Journal of Genetics (2020)99:39
https://doi.org/10.1007/s12041-020-01192-z

RESEARCH ARTICLE

Developing ancestry informative marker panel for Nigeria-Cameroonian chimpanzees

S. ANJANA[1], SAI PRADIVYA SAMMETA[1] and RANAJIT DAS[2]*

[1]*Department of Systems and Computational Biology, School of Life Sciences, University of Hyderabad, Hyderabad 500 046, India*
[2]*Yenepoya Research Centre, Yenepoya (Deemed to be University), Mangalore 575 018, India*
*For correspondence. E-mail: das.ranajit@gmail.com.

Received 11 December 2019; revised 1 February 2020; accepted 17 February 2020

Abstract. Chimpanzees (*Pan troglodytes*), with a dwindling population size, are distributed across sub-Saharan Africa. They are classified into two biogeographical clusters comprising of four subspecies: a western African cluster that includes *P. t. verus* and *P. t. ellioti* and a central/eastern African cluster that includes *P. t. troglodytes* and *P. t. schweinfurthii*. While the genetic distinctness of Nigeria-Cameroonian chimpanzees (*P. t. ellioti*) from western chimpanzees has been known for a while, the fine structures within *P. t. ellioti* population has remained under-studied. In this study, we developed the first ever ancestry informative marker (AIMs) panel that can detect the fine population structure within Nigeria-Cameroonian chimpanzees with high resolution. We compared four commonly used AIMs-determining strategies, namely Infocalc algorithm, Wright's F_{ST}, smart principal component analysis (SmartPCA) and ADMIXTURE to first identify the best approach and then developed an AIMs panel of 435 SNPs employing the consensus of the four approaches (n = 129), with additional supplements from the best two approaches (Infocalc and ADMIXTURE). To the best of our knowledge, we have developed the first-ever AIMs panel for chimpanzees, which can greatly aid in their planned reintroduction to the natural habitat, maintaining their genetic integrity through planned captive breeding, and in tracking illegal trading across the globe.

Keywords. Nigeria-Cameroonian chimpanzee; ancestry informative marker panel; chimpanzee ancestry; Infocalc; ADMIXTURE.

Introduction

Chimpanzees (*Pan troglodytes*) are distributed discontinuously across sub-Saharan Africa from southern Senegal across to the north of the Congo River to western Tanzania and western Uganda (Atlas 2019; List 2019). Across their ranges, chimpanzees show discernible genetic diversity (Gagneux *et al.* 2001) and are broadly classified into two geographical clusters comprising of four subspecies: a western African cluster that includes *P. t. verus* and *P. t. ellioti* and a central/eastern African clusters that includes *P. t. troglodytes* and *P. t. schweinfurthii* (Prado-Martinez *et al.* 2013; Mitchell *et al.* 2015). Both mtDNA and nuclear DNA-based studies have shown the genetic distinctness of Nigeria-Cameroonian chimpanzees (*P. t. ellioti*) from closely related Western chimpanzees (*P. t. verus*) (Gonder *et al.* 2006; Oates *et al.* 2009; Prado-Martinez *et al.* 2013; Mitchell *et al.* 2015). Genomewide analysis on chimpanzee subspecies as part of Great Ape Genome Project revealed presence of discernible genetic variation within populations across all chimpanzee subspecies except the central population, likely can be attributed to the small sample size (Prado-Martinez *et al.* 2013).

While whole genome-based approaches can efficiently identify sub-structures within populations, it is not always cost-effective. An alternative is to develop a panel of highly

Conceptualization and methodology, R. Das. Software, R. Das and S. Anjana. Formal analysis, S. Anjana and S. Pradivya. Investigation, S. Anjana and S. Pradivya. Resources, R. Das. Data curation, S. Anjana and R. Das. Writing – original draft preparation, S. Anjana and S. Pradivya. Writing – review and editing, R. Das. Visualization, R. Das and S. Anjana. Supervision, R. Das. Project administration, R. Das. Funding acquisition, R. Das.

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/s12041-020-01192-z) contains supplementary material, which is available to authorized users.

