the geneticssociety

ARTICLE

Recapitulating whole genome based population genetic structure for Indian wild tigers through an ancestry informative marker panel

Anubhab Khan [1,3], Swathy M. Krishna[2], Uma Ramakrishnan[1] and Ranajit Das [2,3]

Identification of genetic structure within wildlife populations have implications in their conservation and management. Accurately inferring population genetic structure requires whole-genome data across the geographical range of the species, which can be resource-intensive. A cheaper strategy is to employ a subset of markers that can efficiently recapitulate the population genetic structure inferred by the whole genome data. Such ancestry informative markers (AIMs), have rarely been developed for endangered species such as tigers utilizing single nucleotide polymorphisms (SNPs). Here, we first identify the population structure of the Indian tiger using whole-genome sequences and then develop an AIMs panel with a minimum number of SNPs that can recapitulate this structure. We identified four population clusters of Indian tigers with North-East, North-West, and South Indian tigers forming three separate groups, and Terai and Central Indian tigers forming a single cluster. To evaluate the robustness of our AIMs, we applied it to a separate dataset of tigers from across India. Out of 92 SNPs present in our AIMs panel, 49 were present in the new dataset. These 49 SNPs were sufficient to recapitulate the population genetic structure obtained from the whole genome data. To the best of our knowledge, this is the first-ever SNP-based AIMs panel for big cats, which can be used as a cost-effective alternative to whole-genome sequencing for detecting the biogeographical origin of Indian tigers. Our study can be used as a guideline for developing an AIMs panel for the management of other endangered species where obtaining whole genome sequences are difficult.

Heredity (2022) 128:88–96; https://doi.org/10.1038/s41437-021-00477-y

INTRODUCTION

Assessing the population genetic structure of wild species is important for their management (Wultsch et al. 2016). The application includes population assignment (Das and Upadhyai 2019; Kunde et al. 2020), tracking illegal wildlife trade (Frantz et al. 2006), captive breeding programs, planning reintroductions, and informed management (Wasser et al. 2015; Miller et al. 2011; Friar et al. 2001; Laikre et al. 2005; Ballou 1992; Ivy and Lacy 2010; Putnam and Ivy 2014; Jiménez-Mena et al. 2016; Jangtarwan et al. 2019; Lott et al. 2020). Unravelling population structure involves the use of multiple markers from across the genome (Pritchard et al. 2000). Traditionally, upon the advent of Polymerase Chain Reaction (PCR) technology, genetic variation within wild populations used to be determined either employing a handful of neutral microsatellite markers or assessing mitochondrial DNA sequences (Chapman et al. 2009; Kirk and Freeland 2011). However, both aforementioned approaches had several challenges and inaccuracies predominantly due to the scarcity of the markers employed (Chapman et al. 2009). With the advancement in next-generation sequencing (NGS) techniques several thousand to millions of markers across the genome have become available for population genetic analysis, even for non-model organisms (Khan and Tyagi 2021). This has been greatly beneficial in enhancing the power and confidence in the determination of population genetic structure within wild populations (Fuentes-Pardo and Ruzzante 2017; Supple and Shapiro 2018). Sequencing of Restriction site-associated DNA (RAD) markers (RADseq) has made it increasingly possible to perform NGS without having a reference genome in hand (Catchen et al. 2013), however Feng et al. (2020) have analyzed whole-genome sequencing data of birds for studying the genomic features across a phylogeny without reference genomes albeit they had to use reference genomes from phylogenetically related species. Out of 13,505 non-model species, currently recognized as 'threatened' by the International Union for Conservation of Nature (IUCN), only around 1% have their genomes sequenced (Brandies et al. 2019). Consequently, resolving the population structure of most wild and endangered fauna with genome-wide SNP markers has been challenging. This is even more challenging for wide-ranging species like tigers, with populations distributed across countries.

Tigers (*Panthera tigris*) are endangered carnivores. About 60% of the world's wild tigers reside in India (Goodrich et al. 2015). However, it has been an arduous task to identify population genetic structure within the Indian tiger populations, predominantly due to incomplete sampling and the unavailability of genome-wide data. Three distinct genetic clusters of Indian tigers have been identified based on ~10,000 Single Nucleotide Polymorphism (SNPs) (Natesh

[1]National Centre for Biological Sciences, Bangalore, Karnataka, India. [2]Yenepoya Research Centre, Yenepoya (Deemed to be University), Mangalore, Karnataka, India. [3]These authors contributed equally: Anubhab Khan, Ranajit Das. Associate editor Xiangjiang Zhan. email: anubhabkhan@gmail.com; das.ranajit@gmail.com

Received: 29 October 2020 Revised: 30 September 2021 Accepted: 1 October 2021
Published online: 2 December 2021

et al. 2017) and 11 microsatellite markers (Kolipakam et al. 2019). Based on microsatellite analysis, Kolipakam et al. (2019) reported clustering of North-West Indian tigers and Terai tigers. This clustering is consistent with findings of Mondol et al. (2013) based on eight microsatellite markers, who surmised that Terai and North-West Indian tigers shared a common gene pool in the past. Further, Kolipakam et al. (2019) revealed a separate cluster for tigers from the North-East, which is supported by Armstrong et al. (2021), based on genome-wide data. Kolipakam et al. (2019) also observed clustering of tigers from the vicinities of the Western Ghats with the North-West and Terai tigers. However, this has not been supported by any other study. In contrast, Natesh et al. (2017), based on >10,000 SNP markers, unraveled a solitary cluster of Terai and Central Indian tigers, and two distinct, independent clusters of North-Western and South-Indian tigers. This grouping of Indian tigers is supported by Armstrong et al. (2021). Further, Natesh et al. (2017) observed that at high complexity values (K values), North-East Indian tigers form a separate cluster but failed to attain statistical support putatively due to the scarcity of genetic markers. Overall, the employment of only a handful of genetic markers in both SNP and microsatellite-based studies have occluded the identification of true genetic structure within tiger populations across India.

While whole-genome sequencing (WGS) can help to resolve the structure within Indian tiger populations by making a large number of loci available for analysis, it involves sequencing of multiple tiger genomes across its range, which can be both expensive, and time and labor-intensive (Fuentes-Pardo and Ruzzante 2017). A cost-effective approach can be genotyping individual tigers at specific SNP markers that show discernible population-specific allele frequency variation and thus can be more informative about population structure compared to other loci (Rosenberg et al. 2003; Shriver et al. 2003; Nassir et al. 2009; Das et al. 2019). Once such a panel is developed using the WGS of a few individuals from across the range, it would allow researchers to determine the ancestral and biogeographical affiliation of additional individuals without the need for their WGS making it cheaper and cost-effective. Such highly informative SNP markers called Ancestry Informative Markers (AIMs) can significantly aid in genetics-based conservation and management of wildlife populations. Previously, AIMs panels have been developed for species with well-characterized population genetic structure such as humans (Rosenberg et al. 2003; Vongpaisarnsin et al. 2015; Das and Upadhyai 2018; Esposito et al. 2018), honey bees (Muñoz et al. 2015), rhesus monkeys (Kanthaswamy et al. 2014), gorillas (Das et al. 2019), chimpanzees (Anjana et al. 2020) and domestic animals like cattle (Wilkinson et al. 2011), sheep (Somenzi et al. 2020) and pigs (Liang et al. 2019). Development of AIMs panel can be challenging for non-model organisms such as tigers that are largely devoid of the large-scale range-wide whole-genome dataset needed for accurately recapitulating the population structure.

