CHEETAH: Optimized Breeding Programs For Combatting Genetic Bottlenecks

Madison Davis, Avery Park, Sarah Radway

Abstract

Genetic drift poses a dire threat to endangered species like African cheetahs, by increasing the impact of harmful alleles, and thus species susceptibility to disease and extinction. However, it is unclear how stakeholders can increase likelihood of species survival, through effectively overcoming these genetic bottlenecks. In this work, we propose CHEETAH (Cheetah Health Enhancement and Expansion Through Adaptive Heterogenizing), a system which aims to leverage machine learning techniques to provide optimal breeding pairings to increase genetic diversity. Using a dataset of cheetah nucleotides, obtained through a simple blood test, we implement a divisive hierarchical clustering mechanism, allowing us to identify pairings that will most increase genetic diversity. We simulate several rounds of breeding, and ultimately find that CHEETAH is able to successfully increase genetic variations, in contrast with the benchmark case. This will allow for zoos and other breeding conservation programs to increase the likelihood of species survival, for cheetahs and other endangered species.

Introduction

It is clear that genetic bottlenecks pose a dire threat to the survival of endangered species. Genetic bottlenecks are dramatic decreases in the population size of a species, brought on by events such as disease, habitat degradation, or extensive human hunting (Li and Roossinck 2004; Nature 2014). While these bottlenecks can cause harms in many ways, one notable feature is that they may lead to genetic drift. Genetic drift is when changes in the frequency of alleles occur within a population (Andrews 2010). This is harmful, as reduced diversity of alleles leaves species susceptible to extinction by disease, genetic disorders, or environmental influences.

In this paper we focus on one notable case where genetic drift is leading to threats of species extinction: the African cheetah. At the start of the 20th century, there were over 100,000 cheetahs; today there are under 7,000 in the wild (Durant 2023). DNA analysis showcases how cheetahs only retain 0.1-4% of genetic variation shown in most living species (O'Brien et al. 2017). This lack of genetic diversity has begun to impact cheetah health and survival capability: namely, the cheetahs' ability to reproduce (Cohn 1986), high infant mortality (O'Brien, Wildt, and Bush 1986), and poor organ functionality (O'Brien, Wildt, and Bush 1986).

However, genetic bottlenecks are challenging to combat: the less genetically diverse a species becomes, the more limited the potential allele possibilities, and the more infrequent these allele variations become. While one cheetah may have a more unique allele for one gene, it may negatively impact genetic diversity for other genes. It is challenging to ensure increases in the diversity of one allele, without being at the expense of another allele. Understanding what pairings will contribute to the greatest genetic diversity for the species is a difficult problem.

Previous work has taken steps to understand the challenges surrounding genetic diversity in cheetahs, and other genetically-limited species. Namely, works have attempted to use machine learning to predict parental or subspecies lineage (Gutenkunst et al. 2009; Yang et al. 1997) or classify subspecies (Burócziová and Říha 2009). Notably, previous researchers have also collected and evaluated cheetah genetic data, to understand the cause of the genetic bottleneck and attempt to identify potentially harmful genes (Dobrynin et al. 2015). However, we are the first that we know of to investigate a forward-looking mechanism for preventing the cheetah's harmful genetic bottlenecks.

Therefore, we propose CHEETAH (Cheetah Health Enhancement and Expansion Through Adaptive Heterogenizing)", a system which aims to leverage machine learning techniques to provide optimal breeding pairings to increase genetic diversity. CHEETAH leverages machine learning techniques to provide optimal breeding pairings to increase genetic diversity. Using a dataset of cheetah nucleotides, obtained through a simple blood test, we implement a divisive hierarchical clustering mechanism, allowing us to identify pairings that will most increase genetic diversity. We simulate several rounds of breeding, and ultimately find that CHEETAH is able to successfully increase genetic variations, in contrast with the benchmark case. CHEETAH will allow for zoos and other breeding conservation programs to increase the likelihood of species survival, for cheetahs and other endangered species.

Background

Genetic Vocabulary Cheetahs have 19 chromosomal pairs for a total of 38 chromosomes. Each chromosome

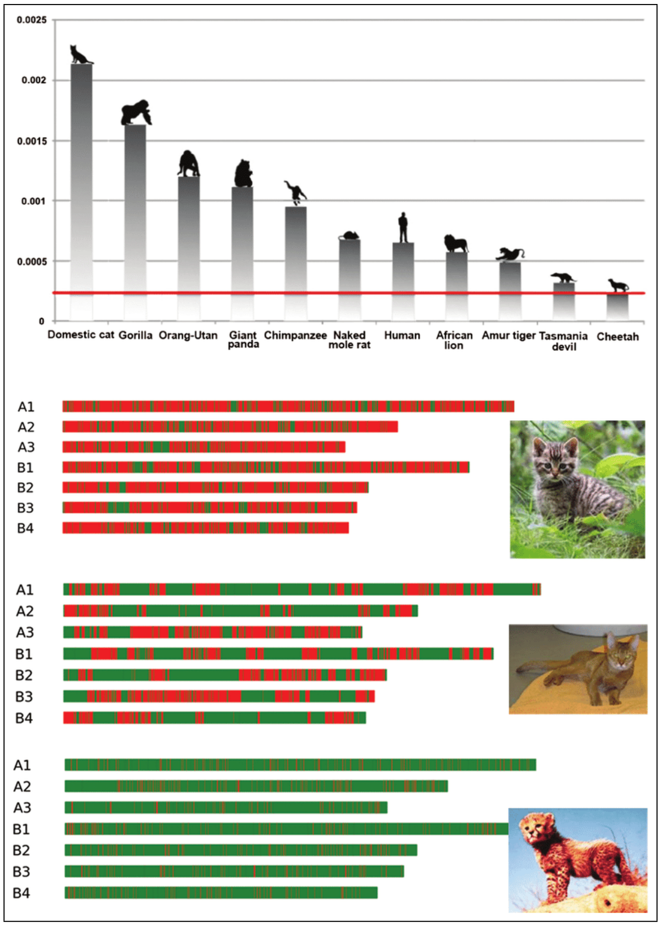


Figure 1: Visual Motivation of Genetic Bottleneck (OBrien et al. 2017)

is a large sequence of genes. Each gene is a sequence of nucleotides, which are organic molecules composed of a nitrogenous base. There are four different kinds of nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C). Adenine pairs with thymine and guanine with cytosine. For a specific gene, the nucleotide sequence can vary ever-so slightly by altering specific nucleotides between their respective pairings. When a specific nucleotide changes compared to the general structure of the gene, this is called a single-nucleotide polymorphism (SNP), and the index or location at which one may look at a specific nuceltoide is called a loci. A specific variation of the gene pattern (due to one or more SNPs) is called an allele, and different genes can have a number of different alleles, some more dominant in the population than others. Alleles are often represented by letter-encodings rather than the entire nucleotide sequence, and they are often upper-cased if they are more dominant. Different genes will show up on different chromosomal pairs. For a specific chromosomal pair, a gene is represented on each chromosome, one copy (or allele) from the mother and one from the father. During breeding, for a specific gene, the two alleles are brought into contact with the alleles from the other individual, where the combination is often represented through a Punnett-square.

