Global and Local Ancestry Estimation in a Large, Captive Baboon Colony

Christopher Kendall[1*], Guilherme Debortoli[2], Amin Nooranikhojasteh, Debbie Christian[4,5], Deborah Newman[4,5], Shelley Cole[4,5], Ken Sayers[5], Esteban Parra[2], Michael Schillaci[3], Bence Viola[1]

[1]*Department of Anthropology, University of Toronto, Toronto, Ontario; *corresponding author (chris.kendall@mail.utoronto.ca); [2]Department of Anthropology, University of Toronto Mississauga, Mississauga, Ontario; [3]Department of Anthropology, University of Toronto Scarborough, Scarborough, Ontario; [4]Texas Biomedical Research Institute, San Antonio, Texas; [5]Southwest National Primate Research Center, San Antonio, Texas*

Introduction

Baboons (genus *Papio*) are Old World monkeys from the subfamily Cercopithecinae that have been classified into 6 currently identified species[1,2,3] or subspecies[4]. Due to similar evolutionary trajectories to modern humans[5,6,7,8], baboons have been extensively studied in various contexts[9,10,11]. Much of this research has been conducted using animals from the Southwest National Primate Research Center (SNPRC). The SNPRC has a large living database of animals and an active breeding program with accompanying pedigree[11] where 95% of the colony is made up of olive (*P. anubis*) and yellow (*P. cynocephalus*) baboons. Previous genetic research on baboons from the SNPRC has shown errors in the pedigree such as improper species/subspecies assignment[12] and admixture within putatively unadmixed founding animals[13].

We examined the whole genomes of 881 olive, yellow, and olive-yellow crosses using high resolution genetic data. Through genotype refinement, imputation, and phasing of low coverage individuals (~5x coverage) using high coverage reference animals (>=15x coverage) we identified global and local ancestry estimates for all animals in our study. Our research goals included: 1) expanding the available genomes from the SNPRC, 2) correcting any further potential pedigree errors for future analysis, and 3) estimating the extent to which putative purebred animals may contain instances of admixed ancestry. Additionally, we were also able to generate genomic resources (data not shown) including Ancestry-Informative-Markers, an updated genetic map, and a list of >27,000 fixed markers between the two species.

Materials & Methods

In this study, we used publicly available whole genome sequencing data from 901 individuals from a pedigreed captive colony at the SNPRC extracted using blood or tissue samples. All raw sequencing data used here are available from the NCBI Sequence Read Archive under BioProject PRJNA433868. Subsets of this complete dataset have been used in analyses for previously published studies[13,14]. Data processing followed the GATK v3 pipeline[15] after alignment to the *Panubis1.0* genome[16]. Hard filtering of variants followed previous workflows[12,17] and were based on statistics from our own data using BCFtools[18]. Repetitive elements were removed using RepeatMasker[19] and Tandem Repeats Finder[20].

After hard filtering we were left with over 96 million variants and 881 samples after removing animals not of putative olive or yellow ancestry. We further filtered using only non-missing variants and those with a minor allele frequency <5% in all animals that were sequenced to high coverage leaving us with more than 6 million markers for analysis. We imputed the low-coverage samples using ShapeIT5[21] and IMPUTE5[22]. Global ancestry was assessed using ADMIXTURE[23], while local ancestry was determined using RFMix[24] with phase-correction done by Tractor[25]. Principal component analysis was done using PLINK[26], plotting of the PCA and global ancestry was completed using R[27] while local ancestry was plotted using haptools[28].

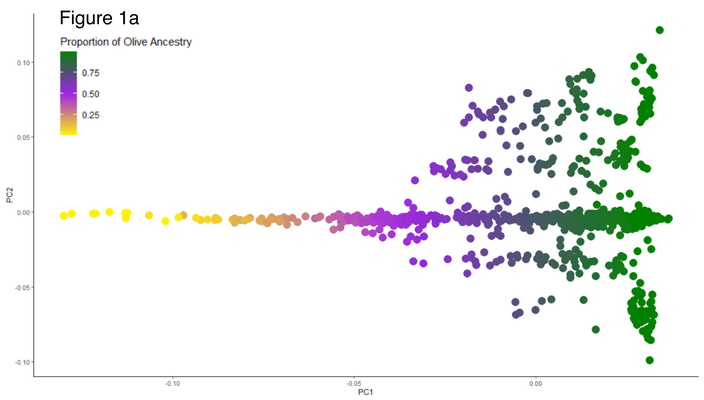

Figure 1a. Principal component analysis of entire 881 sample set. PC1 explains 45.51% of the variation while PC2 represents 32.53% of the variation within the dataset. This highlights a gradient of ancestry from purebred yellow on the left to purebred olive on the right. Colouring is based on the proportion of global olive baboon ancestry identified using ADMIXTURE[23] with k=2 clusters in supervised mode. Green represents purebred olive ancestry and yellow represents purebred yellow ancestry.

