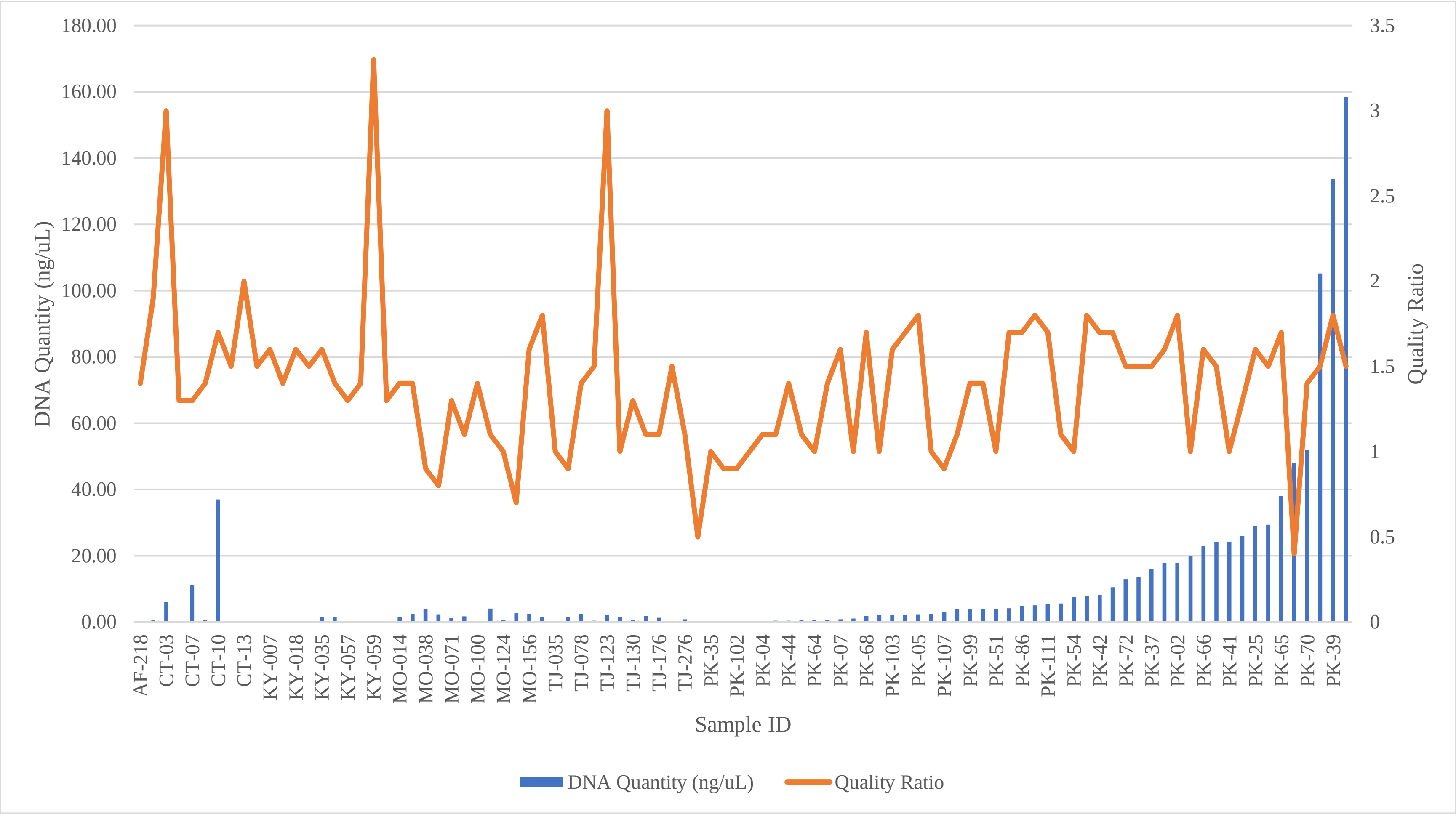

Figure 2.1: DNA quantity in ng/µL (blue bars; primary y-axis) and quality ratio (orange line; secondary y-axis) of DNA extracted from scats of snow leopards.

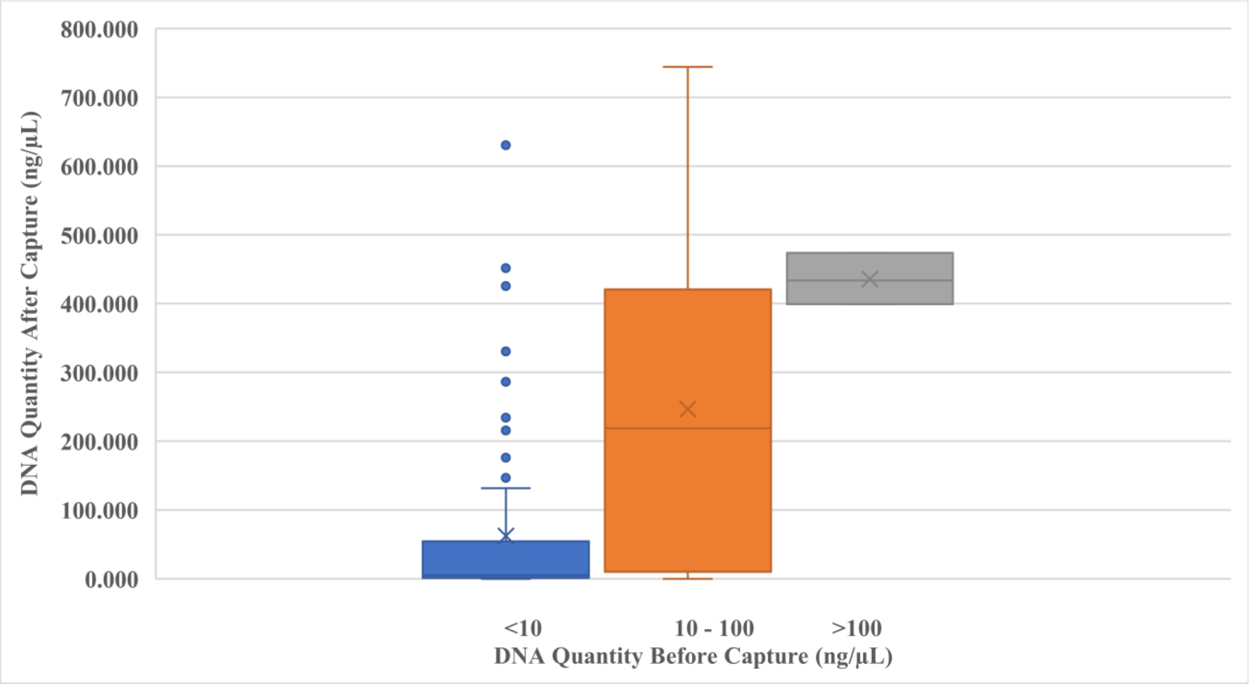

Figure 2.2: Box and whisker plot of initial DNA quantities and DNA concentrations after capture. Samples categorized in lowest DNA quantity (<10 ng/µL) have lowest mean (~50ng/µL) with nine outliers. Means for 10-100 ng/µL and > 100 ng/µL have means ~225 ng/µL and ~425 ng/µL.

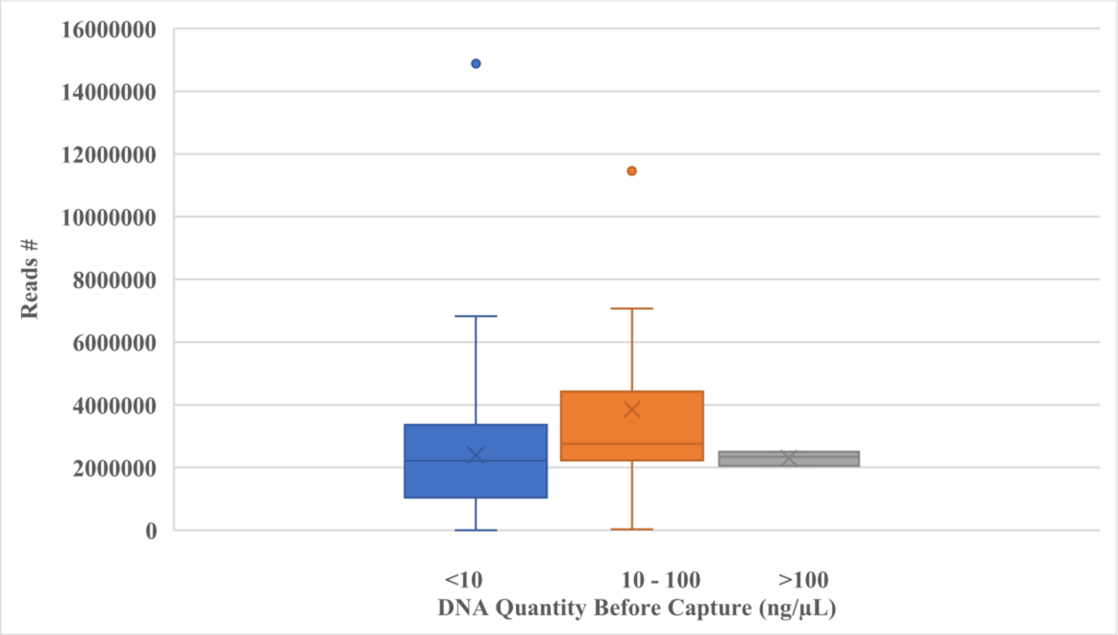

Figure 2.3: Box and whisker plot of initial DNA quantities and number of reads obtained after filtering sequence data. Mean number of reads for all three DNA quantity categories ranges from ~2100000 – 4000000 number of reads.

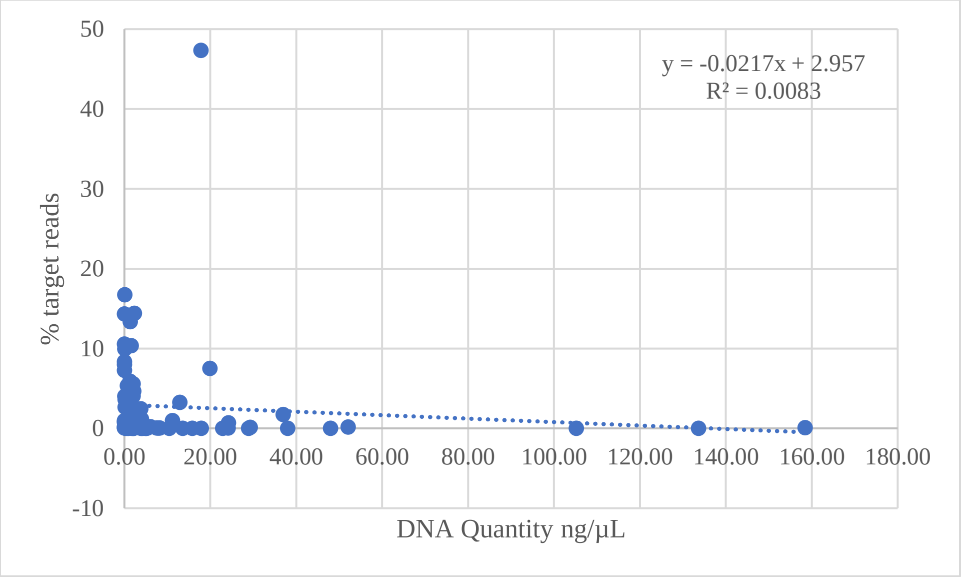

Figure 2.4: Correlation between DNA quantities before capture (ng/µL) and percentage target reads.

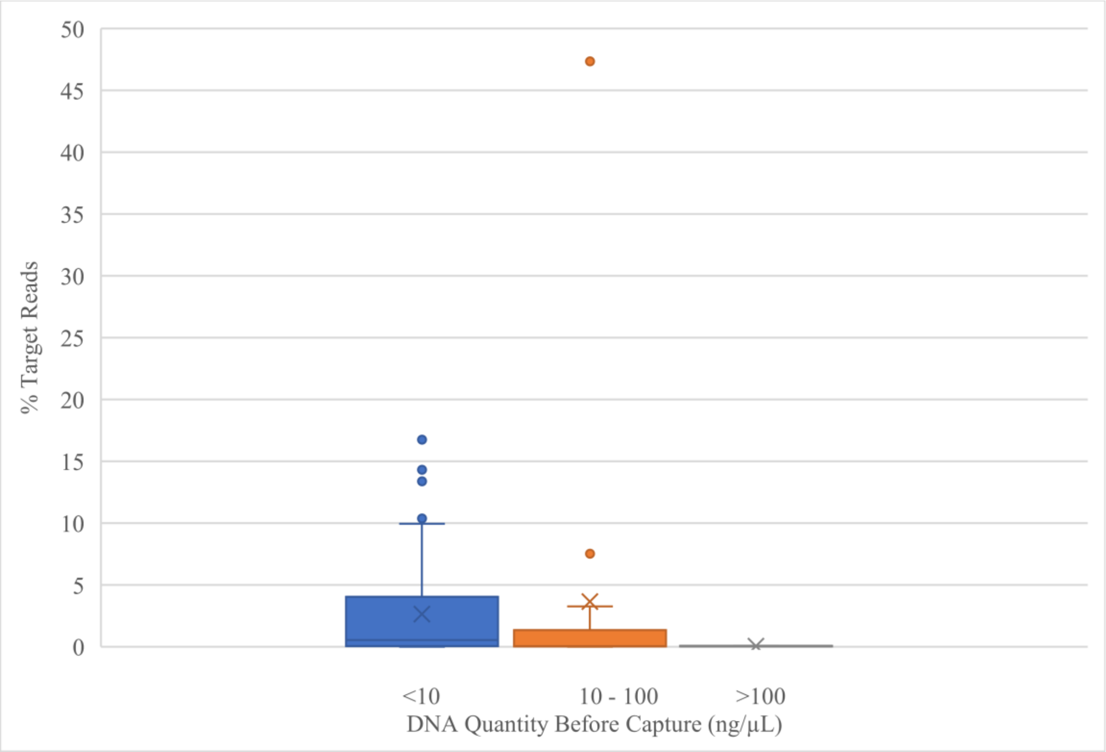

Figure 2.5: Box and whisker plot of initial DNA quantities and percent target reads. Mean % target reads ranges from ~1% to 4%. Highest % target reads were observed in samples in <10 ng/µL DNA group.

2.6. LITERATURE CITED

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J. and Weinstock, G.M. (2007). Direct selection of human genomic loci by microarray hybridization. Nature methods, 4(11), 903-905.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., and Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics, 17(2), 81.

Andrews, K.R., Adams, J.R., Cassirer, E.F., Plowright, R.K., Gardner, C., Dwire, M., Hohenlohe, P.A. and Waits, L.P., (2018). A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RAD seq data. Molecular ecology resources, 18(6), 1263-1281.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Arandjelovic, M., and Vigilant, L. (2018). Non-invasive genetic censusing and monitoring of primate populations. American Journal of Primatology, 80(3), e22743.

Avila-Arcos, M.C., Cappellini, E., Romero-Navarro, J.A., Wales, N., Moreno-Mayar, J.V., Rasmussen, M., Fordyce, S.L., Montiel, R., Vielle-Calzada, J.P., Willerslev, E. and Gilbert, M.T.P. (2011). Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. Scientific reports, 1, 74.

Aylward, M. L., Johnson, S. E., Sullivan, A. P., Perry, G. H., and Louis, E. E. (2018). A novel environmental DNA (eDNA) sampling method for aye-ayes from their feeding traces. bioRxiv, 272153.

Bose, N., Carlberg, K., Sensabaugh, G., Erlich, H., and Calloway, C. (2018). Target capture enrichment of nuclear SNP markers for massively parallel sequencing of degraded and mixed samples. Forensic Science International: Genetics, 34, 186-196.

Buchan, J. C., Archie, E. A., Van Horn, R. C., Moss, C. J., and Alberts, S. C. (2005). Locus effects and sources of error in noninvasive genotyping. Molecular Ecology Notes, 5(3), 680-683.

Carroll, E. L., Bruford, M. W., DeWoody, J. A., Leroy, G., Strand, A., Waits, L., and Wang, J. (2018). Genetic and genomic monitoring with minimally invasive sampling methods. Evolutionary applications, 11(7), 1094-1119.

Chiou, K. L., and Bergey, C. M. (2018). Methylation-based enrichment facilitates low-cost, noninvasive genomic scale sequencing of populations from feces. Scientific reports, 8(1), 1975.

Conant, G. C., and Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nature Reviews Genetics, 9(12), 938-950.

Cosart, T., Beja-Pereira, A., Chen, S., Ng, S. B., Shendure, J., and Luikart, G. (2011). Exome-wide DNA capture and next generation sequencing in domestic and wild species. BMC genomics, 12(1), 347.

Cruz-Dávalos, D.I., Nieves-Colón, M.A., Sockell, A., Poznik, G.D., Schroeder, H., Stone, A.C., Bustamante, C.D., Malaspinas, A.S. and Ávila-Arcos, M.C., (2018). In-solution Y-chromosome capture-enrichment on ancient DNA libraries. BMC genomics, 19(1), 608.

De Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V.C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P. and Schmidt, J.M. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science, 354(6311), 477-481.