Published online: 11 May 2020

informative single-nucleotide polymorphisms (SNPs) from the whole-genome dataset, which can efficiently recapitulate the sub-structures within populations, depicted by the whole-genome data. These highly informative SNPs that exhibit significant difference in allele frequencies across different populations are known as ancestry informative markers (AIMs) (Rosenberg *et al.* 2003; Shriver *et al.* 2003; Kosoy *et al.* 2009). Over the years, several AIMs panels have been developed, mostly for identifying fine structures within and among various human populations (Rosenberg *et al.* 2003; Shriver *et al.* 2003; Kosoy *et al.* 2009; Nassir *et al.* 2009; Kidd *et al.* 2011; Tandon *et al.* 2011; Galanter *et al.* 2012; Huckins *et al.* 2014; Vongpaisarnsin *et al.* 2015; Das and Upadhyai 2018; Esposito *et al.* 2018; Das *et al.* 2019). However, in recent years, AIMs panels have been successfully developed and validated for nonhuman primates such as *Rhesus macaques* (Kanthaswamy *et al.* 2014) and gorillas (Das *et al.* 2019), and commercially important animals such as honey bees (Muñoz *et al.* 2015). However, apart from a panel of 9000 markers that can identify chimpanzees at the subspecies level (Hormozdiari *et al.* 2013), chimpanzee AIMs panel has never been developed, largely due to the unavailability of sufficient number of chimpanzee genomes across various populations.

In this study, we developed the first ever AIMs panel of 435 SNPs for Nigeria-Cameroonian chimpanzees (*P. t. ellioti*) employing the whole-genome data available in Great Ape Genome Project database (Prado-Martinez *et al.* 2013). We undertook the same strategy that was successfully employed to develop AIMs panel for gorillas (Das *et al.* 2019), i.e. first to compare four commonly used strategies used for AIMs determination, namely Infocalc algorithm, Wright's F_{ST}, smart principal component analysis (SmartPCA) and ADMIXTURE to assess the best approach for AIMs determination, and then to develop the AIMs panel employing the consensus of the four approaches with additional supplement from the best approach. Our AIMs panel can not only be useful for identifying substructures within Nigeria-Cameroonian chimpanzee population but also can aid in tracing back parent population of trafficked animals. Further, this AIMs panel can provide precise knowledge about an individual's ancestry which can in turn be used in planned reintroduction of chimpanzees to their natural habitat and in selective breeding in zoos to maintain their genomic integrity.

Materials and methods

Dataset

The dataset employed in this study comprised of 25 chimpanzees available in The Great Ape Genome Project (GAGP) database (Prado-Martinez *et al.* 2013): eastern chimpanzees (n = 6), central chimpanzees (n = 4), western chimpanzees (n = 5) and Nigeria-Cameroonian chimpanzees (n = 10). Due to the absence of desirable sample size (n = 10), eastern, central and western chimpanzee samples could not be employed for further analyses and the AIMs panel was developed only for the Nigeria-Cameroonian chimpanzee population. The initial dataset obtained from GAGP as VCF files, comprised of 53,806,652 markers. VCF files were converted into PLINK format using VCFtools v.0.1.13 (Danecek *et al.* 2011). SNPs, 46,048,193 were pruned out using –indep-pairwise 50 5 0.1 function implemented in PLINK v1.9 (window size 50, step size 5 and r^2 threshold 0.1) (Purcell *et al.* 2007). Remaining 7,758,459 SNPs underwent quality control checks. As a quality control measure SNPs with missing genotype information for $\geq$ 10% of the individuals and minor allele frequency (MAF) $\leq$ 5% were removed. This was performed using –geno 0.1 –maf 0.05 functions implemented in PLINK v1.9. The remaining 2,462,291 SNPs that passed the quality control measures was used for downstream analyses and termed as the complete SNP set (CSS).

Population clustering and admixture analysis on CSS

Principal component analysis (PCA) was performed on the CSS using –pca function implemented in PLINK v1.9 and top two principal components (PCs) were plotted in R v3.6.0 (figure 1a). The ancestry of the chimpanzee genomes was estimated using unsupervised clustering implemented in ADMIXTURE v1.3 (Alexander *et al.* 2009). Admixture analyses were performed for K = 2 – 6. However, K = 4 was selected for further analysis due to its biological relevance. All admixture plots were generated using R v3.6.0 (figure 2a).

Comparison of various AIMs determining approaches

We compared four AIMs determining approaches to determine the best AIMs determination strategy:

***Infocalc*:** Infocalc algorithm implemented in Infocalc v1.1 (Rosenberg *et al.* 2003) determines the informativeness of multiallelic SNPs in determining the ancestry of an individual based on the allele frequencies in the populations (Rosenberg *et al.* 2003). Infocalc v1.1 compatible file was created using –recode-structure function in PLINK v1.9. The output file was sorted based on informativeness defining column (I_n) and the top 10,000 ranking SNPs (n = 35,616) were selected for further analysis.

***Admixture*:** The P output file of ADMIXTURE v1.3, providing information about the differential contribution of study SNPs towards various ancestries was used to determine the candidate AIMs panel. After sorting the SNPs in decreasing order based on column to column variance, top 10,000 ranking SNPs were selected (n = 10,872).