In this study, we investigated the population genetic structure of Indian tigers with WGS based data and then developed the first SNP-based AIMs panel for tigers, also the first among the big cats. We aimed to minimize the number of loci needed to recapitulate the population structure of Indian tigers. Our AIMs panel can be used as a cost-effective alternative to WGS, for detecting population structure within Indian tigers. Such technologies can facilitate conservation and management of endangered species by tracking illegal wildlife trade, studying dispersal patterns, and estimating connectivity among others.

METHODS

Sampling

Tiger blood and tissue samples were collected and sequenced opportunistically, as described in Khan et al. (2021), from the major tiger landscapes in central India, Western ghats, Terai, north-east India, and north-west India. We also sampled a wild-caught individual from Nandankanan zoo in Odisha, India. Thus, our dataset is comprised of 35 wild tigers (Fig. 1).

We divided the samples into two datasets of 17 and 18 individuals as described further for developing and testing AIMs panel respectively.

SNP identification

We trimmed the raw sequencing reads using TRIMMOMATIC (Bolger et al. 2014) to have a mean PHRED-scaled quality of 30 in a sliding window of 15 bp, and any read that was shorter than 36 bp after trimming was removed from further analysis. We aligned these reads to a Bengal tiger reference genome assembly (NCBI accession: JAHFZI000000000) using BOWTIE2 (Langmead and Salzberg 2012). The alignments were then saved in a binary format (BAM) using SAMTOOLS v1.9 (Li et al. 2009). We marked duplicate reads with the Picard Tools 'MarkDuplicates' command (http://broadinstitute.github.io/picard). We called variants from the BAM files using Strelka with default options (Saunders et al. 2012). The variants were filtered with VCFtools v0.1.13 (Danecek et al. 2011) to retain biallelic sites with a minimum quality of 30, a genotype quality of 30, and a minimum depth of 10. We removed all indels and SNPs that are out of Hardy-Weinberg equilibrium (Chi-square test, p-value < 0.001) and required a minimum minor allele count of 3. We further, filtered out the SNPs with frequency <5% and missing genotype rate >30% employing –maf 0.05 and –geno 0.3 flags in PLINK v1.9 (Purcell et al. 2007).

Population genetic structure

For estimating population structure we sampled 17 individuals from central, south, north-west India, north-east India, and Terai regions such that each area had similar sample sizes. This was done to prevent potential bias in the results from poorly sampled areas in our dataset and to increase uniformity in our data. Samples from north-west India are overrepresented in our analysis while those from north-east India are underrepresented. We randomly chose 2–4 individuals from each location to avoid any statistical bias. To re-iterate, we employed genomic data of 17 wild tigers, comprising of 2–4 individuals from each geographic landscape (Armstrong et al. 2021) (Fig. 1) assessing 2,828,619 SNP markers.

Principal Component Analysis (PCA) was performed with the SNPs using –pca function in PLINK v1.9. Plots were generated with principal components (PCs) 1 and 2 taken as x and y coordinates respectively. We applied the model-based unsupervised clustering methods implemented in ADMIXTURE v1.3 (Alexander et al. 2009) to determine the ancestry proportions of the tiger genomes, exploring from K = 2 to K = 7. We plotted the ancestral fractions from the Q file of ADMIXTURE. We estimated genome-wide genetic distance using F_{ST} function as implemented in the program VCFtools v0.1.13.

AIMs determination

To identify ancestry-specific SNP markers for the tiger populations under study, we employed four AIMs determining approaches (1) Infocalc: The algorithm finds the informativeness of multiallelic SNPs in determining the ancestry of an individual based on the allele frequencies in the populations (Rosenberg et al. 2003). Files compatible with Infocalc v1.1 were created using –recode-structure function implemented in PLINK v1.9. The output file was sorted based on the informativeness defining column (I_n) and the top 10,000 ranking SNPs were selected (N = 24, 898). (2) Admixture: The SNP allele frequencies of the ancestral population as obtained from P file output of ADMIXTURE v1.3 was used to determine the AIMs panels (Alexander et al. 2009). We identified the top 10,000 ranking SNPs with the highest column to column variance depicted by the P file (N = 10,000). (3) Wright's F_{ST}: This algorithm measures the degree of differentiation among populations based on the genetic structure of the populations (Wright 1969). Employing –fst function in PLINK v1.9, with family ID as the indicator of the population group, F_{ST} scores were calculated for all 28,28,619 SNPs. Top 10,000 ranking SNPs with highest F_{ST} value were selected (N = 10,043). (4) SmartPCA: SmartPCA finds the weighting of all SNPs for each principal component (PC). SmartPCA algorithm is implemented in EIGENSOFT v7.2.1 (Patterson et al. 2006; Price et al. 2006). The output file depicting the weightage of each SNP is obtained as a *snpwt* file. Top 10,000 ranking SNPs with the highest weight for PC1 were selected (N = 10,011).

Among the aforementioned four approaches, the optimal AIMs determining strategy was assessed by comparing the 10,000 SNPs datasets qualitatively and quantitatively to the Complete SNP Set (CSS). The qualitative comparison was performed employing ADMIXTURE and PCA. For quantitative comparisons, we calculated the Euclidean distance (ED)

