SCHREYER HONORS COLLEGE

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

SEQUENCING THE GENOME OF THE NITTANY LION (*PUMA CONCOLOR* FROM PENNSYLVANIA)

MAYA EVANITSKY
SPRING 2016

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree in Biochemistry and Molecular Biology
with honors in Biochemistry and Molecular Biology

Reviewed and approved* by the following:

George Perry
Assistant Professor of Anthropology and Biology
Thesis Supervisor

Santhosh Girirajan
Assistant Professor of Biochemistry & Molecular Biology and Anthropology
Thesis Supervisor

Teh-hui Kao
Distinguished Professor of Biochemistry & Molecular Biology
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Mountain lions (*Puma concolor)* were previously endemic across Pennsylvania. The species was officially declared regionally extinct in 2011 by the U.S. Fish and Wildlife Service, although the last time that a mountain lion was observed east of the Mississippi River was in Maine in 1938, excluding the current population that resides in Florida. The Northeastern population of *P. concolor* has been almost nonexistent since the early 1800s, most likely due to targeted hunting depredation by farmers to protect livestock, along with habitat destruction and fragmentation. The last documented observation of *P. concolor* in Pennsylvania specifically was in 1874. The primary goal of this study is to sequence the mitochondrial DNA (mtDNA) genomes of multiple Pennsylvania *P. concolor* individuals using Penn State University's Anthropology/Biology ancient DNA (aDNA) laboratory and Huck Institutes of the Life Sciences core sequencing facility. Complete mtDNA genome sequences were obtained from five individuals and analyzed along with previously published *P. concolor* mtDNA sequences from the Western U.S. and Florida to determine the genetic diversity lost with the extinction and update the current phylogeny. Interestingly, no clustering of distinct populations was observed in the phylogeny and one of the four haplotypes found for the entire U.S. mountain lion population was common to all three regions (Northeastern U.S., Western U.S., and Florida). This result further supports previous findings documenting the low diversity found in North American mountain lions and emphasizes the need for continued conservation efforts in the Western U.S. and Florida to prevent further regional extinctions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank everyone in the Perry lab for helping make this project possible, especially Logan Kistler, Richard George, and Stephen Johnson for their advice on conducting experiments and performing analyses. I would also like to thank everyone who helped fundraise for and donate to this project and Rose Carter in particular for all her support. Donations were provided by the Penn State College of the Liberal Arts Undergraduate Research Fund, the Mountain Lion Foundation, Felidae Conservation Fund, the Penn State Huck Institutes of the Life Sciences, and more than 140 other individuals and organizations. Museum specimens were provided by the Penn State All-Sports Museum, the Albright College Biology Dept., the Thomas T. Taber Museum of the Lycoming County Historical Society, Bill and Carol Thomas, and the Pennsylvania State Museum.

Chapter 1

Introduction

Brief History of Mountain Lions

Fossil evidence indicates that the species *Puma concolor* (also referred to as mountain lions, cougars, pumas, and catamounts) originated in North America prior to colonizing South America[1]. *P. concolor* has existed in North America for at least 400,000 years and diverged from *P. yagouaroundi* 4.17 MYA[2,3]. According to a study done by Barnett et al. in 2005, *Miracinonyx trumani*, an extinct felid of North America, is the closest relative to mountain lions, and diverged 3.19MYA[1]. While *Puma concolor* was once endemic throughout North and South America, they are now practically nonexistent east of the Mississippi river, except for a small population in Florida[4]. Of the thirty classical subspecies of *P. concolor* recognized worldwide, thirteen are located in North America north of Mexico[5]. Three of these subspecies are considered endangered (including the Florida panther) while the eastern cougar (*P. concolor couguar)* and the Wisconsin puma (*P. concolor schorgeri*) have been declared extinct[4]. The eastern cougar was officially declared regionally extinct in 2011 by the U.S. Fish and Wildlife Service and was unofficially considered extinct in Pennsylvania since the 1930s[6]. In fact, the last known sighting of a mountain lion east of the Mississippi River was in Maine in 1938[6].

A phylogeographic study by Culver et al. conducted in 2000 analyzed three regions of the mitochondrial genome, the terminal intron of the *Zfy* region on the Y chromosome, and microsatellite data for 315 cougars from across North and South America. Of these individuals,

three were museum specimens from the Northeastern U.S. Culver et al. concluded that the 30 classical subspecies were not genetically distinct and condensed them down to six subspecies, four of which are located in South America. North America and Central America contain only one each and were found to have significantly less genetic diversity, supporting the theory that North America was re-colonized by South American cougars following a mass extinction event in the Late Pleistocene that eliminated approximately 80% of all large mammals[4].

Due to habitat fragmentation and drastic population decreases, the Northeastern cougar, Florida panther, and Western mountain lion have distinct breeding populations with severely restricted gene flow between them since the mid-1850s, although they are not genetically distinct subspecies, and will be referred to as populations throughout this paper. Targeted hunting of mountain lions began with the colonization of North America due to the perceived threat to livestock and European settlement. Even more damaging, colonization led to the mass destruction of forests in the Northeast and the cougar's primary habitat. The introduction of livestock into the Northeastern and Midwestern U.S. led to a marked decrease in available land for white tail deer and other mountain lion prey.

Mountain lions prey mostly on white tail and mule deer in North America, making them dependent on deer populations, which they will migrate to follow[7]. It's been theorized that the over-hunting and drastic population decreases in deer that occurred following the colonization of the Northeastern U.S. is a large factor in the extinction and endangerment of mountain lions[8]. Without sufficient land or prey, the cougar population increasingly dwindled and was hunted to effective extinction by the early 1900s. The Northeastern population of *P. concolor* in particular has been almost nonexistent since the early 1800s, likely due to targeted depredation by farmers

to protect livestock, along with habitat destruction and fragmentation[7]. The last documented observation of *P. concolor* in Pennsylvania specifically was in 1874[9].

Today, mountain lion populations can be found in the Western U.S. and the southern tip of Florida[4,10]. There have been many reported sightings of mountain lions in the Northeastern U.S., and several have left behind genetic evidence in the form of scat and bodies hit by passing cars. None of these reports found evidence of true Northeastern mountain lions; all had travelled great distances from the Western U.S. It has been suggested this is due to a significant expansion in deer populations in the Northeastern U.S. caused by lack of predators and increased regulations in hunting[8].

Several studies have shown that habitat fragmentation has a huge impact on species survival. The separation of a single breeding population into several smaller ones due to highway development and the expansion of human settlements has a large detrimental effect on genetic diversity. Mountain lions in particular are incredibly sensitive to habitat destruction and fragmentation due to their solitary natures and need for large hunting ranges[11]. Mountain lions have vast ranges that can vary from 25 to 400 square miles and have been known to travel great distances in search of food and/or mates. They are solitary predators that do not have overlapping ranges for the most part and practice mutual avoidance in the event that ranges overlap due to prey population density[7].