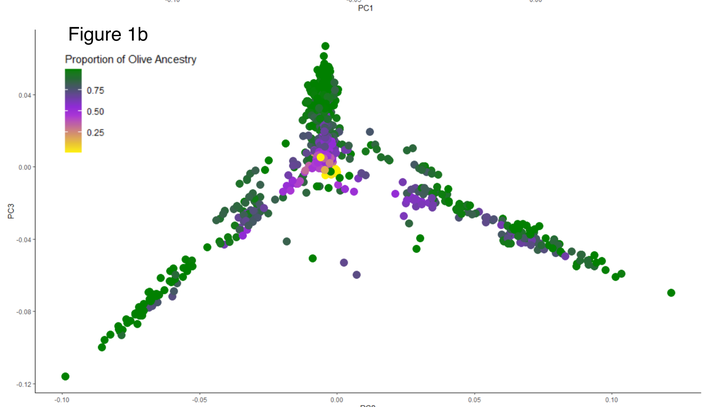

Figure1b. PC3 describes 27.61% of the variation within the dataset. Low diversity of the yellows and high diversity of olives is showcased. Colouring is based on the proportion of global olive baboon ancestry identified using ADMIXTURE[23] with k=2 clusters in supervised mode. Green represents purebred olive ancestry and yellow represents purebred yellow ancestry.

Results & Discussion

Principal component analysis (PCA) of the entire dataset is shown in Figures 1a and 1b. As expected, the individuals separate along a continuum between olive and yellow ancestry in PC1. The PCA also highlights potential population structure within the olives in PC2. PC3 shows the limited diversity among the yellows. Figure 2 showcases the global ancestry proportions as they relate to each putative ancestry, illustrating evidence of pedigree errors. Local ancestry assessment was able to highlight instances of admixture even in putative purebred animals along with tracts of singular ancestry in first-generation hybrids (Figure 3).

Our results align with previous research[3,12,13,14] . Our global and local ancestry estimates identified evidence of admixture in both putative olive and yellow baboons from the SNPRC along with highlighting pedigree errors where the putative animal label does not match what is seen in high resolution genomic data. However, our data does not show the amount of admixture reported previously (upwards of 22% in yellow founders)[13], with our estimates being much more conservative (Figure 2). Our results expanded the number of available baboon genomes available and found good agreement with previous studies.

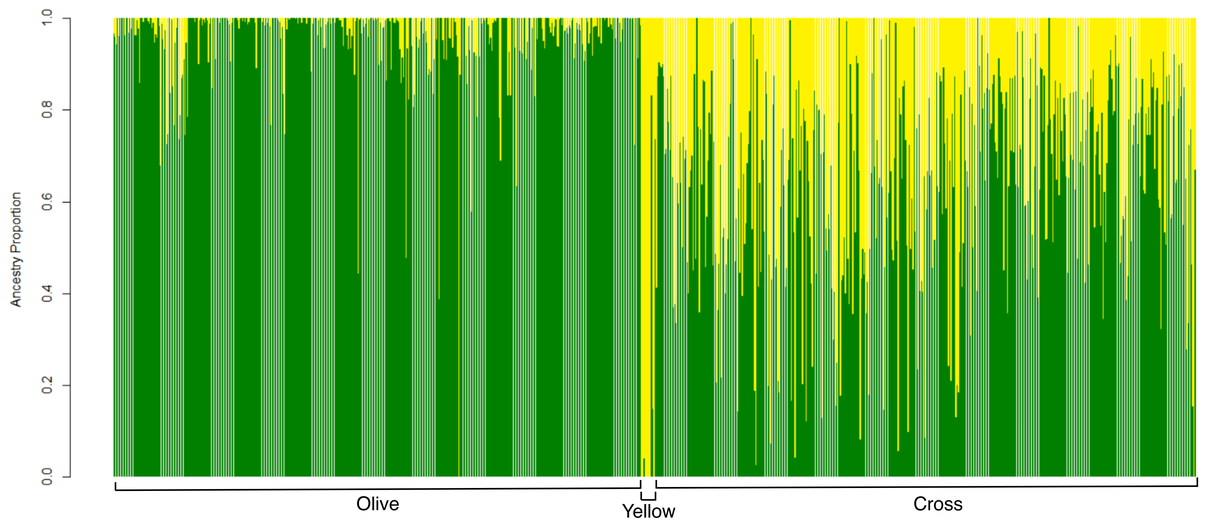

Figure 2. Global ancestry estimation using ADMIXTURE[23] in supervised mode after pruning for linkage disequilibrium. Green and yellow represent proportion of olive or yellow global ancestry, respectively. The olive samples highlight many instances of individuals with yellow ancestors in their past. The olives also exhibit instances of potential cryptic hybridization where likely F1 hybrids were able to pass visually as a purebred olive. Some yellows exhibit substantial olive ancestry contributions to their genome. The crosses have an expected gradient of ancestry contributions. However, there are quite a few examples of extensive yellow ancestry, which is surprising given the small number of yellow founders.

