Unraveling the Genomic Diversity and Evolutionary History of Captive Tigers in the United States

Ellie E. Armstrong[1*&], Jazlyn A. Mooney[1,2*&], Katherine A. Solari[1], Bernard Y. Kim[1], Gregory S. Barsh[3,4], Victoria Grant[1], Gili Greenbaum[5], Christopher B. Kaelin[4], Katya Panchenko[1], Joseph K. Pickrell[6], Noah Rosenberg[1], Oliver A. Ryder[7], Tsuya Yokoyama[1], Uma Ramakrishnan[8], Dmitri A. Petrov[1,9,10], Elizabeth A. Hadly[1,11,12, 13&]

[1]Department of Biology, Stanford University, Stanford, CA, USA.
[2]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA
[3]HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA
[4]Department of Genetics, School of Medicine, Stanford University, Stanford, CA, USA
[5]Department of Ecology, Evolution & Behavior, The Hebrew University of Jerusalem, Jerusalem, Israel
[6]Gencove, Inc., New York, NY, 11101, USA
[7]San Diego Zoo Wildlife Alliance, Escondido, CA, USA,
[8]National Centre for Biological Sciences, TIFR, Bellary Road, Bangalore, India 560065
[9]Chan Zuckerberg BioHub, San Francisco, CA USA.
[10]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA.
[11]Department of Earth System Science, Stanford University, Stanford, CA, USA
[12] Woods Institute for the Environment, Stanford University, Stanford, CA, USA
[13] Center for Innovation in Global Health, Stanford University, Stanford, CA, USA

*Contributed equally
[&]Corresponding authors emails: ellieearmstrong@gmail.com; jazlynmo@usc.edu; hadly@stanford.edu

Abstract

Genetic and genomic studies of rare and endangered species have focused broadly on describing diversity patterns and resolving phylogenetic relationships, with the overarching goal of informing conservation efforts. However, many studies do not consider genetic reserves that are potentially housed in captive populations. For tigers (*Panthera tigris*) in particular, captive individuals vastly outnumber those in the wild, and their diversity remains largely unexplored. Here, we present the first large-scale genetic study of the private (non-zoo) captive tiger population in the United States (U.S.), also known as 'Generic' tigers. We find that the U.S. Generic tiger population has an admixture fingerprint comprising all six extant wild tiger subspecies (*P. t. altaica*, Amur; *P. t. tigris*, Bengal; *P. t. corbetti*, Indochinese; *P. t. jacksoni*,

Malayan; *P. t. amoyensis*, South China; *P. t. sumatrae*, Sumatran). We show that the Generic tiger population has a comparable amount of genetic diversity to most wild subspecies, relatively few private variants, and fewer deleterious mutations. We also observe inbreeding coefficients that are similar to wild populations, suggesting that inbreeding in captive populations is not prevalent as previously thought, although there are some individuals within the Generic population that are quite inbred. Our results reflect the complex demographic history of the Generic tiger population in the U.S. Additionally, we develop a reference panel for tigers and show that it can be used with imputation to accurately distinguish individuals and assign ancestry even with ultra-low coverage (0.25×) data. We anticipate this comprehensive study and panel will propel future research and preservation of tigers in the U.S. and globally.

Main

Tigers (*Panthera tigris*) are one of the most iconic and captivating terrestrial species on the planet and exemplify the severe and rapid decline of population size and connectivity that threatens the survival of many species[1]. There are six extant subspecies of tiger (*P. t. altaica*, Amur; *P. t. tigris*, Bengal; *P. t. corbetti*, Indochinese; *P. t. jacksoni*, Malayan; *P. t. amoyensis*, South China; *P. t. sumatrae*, Sumatran) that have been largely genetically and geographically separated for at least the last 10,000 years[2,3]. The subspecies range from the Russian Far East and northeast China (*P. t. altaica*), Bangladesh, Bhutan, India and Nepal (*P. t. tigris*), Myanmar, Thailand, and Laos (*P. t. corbetti*), the island of Sumatra (*P. t. sumatrae*), and the Malay Peninsula (*P. t. jacksoni*), respectively, while the South China tiger (*P. t. amoyensis*) is extinct in the wild. Although some tiger populations and subspecies have benefitted from recent recovery efforts, others have gone extinct in the wild[4] or have been completely extirpated[5] due to human pressures. All subspecies have undergone severe population bottlenecks through the last century, primarily due to human activities such as hunting and the expansion of agriculture, which have directly reduced tiger numbers, habitat availability, and prey[6].

In addition to wild populations, several captive populations of tigers exist. Some captive populations are regulated, such as those encompassed by World Association of Zoos and Aquariums (WAZA), and some are mostly unregulated, such as farmed and other privately owned tigers. In the United States, the Association of Zoos and Aquariums (AZA), manages several tiger populations as distinct subspecies, specifically the Amur (1950s-present), Sumatran (1950s-present), Malayan (1980s-present), and for a time the “Bengal” (white tigers; 1960s-2011) tiger subspecies. The white tiger (“Bengal”) breeding program was discontinued by the AZA since a single variant is responsible for producing the white phenotype[7,8] (occurring naturally, but rarely, in wild Bengal tiger populations) and breeding was determined to not be relevant to conservation initiatives. Further, the population was intentionally mixed with Amur tigers in an attempt to prevent inbreeding depression before the program was discontinued[6]. Outside of AZA zoos, individuals and entities breed, own, and sell tigers both legally and illegally, and up until the recent passing of federal legislation[9] (Big Cat Public Safety Act), there was previously little oversight of these practices. The formation of the privately-owned, captive population in the U.S. likely coincided with the establishment of zoos and circuses in the early 1900s, since these entities were known to exchange individuals[6]. Over time, the mission of AZA zoos began to run counter to circuses and other private ownership practices, but by the 2000s, estimates put the captive tiger population in the United States at greater than 5,000 individuals[6,10].

The pace of breeding in privately owned tiger populations has prompted concerns of inbreeding, especially since phenotypes (i.e. striped whites, snow whites, and golden tigers) that rely on rare, recessive mutations[7,8] are considered to be of high economic value. A previous study documented signatures of admixture in captive, privately owned tigers from various localities using mitochondrial DNA (mtDNA) and microsatellites, and suggested that this group of individuals may contain unique diversity compared to their wild and captive single-ancestry counterparts[11]. As genetic rescue, translocation, and captive breeding programs are being

implemented more frequently in endangered species for strategic management[12,13], a broader understanding of genetic diversity within captive populations, including the opposing forces of admixture and inbreeding, warrants further investigation. It remains unclear how or if this large population of privately-owned, captive tigers could possibly fit into current management or conservation plans, and/or whether their genomes hold some of this purported unique diversity, or if on the other hand, they show indications of severe inbreeding or high mutation load. A baseline of tiger genomic diversity across captive and wild populations will assist in tracking tiger recovery in the wild, making management decisions regarding rewilding, and identifying source populations of illegally traded individuals and materials.

Here, we use genomic data to reconstruct the admixture history of the captive, privately owned tiger population in the United States (also known as 'Generic' tigers) using whole-genome sequence data obtained from individuals in accredited sanctuaries. These sanctuaries do not breed, sell, or buy tigers, but house tigers previously rescued from such facilities or those that have been forfeited by private owners. We combine our newly sequenced individuals (N = 151) with previously published data for single-subspecies ancestry tigers (N = 105), resulting in a dataset representing 256 unique individuals across all six extant subspecies and U.S. Generic tigers. Our dataset is composed of (N = 69) high coverage (>10×) and (N = 187) low coverage (<10×) individuals. These data allow us to determine how admixture shaped the genomic landscape of the privately owned captive population. We compare the Generic tigers to their potential source populations (tigers of single ancestry, i.e., subspecies), and examine how diversity is partitioned across each group. Then, we test for potential signs of inbreeding and quantify the total amount of deleterious variation in each population. Lastly, we show that ultra-low coverage (0.25×) data can be imputed sufficiently using reference haplotypes from single ancestry tigers to identify ancestry and perform individual identification. Our results demonstrate that low-coverage sequencing and imputation provide a cost-effective alternative to identify the source populations of illegally traded individuals or wildlife materials from tigers and for long-

term monitoring of captive and wild populations.