Ernest et al. conducted a study on California mountain lions and found that mountain lions living in the Santa Ana Mountains, which are surrounded by urbanization or cut off from other mountain lions by a major interstate, had much lower genetic diversity and showed strong evidence of a bottleneck effect and isolation. Although the Santa Anas are in close proximity to the Eastern Peninsular Range, another mountain lion habitat, a ten-lane highway separates the

two. Despite this proximity, mountain lions from the Santa Anas rarely bred with those from the Peninsular Range[11]. This isolation could potentially lead to inbreeding problems, as was seen in Florida panthers[10]. The population bottleneck was estimated to occur between 40-80 years ago, which overlaps with major urban development and the construction of major highways in the area. This study highlights the danger mountain lions still face throughout the U.S. and the need for conservative efforts to protect them from habitat destruction[11].

A similar issue affected Florida panthers before the addition of Texas females to increase genetic diversity. The population was restricted between two major urban centers, Miami and Naples, and had significantly lower genetic diversity than other populations. To save the Florida population, eight female wild Texas panthers were introduced into this habitat in 1995 since gene flow used to occur between the two. This restoration attempt increased the number of panthers by a factor of three, doubled genetic diversity and decreased inbreeding. However, this successful measure did not solve the issues of habitat loss and fragmentation or infectious agents that continue to put the Florida panther at risk[10]. Therefore, conservation efforts are still necessary to protect and maintain the population of Florida panthers, which still face risk of extinction.

Goals

The goals of this study are to sequence the mitochondrial genome of extinct mountain lions from the Northeastern U.S. that have not yet been analyzed, introduce new information about extinct population of mountain lions, and determine the genetic diversity of this extinct population. Skin samples were collected from specimens in Pennsylvania museums that had been

preserved in the late 1800s via taxidermy (Figure 1). Mitochondrial DNA (mtDNA) was chosen over nuclear DNA as mtDNA is more abundant in almost all eukaryotic cells than nuclear DNA. In particular, most skin cells do not contain a nucleus or nuclear DNA and thus only have mitochondrial DNA[12]. This is especially important to consider when processing samples that are upwards of 150 years old and have been exposed to processes that have been shown to significantly damage DNA. Due to the age of the samples and the damage that traditional taxidermy processes can cause genetic material, all were treated according to standard ancient DNA (aDNA) protocols[13]. Although less informative than nuclear DNA, mitochondrial genetic markers are still powerful tools that have been used extensively for felid studies.

Importance

An additional goal of this study was to raise awareness of the importance of extinction risk and processes and endangered species conservation efforts, especially among the broader Pennsylvania and Penn State community, for whom the "Nittany" Lion is a beloved mascot. This was accomplished in part through fundraising efforts and outreach efforts will continue to disperse the results and implications of this study. In addition to raising awareness, this study provides novel complete mitochondrial genome sequences for five mountain lion individuals which can provide a finer genetic structure and a better measure of diversity than select regions of the mitochondrial genome. Furthermore, this study provides information about the quality of endogenous DNA in these samples, which will be useful for complete nuclear genome sequencing. Complete nuclear genome sequence data will be a substantially better measure of

genetic diversity than mitochondrial DNA and of particular import as no complete nuclear genome sequences for mountain lions have been constructed at present.

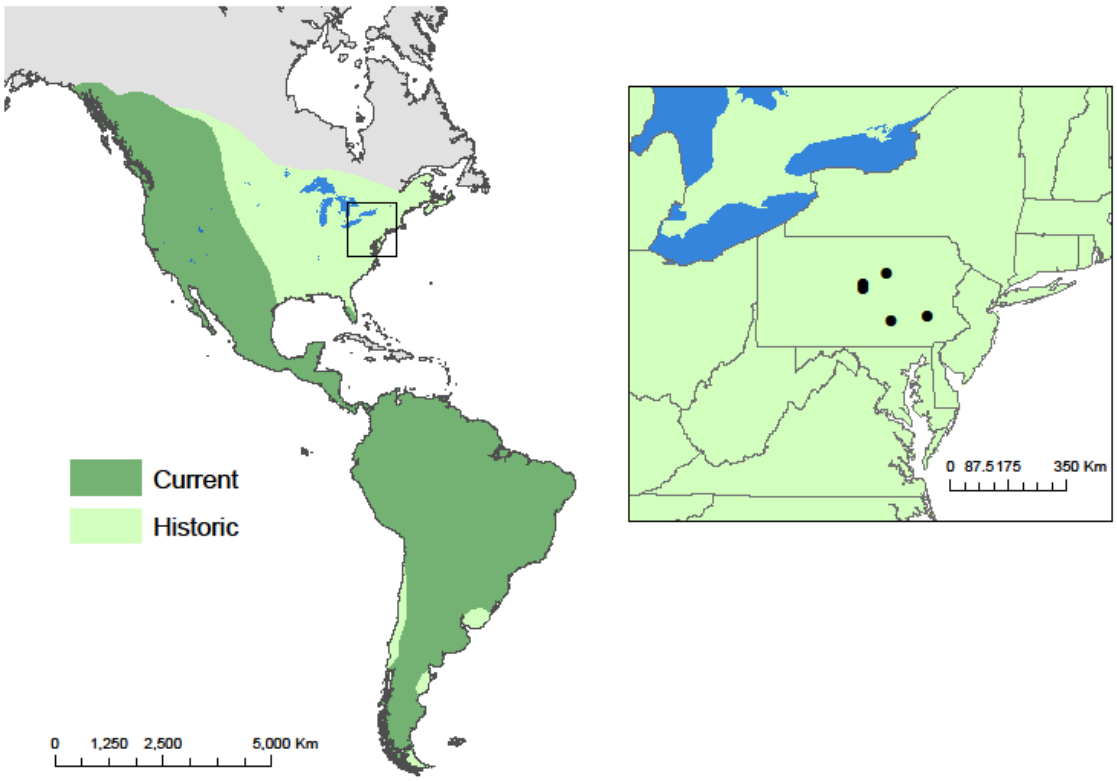

Figure 1: Map of Collection Sites and Mountain Lion Ranges

Samples were collected from Albright College; Bald Eagle State Park; the Pennsylvania State Museum; Howard, PA; and the Thomas T. Taber Museum. The current and historic ranges for mountain lions are also displayed.

Chapter 2

Methods

Sample Collection

Skin samples were obtained from six Northeastern mountain lions preserved using taxidermy prior to the 1900s. Due to the age and morphological importance of the specimens, samples were collected in such a way as to cause as little damage to the specimen as possible. A lab coat, facemask, and gloves were worn to prevent as much modern DNA contamination as possible, as well as to provide personal protection from any harmful chemicals used in the preservation process (Figure 2).

Figure 2: Mountain Lion Specimens and Sampling

Two of the six museum specimens that were sampled for DNA extraction are shown. The left photo is of the sample located at Bald Eagle State Park and the right photo is of the specimen at the Penn State All-Sports museum.