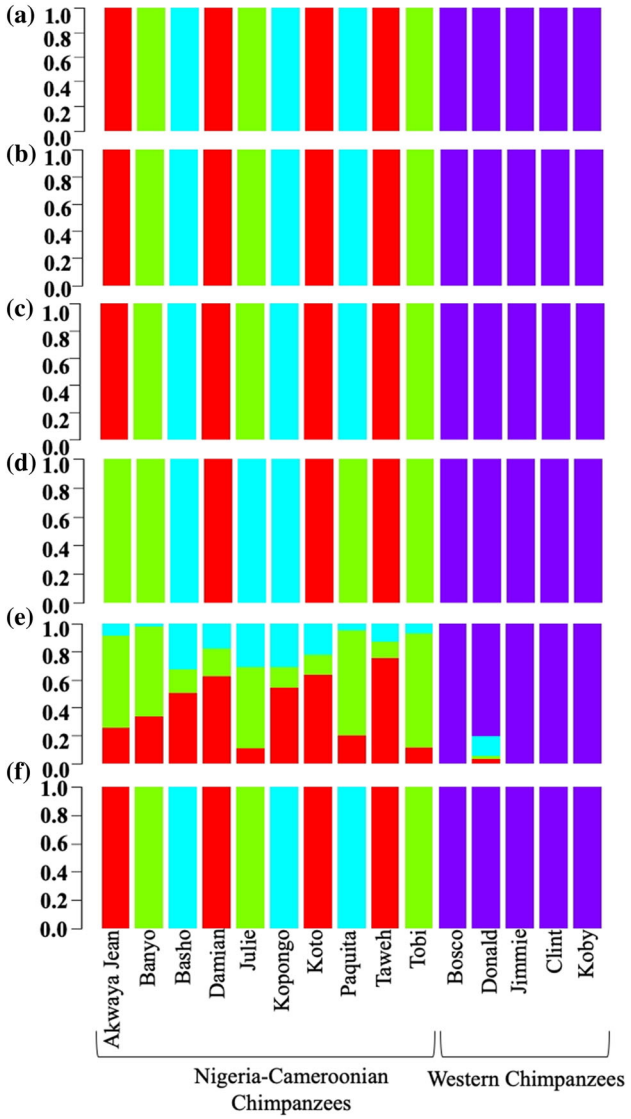

Figure 1. Admixture analysis of data subsets generated through the most informative SNPs detected by various AIMs-determining strategies. Admixture plots showing the ancestry components of gorilla genomes. (a) Admixture analysis of the CSS (2,462,291 SNPs); (b) Admixture analysis of $\text{Infocalc}_{10,000}$ dataset; (c) Admixture analysis of $\text{Admixture}_{10,000}$ dataset; (d) Admixture analysis of $F_{ST10,000}$ dataset; (e) Admixture analysis of $\text{SmartPCA}_{10,000}$ dataset; (f) Admixture analysis of the AIMs (n = 435) dataset. Admixture proportions were generated through an unsupervised admixture analysis at K = 4 using ADMIXTURE v1.3 and plotted in R v3.6.0. Each individual is represented by a vertical line partitioned into coloured segments whose lengths are proportional to the contributions of the ancestral components to the genome of the individual. Purple colour represents western chimpanzees while cyan, green and red represent individual chimpanzees with various Nigeria-Cameroonian ancestry.

Wright's F_{ST}: F_{ST} measures the degree of differentiation among various populations based on the genetic structure of the populations. F_{ST} score was calculated independently for all SNPs under study using $-f_{st}$ function implemented in PLINK v1.9. The family ID (FID) was used as the indicator of the biogeographical affinity of the chimpanzee genomes. Top 10,000 ranking SNPs (n = 11,686) with highest F_{ST} values were selected for further analysis.

***SmartPCA*:** SmartPCA implemented in EIG v7.2.1 (Patterson *et al.* 2006; Price *et al.* 2006) estimates the weightage of all SNPs under study in regards to their ability towards determining population differentiation. The SNP weightage file, obtained as a 'snpwt', determines weightage of SNPs for all PC. Top 10,000 ranking SNPs (n = 12,624) with the highest SNP weightage for PC1 were selected for further analysis.

Among the four 10,000 SNPs datasets, the optimal AIMs developing approach was determined by comparing the datasets qualitatively and quantitatively to the CSS. The PCA and admixture plots were generated using R v3.6.0.

