bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The convoluted evolutionary history of the capped-golden langur lineage (Cercopithecidae: Colobinae) – concatenation versus coalescent analyses

Kunal Arekar[a, *], Abhijna Parigi[a, b], K. Praveen Karanth[a]

[a]Evolving-Phylo Lab, Centre for Ecological Sciences, Indian Institute of Science, Bangalore 560012, India

[b]presently at Integrative Biology Department, Michigan State University, East Lansing, Michigan 48824, USA

*Corresponding author. Email – kunal.cyan@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Abstract

Evolutionary studies have traditionally relied on concatenation based methods to reconstruct relationships from multiple markers. However, due to limitations of concatenation analyses, recent studies have proposed coalescent based methods to address evolutionary questions. Results from these methods tend to diverge from each other under situations where there is incomplete lineage sorting or hybridization. Here we used concatenation as well as multispecies coalescent (MSC) methods to understand the evolutionary origin of capped and golden langur (CG) lineage. Previous molecular studies have retrieved conflicting phylogenies, with mitochondrial tree grouping CG lineage with a largely Indian genus *Semnopithecus,* while nuclear markers support their affinities with a Southeast Asian genus, *Trachypithecus*. However, as pointed by others, the use of nuclear copies of mitochondrial DNA in the above studies might have generated the discordance. Because of this discordance, the phylogenetic position of CG lineage has been much debated in recent times. In this study, we have used nine nuclear and eight mitochondrial markers. Concatenated nuclear as well as the mitochondrial dataset recovered congruent relationships where CG lineage was sister to *Trachypithecus*. However nuclear species tree estimated using different MSC methods were incongruent with the above result, suggesting presence of incomplete lineage sorting (ILS)/hybridisation. Furthermore, CG lineage is morphologically intermediate between *Semnopithecus* and *Trachypithecus.* Based on this evidence, we argue that CG lineage evolved through hybridisation between *Semnopithecus* and *Trachypithecus*. Finally, we reason that both concatenation as well as coalescent methods should be used in conjunction for better understanding of various evolutionary hypotheses.

Keywords: CG lineage; ancient hybridisation; incomplete lineage sorting; *Trachypithecus*; *Semnopithecus*.

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1. Introduction

This day and age most phylogenetic studies use multiple markers to reconstruct the evolutionary relationships between taxa. Typically, such studies employ concatenated analysis which involves building a supermatrix by combining multilocus molecular data (de Querioz & Gatesy, 2006). However, the basic assumption of this approach, which is that all the loci experience same the evolutionary history, is often incorrect (de Querioz & Gatesy, 2006). Processes such as hybridisation and incomplete lineage sorting (ILS) can lead to discrepancy between the gene trees and the true species tree (Degnan James H., Rosenberg, 2006; Maddison, 1997). To overcome this limitation of the concatenation analysis, coalescence-based tree building methods have been proposed as an alternative (Degnan & Rosenberg, 2009; Knowles, L.L, Kubatko, 2010; Song, Liu, Edwards, & Wu, 2012; Zhong, Liu, & Penny, 2014).

ILS arises when two or more gene trees fail to coalesce back in time in the most recent common ancestor (Maddison, 1997) and this is more likely to occur in lineages where not enough time has passed since speciation (Leache & Rannala, 2011) whereas, hybridisation is a process that results in mixing of previously isolated gene pools. The interbreeding of individuals from different gene pools results in a hybrid species that shares genetic information with both the parent species (Mallet, 2005, 2007; Zinner, Arnold, & Roos, 2011). There have been many methods developed to estimate species tree in presence of ILS (Bryant, Bouckaert, Felsenstein, Rosenberg, & Roychoudhury, 2012; Chifman & Kubatko, 2015; Edwards, Liu, & Pearl, 2007; J. Heled & Drummond, 2010; Laura S. Kubatko, Carstens, & Knowles, 2009; Liu & Pearl, 2007; Liu, Yu, & Edwards, 2010; Mirarab & Warnow, 2015; Mossel & Roch, 2008) and hybridisation (Gauthier & Lapointe, 2007; Jin, Nakhleh, Snir, & Tuller, 2006; Olave, Avila, Sites, & Morando, 2017; Posada, 2002; Xu, 2000). However, distinguishing between these two processes is difficult and very few methods have been developed that incorporates both hybridisation and ILS (Buckley, Cordeiro, Marshall, & Simon, 2006; Joly, McLenachan, & Lockhart, 2009; Meng & Kubatko, 2009; Than, Ruths, Innan, & Nakhleh, 2007). Importantly, multispecies coalescent (MSC) models incorporate ILS in species tree estimation but these methods cannot distinguish between hybridisation and ILS (L. S. Kubatko, 2009). Additionally, there is a dearth of studies that compare and contrast concatenated analysis vs. MSC analysis under conditions where ILS/hybridisation is suspected.

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Here we have used multiple nuclear and mitochondrial markers in conjunction with different tree building approaches to better understand the processes that gave rise to a lineage of Asian colobine monkeys – the capped and golden langur (CG) lineage. Colobines are predominantly leaf eating monkeys (Chivers & Hladik, 1980) distributed in the tropical old world. They are a diverse group consisting of 10 genera and over 50 species (Brandon-Jones et al., 2004; Groves, 2001). Among the Asian colobines, *Semnopithecus* and *Trachypithecus*, are the most species rich genera distributed in the Indian subcontinent and Southeast Asia (SEA) respectively. There has been much ambiguity in the taxonomy of these two genera (Brandon-Jones et al., 2004; Davies & Oates, 1994; Groves, 2001); however, with the advent of molecular tools many of these issues have been resolved(Ashalakshmi, Nag, & Karanth, 2014; Karanth, 2008; Karanth, Singh, Collura, & Stewart, 2008). Nevertheless, the taxonomic placement of two species in the genus *Trachypithecus* continues to be in debate; these include the endangered golden langur (*Trachypithecus geei*) and capped langur (*Trachypithecus pileatus*). Golden langur, capped langur and Shortridge's langur (*Trachypithecus shortridgei*) are the three species that are together called as *T. pileatus* group(Wang et al., 2015) However, in this study, we have included *Trachypithecus shortridgei* within CG lineage; recently, *T. shortridgei* was split from capped langur and elevated to species level(Brandon-Jones et al., 2004; Groves, 2001).

Golden langur is distributed in parts of Bhutan and adjoining Indian state of Assam in Northeast India(Ram et al., 2016; Wangchuk, Inouye, & Hare, 2008) whereas, capped langur is more widespread and is distributed in northeast India, parts of Bhutan, Bangladesh, northwest Myanmar and southern China (Kumar & Solanki, 2008; Srivastava & Mohnot, 2001; Zhang, Wang, & Quan, 1981). Molecular data suggests the two species are sister lineages, however their phylogenetic position in relation to *Semnopithecus* and *Trachypithecus* remains unresolved. In the molecular studies undertaken by two separate groups, the mitochondrial Cytochrome *b* (Cyt-*b*) phylogenetic tree placed Capped-Golden (CG) lineage in the *Semnopithecus* clade whereas nuclear autosomal and Y chromosomal markers suggested that they were sister to *Trachypithecus* (Karanth et al., 2008; Osterholz, Walter, & Roos, 2008). This discordance led the authors to hypothesise reticulate evolution of CG lineage due to past hybridisation between *Semnopithecus* and *Trachypithecus*. Osterholz et al. (2008) pointed out that incomplete lineage sorting in either nuclear or mitochondrial markers could also generate this discordance. Interestingly the distribution of CG lineage is flanked by *Semnopithecus* on the west and by *Trachypithecus* in the east

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(Groves 2001; Karanth et al., 2008; Osterholz et al., 2008). The same year another study looked at the phylogenetic position of CG lineage using Cyt-*b* marker (Wangchuk et al., 2008). Surprisingly in their mitochondrial tree CG lineage branched within the *Trachypithecus* clade. Thus, Wangchuk et al. (2008) results were consistent with the retrieved topology using nuclear markers (Karanth et al., 2008; Osterholz et al., 2008) of the aforementioned studies and did not support the mitochondrial-nuclear discordance scenario (Karanth et al., 2008; Osterholz et al., 2008). Accordingly, they rejected the hybridisation hypothesis as a possible reason for the origin of CG lineage.