De Wit, P., Pespeni, M. H., and Palumbi, S. R. (2015). SNP genotyping and population genomics from expressed sequences–current advances and future possibilities. Molecular ecology, 24(10), 2310-2323.

DeSalle, R., and Amato, G. (2004). The expansion of conservation genetics. Nature Reviews Genetics, 5(9), 702-712.

Ekblom, R., and Wolf, J. B. (2014). A field guide to whole-genome sequencing, assembly and annotation. Evolutionary applications, 7(9), 1026-1042.

Flagel, L. E., and Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. New Phytologist, 183(3), 557-564.

Fujii, K., Doi, H., Matsuoka, S., Nagano, M., Sato, H., and Yamanaka, H. (2019). Environmental DNA metabarcoding for fish community analysis in backwater lakes: A comparison of capture methods. PloS one, 14(1), e0210357.

Glenn, T.C., Nilsen, R.A., Kieran, T.J., Finger, J.W., Pierson, T.W., Bentley, K.E., Hoffberg, S.L., Louha, S., García-De León, F.J., del Rio Portilla, M.A. and Reed, K.D. (2016). Adapterama I: universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru and iNext). BioRxiv, 049114.

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. and Gabriel, S. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature biotechnology, 27(2), 182-189.

Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., Angedakin, S., Casals, F., Navarro, A., Vigilant, L. and Kühl, H.S. (2018). The impact of endogenous content, replicates and pooling on genome capture from faecal samples. Molecular ecology resources, 18(2), 319-333.

Janjua, S., Peters, J. L., Weckworth, B., Abbas, F. I., Bahn, V., Johansson, O., and Rooney, T. P. (2020). Improving our conservation genetic toolkit: ddRAD-seq for SNPs in snow leopards. Conservation Genetics Resources, 12, 257-261.

Jones, M. R., and Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. Molecular ecology, 25(1), 185-202.

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

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. Genetica, 115(1), 49-63.

Kohn, M. H., and Wayne, R. K. (1997). Facts from feces revisited. Trends in ecology and evolution, 12(6), 223-227.

Kohn, M. H., Murphy, W. J., Ostrander, E. A., and Wayne, R. K. (2006). Genomics and conservation genetics. Trends in ecology and evolution, 21(11), 629-637.

Lavretsky, P., Janzen, T., and McCracken, K. G. (2019). Identifying hybrids and the genomics of hybridization: Mallards and American black ducks of Eastern North America. Ecology and Evolution.

Li, H., and Durbin, R. (2009). Making the Leap: Maq to BWA. Mass Genomics, 25, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.s (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Lischer, H. E., and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics, 28(2), 298-299.

McKelvey, K. S., and Schwartz, M. K. (2004). Genetic errors associated with population estimation using non-invasive molecular tagging: problems and new solutions. The journal of wildlife management, 68(3), 439-448.

Miotto, R. A., Cervini, M., Begotti, R. A., and Galetti Jr, P. M. (2012). Monitoring a puma (Puma concolor) population in a fragmented landscape in southeast Brazil. Biotropica, 44(1), 98-104.

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., and Portnoy, D. S. (2018). These aren’t the loci you’e looking for: Principles of effective SNP filtering for molecular ecologists. Molecular ecology, 27(16), 3193-3206.

Ouborg, N. J., Pertoldi, C., Loeschcke, V., Bijlsma, R. K., and Hedrick, P. W. (2010). Conservation genetics in transition to conservation genomics. Trends in genetics, 26(4), 177-187.

Paabo, S. (1989). Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proceedings of the National Academy of Sciences, 86(6), 1939-1943.

Perry, G. H., Marioni, J. C., Melsted, P., and Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. Molecular Ecology, 19(24), 5332-5344.

Pickrell, J. K., and Reich, D. (2014). Toward a new history and geography of human genes informed by ancient DNA. Trends in Genetics, 30(9), 377-389.

Piggott, M. P., and Taylor, A. C. (2003). Remote collection of animal DNA and its applications in conservation management and understanding the population biology of rare and cryptic species. Wildlife Research, 30(1), 1-13.

Richards, S.M., Hovhannisyan, N., Gilliham, M., Ingram, J., Skadhauge, B., Heiniger, H., Llamas, B., Mitchell, K.J., Meachen, J., Fincher, G.B. and Austin, J.J. (2019). Low-cost cross-taxon enrichment of mitochondrial DNA using in-house synthesised RNA probes. PloS one, 14(2), e0209499.

Russello, M. A., Waterhouse, M. D., Etter, P. D., and Johnson, E. A. (2015). From promise to practice: pairing non-invasive sampling with genomics in conservation. PeerJ, 3, e1106.

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J.F., Al-Rasheid, K.A., Willerslev, E., Krogh, A. and Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. BMC genomics, 13(1), 178.

Shapiro, B., Hofreiter, M., LA, J., and ER, A. (Eds.). (2012). Ancient DNA: methods and protocols (p. 247). New York: Humana Press.

Stephen, A. M., and Cummings, J. H. (1980). Mechanism of action of dietary fibre in the human colon. Nature, 284(5753), 283-284.

Taberlet, P., Waits, L. P., and Luikart, G. (1999). Noninvasive genetic sampling: look before you leap. Trends in ecology and evolution, 14(8), 323-327.

Vigilant, L., and Guschanski, K. (2009). Using genetics to understand the dynamics of wild primate populations. Primates, 50(2), 105-120.

Wasef, S., Huynen, L., Millar, C.D., Subramanian, S., Ikram, S., Holland, B., Willerslev, E. and Lambert, D.M., (2018). Fishing for Mitochondrial DNA in The Egyptian Sacred Ibis Mummies. bioRxiv, 473454.

Wilcox, T. M., Zarn, K. E., Piggott, M. P., Young, M. K., McKelvey, K. S., and Schwartz, M. K. (2018). Capture enrichment of aquatic environmental DNA: A first proof of concept. Molecular ecology resources, 18(6), 1392-1401.

Wultsch, C., Waits, L. P., and Kelly, M. J. (2014). Noninvasive individual and species identification of jaguars (Panthera onca), pumas (Puma concolor) and ocelots (Leopardus pardalis) in Belize, Central America using cross-species microsatellites and faecal DNA. Molecular ecology resources, 14(6), 1171-1182.

Xia, Z., Zhan, A., Gao, Y., Zhang, L., Haffner, G. D., and MacIsaac, H. J. (2018). Early detection of a highly invasive bivalve based on environmental DNA (eDNA). Biological invasions, 20(2), 437-447.

Chapter 3

DIET ANALYSIS OF SNOW LEOPARDS USING NON-INVASIVE GENETIC SAMPLING

3.1. ABSTRACT

Food and food chain studies are important for planning conservation strategies. Determining prey items for elusive predators, such as snow leopards (*Panthera uncia*), require scat analysis. We obtained non-invasive fecal scats of snow leopards from seven countries, and applied DNA capture, a widely used method for studying human populations, ancient DNA samples, eDNA, and forensics. DNA capture can provide information about genetic diversity at targeted loci, but a large amount of non-target DNA (60-90% of all sequences) are also captured as a byproduct of the method. This non-target DNA is typically filtered out using bioinformatic pipelines. Here, we present how we can use this non-target data, specifically mtDNA, to identify prey items in snow leopard scats that are collected from different regions/locations. We used two methods, method I where we identified prey using five mitochondrial genes (COI, COIII, Cytb, ND2 and ND4) and method II where whole genomes of candidate prey items was used to identify prey items in scats of snow leopards. Overall, wild ungulates (67.5-80%) dominated the snow leopard diet. Domestic prey items constituted 20 - 32.5%. The regional picture of prey items in snow leopard scats will help us understand the feeding ecology of the snow leopard across most of its range and can help us address the conservation and management issues pertaining to this wild cat.

3.2. INTRODUCTION

Snow leopards are elusive predators, inhabiting mountains of Central and South Asia (Jackson *et al.,* 2010). Being apex predators, they play a prominent role in shaping ecosystems and serve as an indicator of healthy eco-regions (Ripple *et al.*, 2001; Dalerum *et al.*, 2008; Ripple and Beschta, 2012; Janecka *et al.*, 2020). As they are difficult to observe and follow, there is a lack of published information on many ecological aspects of snow leopards, including diet, in comparison to other charismatic large carnivores (Lyngdoh *et al.*, 2014; Shehzad *et al.,* 2012 and Anwar *et al.*, 2011). Knowledge of species' diet is key to understanding trophic interactions, and this information is critical for proper conservation strategies. Classical methods using microscopic identification of prey items (Khatoon, 2017; Suryawanshi *et al.*, 2017; Aryal *et al.*, 2014) and DNA barcoding (Chaves *et al.*, 2012) have inherent limitations. The former method is more subjective and highly dependent on the skills of the person identifying the undigested pieces of food items in the samples; furthermore, identification can be limited to the family or genus level and be dependent on species availability data for the area (Shehzad *et al.*, 2012, Lyngdoh *et al.*, 2014 and Janecka *et al.*, 2020). Barcoding is sensitive to the quantity and quality of DNA extracted from scat, which is often very low (Hajibabaei *et al.*, 2006).

Recent advances in next generation sequencing (NGS) enable reliable, high-resolution data (Ekblom and Galindo, 2011 and Cereb *et al.*, 2015). Molecular dietary analysis involving identification of prey, using specific genes (mostly mitochondrial DNA, mtDNA, genes such as cytochrome oxidase I or cytochrome *b*) extracted from predator feces or gut contents (Clare *et al.,* 2009; Eitzinger *et al.*, 2018; Symondson, 2002; Vesterinen *et al.,* 2016), can be useful in this regard. High-throughput sequencing (HTS) enables identification of species by simultaneously sequencing specimens of prey taxa in bulk mixtures (Gibson *et al.,* 2014; Hajibabaei *et al.,* 2006; Meusnier *et al.,* 2008; Pompanon *et al.,* 2012), making the diet analyses faster and cost-effective. Certainly, the use of DNA metabarcoding has the potential to revolutionize ecological studies (Zinger *et al.,* 2019).

Using RNA probes of snow leopard, designed from a ddRAD-seq library (Janjua *et al.,* 2020), ~95-99% of our data were non-targeted sequences, which is a common problem with DNA capture methods (Hawkins *et al.,* 2016). This problem occurs because endogenous DNA in scat samples is very low (~1-5%) compared to exogenous DNA (Hernandez-Rodriguez *et al.,* 2018; Perry *et al.,* 2010). Here we aimed to extract information from non-targeted DNA sequences obtained from our target DNA capture run (Chapter 2). A large quantity (~95-99%) of the non-target sequences was mtDNA, including sequences form both the snow leopard and its prey. Our analysis of mtDNA from scats provides information on diet of snow leopards from seven countries, viz, Afghanistan, China, Kyrgyzstan, Mongolia, Nepal, Pakistan and Tajikistan. In this paper we provide a modified bioinformatic pipeline to extract prey species information from the sequence data available on this identification from different locations/regions.

3.3. MATERIALS AND METHODS

3.3.1. Sampling and DNA extraction

For this study, 63 scat samples were used (Appendix 1: List of samples), 24 collected from different locations of Gilgit-Baltistan (Pakistan) between December 2015 and April 2016, and 39 obtained from the archives at the American Museum of Natural History (China, *N* = 10; Kyrgyzstan, *N* = 9; Mongolia, *N* = 8; Tajikistan, *N* = 10; Nepal, *N*

= 1; and Afghanistan, *N* = 1). QIAamp Fast DNA Stool Mini Kit (Cat No./ID: 19593) was used to extract DNA from the fecal samples. The quantity and quality of extracted DNA was checked using a NanoDrop (Thermo Scientific).

3.3.2. DNA Capture Method

For DNA capture, we used myBaits® (Arbor Biosciences) probes that were designed from the ddRAD-seq library developed from snow leopards (Janjua *et al.,* 2020). These probes are biotinylated oligonucleotides that are complementary to the target sequences and are used to separate DNA of target taxa from nontarget DNA. Two main steps of DNA capture, library preparation and capture, are detailed below.

Step 1: Library preparation

DNA extracted from scats was sheared using Covaris ultra-sonicator (Covaris, MA, USA) to generate DNA fragments of ~250 bp. This sheared dsDNA was used for library preparation using Illumina® TruSeq® Nano DNA Library Prep kits following instructions provided in the user manual. Briefly, sheared DNA was cleaned up using sample purification beads and the ends of fragments were repaired. Then single adenine is added on the 3’ ends of fragments which makes a sticky end DNA fragment, facilitating the ligation of adaptors. Finally, PCR enrichment provided multiple copies of fragments that were quantified using qPCR, which were used for the subsequent step.

Step 2: Capture or hybridization of probes

The enriched library was used for hybridization of probes following protocols provided in the user manual (myBaits® Manual Version: v4 -4.01). Briefly, the library was allowed to hybridize with probes at 55 ^{o}C overnight. These captured fragments were single-stranded DNA; therefore, amplification of the second strand was completed using Dynabeads™ MyOne™ Streptavidin C1 beads with KAPA HiFi DNA polymerase (Kapa Biosystems, USA) under the following conditions; 98 °C for 30 s, followed by 5 cycles of 98 °C for 20 s, 68 °C for 30 s and 72 °C for 45 s, and a final extension for 5 min at 72 °C using primers specific to adaptors attached i.e. i5 and i7 at the ends of the Illumina library (Glenn *et al.,* 2016).

3.3.3. Data Processing

All libraries were dual barcoded and de-multiplexed based on perfect barcode matches only, limiting any potential of barcode sequence jumping. Across samples, Geneious version 10 (http://www.geneious.com; Kearse *et al.,* 2012) was used to merge paired-end files, filter for quality (phred score ≥ 30) per base and total sequence (average phred score ≥ 30), as well as remove PCR duplicates. All samples were then analyzed for per base sequence quality, per sequence GC content, sequence duplication levels, overrepresented sequences, and adapter content in FastQC version 0.11.5 (Andrews 2010). Any samples not passing FastQC quality checks were further filtered or discarded. All pass-filter, merged and unmerged, sequences were retained for downstream processing.Burrows Wheeler Aligner v. 07.12 (bwa; Li and Durbin, 2009) program was used to index and align per sample sequences to our reference – comprised of initial sddRAD sequences used to build the capture array (Janjua *et al.,* 2020) – and using optimized parameters as described in Schubert *et al.* (2012: -T 25 -I 1024). Samples were then sorted and indexed in Samtools v. 1.6 (Li *et al.,* 2009) and then combined using the "mpileup" function with following parameters "-c –A -Q 30 -q 30." Final variant calling was done using program bcftools v. 1.6 (Li *et al.*, 2009). VCF files for each marker, as well as concatenated autosomal and Z-chromosome markers were converted to fasta files using the program PGDspyder v. 2.1.1.2 (Lischer and Excoffier, 2012). In short, base pairs were retained based on sequence coverage of a minimum allele depth of 5x (10x per genotype) and quality Phred scores of ≥30. Using custom scripts, each FASTA file was further filtered for samples having < 50% sequence coverage, and base positions having < 80% of sample representation.

3.3.4. Prey Identification

Method I

Sequences that did not align to our ddRAD-seq reference library were used to identify prey species in scats. We generated five reference databases comprising five barcode genes, commonly used for eukaryotes (COI, COIII, Cytb, ND2 and ND4). Barcode sequences were retrieved from NCBI GenBank using the pipeline developed by Porter *et al*. (2018). Sequences from bacteria, fungi and humans (*Homo sapiens*) were filtered out. Retrieved barcode sequences were used to build local reference databases using BLASTN 2.2.31+ suite. Then a local BLASTN search was done against all the reference databases. Each generated read was matched against barcode sequences with default settings and the threshold expected value (E-value) set to 0.001. Hits were filtered according to the criteria developed by Srivathsan *et al*. (2015), i.e. retaining the hits with pairwise identity higher than 98% and sharing a minimum of 50 bp overlap with reference barcodes. Among all the hits that passed the filter requirements, we inspected the best hit of each read to identify the species. The scripts used for this method are in Appendix 2.

***Method II*:**

We used a reference database generated using complete mitochondrial sequences of potential prey species of snow leopard across its range (List of species with their GenBank accession numbers is in Appendix 3). The list of potential prey was generated *a priori*, and based on potential prey list (Nyhus *et al.*, 2016). Snow leopard mtDNA was also included in this reference. Geneious V. 5.6.5 was to align sample reads with the reference database. Filtration criteria used to identify reads to prey species was:>99 bp length, pairwise identity >97%, E-value 10^{-50}, and >4 reads.

3.4. RESULTS

We obtained 276,124,239 reads from 63 samples, collected from 7 locations/regions. The reads included mtDNA of prey items, nuclear DNA and mtDNA of snow leopard, and bacterial and fungal DNA. Table 3.1 summarizes the number of reads obtained from each gene (method I) and genome (method II) of prey species and snow leopard.

3.4.1. Method I

Using the database generated, for five genes (COI, COIII, Cytb, ND2 and ND4), we found 8,155 confirmed reads for prey species from 59 snow leopard scats. For COI, 35.6% of samples were positive for prey species with 7,099 confirmed COI reads. For both COIII and ND2, 39% of samples were positive for prey items with 93 and 337 reads, respectively. ND4 had 120 reads from 40.7% of samples. Cytb was retrieved from the highest number of samples with 506 reads from 62.7% of samples.