Consensus and negative control datasets

To find a consensus among the top 10,000 ranking SNPs generated through the four AIMs determining approaches, a Venn diagram was plotted among the four datasets and the SNPs common to all methods was termed as the consensus SNP set. PCA and admixture analyses were performed on the consensus SNP set and compared against the CSS. Additionally, to validate the efficiency and utility of the AIMs panel, 20 SNP-sets were generated randomly selecting SNPs from the CSS.

Quantitative analysis

For quantitative comparisons, the Euclidean distances between the admixture fractions of the individuals determined by the CSS and the subsets were calculated. The results are summarized in boxplots plotted in R v3.6.0.

Results

Population clustering and admixture analysis

The ancestry of the chimpanzee genomes was estimated using unsupervised clustering implemented in ADMIXTURE v1.3 and compared against the CSS (figure 1a). The datasets developed though Infocalc and ADMIXTURE approaches performed the best in the admixture analysis, precisely recapitulating the information depicted by the CSS (figure 1, b&c respectively). Both F_{ST} and SmartPCA based approaches failed to replicate the CSS. While the F_{ST} based approach failed to determine the ancestry information of Akwaya Jean, Julie and Paquita (figure 1d), the SmartPCA based dataset completely failed to recapitulate the ancestry information of all Nigeria-Cameroonian chimpanzees and Donald from the western chimpanzee population (figure 1e).