These studies suggested that CG lineage had at least two types of Cyt-*b* sequences, one of which cluster them with *Semnopithecus* and the other with *Trachypithecus*. Multiple mitochondrial sequences when amplified from the same species suggest the presence of nuclear mitochondrial DNA sequences (numts). Numts are mitochondrial sequences incorporated into the nuclear genome(Bensasson, Zhang, Hartl, & Hewitt, 2001; Blanchard & Lynch, 2000; Ford Doolittle, 1998) and behave like nuclear pseudogenes. These nuclear copies of mitochondrial sequences when considered as true mitochondrial sequences and used in analysis can lead to erroneous phylogenetic topologies (Bensasson et al., 2001; Collura & Stewart, 1995). Numts have been reported from both *Semnopithecus* and *Trachypithecus* (Karanth, 2008). A recent study reported the isolation of full length numts from capped langur and *Trachypithecus shortridgei* (Wang et al., 2015). These numt sequences branched with the genus *Semnopithecus* (Karanth et al., 2008; Osterholz et al., 2008), but the mitochondrial sequences from capped langur and *Trachypithecus shortridgei* places them sister to other members of *Trachypithecus*, consistent with the Cyt-*b* phylogeny of Wangchuk et al., (2008). Thus, the authors concluded that "mitochondrial" Cyt-*b* sequences of CG lineage used by Karanth et al. (2008) and Osterholz et al. (2008) were numts.

Taken together, these studies would suggest that CG lineage indeed belongs to *Trachypithecus* clade. Nevertheless, studies on numts by Karanth (2008) and Wang et al. (2015) suggest that hybridisation hypothesis for the origin of CG lineage cannot be rejected. Karanth (2008) isolated numt sequences from capped langur, golden langur and Phayre's leaf monkey (*T. phayrei*). Assuming that CG lineage belongs to *Trachypithecus*, these numts are predicted to be related to the *Trachypithecus* clade or expected to be sister to the clade consisting of both *Semnopithecus* and *Trachypithecus*. Interestingly in their analysis some numts isolated from both capped and golden langurs were also sister to *Semnopithecus*. This

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

unusual placement of some CG numts thus suggested hybridisation. Based on molecular dating analysis Karanth (2008) estimated the hybridisation to have occurred between 7.1 to 3.4 million years ago (mya). Similarly, Wang et al. (2015) proposed that the numts originated in the genus *Semnopithecus* and hypothesised a unidirectional introgression from *Semnopithecus* into the ancestors of capped langur and *T. shortridgei*. They dated the time of hybridisation to be no later than 0.26 mya and no earlier than 3.47 mya. Hybridisation in primates has been reported between species and subspecies(Pastorini, Zaramody, Curtis, Nievergelt, & Mundy, 2009; Rabarivola, Meyers, & Rumpler, 1991; Vasey & Tattersall, 2002), but there are a few studies that have observed it between genera (Davenport et al., 2006; Dunbar & Dunbar, 1974; Jones et al., 2005; Zinner, Chuma, Knauf, & Roos, 2018).

Thus, nuclear and mitochondrial data support placing CG lineage in *Trachypithecus*. Whereas, the numts data suggests past hybridisation between *Trachypithecus* and *Semnopithecus*. Nevertheless, it must be noted that these studies have undertaken phylogenetic analyses of concatenated markers to ascertain the phylogenetic position of CG lineage. Such analyses, as mentioned above, are inappropriate when markers have disparate evolutionary histories. Therefore, here we compare the phylogenetic position of CG lineage determined using both concatenated phylogenetic analyses and species tree approaches (MSC analyses) based on nine nuclear markers to better understand the evolution of this enigmatic taxon. For the nuclear analyses, multiple samples of capped and golden langurs have been used. Furthermore, we also undertook a separate analysis of mtDNA dataset which included a large mitochondrial fragment from a wild caught capped langur to reconfirm its position in the mitochondrial tree. This was important given the past studies either used smaller fragment of mtDNA or included zoo derived specimen of species from CG lineage.

2. Materials and Methods

2.1. Taxon sampling and DNA extraction

Two fecal samples from golden langur and four samples (two fecal and two tissue) from capped langurs were collected non-invasively from different zoos across India (Table S4). Fresh faeces were collected only from wild caught captives, using plant twigs to avoid contamination and preserved in absolute alcohol (Merck, Germany) and stored in -20°C.

For DNA extraction from tissue samples, we used the commercially available DNeasy® Blood & Tissue Kit (QIAGEN Inc.), we followed the manufacturer's protocol. DNA from fecal samples was extracted using the commercially available QIAamp DNA stool mini kit

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(QIAGEN Inc.), following the manufacturer's protocol with slight modifications as mentioned in Mondol et al. (2009), however, we did not add the carrier RNA (Poly A)(Kishore, Reef Hardy, Anderson, Sanchez, & Buoncristiani, 2006; Mondol et al., 2009). Each extraction had a negative control to monitor contamination. The quantity of extracted DNA was measured using a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific Inc).

2.2. PCR Amplification and Sequencing

Nine nuclear markers were selected for this study (Table 1) (Perelman et al., 2011; Karanth et al., 2008). The nuclear markers were amplified using the following conditions. A 25 µl PCR reaction volume was set with 2 µl (total DNA concentration varied between samples ranging from 20 ng/µl to 80 ng/µl) of DNA extract, 0.25 mM of dNTPs (Bangalore Genei, Bangalore), 0.2 µM of each primer (Amnion Biosciences, Bangalore), 1U Taq polymerase (New England BioLabs® Inc.) and a standard 1X reaction buffer premixed with 1.5 mM $MgCl_2$ (New England BioLabs® Inc.). A 'touchdown' PCR program was used with an initial annealing temperature of 60 °C with subsequent decrease by 1 °C per cycle till the specified annealing temperature, 50 °C, is reached. Each reaction was performed for 95 °C for 15 sec, 60-50 °C for 30 sec and 72 °C for 60 sec, with initial denaturation at 95 °C for 10 min. PCR products were purified using the QIAquick® PCR purification kit (QIAGEN Inc.) following the manufacturer's protocol, for some samples we used ExoSAP-IT (Affymetrix®, USB Products). The purified product was outsourced for sequencing to Amnion Biosciences Pvt. Ltd., Bangalore; some of the samples were outsourced to Medauxin, Bangalore.

Additionally, eight mitochondrial markers were amplified from a single capped langur tissue sample (Table 2). First, we used a long-range PCR kit (Expand Long Template PCR System, Roche) to amplify a long fragment of mitochondrial DNA (mtDNA) genome in order to minimise the chances of amplifying a numt sequence. Second, we identified these eight genes by comparing our sequence with the whole mitochondrial (mt) genome of capped langur available in GenBank. Following primer pair was used to amplify the long fragment; 5275F (5' – ACCYCTGTCTTTAGGTTTACAGCCTAATG – 3') and 11718R (5' – CCAATGGATAGCTGTTATCCTTTAAAAGTTGAG – 3') (Raaum, Sterner, Noviello, Stewart, & Disotell, 2005). The PCR cycle conditions followed were as per the manufacturer's protocol (Roche). PCR product was outsourced for purification and

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

sequencing to Amnion Biosciences Pvt. Ltd., Bangalore. We were not able to amplify the long mitochondrial fragment from fecal samples because of the low yield of degraded DNA which is typical of faecal samples. We did not attempt to amplify small mitochondrial fragments for single markers from the fecal samples, to avoid accidental amplification of numt sequences.

2.3. Phylogenetic analysis

The sequence chromatogram file was viewed and edited manually using ChromasLite v2.01 (Technelysium Pty Ltd). We generated a long fragment of mitochondrial genome for one capped langur tissue sample (CES11/299, Table S4). This long fragment of mtDNA consists of eight protein coding loci – Cox1, Cox2, ATP8, ATP6, Cox3, ND3, ND4L, ND4 (Table 2). All the sequences are deposited in GenBank (Table #; sequences will be deposited after the acceptance of manuscript). Each sequence was checked by comparing with orthologous sequences available in GenBank by using Nucleotide BLAST search tool.

The mitochondrial fragment, sequenced from one capped langur (belong to the CG lineage group) tissue sample in this study, was aligned with 19 Asian colobines and one African colobine sequences available in GenBank (Table S1). Downloaded sequences consisted of representatives from odd-nosed monkeys, *Trachypithecus* group, *Semnopithecus* group and *Colobus* used as an African colobine outgroup. Furthermore, the six non-coding tRNAs from the mitochondrial fragment were excluded from the analysis. Nuclear dataset generated in this study i.e. seven sequences from individuals that belong to CG lineage and one sequence from *Semnopithecus hypoleucos*, were aligned with sequences of 13 other Asian and African colobines available in Genbank (Table S2). The nine nuclear genes were concatenated to create a multilocus supermatrix of the nuclear data. The two data sets were aligned separately using Muscle algorithm (Edgar, 2004) incorporated in MEGA v5.2 (Tamura et al., 2011) with default settings. We used p-distance algorithm in MEGA to estimate the evolutionary divergences between sequences.