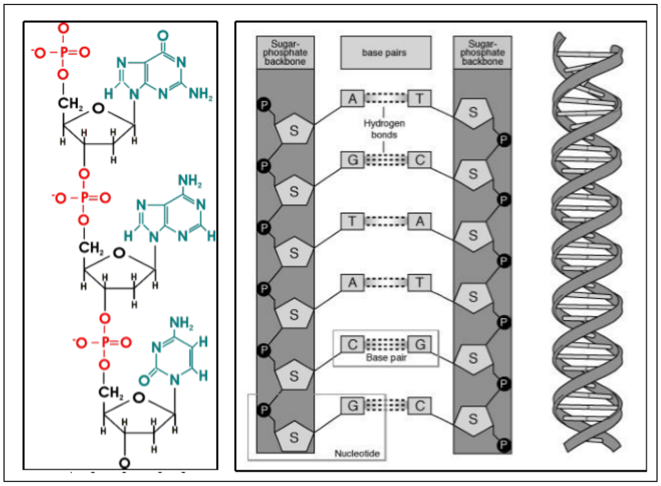


Figure 2: Visual Representation of DNA Sequence (Mac 2023)

Defining genetic diversity Alleles cease to exist when no population member possesses an allele with a certain combination of these nucleotides. Alleles can be introduced through mutations in nucleotides, when a new combination of nucleotides is created. However, when we talk about genetic drift, we are referring to the changes in allele combinations for a given gene due to dramatic decreases in population size. This can lead to the number and distribution of alleles changing greatly, and potentially decreasing in diversity. Our work aims to combat this feature of genetic drift.

Measuring genetic diversity Genetic diversity thus represents the frequency and variations of alleles. In practice, when researchers discuss measuring genetic diversity, they mean measuring the allele variations that are present for each gene in the given genome across the species.

Combating genetic bottlenecks Therefore, when we are seeking to combat genetic bottlenecks, we are aiming to increase genetic diversity, through both (1) increasing the variety of alleles available for a given gene, and (2) diversifying the frequencies of alleles available for a given gene.

Related Work

Previous work has used machine learning to predicted species lineage. In Gutenkunst et al., the authors use a diffusion-based method to produce demographic models using genetic variants to understand population dynamics and evolutionary forces such as migration (Gutenkunst et al. 2009). Yang et al. similarly utilizes maximum likelihood approach to genetic and demographic modeling, presenting PAML, a tool for proposing phylogenetic trees and subsequent evolutionary models (Yang 2007). These works are not forward facing, but seek to reconstruct ***historical*** genetic evolution.

Previous work has used machine learning to classify subspecies. In a similar line of work, Brocziova et al., use various machine learning-based classification methods to perform breed discrimination between different breeds of horses. The authors used 17 microsatellite markers for parentage testing to probabilistically predict an horses's breed. (Burócziová and Říha 2009). These works are not forward facing, but seek to understand ***current*** differentiations between breeds.

Previous work has investigated the cause of the genetic bottleneck, and identified harmful genes. In several works, including Dobrynin et al, Prost et al., and O'Brien et al., the authors assess the genetic diversity of the cheetah genome, using both historical and modern cheetah genetic data (Dobrynin et al. 2015; Prost et al. 2020; O'Brien et al. 2017). Their findings emphasize the importance of promoting genetic diversity, as overcoming genetic bottlenecks isn't enough to fight "against rare recessive genetic abnormalities as well as a hedge against deadly infectious agents" (O'Brien, Wildt, and Bush 1986). These works are not forward facing, but seek to understand how ***past challenges*** contribute to a lack of genetic diversity in the cheetah.

Previous work has investigated crop diversity In other related works, the authors have investigated how to increase the genetic diversity of crops (Allier et al. 2020; Akdemir et al. 2019; Jiang et al. 2020; Khan, Dar, and Dar 2015), such as corn and wheat. The authors propose methods for increasing desired ideotypes that are correlated with better crop outputs. However, these works are focused on optimizing for ***specific traits***, not necessarily for crop survival.

Previous work has investigated cattle Lastly, the authors have investigated the use of genetic testing to prevent inbreeding for cattle (Meuwissen 1998; Day and Grum 2005; Fleming et al. 2018). These works are not intended to reduce the effects of existing genetic bottlenecks, but to ***maintain existing genetic diversity***.

Previous works are not intended to resolve existing genetic bottlenecks, but rather to understand historical trends that led to current bottlenecks, to increase crop or livestock output, or to prevent future harms. To the best of our knowledge, we are the first work to investigate a forward-looking machine learning mechanism for preventing future harmful genetic drift in animals currently facing genetic bottlenecks, such as cheetahs.

Dataset

Our dataset was inspired off of the Cheetah Project 1 dataset, as archived in the National Library of Medicine under accession number PRJNA624893 (Senckenberg 2020). This archive incorporates genomic DNA, ddRAD, mtDNA, and MHC data across a total of 56 specimens from five cheetah subspecies. We focused specifically on the ddRAD results, which came from 55 separate sequence-read archives (SRA). SRA is a repository maintained by the National Center for Biotechnology Information (NCBI) that holds raw sequencing data (in our case, the sequencing data is the sequences of nucleotides that make up the genetic strand of interest). ddRAD, or double digest Restriction Associated DNA, is a kind of sequencing data, though reduced in its complexity. See Figure 3 for a more in-depth visual concerning the process of ddRAD, where arrows indicate restrictions or reductions in complexity. This allows models to more easily analyze the genetic material and potentially spot differences between sequenced data. Any single-nucleotide difference found between two genetic strands (sequenced data) would be labeled as an SNP.