Skin samples removed were no larger than one square inch. Five samples were personally collected by the author using sterilized scalpel and scissors from local Pennsylvania museums,

the visitor center at Bald Eagle State Park, and a private residence while a sixth sample was donated by Albright College (Table 1).

Table 1: Sample Information

Sample ID	Current Location	Museum/Owner	Time of Death
Alb813	Reading, PA	Albright College	1870s
BE813	Howard, PA	Bald Eagle State Park Visitor's Center	unknown
Har813	Harrisburg, PA	Pennsylvania State Museum	1874
RSL813	Howard, PA	Privately Owned	unknown
Wmt813	Williamsport, PA	Thomas T. Taber Museum	1880s
NL813	State College, PA	Penn State All-Sports Museum	1856

DNA Extraction

All sample processing and DNA extractions were performed in a clean lab dedicated to ancient DNA and done according to standard aDNA protocol[13]. The lab is located in the physics building on campus, completely separated from any ongoing modern genetics work, and has positive pressure, HEPA filtration, and stringent work protocols to prevent any modern DNA contamination. DNA was extracted using a phenol: chloroform protocol modified from Casas-Marce et al., Dabney et al and de Moraes-Barros & Morgante[14–16]. All equipment and reagents (as appropriate) were sterilized by UV-irradation for one hour prior to use.

In preparation for digestion, samples were cut into small pieces of approximately 2-3 cm^2. Samples were first immersed in 50% bleach for 30 seconds to remove any contaminant DNA, then rinsed with ethanol and molecular grade water three times each. Each sample was washed three times with 1.5 mL of NTE solution (50 mM tris, pH 9; 20 mM EDTA; 10 mM

NaCl) to remove potential inhibitors. Samples were then hydrated by incubation for 24 hours in 1 mL of TE solution (tris 10 mM, EDTA 1 mM, pH 7.6) and washed with 70% (w/v) ethanol and molecular grade water before a second hydration in TE solution for a further 24 hours. Samples were cut in half prior to digestion using a sterile razor blade.

For the digestion, one extraction blank was used for every six samples. 500 μL lysis buffer (Tris HCl 10 mM, pH 8; NaCl 400 mM; EDTA 2 mM, pH 8.0; SDS 1%), 100 μL of Proteinase K (20mg/mL) and 100 μL of 500 mM Dithiothreitol were added to each sample. Additional lysis buffer was added until the total volume reached 1 mL. Sample tubes were then placed on a rotary mixer and rotated overnight with incubation at 55°C or until sample was fully digested.

One volume phenol was added to each sample, which was then rotated at room temperature for five minutes and centrifuged at maximum speed for 8 minutes until phases separated. The aqueous layer was transferred to a new tube. This was repeated a second time. One volume chloroform was added to the remaining aqueous layer and rotated at room temperature for five minutes. The aqueous layer was transferred to a clean 1.5 mL tube.

Following phenol: chloroform extraction, the DNA was purified and concentrated using a MinElute PCR Purification Kit with modifications to retain short DNA fragments, based on the procedure by Dabney et al.[15]. 6 mL of binding buffer (5 M guanidine hydrochloride, 90 mM sodium acetate, 40% isopropanol, 0.05% Tween-20) was mixed with each sample. Qiagen tube extenders were attached to each MinElute column and placed in 50 mL conical tube. Tube extenders and conical tubes were UV-irradiated prior to use. The extraction-binding buffer mixture was added to a tube extender and centrifuged for four minutes at 1,500 x g in a large centrifuge. The filtrate and tube extenders were discarded and the MinElute column placed in a 2

mL collection tube. The column was dry spun at 3,300 x g in bench centrifuge for one minute. The column was washed twice with 700 μL PE buffer and spun at 3,300 x g in bench centrifuge for one minute. Column was spun once more for one minute at max speed to remove all traces of ethanol. The column was transferred to a clean 1.5 mL tube. 30 μL of TET was added and incubated on the column for 15 minutes at 37°C. The column was centrifuged for one minute at maximum speed to elute.

Library Preparation

Genomic sequencing libraries were constructed according to the Meyer & Kircher protocol with a few modifications for ancient DNA[17]. During this process, any overhanging DNA ends created by fragmentation are repaired and adaptors made up of oligonucleotides are attached to both ends of each DNA fragment. These adaptors allow for almost any type of nucleotide index to be affixed to DNA fragments. Specific indices are used for each sample to allow for parallel sequencing of multiple samples on one lane. Indices allow DNA sequences to be traced back to the source sample following sequencing.

Samples were not sheared prior to end repair as aDNA samples are already heavily degraded and fragmented[13]. For the final PCR amplification step, the samples were split so 10 μl of each sample was used for a 50 μl reaction. Each sample was given a unique nucleotide barcode to allow for pooling samples during sequencing.

MiSeq Test Run

All samples with sufficient concentrations were pooled to equivalent DNA mass and sequenced on the Illumina MiSeq using 2 x 75 bp read length with paired-end reads to test the success of extraction and library construction. Sequence data were first processed and adaptors removed using a custom perl script adapted from Martin Kircher designed for ancient DNA analyses[18]. Sequence reads were then aligned to the complete *Puma concolor* mitochondrial genome available on Genbank (accession number NC_016470) using the Burrows-Wheeler Aligner (BWA)[19,20]. Samples were processed and duplicates removed using the Samtools package. Quality was determined by quantifying the number of total mapped reads, uniquely mapped reads, average depth coverage and determining read length distribution using the Samtools package and commands in awk. C-to-T mismatch was also determined as this is a unique characteristic of aDNA. Cytosine deaminates to uracil at single-stranded overhanging ends of DNA fragments, giving a higher proportion of C-to-T mismatch at the 5' end of DNA fragments[13]. For a DNA samples, the amount of C-to-T substitution has been found to increase with age; one study examining methods to distinguish modern DNA from aDNA found that in samples less than 100 years old less than 10% of cytosine residues appeared as thymines while in samples 500 years or older, more than 10% of cytosines appeared as thymines. It has also been established that aDNA samples have a higher proportion of C-to-T substitutions at the ends of sequence reads than in the middle of reads. For example, a 500 year old horse sample had 20% mismatch at position 1 of a read, but only 5% at position 13, further from the end of the read[21]. The sequencing errors caused by these damaged induced mismatches are typically masked during data analysis. A lower proportion of mismatches were observed for these samples than is typical for aDNA, very likely due to the relatively recent age of the samples. The damage was

sufficiently low that errors in generating consensus sequences due to damage are not expected (Figure 3).