PartitionFinder 1.1.0 (Lanfear, Calcott, Ho, & Guindon, 2012) was used to pick the best partition scheme as well as the model of sequence evolution. Partitioning schemes used for analysis of our data are listed in Table S3 [(A) and (B)]. Phylogenetic reconstruction was performed using Maximum Likelihood (ML) and Bayesian methods. ML analysis was performed in RAxML 7.4.2 incorporated in raxmlGUIversion 1.3 (Stamatakis, 2006). As there is no provision in RAxML for applying multiple models across partitions, we used GTR

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

+ G as the model of substitution for all partitions in our data set (Table S3). We performed 1000 slow bootstrap replicates to assess support for different nodes. The Bayesian analysis was done in MrBayes 3.2.2 (Ronquist et al., 2012). Nuclear data analysis was run for 15 million generations with sampling frequency at 500 and the mitochondrial dataset was run for 15 million generations and sampling frequency at 2000. Analyses had two parallel runs with four chains, convergence of the runs was determined when the standard deviation of split frequencies was <0.01. Using Tracer v1.6 (Rambaut, A., Suchard, M.A., Xie, D., Drummond, 2013) we plotted each parameter against generation time to determine the effective sample size (ESS) value of >200. The first 25% trees were discarded as burn-in.

2.4. Species tree estimation

Evolutionary processes like hybridisation, incomplete lineage sorting (ILS), gene duplication and gene loss make it difficult to estimate species trees from multi-locus nuclear data. These processes can make gene trees incongruent and different from the overall species tree (Maddison, 1997). The individual gene trees from the nine nuclear markers were uninformative due to lack of phylogenetically informative sites; therefore, to infer a nuclear marker based species tree we used two different approaches. First the program ASTRAL II (Mirarab and Warnow, 2015) was used to build species tree for the nuclear dataset. It uses individual nuclear gene trees as input file and estimates a phylogeny that agrees with the maximum quartet trees induced by the gene tree set. The command "java -jar astral.4.10.11.jar –I in.tree -o out.tre" was used to run the input file and to get an output tree file. Input tree file was generated by combining individual gene tree outputs from RAxML.

ASTRAL is a 'summary statistic' method which uses input gene trees to estimate species tree. However, summary methods are known to be sensitive to gene tree estimation error, specifically on short alignments (Chou et al., 2015). The program SVDquartets is a 'single site' method that first builds a quartet tree for all possible quartet combinations; for all the possible trees in a quartet, SVD (single value decomposition) score is calculated and the relationship with the lowest score is picked up as the best tree. It then uses a quartet assembly technique, like QFM (Reaz, Bayzid, & Rahman, 2014) to build a tree. We ran the program with 100000 quartets to evaluate with the option to check all the possible quartets in PAUP* 4.0a157 (Swofford, 2002). Two approaches were used; first species tree analysis was undertaken with "taxa partition" option wherein individual members were assigned to major lineages recovered from the concatenation analysis (Fig. 1 & Fig. 2) and secondly, the

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

analysis was run without taxon partitions where the tips were designated as separate populations rather than grouping them in major lineages. For both the analysis, we used the multispecies coalescent tree model and the analyses were subjected to 100 bootstrap replicates.

2.5. Divergence dating

We used the species tree ancestral reconstruction - *BEAST (Joseph Heled, Bouckaert, Drummond, & Xie, 2013) incorporated in BEAST v2.4.7 (Bouckaert et al., 2014) for divergence time estimation for our nuclear data set. Input files were created in BEAUti v2.4.7 (Bouckaert et al., 2014). The selected partitioning scheme is described in Table S3(A). We used an uncorrelated relaxed lognormal clock for each gene with default values for all the priors. Yule model was used as species tree prior. It is important to set the correct ploidy levels as differences in effective population size influences the coalescence time (Charlesworth, 2009). All the genes in our analysis are autosomal nuclear except for FAM123B, which is an X-linked locus; so, the ploidy levels were selected accordingly. We *a priori* grouped the tips into their respective genera; individuals from CG lineage were assigned to a separate group (Table S5). Two independent analyses were run for 80 million generations, sampling every 5000 generations. Stationarity was assessed in Tracer v1.6 (Rambaut et al., 2013) based on the plots of parameters vs. state or generation and the effective sample sizes (>200). The tree files were combined using LogCombiner 2.4.7 and constructed maximum clade credibility tree with median height using TreeAnnotator v2.4.7 after discarding first 25% trees as burn-in.

We used two fossil calibrations and one secondary calibration for dating the phylogeny. A previous study (Perelman et al., 2011) inferred the most recent common ancestor (MRCA) for Asian colobine at 8.8mya with 95% credible interval (CI) of 6.5 – 11.2 mya. Therefore, the molecular clock prior was set to normal with mean of 8.8. One of the fossils we used for calibration is a primate from middle to late Miocene Eurasia, *Mesopithecus pentelicus* (Pan, Groves, & Oxnard, 2004). To the best of our knowledge, it has never been used as a calibration for any primate phylogeny. According to Pan et al. (2004), this fossil is related to the odd-nosed colobine genus *Rhinopithecus* and *Nasalis*. According to our ML and Bayesian phylogenies, *Pygathrix* is placed sister to the *Rhinopithecus* + *Nasalis* clade, hence we constrained the *Rhinopithecus* + *Nasalis* + *Pygathrix* clade as monophyletic and calibrated this node using lognormal prior distribution with mean = 9.0, offset = 6.0.

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Another fossil we used is a late Pliocene *Semnopithecus gwebinensis* fossil discovered from the Irrawaddy sediments of central Myanmar (Takai et al., 2016). For this calibration, we constrained *Semnopithecus* + *Trachypithecus* + CG lineage as monophyletic and used a lognormal distribution prior with mean = 12.0 and offset = 2.6.

3. Results

3.1. Phylogenetic Analysis

3.1.1. Nuclear phylogeny

Our final concatenated nuclear dataset contained 3307 bp from nine nuclear genes: ABCA1 (324 bp), BCHE (244 bp), BDNF (468bp), DMRT1 (530 bp), ERC2 (333 bp), FAM123B (369 bp), FES (479 bp), LZM (136) and MAPKAP (396 bp) for 20 individuals.

In the Bayesian phylogeny (Fig. 1), the CG lineage was sister to *Trachypithecus* with a strong support (PP = 1). The *Semnopithecus hypoleucos* sequence from this study branched with the rest of the *Semnopithecus* individuals. Furthermore, *Semnopithecus* and *Trachypithecus* were sister to each other to the exclusion of the odd-nosed monkey group. Relationships within these major clades are not completely resolved and lack good support. The ML tree (not shown) also retrieved similar relationships with bootstrap support (BS = 98) for the CG + *Trachypithecus* clade.

3.1.2. Mitochondrial Phylogeny

The mitochondrial dataset consists of a large fragment (5933 bp) containing eight mitochondrial genes (Table 2) from 21 individuals. In the Bayesian tree (Fig. 2), capped langur (from the CG lineage group) was sister to *T. shortridgei* and together these two species are sister to other *Trachypithecus* species from Southeast Asia (PP=1). Furthermore, *Presbytis melalophos*, instead of *Semnopithecus*, was sister to *Trachypithecus*. Here we see discordance between mtDNA and nuclear dataset with respect to relationships between *Semnopithecus*, *Trachypithecus* and *Presbytis* (For further discussion see Roos et al., 2011). The topology in the likelihood tree (BS = 100; not shown) was similar to the Bayesian tree. Thus, both nuclear and mitochondrial markers suggested sister relationship between CG lineage and *Trachypithecus*. We also conducted a separate analysis of the mtDNA dataset that included the previously reported numts by Wang et al. (2015). In this tree, the mtDNA fragment generated in this study did not branch with the numts (tree not shown).

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Additionally, our sequence did not have any of the typical changes associated with numts, such as indels that cause frame shift, nonsense mutations etc.

3.2. Species tree estimation

Our coalescent based multilocus analysis of the nuclear dataset generated contradictory results. In the SVDquartet analysis, where multispecies coalescent model was used without taxon partitions, CG lineage was sister to *Trachypithecus* (BS = 36) with the exclusion of *Semnopithecus*. However, when "taxon partitions" option was used, the position of CG lineage was unresolved with respect to *Trachypithecus* and *Semnopithecus*. Whereas, in the result from the ASTRAL tree, *Semnopithecus* was sister to *Trachypithecus* with the exclusion of CG lineage with low branch support (PP = 0.42). (Figure 3).