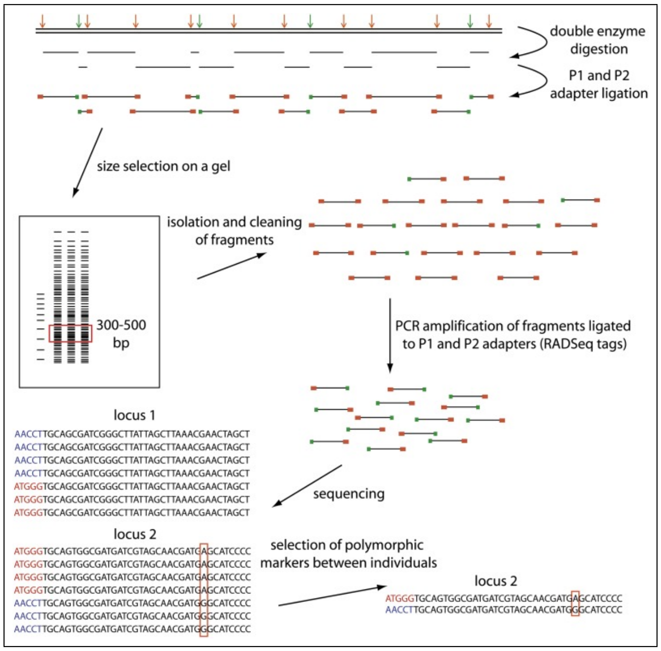
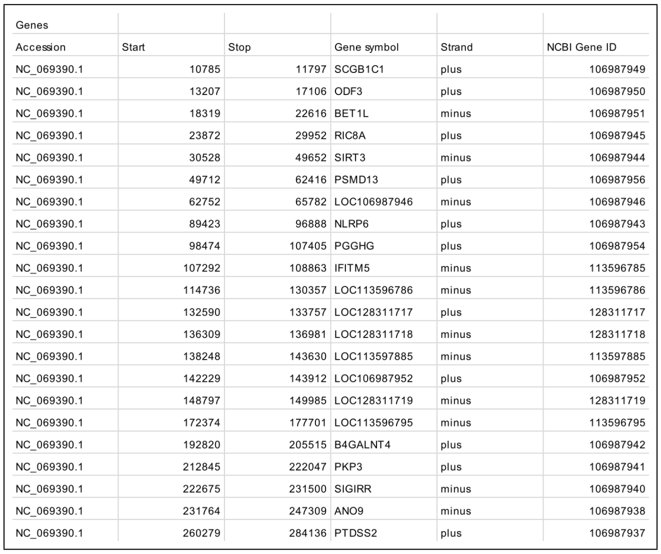

double enzyme digestion
P1 and P2 adapter ligation
size selection on a gel
isolation and cleaning of fragments
300-500 bp
PCR amplification of fragments ligated to P1 and P2 adapters (RADSeq tags)
sequencing
locus 1
AACCTTGCAGCGATCGGGCTTATTAGCTTAAACGAACTAGCT
AACCTTGCAGCGATCGGGCTTATTAGCTTAAACGAACTAGCT
AACCTTGCAGCGATCGGGCTTATTAGCTTAAACGAACTAGCT
AACCTTGCAGCGATCGGGCTTATTAGCTTAAACGAACTAGCT
ATGGGTGCAGCGATCGGGCTTATTAGCTTAAACGAACTAGCT
ATGGGTGCAGCGATCGGGCTTATTAGCTTAAACGAACTAGCT
ATGGGTGCAGCGATCGGGCTTATTAGCTTAAACGAACTAGCT
locus 2
ATGGGTGCAGTGGCGATGATCGTAGCAACGATGAGCATCCCC
ATGGGTGCAGTGGCGATGATCGTAGCAACGATGAGCATCCCC
ATGGGTGCAGTGGCGATGATCGTAGCAACGATGAGCATCCCC
ATGGGTGCAGTGGCGATGATCGTAGCAACGATGAGCATCCCC
AACCTTGCAGTGGCGATGATCGTAGCAACGATGGGCATCCCC
AACCTTGCAGTGGCGATGATCGTAGCAACGATGGGCATCCCC
AACCTTGCAGTGGCGATGATCGTAGCAACGATGGGCATCCCC
selection of polymorphic markers between individuals
locus 2
ATGGGTGCAGTGGCGATGATCGTAGCAACGATGAGCATCCCC
AACCTTGCAGTGGCGATGATCGTAGCAACGATGGGCATCCCC

Figure 3: Visual Representation ddRAD-Seq Model (Recknagel, R, and Elmer 2013)

Our ultimate goal is to locate a strand of genetic material to experiment with in breeding by altering its SNPs through successive generations. To do this, we had to first choose an individual and a part of its genome to analyze. From the project, we chose BioSample SAMN29220483, which incorporated tissue material from an adult wild Cheetah originating from South Sudan (Vetmeduni 2022). The tissue sample was already genetically analyzed into sequence data (ddRAD reduced representation) and was uploaded to the NIH SRA. We copied the first 4,000 SRA reads that contained the ddRAD result, which amounted to 887,888 nucleotides. We did not copy more mainly due to the length constraints in comparing a genetic string with a reference genome. The reference genome is considered to be a healthy cheetah and had its genetic material likewise already sequenced and uploaded to the NIH site (of Veterinary Medicine 2022). Once we had extracted a relevant section of genetic material from the BioSample, we BLASTed it with the reference genome. BLAST is a model that analyzes a unique genetic sequence and identifies where in a reference genome that sequence is most likely to have originated from. We specifically chose the BLAST parameters to optimize for somewhat dissimilar sequences, as the model allows for searches where it can optimize for sequence-comparisons that are highly dissimilar, somewhat dissimilar, or very similar. We chose this middle-ground route because very similar strands have less SNPs and are subsequently less likely to be optimized in our experiments. On the other extreme, we did not decide to look for highly dissimilar sequences, as we wanted to get an average-baseline for the number of SNPs within a given region and see on average how well the variance would improve in the breeding program.