Prey items identified from different regions are shown in Table 3.2. Of 10 samples from China, 3 had no prey items, 4 had a single prey item (*Pseudois nayaur*, blue sheep, *N* = 3; *Capra hircus*, goat, *N* = 1), while two prey items were identified in three samples (*Bos taurus* cow + *P. nayaur*, *N* = 2; and *Capra falconeri*, markhor + cow, *N* = 1). Out of 9 samples from Kyrgyzstan, 4 had no prey items identified, and 5 had a single prey item: *Capra sibirica*, Siberian ibex, *N* =1;, *Aythya ferina*, common pochard (a duck), *N* = 1; and plants, *N* = 3. In 8 Mongolian samples, 5 had no prey items identified, Single prey was identified in one (*C. sibirica*, Siberian ibex, N =1) and two prey items were identified in two samples (*Bos taurus*, cow + *Capra falconeri*, markhor, N = 2). No prey item was detected in the sample from Afghanistan. The Nepalese sample had *Vulpex vulpex* (red fox). In 10 of the 24 samples from Pakistan, no animal prey items were detected, but plants were confirmed in 5 samples. Single animal prey was identified 7 samples (*B. mutus,* yak, N = 4; *B. taurus*, cow, N = 2 and *Ovis aries,* domestic sheep, N = 1), while two prey items were identified in 7 samples (*B. taurus,* cow + *C. falconeri*, markhor, N = 5; *V. vulpex,* fox + *C. falconeri,* markhor, N = 1, *C. ibex,* ibex + *B. taurus,* cow, N = 1). Out of 7 Tajikistan samples 4 had no prey identified, three had single prey items (*Marmota himalayana,* Himalayan marmot, N = 1; *P. nayaur,* blue sheep, N = 1 and *B. mutus,* yak, N = 1) and plants were identified in one.

Overall (Figure 3.1) no prey item is identified in 31.3% of the samples, 14.9% had plants. Of the identified prey species 76% were wild species and 24% livestock species. Two unusual prey items (common pochard, a duck; and a fish species, not native to the region) were identified in two samples.

3.4.2. Method II

Prey identified in different regions are shown in Table 3.2. Out of ten Chinese samples, two had no prey items identified, four had a single prey identified (*Pseudois nayaur*, blue sheep, N = 4). Two prey items were identified in four samples (*Pseudois nayaur,* blue sheep + *Bos taurus,* cow, N = 2; *Pseudois nayaur,* blue sheep + *Martes foina*, stone marten, N = 1 and *Pseudois nayaur,* blue sheep + *Ovis aries,* domestic sheep, N = 1). From nine Kyrgyzstan samples, only one had *Capra sibirica* (Siberian ibex) while the remaining eight had no prey item identified. Five out of eight Mongolian samples had no prey items identified. While single prey was identified in two samples (*Capra sibirica*, Siberian ibex, N = 2) and two prey items were identified in one sample (*Capra sibirica*, Siberian ibex + *Bos taurus*, cow, N = 1). No prey was identified in the Afghanistan samples, and red fox was present in the sample from Nepal. From 24 samples from Pakistan, five had no prey items identified, single prey items was identified in 15 samples (*Capra falconeri,* markhor N = 8; *Bos taurus,* cow, N = 2; *Bos mutus,* yak, N = 4 and *Ovis aries* domestic sheep, N = 1). Four samples had more than one prey item: *Bos taurus,* cow + *Capra falconeri*, markhor, N = 1; , *Capra ibex,* alpine ibex + *Bos taurus*, cow, N = 1; *Bos taurus,* cow + *Capra falconeri,* markhor, N = 1 and *Vulpes Vulpes,* red fox + *Capra falconeri*, markhor, N = 1). Three out of seven Tajikstan samples had no prey items identified. Single prey was identified in three samples (*Capra sibirica,* Siberian ibex, N = 1; *Bos mutus*, yak, N = 1 and *Marmota himalayana,* Himalayan marmot, N =1). One sample had two prey items: *Ovis ammon,* Argali + *Ovis orientalis,* Asian mouflon, N = 1.

Overall (Figure 3.2) no prey item was identified in 31.9% of samples. Of the identified prey items 78.0% were wild species whereas the remaining 22% were domestic livestock species. In comparison (Figure 3.3), blue sheep, domestic cow, domestic goat, ibex, markhor, marmot, red fox, yak and domestic sheep were identified using both methods, while duck, fish, goat and plants were unique to method I, and argali, Asian mouflon and stone marten were unique to method II.

3.5. DISCUSSION

3.5.1. Snow Leopard Diet

Information about the diet is very important for understanding the biology of a species and its role in ecosystems. The results from this study are in accord with previous studies (Table 3.3); ungulates (wild and domestic) are the major prey items of the snow leopard (method I: 71.4% and method II: 91.1%). Wild prey items are in higher proportion (67.5%-80.0%) than domestic prey items (32.5%-20%). Among ungulates, 62.9% and 78.0% are wild, whereas 37.1% and 21.9% are domestic (method I and method II, respectively). Wild ungulates include blue sheep from China, markhor from Pakistan, Siberian ibex from Kyrgyzstan, Mongolia and Tajikistan, yak from Pakistan and Tajikistan, Alpine ibex from Pakistan, and Argali and Asian mouflon from Tajikistan. The diets of snow leopards from China, Pakistan and Mongolia also include domestic cows. Domestic goat and sheep are present in samples from China and Pakistan, respectively. We don’t have information on prey item density in our sampling areas, but higher frequency of wild prey items in scats is indicative of snow leopards preferring wild prey compared to domestic prey, perhaps because domestic animals are often guarded by humans and pose risk to snow leopards. Furthermore, a major portion of the snow leopard population is present in arduous snow-covered tracts that are difficult to survey, and therefore, we expect the scat sampling to be biased towards gentler areas falling close to the human habitation.

This may suggest that wild species probably contribute a higher proportion of the snow leopard's diet than suggested by our analyses. The snow leopard may also consume dead animal or parts of dead animals, present in the garbage close to the human settlement or carcasses of livestock dying in accidents.

Small mammals (marmots, red fox), fish and duck have appeared as prey species of snow leopard. Bird species have been noted as snow leopard prey items in different studies (Mongolia: Shehzad *et al.,* 2012; Lhagvasuren and Mynkhtsog, 2000, Pakistan: Anwar *et al.*, 2011, India: Bagchi and Mishra, 2006; Chundawat and Rawat, 1994, Nepal: Oli *et al.,* 1994). Likewise, the presence of red fox in samples from Pakistan and Nepal is also consistent with previous reports from Nepal (Oli *et al.*, 1994) and India (Chundawat and Rawat, 1994). Presence of such animals in the diet of snow leopard scats/ diet probably represent opportunistic consumption of animals that are not the regular food of this predator species. Large and medium sized wild ungulates present at higher altitudes is the main source of food for snow leopard, which fits with the expectations of optimized energy gained and energy spent during hunting efforts.

In the majority (60.7%–73.0%) of samples, a single prey item was detected, supporting previous studies showing that individual snow leopards tend to focus primarily on a single prey item (Shehzad *et al.,* 2012, Oli *et al*. 1994 and Chundawat and Rawat, 1994) that provides sufficient resources for a few days. Snow leopard is reported to kill large prey every 10-15 days (Shehzad *et al.,* 2012 and McCarthy and Chapron, 2003) and consume it until its next kill. However, in 27-39% of samples, we detected multiple prey species in a single scat. These findings might represent instances where the snow leopard finished consuming one prey item and then switched to a second prey, or perhaps instances where the snow leopard opportunistically included a second prey item. Although possible, it seems unlikely that an individual snow leopard would feed from two major kills concurrently.

In concordance with previous reports from different regions (Pakistan: Anwar *et al.,* 2011, Nepal: Oli *et al.,* 1994, India: Chundawat and Rawat, 1994; Bagchi and Mishra, 2006, Mongolia: Lhagvasuren and Mynkhtsog, 2000), we observed plant material in samples coming from Pakistan, Tajikistan and Kyrgyzstan. It has been reported that carnivores eat plants to fulfill their needs for minerals and vitamins (Shehzad *et al.,* 2012), facilitate movement of food in gut, and for some medicinal purposes (Huffman, 2003). However, some studies (Bagchi and Mishra, 2006; Oli *et al.,* 1994; Lhagvasuren and Mynkhtsog, 2000) suggested that this ingestion is accidental while eating prey. Chances of contamination of wet and sticky snow leopard scats with plant material can also not be excluded.

No prey items were identified in 31.3% (method I) and 31.9% (methods II) of samples. The reasons for this could be that the DNA of prey items is very low and did not pass the set threshold to call it confirmed prey item. Alternatively, during intervals between meals (McCarthy and Chapron, 2003) scats would mainly contain hair from grooming and its own metabolic wastes.

3.5.2. NGS Based Methods

DNA-based techniques are useful in studying different aspects of wild and cryptic species like the snow leopard. Due to the high-resolution power of NGS based analysis, it is becoming popular for diet analysis of different vertebrate (Pompanon *et al.,* 2012;

Srivathsan *et al.,* 2015) and invertebrate species (Brown *et al.,* 2012). Most of these methods are robust and provide highly accurate and reliable information about the diet of species (Pompanon *et al.,* 2012). Traditional methods for diet analysis through scats were either unable to identify prey items or can only identify them to the genus level in the majority of cases. Especially distinguishing closely related wild ungulates like argali and sheep, wild yak and domestic yak, Capra spp., etc (Shehzad *et al.,* 2012). Here we used DNA capture method and utilized non-targeted sequences that were incidental captures. Both of our approaches for analyzing these data have advantages over previously used methods. Method I, where we used five barcoding genes (COI, COIII, Cytb, ND2 and ND4) was faster compared to method II, and included all the species present in the GenBank database. This helped in identifying the presence of unusual prey items like common pochard and gold mandarin fish sp. in some samples, which otherwise would not be possible when using species-specific primers for each prey item (Shehzad *et al.,* 2012). Method II was used keeping in mind the probability that due to low quantity and quality of DNA the genes used in method I are not amplified in quantity that can reach our set threshold for prey identification. This method required more time, on average two days per sample, and was dependent on our *a priori* reference database of mtDNA genomes available in GenBank. This method helps to find additional prey items that were not detected by method I, including stone marten, argali and Asian mouflon. This is because GenBank does not have separate entries for each of the five genes used in method I for these species. Thus, both analysis methods complement each other, with their gene specific and genome wide approach, and in combination, provide a wider breadth of information about the diet of snow leopards.

It is also important to mention here that samples from Mongolia, Kyrgyzstan and Tajikistan were collected in 2008 -2012 as part of different research projects. We were successfully able to identify prey items in these samples, indicating the high sensitivity of our method. However, recently collected samples from Pakistan have low percentage of unidentified prey (no prey) items compared to other regions (Table 3.3).

3.5.3. Conservation Implications

It is very important to have a clear understanding of feeding habits of snow leopard, an apex predator and species of conservation importance, for effective management and conservation planning. On the one hand, identifying the major diet components responsible for sustaining predator's population is important, by helping the understanding of relationship of predator-prey densities in the region, competition with other predators sharing the habitat and resources and most important interaction with local communities; on other hand, it is equally important to know prey species that might be threatened. We need to understand and communicate to the communities the importance of this apex predator. Predator-prey relationship is not just killing of one animal to feed another but is regulating the ecosystem. At altitudes where vegetation is scarce and cannot sustain large herbivore populations, predators are helping prey by lowering their intraspecific competition for limited resources and adding nutrients, in form of feces and decaying remains of kill, back to the soil for plant growth. In addition, often weak animals are killed by predators, and so predation is facilitating the pooling of stronger genes in prey populations. Despite a lot of effort, diet of this carnivore is not well assessed, largely as a result of the problems associated with classical methods involving histological examination of fecal content and failure of amplification due to low quality and quantity of DNA in barcoding approach. The high number of reads for each prey obtained through high-throughput sequencing makes it is easier to accurately identify closely related domestic and wild prey species (e.g. wild and domestic yak), which is particularly important for proper conservation plans focused on human-cat conflict resulting from livestock depredation.

3.5.4. Human-Snow leopard Conflicts

Human-carnivore conflicts are common worldwide and are especially concerning in areas were both share the habitat, like for snow leopard, where its habitat is extensively used for livestock grazing. The general negative behavior towards carnivores in conflicts is one of the biggest hurdles in conservation of species (Woodroffe *et al.,* 2005). It's very difficult to understand and address the problem, where on one hand livestock killing is catastrophic for owner's household, which is almost equivalent to half of the regional per capita income (Mishra, 1997) and on other hand is posing threat to snow leopard population in form of retaliatory killing (Hussain, 2003; Ikeda, 2004; Mishra, 1997; Mishra and Fitzherbert, 2004; Sangay and Vernes, 2008).

For proper management we need to understand the actual extent of this problem and factors that are contributing. Through number of studies we know that snow leopard abundance is dependent on the abundance of wild ungulates (Shehzad *et al.,* 2012; Lhagvasuren and Mynkhtsog, 2000; Anwar *et al.,* 2011; Bagchi and Mishra, 2006; Chundawat and Rawat, 1994, Nepal: Oli *et al.,* 1994; Johansson *et al.,* 2015; Lovari *et al.,* 2013). Even the areas with more livestock than wild ungulates, snow leopards preferred preying upon wild ungulates (Johansson *et al.,* 2015). But occasional opportunistic livestock killings mainly resulted from chance encounters, where free-ranging livestock got left behind in pastures overnight, or grazing animals are out of herder’s sight (Johansson *et al.,* 2015). Also, in some regions due to unavailability of wild prey snow leopard attack livestock (Lovari *et al.,* 2009; Sharma *et al.,* 2015). One of the reasons of decline in wild ungulates is competition with livestock for resources (Sharma *et al.*, 2015; Suryawanshi *et al.,* 2017). Overall number of intricate factors are contributing to this problem and multipronged approach is required to address this problem for the benefit of all.

Community based efforts proved to be useful for reducing livestock depredation. For example, in Ladakh improving the livestock pens prevented the mass attacks on livestock by snow leopards (Jackson and Wangchuk, 2004). Livestock compensation programs are used in some regions like China, India and Mongolia but are proved to be short term solution to the problem (Mishra, 1997; Nyhus *et al.*, 2003; Pettigrew *et al.*, 2012). Which were later replaced by more successful long-term solution through comprehensive livestock insurance (Gurung *et al.*, 2011; Rosen *et al.*, 2012). Livestock vaccination initiative under Ecosystem Health Program (EHP) in Pakistan was indirect approach to address the problem. This initiative targets to minimize loss of livestock to disease thus economically empowering the communities through healthy but small livestock (Nawaz *et al.*, 2016). These community-based conservation efforts successfully used in one region are not suitable in other region (Mishra, 1997) indicating that we need to understand social factors influencing these programs.

In conclusion, effective conservation plan not just need better understanding of the diet of the snow leopard but also knowledge of social, cultural religious, economical and psychological factors of communities that can be influence the plan in specific region.

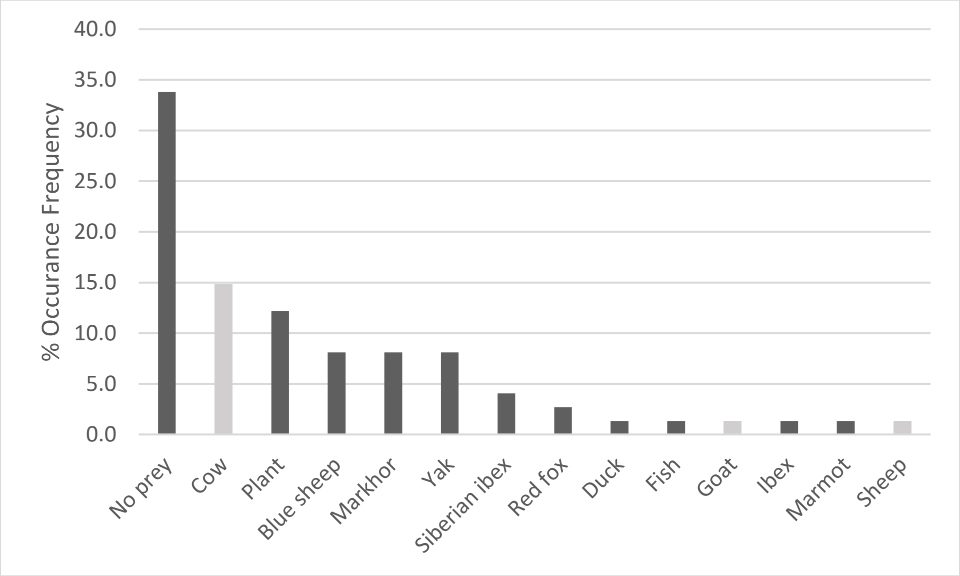

Figure 3.1: Relative frequency of prey items identified using five barcoding genes. Light grey bars indicate domestic prey species.