3.3. Divergence Dating

Molecular dates are shown in Fig. 4. These age estimates show the split of African and Asian colobines to be 9.5 mya with CI of 12.6 – 7.09 mya. Within the Asian colobines, the odd-nosed monkeys are estimated to have diverged from the langurs in late Miocene between 9.8 and 6.7 mya. The split between *Semnopithecus* and *Trachypithecus* is estimated to be between 5.2 and 2.7 mya. The date estimates in this study are younger than estimated in an older study using nuclear concatenated dataset (Perelman et al., 2011). Divergence of CG lineage from the rest of *Trachypithecus* occurred 2.5 mya with CI of 4.0 -1.2 mya. The ESS values for all the parameters were >>200 for the two independent BEAST runs suggesting stationarity.

4. Discussion

Previous studies that have attempted to resolve the phylogenetic placement of CG lineage were based on single nuclear and / or mitochondrial markers (Karanth et al., 2008; Osterholz et al., 2008; Wangchuk et al., 2008). Among these studies, Karanth et al. (2008) and Osterholz et al. (2008) reported incongruence between nuclear and mitochondrial markers which was taken as evidence for hybrid origin of CG lineage. However, the study by Wang et al. (2015) suggested that this incongruence might be due to the use of numts instead of true mitochondrial sequences. In the present study, we used multiple nuclear and mitochondrial markers to resolve the phylogenetic position of CG lineage. Additionally, we also implemented both concatenated analysis and coalescent based species tree building methods to better understand the origin of CG lineage (Fig. 3).

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Phylogenetic tree obtained from the nuclear dataset showed the CG lineage to be monophyletic. However, within CG lineage, neither capped or golden langur was found to be monophyletic. This could be because these are recently diverged taxa and the nuclear DNA does not contain enough information to separate out these two lineages. The relationship within CG lineage can further be resolved by using fast evolving nuclear markers, such as microsatellites. Additionally, adding more samples to the phylogeny might also help resolve the relationship within the CG lineage.

The nuclear as well as the mitochondrial concatenated analyses placed CG lineage sister to *Trachypithecus* with high support. However, the nuclear species trees supported conflicting topologies. Among them only SVDquartet analysis (without taxon partitions) recovered the above relationship but with very low support (Fig 3, A). Whereas ASTRAL and SVDquartet (with taxon partitions) supported alternate scenarios again with low support. Thus, the species trees are unable to resolve the relationship between *Trachypithecus*, *Semnopithecus* and CG lineage. The low support for relevant nodes in the species trees is suggestive of different evolutionary histories of these nuclear markers due to processes like incomplete lineage sorting. Whereas the high support for CG lineage + *Trachypithecus* monophyly in the nuclear concatenated analysis might be due to few highly variable markers overwhelming the dataset. Additionally, under conditions of low ILS, the species trees are identical to trees based on concatenated analysis (Chou et al., 2015) which is not the case with our results.

Collectively, these results suggest that the phylogenetic position of CG lineage remains unresolved. The convoluted evolutionary history of the CG nuclear makers, we believe, is due to hybridization rather than ILS. This is because, multispecies coalescent models incorporate ILS in species tree estimation (Chou et al., 2015) but these methods cannot distinguish between hybridisation and ILS (Kubatko, 2009). There are three additional lines of evidence that support this scenario. First the numts analyses by Karanth (2008) and Wang et al. (2015) strongly support past hybridization between *Semnopithecus* and *Trachypithecus*. Secondly, the CG lineage is distributed in an area that abuts the eastern and western distributional limits of *Semnopithecus* and *Trachypithecus* respectively. Recent discovery of the late Pliocene fossil of *Semnopithecus gwebinensis* from central Myanmar (Takai et al., 2015) suggests that *Semnopithecus* langurs were distributed till central Myanmar in late Pliocene, whereas today the eastern most distribution of *Semnopithecus* is Bangladesh and Bhutan. This implies that there was an overlap in the distribution of

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Semnopithecus and *Trachypithecus* in the past. Thirdly, the mean body weight (Wang et al., 2015), and body size and skull measurements of CG lineage is intermediate between *Semnopithecus* and *Trachypithecus* (Table S6). Presence of morphologically intermediate forms can be an indication of possible hybridization (Zinner et al., 2011). Such intermediate forms have been seen in other primates like baboons where hybridization has been reported (Phillips-conroy & Jolly, 1981).

Wang et al. (2015) suggest that this hybridization could have been initiated by the bigger *Semnopithecus* males which have an advantage over the smaller *Trachypithecus* males for access to *Trachypithecus* females. The genetic data is consistent with this scenario wherein the mtDNA of CG lineage is related to *Trachypithecus* whereas the nuclear DNA appears to be of both *Semnopithecus* and *Trachypithecus* origin. Our dating analysis (Fig. 4) shows the divergence between *Semnopithecus* and the *Trachypithecus* + CG clade to be around early to mid-Pliocene (3.9 mya; CI – 5.2 mya to 2.7 mya) and the divergence between *Trachypithecus* and CG lineage occurred in late Pliocene (2.6 mya; CI – 4.0 mya to 1.2 mya). Thus, this hybridisation event between lineages of *Semnopithecus* and *Trachypithecus* might have occurred during the Pliocene to early Pleistocene period (5.2 mya to 1.2 mya).

5. Conclusion

Concatenation based analysis of nuclear markers yielded tree with good node support for the placement of CG lineage sister to *Trachypithecus*. However, results from different MSC method based analyses were incongruent with each other and with the concatenated tree suggesting ILS/ hybridization in the data set. The low node support in the trees for the node of interest (Fig. 3) could also arise due to gene tree estimation error owing to splitting the overall dataset into small partitions (Gatesy & Springer, 2013; Meredith et al., 2011; Simmons & Gatesy, 2015). Further analysis using longer gene sequences to build individual gene trees, can help resolve this issue. Nevertheless, we suggest the use of both the concatenation and coalescent methods for a robust understanding of various questions in evolutionary biology and systematics.

Overall results from this study and previous studies, suggests that CG lineage might have evolved as a result of hybridisation between *Semnopithecus* and *Trachypithecus*. The genetic evidence for the reticulate evolution of CG lineage needs to be further fortified with more nuclear markers. Nevertheless, the discordance between concatenated and species tree approaches with respect to the position of CG lineage implies that this taxon has a unique

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

evolutionary history. Currently, the CG lineage is placed in the genus *Trachypithecus*, but more studies need to be undertaken to resolve its taxonomic position.

Supplementary Information

Supporting data is in the file Supplementary information.docx

Acknowledgments

We thank the Department of Biotechnology, Govt. of India and the Ministry of Environment Forest and Climate Change for the funds that covered the molecular work and a part of fieldwork for this study. The forest department of Assam provided the permits to collect samples. We thank Guwahati zoo for allowing us to collect the samples. Dr. Senthil Kumar and Dr. H.T. Lalremsanga from Mizoram University for providing samples from Tripura zoo and Mizoram zoo. Dr. G. Umapathy from LaCONES, CCMB for help in sample collection from Hyderabad zoo. KA would like to thank Mr. Rajani Deka and Mr. Firoz Ahmed for support during field work. Aniruddha Datta-Roy, Ishan Agarwal and V. Deepak for comments on the manuscript. Bhavani from CES for help in accounting.

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

Ashalakshmi, N. C., Nag, K. S. C., & Karanth, K. P. (2014). Molecules support morphology: species status of South Indian populations of the widely distributed Hanuman langur. *Conservation Genetics*, *16*(1939), 43–58. https://doi.org/10.1007/s10592-014-0638-4

Bensasson, D., Zhang, D. X., Hartl, D. L., & Hewitt, G. M. (2001). Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology and Evolution*, *16*(6), 314–321. https://doi.org/10.1016/S0169-5347(01)02151-6

Blanchard, J. L., & Lynch, M. (2000). Organellar genes why do they end up in the nucleus? *Trends in Genetics*, *16*(7), 315–320. https://doi.org/10.1016/S0168-9525(00)02053-9

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., … Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, *10*(4), 1–6. https://doi.org/10.1371/journal.pcbi.1003537

Brandon-Jones, D., Eudey, A. A., Geissmann, T., Groves, C. P., Melnick, D. J., Morales, J. C., … Stewart, C.-B. B. (2004). Asian Primate Classification. *International Journal of Primatology*, *25*(1), 97–164. https://doi.org/10.1023/B:IJOP.0000014647.18720.32

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & Roychoudhury, A. (2012). Inferring Species Trees Directly from Biallelic Genetic Markers□: Bypassing Gene Trees in a Full Coalescent Analysis Research article. *Molecular Biology and Evolution*, *29*(1998), 1917–1932. https://doi.org/10.1093/molbev/mss086

Buckley, T. R., Cordeiro, M., Marshall, D. C., & Simon, C. (2006). Differentiating between Hypotheses of Lineage Sorting and Introgression in New Zealand Alpine Cicadas (Maoricicada Dugdale). *Syst. Biol.*, *55*(3), 411–425.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205. https://doi.org/10.1038/nrg2526