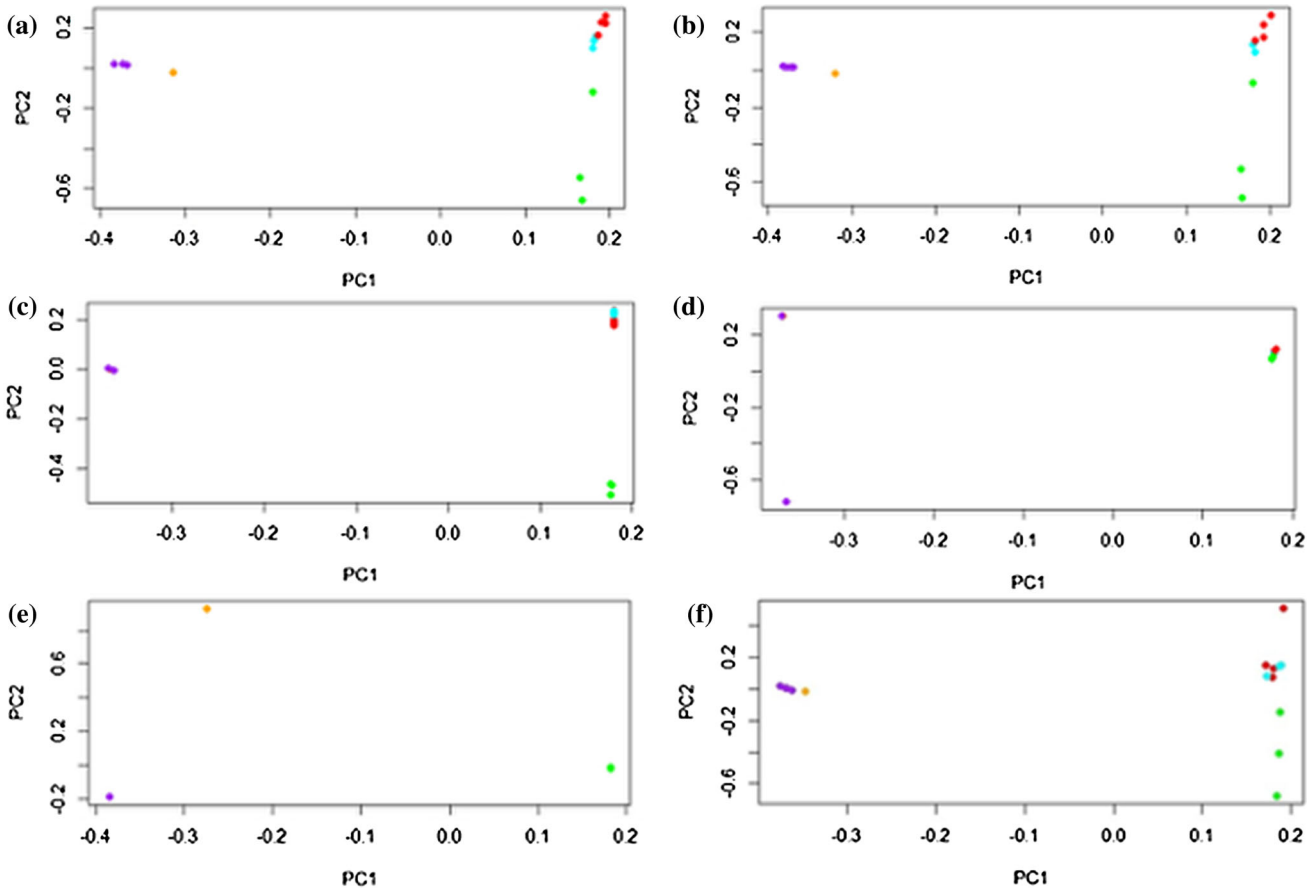

Figure 2. PCA of chimpanzee genomes. PCA plots showing genetic differentiation among query chimpanzee genomes. The data subsets were generated using the informative SNPs detected through various AIMs determining approaches. (a) PCA of the CSS (2,462,291 SNPs); Here, the *x*-axis (PC1) explained 47% variance while the *y*-axis (PC2) explained 20% variance of the data. (b) PCA of $\text{Infocalc}_{10,000}$; In this case too, the *x*-axis (PC1) explained 47% variance while the *y*-axis (PC2) explained only 20% variance of the data. (c) PCA of $\text{Admixture}_{10,000}$; in this case, the *x*-axis (PC1) explained 98% variance while the *y*-axis (PC2) explained only 1% variance of the data. (d) PCA of $F_{\text{ST}10,000}$; in this case, the *x*-axis (PC1) explained 99% variance while the *y*-axis (PC2) explained only 1% variance of the data. (e) PCA of $\text{SmartPCA}_{10,000}$; here too, the *x*-axis (PC1) explained 98% variance while the *y*-axis (PC2) explained only 2% variance of the data. (f) PCA of the AIMs dataset (n = 435); In this case, the *x*-axis (PC1) explained 76% variance while the *y*-axis (PC2) explained 9% variance of the data. The colour of the individuals is same as figure 1. The western chimpanzee Donald is coded in orange. $\text{Infocalc}_{10,000}$ and our AIMs panel (n = 435) could identify its genetic distinctness. In all cases, PCA was performed in PLINK v1.9 and the top four PCs were extracted. Top two PCs (PC1 and PC2), explaining the highest variance of the data were plotted in R v3.6.0.

PCA was performed using –pca function implemented in PLINK v1.9. In PCA, the dataset generated through the top 10,000 SNPs obtained from Infocalc, recapitulated the fine structures within Nigeria-Cameroonian chimpanzee population depicted by the CSS (figure 2a) with the highest precision (figure 2b). The dataset generated through the top 10,000 SNPs from ADMIXTURE was distant second, portraying more stringent clustering among individuals compared to the CSS (figure 2c). The other two approaches (F_{ST} and SmartPCA) completely failed to replicate the clustering depicted by the CSS (figure 2, d&e respectively).

Overall, Infocalc emerged as the best ancestry-determining approach in both PCA and admixture analysis, followed by the ADMIXTURE based strategy.

Consensus and negative control datasets

To find a consensus among the top 10,000 ranking SNPs generated through the four AIMs determining approaches: Infocalc, SmartPCA, ADMIXTURE and F_{ST}, a Venn diagram was plotted among the four datasets (figure 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/) and the SNPs common to all methods was selected. The consensus panel was comprised of 129 SNPs. The admixture and PCA plots generated through the consensus dataset failed to recapitulate the fine population structure depicted by the CSS (figures 2 and 3 in electronic supplementary material, respectively). Since the consensus SNP set failed to replicate the CSS, likely due to the absence of essential private alleles, we supplemented this SNP set with

SNPs determined by the best two approaches: Infocalc and admixture. We thus generated six sets of top-ranking SNPs from admixture SNP sets: 5, 20, 50, 100, 200 and <500. In the case of Infocalc SNP set, all 10,000 ranking SNPs had the same informativeness (I_n value). Thus, we picked a random set, each of 100, 200, 300, 400 and 500 SNPs from the 10,000 SNP dataset. Employing the newly generated aforesaid datasets and the consensus SNP set in different combinations, 71 different datasets were generated. PCA and admixture analyses were performed on these datasets and compared them against the CSS to find the best combination.

Among the 71 datasets, that which recapitulated the information given by CSS with the highest precision and with the least number of SNPs comprised of the consensus dataset (n = 129), along with top five ranking SNPs (n = 6) determined by the ADMIXTURE and 300 SNPs from the Infocalc (n = 435) (figures 1f and 2f, respectively in admixture and PCA). We considered this panel of 435 SNPs as the AIMs panel. The average MAF of the AIMs panel was 0.192 with a range of 0.0667 to 0.5 (table 1 in electronic supplementary material).