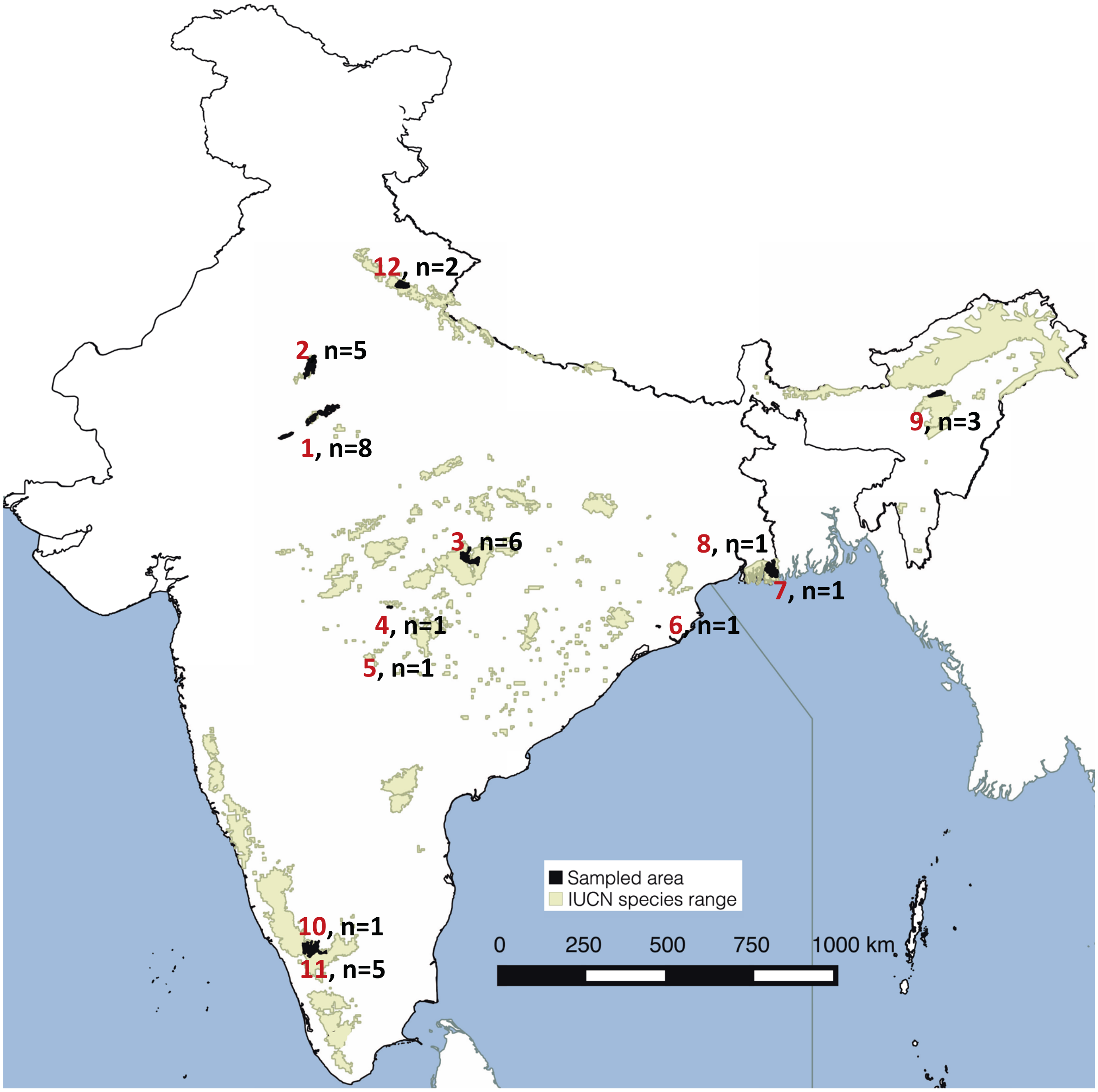


Fig. 1 Sample locations of tiger from across India. The numbers in the map represent the geographical location of the sample: 1 = Ranthambore Tiger Reserve, 2 = Sariska Tiger Reserve, 3 = Kanha Tiger Reserve, 4 = Bor Tiger Reserve, 5 = Chandrapur, 6 = Nandankanan, 7 = Sunderban Tiger Reserve, 8 = Lalgarh Forest Range, 9 = Kaziranga Tiger Reserve, 10 = Bandipur Tiger Reserve, 11 = Wayanad Wildlife Sanctuary, 12 = Corbett Tiger Reserve.

between the four putative ancestral components (North-East (NE), Central (CI), North-West (NW) and South (SI)) of all data subsets and the CSS following the formula mentioned below:

$$ED = \begin{pmatrix} (NE_{CSS} - NE_{subset})^2 + (CI_{CSS} - CI_{subset})^2 \\ + (NW_{CSS} - NW_{subset})^2 + (SI_{CSS} - SI_{subset})^2 \end{pmatrix}^{0.5}$$

The approach(es) that generated the minimum Euclidean distance (ED → 0) between the ancestry fractions of CSS and the subsets, was/were considered to be the best AIMs determining strategy.

Consensus AIMs

To find a consensus between the four AIMs determining approaches, a Venn diagram was plotted (http://bioinformatics.psb.ugent.be/webtools/Venn/) between the four datasets generated by Infocalc, ADMIXTURE, F_{ST} and SmartPCA, and common loci were selected based on the most accurate approaches.

Testing AIMs

To assess whether our AIMs panel performs statistically significantly better than any equal-sized randomly generated SNP sets, we developed 100 SNP panels, each comprising of equinumerous randomly generated SNPs ($N = 92$) from the CSS and clustered individuals using ADMIXTURE. Subsequently, the coefficients of determination (r^2), as implemented in GraphPad Prism v9, was calculated between the four ancestry fractions, namely South India, North West Indian, North East Indian, and Central of the CSS, and the AIMs panels and the random panels. We further evaluated the efficiency of the ancestry assignments obtained using our AIMs panel ($N = 92$) considering the null hypothesis that similar ancestry fractions can be obtained using any equal-sized SNP sets, chosen at random ($p < 0.05$). To note, for the random sets, the mean ancestry fraction of the 100 random panels were employed for each individual.

To assess the efficiency, versatility, and replicability of our AIMs panel, we used SNPs from a dataset of 18 individuals from various geographic locations across India (Fig. 1) that for not involved in the panel development. Subsequently, we estimated the structure within the test populations employing the complete set of SNPs and also our AIMs panel. All PCA and ADMIXTURE plots were generated as before using R v3.5.1.

RESULTS

Population genetic structure of Indian tigers

PCA. The first two principal components explained 25% of the variance between them. Along PC1 the populations depicted an east to west cline. The Central Indian tigers from Kanha Tiger Reserve and Chandrapur formed a cluster at the centre with the tigers from Terai (Corbett Tiger Reserve), Lalgarh, and Nandankanan Zoo. Tigers from the vicinities of Western Ghats (Wayanad Wildlife Santuary), North-East Indian tigers (Kaziranga Tiger Reserve), and North-West Indian tigers (Ranthambore Tiger Reserve) formed separate and distinct clusters (Fig. 2).