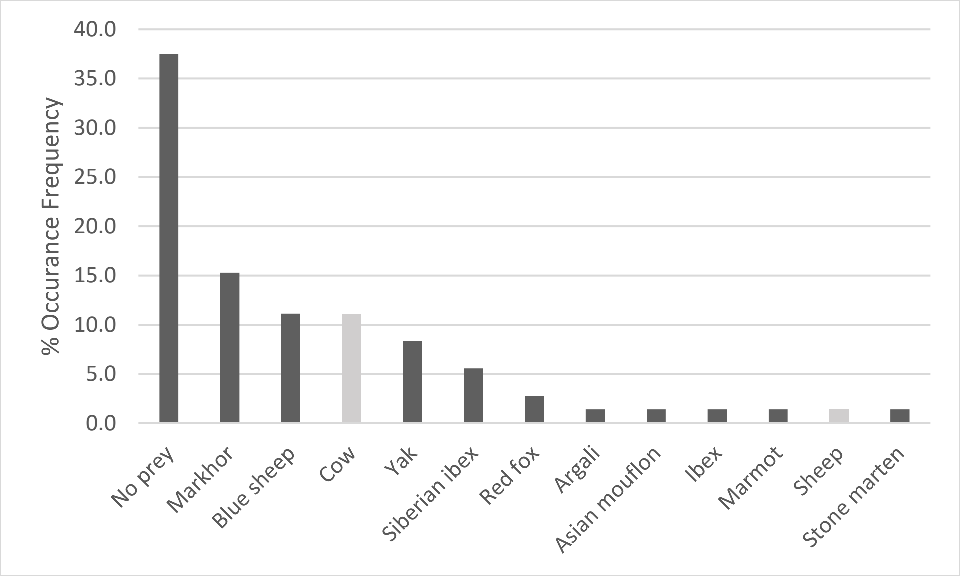

Figure 3.2: Prey items percentages identified in different samples using complete mtDNA of potential prey species of snow leopard. Light grey bars indicate the domestic prey species.

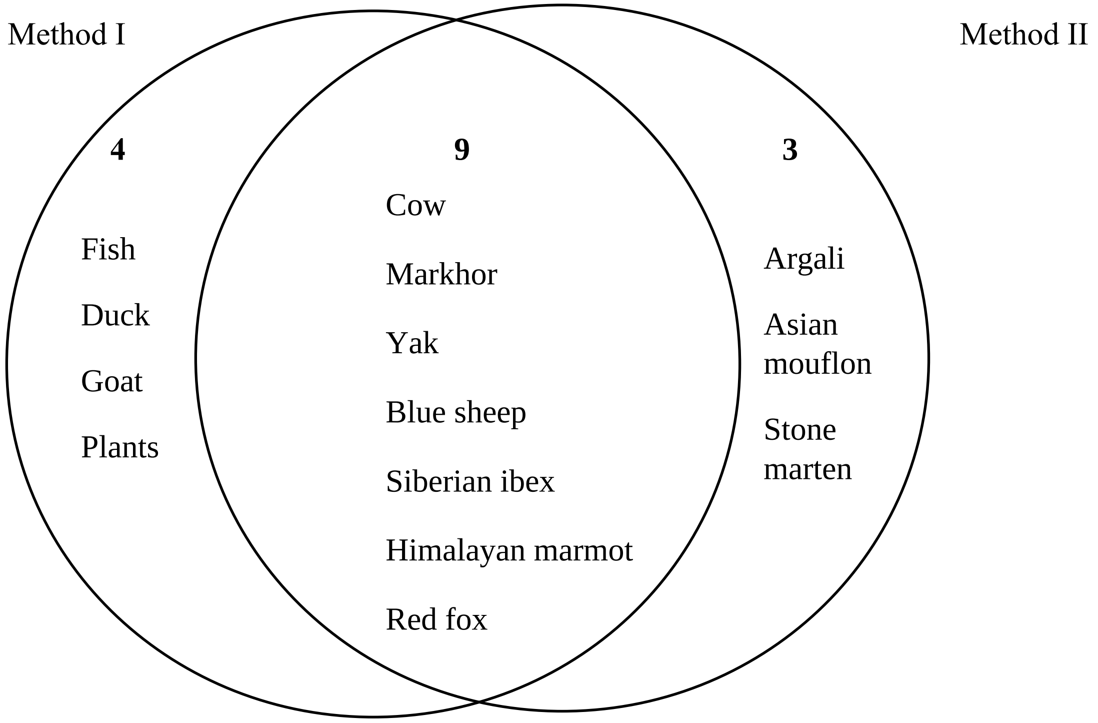

Figure 3.3: Venn diagram showing the number of common and unique prey items in all samples identified using both methods I and II. Method I had four unique prey identified, fish, duck, goat and plants. Method II had three including, argali, Asian mouflon and stone marten.

Table 3.1: Number of reads for each species obtained from using Method I and Method II.

	Method I						Method II
	COI (reads #)	COIII (reads #)	Cytb (reads #)	ND2 (reads #)	ND4 (reads #)	Total	Reads #
Snow leopard	4556	1265	12429	7116	5900	31266	42220
Prey	7099	93	506	337	120	8155	261997
Cow	3388	0	715	0	0	4103	1011
Markhor	66	37	44	1	41	189	578
Yak	63	15	10	23	9	120	684
Blue sheep	103	4	22	0	8	137	1170
Siberian ibex	1	0	11	0	0	12	1677
Himalayan marmot	10	0	1	0	0	11	59
Red fox	49	3	0	2	0	54	120
Sheep	52	0	15	0	0	67	135
Alpine Ibex	0	0	9	0	0	9	89
Fish	999	0	0	0	0	999	-
Common pochard	0	0	0	4	4	8	-
Goat	3	0	8	5	1	17	-
Argali	-	-	-	-	-	-	56
Asian mouflon	-	-	-	-	-	-	67
Stone marten	-	-	-	-	-	-	215

Table 3.2: Prey items identified from samples of different countries/regions using Method I and Method II.

Location	Prey	Frequency		% Freq. of Occurrence	
		Method I	Method II	Method I	Method II
China (n=10)	Blue sheep	6	8	50.0	61.5
	Cow	2	2	16.7	15.4
	Goat	1	0	8.3	0
	Stone marten	0	1	0	7.7
	No prey	3	2	25.0	15.4
	Total	12	13		
Kyrgyzstan (n=9)	Siberian ibex	1	1	11.1	11.1
	Duck	1	N/A	11.1	N/A
	Plants	3	N/A	33.3	N/A
	No prey	4	8	44.4	88.9
	Total	9	9		
Mongolia (n=8)	Siberian ibex	2	3	22.2	33.3
	Cow	1	1	11.1	11.1
	No prey	6	5	66.7	55.6
	Total	9	9		
Pakistan (n=24)	Cow	8	5	25.0	17.9
	Markhor	6	11	18.8	39.3
	Ibex	1	1	3.1	3.6
	Sheep	1	1	3.1	3.6
	Red fox	1	1	3.1	3.6
	Yak	4	4	12.5	14.3
	Fish	1	N/A	3.1	0.0
	Plant	5	N/A	15.6	0.0
	No prey	5	5	15.6	17.9
	Total	32	28		
Tajikistan (n=10)	Yak	2	2	20	18.2
	Marmot	1	1	10	9.1
	Plants	1	N/A	10	0.0
	Argali	0	1	0	9.1
	Asian mouflon	0	1	0	9.1
	No prey	6	6	60	54.5
	Total	10	11		
Nepal (n=1)	Red fox	1	1	N/A	N/A
Afghanistan (n=1)	No prey	1	1	N/A	N/A

Table 3.3: A comparison of frequency of occurrence (%) of prey items in scat of snow leopard from different regions across its range.

Prey	China		Kyrgyzstan			Mongolia				Pakistan				Tajikistan	
	Present Study (n=10)		Present Study (n=9)		Jumabay-Uulu *et al.*, 2014 (n= 39)	Present Study (n=8)		Shehzad *et al.*, 2012 (n=81)	Lhagvasuren and Munkhtsog, 2000 (n=168)	Present Study (n=24)		Anwar *et al.*, 2011 (n=49)	Khatoon 2017 (n-56)	Present Study (n=10)	
	M1	M2	M1	M2		M1	M2			M1	M2			M1	M2
Ibex			11.1	11.1	8	22.2	33.3			3.1	3.6				
Argali					12			8.6							9.1
Birds			11.1					1.2	2.4			2.2			
Blue sheep	50	61.5													
Civet													16.8		
Cow	16.7	15.4				11.1	11.1			25	17.9				
Donkey															
Fish										3.1					
Goat	8.3							17.3	3.6			11.8	8.8		
Goitered gazelle									3.6						
Hare									1.2				4		
Horse									5.4						
Insects									2.4						
Ladakh urial															
Markhor										18.8	39.3	3.2	4.8		
Marmots									1.2				8.8	10	9.1
Marten		7.7													
Mouflon															9.1
Pika									5.9				3.2		
Plant matter			33.3						14.9	15.6		31.2		10	
Red deer									2.4						
Roe deer									0.6						
Red fox													2.4		
Rodents									0.6						
Vole													1.6		
Sheep		6.3						2.5	17.3	3.1		16.1	12		
Ibex								70.4	38.7			9.7	2.4		
Tahr															

Weasel															
Yak										12.5	14.3		5.6	20	18.2
Unidentified	25	15.4	44.4	88.9		66.7	55.6			15.6	17.9			60	54.5

3.6. LITERATURE CITED

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Anwar, M.B., Jackson, R., Nadeem, M.S., Janečka, J.E., Hussain, S., Beg, M.A., Muhammad, G. and Qayyum, M. (2011). Food habits of the snow leopard Panthera uncia (Schreber, 1775) in Baltistan, Northern Pakistan. European Journal of Wildlife Research, 57(5), 1077-1083.

Aryal, A., Brunton, D., Ji, W., Barraclough, R. K., and Raubenheimer, D. (2014). Human–carnivore conflict: ecological and economical sustainability of predation on livestock by snow leopard and other carnivores in the Himalaya. Sustainability Science, 9(3), 321-329.

Bagchi, S., and Mishra, C. (2006). Living with large carnivores: predation on livestock by the snow leopard (Uncia uncia). Journal of zoology, 268(3), 217-224.

Brown, D. S., Jarman, S. N., and Symondson, W. O. (2012). Pyrosequencing of prey DNA in reptile faeces: analysis of earthworm consumption by slow worms. Molecular Ecology Resources, 12(2), 259-266.

Cereb, N., Kim, H. R., Ryu, J., and Yang, S. Y. (2015). Advances in DNA sequencing technologies for high resolution HLA typing. Human immunology, 76(12), 923-927.

Chaves, P. B., Graeff, V. G., Lion, M. B., Oliveira, L. R., and Eizirik, E. (2012). DNA barcoding meets molecular scatology: short mtDNA sequences for standardized species assignment of carnivore noninvasive samples. Molecular Ecology Resources, 12(1), 18-35.

Chundawat, R.S., and Rawat, G. S. (1994) Food habits of the snow leopard in Ladakh.In: Fox JL, Jizeng D, eds. International Snow Leopard Trust and Northwest Plateau Institute of Biology. Seattle: International Snow Leopard Trust. pp 127–132.

Clare, E. L., Fraser, E. E., Braid, H. E., Fenton, M. B., and Hebert, P. D. (2009). Species on the menu of a generalist predator, the eastern red bat (Lasiurus borealis): using a molecular approach to detect arthropod prey. Molecular ecology, 18(11), 2532-2542.

Dalerum F, Somers MJ, Kunkel KE, Cameron EZ (2008) The potential for large carnivores to act as biodiversity surrogates in southern Africa. Biodiversity and Conservation 17: 2939–2949.

Eitzinger, B., Rall, B. C., Traugott, M., and Scheu, S. (2018). Testing the validity of functional response models using molecular gut content analysis for prey choice in soil predators. Oikos, 127(7), 915-926.

Ekblom, R., and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107(1), 1-15.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., Hallwachs, W. and Hajibabaei, M., (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proceedings of the National Academy of Sciences, 111(22), 8007-8012.

Glenn, T.C., Nilsen, R.A., Kieran, T.J., Finger, J.W., Pierson, T.W., Bentley, K.E., Hoffberg, S.L., Louha, S., García-De León, F.J., del Rio Portilla, M.A. and Reed, K.D. (2016). Adapterama I: universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru and iNext). BioRxiv, 049114.

Gurung, G. S., Thapa, K., Kunkel, K., Thapa, G. J., Kollmair, M., and Müller-Böker, U. (2011). Enhancing herders' livelihood and conserving the snow leopard in Nepal. Cat News, 55, 17-21.

Hajibabaei, M., Smith, M. A., Janzen, D. H., Rodriguez, J. J., Whitfield, J. B., and Hebert, P. D. (2006). A minimalist barcode can identify a specimen whose DNA is degraded. Molecular Ecology Notes, 6(4), 959-964.

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

Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., Angedakin, S., Casals, F., Navarro, A., Vigilant, L. and Kühl, H.S. (2018). The impact of endogenous content, replicates, and pooling on genome capture from faecal samples. Molecular ecology resources, 18(2), 319-333.

Huffman, M. A. (2003). Animal self-medication and ethno-medicine: exploration and exploitation of the medicinal properties of plants. Proceedings of the Nutrition Society, 62(2), 371-381.

Hussain, S. (2003) The status of the snow leopard in Pakistan and its conflict with local farmers. Oryx 37: 26–33.

Ikeda, N. (2004). Economic impacts of livestock depredation by snow leopard Uncia uncia in the Kanchenjunga Conservation Area, Nepal Himalaya. Environmental Conservation, 31(4), 322-330.

Jackson, R. M., and Wangchuk, R. (2004). A community-based approach to mitigating livestock depredation by snow leopards. Human dimensions of wildlife, 9(4), 1-16.

Jackson, R. M., Mishra, C., McCarthy, T., Ale, S. B., Macdonald, D. W., and Loveridge, A. J. (2010). Snow leopards, conservation, and conflict. MacDonald, D. and Loveridge, A.(Eds.), 417-430.

Janecka, J.E., Hacker, C., Broderick, J., Pulugulla, S., Auron, P., Ringling, M., Nelson, B., Munkhtsog, B., Hussain, S., Davis, B. and Jackson, R. (2020). Noninvasive Genetics and Genomics Shed Light on the Status, Phylogeography, and Evolution of the Elusive Snow Leopard. In Conservation Genetics in Mammals (pp. 83-120). Springer, Cham.

Janjua, S., Peters, J. L., Weckworth, B., Abbas, F. I., Bahn, V., Johansson, O., and Rooney, T. P. (2020). Improving our conservation genetic toolkit: ddRAD-seq for SNPs in snow leopards. Conservation Genetics Resources, 12, 257-261.

Johansson, Ö., McCarthy, T., Samelius, G., Andrén, H., Tumursukh, L., and Mishra, C. (2015). Snow leopard predation in a livestock dominated landscape in Mongolia. Biological Conservation, 184, 251-258.

Jumabay-Uulu, K., Wegge, P., Mishra, C., and Sharma, K. (2014). Large carnivores and low diversity of optimal prey: A comparison of the diets of snow leopards Panthera uncia and wolves Canis lupus in Sarychat-Ertash Reserve in Kyrgyzstan. Oryx, 48(4), 529-535. doi:10.1017/S0030605313000306

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

Khatoon, R., Hussain, I., Anwar, M., and Nawaz, M. A. (2017). Diet selection of snow leopard (Panthera uncia) in Chitral, Pakistan. Turkish Journal of Zoology, 41(5), 914-923.

Lhagvasuren, B., and Munkhtsog, B. (2002). The yak population in Mongolia and its relation with snow leopards as a prey species. Yak production in central Asian highlands, 69.

Li, H., and Durbin, R. (2009). Making the Leap: Maq to BWA. Mass Genomics, 25, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.s (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Lischer, H. E., and Excoffier, L. (2012). PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics, 28(2), 298-299.

Lovari, S., Minder, I., Ferretti, F., Mucci, N., Randi, E., and Pellizzi, B. (2013). Common and snow leopards share prey, but not habitats: competition avoidance by large predators?. Journal of Zoology, 291(2), 127-135.

Lyngdoh, S., Shrotriya, S., Goyal, S. P., Clements, H., Hayward, M. W., and Habib, B. (2014). Prey preferences of the snow leopard (Panthera uncia): regional diet specificity holds global significance for conservation. PloS one, 9(2).

McCarthy, T. M., and Chapron, G. (2003). Snow leopard survival strategy. International Snow Leopard Trust and Snow Leopard Network, Seattle, USA, 105.

Meusnier, I., Singer, G. A., Landry, J. F., Hickey, D. A., Hebert, P. D., and Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. BMC genomics, 9(1), 1-4.

Mishra, C. (1997) Livestock depredation by large carnivores in the Indian trans- Himalaya: conflict perceptions and conservation prospects. Environ Conserv 24: 338–343.

Mishra, C., and Fitzherbert, A. (2004). War and wildlife: a post-conflict assessment of Afghanistan's Wakhan Corridor. Oryx, 38(1), 102-105.

Nawaz, M. A., Din, J.U., and Buzdar., H. (2016). The Ecosystem Health Program: A Tool to Promote the Coexistence of Livestock Owners and Snow Leopards. In McCarthy, T., Mallon, D., and Nyhus, P. J. (Eds.). Snow leopards. Academic Press. Pp 188-195.

Nyhus, P. J., Fisher, H., Osofsky, S., and Madden, F. (2003). Taking the bite out of wildlife damage: the challenges of wildlife compensation schemes. Conservation Magazine (formerly Conservation in Practice), 37.