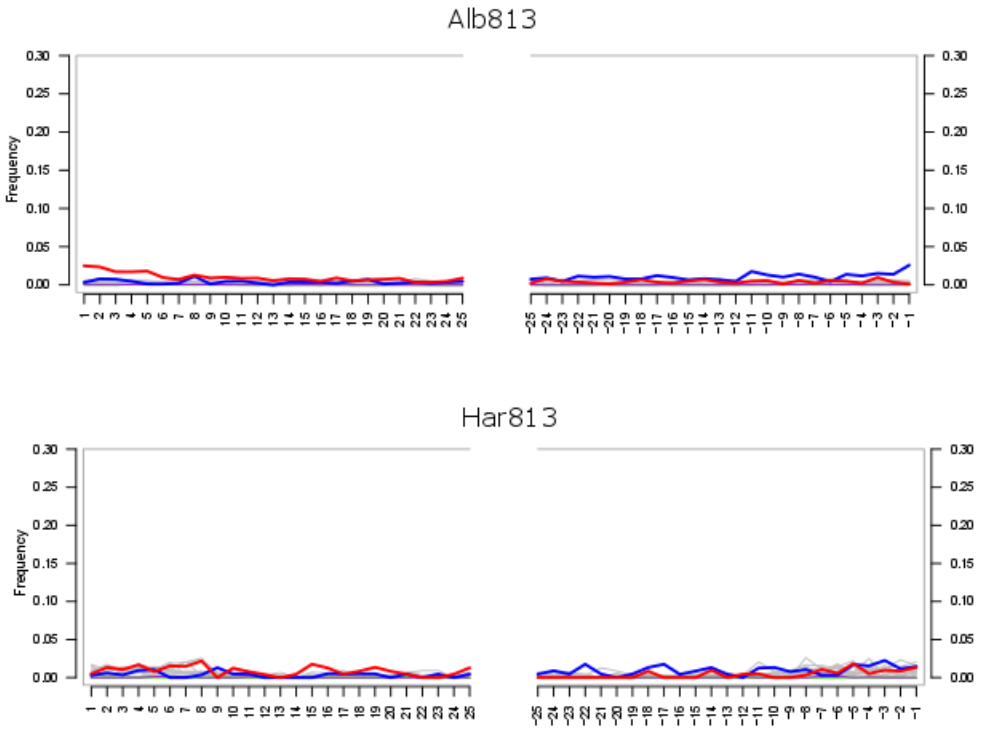

Figure 3: C-to-T Misincorporation Per Base Pair Position

The plot indicates the frequency of the 5' C-to-T and 3' G-to-A substitutions per base pair position on fragments for two samples: Alb813 and Har813. All five samples had a misincorporation frequency less than 0.05.

Three out of the six samples were successfully sequenced to adequate coverage (with each base pair covered more than three times on average), while two others had evidence of mountain lion DNA present but not yet sufficient for reconstructing the complete mtDNA genome (Table 2).

Table 2: MiSeq Sequence Data Quality Analysis

Sample ID	Total Reads	Non-redundant Mapped Reads*	Non-redundant Uniquely Mapped Reads*	Average Depth Coverage*
Alb813	3698194	74	25	0.0683
BE813	9381398	6863	6705	32.1
Har813	5164026	268	71	0.224
NL813	402	2	0	0
RSL813	680853	1260	1178	4.58
Wmt813	1722516	1137	977	3.15

*Calculated after removing duplicate reads using Samtools

Bait Capture Reaction

Due to the low amount of sequence reads, one sample (Table 1: NL813) was filtered using Microcon 30 kDa centrifuge tubes following DNA extraction in an attempt to remove any potential arsenic residue before constructing a new library and proceeding to DNA enrichment. Five out of the six samples underwent further processing before being sequenced on the Illumina HiSeq platform. Sample BE813 was not enriched due to the high sequencing quality obtained on the MiSeq test run (Table 2) so as to preserve materials for future experiments. The five samples were enriched for endogenous DNA using MYcroarray Mybaits kit. The Mybaits kit works by using RNA oligonucleotides designed to bind endogenous DNA, with the RNA probes designed to be complementary to the DNA sequence of interest[22]. These probes are biotinylated and can be bound to streptavidin-coated magnetic beads, which are used to pull probes and bound DNA from the solution with a powerful magnet. Any unbound DNA is removed by washing. After the exogenous DNA is removed, the bound DNA is released from the beads and amplified.

Biotinylated RNA probes were designed using the *P. concolor* reference mitochondrial genome[20]. Protocol version 3.0 was followed, using the manufacturer's suggested modifications

for aDNA. Baits and DNA sequencing libraries were incubated together for 40 hours, rather than the standard 16-24 hours suggested for modern DNA. Following binding bait to the beads, 180µl of wash buffer was used to wash samples for a total of four times. Final amplification of the captured libraries was repeated for a total of 24 cycles.

Sequencing on the Illumina HiSeq

All six samples (five enriched and one non-enriched library) were sequenced in parallel one a single lane using the Illumina HiSeq 2500 with the same parameters as the MiSeq run. Samples were analyzed for quality according to the same protocol as described above. Five out of the six samples were successfully sequenced. One sample (NL813) failed on both the MiSeq and HiSeq runs, indicating either the extraction failed or the arsenic reported to be in the sample interfered significantly with library construction (Table 3).

Table 3: HiSeq Sequencing Data Analysis

Sample ID	Total Reads	Non-redundant Mapped Reads*	Non-redundant Uniquely Mapped Reads*	Depth Coverage*
Alb813	25450196	4523	4079	16.6636
BE813	18691338	9331	8819	41.7265
Har813	21357289	1611	1099	4.83822
NL813	288	0	0	0
RSL813	27947047	8183	7579	30.7857
Wmt813	31508515	4427	3684	13.5467

*Calculated after removing duplicate reads using Samtools

Consensus sequences were constructed by using Samtools mpileup and bcftools to call and filter variants. Variants with an overall quality lower than 25 and less than 5X read coverage were removed. The consensus was constructed by applying variants to the *P. concolor* reference sequence using bcftools. Genetic analyses were conducted by comparing consensus sequences for Northeastern mountain lions with previously published mitochondrial sequences for modern populations of mountain lions in the Western U.S. and Florida[4,20,23]. Sequences were aligned using MAFFT and verified visually.

The 16s rRNA, ATPase8, and NADH dehydrogenase subunit 5 (NAD5) gene regions were used for comparing Northeastern cougars to Western and Florida panthers due to the availability of published data and limited availability of contemporary samples for DNA extraction and sequencing. Overall genetic distance and phylogenetic tree analyses were constructed using MEGA version 7. Phylogenetic trees were constructed using maximum-likelihood and Neighbor-joining methods[24]. Sequences were further analyzed via AMOVA and FST using Arlequin version 3.5.2.2.

Chapter 3

Results

Mitochondrial DNA was extracted and sequenced (following RNA bait capture on five samples) from six samples obtained from museums and colleges in Pennsylvania (Figure 1). Of those samples, five were successfully sequenced for the entire mitochondrial genome. mtDNA sequences were compared to genetic data from Culver et al.[4] (201 individuals from North America) and the *P. concolor* reference[20] sequence on Genbank for analyses of select regions. Analyses of the entire mitochondrial genome involved all five Northeastern U.S. samples from this study and the *P. concolor* reference sequence, which was obtained from a Texas mountain lion. Across the mitochondrial genome, 101 positions were found to be variable among Northeastern mountain lions and 112 positions were variable between Northeastern mountain lions and the reference sequence from Texas. Pairwise distances between Northeastern mountain lions and the reference sequence were calculated using MEGA7 (Table 4). 891 base pairs in the 16s rRNA, ATPase8, and NAD5 genes were analyzed to determine AMOVA and FST using Arlequin V3.5.2.2. In those 891 base pairs, four single nucleotide polymorphisms were found, generating four distinct haplotypes (Table 5).