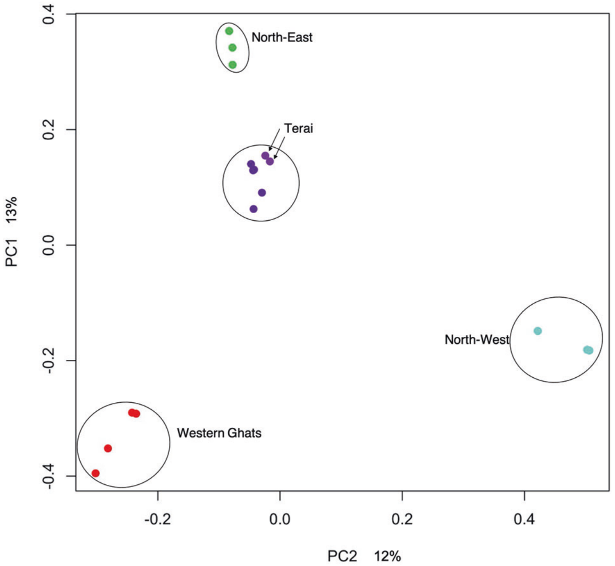


Fig. 2 PCA plot of Indian tigers based on complete SNP set (CSS). The percentage value on each axis is the amount of variance explained by that axis.

Admixture. We aimed at discerning the optimum number of ancestral components (K) by minimizing the cross-validation error (CVE; Alexander et al. 2009) implemented in ADMIXTURE v1.3 using the –cv flag in the ADMIXTURE command line. However, we were unable to find a single K when the genomes exhibited homogeneous admixture patterns. For instance, admixture analysis revealed that at complexity (K) values of 2 and 3, while the North-West Indian and Western Ghats tigers formed separate clusters, no distinct cluster was formed for the North-East tigers, which grouped with the Central tigers. The most robust population structure was found for the K value of 4, where the North-East tigers form a separate cluster from the rest of the Indian tigers (Fig. 3). We note here that K values more than 4 do not yield biologically meaningful results. Overall, our PCA and admixture analysis with tiger WGS data across India indicated four population clusters for Indian tigers.

F_{ST}. We obtained the lowest F_{ST} between the Terai and Central Indian tigers, and the highest between the Terai and North-West Indian tigers (Balloux and Lugon-Moulin 2002) (Fig. 4). It is noteworthy that North-West tigers in general have higher F_{ST} with tigers from other populations.

AIMs Panel

We employed both qualitative (Admixture and PCA plots, Supplemental Fig. 1a, b respectively) and quantitative (Fig. 5) measures to assess whether the putative AIMs panels identified by INFOCALC, ADMIXTURE, SmartPCA, and F_{ST} are able to recapitulate

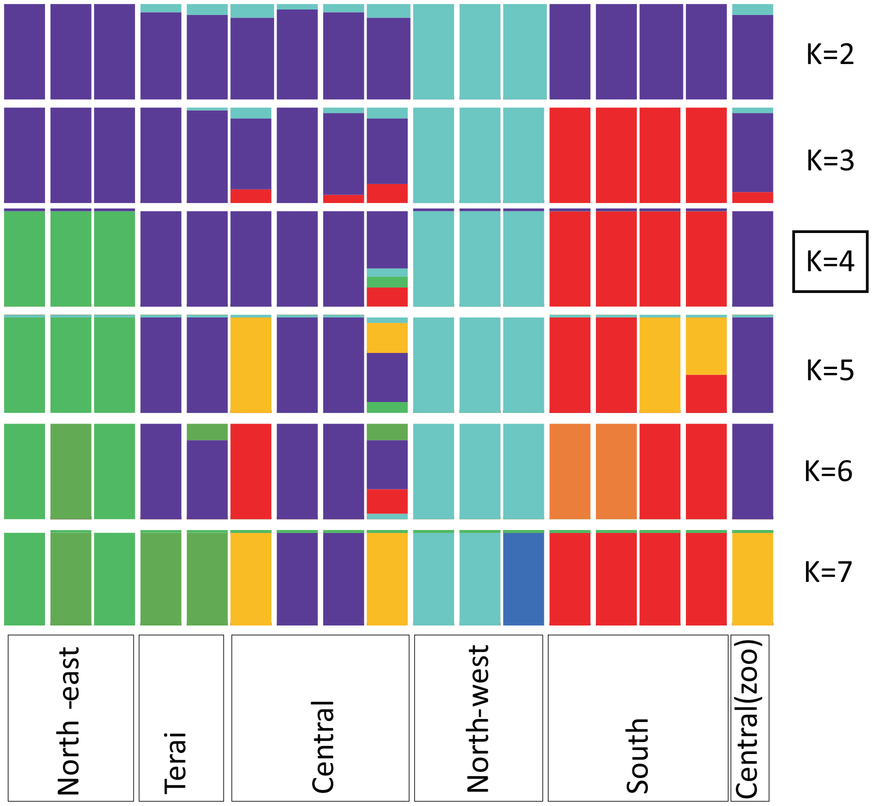
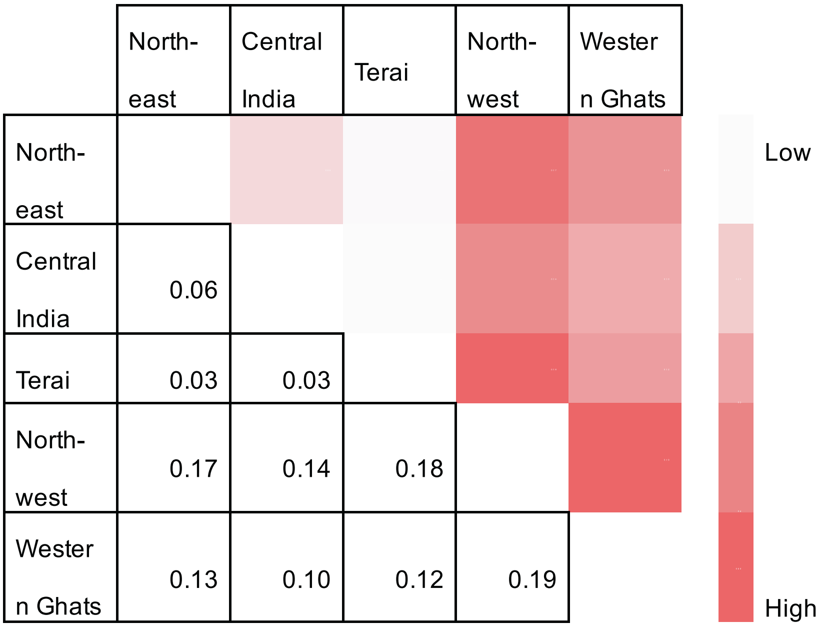


Fig. 3 Comparison of Admixture plots K = 2–7. Plot for K = 4, putatively optimum K, is highlighted with a box.



	North-east	Central India	Terai	North-west	Western Ghats
North-east					
Central India	0.06				
Terai	0.03	0.03			
North-west	0.17	0.14	0.18		
Western Ghats	0.13	0.10	0.12	0.19	

Fig. 4 Population differentiation. Pairwise F_{ST} between tigers from various geographic regions.