Results

Population structure and ancestry in Generic tigers

We investigated population structure and ancestry of the Generic tigers using a combination of, principal component analysis (PCA), supervised ADMIXTURE, mitochondrial haplotyping, and local ancestry segments. Both the supervised ADMIXTURE analysis (Fig. 1A and *Supplementary Notes*) and PCA (Supplementary Fig. 1), show that Generic tigers are admixed between the six extant tiger subspecies, with individuals varying in the proportion of their genome derived from each ancestral population (Fig. 1A). This result is also mirrored in their mitochondrial haplotypes (Supplementary Fig. 2-4) and even more extensively when examining local ancestry (Supplementary Fig. 5). We found no Generic tigers to have single-subspecies ancestry. Generic individuals, on average, are primarily admixed between the Amur and Bengal tiger subspecies (mean: 38.6% and 28.2% of ancestry derived from each subspecies, respectively) and derive the least ancestry from the Sumatran and Malayan populations (mean: 5.0% and 11.1%, respectively; Fig. 1A). Our clustering analyses and PCA results showed that the captive population clusters together by their dominant ancestry in both PC space and the identity-by-state (IBS) tree (Fig. 1B; Supplementary Fig. 6C). We also observe some clustering based on putative geographic origin in North America (Fig. 1C), though it is less distinguishable than the clustering by ancestry. These results suggest that structure of the Generic population is primarily caused by the breeding events that established the captive population and there is weak evidence for subsequent geographic structuring. We see further evidence of historical admixture based on mitochondrial haplotype results, where Generic individuals cluster with Amur, Indochinese, and Bengal tiger subspecies. We observed only one Generic sample (SRR836354) grouping with Malayan tigers in mitochondrial haplotype

analyses, and none grouping with Sumatran or South China tigers (Supplementary Fig. 2-4). Further information can be found in *Supplementary Notes*.

Interestingly, our analyses also identified several individuals previously labeled as single-ancestry subspecies to be Generic individuals. These individuals were originally labeled as Amur (SRR7651464-67, SRR7651470) and Bengal (SRR836354) subspecies in public repositories, but we observe that they all have a mixed ancestry background based on both PCA (Supplementary Fig. 17) and subsequent admixture analysis (Fig. 1A). These individuals lack additional metadata, but are presumed to have originated in Asia (Supplementary Table 1) and thus are labeled as ‘Asia’ in the admixture plot (Fig 1A).

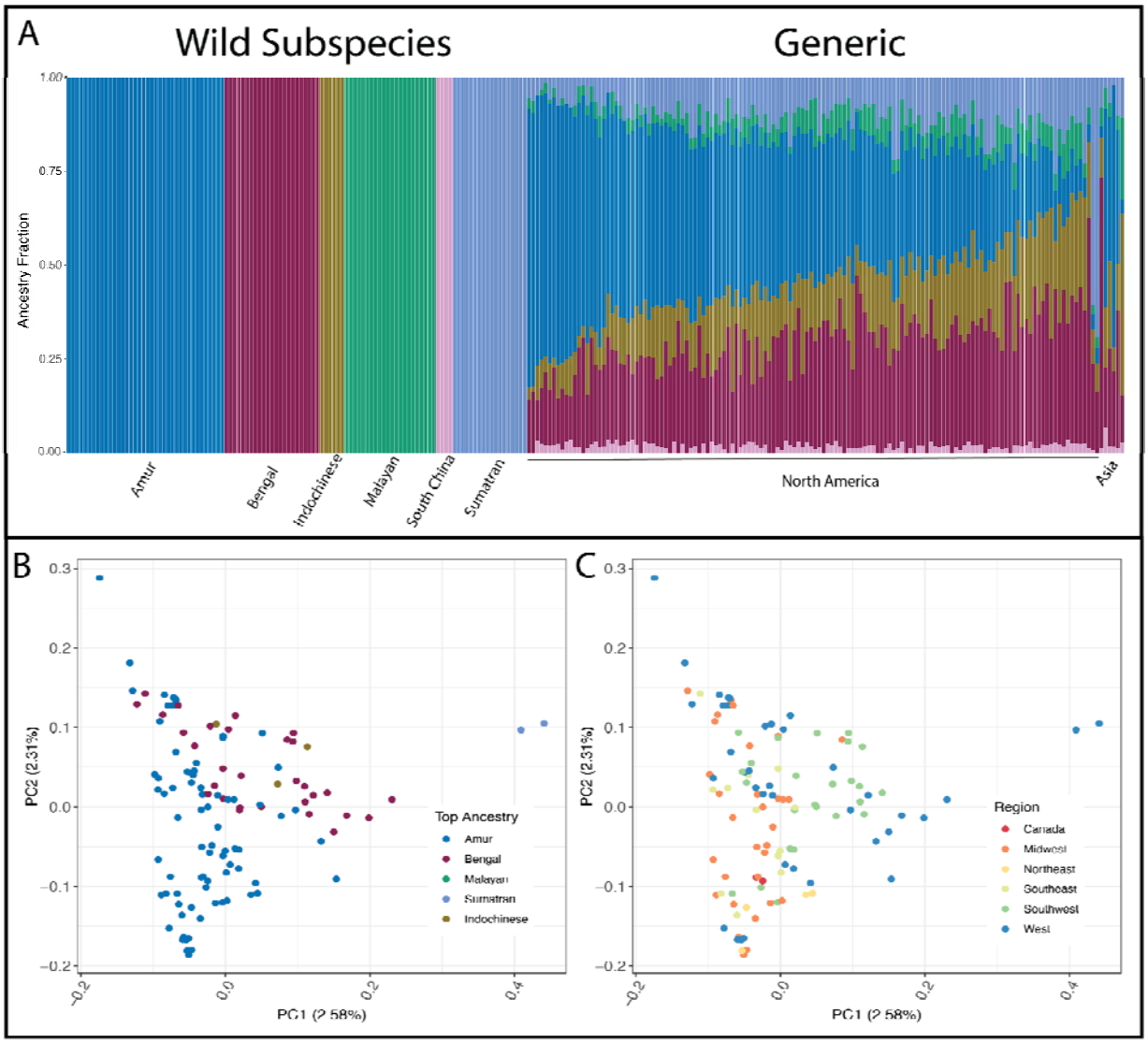


Fig. 1: Ancestral diversity in Generic tigers. **A)** ADMIXTURE analysis showing ancestry of all individuals, wild and Generic. **B)** PCA of North American Generic tigers where color corresponds to an individual's major ancestry component. **C)** PCA of North American Generic tigers where color corresponds to the region of North America from which the individual originated. For panels B & C, one individual was removed as an outlier, details can be found in *Supplementary Notes*.

Tiger subspecies diversity and population history

Since the Generic tigers form a well-admixed population, we next investigated their genetic diversity, inbreeding, and mutation load relative to the single-ancestry, wild subspecies. Primarily, we wanted to determine whether Generic tigers were potential reservoirs of genetic diversity that is absent in the wild. Given that bottlenecks have occurred both in wild populations

and during the process of founding the various captive tiger populations, the surviving diversity in the Generic population in comparison to wild populations is unknown. To address this, we estimated the number of distinct alleles per-locus, calculated the proportion of sites that are variant and segregating in the Generic tigers, but fixed in the wild subspecies, and calculated observed heterozygosity for each individual (Fig. 2 & Supplementary Figs. 7-8). ADZE trajectories for allelic diversity suggested the Indochinese subspecies (Supplementary Fig. 7) contained the most diversity, followed by the Bengal subspecies. However, due to the small sample size for Indochinese tigers (N = 6), it is unclear whether their allelic diversity scales with sample size as it does in the Bengal tiger subspecies (Supplementary Fig. 8).

Next, we asked whether there might be an enrichment of variant sites that are solely segregating within the Generic population but fixed within any reference wild subspecies (Fig. 2A). We found that most sites that were segregating within the Generic tigers are also segregating in at least one of the wild subspecies (Fig. 2A). Specifically, less than 10% of sites are exclusively segregating within the Generic tigers (Fig. 2A). In other words, approximately 90% of variant sites in Generic tigers are also variable within any one of the wild subspecies (Fig. 2A).