Chifman, J., & Kubatko, L. (2015). Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology*, *374*, 35–47. https://doi.org/10.1016/j.jtbi.2015.03.006

Chivers, D. J., & Hladik, C. M. (1980). Morphology of the gastrointestinal tract in primates□: Comparisons with other mammals in relation to diet. *Journal of Morphology*, *166*(3), 337–386. https://doi.org/10.1002/jmor.1051660306

Chou, J., Gupta, A., Yaduvanshi, S., Davidson, R., Nute, M., Mirarab, S., & Warnow, T. (2015). A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, *16*(Suppl 10), S2. https://doi.org/10.1186/1471-2164-16-S10-S2

Collura, R. V, & Stewart, C.-B. (1995). Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. *Nature*, *378*(6556), 485–489. Retrieved from http://dx.doi.org/10.1038/378485a0

Davenport, T. R. B., Stanley, W. T., Sargis, E. J., De Luca, D. W., Mpunga, N. E., Machaga, S. J., & Olson, L. E. (2006). A new genus of african monkey, \emphRungwecebus: morphology, ecology, and molecular phylogenetics. *Science*, *312*(June), 1372–1374. https://doi.org/10.1126/science.1125631

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Davies, G., & Oates, J. (1994). *Colobine Monkeys: Their Ecology, Behaviour and Evolution*. Cambridge University Press. Retrieved from https://books.google.co.in/books?id=xfMT7KD5STMC

de Querioz, A., & Gatesy, J. (2006). The supermatrix approach to systematics. *Trends in Ecology & Evolution*, *22*(1), 34–41. https://doi.org/10.1016/j.tree.2006.10.002

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, *24*(6), 332–340. https://doi.org/10.1016/j.tree.2009.01.009

Degnan James H., Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees: A unifying principle. *PLoS Genetics*, *3*, 762–768. https://doi.org/10.1093/molbev/mst160

Dunbar, R. I. M., & Dunbar, P. (1974). On hybridization between Theropithecus gelada and Papio anubis in the wild. *Journal of Human Evolution*, *3*(3), 187–192. https://doi.org/10.1016/0047-2484(74)90176-6

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edwards, S. V., Liu, L., & Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences*, *104*(14), 5936–5941. https://doi.org/10.1073/pnas.0607004104

Ford Doolittle, W. (1998). You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics*, *14*(8), 307–311. https://doi.org/10.1016/S0168-9525(98)01494-2

Gatesy, J., & Springer, M. S. (2013). Concatenation versus coalescence versus "concatalescence." *Proceedings of the National Academy of Sciences*, *110*(13), E1179–E1179. https://doi.org/10.1073/pnas.1221121110

Gauthier, O., & Lapointe, F. (2007). Hybrids and Phylogenetics Revisited□: A Statistical Test of Hybridization Using Quartets Published by□: American Society of Plant Taxonomists Stable URL□: http://www.jstor.org/stable/25064223 REFERENCES Linked references are available on JSTOR for this ar. *Systematic Botany*, *32*(1), 8–15.

Groves, C. P. (2001). *Primate Taxonomy*. Smithsonian Institution Press.

Heled, J., Bouckaert, R., Drummond, A. J., & Xie, W. (2013). * BEAST in BEAST 2 . 0 Estimating Species Trees from Multilocus Data, 1–18.

Heled, J., & Drummond, A. J. (2010). Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution*, *27*(3), 570–580. https://doi.org/10.1093/molbev/msp274

Jin, G., Nakhleh, L., Snir, S., & Tuller, T. (2006). Maximum likelihood of phylogenetic networks. *Bioinformatics*, *22*(21), 2604–2611. https://doi.org/10.1093/bioinformatics/btl452

Joly, S., McLenachan, P. A., & Lockhart, P. J. (2009). A Statistical Approach for Distinguishing Hybridization and Incomplete Lineage Sorting. *The American Naturalist*, *174*(2), E54–E70. https://doi.org/10.1086/600082

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Jones, T., Ehardt, C. L., Butynski, T. M., Davenport, T. R. B., Mpunga, N. E., Machaga, S. J., & De Luca, D. W. (2005). The Highland Mangabey Lophocebus kipunji□: A New Species of African Monkey. *Science*, *308*, 1161–1165. https://doi.org/10.1126/science.1109191

Karanth, K. P. (2008). Primate numts and reticulate evolution of capped and golden leaf monkeys (Primates: Colobinae). *J. Biosci*, *33*, 761–770.

Karanth, K. P., Singh, L., Collura, R. V., & Stewart, C.-B. (2008). Molecular phylogeny and biogeography of langurs and leaf monkeys of South Asia (Primates: Colobinae). *Molecular Phylogenetics and Evolution*, *46*(2), 683–694. https://doi.org/10.1016/j.ympev.2007.11.026

Kishore, R., Reef Hardy, W., Anderson, V. J., Sanchez, N. A., & Buoncristiani, M. R. (2006). Optimization of DNA extraction from low-yield and degraded samples using the BioRobot EZ1 and BioRobot M48. *Journal of Forensic Sciences*, *51*(5), 1055–1061. https://doi.org/10.1111/j.1556-4029.2006.00204.x

Knowles, L.L, Kubatko, L. S. (2010). *Estimating Species Trees: Practical and Theoretical Aspects*. (L. S. Knowles, L.L, Kubatko, Ed.). Wiley & Sons, Hoboken, New Jersey.

Kubatko, L. S. (2009). Identifying Hybridization Events in the Presence of Coalescence via Model Selection. *Syst. Biol.*, *58*(5), 478–488. https://doi.org/10.1093/sysbio/syp055

Kubatko, L. S., Carstens, B. C., & Knowles, L. L. (2009). STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, *25*(7), 971–973. https://doi.org/10.1093/bioinformatics/btp079

Kumar, A., & Solanki, G. S. (2008). Population Status and Conservation of Capped Langurs (Trachypithecus pileatus) in and Around Pakke Wildlife Sanctuary, Arunachal Pradesh, India. *Primate Conservation*, *23*(1), 97–105. https://doi.org/10.1896/052.023.0111

Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, *29*(6), 1695–1701. https://doi.org/10.1093/molbev/mss020

Leache, A. D., & Rannala, B. (2011). Society of Systematic Biologists The Accuracy of Species Tree Estimation under Simulation□: A Comparison of Methods. *Syst. Biol.*, *60*(2), 126–137.

Liu, L., & Pearl, D. K. (2007). Species Trees from Gene Trees□: Reconstructing Bayesian Posterior Distributions of a Species Phylogeny Using Estimated Gene Tree Distributions. *Syst. Biol.*, (56), 504–514.

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 25–27.

Maddison, W. (1997). Gene Trees in Species Trees. *Syst. Biol.*, *46*(3), 523–536. https://doi.org/10.2307/2413694

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology and Evolution*, *20*(5), 229–237. https://doi.org/10.1016/j.tree.2005.02.010

Mallet, J. (2007). Hybrid speciation. *Nature*, *446*(7133), 279–283. https://doi.org/10.1038/nature05706

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Meng, C., & Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology*, *75*(1), 35–45. https://doi.org/10.1016/j.tpb.2008.10.004

Meredith, R. W., Janec, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., … Murphy, W. J. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg Extinction on Mammal Diversification. *Science*, *334*, 521–525.

Mirarab, S., & Warnow, T. (2015). ASTRAL-II□: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*, 44–52. https://doi.org/10.1093/bioinformatics/btv234

Mondol, S., Ullas Karanth, K., Samba Kumar, N., Gopalaswamy, A. M., Andheria, A., & Ramakrishnan, U. (2009). Evaluation of non-invasive genetic sampling methods for estimating tiger population size. *Biological Conservation*, *142*(10), 2350–2360. https://doi.org/10.1016/j.biocon.2009.05.014

Mossel, E., & Roch, S. (2008). Incomplete lineage sorting: Consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *7*(1), 166–171. https://doi.org/10.1109/TCBB.2008.66

Olave, M., Avila, L. J., Sites, J. W., & Morando, M. (2017). Detecting hybridization by likelihood calculation of gene tree extra lineages given explicit models. *Methods in Ecology and Evolution*, *2017*(June), 1–13. https://doi.org/10.1111/2041-210X.12846

Osterholz, M., Walter, L., & Roos, C. (2008). Phylogenetic position of the langur genera Semnopithecus and Trachypithecus among Asian colobines, and genus affiliations of their species groups. *BMC Evolutionary Biology*.