Table 4: Estimates of Nucleotide Sequence Distance Among U.S. Mountain Lions

	***Puma concolor* reference (Texas)**	**Alb813**	**BE813**	**Har813**	**RSL813**	**Wmt813**
***Puma concolor* reference**		0.0005	0.0004	0.0002	0.0004	0.0004
Wmt813	0.0042		0.0003	0.0005	0.0004	0.0003
RSL813	0.0045	0.0020		0.0004	0.0003	0.0002
Har813	0.0009	0.0036	0.0037		0.0004	0.0004

BE813	0.0052	0.0033	0.0016	0.0048		0.0003
Alb813	0.0037	0.0013	0.0008	0.0029	0.0023	

The number of base substitutions per site from between sequences are shown. Standard error estimate(s) are shown above the diagonal and were obtained by a bootstrap procedure (500 replicates). Analyses were conducted using the Tamura-Nei model [1]. The analysis involved 6 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 17149 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [2].

Table 5: Geographic Location and Distribution of Haplotypes in U.S. Mountain Lions

	16s rRNA[a]			**NAD5**		
Haplotype	2959[b]	2980	3078	12955	N[c]	Geographic Location
H1	T	C	G	T	1	Texas
H2	C	T	·	·	197	Western U.S., Northeastern U.S., Florida
H3	·	T	·	·	1	Pennsylvania
H4	C	T	·	C	4	Washington State

[a]Nucleotides identical to hapotype 1 are indicated by a period
[b]From the *Puma concolor* mtDNA sequence[20]
[c]Number of samples

The SNPs responsible for these haplotypes were verified visually by examining the depth of coverage and fraction of reads that supported the base substitution at each position where a SNP occurred (see Figure 4 for one example region). Each SNP variant call was supported by almost all reads and by all high quality reads. Instances where one or two reads out of several indicated a base pair substitution were considered sequencing errors and ignored.

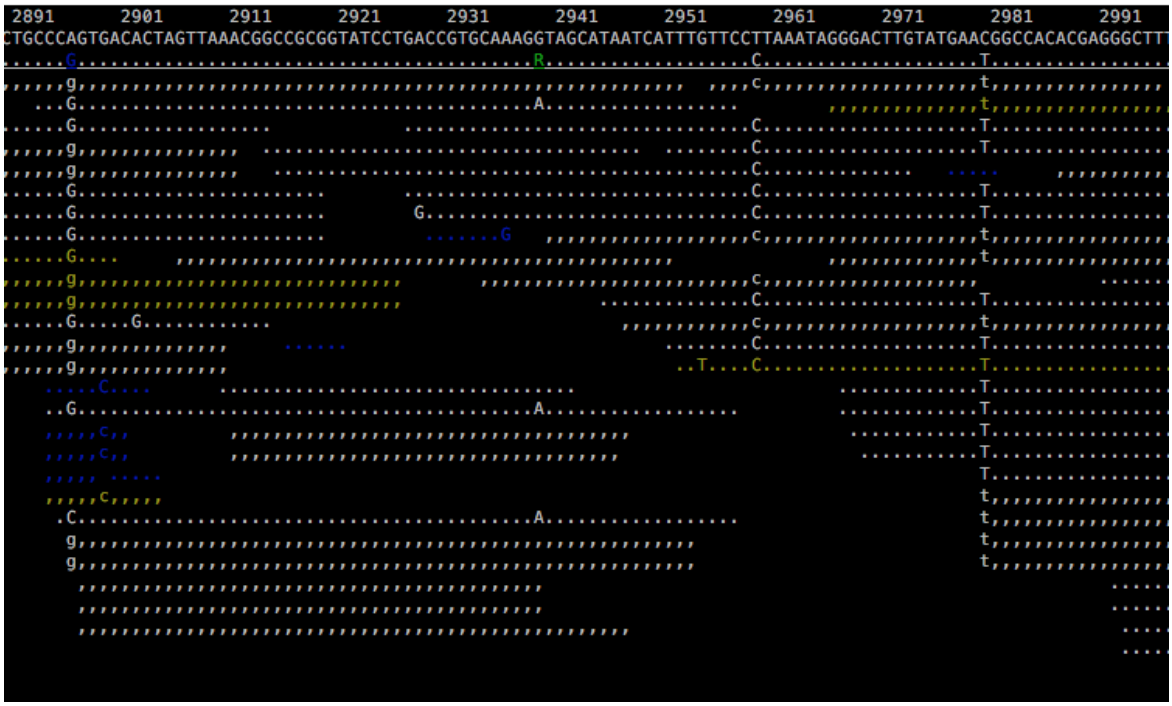

Figure 4: Visualization of SNPs for Wmt813

Alignment was visualized using the Samtools package. Positions containing SNPs are 2896, 2959, and 2980. Colors reflect the quality of the reads, white being the highest quality and blue the lowest.

A haplotype network was generated using PopArt 1.7 (Figure 5). Haplotype two (H2) was common to all three populations (Florida panthers, Western mountain lions and Northeastern cougars) while H1, H3, and H4 were specific to samples from Texas, Williamsport, and Washington State respectively. The Florida population of mountain lions appears to be the least diverse with only one haplotype while the Western population is the most diverse with three distinct haplotypes.

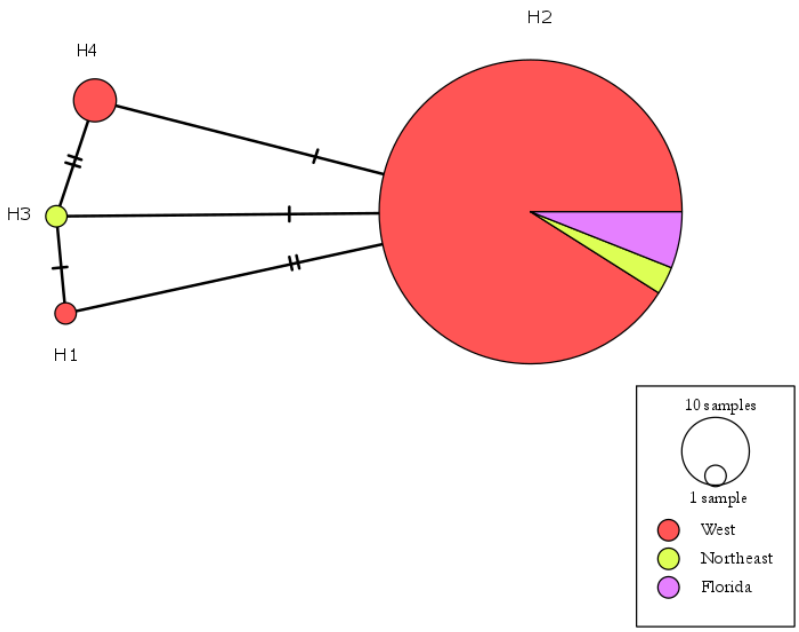

Figure 5: Median-joining Network Based on 3 Gene Regions for U.S. Mountain Lions

Haplotype network was generated using PopArt 1.7 and the Median-Joining Network method. The area of each circle is proportional to the number of samples in each haplotype. Each circle represents a single haplotype and each color represents a population. Hatch marks represent the number of single nucleotide polymorphisms differentiating each haplotype.