Finally, we examined heterozygosity. Generic tiger heterozygosity fell well within the range of other wild subspecies (Fig. 2B). Like what we observed with ADZE, Bengal and Indochinese tigers had the highest heterozygosity. Taken together, our analyses suggest that the Generic population is not a major reserve of unique genetic diversity, rather it contains diversity that is already present in the wild.

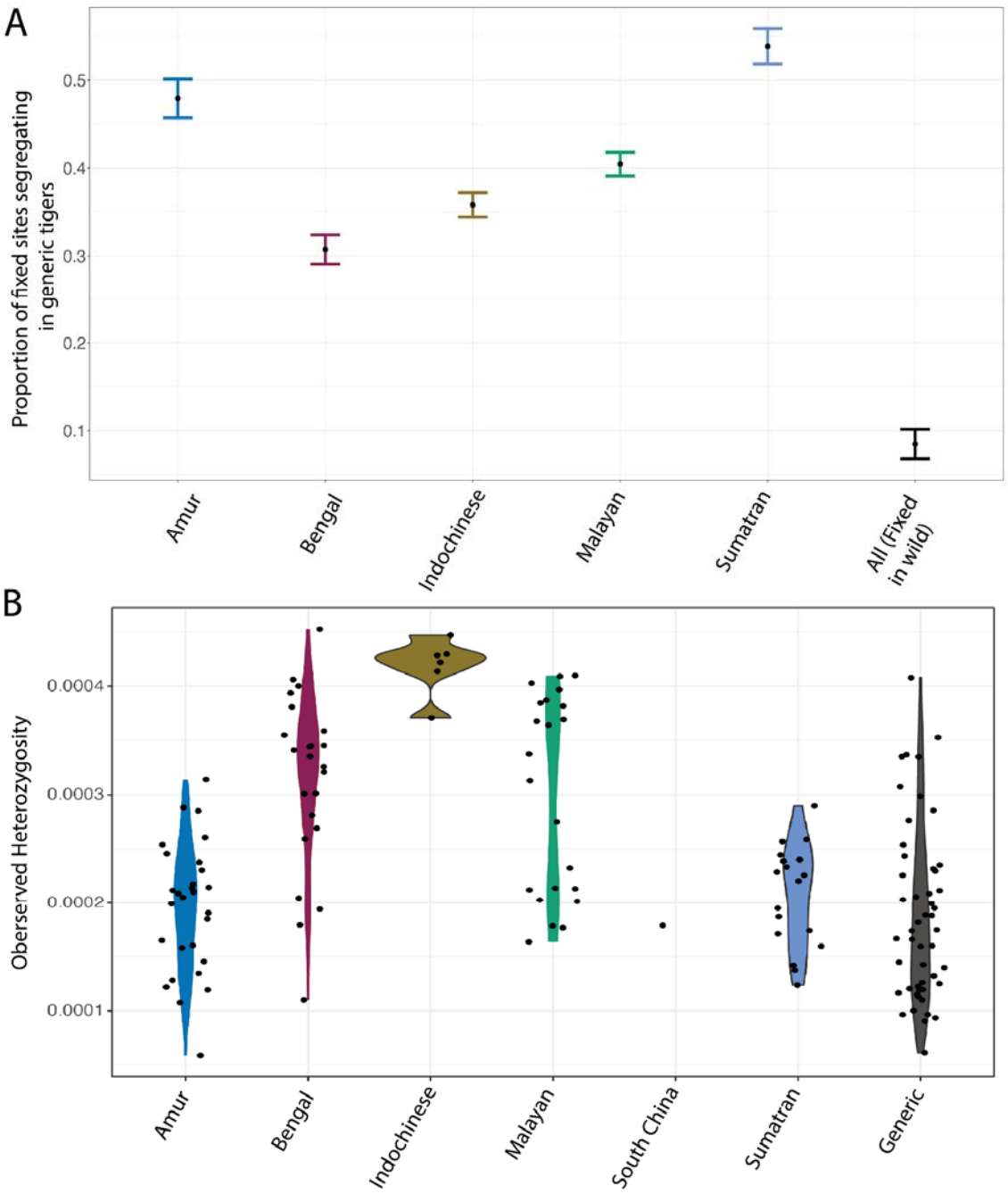


Fig. 2: Measurements of shared and within subspecies genetic diversity across tigers. **A)** Proportion of segregating sites that are fixed in each wild subspecies compared to Generic tigers. The mean and standard deviation are computed across 10 replicates of captive individuals (unimputed individuals). South China is not included since there is only one individual. The designation of "All" indicates the site is fixed across all wild subspecies, but segregating in the Generic population. **B)** Observed heterozygosity distributions of unimputed individuals (See Supplementary Methods and Imputation and filtering for details).

Due to captive breeding practices and a small founder population, we wondered whether the Generic tiger population showed signs of inbreeding. Thus, we examined the extent of inbreeding using multiple metrics, including runs of homozygosity (ROH), shared identity-by-descent (IBD) segments, and inbreeding coefficients from both SNPs (F_{SNP}) and ROH (F_{ROH}).

We observed that the Amur population had, on average, the most Type C (i.e., long runs, informative about recent inbreeding) ROH and highest IBD-sharing (Fig. 3 and Supplementary Figs. 9-10), suggesting that Amur tigers experienced more severe inbreeding than the Generic population. Further, Amur tigers also had the largest amount of Type B (i.e., intermediate length) ROH, which indicates a long-term, small population size[14]. Indeed, Amur tigers were documented as having experienced an extreme bottleneck in the 1930s/1940s[15,16], possibly explaining these extreme patterns compared to other subspecies.

When examining two other metrics of inbreeding (F_{SNP} and F_{ROH}) we observed that the Generic tiger population was once again not an outlier. Instead, most Generic individuals had similar inbreeding values to individuals in the other subspecies (Fig. 3). However, there are a few individuals within the Generic population that are quite inbred. The variable levels of inbreeding matched previous analysis of heterozygosity, where we observed variable levels of genome-wide heterozygosity across the Generic population (Fig. 2B). Additionally, we also found that when examining the site frequency spectra (SFS), the Generic tigers also had the largest proportion of singletons (Supplementary Fig. 11). The variable inbreeding coefficients (Type C ROH, F_{ROH}, and F_{SNP}) suggest that in the case of Generic tigers, admixture between distinct subspecies likely increased heterozygosity in some individuals, but inbreeding within individual facilities eroded it[17]. Taken together, our results show that contrary to previous intuition, the Generic population is not any more inbred than wild tiger populations, nor is it more diverse.