Pan, R., Groves, C., & Oxnard, C. (2004). Relationships between the Fossil Colobine Mesopithecus pentelicus and Extant Cercopithecoids, Based on Dental Metrics. *American Journal of Primatology*, *62*(4), 287–299. https://doi.org/10.1002/ajp.20022

Pastorini, J., Zaramody, A., Curtis, D. J., Nievergelt, C. M., & Mundy, N. I. (2009). Genetic analysis of hybridization and introgression between wild mongoose and brown lemurs. *BMC Evolutionary Biology*, *9*, 32. https://doi.org/10.1186/1471-2148-9-32

Perelman, P., Johnson, W. E., Roos, C., Seuánez, H. N., Horvath, J. E., Moreira, M. A. M., … Pecon-Slattery, J. (2011). A Molecular Phylogeny of Living Primates. *PLoS Genetics*, *7*(3), e1001342. https://doi.org/10.1371/journal.pgen.1001342

Phillips-conroy, J. E., & Jolly, C. J. (1981). Sexual Dimorphism in Two Subspecies of Ethiopian Baboons (Papio hamadryas) and Their Hybrids, *129*, 115–129.

Posada, D. (2002). Evaluation of Methods for Detecting Recombination from DNA Sequences: Empirical Data. *Molecular Biology and Evolution*, *19*(5), 708–717. https://doi.org/10.1093/oxfordjournals.molbev.a004129

Raaum, R. L., Sterner, K. N., Noviello, C. M., Stewart, C.-B., & Disotell, T. R. (2005). Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. *Journal of Human Evolution*, *48*(3), 237–257. https://doi.org/10.1016/j.jhevol.2004.11.007

Rabarivola, C., Meyers, D., & Rumpler, Y. (1991). Distribution and morphological characters of intermediate forms between the black lemur (Eulemur macaco macaco) and the Sclater's lemur (E. m. flavifrons). *Primates*, *32*(2), 269–273.

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

https://doi.org/10.1007/BF02381186

Ram, M. S., Kittur, S. M., Biswas, J., Nag, S., Shil, J., & Umapathy, G. (2016). Genetic Diversity and Structure among Isolated Populations of the Endangered Gees Golden Langur in Assam , India. *PLoS ONE*, *11*(8), e0161866. https://doi.org/10.1371/journal.pone.0161866

Rambaut, A., Suchard, M.A., Xie, D., Drummond, A. J. (2013). Tracer v1.6. <http://beast.bio.ed.ac.uk/software/tracer/>

Reaz, R., Bayzid, M. S., & Rahman, M. S. (2014). Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLoS ONE*, *9*(8). https://doi.org/10.1371/journal.pone.0104008

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., … Huelsenbeck, J. P. (2012). Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, *61*(3), 539–542. https://doi.org/10.1093/sysbio/sys029

Roos, C., Zinner, D., Kubatko, L. S., Schwarz, C., Yang, M., Meyer, D., … Ziegler, T. (2011). Nuclear versus mitochondrial DNA□: evidence for hybridization in colobine monkeys. *BMC Evolutionary Biology*, *11*(1), 77. https://doi.org/10.1186/1471-2148-11-77

Simmons, M. P., & Gatesy, J. (2015). Coalescence vs. concatenation: Sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular Phylogenetics and Evolution*, *91*, 98–122. https://doi.org/10.1016/j.ympev.2015.05.011

Song, S., Liu, L., Edwards, S. V, & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sci U S A*, *109*(37), 14942–14947. https://doi.org/10.1073/pnas.1518753112

Srivastava, A., & Mohnot, S. M. (2001). Distribution, Conservation Status and Priorities for Primates in Northeast India. *Envis Bulletin: Wildlife and Protected Areas*, *1*, 102-108.

Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), 2688–2690. https://doi.org/10.1093/bioinformatics/btl446

Swofford, D. (2001). PAUP*—Phylogenetic Analysis Using Parsimony and Other Methods, Version 4. Sinauer Associates, Sunderland, MA.

Takai, M., Nishioka, Y., Thaung-Htike, Maung, M., Khaing, K., Zin-Maung-Maung-Thein, … Egi, N. (2016). Late Pliocene Semnopithecus fossils from central Myanmar: rethinking of the evolutionary history of cercopithecid monkeys in Southeast Asia. *Historical Biology*, *28*(1–2), 172–188. https://doi.org/10.1080/08912963.2015.1018018

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*(10), 2731–2739. https://doi.org/10.1093/molbev/msr121

Than, C., Ruths, D., Innan, H., & Nakhleh, L. (2007). Confounding Factors in HGT Detection: Statistical Error, Coalescent Effects, and Multiple Solutions. *Journal of Computational Biology*, *14*(4), 517–535. https://doi.org/10.1089/cmb.2007.A010

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Vasey, N., & Tattersall, I. (2002). Do ruffed lemurs form a hybrid zone? Distribution and discovery of Varecia, with systematic and conservation implications. *American Museum Novitates*, *3376*(3376), 1–26. https://doi.org/10.1206/0003-0082(2002)376<0001:DRLFAH>2.0.CO;2

Wang, B., Zhou, X., Shi, F., Liu, Z., Roos, C., Garber, P. A., … Pan, H. (2015). Full-length *Numt* analysis provides evidence for hybridization between the Asian colobine genera *Trachypithecus* and *Semnopithecus*. *American Journal of Primatology*, *77*(8), 901–910. https://doi.org/10.1002/ajp.22419

Wangchuk, T., Inouye, D. W., & Hare, M. P. (2008). The Emergence of an Endangered Species: Evolution and Phylogeny of the Trachypithecus geei of Bhutan. *International Journal of Primatology*, *29*, 565-582.

Xu, S. (2000). Phylogenetic Analysis Under Reticulate Evolution. *Mol Biol Evol*, *17*(6), 897–907. Retrieved from http://mbe.oxfordjournals.org/cgi/content/abstract/17/6/897

Zhang, Y. Z., Wang, S., & Quan, G. Q. (1981). On the geographical distribution of primates in China. *Journal of Human Evolution*, *10*(3), 219–226. https://doi.org/10.1016/S0047-2484(81)80060-7

Zhong, B., Liu, L., & Penny, D. (2014). The multispecies coalescent model and land plant origins: a reply to Springer and Gatesy. *Trends in Plant Science*, *19*(5), 270–272. https://doi.org/10.1016/j.tplants.2014.02.011

Zinner, D., Arnold, M. L., & Roos, C. (2011). The strange blood: Natural hybridization in primates. *Evolutionary Anthropology*, *20*(3), 96–103. https://doi.org/10.1002/evan.20301

Zinner, D., Chuma, I. S., Knauf, S., & Roos, C. (2018). Inverted intergeneric introgression between critically endangered kipunjis and yellow baboons in two disjunct populations. *Biology Letters*, *14*(1). https://doi.org/10.1098/rsbl.2017.0729

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure legends

Fig. 1: Bayesian phylogeny of Asian colobines for the concatenated nuclear data set. The numbers on the nodes are the posterior probability values/Maximum-likelihood bootstrap support. Support is not shown for those nodes which had both the Posterior probability and ML bootstrap less than 0.75 and 75, respectively. (* indicates Posterior probability or ML bootstrap values less than 0.75 or 75 respectively for that particular node only) CL = capped langur; GL = golden langur. Support for all the nodes within the CL and GL clade was below 0.75/75.

Fig. 2: Bayesian phylogeny of Asian colobines for the mitochondrial concatenated data set. The values at the nodes denote the posterior probability values/Maximum-likelihood bootstrap support. Support is not shown for those nodes which had both the Posterior probability and ML bootstrap less than 0.75 and 75, respectively. (* indicates Posterior probability or ML bootstrap values less than 0.75 or 75 respectively for that particular node only). In this tree, the capped langur sequence was generated in this study and the *T. pileatus* was downloaded from GenBank (KF680163)

Fig. 3: The phylogenetic position of CG lineage with respect to *Semnopithecus* and *Trachypithecus* – from different analyses. (A) species-tree topology from SVDq without taxa partitions (for details, see text), (B) topology estimated with SVDq with taxa partitions (for details, see text), (C) species-tree topology estimated with ASTRAL-II. Numbers at the node are support for that node (support values only for node of interest are shown). Panel A & B shows the BS values; panel C shows PP values.

Fig. 4: Divergence time tree constructed using coalescent approach in *BEAST using the nuclear data set. The value above each node indicates the mean divergence time and the grey bars indicate the 95% credible intervals. PP support for all the nodes was above 0.95 (except where indicated).

701 **Tables**

702 **Table 1:** Nuclear markers used in this study with primers and source. * indicates primer pairs that have not been redesigned.