BLAST Result After the BLAST, we received 100 sequences (or large strands of genetic material, where each sequence is based on a different chromosome or scaffolded region) that the model identified were somewhat dissimilar with the reference genome. We focused on the sequence that analyzed genetic material in the D1 chromosome. This is because the D1 chromosome is particularly useful and known to be a regular baseline for cheetah population health. Many of its genes focus on immune system vitality, nervous system function, and general disease resistance of the major histocompatibility complex (MHC), which the cheetah has been especially vulnerable to in recent years (Aines, Bettina, and Simone 2010). A select sample of the genes identified in the cross-comparison between the reference genome and the sample are shown in Figure 4. To focus on a few:

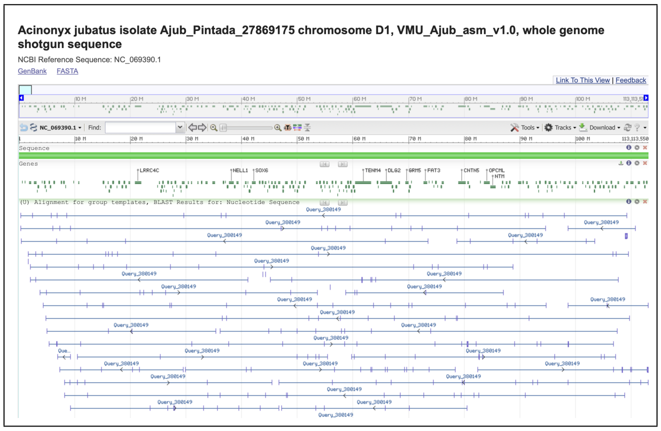
1. NELL-1/Nell-1: this gene promotes orthotopic bone regeneration (X et al. 2010)
2. SOX6: this gene acts as a tumor-suppressor (NCBI 2024)
3. LRRC4C: this gene promotes the growth of thalamic neurons in the central nervous system (of Science 2024)



Genes					
Accession	Start	Stop	Gene symbol	Strand	NCBI Gene ID
NC_069390.1	10785	11797	SCGB1C1	plus	106987949
NC_069390.1	13207	17106	ODF3	plus	106987950
NC_069390.1	18319	22616	BET1L	minus	106987951
NC_069390.1	23872	29952	RIC8A	plus	106987945
NC_069390.1	30528	49652	SIRT3	minus	106987944
NC_069390.1	49712	62416	PSMD13	plus	106987956
NC_069390.1	62752	65782	LOC106987946	minus	106987946
NC_069390.1	89423	96888	NLRP6	plus	106987943
NC_069390.1	98474	107405	PGGHG	plus	106987954
NC_069390.1	107292	108863	IFITM5	minus	113596785
NC_069390.1	114736	130357	LOC113596786	minus	113596786
NC_069390.1	132590	133757	LOC128311717	plus	128311717
NC_069390.1	136309	136981	LOC128311718	minus	128311718
NC_069390.1	138248	143630	LOC113597885	minus	113597885
NC_069390.1	142229	143912	LOC106987952	plus	106987952
NC_069390.1	148797	149985	LOC128311719	minus	128311719
NC_069390.1	172374	177701	LOC113596795	minus	113596795
NC_069390.1	192820	205515	B4GALNT4	plus	106987942
NC_069390.1	212845	222047	PKP3	plus	106987941
NC_069390.1	222675	231500	SIGIRR	minus	106987940
NC_069390.1	231764	247309	ANO9	minus	106987938
NC_069390.1	260279	284136	PTDSS2	plus	106987937

Figure 4: Select Sample of Genes Identified in BLAST

The D1 sequence-comparison was done by a series of queries, where each query focused on smaller comparisons between genetic material. The queries can be visually seen in Figure 5. Each blue line indicates a single query. Each query is aligned in parallel with where it is predicted to have matched with the reference genome, as shown in bold green. The smaller strips of green lines underneath the bold green line indicate genes that have been identified and tagged by research for the cheetah species.



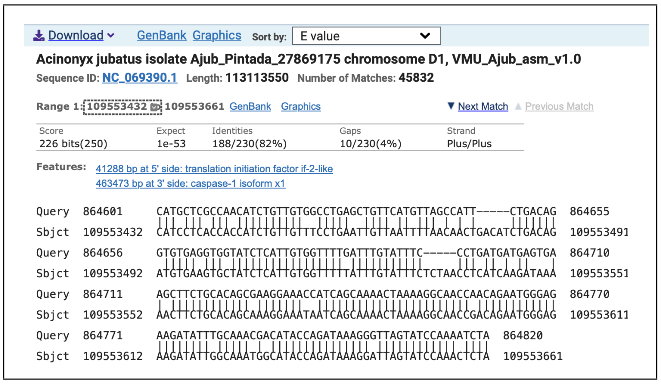
Acinonyx jubatus isolate Ajub_Pintada_27869175 chromosome D1, VMU_Ajub_asm_v1.0, whole genome shotgun sequence

NCBI Reference Sequence: NC_069390.1

GenBank FASTA

Figure 5: D1 Full Comparison with BLAST Sequence Read

Each query sample involved finding gaps between the genetic sequence and the reference, which subsequently gave way to identifying SNPs. Figure 6 demonstrates a sample alignment where the query (a sample from our genetic sequence) and subject (the reference genome) were compared per nucleotide and where the SNPs were found.



Acinonyx jubatus isolate Ajub_Pintada_27869175 chromosome D1, VMU_Ajub_asm_v1.0

Sequence ID: NC_069390.1 Length: 113113550 Number of Matches: 45832

Range 1: 109553432 to 109553661 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
226 bits(250)	1e-53	188/230(82%)	10/230(4%)	Plus/Plus

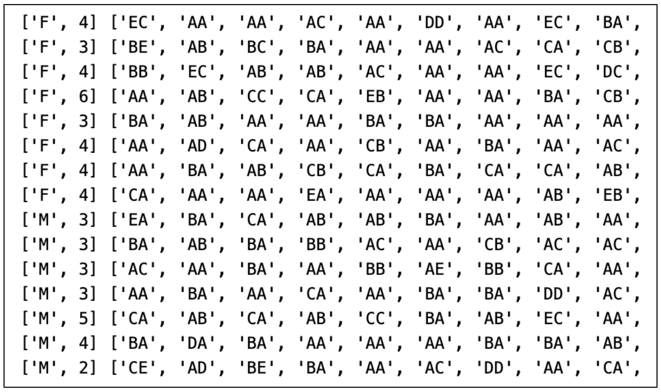
Features: 41288 bp at 5' side: translation initiation factor if-2-like
463473 bp at 3' side: caspase-1 isoform x1

```
Query  864601     CATGCTCGCCAACATCTGTTGTGGCCTGAGCTGTTCATGTTAGCCATT-----CTGACAG  864655
Sbjct  109553432  CATCCTCACCACCATCTGTTGTTTCCTGAATTGTTAATTTTAACAACTGACATCTGACAG  109553491
Query  864656     GTGTGAGGTGGTATCTCATTGTGGTTTTGATTTGTATTTC-----CCTGATGATGAGTGA  864710
Sbjct  109553492  ATGTGAAGTGCTATCTCATTGTGGTTTTTATTTGTATTTCTCTAACCTCATCAAGATAAA  109553551
Query  864711     AGCTTCTGCACAGCGAAGGAAACCATCAGCAAAACTAAAAGGCAACCAACAGAATGGGAG  864770
Sbjct  109553552  AACTTCTGCACAGCAAAGGAAATAATCAGCAAAACTAAAAGGCAACCGACAGAATGGGAG  109553611
Query  864771     AAGATATTTGCAAACGACATACCAGATAAAGGGTTAGTATCCAAAATCTA  864820
Sbjct  109553612  AAGATATTGGCAAATGGCATACCAGATAAAGGATTAGTATCCAAACTCTA  109553661
```