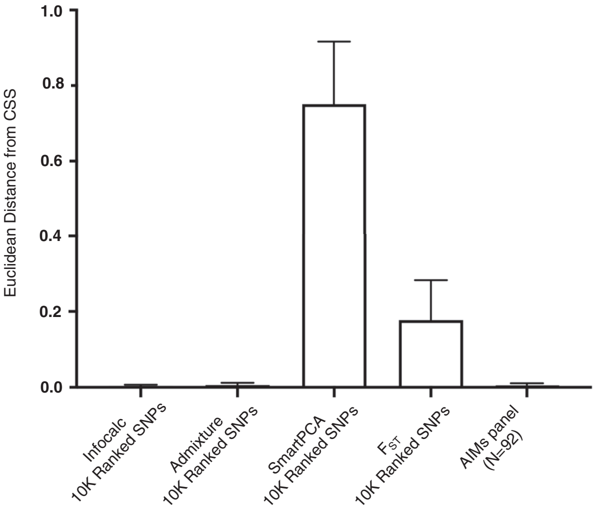


Fig. 5 Quantitative comparison of four AIMs determination strategies: Infocalc-based, ADMIXTURE-based, SmartPCA-based and F_{ST}-based, and the AIMs panel. 10 K Ranked SNPs indicated the top 10,000 candidate SNPs selected by the method.

the population genetic structure of Indian tigers depicted by the Complete SNP set (CSS).

While the SmartPCA and F_{ST} based approaches largely failed to recapitulate the population genetic structure depicted by CSS, both qualitatively and quantitatively, ADMIXTURE and INFOCALC based strategies performed the best, replicating the population genetic structure depicted by the CSS with the highest resolution (Supplemental Figs. 1 and 5). All four datasets depicted a high number of private alleles with Infocalc, Admixture, F_{ST} and SmartPCA having 24,668, 6573, 6541, and 8663 private SNPs, respectively. Seven SNPs were found to be common among the four methods and only 678 were common to at least three of the four approaches. Because of the failure of F_{ST} and SmartPCA based approaches, we ignored the SNPs identified by these two approaches and developed our AIMs panel using the SNPs that were detected by both Infocalc and Admixture-based approaches ($N = 92$).

We subsequently compared our AIMs panel ($N = 92$) qualitatively using Admixture (Fig. 6), PCA (Supplemental Fig. 2), and quantitatively (Fig. 5) to the CSS. The AIMs panel could recapitulate the population genetic structure depicted by the CSS with a high resolution both qualitatively and quantitatively and was found to be statistically similar to the 10,000 Infocalc and Admixture datasets (Tukey's post hoc analysis, $P = 0.99$).

We compared our AIMs panel with 100 panels of randomly sampled 92 SNPs. As envisioned, all 100 SNP panels failed to recapitulate the population genetic structure depicted by the CSS with any precision (20 panels are shown in Supplemental Figure 1). The only population genetic structure that could be identified by most random panels was North-West.

We further calculated the assignment proportion for each individual in the random sets to investigate the proportion of times (%) they were correctly assigned to the four major assignment zones (Central, South, Northeast, or North-West) defined by the CSS. Consistent with the admixture plots (Supplemental Fig. 3), randomly generated SNPs could successfully assign tigers from North-West and North-East to their actual zones of origin >85 and >55% times respectively, but completely failed to do so for Central and Southern tigers (Fig. 7).

The coefficient of determination (r^2) computed between the ancestry fractions of the CSS and both AIMs panels indicated a strong association between the CSS and both AIMs panels for all ancestry assignments ($r^2 = 0.95–1$, $P < 0.0001$) (Table 1). As predicated, while r^2 was significantly high for the North West ancestry assignment between CSS and the random panels ($r^2 = 0.97$, $P < 0.0001$), the same for the Central Indian assignment was discernibly low ($r^2 = 0.03$, $P = 0.49$) (Table 1). The null hypothesis that similar ancestry fractions can be obtained using equal-sized SNP panels, was rejected since the AIMs

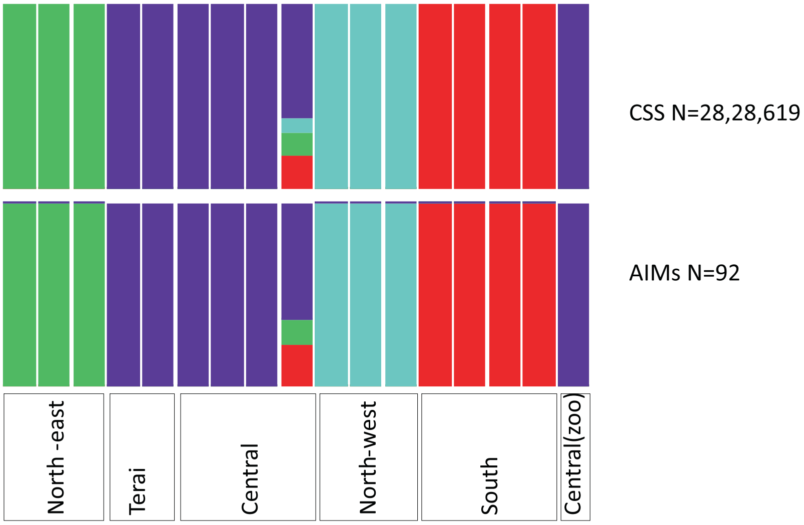


Fig. 6 Structure from whole genome vs AIMs. Qualitative comparison of the admixture plots generated by CSS and the AIMs panel.

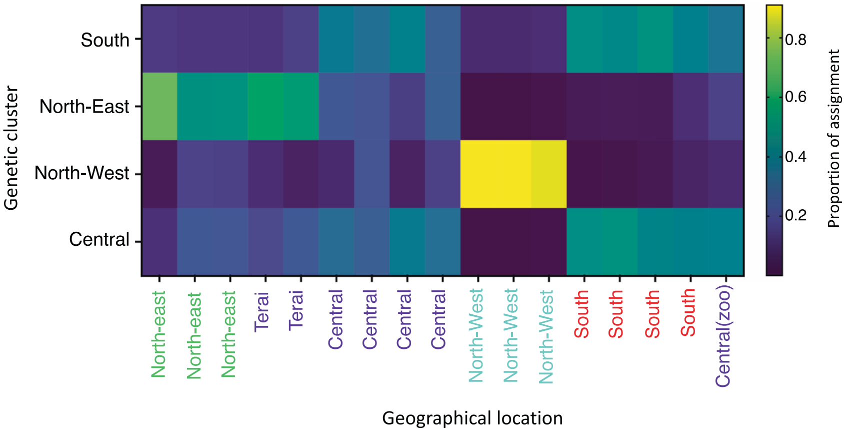


Fig. 7 Assignment proportions of individual tigers to their respective zones based on 92 random SNPs. The bars indicate the proportion of times an individual tiger was assigned to one the four major assignment zones (Central, South, North-east or North-West).

Table 1. Coefficient of determination (r^2) computed for various ancestry assignments.