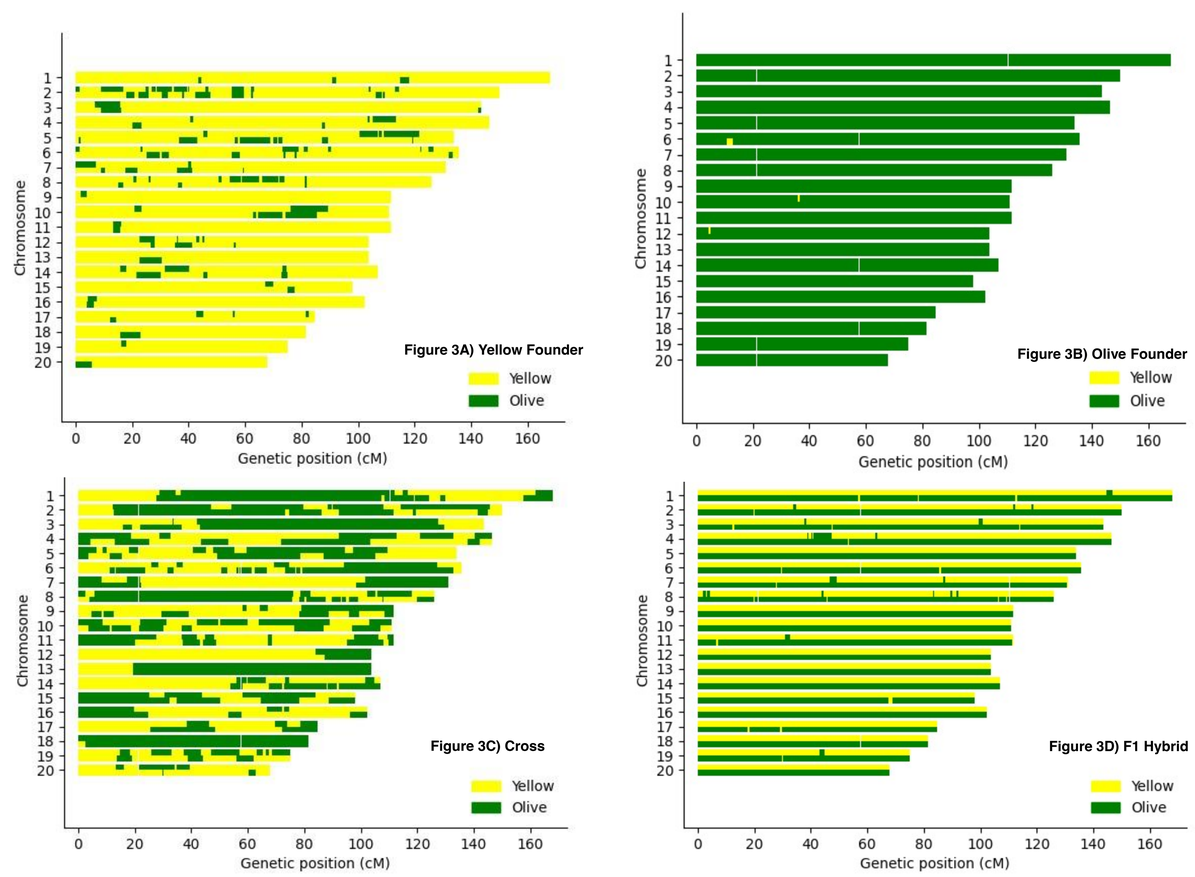

Figure 3. Local ancestry estimation using RFMix[24] and plotted using haptools[28]. **A)** 1X0102, labelled as a putative yellow in the SNPRC pedigree. **B)** 1X0026, labelled as a putative olive in the SNPRC pedigree. **C)** 28037, crossed sample of unknown degree. **D)** 10488, F1 hybrid of putative purebred olive and yellow. While there may be some ancestry misassignment on all our samples due to software errors, the more probable explanation is that admixture occurred in the past readily between these groups. We also note evidence of possible phase-switch errors persisting even after correcting for these using Tractor[25].

Acknowledgements & References

We would like to thank Dr. Jacqueline Robinson and Dr. Jeff Wall for providing their data for analysis. CK was supported by an NSERC CGS-D grant (CGSD2 - 535025 - 2019) for the duration of this research. Thank you to the CABA-ACAB organizing committee for allowing us to present this work and for organizing the event. The establishment, maintenance, and biological characterization of the pedigreed baboon colonies at the Southwest National Primate Research Center of Texas Biomedical Research Institute (SNPRC at Texas Biomed) has been supported in large part by grants to Texas Biomed Investigators by the National Institutes of Health (P51 RR013986, P01 HL028972). Please scan the QR code for access to the references and a digital copy of this poster.