Figure 6: Sample Alignment

Making the First Generation Once this BLAST result was generated, we needed to then create our first foundational generation of cheetahs for the breeding program. We first downloaded the BLAST result as a txt file. We then parsed the input so as to extract the query and subject nucleotides separately and then placed this information into two separate arrays, one for our cheetah sample and the other for our reference genome. We later identified where in these arrays the sequences differed per nucleotide. We had two variations of differences: one where the sample individual had a nucleotide different than the reference genome, and one where the sample individual had omitted nucleotides, indicated as a dash. In the second case, due to lack of information, we omitted those nucleotides and did not consider them to be plausible sites for SNPs. After extracting all of the indexes, we cut down on the size of the arrays further for computational complexity. In the end, we had 15,027 SNPs (or indexes) to examine. For the remainder of this experiment, we focused specifically on these 15,027 indexes. For our first generation, we assumed that these SNPs were broken up into 100 genes of roughly equal length. Therefore, each gene would have approximately 150 SNPs to consider. We looked at the SNPs that fell within the range of each gene and created allele letter-encodings for variations of those SNPs. For our baseline testing, we made 5 different variations of these SNP combinations, or alleles, by which we would construct the genetic pairings of future generations. Therefore, there were 25 plausible 2-allele combinations per gene for each cheetah individual we generated. We then decided that 80 percent of the genes would model a skewed distribution such that the majority of the alleles would not necessarily be well-represented and only 1 to 2 would be more prevalent. The other 20 percent of the genes would model a lesser skewed distribution such that the spread of the alleles was more even.

For each cheetah, their data would be represented as an array holding both metadata and the generated alleles for each of the 100 genes, based on their aforementioned skewed distributions. For the program, we decided to have 50 initial members, 25 female and 25 male. Because cheetahs can only breed within a certain window of their life, we also incorporated into the metadata a parameter for age. The beginning baseline was assumed to be on the younger side, and so their ages ranged uniformly from 2 to 6. An example output for the starting generation is shown in Figure 7, where the number of genes were truncated for legibility.



```
['F', 4] ['EC', 'AA', 'AA', 'AC', 'AA', 'DD', 'AA', 'EC', 'BA',
['F', 3] ['BE', 'AB', 'BC', 'BA', 'AA', 'AA', 'AC', 'CA', 'CB',
['F', 4] ['BB', 'EC', 'AB', 'AB', 'AC', 'AA', 'AA', 'EC', 'DC',
['F', 6] ['AA', 'AB', 'CC', 'CA', 'EB', 'AA', 'AA', 'BA', 'CB',
['F', 3] ['BA', 'AB', 'AA', 'AA', 'BA', 'BA', 'AA', 'AA', 'AA',
['F', 4] ['AA', 'AD', 'CA', 'AA', 'CB', 'AA', 'BA', 'AA', 'AC',
['F', 4] ['AA', 'BA', 'AB', 'CB', 'CA', 'BA', 'CA', 'CA', 'AB',
['F', 4] ['CA', 'AA', 'AA', 'EA', 'AA', 'AA', 'AA', 'AB', 'EB',
['M', 3] ['EA', 'BA', 'CA', 'AB', 'AB', 'BA', 'AA', 'AB', 'AA',
['M', 3] ['BA', 'AB', 'BA', 'BB', 'AC', 'AA', 'CB', 'AC', 'AC',
['M', 3] ['AC', 'AA', 'BA', 'AA', 'BB', 'AE', 'BB', 'CA', 'AA',
['M', 3] ['AA', 'BA', 'AA', 'CA', 'AA', 'BA', 'BA', 'DD', 'AC',
['M', 5] ['CA', 'AB', 'CA', 'AB', 'CC', 'BA', 'AB', 'EC', 'AA',
['M', 4] ['BA', 'DA', 'BA', 'AA', 'AA', 'AA', 'BA', 'BA', 'AB',
['M', 2] ['CE', 'AD', 'BE', 'BA', 'AA', 'AC', 'DD', 'AA', 'CA',
```

Figure 7: Partition of the First Cheetah Generation

Methods

The following sections detail the methodology of the project, from feature selection to clustering and breeding [1]

Feature Selection

Because we will undergo several rounds of clustering over the generations, we sought to focus our efforts each round on genes that exhibit low genetic diversity. For example, say a gene had allele combinations AA, AB, and BB. When the simulations begin, variance within the gene might be low. For example, we may reference Figure 8. For Gene 1, AA, AB, and BB may have 33% of the population each. For another gene, with the same allele combinations,AA makes up 80% of the population, and AB and BB make up 10% each respectively. Our algorithm focuses on genes that resemble the second example gene (Gene 2).

Therefore, in order to select features, we evaluate the "variance of frequencies" of all alleles, representing the extent to which the individual frequencies in a frequency distribution deviate from the mean frequency. We then order alleles by their genetic diversity under this metric, and for each given round of clustering, we repeat this evaluation of diversity across genes. We assume that each gene can have five potential alleles. This allows us to focus on increasing genetic diversity where it is most needed.

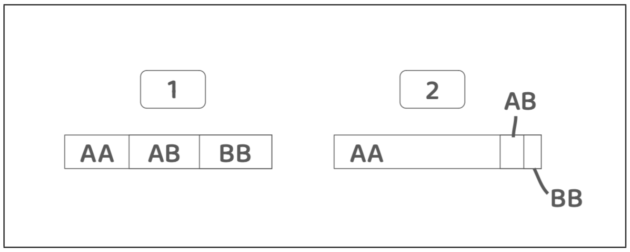


Figure 8: The frequency of three different alleles for two genes (1 and 2). Gene 1 is of high variance, Gene 2 is of low variance.