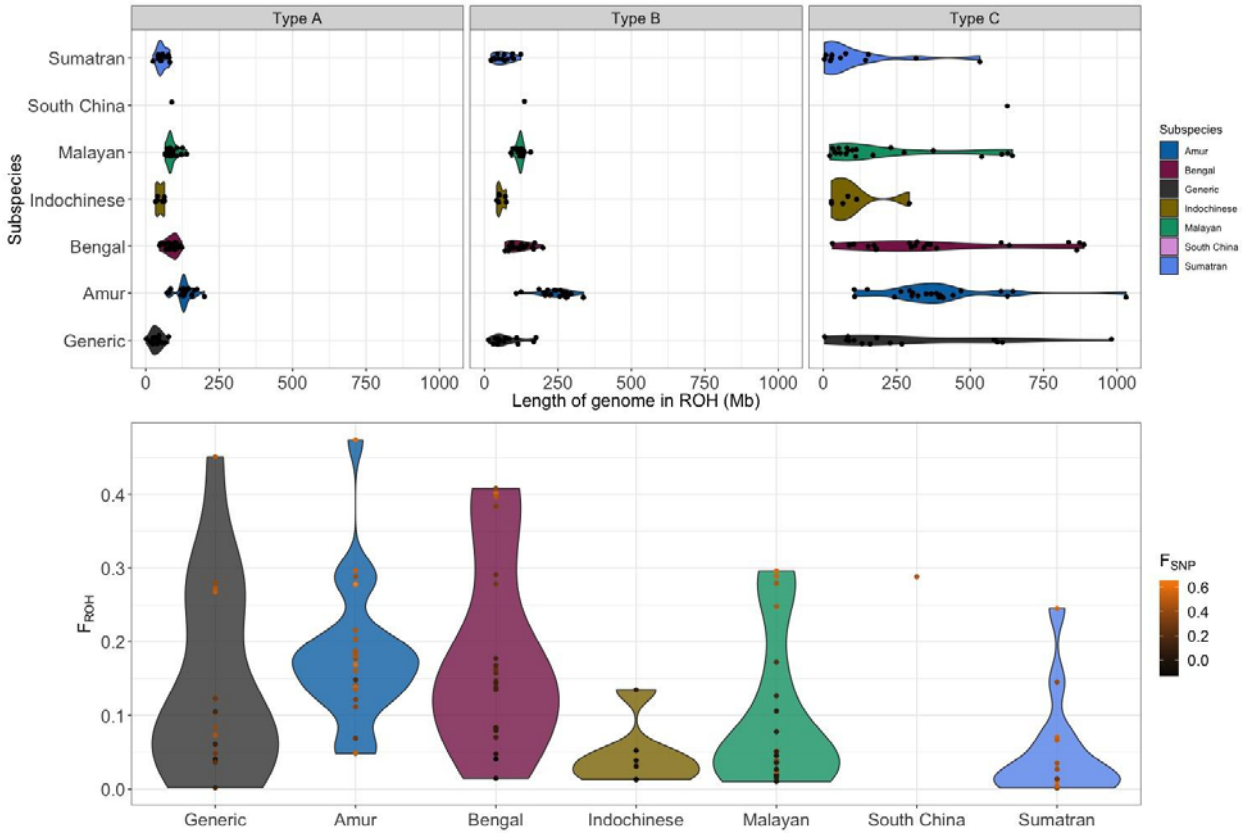


Fig. 3: Quantification of different types of ROH across tiger subspecies and measurements of inbreeding. **(Top)** Length of different classes of ROH (A; short, B; intermediate, C; long) for each tiger subspecies and Generic tigers. **(Bottom)** Total proportion of genome in Type C ROH (F_{ROH}) for each individual per population. We have labeled each individual with their corresponding inbreeding coefficient measured as F_{SNP}, the darkest color (black) is the lowest value and the brightest color (orange) corresponds to the largest value.

Tiger subspecies mutation load

We last sought to characterize the prevalence of putatively deleterious variation in the captive population, relative to each subspecies. Variants were categorized as putatively deleterious if they were nonsynonymous, derived mutations with a SIFT score less than 0.05 (see *Supplementary Methods* for more detail). We implemented multiple approaches to count deleterious variants in the genome: 1) counting homozygous derived genotypes; 2) counting the total number of homozygous and heterozygous derived genotypes (counting variants); and 3)

counting alleles where we tabulate twice the number of homozygous derived genotypes plus heterozygous genotypes (counting alleles).

Given that the Generic population primarily derived its ancestry from the Amur and Bengal subspecies (Fig. 1A), we expected that the distribution of deleterious variation in Generic tigers would be comparable to the Bengal and Amur subspecies. Indeed, the observed distribution of both neutral and deleterious variants in Generic tigers was more similar to the Amur tigers and Bengal tigers than other wild subspecies (Fig. 4 & Supplementary Fig. 12). Overall, we observed similar levels of deleterious variation across Sumatran, Malayan, Indochinese, and Bengal subspecies. Though the mean number of derived deleterious variants was slightly, but consistently, higher in the Sumatran subspecies for all counting methods, the Sumatran subspecies distribution overlaps with at least one of the four other subspecies for all counting methods (Supplementary Fig. 12). We found that irrespective of subspecies, individuals with the most putatively deleterious derived homozygotes also tended to have the largest inbreeding coefficients, quantified with either F_{SNP} or F_{ROH} (Supplementary Fig. 14). Thus, inbreeding likely led to an overall decrease in fitness for all tigers, irrespective of their ancestry. Our observation that the Generic tigers had the lowest average numbers of derived deleterious mutations, suggests that the Generic tigers may have been able to either purge some of the variants that are deleterious in homozygous form being carried by the other subspecies, or the admixture process reduced their frequency as homozygotes.

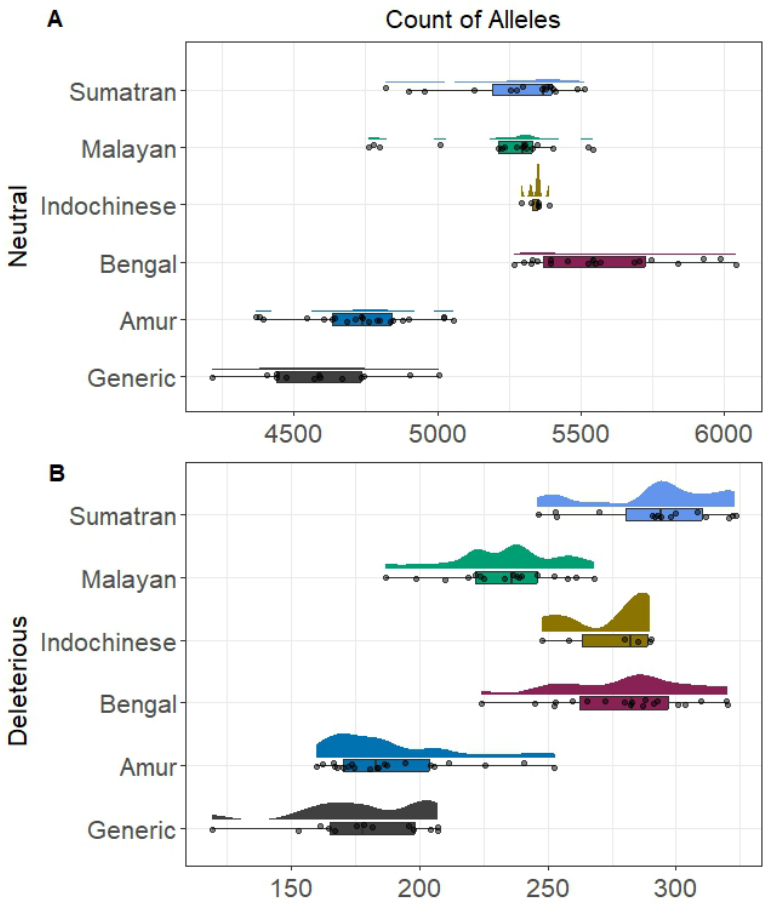


Fig. 4: Quantifying genetic load in each subspecies by counting derived neutral and deleterious variants. **A)**: Scaled counts of derived neutral alleles in each tiger subspecies and Generic tigers. **B)**: Scaled counts of derived deleterious alleles in each tiger subspecies and Generic tigers. Extended figure (Supplementary Fig. 12) contains the count of homozygotes and count of variants.

Creation of a reference panel for future tiger conservation

Given that we collated the largest dataset of tigers to date, we also decided to test whether the data was sufficient to create an accurate reference panel for imputation, which would be particularly useful in wildlife forensics. The primary goal of the reference panel was to assess whether we could impute ultra-low coverage data and accurately identify individuals and their ancestry. The full data was split into a training (reference panel) set composed of 106 single-ancestry individuals[2,3] with coverage between 2×-43×. Then, imputation was performed using *lowimpute* on 86 low- and ultra-low coverage (0.25×-6×) samples (Supplementary Fig. 1).

We tested the accuracy of the imputation pipeline by comparing variant calls, ancestry predictions, and the ability to uniquely identify individuals. We tested accuracy in a set of nine individuals (six unimputed and three imputed), where four of the individuals were also represented in the reference panel (See Supplementary Notes for additional details).

We found that as sequencing depth increased, the accuracy of variant calls, as measured by the non-reference discordance rate (NDR), also increased (Supplementary Table 1). NDR ranged from 7.71% at the highest depth (6×) to 14.49% at the lowest depth (0.25×). We also observed variance in NDR depending on whether a sample was included in the reference panel, and predictably, samples from the reference panel had the lowest NDR (Supplementary Table 1). Despite the variance in NDR, ancestry (Supplementary Fig. 15) and the ability to identify individuals (Supplementary Table 2), even at ultra-low coverage (0.25×), remained accurate. Our results show that low-coverage sequencing and imputation are a sufficient and low-cost alternative to high-coverage sequencing or genotyping using this reference panel for tigers.