Nyhus, P. J., Mccarthy, T., and Mallon, D. (2016). Snow leopards: biodiversity of the world: conservation from genes to landscapes. Academic Press.

Oli, M. K., Taylor, I. R., and Rogers, M. E. (1994). Snow leopard Panthera uncia predation of livestock: an assessment of local perceptions in the Annapurna Conservation Area, Nepal. Biological Conservation, 68(1), 63-68.

Perry, G. H., Marioni, J. C., Melsted, P., and Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. Molecular Ecology, 19(24), 5332-5344.

Pettigrew, M., Xie, Y., Kang, A., Rao, M., Goodrich, J., Liu, T., and Berger, J. (2012). Human–carnivore conflict in China: a review of current approaches with recommendations for improved management. Integrative Zoology, 7(2), 210-226.

Pompanon, F., Deagle, B. E., Symondson, W. O., Brown, D. S., Jarman, S. N., and Taberlet, P. (2012). Who is eating what: diet assessment using next generation sequencing. Molecular ecology, 21(8), 1931-1950.

Porter, T. M., and Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. PloS one, 13(9).

Ripple, W. J., and Beschta, R. L. (2012). Trophic cascades in Yellowstone: the first 15 years after wolf reintroduction. Biological Conservation, 145(1), 205-213.

Ripple, W. J., Larsen, E. J., Renkin, R. A., and Smith, D. W. (2001). Trophic cascades among wolves, elk and aspen on Yellowstone National Park’s northern range. Biological conservation, 102(3), 227-234.

Rosen, T., Hussain, S., Mohammad, G., Jackson, R., Janecka, J. E., and Michel, S. (2012). Reconciling sustainable development of mountain communities with large carnivore conservation. Mountain Research and Development, 32(3), 286-293.

Sangay, T., and Vernes, K. (2008). Human–wildlife conflict in the Kingdom of Bhutan: patterns of livestock predation by large mammalian carnivores. Biological Conservation, 141(5), 1272-1282.

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J.F., Al-Rasheid, K.A., Willerslev, E., Krogh, A. and Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. BMC genomics, 13(1), 178.

Sharma, R. K., Bhatnagar, Y. V., and Mishra, C. (2015). Does livestock benefit or harm snow leopards?. Biological Conservation, 190, 8-13.

Shehzad W, Mccarthy TM, Pompanon F, Purevjav L, Riaz T (2012) Prey Preference of Snow Leopard (*Panthera uncia*) in South Gobi, Mongolia. PloS ONE 7: 1–8.

Srivathsan, A., Sha, J. C., Vogler, A. P., and Meier, R. (2015). Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (P ygathrix nemaeus). Molecular Ecology Resources, 15(2), 250-261.

Suryawanshi, K. R., Redpath, S. M., Bhatnagar, Y. V., Ramakrishnan, U., Chaturvedi, V., Smout, S. C., and Mishra, C. (2017). Impact of wild prey availability on livestock predation by snow leopards. Royal Society open science, 4(6), 170026.

Symondson, W. O. C. (2002). Molecular identification of prey in predator diets. Molecular ecology, 11(4), 627-641.

Vesterinen, E.J., Ruokolainen, L., Wahlberg, N., Peña, C., Roslin, T., Laine, V.N., Vasko, V., Sääksjärvi, I.E., Norrdahl, K. and Lilley, T.M. (2016). What you need is what you eat? Prey selection by the bat Myotis daubentonii. Molecular ecology, 25(7), 1581-1594.

Woodroffe, R., Thirgood, S., and Rabinowitz, A. (Eds.). (2005). People and wildlife, conflict or co-existence? (No. 9). Cambridge University Press.

Zinger, L., Bonin, A., Alsos, I.G., Bálint, M., Bik, H., Boyer, F., Chariton, A.A., Creer, S., Coissac, E., Deagle, B.E. and De Barba, M. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. Molecular ecology, 28(8), 1857-1862.

Chapter 4

SNP IDENTIFICATION IN MITOCHONDRIAL GENES OF SNOW LEOPARDS FROM SAMPLES COLLECTED FROM DIFFERENT REGIONS ACROSS ITS RANGE

4.1. ABSTRACT

Mitochondrial DNA (mtDNA) based studies have provided important insights into taxonomy, population genetics, population structure, and behavior of wild populations of number of species. mtDNA is a valuable marker because it does not undergo recombination, and it has a high mutation rate and fast sorting rate. Therefore, patterns of spatial variation accumulate rapidly, enabling scientists to answer important questions related to evolution, population genetics and structure. Here we used by-product data of a DNA capture run (detailed in chapter 2) to identify SNPs in mitochondrial genes: Cytb, COI and COII. Collectively 3369 bp were used for this purpose and we were able to identify 22 parsimony informative sites that can be useful for future mitochondrial gene-based population genetics and structure studies of snow leopards.

4.2. INTRODUCTION

Snow leopard (*Panthera uncia*) is one of the least studied big cats, mainly due to its elusive behavior. Collecting tissue or blood samples can be challenging due to the difficulty of finding individuals, the risk associated with capturing animals, and the regulations imposed by the Convention on International Trade in Endangered Species (CITES). The most feasible option for genetic studies of such animals is non-invasive samples, including fecal scats. DNA extracted from scats has helped us understand aspects of wild populations, including population genetics and population structure (Mondol *et al.*, 2009; Frosch *et al.,* 2014; Janecka *et al.,* 2 017), behavior (Schwarzenberger, 2007), and dispersal (Tiedemann *et al.,* 2000; de Oliveira *et al.,* 2012). However, low quality and quantity of DNA extracted from non-invasive samples is one of the major impediments for large-scale genetic studies of wild populations (Taberlet *et al*. 1999).

To date, genetic studies mostly have been limited to using microsatellites designed originally for domestic cats (Waits *et al.,* 2007) and later modified for specificity to snow leopard (Janecka *et al.,* 2008). Most of these initial studies (Janecka *et al.,* 2008, Karmacharya *et al.,* 2011; Aryal *et al.,* 2014) were optimization studies and were not informative in terms of snow leopard population genetics and structure. However, Janecka *et al*. (2017) reported three separate genetic clusters of snow leopards based on 33 microsatellites. Based on those data, Janecka *et al*. (2017) proposed the recognition of three separate subspecies in the northern, central, and western parts of the snow leopard's range. However, 683 base pairs (bp) of mitochondrial DNA (mtDNA) lacked informative variation (Janecka *et al.,* 2017), and did not support sub-species status (see Moritz, 1994). It is unclear whether the size of this mtDNA fragment was insufficient for capturing informative variation or if actual mitonuclear discordance (Toews and Brelsford 2012) is present in the species. Additional genetic work is needed to test these hypotheses.

Next generation sequencing technologies enable us to rapidly sequence DNA and RNA samples and thus support a wide range of applications in medicine (Kotelnikova *et al.,* 2016), forensics (Børsting and Morling, 2015), genetics (Strafella *et al.,* 2020), etc. It also holds great promise in the field of wildlife conservation genetics, especially for non-model organisms and wild populations of endangered species, such as snow leopard. Non-invasive samples are often the only option available for studying large mammals like snow leopard, but such samples have highly degraded DNA, which can result in incomplete/reduced nuclear representation (Blåhed *et al.*, 2019). In addition, fecal DNA has chemicals that inhibit PCR, which is necessary for most genetic methods (Kohn and Wayne 1997; Nechvatal *et al*. 2008). Another problem associated with fecal DNA is an overwhelming quantity of exogenous DNA, including gut microbiota, environmental contaminants, and prey species (Hernandez-Rodriguez *et al.*, 2018). This exogenous DNA significantly reduces the chances of obtaining good coverage of nuclear DNA from high-throughput methods.

In Chapter 3, I reported that myBAITS RNA probes, designed from a ddRAD-seq library of snow leopard (Janjua *et al.*, 2020), captured ~95-99% non-targeted sequences. This is because scat DNA has just 1-5% of endogenous DNA (Hernandez-Rodriguez *et al.*, 2018; Perry *et al.*, 2010). We found that the non-targeted sequences included a large amount of mtDNA, which has a much higher copy number than nuclear DNA, from snow leopards. I extracted cytochrome b (Cytb), cytochrome oxidase subunit I (COI), and cytochrome oxidase subunit II (COII) gene sequences from the snow leopard DNA capture results and identified SNPs that can be useful for studying population genetics of snow leopards.

4.3. MATERIALS AND METHODS

4.3.1. Sampling

We began with 96 scat samples of snow leopards collected from different regions (Detailed in section 2.3.1). Briefly, we had 53 samples from Pakistan (PK), 11 samples from Tajikistan (TJ), 11 from Mongolia (MO), 10 from China (CT and LXC), 9 from

Kyrgyzstan (KY), and 1 from Nepal (NE) and Afghanistan (AF) each. We searched genomic data obtained using MyBAITS RNA probes (described in Chapter 2) for mtDNA sequences that were captured as a by-product for each of these samples.

4.3.2. Snow Leopard Mitochondrial DNA Extraction

Geneious V. 5.6.5 was used to align sequences from the MyBAITS capture with a snow leopard reference mitochondrial genome (Accession Number: NC_010638.1). Sequences that were greater than 150 bp, were represented by at least five copies, and had e-values less than or equal to 10^{-7} were retained. To account for PCR and sequencing errors, we required that the same nucleotide was present in ≥ 75% of sequences at each site. If fewer than 75% of nucleotides matched, we assigned ambiguity codes to those sites. Three genes, including cytochrome b (Cytb), cytochrome oxidase subunit I (COI), and cytochrome oxidase subunit II (COII), were examined. MEGA (Molecular Evolutionary Genetics Analysis) X (Kumar *et al.*, 2018) was used to align mtDNA sequences from different individuals.

4.3.3. Data Analysis

Pairwise genetic distance (*d*xy), the number of segregating sites (*S*), the number of parsimony informative sites (PI), nucleotide diversity (π) and Tajima's *D* (Tajima, 1989) were calculated for each population using MEGA-X. The fixation index, F_{ST} (a measure of the proportion of genetic diversity explained by differences among populations) was calculated for parsimony informative sites using the equation: $FST = \frac{HT-HS}{HT}$, where H_S is the average expected heterozygosity of subpopulations and H_T is the expected heterozygosity across all populations (Nei and Chesser, 1983). F_{ST} was calculated for each PI site and averaged across the gene to obtain a composite value. Referring to the criterion for genetic differentiation by Wright (1978), we defined genetic differentiation as low for F_{ST} <0.05, moderate for 0.05< F_{ST} <0.15, high for 0.15< F_{ST} <0.25, and very high for F_{ST} >0.25.

4.4. RESULTS

4.4.1. Snow Leopard mtDNA

Using our filtration criteria, we obtained 755 bp to 16,133 bp of mtDNA from 58 samples. We focused our analysis on three genes that are widely used for studying population genetics, including 1140 bp of cytochrome b (Cytb; N = 44 samples), 1545 bp of cytochrome oxidase subunit I (COI; N = 39 samples), and 684 bp of cytochrome oxidase subunit II (COII; N = 28 samples). The number of base pairs obtained for each gene and for each individual are provided in Table 4.1.

4.4.2. Genetic Diversity

The nucleotide diversity for Cytb, COI, and COII is 0.0016, 0.0016, and 0.0027, respectively (Table 4.2). Parsimony informative sites for each gene were Cytb = 5, COI = 11 and COII = 6. Among all pairwise comparisons (Table 4.3), F_{ST} values were consistently high for MO versus CT (range = 0.519 to 0.803) and MO versus KY (range = 0.728 to 0.989). On the other hand KY and TJ had the lowest F_{ST} values, ranging from <0 to 0.182. We also found moderate to high F_{ST} (range = 0.124 to 0.728) values among the remaining sites. All three genes had negative Tajima's D values (Table 4.2).

4.5. DISCUSSION

From snow leopard mtDNA sequences obtained as non-targeted sequences from DNA capture, we extracted three mitochondrial genes, *viz* Cytb (1140 bp), COI (1545 bp) and COII (648 bp). Based on 44 samples from Cytb, 39 for COI, and 28 for COII, and assuming each sample is a separate individual, pairwise genetic distance and F_{ST} was calculated that ranges from low to very high according to criteria put forth by Wright (1978). However, due to limited sample size for each collection site and lack of individual identification, these numbers are not informative enough to draw conclusions about population spatial structure. Regardless, the SNPs identified will be useful for future mitochondrial gene-based population genetics and structure studies of snow leopards.

Consistent with Janecka *et al*. (2008, 2017), we found low nucleotide diversity in mtDNA ($\pi = 0.0016$, 0.0016, and 0.0027, for Cytb, COI, and COII, respectively). However, in contrast to the previous study, we found 32 segregating sites and five parsimony informative sites in the Cytb gene. This difference could be because Janecka *et al*. (2008, 2017) sequenced a 96-100 bp fragment of Cytb, whereas we used the complete 1140 bp gene sequence.

An accurate understanding of population genetics and structure is very important for species conservation. Genetic variation is very crucial for species survival and adaptation to environmental changes (Forrest *et al.*, 2012; Pauls *et al.*, 2013; Aryal *et al.*, 2016). Population structure suggests isolation of populations and/or interrupted gene flow, that increases the chances of inbreeding and ultimately loss of genetic diversity, but individual identification is necessary before these values can be interpreted conclusively. Low genetic diversity and low gene flow across its range suggests the snow leopard populations are vulnerable. Concerted conservation efforts are needed to work across country boundaries to save this charismatic species.

Table 4.1: Number of nucleotides retrieved for each reference gene for each sample.

Sample	Retrieved gene sequence size (bp)		
	Cytb: 1140	COI: 1545	COII: 684
AF	--	157	--
NE	1140	1545	684
CT-2	753	305	268
CT-3	702	1335	499
CT-4	753	943	241
CT-7	753	947	365
CT-10	1120	1442	610
CT-12	1140	1505	684
CT-13	188	543	311
CT-17	1140	1442	632
LXC	--	151	--
KY-007	972	518	345
KY-015	1140	945	204
KY-018	1140	1519	581
KY-029	1140	1545	684
KY-035	1140	923	494
KY-045	1140	1544	684
KY-058	1140	1498	684
MO-156	1140	1545	684
MO-038	1081	191	151
MO-071	151	--	--
MO-106	416	--	--
MO-034	151	--	--
TJ-043	1059	333	151
TJ-078	1140	1539	654
TJ-082	1053	1230	546
TJ-123	328	302	--
TJ-127	1111	845	527
TJ-130	302	302	--
TJ-176	1140	555	249
TJ-269	1140	1539	684
PK-14	587	158	--
PK-35	--	364	--
PK-42	650	659	--
PK-54	554	613	--
PK-55	--	302	--
PK-64	1140	1545	684
PK-66	523	1047	--
PK-67	422	--	--
PK-68	1140	1545	684

PK-69	1140	1519	677
PK-72	1140	1545	684
PK-86	504	--	--
PK-84	480	--	--
PK-99	338	--	--
PK-103	354	302	--
PK-104	302	--	--
PK-108	302	--	--

Table 4.2: Parameters of genetic diversity and neutral test based on three mitochondrial gene sequences

Genes	*N*	*S*	PI	*Π*	*D*
Cytb	44	32	5	0.0016	-2.55
COI	39	58	11	0.0016	-2.94
COII	28	20	6	0.0027	-2.28
N = number of sequences, *S* = Number of segregating sites, PI = parsimony informative sites, π = nucleotide diversity, and *D* is the Tajima's *D* test statistic.					

Table 4.3: Pairwise genetic distance (below diagonal) and F_{ST} (above diagonal) based on three mitochondrial gene sequences data of five sampling locations. Population codes are as given in Table 4.1.

Cytochrome b (Cytb)					
	TJ	PK	MO	KY	CT
TJ		0.121	0.346	0.183	0.209
PK	0 0050		0.421	0.325	0.258
MO	0.0014	0.0055		n/c*	0.803
KY	0.0077	0.0039	0.00019		0.189
CT	0.0022	0.0054	0.0024	0.0012	
Cytochrome Oxidase Subunit I (COI)					
	TJ	PK	MO	KY	CT
TJ		0.284	0.694	-0.022	0.330
PK	0.0022		0.437	0.309	0.226
MO	0.0024	0.0062		0.728	0.519
KY	0.0021	0.0037	0.0034		0.206
CT	0.0050	0.0057	0.0084	0.00494	
Cytochrome Oxidase Subunit II (COII)					
	TJ	PK	MO	KY	CT
TJ		0.296	0.661	0.012	0.124
PK	0.0050		0.384	0.427	0.213
MO	0.0072	0.0013		0.989	0.593
KY	0.0032321q`24	0.0039	0.0048		0.178
CT	0.0067	0.0068	0.0046	0.0052	
*n/c is value not calculated					

4.6. LITERATURE CITED

Aryal, A., Brunton, D., Ji, W., Karmacharya, D., McCarthy, T., Bencini, R., and Raubenheimer, D. (2014). Multipronged strategy including genetic analysis for assessing conservation options for the snow leopard in the central Himalaya. Journal of Mammalogy, 95(4), 871-881.

