METHODS

Sequencing and assembly of the sooty mangabey genome. DNA from a female sooty mangabey (*C. atys*) born and maintained at the Yerkes National Primate Research Center was extracted from whole blood. The animal selected for sequencing was one of the original dams of a large matrilineal line of the colony. In addition, she possessed the most common MHC haplotype observed within the group. As such, her genetic constitution within the closed population was thought to be the most representative of any single animal. All animals were housed at the Yerkes National Primate Research Center of Emory University and maintained in accordance with US NIH guidelines. All studies were approved by the Emory University Institutional Animal Care and Usage Committee. Following quality control to ensure purity and molecular weight, a series of Illumina sequencing libraries were prepared using standard procedures. Paired-end libraries with nominal insert sizes 180 bp and 500 bp were produced. In brief, 1 μg of DNA was sheared to the desired size using a Covaris S-2 system. Sheared fragments were purified with Agencourt AMPure XP beads, end-repaired, dA-tailed and ligated to Illumina universal adaptors. After adaptor ligation, DNA fragments were further size selected by agarose gel and PCR amplified for six to eight cycles using Illumina P1 and Index primer pair and Phusion High-Fidelity PCR Master Mix (New England Biolabs). The final library was purified using Agencourt AMPure XP beads and quality assessed by Agilent Bioanalyzer 2100 (DNA 7500 kit) to determine library quantity and fragment size distribution before sequencing.

Long mate-pair libraries with 2-kb, 3-kb, 5-kb and 8-kb insert sizes were constructed according to the manufacturer's protocol (Mate Pair Library v.2 Sample Preparation Guide 15001464 Rev. A Pilot Release). In brief, 5 μg (for 2- and 3-kb size libraries) or 10 μg (5- and 8-kb libraries) of genomic DNA was sheared to the desired size by Hydroshear (Digilab), then end-repaired and biotinylated. Fragment sizes between 1.8–2.5 kb (2 kb), 3.0–3.7 kb (3 kb), 4.5–6.0 kb (5 kb) or 8–10 kb (8 kb) were purified from a 1% low-melting agarose gel and circularized by blunt-end ligation. These size-selected circular DNA fragments were then sheared to 400 bp (Covaris S-2), purified using Dynabeads M-280 Streptavidin Magnetic Beads, end-repaired, dA-tailed and ligated to Illumina PE sequencing adapters. DNA fragments with adaptor molecules on both ends were amplified for 12 to 15 cycles with Illumina P1 and Index primers. Amplified DNA fragments were purified with Agencourt AMPure XP beads. Quantification and size distribution of the final library was determined as described above before sequencing.

Sequencing was performed on Illumina HiSeq 2000 instruments, generating 100-bp paired-end reads. Raw sequences have been deposited in NCBI under Bioproject PRJNA157077. Reads were assembled using ALLPATHS-LG and further scaffolded and gap-filled using in-house tools Atlas-Link (v.1.0) and Atlas GapFill (v.2.2) (https://www.hgsc.bcm.edu/software/)[23]. Atlas-link is a scaffolding or super-scaffolding method that uses all unused mate pairs to increase scaffold sizes and create new scaffolds in draft-quality assemblies. Those modified scaffolds are then ordered and oriented. Atlas GapFill is run on a super-scaffolded assembly. Regions with gaps are identified and reads mapping within or across those gaps are locally assembled using different assemblers (Phrap, Newbler and Velvet) in order to bridge the gaps with the most conservative assembly of previously unincorporated reads.

PBJelly (v.14.9.9) is a pipeline that improves the contiguity of draft assemblies by filling gaps, increasing contig sizes and super scaffolding by making use of long reads[24]. We used 12.3× coverage of long Pacific Biosciences RSI and RS II sequences, along with the gap-filled Illumina read assembly, as input into PBJelly to produce the final *C. atys* hybrid Illumina–PacBio assembly. This assembly is available at NCBI as Caty1.0 (RefSeq accession GCF_000955945.1).

The total size of the assembled *C. atys* genome is around 2.85 Gb, with a contig N50 size of 112.9 kb and scaffold N50 size of 12.85 Mb (Table 1). By comparison, this contig N50 size is greater than equivalent values for 22 of the 26 other nonhuman primate genome assemblies currently available. To assess completeness, we mapped 21,772 human protein-coding canonical transcripts to Caty_1.0 and found that 94.9% map to this *C. atys* genome with lengths of 95–100% (97.3% of transcripts map at length 70% or greater). As a more stringent test, we mapped 3023 Benchmarking Universal Single-Copy Orthologues (BUSCO) genes and found that over 95% are present in Caty_1.0 (88.8% complete single copy and the others present but duplicated or fragmented)[25].

Genome annotation was performed through the NCBI Genome Annotation Pipeline, which generated models for genes, transcripts and proteins[26]. To aid accurate transcript annotation, the NCBI pipeline incorporated RNA-seq data from a sooty mangabey pooled tissue reference sample, and data from 14 separate tissues produced through a joint effort by the Nonhuman Primate Reference Transcriptome Resource (NHPRTR; http://www.nhprtr.org/)[27] and the Human Genome Sequencing Center (HGSC) of Baylor College of Medicine. The NCBI process also used human RefSeq and GenBank transcripts along with other primate protein data.