The variation within populations was found to account for 94% of the total variation and the FST for all populations was found to be 0.05707 (Table 6). The overall genetic distance for U.S. mountain lions over 891 base pairs was found to be 0.0010 (standard error 0.0006) when calculated using the absolute number of base pair differences (p-distance) and the 2-parameter Kimura model. Across the entire genome, the overall genetic distance for Northeastern and Western mountain lions was found to be 0.00299 (standard error 0.00024).

Table 6: AMOVA of 3 Gene Regions for U.S. Mountain Lions

Source of Variation	Percent Variation
Among populations	5.71
Within populations	94.29
Fixation Index: FST	0.05707
P-value	0.16031

Phylogenetic relationships were determined using maximum-likelihood composition (ML) and neighbor-joining (NJ) methods for two different data sets. Results were concurrent between the two methods. The first data set consisted of all five samples and the *P. concolor* reference sequence[20] and examined relationships across the entire mitochondrial genome (Figure 6). The Texas reference sequence and Har813 are the most divergent samples. Overall, no clustering by population is observed in the phylogenetic analysis. The second data set included data from Culver et al.[4] and included samples from Central and South America and featured 891 base pairs in three gene regions (Figure 7). As expected, distinct clustering is observed between North American and South American populations. Similar to the previous phylogeny, no regional clustering occurs for U.S. mountain lions. However, the latter phylogeny indicates the Texas sample is equally related to all other North American samples and all Northeastern samples, except Har813, are shown to be identical to each other and haplotype M, which has been found across North America. This discrepancy is most likely due to the small portion of the mitochondrial genome analyzed in Figure 7.

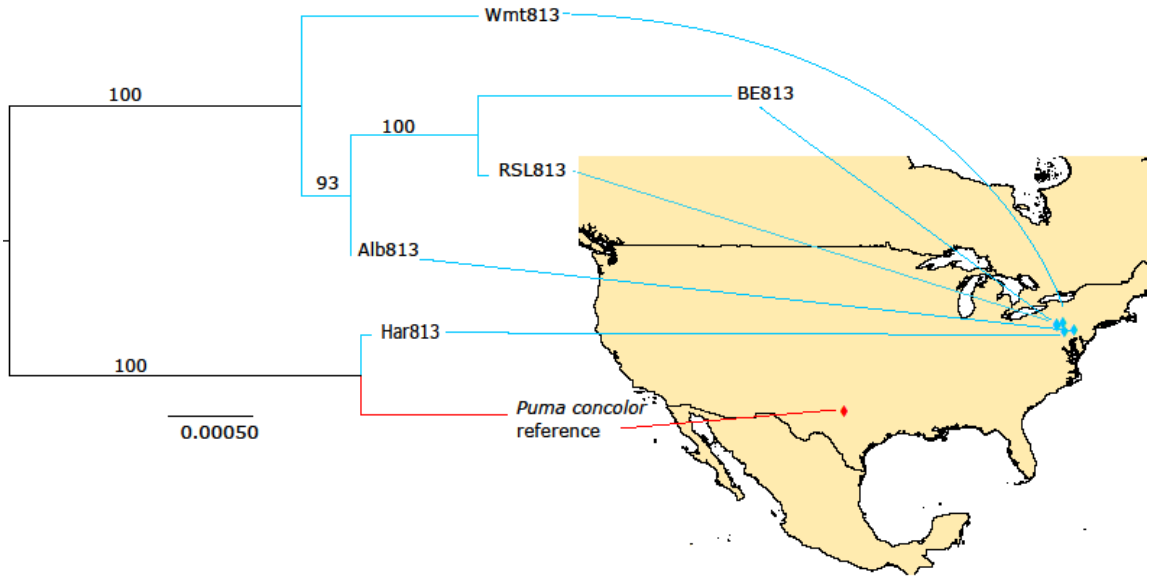

Figure 6: Molecular Phylogenetic Analysis Across the Mitochondrial Genome

The evolutionary history was inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model[25]. The tree with the highest log likelihood (-24013.6555) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 6 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 17149 positions in the final dataset. Evolutionary analyses were conducted in MEGA7[26].