Ancestry Assignment	CSS vs. AIMs panel ($N = 92$)	CSS vs. AIMs panel ($N = 49$)	CSS vs. Random panels ($N = 92$)	AIMs panel ($N = 92$) vs. Random panels ($N = 92$)
Central India	0.9999	0.9473	0.0319	0.0322
South India	0.9992	0.926	0.5371	0.5378
North West India	0.9979	0.9856	0.9745	0.9719
North East India	1	NA	0.4426	0.4429

panel performed significantly better than the random panels ($P < 0.0001$).

To assess the robustness of our AIMs panel in recapitulating the ancestry information of Indian tigers, we applied our AIMs panel to a separate whole-genome dataset of 18 tigers assessing 1,538,042 SNPs. Out of 92 SNPs present in our AIMs panel, 49 were present in the new dataset. Interestingly these 49 SNPs could efficiently recapitulate the population genetic structure within the Indian tiger populations assessed indicating these 49 SNPs might be sufficient to identify the biogeographical affinity of an individual tiger (Fig. 8).

We further performed power analyses for both AIMs panels ($N = 92$ and $N = 49$). Power analyses was performed using G*Power v3.1. The power ($1-β$) was calculated from the Coefficient of determination (r^2) given $α = 0.01$ and sample sizes of 17 and 18 respectively for the two AIMs panels. The average r^2 for the larger AIMs panel ($N = 92$) was 0.99 (Table 1) and as a consequence the effect size ($|r|$) was estimated to be 0.995. The average r^2 for the smaller AIMs panel ($N = 49$) was 0.96 (Table 1) and as a consequence the effect size ($|r|$) was estimated to be 0.979. for both AIMs panels we obtained absolute statistical powers ($1 - β) = 1$, indicating the statistical efficiency of both

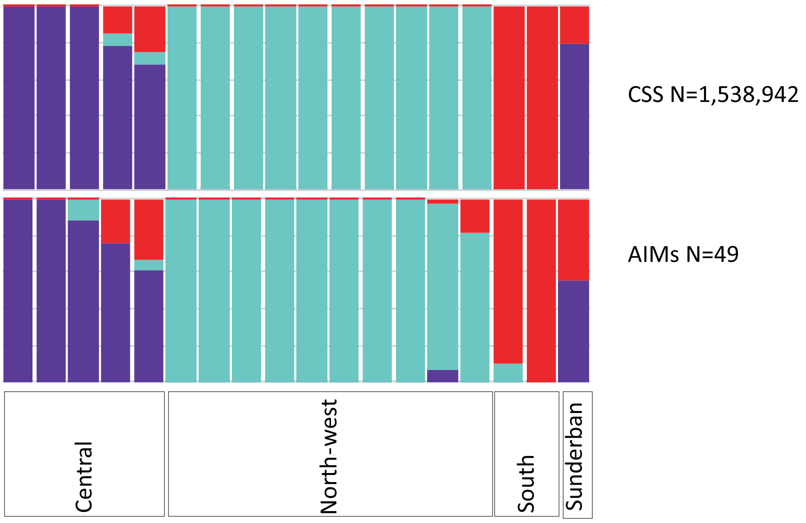


Fig. 8 Testing structure from developed AIMs panel. CSS vs. AIMs panel comparison for an independent dataset.

AIMs panels in replicating the ancestry information of the CSS. The statistical power of the random panels varied greatly among the four ancestry assignments according to r^2 (Table 1). While the power was only 0.03 for the Central Indian tigers (effect size (|r | = 0.179)), it was discernibly high for the tigers from the North East (effect size (|r | = 0.665), power = 0.75) and the South (effect size (| r | = 0.733), power = 0.91), and found to be absolute (effect size (| r | = 0.987), power = 1) for the North-West tigers.

DISCUSSION

Population genetic structure of Indian tigers

Tiger whole-genome data employed in this study depicted four distinct population groups of Indian tigers. Both clustering algorithms (PCA and ADMIXTURE) and the measurements of F_{ST} did not show any support for considering Terai and Central Indian tigers as separate clusters. Our analysis of cross-validation error (CVE) for ADMIXTURE remains inconclusive. Recently, biologically meaningful complexity values have been favoured over ones suggested by CVE (for example vervet monkeys (Svardal et al. 2017) and humans (Esposito et al. 2018)). We note here that our results are consistent with the previous SNP-based study by Natesh et al. (2017) and the patterns observed in Armstrong et al. (2021) with less extensive sampling of Indian tigers. Further, our results are partially consistent with Kolipakam et al. (2019) such that both found a separate cluster for North-East Indian tigers. However, unlike Kolipakam et al. (2019), we did not find any evidence of Terai and Central Indian tigers being separate clusters, and Western-Ghats, Terai and North-West tigers share the same gene pool (Kolipakam et al. 2019; Mondol et al. 2013). It is noteworthy that even at higher K values, ADMIXTURE plots could not identify separate clusters for the Terai tigers, and they were found to be virtually indistinguishable from the Central Indian tigers throughout the analysis. Both PCA and ADMIXTURE analysis extended support for four cluster model of Indian tigers, depicting separate clusters for the South, North-West, North-East tigers, and a combined Terai and Central Indian cluster.

Ancestry Informative Markers

We developed a set of 92 AIMs that can optimally replicate the population genetic structure of Indian tigers. However, individuals belonging to small and isolated populations such as those from North-West India (Alcala et al. 2019; Natesh et al. 2017; Khan et al. 2021), can mostly get correctly assigned using any random subset of loci (Fig. 7, Supplemental Fig. 3). Such populations can be presumed to have multiple loci with fixed alleles as a consequence of genetic drift and inbreeding across multiple generations (Dudash and Fenster 2000; Schlaepfer et al. 2018). Therefore, we suggest that these populations can be easily identified as separate clusters due to high differentiation from other populations. In contrast, individuals belonging to recently isolated populations with fairly high population sizes or incomplete lineage sorting can be very difficult to identify as separate clusters due to low differentiation among each other, for example, individuals from Sunderban Tiger Reserve (Fig. 8). We surmise here that while the AIMs panel might not be robust enough to detect fine and subtle population structure among tigers belonging to the same geographical landscape (i.e. individual-level variation) due to the presence of only a handful of markers, it can detect the major ancestral components in each individual, significantly similar to the CSS. Thus, the AIMs panel can be successfully utilized to detect inter-population variation based on the variation in these ancestral components. For example, the presence of South Indian and Central Indian ancestry components in the lone Sundarban tiger, employed in this study, made it unique and genetically distinguishable from other tigers (Fig. 8). Similarly, we could detect discernible variation among individuals of Central Indian genetic cluster albeit from different geographical locations (Figs. 6 and 8). To summarize, similar to the CSS, our AIMs panel can detect structure among populations.