Marker	Name	Product Size (bp)	Forward Primer (5' to 3')	Reverse Primer (5' to 3')	Source
ABCA1	ATP□binding cassette, sub□family A (ABC1), member 1	400	CCTCCATCTTTTCAGCTCTACCTAC	GGGTTCTAGAGCCTCAATCACAG	Perelman et al., 2011
BCHE	butyrylcholinesterase	244	TGGCTTCAGCAAAGACAACA	GGCATTATTTCCCCATTCTG	Perelman et al., 2011
BDNF*	brain□derived neurotrophic factor	570	CATCCTTTTCCTTACTATGGTT	TTCCAGTGCCTTTTGTCTATG	Perelman et al., 2011
DMRT1*	doublesex and mab□3 related transcription factor 1	560	ATCCCTTGTTCTGAGTGCCA	ACATTGCAAAGACCCCTGAC	Perelman et al., 2011
ERC2	ELKS/RAB6-interacting/CAST family member 2	400	TGTCTTAGGTCTGCTTGTCACAT	ATTCGGTCTGCAAACTCAGG	Perelman et al., 2011
FAM123B	family with sequence similarity 123B	388	TGGTGGGAAAAGCTGAGGTA	AACGGCTAGTGACCATCCAG	Perelman et al., 2011
FES*	feline sarcoma oncogene	500	GGGGAACTTTGGCGAAGTGTT	TCCATGACGATGTAGATGGG	Perelman et al., 2011
LZM	lysozyme	500	GAATCAGTAGATCAATACACAGT	CCTAGCACTAGAACAGTACA	Karanth et al., 2008
MAPKAP	mitogen□activated protein kinase associated protein 1	400	TCAAGTGATAACGGCTGTAGCTT	GAGAGATAGAGCTCCTGACATCAA	Perelman et al., 2011

bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

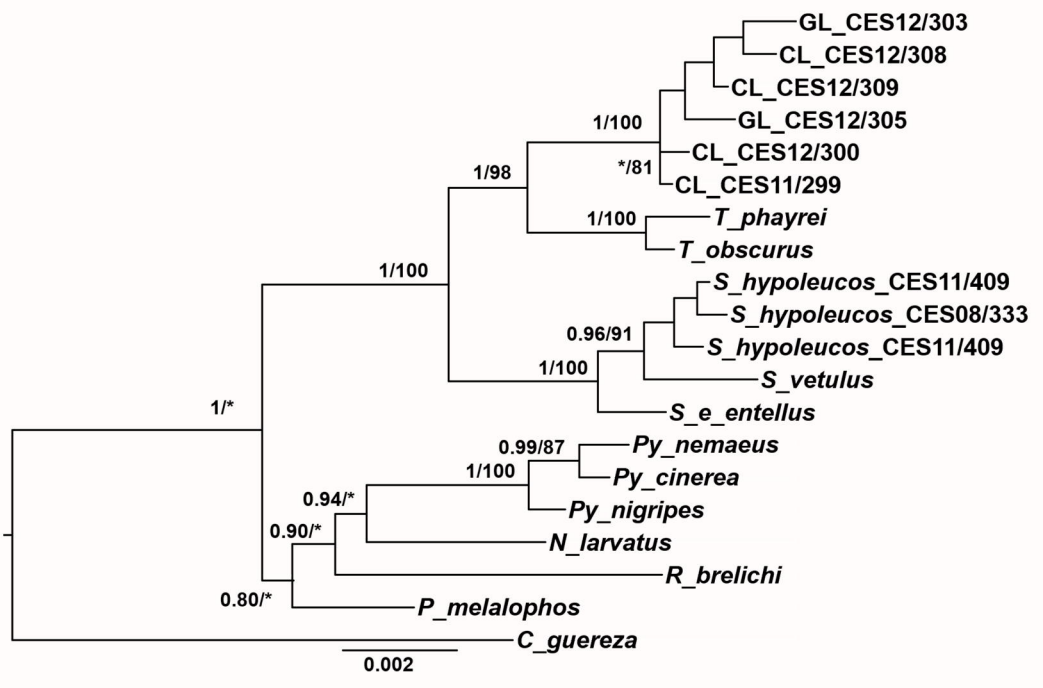
bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

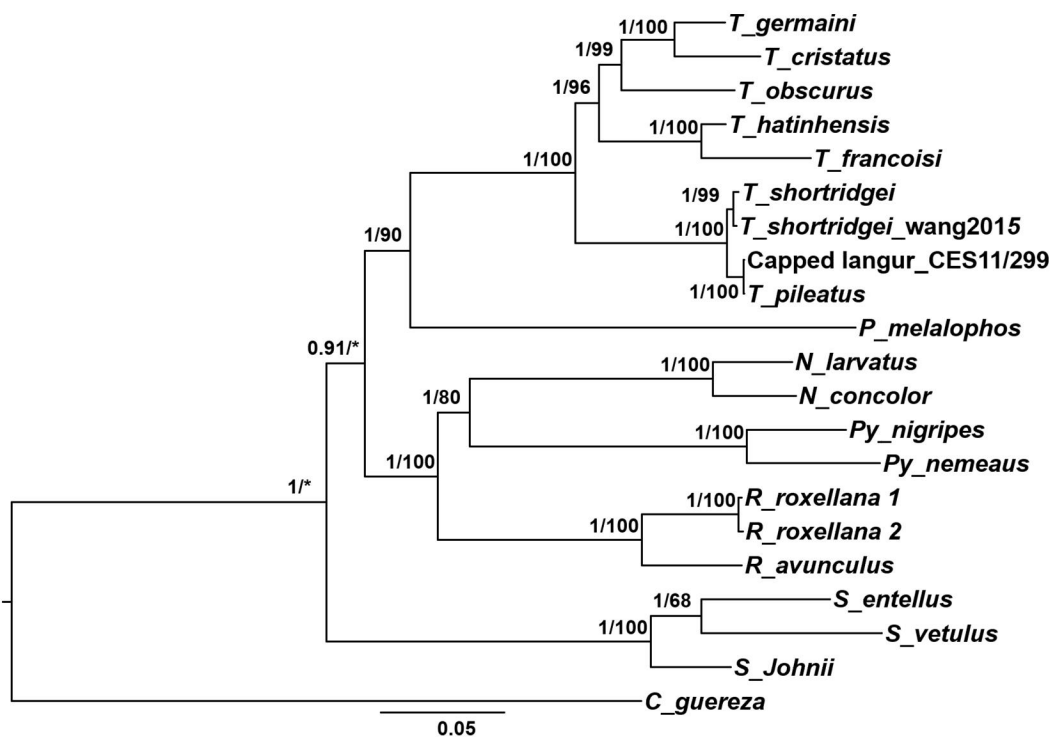
Table 2: Eight genes that are amplified in the long mitochondrial DNA (mtDNA) fragment (5933 bp).

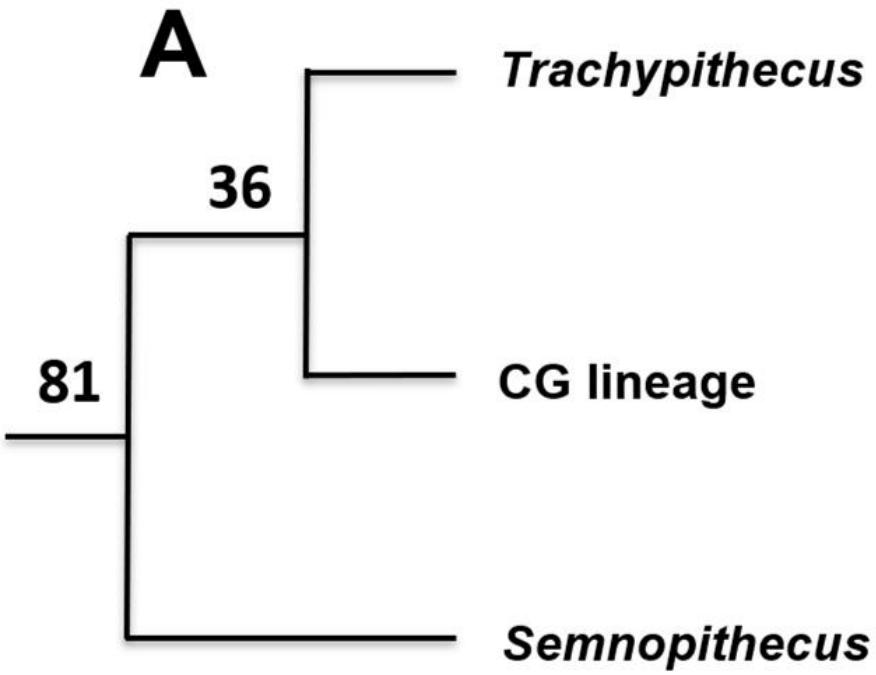
Gene	Name	Length (bp)
COX1	Cytochrome c oxidase subunit 1	1546
COX2	Cytochrome c oxidase subunit 2	688
ATP8	ATP synthase protein 8	211
ATP6	ATP synthase protein 6	681
COX3	Cytochrome c oxidase subunit 3	785
ND3	NADH dehydrogenase 3	346
ND4L	NADH-ubiquinone oxidoreductase chain 4L	297
ND4	NADH dehydrogenase 4	1379