To establish the efficiency and precision in recapitulating the ancestry information of the chimpanzee genomes depicted by the CSS, we next generated 20 negative control datasets, randomly selecting 435 SNPs from the CSS. Since all of them failed to recapitulate the fine population structure depicted by the complete SNP set both qualitatively and quantitatively, we did not see the necessity of developing additional random negative control datasets. Our results indicated the superiority of AIMs panels over randomly selected SNP-sets in replicating the ancestry information.

Quantitative analysis

For quantitative comparisons, the Euclidean distances between the admixture fractions of the individuals determined by the CSS and the subsets were calculated. Quantitatively, we did not find any significant difference among the 10,000 SNP panels developed through Infocalc, ADMIXTURE and F_{ST} based approaches (figure 3). However, the SmartPCA based approach performed significantly worse than the others. Among the other smaller datasets, consensus SNP set performed worse. There was no significant difference among most combination datasets, except the one comprised of the consensus and 200 Infocalc SNPs (figure 3).

Discussion

Over the years, chimpanzees with dwindling population size, are faced with a serious survival threat. One of the measures to save chimpanzee gene pools and to maintain their genetic integrity can be planned captive breeding programmes. However, proper knowledge of the ancestry of the participating individuals can be essential for captive breeding programmes to avoid inbreeding among genetically related/similar individuals and to increase genetic diversity. Although, whole-genome based approaches can precisely determine the ancestry of participating individuals, it is not always cost-effective. An alternative cost-effective strategy is to develop an AIMs panel from the available whole-genome datasets and then to employ the AIMs panel for determining ancestry of other individuals of the population.

In this study, we sought to develop the first-ever AIMs panel for chimpanzees, which can recapitulate the ancestry information with high efficiency and precision. Since only Nigeria-Cameroonian chimpanzee population had 10 samples and others had <10 individuals, we could only develop the AIMs panel for the former. We first compared four commonly used AIMs determining approaches: Infocalc, ADMIXTURE, SmartPCA and F_{ST}. As mentioned previously, Infocalc emerged as the best ancestry-determining approach in both PCA and admixture analysis, followed by the ADMIXTURE-based strategy. The datasets developed through Infocalc and ADMIXTURE based strategies could precisely recapitulate the ancestry information of the chimpanzee genomes depicted by the CSS in the admixture analysis (figure 1, b&c respectively). On the contrary, both F_{ST} and SmartPCA based approaches failed to replicate the ancestry information depicted by the CSS. While the SmartPCA based dataset completely failed to recapitulate the ancestry information of all Nigeria-Cameroonian chimpanzees and Donald from the western chimpanzee population (figure 1e), the F_{ST} based approach, although performed a little better than the former, failed to determine the ancestry information of Akwaya Jean, Julie and Paquita (figure 1d). In PCA, Infocalc performed significantly better than the ADMIXTURE-based approach. It recapitulated the fine structures within Nigeria-Cameroonian chimpanzee population depicted by the CSS (figure 2a) with the highest precision (figure 2b). The ADMIXTURE based approach (figure 2c), although performed discernibly better than the other two strategies (F_{ST} and SmartPCA) failed to capture the essence of population clustering depicted by the CSS. However, in the quantitative analysis, while SmartPCA based strategy came out to be the least effective AIMs determining approach similar to the qualitative analyses, we did not find any significant difference among the datasets developed through Infocalc, ADMIXTURE and F_{ST} based approaches (figure 3). It reiterates the importance of both qualitative and quantitative analyses while developing AIMs panel, especially when the sample size is small. Interestingly, similar trend was observed in case of gorillas, Infocalc worked the best in recapitulating the ancestry information and SmartPCA was the least effective strategy in doing so (Das *et al.* 2019). Further, 99.19% SNPs determined through the Infocalc based approach were private alleles, depicting the uniqueness of the approach compared to the other three. On the contrary, only 13.07% private alleles were found in case of the ADMIXTURE based strategy, followed by 19.07% depicted by the F_{ST}.