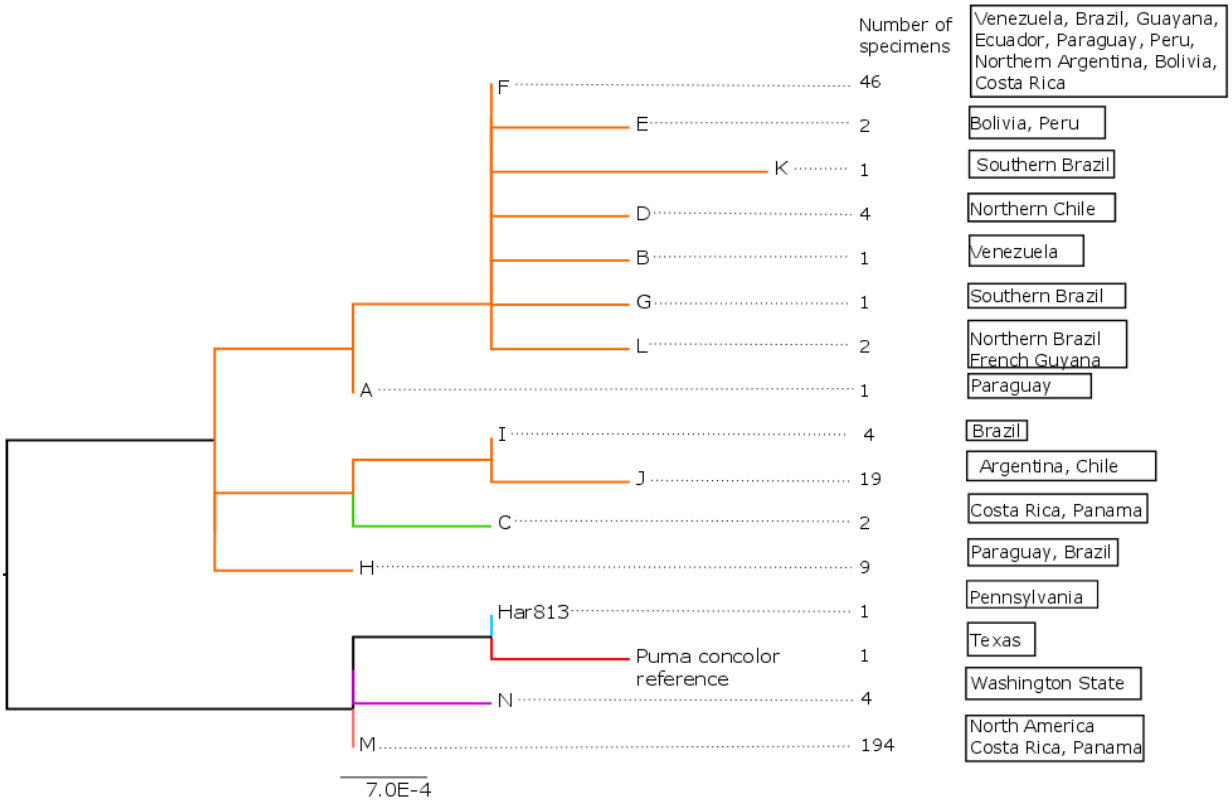

Figure 7: Molecular Phylogenetic Analysis for 3 Gene Regions

The evolutionary history was inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model[25]. The tree with the highest log likelihood (-1337.6891) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 20 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 891 positions in the final dataset. Data were from the five samples, Culver et al.[4] and the *Puma concolor* mtDNA reference sequence[20]. The regions analyzed included 16s rRNA, ATPase8, and NAD5 gene regions. Evolutionary analyses were conducted in MEGA7[26]. The letters reflect the haplotype groups used by Culver et al. The four samples other than Har813 were included with haplotype group M as all were four were identical to this haplotype group for these regions.

Chapter 4

Discussion

This study supports previous findings that U.S. mountain lions can be condensed into one phylogeographic group based on mitochondrial DNA, despite the fact that all three populations have been mostly isolated from one another for over 150 years. While there have been instances in the past century in which Western mountain lions travelled several hundred miles to the Eastern U.S., these are rare occurrences and likely not indicators of consistent gene flow between past populations[8]. The presence of a common haplotype among all three populations supports the levels of low diversity found in North American mountain lions (Figure 5). Additionally, two of the three haplotypes (H2 and H3) specific to a single population were each present in only one individual and are not indicative of overall differentiation between the three populations. Similarly, the low FST value of 0.06435 indicates that almost all variation is caused by variation between individuals rather than between populations (Table 6). This concurs with the distribution of haplotypes seen in Figure 4.

With regards to phylogenetic relationships, the lack of geographical clustering is likely due to the overall low genetic diversity and the ability of mountain lions to travel great distances in search of a mate[7] (Figure 6). Prior to the effects of human colonization, mountain lions were widespread throughout the U.S. and populations were probably not as separated as they are today. This, combined with the low diversity and relatively little evolutionary time since the separation, likely accounts for the existing similarity between populations. Population analyses are also confounded by the forced admixture of Texas and Florida panthers, which was made necessary due to the extremely low diversity and high rate of inbreeding seen in Florida

panthers[10]. Therefore, a conclusion as to whether Northeastern mountain lions were more closely related to western or Florida panthers cannot be made. Perhaps a future analysis sequencing the complete mitochondrial genome of historic Florida samples that existed prior to the addition of Texas panthers into the Florida breeding population could determine this.

Although only Florida panthers are listed on the endangered species list by the U.S. Fish and Wildlife Service, all mountain lions are an important aspect of each ecosystem they inhabit[27]. The loss of such an apex predator from the forests of the Northeast caused vast changes in the environment and has lead to the overpopulation of deer. However, there has been recent evidence that cougars are making a return eastward due to these increasing deer populations[8]. While cougars are adaptable to a variety of habitats, having detailed genetic data about Northeastern mountain lions to compare to Western ones could be of use in conservation and restoring the ecology of the Northeast.

Of course, the methods by which data are analyzed are equally important. There is a significant difference in phylogenetic relationships for U.S. mountain lions depending on how many base pairs are used (Figures 6 and 7). In Figure 6, the sample from Texas and Har813 are shown to be divergent from the rest of the U.S. samples, while in Figure 7, the Texas sample is shown as equally related to all other U.S. samples and all Northeastern samples except Har813 are shown to be identical. These discrepancies demonstrate the impact of using too few base pairs and how it can underestimate genetic diversity. While 891 base pairs may seem to be a reasonable subset, the *P. concolor* mitochondrial genome consists of 17,153 bases, meaning that 891 bases only covers 5% of the mitochondrial genome and may not be sufficient for accurately determining genetic diversity.

Additionally, not sequencing the entire mitochondrial genome can cause issues with results when using different genes. In a paper focusing on mountain lions from Central and South America, Caragiulo et al.[23] used the 16s rRNA, Cytochrome b, 12s rRNA and ATPase-6 gene regions while Culver et al. used the 16s rRNA, ATPase-8 and NAD5 genes and obtained different results as to the phylogeographic groupings for North and South American mountain lions. The former found only four major groups while the latter found six major groups across North and South America. Caragiulo et al. concluded this was likely due to lack of sufficient sampling for geographic regions, however, it could also be partly due to the different regions selected for sequencing. Some regions of the genome are more diverse than others, as indicated by another study focusing more on Central and South American cougars, which sequenced a larger section of the NADH5 region as this one was found to have the most diversity in the Culver et al. 2000 study[3]. Without standards as to which regions of the mitochondrial genome to use, it is difficult to know which diversity estimate is accurate when different regions are analyzed.

This also causes a problem when combining data sets for further analyses. At present, it is difficult to obtain a true representation of the genetic diversity of mountain lions with the current limited data as only the three aforementioned papers have analyzed the diversity of mountain lions in North and South America. Of these three papers, there is no common gene found between all three and none of them have more than one gene region in common, making it difficult to perform meta-analyses and combine data sets to improve sample numbers without losing important sequence information. The main reason for such an issue is that the cost of sequencing the entire mitochondrial genome of numerous samples was too high.

However, massively parallel sequencing, as used in this study, has significantly decreased the cost of sequencing the complete mitochondrial genome – and even eventually the nuclear genome. Therefore, researchers should make an effort to sequence the entire mitochondrial genome in future analyses. This would further enhance our current knowledge and allow others to combine available data more easily without having to choose between including more loci or more samples for their analyses. With the diversity of North American mountain lions as low as it is, not analyzing the entire mitochondrial genome can cause significant errors when calculating diversity.

Furthermore, no one has completely sequenced the nuclear genome of *Puma concolor* as of yet. As a follow-up to this project, the five successfully sequenced samples will be further analyzed using nuclear genome sequencing. The samples with the highest average depth coverage (particularly BE813 and RSL813) will likely yield the best results in terms of sequencing quality. While useful and cost-effective, mitochondrial DNA can only provide so much information about the genetic diversity and potential conservation risk factors. Obtaining complete nuclear genome sequences from multiple mountain lion individuals would be incredibly useful for future conservation efforts and could potentially provide genetic data about how mountain lions have adapted to their various habitats. Furthermore, the determination of several nuclear genomes would allow for analyses of nuclear genome diversity and population level genetic data.