We have obtained tiger samples from several geographical locations across India and identify the broad population genetic structure of Indian tigers. Bengal tigers residing outside India remain under-sampled in literature for genomics scale studies. These populations need to be sampled and sequenced for their conservation. It might be useful to assign individuals to the specific protected area or landscape they belong to for effective conservation measures. However, detection of these subtle genetic structure depends on extensive sampling of individuals from each of these protected areas and building a whole-genome dataset to discover fine-scale structures. This whole-genome data can subsequently be used to develop AIMs panels. It might be expected that the minimum number of ancestry informative SNPs required for the identification of fine-scale population genetic structure within a landscape consisting of several protected areas might be much higher. Conversely, if there are certain mutations that are fixed in a landscape due to drift or selection, they can be used effectively to assign individuals. Examples could include locally high-frequency alleles like that for the pseudo-melanism in Similipal Tiger Reserve (Sagar et al. 2021).

Out of the 92 SNPs present in our AIMs panel, we could apply 49 SNPs in the test dataset due to high missingness in the other loci post quality filtering. These 49 AIMs sufficed to accurately model the population genetic structure of the test set. However, our test dataset did not have individuals from North-Eastern India due to a deficiency in sampling. Whether these 49 AIMs will suffice to model the population genetic structure of Indian tigers with North-East Indian tigers remains to be evaluated. Additional testing could be conducted using large sample sizes and better geographically distributed non-invasive DNA samples. Further, as Northeast individuals do show signatures of possible admixture with individuals outside India (Armstrong et al. 2021), testing our AIMs on global tiger datasets will be important.

The AIMs panel for Indian tigers that we developed in this study can be added to the set of 126 SNPs panel, currently in use for tigers (Natesh et al. 2019). These AIMs will increase the accuracy with which Indian tigers can be assigned to a geographical area while keeping the costs low. This will aid tremendously in forensics, keeping strict tabs on wildlife trade, studying individual dispersals from one genetic cluster to another, and management of populations.

Big cats are charismatic and are of global conservation concern. They are also trafficked for various illegal activities. Such traffic can be monitored using genetic tools, such as population assignment (as done for elephants for example (Wasser et al. 2015)). Discerning population genetic structure may be difficult in species with large geographic distributions, but our AIMs panel can be used effectively in such cases to identify the origin of a sample down to their most likely protected areas. Such information can facilitate the protection and management of trade. Further, zoo-bred individuals have often been found to have admixed ancestries (Luo et al. 2008). Our AIMs panel can aid in discerning the various ancestry proportions of the zoo individuals. Discerning ancestry for captive tigers through AIMs can help to avoid breeding between individuals with the similar genetic tapestry. Such planned breeding programs can be monumental in both preserving the genetic integrity of the captive animals and increase in their genetic variation through a planned admixture of animals of different ancestral make-up.

We suggest that a denser sampling of individuals from various protected areas across India will allow us to develop a more versatile AIMs panel that will be able to assign individuals even closer to its true biogeographical origin. Studies such as the one presented here are needed for all big cats across their range for providing viable genetics-based management solutions.

Data archiving

Sequencing data can be accessed from BioProject accession numbers PRJNA728665, PRJNA693788, PRJNA559670, PRJNA749163 in NCBI.

REFERENCES

Alcala N, Goldberg A, Ramakrishnan U, Rosenberg NA (2019) Coalescent theory of migration network motifs. Mol Biol Evol 36:2358–2374

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664

Anjana S, Sammeta SP, Das R (2020) Developing ancestry informative marker panel for Nigeria-Cameroonian chimpanzees. J Genet 99:1–7

Armstrong E, Khan A, Taylor RW, Gouy A, Greenbaum G, Thiéry A et al. (2021) Recent evolutionary history of tigers highlights contrasting roles of genetic drift and selection. Mol Biol Evol 38:2366–2379

Ballou JD (1992) Genetic and demographic considerations in endangered species captive breeding and reintroduction programs. In: Wildlife 2001: populations, Springer: Dordrecht, pp. 262–275

Balloux F, Lugon-Moulin N (2002) The estimation of population differentiation with microsatellite markers. Mol Ecol 11:155–165

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Brandies P, Peel E, Hogg CJ, Belov K (2019) The value of reference genomes in the conservation of threatened species. Genes (Basel) 10:846

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140

Chapman JR, Nakagawa S, Coltman DW, Slate J, Sheldon BC (2009) A quantitative review of heterozygosity–fitness correlations in animal populations. Mol Ecol 18:2746–2765

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA et al. (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Das R, Roy R, Venkatesh N (2019) Using ancestry informative markers (AIMs) to detect fine structure within gorilla populations. Front Genet 10:43

Das R, Upadhyai P (2018) An ancestry informative marker set which recapitulates the known fine structure of populations in South Asia. Genome Biol Evol 10:2408–2416

Das R, Upadhyai P (2019) Application of the geographic population genetic structure (GPS) algorithm for biogeographical analyses of wild and captive gorillas. BMC Bioinforma 20:35

Dudash MR, Fenster CB (2000) Inbreeding and outbreeding depression in fragmented populations. Conserv Biol Ser 35–54

Esposito U, Das R, Syed S, Pirooznia M, Elhaik E (2018) Ancient ancestry informative markers for identifying fine-scale ancient population genetic structure in Eurasians. Genes (Basel) 9:625

Feng S, Stiller J, Deng Y, Armstrong J, Fang Q, Reeve AH et al. (2020) Dense sampling of bird diversity increases power of comparative genomics. Nature 587:252–257

Frantz AC, Pourtois JT, Heuertz M, Schley L, Flamand MC, Krier A et al. (2006) Genetic structure and assignment tests demonstrate illegal translocation of red deer (Cervus elaphus) into a continuous population. Mol Ecol 15:3191–3203

Friar EA, Boose DL, LaDoux T, Roalson EH, Robichaux RH (2001) Population genetic structure in the endangered Mauna Loa silversword, Argyroxiphium kauense (Asteraceae), and its bearing on reintroduction. Mol Ecol 10:1657–1663

Fuentes-Pardo AP, Ruzzante DE (2017) Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol Ecol 26:5369–5406

Goodrich J, Lynam A, Miquelle D, Wibisono H, Kawanishi K, Pattanavibool A et al. (2015) Panthera tigris. In: The IUCN Red List of Threatened Species 2015, p. 15955 50659951

Ivy JA, Lacy RC (2010) Using molecular methods to improve the genetic management of captive breeding programs for threatened species. In: Molecular approaches in natural resource conservation and management, pp. 267–295

Jangtarwan K, Koomgun T, Prasongmaneerut T, Thongchum R, Singchat W, Tawichasri P et al. (2019) Take one step backward to move forward: Assessment of genetic diversity and population structure of captive Asian woolly-necked storks (Ciconia episcopus). PLoS One 14:e0223726