Discussion

Using whole genome data from 256 tigers, we investigated the diversity and ancestry of the privately owned, Generic, captive tiger population in the United States and how they compare to the six subspecies of tiger. Additionally, we built a reference panel for tigers that can be used to impute low-coverage data, reliably distinguish individuals, and determine individual ancestry even at ultra low-coverages (0.25× and above).

We found that all Generic tigers sampled from the U.S. are admixed and most contain ancestry from all extant tiger subspecies (Fig. 1A). Further, we identified several tiger genomes in public data repositories originating in Asia that are incorrectly identified as having ancestry from a single tiger subspecies (Fig. 1A, Supplementary Figs. 17-18, *Supplementary Notes*). Most Generic tigers derive a majority of their ancestry from the Amur and Bengal subspecies

(Fig. 1A), but individuals varied in the proportion of ancestry derived from any one subspecies (Fig. 1A, Supplementary Fig. 5). The length distribution of local ancestry segments (Supplementary Fig. 5) indicated that there are likely very few individuals in the extant U.S. Generic tiger population that have directly descended from recently introduced, single ancestry sources.

Our findings differ from the results obtained from Luo et al. 2008, where microsatellite and mitochondrial data were used to investigate samples from captive tigers obtained between 1982 and 2002 from 14 countries (including the U.S.); and found that almost half of the individuals tested from captivity were assigned to single subspecies ancestry[11]. Luo and colleagues (2008) did not describe the origin of the sampled individuals in detail, but it is possible that individuals of single-subspecies ancestry were being actively mixed into the privately-owned captive tiger population when those samples were collected. Alternatively, it is possible that the small number of microsatellites used in the previous study have comparatively limited power to distinguish the complex ancestries of the Generic tigers. Interestingly, Luo and colleagues (2008) found that captive tigers with mixed subspecies ancestry contained primarily ancestry from the Indochinese tiger lineage, whereas the tigers analyzed here contained mostly Amur and Bengal ancestry. Additional genomes for captive, non-zoo individuals from different countries would provide insight into whether tiger ancestry proportions are a useful fingerprint distinguishing continental or country-specific tiger populations that could be leveraged for wildlife investigations.

There has been ongoing discussion regarding how captive populations (both zoo and Generic tigers) might contribute to conservation, and particularly whether they harbor unique alleles beneficial to the survival of the species. While there is no doubt that the Generic tiger population carries some unique variation simply because of the number of individuals in captivity (i.e., census size), we have shown here that the population does not contain more or more unique genetic variation than its single-ancestry counterparts (Fig. 2, Supplementary Figs.

7-8). In other words, Generic tigers have average heterozygosity (Fig. 2B), very few unique segregating sites (Supplementary Fig. 7-8), and when compared to other subspecies, most sites that are segregating in the Generic population are also segregating in the wild populations (Fig. 2A). While there may be some beneficial alleles circulating in the Generic population, predicting which loci are adaptive is a challenging task for even the most well-studied species, and implementing gene-based management strategies can also bring about harmful side-effects[18,19].

The Generic population contains less deleterious recessive homozygous variation and relatively low amounts of Class C ROH relative to wild populations (Fig. 3 and Fig. 4). Given the degree of admixture in the Generics, the low frequency of deleterious mutations is expected, yet surprising given the purported inbreeding of 'farmed' tigers[20]. Thus, the impact of admixture among Generic tigers on the genomic landscape are, in this case, not obvious. Heterozygosity and load are similar in the other subspecies and the Generic population, likely due to a composite effect of admixture between distinct lineages increasing heterozygosity, followed by severe bottlenecks of repeated founding events and deliberate inbreeding. The observed outcome of admixture in the Generics is the result of chance, but is not necessarily predictive of other metrics of inbreeding, load, or heterozygosity, and also does not negate possible effects of outbreeding, which we did not examine here. This result has also been observed in human populations, specifically in admixed Latin American founder populations[17]. Several notable zoo populations, such as giraffes[21] and orangutans[22] have sourced animals for captive breeding programs from multiple, genetically distinct populations, subspecies, or species. If we continue to consider captive populations as diversity reservoirs for wild species, additional investigation into the potential outcomes and interplay of such processes (admixture, inbreeding vs outbreeding depression) is extremely important. We further caution that deleterious variation analyses, especially in non-model organisms, should be viewed skeptically. Deleterious variation analyses generally rely on annotations from other organisms and the accurate inference of an ancestral state, both of which can be imprecise analyses even in the most well-

studied model organisms[23–25]. Until we have a way to more confidently link variants to their functional consequences in the wild, this information should be integrated into management plans with extreme caution.

The last objective of this study was to compile a reference dataset and imputation panel that was capable of identifying the ancestry of unknown individuals. Our panel has already proven useful, since we revealed that several tigers publicly identified as single-ancestry are in fact admixed between several subspecies (See *Supplementary Information*, Fig. 1A, Supplementary Figs. 16-17). As more tiger genomes are sequenced and tiger genome assemblies improve, the tiger reference database developed here can be updated to improve its accuracy. The reference dataset and imputation pipeline from our study provides a robust way to identify individuals and their ancestry rapidly and cost-effectively, even from very low-coverage samples.

Cumulatively, our analyses are concordant with the known history of tiger subspecies in terms of historic bottlenecks and recent inbreeding, yet we also reveal the unique history of the large population of Generic tigers in the United States and provide a comprehensive overview of all extant subspecies and available data. Most Generic tigers contain ancestry from all six wild tiger subspecies in their genomes. Contrary to previous intuition, most Generic tigers do not show signs of severe, recent inbreeding, nor do they hold unique diversity. Thus the role they might play (if any) in conservation remains to be seen. Indeed, further investigation into the single-ancestry captive populations will further elucidate the genetic resources available for tiger conservation and the formation of these unique captive populations. However, this Generic population is the product of unique evolutionary admixture events which may help us better understand the consequences of mixing divergent, wild lineages in the future. To encourage future study and aid illegal trade and trafficking investigations, we also present a reference panel that quickly identifies tigers and their ancestry using even ultra low-coverage sequencing. Taken together, our results are not only informative for the conservation management of tigers,

but also provide a cost-effective path for others in the conservation community to follow for generating a reference panel when using low-coverage sequencing for identification of individuals and their ancestral origins.

Methods

Sample collection and sequencing

A total of 155 tiger samples were collected opportunistically during routine vet care from sanctuary facilities (Supplementary Table 1) by vet and sanctuary staff. All samples were extracted using a Qiagen DNeasy kit (Cat. No. 69504) and samples prepared using a modified Nextera library prep protocol[26]. 78 of these samples (listed as 'Unimputed' in Supplementary Data 1) were sequenced between approximately 2× and 5× depth. The remaining 77 samples ('Imputed' in Supplementary Data 1) were sequenced at approximately 0.25×. Tigers collected in sanctuaries or with unknown ancestry are labeled as "Generic" for the purposes of this manuscript. We additionally collected data on the putative location of birth of each tiger from the sanctuaries in North America. These locations were then translated into one of six regions: West (Washington, Oregon, California, Nevada, Montana, Idaho, Wyoming, Colorado, Utah), Southwest (Arizona, New Mexico, Texas, Oklahoma), Midwest (North Dakota, South Dakota, Nebraska, Kansas, Minnesota, Iowa, Missouri, Wisconsin, Illinois, Michigan, Indiana, Ohio), Southeast (Arkansas, Louisiana, Tennessee, Mississippi, Alabama, Georgia, Florida, Kentucky, West Virginia, Virginia, North Carolina, South Carolina), and Northeast (Maine, New Hampshire, Massachusetts, New York, Pennsylvania, Maryland, Delaware, Connecticut, Vermont, Rhode Island, New Jersey) and Canada.