Sequencing and polymorphism screen of 10 sooty mangabeys. DNA was prepared from blood or liver samples from 10 sooty mangabeys from the YNPRC colony. Ten sooty mangabey breeder animals were selected in consultation with the YNPRC Breeding Manager representing at least 90% of colony diversity based on the pedigree of the colony. Illumina paired-end libraries (300-bp insert size) were prepared as described above for 500-bp paired-end libraries. These libraries were sequenced (100 bp reads) on a HiSeq2000 instrument, producing an average of 30× whole-genome coverage across individuals. These reads were mapped to the *C. atys* assembly using BWA-mem and single-nucleotide variants were called using GATK (https://software.broadinstitute.org/gatk/). A gVCF file was created for each animal, and variation in the regions of interest for *TLR4* and *ICAM2* were identified in those files.

Polymorphism screen among rhesus macaques. To assess variation in *TRL4* and *ICAM2* among rhesus macaques, we used our database of whole-genome sequence data from 133 individuals of this species. The details of sequencing and single-nucleotide variants discovery for this population have previously been described[8]. The population-level VCF file for this study was examined for relevant variation in these two genes.

Targeted re-sequencing of *ICAM2* and *TLR4* in rhesus macaques and sooty mangabeys. To test the validity of the apparent species differences in *ICAM2* and *TLR4* between rhesus macaques and sooty mangabeys, primers were designed to flank three areas of interest (see Extended Data Figs 3a, 5b), PCR was performed using genomic DNA from two rhesus macaques and two sooty mangabeys (including FAK, the animal used for the Caty_1.0 reference genome) and the PCR product was subjected to Sanger sequencing. PCR primers were designed using Primer3 with default settings with the exception that the human mis-priming library was selected (http://bioinfo.ut.ee/primer3/)[28,29]. Primers were tailed with M13 sequences to facilitate Sanger sequencing.

PCR primer pairs (gene specific sequences are underlined): *ICAM2*_Ex2_F GTAAAACGACGGCCAGTATGTGCAGGTGGAGTGTGAT; *ICAM2*_Ex2_R GGAAACAGCTATGACCATGGCTCGAACAGACTCAGTGGA; *ICAM2*_Ex3_F GTAAAACGACGGCCAGTAAGCAGAGCAGGACAGATGT; *ICAM2*_Ex3_R GGAAACAGCTATGACCATGACTCTGCACAGTCAGACCTT; *TLR4*_SL_F GTAAAACGACGGCCAGTACCATGGAATGACTTGCCCT; *TLR4*_SL_R GGAAACAGCTATGACCATGCCTTTCAGCTCTGCCTTCAC.

AmpliTaq Gold 360 DNA Polymerase (Applied Biosystems) was used to amplify PCR products using the following protocol: 95 °C for 10 min; 95 °C for 30 s, 65 °C for 30 s, 72 °C for 30 s, 10 cycles (annealing temperature is decreased by 1 °C per cycle); 94 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s, 30 cycles; 72 °C for 10 min. PCR products were subjected to Sanger sequencing (in both directions) using M13 primers. PCR and Sanger sequencing was performed at ACGT. Traces (see Fig. 2a for examples) were inspected and consensus sequences obtained for each PCR product. Primer sequences were trimmed and consensus sequences were deposited in GenBank (accession numbers: MF468275–MF468286).

Sequencing and *de novo* assembly of RNA-seq transcripts. Transcripts for sooty mangabey were assembled *de novo* from RNA-seq reads using Trinity on XSEDE's Blacklight supercomputer[30]. The RNA-seq reads were pooled from 12 different tissues and were prepared by the standard mRNA-seq with the uracil DNA glycosylase protocol (Illumina kit Part RS-122-2303) and are publicly available from the Nonhuman Primate Reference Transcriptome Resource (NCBI SRA accession numbers SRX270666 and SRX270667)[27]. We performed a number of filtering steps to prepare threads for *de novo* assembly, which included removing adapters, filtering for quality, removing poly A/T tails and removing mtDNA and common mammalian rRNA[27,31]. After filtering, we used an input of 1,635,074,685 RNA-seq reads as the basis for the transcriptome assembly. Using around 550 mostly continuous compute hours on Blacklight, we partitioned the computational job into three phases described by the Trinity algorithm: Inchworm (around 100 h × 64 cores), Chrysalis (around 400 h × 128 cores), and Quantify Graph and Butterfly (around 50 h × 64 cores). To circumvent the large amount of I/O generated in the Quantify Graph phase, we ran Trinity directly from the RAM disk for this phase. Using Trinity (version r2012-10-05), the following options were selected:
Trinity.pl–JM 512G–no_run_chrysalis–seqType fa–single, reads.fasta–run_as_paired–CPU 16, Trinity.pl–JM 512G–no_run_quantifygraph–seqType fa–single, reads.fasta–run_as_paired–CPU 16–bflyGCThreads 4, Trinity.pl–JM 512G–no_run_butterfly–seqType fa–single
reads.fasta–run_as_paired–CPU 16., Trinity.pl–JM 512G–bflyGCThreads 16–bflyCPU 32–seqType fa, –single reads.fasta–run_as_paired–CPU 16.
The large N25 (6,431 bp), N50 (3,483 bp) and N75 (1,116bp) values of the resulting assembly were indicative of its success.

Pipeline for finding divergent sooty mangabey proteins. *C. atys* assembly Caty_1.0 protein model predictions were screened against the curated *M. mulatta* MacaM protein models by alignment with BLASTp (v.2.2.28+)[22]. The *C. atys* protein model alignment with the lowest *e* value or highest bitscore (for equal *e* values)