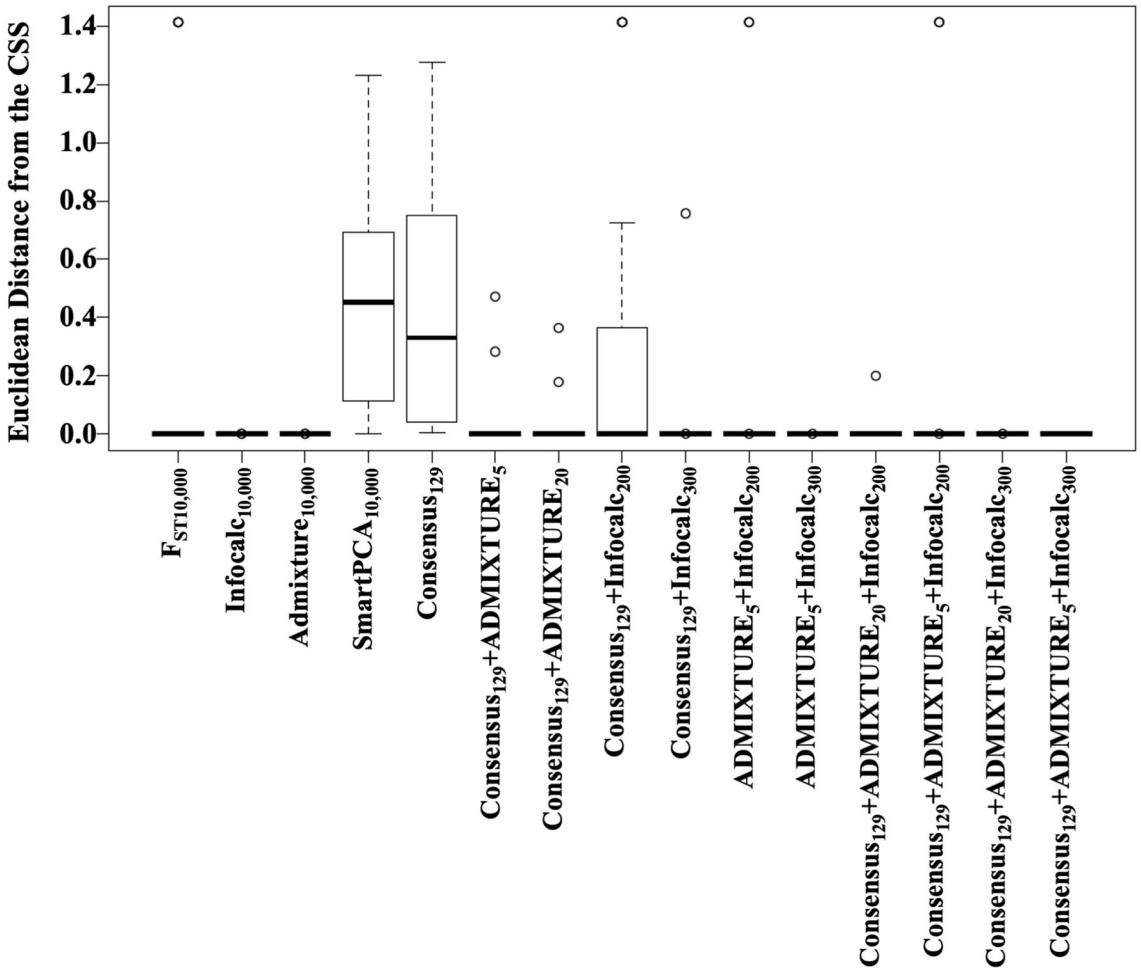

Figure 3. Box and whisker plots comparing the Euclidean distances between the admixture proportions of the chimpanzee genomes obtained from the CSS and those obtained from the reduced datasets. The box and whisker plot was generated in R v3.6.0. The number of SNPs the datasets are comprised of is mentioned in their nomenclature. There was no significant difference among the 10,000 SNP panels developed through Infocalc, ADMIXTURE and F_{ST} based approaches but the SmartPCA based approach performed significantly worse. Among the other smaller datasets, consensus SNP set (n = 129) performed the worst. There was no significant difference among combination datasets with the exception of consensus (n = 129) + 200 Infocalc SNPs dataset.

Subsequently, we developed a consensus SNP set taking the intersection of the aforesaid approaches. However, due to the inadequacy of the consensus panel to recapitulate the ancestry information with high precision, we supplemented it with private SNPs determined by the best two approaches: Infocalc and ADMIXTURE. We thus developed an AIMs panel of 435 SNPs, which could recapitulate the ancestry information of the Nigeria-Cameroonian chimpanzees with high precision.

Here we note that the western chimpanzees were used as outgroup in all analyses; they were not used for the AIMs determination. Our study solely focussed on developing an AIMs panel that can depict the fine population structure within Nigeria-Cameroonian chimpanzees. However, the genomic distinctness of the western chimpanzee Donald from the other westerns, depicted by the CSS (figure 2a), was precisely captured by our AIMs panel (figure 2f). It indicates that, despite being developed only for the Nigeria-Cameroonian chimpanzees, our AIMs panel is versatile enough to detect fine structures within other chimpanzee populations.

Overall, to the best of our knowledge, we have developed the first-ever AIMs panel for Nigeria-Cameroonian chimpanzees, which can be monumental in maintaining the genetic integrity of these animals through planned captive breeding and in tracking illegal trading across the globe.