Variant-calling and reference panel construction

An additional 99, publicly available (as of December 2019) tiger genome samples were downloaded from NCBI. Reads were mapped to the GenTig1.0 genome[27] using BWA-MEM v0.7.17[28] and variant calling was subsequently performed by Gencove using the Genome

Analysis Toolkit (GATK) v4.1.4.1[29] according to best practices. Initial variant calling was performed on all samples available at the time, excluding those sequenced at 0.25×, for a total of 177 individuals. All of these samples are referred to as 'unimputed' for the purposes of this manuscript. Initial variant calling yielded a total of 23,579,569 variable sites across the entire genome. We restricted calls to biallelic sites using BCFtools v1.6[30], and subsequently filtered for quality, missingness, and depth. See *Supplementary Methods* for additional details. The final data set contained a total of 7,519,430 sites across unimputed individuals and a callable genome of 2,174,711,735 base pairs.

In order to select individuals to build the reference panel and accurately split individuals into groups for kinship estimation, we conducted Principal Component Analysis (PCA) to ensure that all individuals in the unimputed dataset were clustering according to subspecies using PLINK v2[31]. Individuals were subsequently split into ancestry groups to form the reference panel, which included representatives from all six tiger subspecies. Further, we tested several methods for detecting relatedness using pedigreed individuals in the dataset, which were subsequently used to identify and remove duplicates. Additional information can be found in *Supplementary Methods*.

Imputation and filtering

Using only single-subspecies ancestry individuals verified here and by previous studies, we developed a reference panel to impute variants for an additional 86 individuals (labeled as 'imputed' in Supplementary Data 1) through the *loimpute* pipeline developed by Gencove and available at www.gencove.com[32]. The 86 imputed individuals were composed primarily of individuals sequenced at ultra-low coverage (N = 75; 0.25×), but also included two individuals from a Canadian Zoo sequenced at approximately 3×, and an additional 9 samples that became publicly available after the initial variant calling had been performed (see Supplementary Data 1 for details).

We combined the VCFs for imputed individuals with the unimputed individuals using BCFtools *merge*. Because imputation emits a call for every site in the reference pipeline, we restricted the merged sites VCF to retain only the quality sites identified after initial variant calling and filtering. We further checked for imputation accuracy using concordance measures, specifically non-reference discordance[33], and examined the accuracy of ancestry and relatedness measures over a variety of coverages. Additional information can be found in *Supplementary Methods*.

Genetic diversity

We used several approaches to investigate genetic diversity across tigers. The first approach was to create equal sized groups of (N=10) individuals and tabulate the proportion of sites that were SNPs in the Generic population and fixed in each wild subspecies. We used the same reference groups from each subspecies, and generated 10 replicate samples (with replacement) of the Generic tigers to check whether these proportions varied across the Generic population.

Heterozygosity was also calculated as the total number of heterozygous sites divided by the number of callable sites (2,174,711,735) in the genome. Observed homozygous sites were counted in each subspecies using VCFtools[34] using the '--het' flag and heterozygous sites were calculated by subtracting the (O)HOM (observed homozygosity) column from the NSITES (the total sites queried) column. The number of callable sites was determined as the total number of basepairs minus the sites with mappability scores < 1 for autosomal scaffolds. We additionally tested to see if heterozygosity was correlated with missingness, since most missing calls would presumably be heterozygous given the data was imputed. See *Supplementary Methods* for details.

Runs of homozygosity (ROH)

GARLIC[35] was used to detect ROH. The error was set at 0.001, the window size at 700, and centromeres were set as 0,0 since no centromere information was available. To ensure that ROH were not mistakenly called on regions with an excess of missing calls, each ROH was then intersected with the number of callable sites in each window (see *Heterozygosity* section for details). ROH larger than 100kb and containing callable sites within one standard deviation ($\pm$0.066) of the mean coverage (0.655) were retained. F_{ROH} was computed as the total fraction of the genome within a type C ROH.

Identity-by-descent segments (IBD)

IBD segments were called using TRUFFLEv1.38 with parameters "--segments --missing 1 --maf 0 –nofiltering[36]. Segments greater than 2Mb where the fraction of coverage by callable sites (count listed above) was within one standard deviation ($\pm$0.032) of the mean coverage (0.660) were retained. IBD scores were computed using the same approach from Nakatsuka et al.[37], where we normalize shared IBD by the number of sampled individuals.

Putatively neutral and deleterious variation

We polarized and annotated sites by first subsetting the data to scaffolds that corresponded to main chromosomes from felCat8[38],GCF_000181335.2). Coordinates were identified using liftOver[39]. The remaining sites, with felCat8 coordinates, were annotated with an impact and consequence using VEP v92. We removed intergenic sites, splice acceptor, splice donor, splice region annotations, and selected the most damaging impact for a given transcript. Sites were coded as nonsynonymous (NS), synonymous (SYN), or loss of function (LOF) and SIFT scores were added [40]. Information from VEP was combined with a SIFT score to find putatively neutral (SYN with SIFT score greater than 0.05) and putatively deleterious sites (NS or LOF with SIFT score less than 0.05). See *Supplementary Methods* for details about annotations and SIFT scores.

Ancestral bases were identified with Progressive CACTUS[41] alignment of the GenTig1.0 genome with 9 other felid genomes; further details can be found in *Supplementary Methods*. To assess load in each subspecies, we used only unimputed individuals with at least 5× coverage and less than 5% missing data. Lastly, we scaled the number of sites per individual by subtracting the total number of variant sites across all individuals from missing sites to get the total number of called sites. Each count was divided by the number of callable sites for that individual; the proportion was multiplied by the average number of callable sites across all subspecies.

Data Availability

Data associated with this stud has been deposited into bioproject number PRJNA976043 and will be released upon acceptance.

Code Availability

All code associated with this project can be found at https://github.com/jaam92/Tigers/.