Clustering

For the clustering method, we start with the alleles from feature selection. Our main approach is to use divisive hierarchical clustering, which has been highly used in genetic data given its underlying hierarchical structure (Liang et al. 2018). A graphical output of divisive hierarchical clustering is the dendrogram, where future iterations of the cluster model are shown by splits from the original large cluster. This can be more clearly seen in Figure 9. We use divisive clustering given our goal of increasing genetic diversity, as we start with all individuals in the initial cluster and ideally ending with as many clusters as possible, where each cluster represents individuals that have similar genetic composition. A key difference in our approach is that we do not formally use an existing divisive clustering algorithm, like Divisive ANAlysis Clustering (DIANA) that starts with all observations in a cluster and then chooses the cluster to split based on some criterion (Kaufman and Rousseeuw 1990). The rationale behind not using DIANA was for two reasons:

1. We are not looking for a cluster to split at each step as this would likely split the most diverse cluster to reduce its size, but this would not necessarily give us the optimal parents to breed.
2. The divisive process tends to focus more on the structure within the cluster and not the relationships between the

[1] All code is publicly available at this Github repository

clusters. For this project, we want to focus more on how different parents could be paired across clusters.

Instead, we choose to simulate divisive clustering by using an agglomerative clustering algorithm, which we explain in the next section. Lastly, we note that we use the alleles from feature selection to do our clustering, but when we do breeding, we use all of the parents' alleles to generate the offspring.

For the algorithm itself, we start by finding the pairwise distance matrix across alleles, where the distance between two specimens would be the number of alleles that differ, representing genetic dissimilarity. We note that our approach generalizes the distance between every pairing of alleles to be 1, regardless of which alleles we are using, whereas more sophisticated measures might consider allele frequencies.

To run our version of divisive hierarchical clustering, we start with the entire dataset as one cluster, which is implicit in our use of the agglomerative algorithm as we do not have labels yet. We then use the `linkage` and `fcluster` functions from the scipy.cluster.hierarchy library (Virtanen et al. 2020). We input the custom distance matrix from the first step and complete linkage (maximum distance) into the `linkage` function to then get our linkage matrix. We chose complete linkage so that we would create more homogenous clusters by focusing on the most dissimilar specimens and being less influenced by noise and outliers when compared to other linkage criterion. The `fcluster` function then extracts the cluster labels from the linkage matrix, where we define the maximum number of clusters based on the expected number of specimens in the program. This number is calculated as the total number of specimens plus the expected number of offspring minus the expected number of specimens that phase out of the program. On each iteration, we increase the number of clusters that are passed into `fcluster` since we are increasing the offspring, which forces the creation of new groups from the existing clusters and allows us to mimic the top-down approach of a divisive algorithm.

Before passing our cluster labels to the breeding algorithm, we also must ensure that the specimens we want to pair are of an age in which they can breed. We add two years to each of the specimen's age value, since we set each iteration to be every two years, and then remove any specimens above the age of ten (cheetah's average life expectancy).

Breeding

The goal of the breeding component of our methodology is to find optimal parents to pair that are as genetically different from each other as possible, so that we can increase the genetic diversity of the next generation via the offspring.

Using the distance matrix, we iterate through all possible combinations of female and male pairings. For each possible pair, we have the following criteria required for breeding.

- The parents must be from different clusters. This allows us to incorporate our clustering algorithm and focus on the alleles we want to improve on.
- The parents have to be of opposite gender. We account for this by storing a 'metadata' array for each specimen which has their gender and age.
- The distance between the pairing of the parents must exceed our current maximum distance found. The maximum distance metric is increased to be the value of the largest distance of all the previous pairings, so that we ensure each additional pairing would be beneficial.

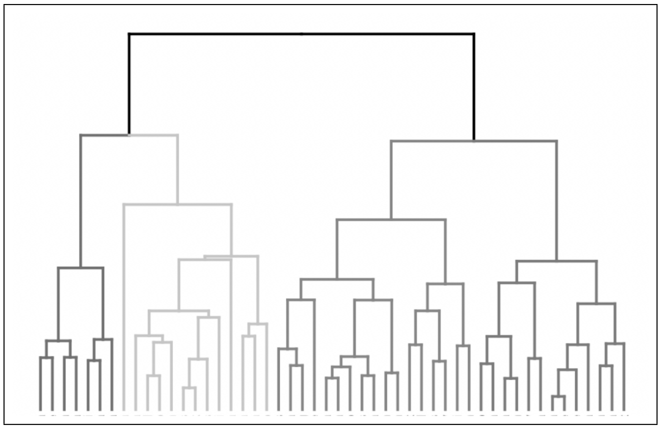


Figure 9: Example of Dendrogram Output Using Divisive Hierarchical Clustering: Successive Iterations Produce More Clusters

For each pair, we then use a simple Punnett square approach to generate four children, which is the approximate average of cheetahs per litter, based on all of the alleles of the parents. We then randomly assign a gender to each child and set their age to 2, as we had updated all of the specimens in the population prior to breeding. After generating the offspring's data, we then add them to our population so that they can be included in the next iteration's breeding cycle. This is because the timeline of each iteration is around the same amount of time required for the offspring to reach adolescence.

Results

Experiments

CHEETAH was conducted on an initial generation of 50 cheetahs, 25 male and 25 female. To benchmark our feature selection and clustering methodology, we compared these results with an experiment that simply randomized parental combinations. Females were randomly assigned to a single male who had yet to be bred, and subsequently the pairings were one-to-one, similar to our project's proposal. The benchmark started with the same 50 individuals as our methodology, and the number of total children produced per generation in the benchmark was the same as the total number of children produced per generation in the project. Thus, we conclude our algorithm with the addition of new offspring, and from here, we repeat the entire methodology with the updated population.

Result Discussions

The main experiment conducted was calculating variance per gene and per generation, specifically only among the new children born. This aligns with the vision of genetic diversity through polymorphism. The experiments were conducted looking at several parameters of the model, namely the number of genes to look at for feature extraction, the number of iterations of the model (that is, the number of breedings taken place), and the parental pairing type (for our baseline case, this was assumed to be one to one).

Number of Genes for Features and Clusters Figure 10 and 11 look at experiments for where we varied the number of genes selected for feature and clustering extraction. For legibility, the first 20 genes were shown in the figures, however the model manipulated all 100 genes. This case used 10 iterations and one-to-one pairings. In the figures, per gene, the variances towards the beginning iterations are shown in red, and successive iterations are colored closer to blue. Over these iterations, we notice that the variations indeed become more widespread compared to the Benchmark case. This difference is even more noticeable if we consider more genes for feature and clustering extraction.