BIBLIOGRAPHY

1. Barnett, R. *et al.* Evolution of the extinct Sabretooths and the American cheetah-like cat. *Curr. Biol.* **15,** R589–90 (2005).
2. Johnson, W. E. *et al.* The Late Miocene Radiation of Modern Felidae: A Genetic Assessment. *Science (80-.).* **311,** 73–77 (2006).
3. Matte, E. M. *et al.* Molecular evidence for a recent demographic expansion in the puma (Puma concolor) (Mammalia, Felidae). *Genet. Mol. Biol.* **36,** 586–97 (2013).
4. Culver, M., Johnson, W. E., Pecon-Slattery, J. & O'Brien, S. J. Genomic ancestry of the American puma (Puma concolor). *J. Hered.* **91,** 186–97 (2000).
5. Hall, E. R. *The mammals of North America*. (John Wiley and Sons., 1981).
6. Mccollough, M., Miller, M. & Racey, M. *U.S. Fish and Wildlife Service Concludes Eastern Cougar Extinct*. (2011). at <http://www.fws.gov/northeast/ecougar/newsreleasefinal.html>
7. Hansen, K. & Redford, R. *Cougar: The American Lion*. (Northland publishing, 1992).
8. Dybas, C. L. Phantom of the Forest: Could the Cougar Again Haunt Eastern U.S. Woodlands? *National Geographic* (2015). at <http://voices.nationalgeographic.com/2015/11/04/phantom-of-the-forest-could-the-cougar-again-haunt-eastern-u-s-woodlands/?utm_source=Twitter&utm_medium=Social&utm_content=link_tw20151105voices-cougar&utm_campaign=Content&sf14928954=1>
9. Mcginnis, H. Anson Panther Re-discovered: Donated to State Museum of Pennsylvania in

Harrisburg. *Eastern Cougar Foundation Newsletter* 12 (2005). at <http://www.easterncougar.org/newltr_pdf/ecfnew_autumn05w.pdf>

10. Johnson, W. E. *et al.* Genetic restoration of the Florida panther. *Science* **329,** 1641–1645 (2010).
11. Ernest, H. B. *et al.* Genetic structure of mountain lion (Puma concolor) populations in California. *Conserv. Genet.* **4,** 353–366 (2003).
12. Alberts, B. *et al. Molecular Biology of the Cell*. *Igarss 2014* (Garland Science, 2008). doi:10.1007/s13398-014-0173-7.2
13. Shapiro, B. & Hofreiter, M. *Ancient DNA: Methods and Protocols*. *Life Sciences* **531,** (Humana Press, 2012).
14. Casas-Marce, M., Revilla, E. & Godoy, J. A. Searching for DNA in museum specimens: A comparison of sources in a mammal species. *Mol. Ecol. Resour.* **10,** 502–507 (2010).
15. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 15758–63 (2013).
16. de Moraes-Barros, N. & Morgante, J. S. A simple protocol for the extraction and sequence analysis of DNA from study skin of museum collections. *Genet. Mol. Biol.* **30,** 1181–1185 (2007).
17. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5448 (2010).
18. Kircher, M. Analysis of high-throughput ancient DNA sequencing data. *Methods Mol. Biol.* **840,** 197–228 (2012).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler

transform. *Bioinformatics* **25,** 1754–1760 (2009).

20. McCulloch, D. L. Puma concolor mitochondrion, complete genome. *Genbank* (2011). at <http://www.ncbi.nlm.nih.gov/nuccore/364284064>

21. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Pääbo, S. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. *PLoS One* **7,** e34131 (2012).

22. Gnirke, A. *et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27,** 182–189 (2009).

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

24. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).

25. Hasegawa M., K. H. & T., Y. Dating the human-ape split by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22,** 160–174 (1985).

26. Kumar S., S. G. & K., T. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **(submitted,** (2015).

27. *Species Profile for Florida panther (Puma (=felis) concolor coryi)*. *Environmental Conservation Online System* at <http://ecos.fws.gov/tess_public/profile/speciesProfile?spcode=A008>

ACADEMIC VITA

Maya Evanitsky
mevanitsky@gmail.com

Education

Pennsylvania State University, University Park, PA
Degree: B.S. in Biochemistry and Molecular Biology
Minors: Anthropology, Microbiology

Honors: Biochemistry and Molecular Biology

Thesis Title: Sequencing the Genome of the Nittany Lion (*Puma Concolor* from Pennsylvania)

Thesis Supervisor: George Perry, Santosh Giririjan

Work Experience

Carpenter Building, Penn State University, University Park, PA
Research Assistant Fall 2013- Spring 2016

- Extract DNA from various tissues using techniques for modern and ancient DNA
- Prepare DNA sequence libraries using the Meyer & Kircher protocol
- Analyze DNA sequence data using bioinformatics tools

NIDDK, National Institutes of Health, Bethesda, MD
Laboratory Intern Summer 2014

- Conducted a research project analyzing the effect of MSH3 mutations on repeat expansions in Fragile X premutation mouse models
- Extracted DNA from mouse tails and organs
- Analyzed DNA using PCR and gel electrophoresis
- Assisted in mouse dissection
- Analyzed DNA allele size using statistical models

Althouse Lab, Penn State University, University Park, PA
Work Study Assistant Fall 2012-Spring 2014

- Cleaned and sterilized lab equipment
- Inventoried chemical waste and lab equipment
- Assisted with research projects in the lab by counting megakaryocytes in bone slices of mouse models

Grants Received

Schreyer Honors College Thesis Research Grant Summer 2015

Awards

Schreyer Honors College Scholarship 2012-2016
Dean's List 2012 - Spring 2013, 2014 - 2015

Publications

Zhao, X., Kumari, D., Gupta, S., Wu, D., Evanitsky, M., Yang, W., Usdin, K. (2015). MutSβ generates both expansions and contractions in a mouse model of the Fragile X-associated disorders. *Human Molecular Genetics*. 24(24), 7087-96

Presentations

Evanitsky, M. (November 2015). Sequencing the Nittany Lion Genome. Westinghouse Science Honors Institute, State College, PA

Evanitsky, M., Zhao, X., Usdin, K. (August 2015). The Effects of Msh3 Mutations on Repeat Expansion in a Fragile X Premutation Mouse Model. Poster session presented at NIH Annual Poster Presentation, Bethesda, MD

Computational Languages and Programs

R Statistical Package	Arlequin
Burrows-Wheeler Aligner	MEGA
MAFFT	Samtools

Extracurricular Activities

Active member of Penn State Forensic Science Club
Active member of Penn State Springfield
Active member of Penn State LGBTA Student Coalition
Secretary of Penn State Club Yoga