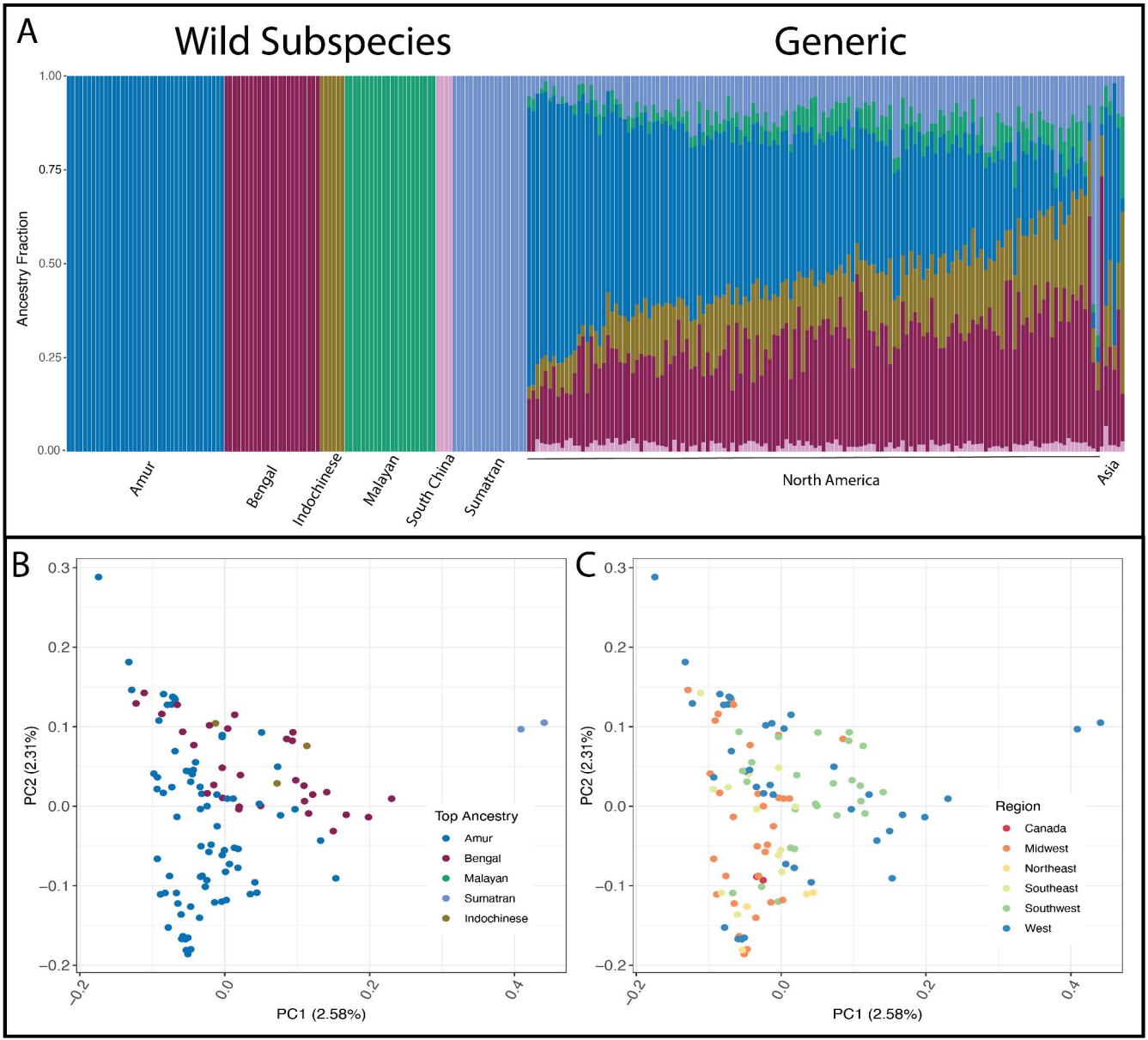
Acknowledgements

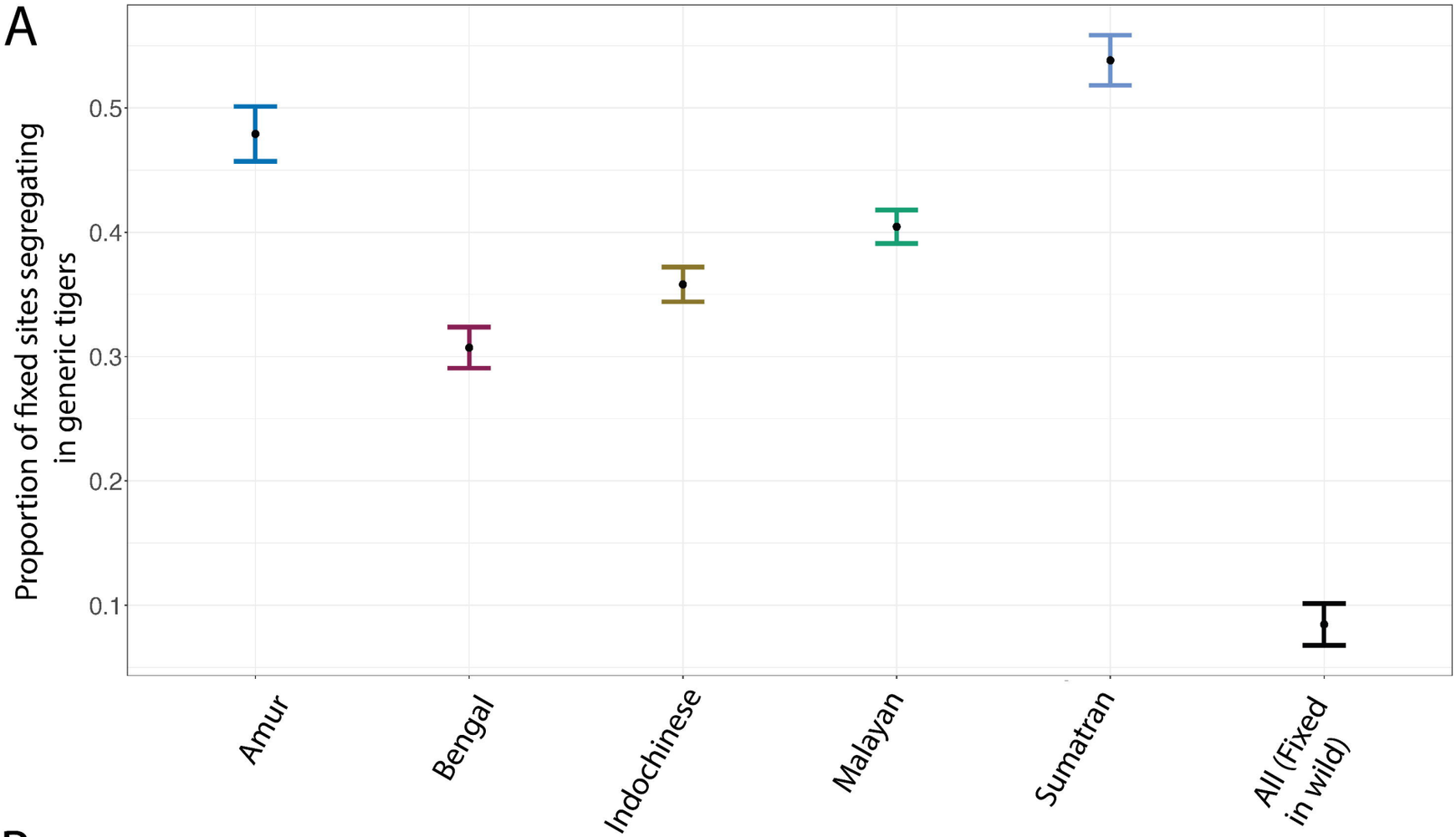
We would like to acknowledge each of the sanctuaries and zoos who contributed samples, knowledge, and time to this study, specifically the Performing Animal Welfare Society, the Exotic Feline Rescue Center, In-Sync Exotics, Turpentine Creek Wildlife Refuge, Carolina Tiger, and others. We especially acknowledge Joe Taft and Rebecca Rizzo of EFRC, Jackie Gai and Ed Stewart of PAWS, and Vicky Keahy of In-Sync, who have spent many hours discussing this project with us. This project would not have been possible without Bill Nimmo and Kizmin Reeves of Tigers in America. In addition, we thank the San Diego Frozen Zoo for access to their collections and their collaborative spirit. We also thank Sara Huston Katsanis for procuring the Carolina Tiger Rescue and Tiger World specimens and supporting the mentoring grant for V. Grant.