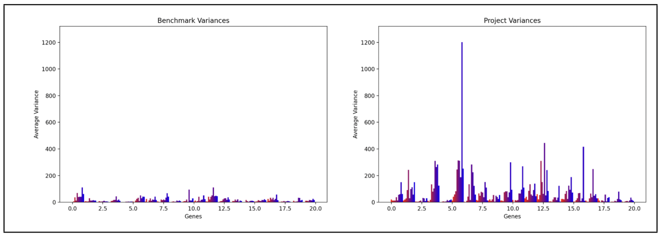


Figure 10: Considering 25 Genes for Features and Clustering

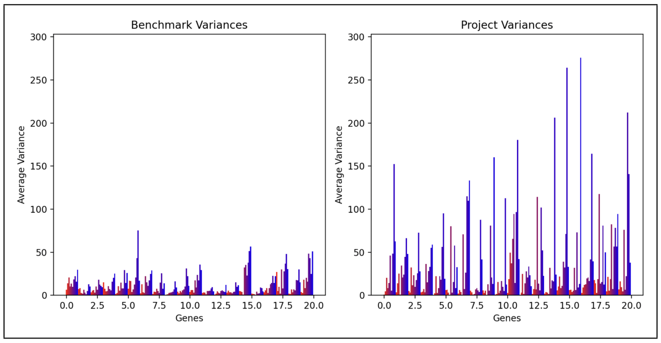


Figure 11: Considering 75 Genes for Features and Clustering

Number of Iterations Figure 12 and 13 look at experiments for where we increased the number of iterations. Again, for legibility, the first 20 genes were shown in the figures, however the model manipulated all 100 genes. This case used the 75-gene feature clustering parameter and one-to-one pairings. In the figures, per gene, the variances towards the beginning iterations are shown in red, and successive iterations are colored closer to blue. Over these iterations, we notice that the variations intuitively increase across all genes. Some variances increase at a faster rate compared to others. This is likely due to the side effects of pairing across many genes. Even if the model itself is not necessarily focused on perfecting some gene at a specific iteration, that gene may be indirectly yet positively affected by breeding for other genes and their respective variances.

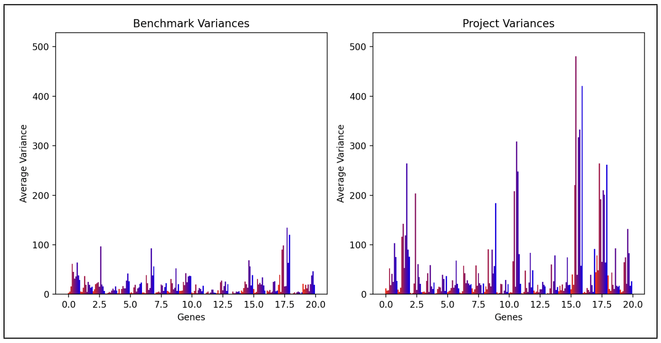


Figure 12: 10 Iterations

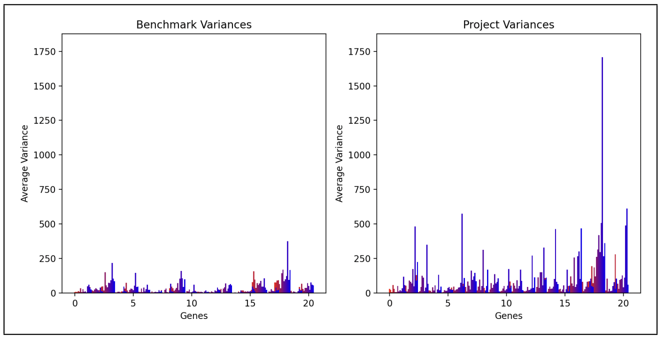


Figure 13: 15 Iterations

Further Work and Considerations

This project has ample room for growth in terms of its methodology and applicability. In terms of the former, the project can add on additional points of metadata for a more robust prediction model. For example, one could incorporate metadata relating to familial ties and ensuring that if a lineage were kept in the breeding program that subsequent breedings were not paired against the same lineage or lineages that have recently bred in the past. In addition, we also want to address how cheetahs can have multiple fathers for a single litter. We note that this might increase the genetic diversity per generation more than our implementation, since there would presumably be around two fathers with different genetic composition per paired breeding, and this allows for more allele differences in the offspring. An approach to this could be finding the two most distant fathers from the mother; however, we would need to consider how different those fathers are from each other as well and could implement this such that all three specimens are in different clusters.

Beyond methodology, the project itself can be updated to incorporate more real-life use-cases of breeding predictions.

For instance, the project can include a browser-like experience where scientists can input sequenced genetic material. The genetic material may have multiple genes of interest of varying lengths, which is unlike our experiments where we assumed (for simplicity in allele generation) the lengths of genes were roughly identical. Moreover, more careful notation can be fed into the model concerning what kinds of alleles may show up per gene, which is unlike our baseline version where we randomly presuppose some permutation of the SNPs to be an allele.

This project, once perfected for cheetahs, can also be applied to other species facing genetic bottlenecks. Pandas and golden snub-nosed monkeys would be potential candidates, as the panda population was believed to have undergone a genetic bottleneck 43,000 years ago and the snub-nosed monkey has seen population decline due to environmental factors (Zhang et al. 2002). Other species include the pinnated grouse and the Northern elephant seal due to habitat destruction and overhunting, respectively (Mussmann, Douglas, and Anthonysamy 2017).

Lastly, we want to address some ethical considerations, given that the project is centered around breeding. One of the more prominent examples of feedback for this project is to note that suggested paired breedings may not be fully realized, as a pairing's success is highly dependent on whether the two individuals are compatible with one another behaviorally. While artificial insemination (AI) may be able to get around this, in many species AI has not necessarily seen large breakthroughs in wild populations and is often relatively expensive (Orzechowski 2015). Similarly, we do not consider the specific alleles' significance in the biological context and might end up increasing the frequency of less advantageous alleles. While this might be evened out via natural selection over time, this might suggest a sub-optimal solution in our pairings, as there might have been better alleles to focus on. We do address the risk of recessive mutant disorders in our approach of choosing optimal parent pairings based on different cluster labels and maximum distance such that the likelihood of inbreeding would decrease.

Conclusion

In this paper, we introduce a forward-looking approach for a cheetah breeding program using the cheetah dataset from the National Library of Medicine. By using feature selection on the alleles and a variation of a divisive hierarchical clustering algorithm, we observe a higher variance in the model's gene output when compared to the benchmark, which was even more notable in the case where we increase the number of genes. The increase in variance suggests that we are able to broaden the genetic compensation among the cheetah population, which is especially significant to this species, since it has a much lower genetic variance than most. Moving forward, we would want to address some of the ethical concerns with breeding programs and also consider more complex genetic traits / environmental considerations to enable higher survival rates of cheetahs. We hope that this approach can be useful for zoos or other environments in which cheetahs can be bred in a safe setting as well as other species that might also benefit from breeding programs.