References

Alexander D. H., Novembre J. and Lange K. 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genom. Res.* **19**, 1655–1664.

Atlas W. 2019 World Atlas of great apes and their conservation by Julian Caldecott, Lera Miles. University of California Press, Hardcover.

Danecek P., Auton A., Abecasis G., Albers C. A., Banks E., DePristo M. A. *et al.* 2011 The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158.

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

Das R., Roy R. and Venkatesh N. 2019 Using ancestry informative markers (AIMs) to detect fine structures within gorilla populations. *Front. Gen.* **10**, 43.

Esposito U., Das R., Syed S., Pirooznia M. and Elhaik E. 2018 Ancient ancestry informative markers for identifying fine-scale ancient population structure in Eurasians. *Genes* **9**.

Gagneux P., Gonder M. K., Goldberg T. L. and Morin P. A. 2001 Gene flow in wild chimpanzee populations: what genetic data tell us about chimpanzee movement over space and time. *Philos. Trans. R. Soc. Lond. Sci. B* **356**, 889–897.

Galanter J. M., Fernandez-Lopez J. C., Gignoux C. R., Barnholtz-Sloan J., Fernandez-Rozadilla C., Via M. *et al.* 2012 Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet.* **8**, e1002554.

Gonder M. K., Disotell T. R. and Oates J. F. 2006 New genetic evidence on the evolution of chimpanzee populations and implications for taxonomy. *Int. J. Primat.* **27**, 1103.

Hormozdiari F., Konkel M. K., Prado-Martinez J., Chiatante G., Herraez I. H., Walker J. A., *et al.* 2013 Rates and patterns of great ape retrotransposition. *Proc. Natl. Acad. Sci USA* **110**, 13457–13462.

Huckins L. M., Boraska V., Franklin C. S., Floyd J. A. B., Southam L., Sullivan P. F. *et al.* 2014 Using ancestry-informative markers to identify fine structure across 15 populations of European origin. *Eur. J. Hum. Gen.* **22**, 1190–1200.

Kanthaswamy S., Johnson Z., Trask J. S., Smith D. G., 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. Primat.* **76**, 1105–1113.

Kidd J. R., Friedlaender F. R., Speed W. C., Pakstis A. J., De La Vega, F. M. and Kidd K. K. 2011 Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Inves. Genet.* **2**, 1.

Kosoy R., Nassir R., Tian C., White P. A., Butler L. M., Silva G., *et al.* 2009 Ancestry Informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mut.* **30**, 69–78.

List I. R. 2019 Pan troglodytes (Chimpanzee).

Mitchell M. W., Locatelli S., Ghobrial L., Pokempner A. A., Sesink Clee P. R., Abwe E. E. *et al.* 2015 The population genetics of wild chimpanzees in Cameroon and Nigeria suggests a positive role for selection in the evolution of chimpanzee subspecies. *BMC Evol. Biol.* **15**, 3.

Muñoz I., Henriques D., Johnston J. S., Chávez-Galarza J., Kryger P. and Pinto M. A. 2015 Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). *PLoS One* **10**.

Nassir R., Kosoy R., Tian C., White P. A., Butler L. M., 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.

Oates J. F., Groves C. P. and Jenkins P. D. 2009 The type locality of Pan troglodytes vellerosus (Gray, 1862), and implications for the nomenclature of West African chimpanzees. *Primates* **50**, 78–80.

Patterson N., Price A. L. and Reich D. 2006 Population structure and eigenanalysis. *PLoS Genet.* **2**, e190.

Prado-Martinez J., Sudmant P. H., Kidd J. M., Li H., Kelley J. L., Lorente-Galdos B. *et al.* 2013 Great ape genetic diversity and population history. *Nature* **499**, 471–475.

Price A. L., Patterson N. J., Plenge R. M., Weinblatt M. E., Shadick N. A. and Reich D. 2006 Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Gen.* **38**, 904–909.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M. A. R., 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.

Rosenberg N. A., Li L. M., Ward R. and Pritchard J. K. 2003 Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73**, 1402–1422.

Shriver M. D., Parra E. J., Dios S., Bonilla C., Norton H., Jovel C. *et al.* 2003 Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum. Genet.* **112**, 387–399.

Tandon A., Patterson N. and Reich D. 2011 Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays. *Genet. Epidemiol.* **35**, 80–83.

Vongpaisarnsin K., Listman J. B., Malison R. T. and Gelernter J. 2015 Ancestry informative markers for distinguishing between Thai populations based on genome-wide association datasets. *Leg. Med* (Tokyo, Japan) **17**, 245–250.

Corresponding editor: H. A. Ranganath