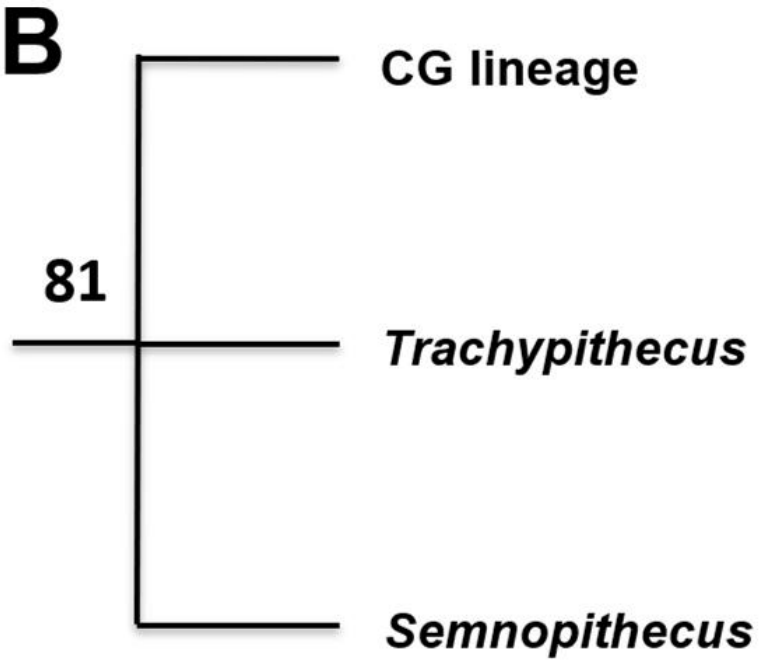
bioRxiv preprint doi: https://doi.org/10.1101/508929; this version posted March 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

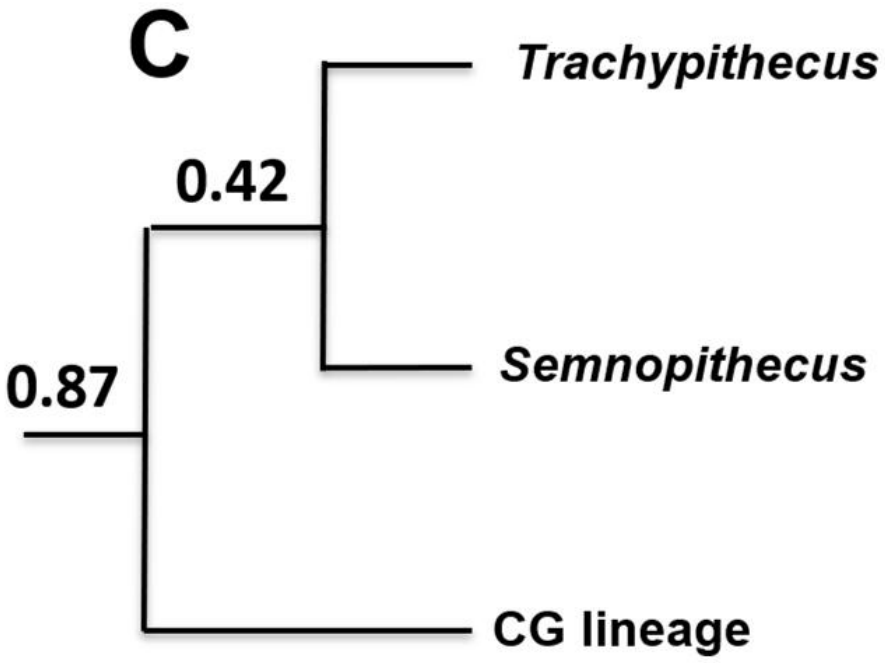
705

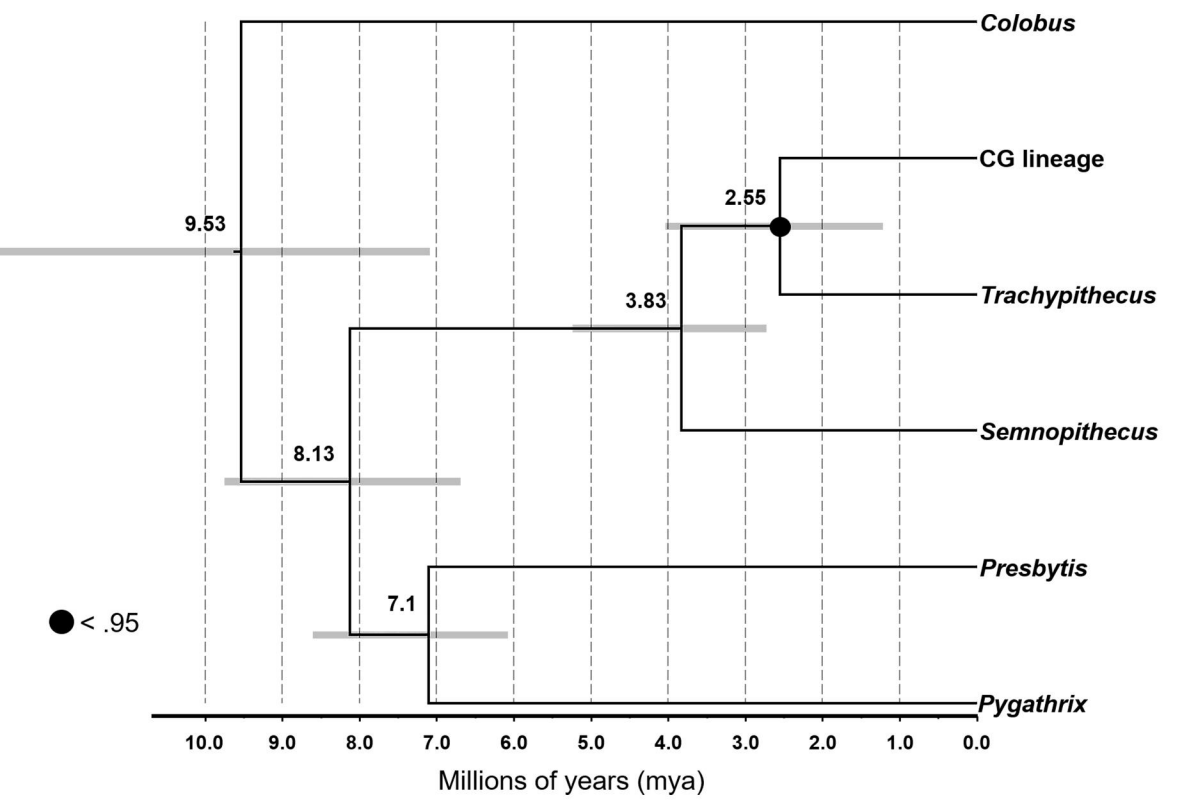

GL_CES12/303
CL_CES12/308
CL_CES12/309
GL_CES12/305
CL_CES12/300
CL_CES11/299
T_phayrei
T_obscurus
S_hypoleucos_CES11/409
S_hypoleucos_CES08/333
S_hypoleucos_CES11/409
S_vetulus
S_e_entellus
Py_nemaeus
Py_cinerea
Py_nigripes
N_larvatus
R_brelichi
P_melalophos
C_guereza
1/100
*/81
1/98
1/100
1/100
0.96/91
1/100
1/*
0.99/87
1/100
0.94/*
0.90/*
0.80/*
0.002


1/100
T_germaini
1/99
T_cristatus
1/96
T_obscurus
1/100
T_hatinhensis
1/100
T_francoisi
1/99
T_shortridgei
1/100
T_shortridgei_wang2015
Capped langur_CES11/299
1/100
T_pileatus
1/90
P_melalophos
0.91/*
1/100
N_larvatus
1/80
N_concolor
1/100
Py_nigripes
1/100
Py_nemeaus
1/*
1/100
R_roxellana 1
1/100
R_roxellana 2
R_avunculus
1/68
S_entellus
1/100
S_vetulus
S_Johnii
C_guereza
0.05


A
Trachypithecus
36
81
CG lineage
Semnopithecus


B
CG lineage
81
Trachypithecus
Semnopithecus


C
Trachypithecus
0.42
Semnopithecus
0.87
CG lineage


Colobus
CG lineage
Trachypithecus
Semnopithecus
Presbytis
Pygathrix
9.53
2.55
3.83
8.13
7.1
< .95
10.0
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
Millions of years (mya)