Contributions

All individuals contributed equally to the project.

1. Sarah Radway: Essay contribution focused on Abstract, Introduction, Background: Related Work, Methodology: Feature Selection. Code contribution focused on Feature Selection.
2. Avery Park: Essay contribution focused on Methodology: Clustering and Breeding, Further Work and Considerations, Conclusion. Code contribution focused on Clustering, Pairing, and Breeding.
3. Madison Davis: Essay contribution focused on Background: Genetic Vocabulary, Dataset, Results, and Future Work and Considerations. Code contribution focused on Dataset Preparation, Main Function, and Benchmark Tool.

Acknowledgements

We thank Professor Andrew Berry, as well as the CS288 staff for their help in developing this line of work.

References

Aines, C.-P.; Bettina, W.; and Simone, S. 2010. Cheetah Paradigm Revisited: MHC Diversity in the World's Largest Free-Ranging Population. *Molecular Biology and Evolution.*

Akdemir, D.; Beavis, W.; Fritsche-Neto, R.; Singh, A. K.; and Isidro-Sánchez, J. 2019. Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity*, 122(5): 672–683.

Allier, A.; Teyssèdre, S.; Lehermeier, C.; Moreau, L.; and Charcosset, A. 2020. Optimized breeding strategies to harness genetic resources with different performance levels. *BMC genomics*, 21: 1–16.

Andrews, C. A. 2010. Natural selection, genetic drift, and gene flow do not act in isolation in natural populations. *Nature Education Knowledge*, 3(10): 5.

Burócziová, M.; and Říha, J. 2009. Horse breed discrimination using machine learning methods. *Journal of applied genetics*, 50: 375–377.

Cohn, J. P. 1986. Surprising cheetah genetics. *Bioscience*, 36(6): 358–362.

Day, M. L.; and Grum, D. E. 2005. Breeding strategies to optimize reproductive efficiency in beef herds. *Veterinary Clinics: Food Animal Practice*, 21(2): 367–381.

Dobrynin, P.; Liu, S.; Tamazian, G.; Xiong, Z.; Yurchenko, A. A.; Krasheninnikova, K.; Kliver, S.; Schmidt-Küntzel, A.; Koepfli, K.-P.; Johnson, W.; et al. 2015. Genomic legacy of the African cheetah, Acinonyx jubatus. *Genome biology*, 16: 1–20.

Durant, S. 2023. Cheetah conervation in Africa.

Fleming, A.; Abdalla, E. A.; Maltecca, C.; and Baes, C. F. 2018. Invited review: Reproductive and genomic technologies to optimize breeding strategies for genetic progress in dairy cattle. *Archives Animal Breeding*, 61(1): 43–57.

Gutenkunst, R. N.; Hernandez, R. D.; Williamson, S. H.; and Bustamante, C. D. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*, 5(10): e1000695.

Jiang, S.; Cheng, Q.; Yan, J.; Fu, R.; and Wang, X. 2020. Genome optimization for improvement of maize breeding. *Theoretical and Applied Genetics*, 133: 1491–1502.

Kaufman, L.; and Rousseeuw, P. J. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley Sons.

Khan, M. H.; Dar, Z. A.; and Dar, S. A. 2015. Breeding strategies for improving rice yield—a review. *Agricultural Sciences*, 6(5): 467–478.

Li, H.; and Roossinck, M. J. 2004. Genetic bottlenecks reduce population variation in an experimental RNA virus population. *Journal of virology*, 78(19): 10582–10587.

Liang, X.; Sha, Q.; Rho, Y.; and Zhang, S. 2018. A hierarchical clustering method for dimension reduction in joint analysis of multiple phenotypes. *Genetic Epidemiology*, 42(4): 344–353.

Mac, B. C. 2023. Nucleic Acid Structure Theory.

Meuwissen, T. 1998. Optimizing pure line breeding strategies utilizing reproductive technologies. *Journal of dairy science*, 81: 47–54.

Mussmann, S.; Douglas, M.; and Anthonysamy, W. 2017. Genetic rescue, the greater prairie chicken and the problem of conservation reliance in the Anthropocene. *R Soc Open Sci.*

Nature. 2014. Population Bottleneck.

NCBI. 2024. SOX6 SRY-box transcription factor 6.

OBrien, S.; Johnson, W.; Driscoll, C.; and Dobrynin. 2017. Conservation Genetics of the Cheetah: Lessons Learned and New Opportunities.

O'Brien, S. J.; Wildt, D. E.; and Bush, M. 1986. The cheetah in genetic peril. *Scientific american*, 254(5): 84–95.

of Science, W. I. 2024. LRRC4C Gene - Leucine Rich Repeat Containing 4C.

of Veterinary Medicine, U. 2022. Acinonyx Jubatus Reference Genome.

Orzechowski, K. 2015. Endangered Species and Artificial Insemination: A Natural Fit?

O'Brien, S. J.; Johnson, W. E.; Driscoll, C. A.; Dobrynin, P.; and Marker, L. 2017. Conservation genetics of the cheetah: lessons learned and new opportunities. *Journal of Heredity*, 108(6): 671–677.

Prost, S.; Machado, A. P.; Zumbroich, J.; Preier, L.; Mahtani-Williams, S.; Meissner, R.; Guschanski, K.; Brealey, J. C.; Fernandes, C.; Vercammen, P.; et al. 2020. Conservation Genomic Analyses of African and Asiatic Cheetahs (Acinonyx jubatus) across Their Current and Historical Species Range. *BioRxiv*, 2020–02.

Recknagel, H.; R, K. R.; and Elmer, A. 2013. A Hybrid Genetic Linkage Map of Two Ecologically and Morphologically Divergent Midas Cichlid Fishes (Amphilophus spp.) Obtained by Massively Parallel DNA Sequencing (ddRAD-Seq).

Senckenberg. 2020. Cheetah Project 1.

Vetmeduni. 2022. 1 ILLUMINA (Illumina HiSeq 2500) ddRAD Run.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3): 261–272.

X, Z.; J., Z.; R.K., S.; K., T.; and C., S. 2010. The Role of NELL-1, a Growth Factor Associated with Craniosynostosis, in Promoting Bone Regeneration. *J Dent Res*.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8): 1586–1591.

Yang, Z.; et al. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences*, 13(5): 555–556.

Zhang, Y.-p.; Wang, X.-x.; Ryder, O.; A., L.; Hai-peng, Z.; He-ming, Y.; and Wang, P.-y. 2002. Genetic diversity and conservation of endangered animal species. *Pure and Applied Chemistry*.