Aryal, A., Shrestha, U.B., Ji, W., Ale, S.B., Shrestha, S., Ingty, T., Maraseni, T., Cockfield, G. and Raubenheimer, D. (2016). Predicting the distributions of predator (snow leopard) and prey (blue sheep) under climate change in the Himalaya. Ecology and Evolution, 6(12), 4065-4075.

Blåhed, I. M., Ericsson, G., and Spong, G. (2019). Noninvasive population assessment of moose (Alces alces) by SNP genotyping of fecal pellets. European Journal of Wildlife Research, 65(6), 96.

Børsting, C., and Morling, N. (2015). Next generation sequencing and its applications in forensic genetics. Forensic Science International: Genetics, 18, 78-89.

de Oliveira, L. R., Loizaga De Castro, R., Cárdenas-Alayza, S., and Bonatto, S. L. (2012). Conservation genetics of South American aquatic mammals: an overview of gene diversity, population structure, phylogeography, non-invasive methods and forensics. Mammal Review, 42(4), 275-303.

Forrest, J.L., Wikramanayake, E., Shrestha, R., Areendran, G., Gyeltshen, K., Maheshwari, A., Mazumdar, S., Naidoo, R., Thapa, G.J. and Thapa, K. (2012). Conservation and climate change: Assessing the vulnerability of snow leopard habitat to treeline shift in the Himalaya. Biological Conservation, 150(1), 129-135.

Frosch, C., Dutsov, A., Zlatanova, D., Valchev, K., Reiners, T. E., Steyer, K., ... and Nowak, C. (2014). Noninvasive genetic assessment of brown bear population structure in Bulgarian mountain regions. Mammalian Biology, 79(4), 268-276.

Hernandez-Rodriguez, J., Arandjelovic, M., Lester, J., de Filippo, C., Weihmann, A., Meyer, M., Angedakin, S., Casals, F., Navarro, A., Vigilant, L. and Kühl, H.S. (2018). The impact of endogenous content, replicates, and pooling on genome capture from faecal samples. Molecular ecology resources, 18(2), 319-333.

Janecka, J. E., Jackson, R., Yuquang, Z., Diqiang, L., Munkhtsog, B., Buckley-Beason, V., and Murphy, W. J. (2008). Population monitoring of snow leopards using noninvasive collection of scat samples: a pilot study. Animal Conservation, 11(5), 401-411.

Janečka, J. E., Jackson, R., Yuquang, Z., Diqiang, L., Munkhtsog, B., Buckley-Beason, V., and Murphy, W. J. (2008). Population monitoring of snow leopards using noninvasive collection of scat samples: a pilot study. Animal Conservation, 11(5), 401-411.

Janecka, J. E., Zhang, Y., Li, D., Munkhtsog, B., Bayaraa, M., Galsandorj, N., Wangchuk T. R., Karmacharya, D., Li, J., Lu, Z., Uulu, K. Z. Gaur, A., Kumar, S., Kumar, K., Hussain, S., Muhammad, G., Jevit, M., Hacker, C., Burger, P., Wultsch, C., Janecka, M. J. Helgen, K., Murphy, W. J., Jackson, R. (2017). Range-wide snow leopard phylogeography supports three subspecies. Journal of Heredity, 108(6), 597-607.

Janjua, S., Peters, J. L., Weckworth, B., Abbas, F. I., Bahn, V., Johansson, O., and Rooney, T. P. (2020). Improving our conservation genetic toolkit: ddRAD-seq for SNPs in snow leopards. Conservation Genetics Resources, 12(2), 257-261.

Karmacharya, D. B., Thapa, K., Shrestha, R., Dhakal, M., and Janecka, J. E. (2011). Noninvasive genetic population survey of snow leopards (Panthera uncia) in Kangchenjunga conservation area, Shey Phoksundo National Park and surrounding buffer zones of Nepal. BMC research notes, 4(1), 516.

Kohn, M. H., and Wayne, R. K. (1997). Facts from feces revisited. Trends in ecology and evolution, 12(6), 223-227.

Kotelnikova, E. A., Pyatnitskiy, M., Paleeva, A., Kremenetskaya, O., and Vinogradov, D. (2016). Practical aspects of NGS-based pathways analysis for personalized cancer science and medicine. Oncotarget, 7(32), 52493.

Kumar S., Stecher G., Li M., Knyaz C., and Tamura K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution 35:1547-1549.

Mondol, S., Karanth, K. U., Kumar, N. S., Gopalaswamy, A. M., Andheria, A., and Ramakrishnan, U. (2009). Evaluation of non-invasive genetic sampling methods for estimating tiger population size. Biological Conservation, 142(10), 2350-2360.

Moritz, C. (1994). Defining ‘evolutionarily significant units’ for conservation. Trends in ecology and evolution, 9(10), 373-375.

Nechvatal, J.M., Ram, J.L., Basson, M.D., Namprachan, P., Niec, S.R., Badsha, K.Z., Matherly, L.H., Majumdar, A.P. and Kato, I. (2008). Fecal collection, ambient preservation, and DNA extraction for PCR amplification of bacterial and human markers from human feces. Journal of microbiological methods, 72(2), 124-132.

Nei, M., and Chesser, R. K. (1983). Estimation of fixation indices and gene diversities. Annals of human genetics, 47(3), 253-259.

Pauls, S. U., Nowak, C., Bálint, M., and Pfenninger, M. (2013). The impact of global climate change on genetic diversity within populations and species. Molecular ecology, 22(4), 925-946.

Perry, G. H., Marioni, J. C., Melsted, P., and Gilad, Y. (2010). Genomic-scale capture and sequencing of endogenous DNA from feces. Molecular Ecology, 19(24), 5332-5344.

Schwarzenberger, F. (2007). The many uses of non-invasive faecal steroid monitoring in zoo and wildlife species. International Zoo Yearbook, 41(1), 52-74.

Strafella, C., Caputo, V., Campoli, G., Galota, R. M., Mela, J., Zampatti, S., ... and Cascella, R. (2020). Genetic Counseling and NGS Screening for Recessive LGMD2A Families. High-Throughput, 9(2), 13.

Taberlet, P., Waits, L. P., and Luikart, G. (1999). Noninvasive genetic sampling: look before you leap. Trends in ecology and evolution, 14(8), 323-327.

Tajima F. (1989). Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Tiedemann, R., Hardy, O., Vekemans, X., and Milinkovitch, M. C. (2000). Higher impact of female than male migration on population structure in large mammals. Molecular Ecology, 9(8), 1159-1163.

Toews, D. P., and Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. Molecular Ecology, 21(16), 3907-3930.

Waits, L. P., Buckley-Beason, V. A., Johnson, W. E., Onorato, D., and McCarthy, T. O. M. (2007). A select panel of polymorphic microsatellite loci for individual identification of snow leopards (Panthera uncia). Molecular Ecology Notes, 7(2), 311-314.

Wright, S. (1978). Evolution and the genetics of populations: a treatise in four volumes: Vol. 4: variability within and among natural populations. University of Chicago Press.

APPENDICES

Appendix 1: List of scat samples used.

#	Sample ID	Sampling Site/Country	#	Sample ID	Sampling Site/Country
1.	AF-218	Afghanistan	49.	PK-31	Pakistan
2.	CT-2	China	50.	PK-35	Pakistan
3.	CT-10	China	51.	PK-37	Pakistan
4.	CT-3	China	52.	PK-39	Pakistan
5.	CT-7	China	53.	PK-41	Pakistan
6.	CT-17	China	54.	PK-42	Pakistan
7.	CT-12	China	55.	PK-44	Pakistan
8.	LXC	China	56.	PK-48	Pakistan
9.	CT-8	China	57.	PK-51	Pakistan
10.	CT-4	China	58.	PK-52	Pakistan
11.	CT-13	China	59.	PK-53	Pakistan
12.	KY-007	Kyrgyzstan	60.	PK-54	Pakistan
13.	KY-029	Kyrgyzstan	61.	PK-55	Pakistan
14.	KY-018	Kyrgyzstan	62.	PK-59	Pakistan
15.	KY-015	Kyrgyzstan	63.	PK-60	Pakistan
16.	KY-057	Kyrgyzstan	64.	PK-63	Pakistan
17.	KY-045	Kyrgyzstan	65.	PK-64	Pakistan
18.	KY-058	Kyrgyzstan	66.	PK-65	Pakistan
19.	KY-035	Kyrgyzstan	67.	PK-66	Pakistan
20.	KY-059	Kyrgyzstan	68.	PK-67	Pakistan
21.	MO-038	Mongolia	69.	PK-68	Pakistan
22.	MO-147	Mongolia	70.	PK-69	Pakistan
23.	MO-124	Mongolia	71.	PK-71	Pakistan
24.	MO-071	Mongolia	72.	PK-72	Pakistan
25.	MO-014	Mongolia	73.	PK-75	Pakistan
26.	MO-034	Mongolia	74.	PK-76	Pakistan
27.	MO-087	Mongolia	75.	PK-81	Pakistan
28.	MO-040	Mongolia	76.	PK-83	Pakistan
29.	MO-156	Mongolia	77.	PK-84	Pakistan
30.	MO-106	Mongolia	78.	PK-86	Pakistan
31.	NE-006	Nepal	79.	PK-99	Pakistan
32.	PK-01	Pakistan	80.	PK-102	Pakistan
33.	PK-02	Pakistan	81.	PK-103	Pakistan
34.	PK-03	Pakistan	82.	PK-104	Pakistan
35.	PK-04	Pakistan	83.	PK-107	Pakistan
36.	PK-05	Pakistan	84.	PK-108	Pakistan
37.	PK-06	Pakistan	85.	PK-109	Pakistan
38.	PK-07	Pakistan	86.	PK-111	Pakistan
39.	PK-09	Pakistan	87.	TJ-078	Tajikistan
40.	PK-11	Pakistan	88.	TJ-127	Tajikistan

41.	PK-12	Pakistan	89.	TJ-130	Tajikistan
42.	PK-14	Pakistan	90.	TJ-043	Tajikistan
43.	PK-15	Pakistan	91.	TJ-123	Tajikistan
44.	PK-19	Pakistan	92.	TJ-161	Tajikistan
45.	PK-23	Pakistan	93.	TJ-176	Tajikistan
46.	PK-25	Pakistan	94.	TJ-035	Tajikistan
47.	PK-29	Pakistan	95.	TJ-276	Tajikistan
48.	PK-30	Pakistan	96.	TJ-082	Tajikistan