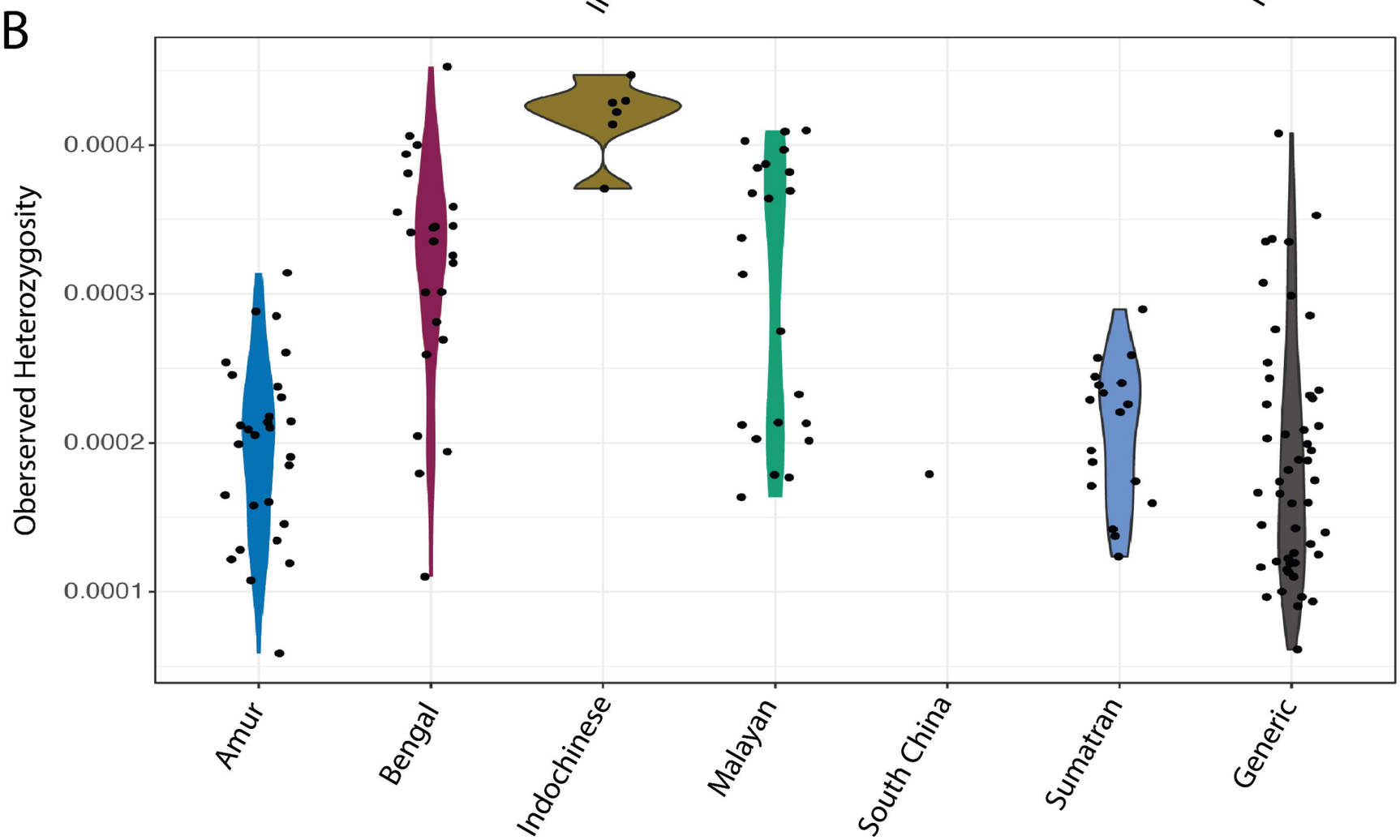
References

1. Walston, J. *et al.* Bringing the Tiger Back from the Brink—The Six Percent Solution. *PLoS Biol.* **8**, e1000485 (2010).
2. Liu, Y.-C. *et al.* Genome-wide evolutionary analysis of natural history and adaptation in the world's tigers. *Curr. Biol.* **28**, 3840–3849 (2018).
3. Armstrong, E. E. *et al.* Recent Evolutionary History of Tigers Highlights Contrasting Roles of Genetic Drift and Selection. *Mol. Biol. Evol.* **38**, 2366–2379 (2021).
4. Tilson, R., Traylor-Holzer, K. & Jiang, Q. M. The decline and impending extinction of the South China tiger. *Oryx* **31**, 243–252 (1997).
5. Xue, H.-R. *et al.* Genetic ancestry of the extinct Javan and Bali tigers. *J. Hered.* **106**, 247–257 (2015).
6. Tilson, R. & Nyhus, P. J. *Tigers of the World: The Science, Politics and Conservation of Panthera tigris*. (Academic Press, 2009).
7. Xu, X. *et al.* The genetic basis of white tigers. *Curr. Biol.* **23**, 1031–1035 (2013).
8. Xu, X. *et al.* The genetics of tiger pelage color variations. *Cell Res.* **27**, 954–957 (2017).
9. Grijalva, R. M. *et al.* Big Cat Public Safety Act. [H. rept. 116-430 for H.r. 1380]. (2020).
10. Garner, R. A Census of Federally Regulated Big Cat Populations within the United States as of December 2020. *Journal of Zoological and Botanical Gardens* **2**, 517–530 (2021).
11. Luo, S.-J. *et al.* Subspecies genetic assignments of worldwide captive tigers increase conservation value of captive populations. *Curr. Biol.* **18**, 592–596 (2008).
12. Whiteley, A. R., Fitzpatrick, S. W., Funk, W. C. & Tallmon, D. A. Genetic rescue to the rescue. *Trends Ecol. Evol.* **30**, 42–49 (2015).
13. Witzenberger, K. A. & Hochkirch, A. Ex situ conservation genetics: a review of molecular studies on the genetic consequences of captive breeding programmes for endangered animal species. *Biodivers. Conserv.* **20**, 1843–1861 (2011).
14. Pemberton, T. J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).
15. Alasaad, S., Soriguer, R. C., Chelomina, G., Sushitsky, Y. P. & Fickel, J. Siberian tiger's recent population bottleneck in the Russian Far East revealed by microsatellite markers. *Mamm. Biol.* **76**, 722–726 (2011).
16. Henry, P. *et al.* In situ population structure and ex situ representation of the endangered Amur tiger. *Mol. Ecol.* **18**, 3173–3184 (2009).
17. Mooney, J. A. *et al.* Understanding the Hidden Complexity of Latin American Population Isolates. *Am. J. Hum. Genet.* **103**, 707–726 (2018).
18. Kardos, M. & Shafer, A. B. A. The Peril of Gene-Targeted Conservation. *Trends Ecol. Evol.* **33**, 827–839 (2018).
19. Miller, P. S. & Hedrick, P. W. MHC Polymorphism and the Design of Captive Breeding Programs: Simple Solutions Are Not the Answer. *Conserv. Biol.* **5**, 556–558 (1991).
20. Guo, J. Tigers in trouble: year of the tiger. *Nature* **449**, 16–18 (2007).
21. Lackey, L. B. Giraffe studbook Giraffa camelopardalis North American regional/global. *Association of Zoos and Aquariums and Disney's*.
22. Banes, G. L., Galdikas, B. M. F. & Vigilant, L. Reintroduction of confiscated and displaced mammals risks outbreeding and introgression in natural populations, as evidenced by orang-utans of divergent subspecies. *Sci. Rep.* **6**, 22026 (2016).
23. Pejaver, V. *et al.* Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *Am. J. Hum. Genet.* **109**, 2163–2177 (2022).

24. Liu, L. *et al.* Biological relevance of computationally predicted pathogenicity of noncoding variants. *Nat. Commun.* **10**, 330 (2019).
25. Li, H. *et al.* Exome variant discrepancies due to reference-genome differences. *Am. J. Hum. Genet.* **108**, 1239–1250 (2021).
26. Baym, M. *et al.* Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* **10**, e0128036 (2015).
27. Armstrong, E. E. *et al.* Genome Report: Chromosome-level draft assemblies of the snow leopard, African leopard, and tiger (Panthera uncia, Panthera pardus pardus, and Panthera tigris). *bioRxiv* 2022.04.26.489474 (2022) doi:10.1101/2022.04.26.489474.
28. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
29. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
30. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
31. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
32. Wasik, K. *et al.* Comparing low-pass sequencing and genotyping for trait mapping in pharmacogenetics. *BMC Genomics* **22**, 197 (2021).
33. Arthur, R., O'Connell, J., Schulz-Trieglaff, O. & Cox, A. J. Rapid genotype refinement for whole-genome sequencing data using multi-variate normal distributions. *Bioinformatics* **32**, 2306–2312 (2016).
34. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* vol. 27 2156–2158 Preprint at https://doi.org/10.1093/bioinformatics/btr330 (2011).
35. Szpiech, Z. A., Blant, A. & Pemberton, T. J. GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification. *Bioinformatics* **33**, 2059–2062 (2017).
36. Dimitromanolakis, A., Paterson, A. D. & Sun, L. Fast and Accurate Shared Segment Detection and Relatedness Estimation in Un-phased Genetic Data via TRUFFLE. *Am. J. Hum. Genet.* **105**, 78–88 (2019).
37. Nakatsuka, N. *et al.* The promise of discovering population-specific disease-associated genes in South Asia. *Nat. Genet.* **49**, 1403–1407 (2017).
38. Pontius, J. U. *et al.* Initial sequence and comparative analysis of the cat genome. *Genome Res.* **17**, 1675–1689 (2007).
39. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).
40. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
41. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).


A
Wild Subspecies
Generic
Ancestry Fraction
1.00
0.75
0.50
0.25
0.00
Amur
Bengal
Indochinese
Malayan
South China
Sumatran
North America
Asia
B
PC2 (2.31%)
PC1 (2.58%)
Top Ancestry
Amur
Bengal
Malayan
Sumatran
Indochinese
C
PC2 (2.31%)
PC1 (2.58%)
Region
Canada
Midwest
Northeast
Southeast
Southwest
West


A
Proportion of fixed sites segregating in generic tigers
0.5
0.4
0.3
0.2
0.1
Amur
Bengal
Indochinese
Malayan
Sumatran
All (Fixed in wild)


B
Oberserved Heterozygosity
0.0004
0.0003
0.0002
0.0001
Amur
Bengal
Indochinese
Malayan
South China
Sumatran
Generic