Jiménez-Mena B, Schad K, Hanna N, Lacy RC (2016) Pedigree analysis for the genetic management of group-living species. Ecol Evol 6:3067–3078

Kanthaswamy S, Johnson Z, Trask JS, Smith DG, Ramakrishnan R, Bahk J et al. (2014) Development and validation of a SNP-based assay for inferring the genetic ancestry of rhesus macaques (Macaca mulatta). Am J Primatol 76:1105–1113

Khan A, Patel K, Bhattacharjee S, Sharma S, Chugani AN, Sivaraman K et al. (2020) Are shed hair genomes the most effective noninvasive resource for estimating relationships in the wild? Ecol Evol 10:4583–4594

Khan A, Patel K, Shukla H, Viswanathan A, van der Valk T, Borthakur U, et al. (2021). Genomic evidence for inbreeding depression and purging of deleterious genetic variation in Indian tigers. bioRxiv

Khan A, Tyagi A (2021) Considerations for initiating a wildlife genomics research project in South and South-East Asia. J Indian Inst Sci 101:243–256. https://doi.org/10.1007/s41745-021-00243-3

Kirk H, Freeland JR (2011) Applications and implications of neutral versus non-neutral markers in molecular ecology. Int J Mol Sci 12:3966–3988

Kolipakam V, Singh S, Pant B, Qureshi Q, Jhala YV (2019) Genetic structure of tigers (Panthera tigris tigris) in India and its implications for conservation. Glob. Ecol Conserv 20:710

Kunde MN, Martins RF, Premier J, Fickel J, Förster DW (2020) Population and landscape genetic analysis of the Malayan sun bear Helarctos malayanus. Conserv Genet 21:123–135

Laikre L, Palm S, Ryman N (2005) Genetic population genetic structure of fishes: implications for coastal zone management. AMBIO A J Hum Environ 34:111–119

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357

Liang Z, Bu L, Qin Y, Peng Y, Yang R, Zhao Y (2019) Selection of optimal ancestry informative markersfor classification and ancestry proportion estimation in pigs. Front genet 10:183

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079

Lott MJ, Wright BR, Kemp LF, Johnson RN, Hogg CJ (2020) Genetic Management of Captive and Reintroduced Bilby Populations. J Wildl Manag 84:20–32

Luo SJ, Johnson WE, Martenson J, Antunes A, Martelli P, Uphyrkina O et al. (2008) Subspecies genetic assignments of worldwide captive tigers increase conservation value of captive populations. Curr Biol 18:592–596

Miller W, Hayes VM, Ratan A, Petersen DC, Wittekindt NE, Miller J et al. (2011) Genetic diversity and population structure of the endangered marsupial Sarcophilus harrisii (Tasmanian devil) Proc Natl Acad Sci 108:12348–12353

Mondol S, Bruford MW, Ramakrishnan U (2013) Demographic loss, genetic structure and the conservation implications for Indian tigers. Proc R Soc B Biol Sci 280:20130496

Muñoz I, Henriques D, Johnston JS, Chávez-Galarza J, Kryger P, Pinto MA (2015) Reduced SNP Panels for Genetic Identification and Introgression Analysis in the Dark Honey Bee (Apis mellifera mellifera) PLOS One 10:e0124365

Nassir R, Kosoy R, Tian C, White PA, Butler LM, Silva G et al. (2009) An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. BMC Genet 10:39

Natesh M, Atla G, Nigam P, Jhala YV, Zachariah A, Borthakur U et al. (2017) Conservation priorities for endangered Indian tigers through a genomic lens. Sci Rep. 7:1–11

Natesh M, Taylor RW, Truelove NK, Hadly EA, Palumbi SR, Petrov DA et al. (2019) Empowering conservation practice with efficient and economical genotyping from poor quality samples. Methods Ecol evolution 10:853–859

Patterson N, Price AL, Reich D (2006) Population genetic structure and eigenanalysis. PLoS Genet 2:190

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population genetic structure using multilocus genotype data. Genetics 155:945–959

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575

Putnam AS, Ivy JA (2014) Kinship-based management strategies for captive breeding programs when pedigrees are unknown or uncertain. J Hered 105:303–311

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402–1422

Sagar V, Kaelin CB, Natesh M, Reddy PA, Mohapatra RK, Chhattani H et al. (2021) High frequency of an otherwise rare phenotype in a small and isolatedtiger population. Proc Natl Acad Sci p. 118

Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK (2012) Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics 28:1811–1817

Schlaepfer DR, Braschler B, Rusterholz HP, Baur B (2018) Genetic effects of anthropogenic habitat fragmentation on remnant animal and plant populations: a meta-analysis. Ecosphere 9:2488

Shriver MD, Parra EJ, Dios S, Bonilla C, Norton H, Jovel C et al. (2003) Skin pigmentation, biogeographical ancestry and admixture mapping. Hum Genet 112:387–399

Somenzi E, Ajmone-Marsan P, Barbato M (2020) Identification of ancestry informative marker (AIM) panels to assess hybridisation between feral and domestic sheep. Animals 10:582

Supple MA, Shapiro B (2018) Conservation of biodiversity in the genomics era. Genome Biol 19:1–12

Svardal H, Jasinska AJ, Apetrei C, Coppola G, Huang Y, Schmitt CA et al. (2017) Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. Nat genet 49:1705–1713

Vongpaisarnsin K, Listman JB, Malison RT, Gelernter J (2015) Ancestry informative markers for distinguishing between Thai populations based on genome-wide association datasets. Leg Med 17:245–250

Wasser SK, Brown L, Mailand C, Mondol S, Clark W, Laurie C et al. (2015) Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots. Science (80-) 349:84–87

Wilkinson S, Wiener P, Archibald AL, Law A, Schnabel RD, McKay SD et al. (2011) Evaluation of approaches for identifying population informative markers from high density SNP chips. BMC genetics 12:1–14

Wright S (1969) Evolution and the genetics of populations

Wultsch C, Caragiulo A, Dias-Freedman I, Quigley H, Rabinowitz S, Amato G (2016) Genetic diversity and population genetic structure of Mesoamerican jaguars (Panthera onca): implications for conservation and management. PLoS One 11:162377

ACKNOWLEDGEMENTS

Gratitude to NCBS IT team for support during data analysis on the NCBS cluster. NCBS data cluster used is supported under project no. 12-R&D-TFR-5.04-0900, Department of Atomic Energy, Government of India.

AUTHOR CONTRIBUTIONS

AK, URK, and RD conceived the project and designed the study. AK, SK, and RD performed the analysis. AK, URK, and RD wrote the paper. All authors have read and approved the paper.

CONFLICT OF INTEREST

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41437-021-00477-y.

Correspondence and requests for materials should be addressed to Anubhab Khan or Ranajit Das.

Reprints and permission information is available at http://www.nature.com/reprints