Appendix 2: Script used for method I

```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 24,
   "metadata": {},
   "outputs": [],
   "source": [
    "import numpy as np\n",
    "import pandas as pd\n",
    "from Bio import SeqIO\n",
    "from datetime import datetime\n",
    "from os import path\n",
    "from Bio.Blast.Applications import NcbiblastnCommandline\n"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "# Check COI sequences collected from COI_classifierv4"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 3,
   "metadata": {},
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "Total number of COI = 837845\n",
      "Maximum length of COI = 9861\n",
      "Minimum length of COI = 200\n",
      "Number of species = 147846\n"
     ]
    }
   ],
   "source": [
    "filehandle = open(\"gbCOI.fasta\", \"r\")\n",
    "coi_lens   = [len(record.seq) for record in SeqIO.parse(filehandle, \"fasta\")]\n",
    "print('Total number of COI =', len(coi_lens))\n",
    "print('Maximum length of COI =', max(coi_lens))\n",
    "print('Minimum length of COI =', min(coi_lens))\n",
    "\n",
    "filehandle = open(\"gbCOI.fasta\", \"r\")\n",
   "species    = [record.description.split()[1] for record in SeqIO.parse(filehandle, 'fasta')]\n",
   "print('Number of species =', len(set(species)))\n"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "---\n",
   "### Prepare COIref Blast local database"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 27,
  "metadata": {},
  "outputs": [
   {
    "name": "stdout",
    "output_type": "stream",
    "text": [
     "\n",
     "\n",
     "Building a new DB, current time: 10/29/2019 11:32:41\n",
     "New DB name:   /home/skho/SnowLeopard/GBcoi/gbCOI.fasta\n",
     "New DB title:  COIref\n",
     "Sequence type: Nucleotide\n",
     "Keep MBits: T\n",
     "Maximum file size: 1000000000B\n",
     "Adding sequences from FASTA; added 837845 sequences in 18.8679 seconds.\n"
    ]
   }
  ],
  "source": [
   "# only need to build once\n",
   "!makeblastdb -in gbCOI.fasta -parse_seqids -title COIref -dbtype nucl"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "### Prepare query.fasta file"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 2,
   "metadata": {},
   "outputs": [
    {
     "data": {
      "text/plain": [
       "96"
      ]
     },
     "execution_count": 2,
     "metadata": {},
     "output_type": "execute_result"
    }
   ],
   "source": [
    "samples = ['AF-218', 'CT-2', 'CT-3', 'CT-4', 'CT-7', 'CT-8', 'CT-10', 'CT-12', 'CT-13', 'CT-17']\n",
    "samples = samples + ['KY-007', 'KY-015', 'KY-018', 'KY-029', 'KY-035', 'KY-045', 'KY-057', 'KY-
058', 'KY-059', 'M0-100', 'MO-034', 'MO-038', 'MO-087', 'MO-106', 'MO-124', 'MO-156', 'NE-
006']\n",
    "samples = samples + ['PK-01', 'PK-02', 'PK-03', 'PK-04', 'PK-05', 'PK-07', 'PK-09', 'PK-11', 'PK-
12', 'PK-14', 'PK-15', 'PK-25', 'PK-30', 'PK-31', 'PK-35', 'PK-37', 'PK-39', 'PK-42', 'PK-44', 'PK-48']\n",
    "samples = samples + ['PK-51', 'PK-52', 'PK-54', 'PK-55', 'PK-59', 'PK-60', 'PK-63', 'PK-64', 'PK-
65']\n",
    "samples = samples + ['PK-66', 'PK-67', 'PK-68', 'PK-69', 'PK-70', 'PK-72', 'PK-75', 'PK-76', 'PK-
81', 'PK-84', 'PK-86', 'PK-99']\n",
    "samples = samples + ['PK-103', 'PK-104', 'PK-107', 'PK-108', 'PK-111']\n",
    "samples = samples + ['TJ-043', 'TJ-078', 'TJ-082', 'TJ-123', 'TJ-127', 'TJ-130', 'TJ-161', 'TJ-176',
'TJ-269', 'TJ-276']\n",
    "samples = samples + ['LXC', 'MO-014', 'MO-040', 'MO-071', 'MO-147', 'PK-19', 'PK-41', 'PK-53',
'PK-71', 'PK-83', 'PK-102', 'PK-109', 'TJ-035']\n",
    "len(samples)\n"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "---\n",
    "# Run blastn locally"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 4,
   "metadata": {
    "scrolled": true
   },
   "outputs": [
    {
```

```
    "name": "stdout",
    "output_type": "stream",
    "text": [
     "Archive:  ../zipfa/LXC.fa.zip\n",
     "  inflating: LXC.fa                  \n",
     "blastn -out ../blastCOI/LXC.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length mismatch gapopen qstart qend sstart send evalue bitscore\" -query LXC.fa -db COIref.fasta -evalue 0.001\n",
     "0:00:08.238890\n",
     "Archive:  ../zipfa/MO-014.fa.zip\n",
     "  inflating: MO-014.fa               \n",
     "blastn -out ../blastCOI/MO-014.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length mismatch gapopen qstart qend sstart send evalue bitscore\" -query MO-014.fa -db COIref.fasta -evalue 0.001\n",
     "0:00:09.356412\n",
     "Archive:  ../zipfa/MO-040.fa.zip\n",
     "  inflating: MO-040.fa               \n",
     "blastn -out ../blastCOI/MO-040.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length mismatch gapopen qstart qend sstart send evalue bitscore\" -query MO-040.fa -db COIref.fasta -evalue 0.001\n",
     "0:00:01.936383\n",
     "Archive:  ../zipfa/MO-071.fa.zip\n",
     "  inflating: MO-071.fa               \n",
     "blastn -out ../blastCOI/MO-071.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length mismatch gapopen qstart qend sstart send evalue bitscore\" -query MO-071.fa -db COIref.fasta -evalue 0.001\n",
     "0:00:03.590793\n",
     "Archive:  ../zipfa/MO-147.fa.zip\n",
     "  inflating: MO-147.fa               \n",
     "blastn -out ../blastCOI/MO-147.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length mismatch gapopen qstart qend sstart send evalue bitscore\" -query MO-147.fa -db COIref.fasta -evalue 0.001\n",
     "0:00:06.774942\n",
     "Archive:  ../zipfa/PK-19.fa.zip\n",
     "  inflating: PK-19.fa                \n",
     "blastn -out ../blastCOI/PK-19.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length mismatch gapopen qstart qend sstart send evalue bitscore\" -query PK-19.fa -db COIref.fasta -evalue 0.001\n",
     "0:00:01.466981\n",
     "Archive:  ../zipfa/PK-41.fa.zip\n",
     "  inflating: PK-41.fa                \n",
     "blastn -out ../blastCOI/PK-41.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length mismatch gapopen qstart qend sstart send evalue bitscore\" -query PK-41.fa -db COIref.fasta -evalue 0.001\n",
     "0:00:04.558072\n",
     "Archive:  ../zipfa/PK-53.fa.zip\n",
     "  inflating: PK-53.fa                \n",
      "blastn -out ../blastCOI/PK-53.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length
mismatch gapopen qstart qend sstart send evalue bitscore\" -query PK-53.fa -db COIref.fasta -
evalue 0.001\n",
      "0:04:34.849008\n",
      "Archive:  ../zipfa/PK-71.fa.zip\n",
      "  inflating: PK-71.fa               \n",
      "blastn -out ../blastCOI/PK-71.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length
mismatch gapopen qstart qend sstart send evalue bitscore\" -query PK-71.fa -db COIref.fasta -
evalue 0.001\n",
      "0:00:01.697801\n",
      "Archive:  ../zipfa/PK-83.fa.zip\n",
      "  inflating: PK-83.fa               \n",
      "blastn -out ../blastCOI/PK-83.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length
mismatch gapopen qstart qend sstart send evalue bitscore\" -query PK-83.fa -db COIref.fasta -
evalue 0.001\n",
      "0:00:04.623222\n",
      "Archive:  ../zipfa/PK-102.fa.zip\n",
      "  inflating: PK-102.fa              \n",
      "blastn -out ../blastCOI/PK-102.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length
mismatch gapopen qstart qend sstart send evalue bitscore\" -query PK-102.fa -db COIref.fasta -
evalue 0.001\n",
      "0:00:01.800242\n",
      "Archive:  ../zipfa/PK-109.fa.zip\n",
      "  inflating: PK-109.fa              \n",
      "blastn -out ../blastCOI/PK-109.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length
mismatch gapopen qstart qend sstart send evalue bitscore\" -query PK-109.fa -db COIref.fasta -
evalue 0.001\n",
      "0:00:03.113963\n",
      "Archive:  ../zipfa/TJ-035.fa.zip\n",
      "  inflating: TJ-035.fa              \n",
      "blastn -out ../blastCOI/TJ-035.tsv -outfmt \"6 qseqid qlen sseqid stitle pident length
mismatch gapopen qstart qend sstart send evalue bitscore\" -query TJ-035.fa -db COIref.fasta -
evalue 0.001\n",
      "0:00:07.080177\n"
     ]
    }
   ],
   "source": [
    "for sample in samples:\n",
    "    !unzip zipfa/{sample}.fa.zip\n",
    "    \n",
    "    blastx_cline = NcbiblastnCommandline(query=sample+'.fa', db='gbCOI.fasta',
evalue=0.001, outfmt=\"6 qseqid qlen sseqid stitle pident length mismatch gapopen qstart qend
sstart send evalue bitscore\", out='blastCOI/'+sample+'.tsv')\n",
    "    print(blastx_cline)\n",
    "\n",
    "    t0 = datetime.now()\n",
    "    stdout, stderr = blastx_cline()\n",
   "    t1 = datetime.now()\n",
   "    print(t1 - t0)\n",
   "\n",
   "    !rm {sample}.fa\n"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "---\n",
   "## Parse blastn results"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 9,
  "metadata": {
   "scrolled": true
  },
  "outputs": [
   {
    "name": "stdout",
    "output_type": "stream",
    "text": [
     "Sample AF-218 - Number of reads with hits = 257\n",
     "Total results = (9255, 14)\n",
     "Archive:  ../zipfa/AF-218.fa.zip\n",
     "  inflating: AF-218.fa               \n",
     "Sample CT-2 - Number of reads with hits = 26\n",
     "Total results = (12512, 14)\n",
     "Archive:  ../zipfa/CT-2.fa.zip\n",
     "  inflating: CT-2.fa                 \n",
     "Sample CT-3 - Number of reads with hits = 91\n",
     "Total results = (19352, 14)\n",
     "Archive:  ../zipfa/CT-3.fa.zip\n",
     "  inflating: CT-3.fa                 \n",
     "Sample CT-4 - Number of reads with hits = 77\n",
     "Total results = (23149, 14)\n",
     "Archive:  ../zipfa/CT-4.fa.zip\n",
     "  inflating: CT-4.fa                 \n",
     "Sample CT-7 - Number of reads with hits = 2028\n",
     "Total results = (151664, 14)\n",
     "99999\n",
     "Archive:  ../zipfa/CT-7.fa.zip\n",
     "  inflating: CT-7.fa                 \n",
     "Sample CT-8 - Number of reads with hits = 128\n",
     "Total results = (10050, 14)\n",
"\tSample CT-8 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample CT-10 - Number of reads with hits = 36\n",
"Total results = (8939, 14)\n",
"Archive:  ../zipfa/CT-10.fa.zip\n",
"  inflating: CT-10.fa              \n",
"Sample CT-12 - Number of reads with hits = 76\n",
"Total results = (21598, 14)\n",
"Archive:  ../zipfa/CT-12.fa.zip\n",
"  inflating: CT-12.fa              \n",
"Sample CT-13 - Number of reads with hits = 346\n",
"Total results = (96211, 14)\n",
"Archive:  ../zipfa/CT-13.fa.zip\n",
"  inflating: CT-13.fa              \n",
"Sample CT-17 - Number of reads with hits = 30\n",
"Total results = (8680, 14)\n",
"Archive:  ../zipfa/CT-17.fa.zip\n",
"  inflating: CT-17.fa              \n",
"Sample KY-007 - Number of reads with hits = 828\n",
"Total results = (55596, 14)\n",
"Archive:  ../zipfa/KY-007.fa.zip\n",
"  inflating: KY-007.fa             \n",
"Sample KY-015 - Number of reads with hits = 197\n",
"Total results = (78191, 14)\n",
"Archive:  ../zipfa/KY-015.fa.zip\n",
"  inflating: KY-015.fa             \n",
"Sample KY-018 - Number of reads with hits = 67\n",
"Total results = (20238, 14)\n",
"Archive:  ../zipfa/KY-018.fa.zip\n",
"  inflating: KY-018.fa             \n",
"Sample KY-029 - Number of reads with hits = 231\n",
"Total results = (70601, 14)\n",
"Archive:  ../zipfa/KY-029.fa.zip\n",
"  inflating: KY-029.fa             \n",
"Sample KY-035 - Number of reads with hits = 214\n",
"Total results = (13086, 14)\n",
"Archive:  ../zipfa/KY-035.fa.zip\n",
"  inflating: KY-035.fa             \n",
"Sample KY-045 - Number of reads with hits = 378\n",
"Total results = (33964, 14)\n",
"Archive:  ../zipfa/KY-045.fa.zip\n",
"  inflating: KY-045.fa             \n",
"Sample KY-057 - Number of reads with hits = 1058\n",
"Total results = (74943, 14)\n",
"\tSample KY-057 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample KY-058 - Number of reads with hits = 97\n",
"Total results = (27062, 14)\n",
"Archive:  ../zipfa/KY-058.fa.zip\n",
"  inflating: KY-058.fa             \n",
"Sample KY-059 - Number of reads with hits = 301\n",
"Total results = (7142, 14)\n",
"\tSample KY-059 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample M0-100 - Number of reads with hits = 583\n",
"Total results = (98047, 14)\n",
"Archive:  ../zipfa/M0-100.fa.zip\n",
"  inflating: M0-100.fa               \n",
"Sample MO-034 - Number of reads with hits = 61\n",
"Total results = (12278, 14)\n",
"\tSample MO-034 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample MO-038 - Number of reads with hits = 32\n",
"Total results = (7324, 14)\n",
"Archive:  ../zipfa/MO-038.fa.zip\n",
"  inflating: MO-038.fa               \n",
"Sample MO-106 - Number of reads with hits = 258\n",
"Total results = (5715, 14)\n",
"Archive:  ../zipfa/MO-106.fa.zip\n",
"  inflating: MO-106.fa               \n",
"Sample MO-124 - Number of reads with hits = 1\n",
"Total results = (208, 14)\n",
"\tSample MO-124 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample MO-156 - Number of reads with hits = 642\n",
"Total results = (215746, 14)\n",
"99999\n",
"199999\n",
"Archive:  ../zipfa/MO-156.fa.zip\n",
"  inflating: MO-156.fa               \n",
"Sample NE-006 - Number of reads with hits = 99\n",
"Total results = (24042, 14)\n",
"Archive:  ../zipfa/NE-006.fa.zip\n",
"  inflating: NE-006.fa               \n",
"Sample PK-01 - Number of reads with hits = 1268\n",
"Total results = (33632, 14)\n",
"Archive:  ../zipfa/PK-01.fa.zip\n",
"  inflating: PK-01.fa                \n",
"Sample PK-02 - Number of reads with hits = 1408\n",
"Total results = (159449, 14)\n",
"99999\n",
"Archive:  ../zipfa/PK-02.fa.zip\n",
"  inflating: PK-02.fa                \n",
"Sample PK-03 - Number of reads with hits = 535\n",
"Total results = (61405, 14)\n",
"Archive:  ../zipfa/PK-03.fa.zip\n",
"  inflating: PK-03.fa                \n",
"Sample PK-04 - Number of reads with hits = 3627\n",
"Total results = (338704, 14)\n",
"99999\n",
"199999\n",
"299999\n",
"Archive:  ../zipfa/PK-04.fa.zip\n",
"  inflating: PK-04.fa                \n",
"Sample PK-05 - Number of reads with hits = 1372\n",
"Total results = (95213, 14)\n",
"Archive:  ../zipfa/PK-05.fa.zip\n",
"  inflating: PK-05.fa                \n",
"Sample PK-07 - Number of reads with hits = 114\n",
"Total results = (3984, 14)\n",
"\tSample PK-07 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample PK-09 - Number of reads with hits = 297\n",
"Total results = (26402, 14)\n",
"\tSample PK-09 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample PK-11 - Number of reads with hits = 968\n",
"Total results = (29301, 14)\n",
"Archive:  ../zipfa/PK-11.fa.zip\n",
"  inflating: PK-11.fa                \n",
"Sample PK-12 - Number of reads with hits = 799\n",
"Total results = (75435, 14)\n",
"Archive:  ../zipfa/PK-12.fa.zip\n",
"  inflating: PK-12.fa                \n",
"Sample PK-14 - Number of reads with hits = 525\n",
"Total results = (42450, 14)\n",
"Archive:  ../zipfa/PK-14.fa.zip\n",
"  inflating: PK-14.fa                \n",
"Sample PK-15 - Number of reads with hits = 995\n",
"Total results = (47724, 14)\n",
"Archive:  ../zipfa/PK-15.fa.zip\n",
"  inflating: PK-15.fa                \n",
"Sample PK-25 - Number of reads with hits = 249\n",
"Total results = (18507, 14)\n",
"Archive:  ../zipfa/PK-25.fa.zip\n",
"  inflating: PK-25.fa                \n",
"Sample PK-30 - Number of reads with hits = 1015\n",
"Total results = (90318, 14)\n",
"Archive:  ../zipfa/PK-30.fa.zip\n",
"  inflating: PK-30.fa                \n",
"Sample PK-31 - Number of reads with hits = 969\n",
"Total results = (70087, 14)\n",
"Archive:  ../zipfa/PK-31.fa.zip\n",
"  inflating: PK-31.fa                \n",
"Sample PK-37 - Number of reads with hits = 284\n",
"Total results = (23132, 14)\n",
"Archive:  ../zipfa/PK-37.fa.zip\n",
"  inflating: PK-37.fa                \n",
"Sample PK-39 - Number of reads with hits = 648\n",
"Total results = (33792, 14)\n",
"Archive:  ../zipfa/PK-39.fa.zip\n",
"  inflating: PK-39.fa               \n",
"Sample PK-42 - Number of reads with hits = 303\n",
"Total results = (38017, 14)\n",
"Archive:  ../zipfa/PK-42.fa.zip\n",
"  inflating: PK-42.fa               \n",
"Sample PK-44 - Number of reads with hits = 323\n",
"Total results = (32385, 14)\n",
"Archive:  ../zipfa/PK-44.fa.zip\n",
"  inflating: PK-44.fa               \n",
"Sample PK-48 - Number of reads with hits = 635\n",
"Total results = (32004, 14)\n",
"Archive:  ../zipfa/PK-48.fa.zip\n",
"  inflating: PK-48.fa               \n",
"Sample PK-51 - Number of reads with hits = 292\n",
"Total results = (33404, 14)\n",
"Archive:  ../zipfa/PK-51.fa.zip\n",
"  inflating: PK-51.fa               \n",
"Sample PK-52 - Number of reads with hits = 3840\n",
"Total results = (168275, 14)\n",
"99999\n",
"Archive:  ../zipfa/PK-52.fa.zip\n",
"  inflating: PK-52.fa               \n",
"Sample PK-54 - Number of reads with hits = 477\n",
"Total results = (25875, 14)\n",
"Archive:  ../zipfa/PK-54.fa.zip\n",
"  inflating: PK-54.fa               \n",
"Sample PK-55 - Number of reads with hits = 351\n",
"Total results = (30890, 14)\n",
"Archive:  ../zipfa/PK-55.fa.zip\n",
"  inflating: PK-55.fa               \n",
"Sample PK-59 - Number of reads with hits = 2786\n",
"Total results = (189798, 14)\n",
"99999\n",
"Archive:  ../zipfa/PK-59.fa.zip\n",
"  inflating: PK-59.fa               \n",
"Sample PK-60 - Number of reads with hits = 993\n",
"Total results = (73166, 14)\n",
"Archive:  ../zipfa/PK-60.fa.zip\n",
"  inflating: PK-60.fa               \n",
"Sample PK-63 - Number of reads with hits = 1392\n",
"Total results = (70886, 14)\n",
"Archive:  ../zipfa/PK-63.fa.zip\n",
"  inflating: PK-63.fa               \n",
"Sample PK-64 - Number of reads with hits = 1325\n",
"Total results = (219520, 14)\n",
"99999\n",
"199999\n",
"Archive:  ../zipfa/PK-64.fa.zip\n",
"  inflating: PK-64.fa              \n",
"Sample PK-65 - Number of reads with hits = 359\n",
"Total results = (15800, 14)\n",
"\tSample PK-65 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample PK-66 - Number of reads with hits = 1185\n",
"Total results = (80437, 14)\n",
"Archive:  ../zipfa/PK-66.fa.zip\n",
"  inflating: PK-66.fa              \n",
"Sample PK-67 - Number of reads with hits = 162\n",
"Total results = (1577, 14)\n",
"\tSample PK-67 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample PK-68 - Number of reads with hits = 639\n",
"Total results = (99717, 14)\n",
"Archive:  ../zipfa/PK-68.fa.zip\n",
"  inflating: PK-68.fa              \n",
"Sample PK-69 - Number of reads with hits = 1123\n",
"Total results = (252599, 14)\n",
"99999\n",
"199999\n",
"Archive:  ../zipfa/PK-69.fa.zip\n",
"  inflating: PK-69.fa              \n",
"Sample PK-70 - Number of reads with hits = 1972\n",
"Total results = (87058, 14)\n",
"\tSample PK-70 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample PK-72 - Number of reads with hits = 193\n",
"Total results = (36378, 14)\n",
"Archive:  ../zipfa/PK-72.fa.zip\n",
"  inflating: PK-72.fa              \n",
"Sample PK-75 - Number of reads with hits = 292\n",
"Total results = (20180, 14)\n",
"Archive:  ../zipfa/PK-75.fa.zip\n",
"  inflating: PK-75.fa              \n",
"Sample PK-76 - Number of reads with hits = 626\n",
"Total results = (104133, 14)\n",
"99999\n",
"Archive:  ../zipfa/PK-76.fa.zip\n",
"  inflating: PK-76.fa              \n",
"Sample PK-81 - Number of reads with hits = 375\n",
"Total results = (64563, 14)\n",
"\tSample PK-81 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample PK-84 - Number of reads with hits = 892\n",
"Total results = (42575, 14)\n",
"\tSample PK-84 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample PK-86 - Number of reads with hits = 587\n",
"Total results = (32244, 14)\n",
"Archive:  ../zipfa/PK-86.fa.zip\n",
"  inflating: PK-86.fa              \n",
"Sample PK-99 - Number of reads with hits = 1117\n",
"Total results = (65546, 14)\n",
"Archive:  ../zipfa/PK-99.fa.zip\n",
"  inflating: PK-99.fa               \n",
"Sample PK-103 - Number of reads with hits = 221\n",
"Total results = (9383, 14)\n",
"Archive:  ../zipfa/PK-103.fa.zip\n",
"  inflating: PK-103.fa              \n",
"Sample PK-104 - Number of reads with hits = 387\n",
"Total results = (4704, 14)\n",
"\tSample PK-104 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample PK-107 - Number of reads with hits = 166\n",
"Total results = (8331, 14)\n",
"\tSample PK-107 do not have reads hit Panthera nor others with identity > 98%\n",
"Sample PK-108 - Number of reads with hits = 7569\n",
"Total results = (424959, 14)\n",
"99999\n",
"199999\n",
"299999\n",
"399999\n",
"Archive:  ../zipfa/PK-108.fa.zip\n",
"  inflating: PK-108.fa              \n",
"Sample PK-111 - Number of reads with hits = 1376\n",
"Total results = (88579, 14)\n",
"Archive:  ../zipfa/PK-111.fa.zip\n",
"  inflating: PK-111.fa              \n",
"Sample TJ-043 - Number of reads with hits = 939\n",
"Total results = (135983, 14)\n",
"99999\n",
"Archive:  ../zipfa/TJ-043.fa.zip\n",
"  inflating: TJ-043.fa              \n",
"Sample TJ-078 - Number of reads with hits = 185\n",
"Total results = (60006, 14)\n",
"Archive:  ../zipfa/TJ-078.fa.zip\n",
"  inflating: TJ-078.fa              \n",
"Sample TJ-082 - Number of reads with hits = 267\n",
"Total results = (34977, 14)\n",
"Archive:  ../zipfa/TJ-082.fa.zip\n",
"  inflating: TJ-082.fa              \n",
"Sample TJ-123 - Number of reads with hits = 38\n",
"Total results = (10752, 14)\n",
"Archive:  ../zipfa/TJ-123.fa.zip\n",
"  inflating: TJ-123.fa              \n",
"Sample TJ-127 - Number of reads with hits = 371\n",
"Total results = (130555, 14)\n",
"99999\n",
"Archive:  ../zipfa/TJ-127.fa.zip\n",
"  inflating: TJ-127.fa              \n",
"Sample TJ-130 - Number of reads with hits = 23\n",
   "Total results = (2533, 14)\n",
   "Archive:  ../zipfa/TJ-130.fa.zip\n",
   "  inflating: TJ-130.fa               \n",
   "Sample TJ-161 - Number of reads with hits = 402\n",
   "Total results = (32618, 14)\n",
   "\tSample TJ-161 do not have reads hit Panthera nor others with identity > 98%\n",
   "Sample TJ-176 - Number of reads with hits = 436\n",
   "Total results = (171626, 14)\n",
   "99999\n",
   "Archive:  ../zipfa/TJ-176.fa.zip\n",
   "  inflating: TJ-176.fa               \n",
   "Sample TJ-269 - Number of reads with hits = 279\n",
   "Total results = (99893, 14)\n",
   "Archive:  ../zipfa/TJ-269.fa.zip\n",
   "  inflating: TJ-269.fa               \n",
   "Sample TJ-276 - Number of reads with hits = 108\n",
   "Total results = (652, 14)\n",
   "\tSample TJ-276 do not have reads hit Panthera nor others with identity > 98%\n",
   "Sample LXC - Number of reads with hits = 5\n",
   "Total results = (959, 14)\n",
   "Archive:  ../zipfa/LXC.fa.zip\n",
   "  inflating: LXC.fa                  \n",
   "Sample MO-014 - Number of reads with hits = 2\n",
   "Total results = (143, 14)\n",
   "\tSample MO-014 do not have reads hit Panthera nor others with identity > 98%\n",
   "Sample MO-040 - Number of reads with hits = 1\n",
   "Total results = (164, 14)\n",
   "\tSample MO-040 do not have reads hit Panthera nor others with identity > 98%\n",
   "Sample PK-19 - Number of reads with hits = 1\n",
   "Total results = (145, 14)\n",
   "\tSample PK-19 do not have reads hit Panthera nor others with identity > 98%\n",
   "Sample PK-41 - Number of reads with hits = 7\n",
   "Total results = (7, 14)\n",
   "\tSample PK-41 do not have reads hit Panthera nor others with identity > 98%\n",
   "Sample PK-53 - Number of reads with hits = 459\n",
   "Total results = (29911, 14)\n",
   "Archive:  ../zipfa/PK-53.fa.zip\n",
   "  inflating: PK-53.fa                \n",
   "Sample PK-71 - Number of reads with hits = 5\n",
   "Total results = (762, 14)\n",
   "\tSample PK-71 do not have reads hit Panthera nor others with identity > 98%\n",
   "Sample PK-109 - Number of reads with hits = 2\n",
   "Total results = (132, 14)\n",
   "\tSample PK-109 do not have reads hit Panthera nor others with identity > 98%\n"
  ]
 }
],
"source": [
    "numreads = dict()\n",
    "categories = ['human', 'panther', 'puncia', 'bacteria', 'fungi', 'neither']\n",
    "for numkey in ['sample', 'total']+categories :\n",
    "    numreads[numkey] = []\n",
    "\n",
    "#sample CT-35 has a huge file and it takes a while to process it\n",
    "for sample in samples:\n",
    "\n",
    "    numreads['sample'].append(sample)\n",
    "    lines = !wc -l ../blastCOI/{sample}.tsv\n",
    "    if lines[0].split()[0] == '0':\n",
    "        for key in numreads.keys():\n",
    "            if key != 'sample':    numreads[key].append(0)\n",
    "        continue\n",
    "    \n",
    "    blastn = pd.read_csv('../blastCOI/'+sample+'.tsv', sep='\\t', header=None)\n",
    "    blastn.columns = 'qseqid qlen sseqid stitle pident length mismatch gapopen qstart qend
sstart send evalue bitscore'.split(' ')\n",
    "    reads = set(blastn['qseqid'])\n",
    "    numreads['total'].append(len(reads))\n",
    "    print('Sample', sample, '- Number of reads with hits =', len(reads))\n",
    "    print('Total results =', blastn.shape)\n",
    "\n",
    "    punciaidx  = []\n",
    "    pantheridx = []\n",
    "    neitheridx = []\n",
    "    hitread = dict()\n",
    "    for category in categories:\n",
    "        hitread[category] = set()\n",
    "    \n",
    "    prev = ''\n",
    "    for idx, eachrow in blastn.iterrows():\n",
    "        if (idx+1) % 100000 == 0:    print(idx)\n",
    "        if 'Homo_sapiens' in eachrow['stitle']:\n",
    "            hitread['human'].add(eachrow['qseqid'])\n",
    "            continue\n",
    "        if 'Bacteria' in eachrow['stitle']:\n",
    "            hitread['bacteria'].add(eachrow['qseqid'])\n",
    "            continue\n",
    "        if 'Fungi' in eachrow['stitle']:\n",
    "            hitread['fungi'].add(eachrow['qseqid'])\n",
    "            continue\n",
    "\n",
    "        read = eachrow['qseqid']\n",
    "        if read != prev:\n",
    "            prev = read\n",
    "            if 'Panthera_uncia' in eachrow['stitle']:\n",
    "                category = 'puncia'\n",
    "                hitread['puncia'].add(eachrow['qseqid'])\n",
    "            elif 'Panthera' in eachrow['stitle']:\n",
    "                category = 'panther'\n",
    "                hitread['panther'].add(eachrow['qseqid'])\n",
    "            else:\n",
    "                category = 'neither'\n",
    "                hitread['neither'].add(eachrow['qseqid'])\n",
    "\n",
    "        if category == 'puncia':\n",
    "            punciaidx.append(idx)\n",
    "        elif category == 'panther':\n",
    "            pantheridx.append(idx)\n",
    "        elif category == 'neither':\n",
    "            neitheridx.append(idx)\n",
    "\n",
    "    #print('Reads those hit Human =', len(sapiens))\n",
    "    #print('Reads those hit Panthera =', len(panther)+len(puncia)) \n",
    "    #print('\\tAmong them, reads that hit Panthera_uncia =', len(puncia))\n",
    "    #print('Reads those hit Bacteria =', len(bacteria))\n",
    "    #print('Reads those hit Fungi    =', len(fungi))\n",
    "    #print('Reads those hit neither  =', len(neither))\n",
    "    for category in categories:\n",
    "        numreads[category].append(len(hitread[category]))\n",
    "    \n",
    "    \n",
    "    ############# preparing to write\n",
    "    subset = blastn.loc[punciaidx+pantheridx+neitheridx, :]\n",
    "\n",
    "    towrite= []\n",
    "    for idx, eachrow in subset.iterrows():\n",
    "        qlen   = eachrow[1]\n",
    "        alen   = eachrow[5]\n",
    "        pident = eachrow[4]\n",
    "        if alen > 50 and pident > 98:\n",
    "            towrite.append(idx)\n",
    "\n",
    "    if len(towrite) == 0:\n",
    "        print('\\tSample', sample, 'do not have reads hit Panthera nor others with identity > 
98%')\n",
    "        continue\n",
    "\n",
    "    subset = subset.loc[towrite,:]\n",
    "    qseqids = list(set(list(subset['qseqid'])))\n",
    "    qseqdict = dict()\n",
    "    \n",
    "    !unzip ../zipfa/{sample}.fa.zip\n",
    "    zipfa = SeqIO.parse(sample+'.fa', 'fasta')\n",
    "    for read in zipfa:\n",
   "        if read.id in qseqids:\n",
   "            qseqdict[read.id] = read.seq\n",
   "    !rm {sample}.fa\n",
   "\n",
   "    qseqs = []\n",
   "    for idx, eachrow in subset.iterrows():\n",
   "        qseqs.append(str(qseqdict[eachrow['qseqid']]))\n",
   "    subset['qseq'] = qseqs\n",
   "    subset.to_csv('../blastCOIhit/'+sample+'-ident98.tsv', sep='\\t')\n",
   "    \n",
   "pd.DataFrame.from_dict(numreads).to_csv('numreads_stat.csv', sep='\\t')"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "---\n",
   "## Extract top 5 and top 1\n",
   "### Counting number of results"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 29,
  "metadata": {},
  "outputs": [
   {
    "name": "stdout",
    "output_type": "stream",
    "text": [
     "Time for PK-35 0:00:18.185733\n"
    ]
   }
  ],
  "source": [
   "t0 = datetime.now()\n",
   "for sample in samples:\n",
   "    if not path.exists('../blastCOIhit/'+sample+'-iden98.tsv'):    continue\n",
   "    ident = pd.read_csv('../blastCOIhit/'+sample+'-iden98.tsv', sep='\\t', index_col=0)\n",
   "\n",
   "    qseqids = []\n",
   "    best5 = []\n",
   "    best1 = []\n",
   "\n",
   "    for idx, eachrow in ident.iterrows():\n",
   "        qseqid = eachrow['qseqid']\n",
   "        if qseqid in qseqids:    continue\n",
    "\n",
    "        qseqids.append(qseqid)\n",
    "\n",
    "        subset = ident[ident['qseqid'] == qseqid]\n",
    "        subset = subset.sort_index('index')\n",
    "        ## the index is generated by blastn output\n",
    "        ## the index is arranged according to qseqid, then e-value\n",
    "        [best5.append(x) for x in subset.index[:5]]\n",
    "        [best1.append(x) for x in subset.index[:1]]\n",
    "\n",
    "    ident.loc[best5, :].to_csv('../blastCOIhit/'+sample+'-top5.tsv', sep='\\t')\n",
    "    ident.loc[best1, :].to_csv('../blastCOIhit/'+sample+'-top1.tsv', sep='\\t')\n",
    "\n"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 31,
   "metadata": {},
   "outputs": [],
   "source": [
    "numfilter = dict()\n",
    "keys = ['num_iden98', 'panther_iden98', 'puncia_iden98', 'other_iden98', 'num_top5',
'panther_top5', 'puncia_top5', 'other_top5', 'num_top1', 'panther_top1', 'puncia_top1',
'other_top1']\n",
    "for key in ['sample'] + keys:\n",
    "    numfilter[key] = []\n",
    "\n",
    "    \n",
    "for sample in samples:\n",
    "    numfilter['sample'].append(sample)\n",
    "    if not path.exists('../blastCOIhit/'+sample+'-iden98.tsv'):\n",
    "        for key in keys:\n",
    "            numfilter[key].append(0)\n",
    "        continue\n",
    "        \n",
    "    for filterstep in ['iden98', 'top5', 'top1']:\n",
    "        pantheridx = []\n",
    "        punciaidx  = []\n",
    "        neitheridx = []\n",
    "\n",
    "        ident = pd.read_csv('../blastCOIhit/'+sample+'-'+filterstep+'.tsv', sep='\\t',
index_col=0)\n",
    "        numfilter['num_'+filterstep].append(ident.shape[0])\n",
    "        for idx, eachrow in ident.iterrows():\n",
    "            if 'Panthera_uncia' in eachrow['stitle']:\n",
    "                punciaidx.append(idx)\n",
    "            elif 'Panthera' in eachrow['stitle']:\n",
    "                pantheridx.append(idx)\n",
    "            else:\n",
    "                neitheridx.append(idx)\n",
    "\n",
    "        numfilter['puncia_'+filterstep].append(len(punciaidx))\n",
    "        numfilter['panther_'+filterstep].append(len(pantheridx))\n",
    "        numfilter['other_'+filterstep].append(len(neitheridx))\n",
    "\n",
    "pd.DataFrame.from_dict(numfilter).to_csv('numfilter_stat.csv', sep='\\t')\n",
    "\n"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "### Counting number of reads per prey"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 42,
   "metadata": {
    "scrolled": true
   },
   "outputs": [],
   "source": [
    "allprey = dict()\n",
    "\n",
    "for idx, sample in enumerate(samples):\n",
    "    if not path.exists('../blastCOIhit/'+sample+'-top1.tsv'):    continue\n",
    "\n",
    "    ident = pd.read_csv('../blastCOIhit/'+sample+'-top1.tsv', sep='\\t', index_col=0)\n",
    "    preys = list(ident.stitle)\n",
    "    preys = [x.split(';')[-1] for x in preys]\n",
    "\n",
    "    for prey in preys:\n",
    "        if prey not in allprey.keys():    allprey[prey] = [0]*96\n",
    "        allprey[prey][idx] += 1\n",
    "\n",
    "allprey = pd.DataFrame.from_dict(allprey, orient='index')\n",
    "allprey.columns = samples\n",
    "allprey.to_csv('species_cnt.csv', sep='\\t')"
   ]
  }
 ],
 "metadata": {
  "kernelspec": {
   "display_name": "Python 3",
   "language": "python",
   "name": "python3"
  },
  "language_info": {
   "codemirror_mode": {
    "name": "ipython",
    "version": 3
   },
   "file_extension": ".py",
   "mimetype": "text/x-python",
   "name": "python",
   "nbconvert_exporter": "python",
   "pygments_lexer": "ipython3",
   "version": "3.5.2"
  }
 },
 "nbformat": 4,
 "nbformat_minor": 4
}
```

Appendix 3: List of species with their GenBank accession numbers used for prey identification Method II.

	Accession #	Scientific name	Common name
1.	MF784603.1	*Alces alces*	Moose
2.	NC_006380.3	*Bos grunniens*	Domestic yak
3.	NC_006853.1	*Bos taurus*	Cow
4.	NC_008092.1	*Canis lupus*	Gray wolf
5.	FJ207525.1	*Capra falconeri*	Markhor
6.	NC_005044.2	*Capra hircus*	Domestic goat
7.	FJ207526.1	*Capra ibex*	Siberian ibex
8.	NC_025271.1	*Capreolus pygargus*	Siberian roe deer
9.	HM049636.1	*Cervus albirostris*	Thorold's deer
10.	KP172593.1	*Cervus elaphus*	Red deer
11.	NC_001640.1	*Equus caballus*	Horse
12.	HM118851.1	*Equus hemionus*	Wild ass
13.	NC_024030.1	*Equus przewalskii*	Przewalskii horse
14.	JN632643.1	*Gazella subgutturosa*	Goitered gazelle
15.	FJ207531.1	*Hemitragus jemlahicus*	Himalayan tahr
16.	NC_034002.1	*Lagopus muta*	Rock ptarmigan
17.	KR019013.1	*Lepus timidus*	Mountain hare
18.	NC_025748.1	*Lepus tolai*	Tolai hare
19.	NC_018367.1	*Marmota himalayana*	Himalayan marmot
20.	KM347744.1	*Martes flavigula*	Yellow-throated marten
21.	HM106325.1	*Martes foina*	Stone marten
22.	NC_011125.1	*Meles meles*	Eurasian badger
23.	NC_012694.1	*Moschus berezovskii*	Forest musk deer
24.	KC425457.1	*Moschus chrysogaster*	Alpine musk deer
25.	NC_042604.1	*Moschus leucogaster*	Himalayan musk deer
26.	KT337321.1	*Moschus moschiferus*	Siberian musk deer
27.	NC_025269.1	*Mus cervicolor*	Fawn-colored mouse
28.	NC_005089.1	*Mus musculus*	House mouse
29.	NC_020639.1	*Mustela nivalis*	Least weasel
30.	EF535828.1	*Ochotona curzoniae*	Plateau pika
31.	NC_044120.1	*Ochotona dauurica*	Pika
32.	NC_037186.	*Ochotona erythrotis*	Chinese red pika
33.	NC_039987.1	*Ochotona koslowi*	Pika
34.	KT781689.1	*Ovis ammon*	Argali
35.	NC_001941.1	*Ovis aries*	Domestic sheep
36.	NC_026063.1	*Ovis orientalis breed*	Asian mouflon
37.	NC_026064.1	*Ovis vignei*	Urial
38.	NC_039591.1	*Paradoxurus hermaphroditus*	Asian palm civet
39.	NC_016707.1	*Przewalskium albirostris*	White-lipped deer
40.	FJ207537.1	*Pseudois nayaur*	Blue sheep
41.	KU962990.1	*Sciurus vulgaris*	Red squirrel
42.	KX146493.1	*Sus scrofa*	Wild boar

43.	NC_027279.1	*Tetraogallus himalayensis*	Himalayan snow cock
44.	NC_003427.1	*Ursus arctos*	Himalayan brown bear
45.	JN711443.1	*Vulpes vulpes*	Red fox