**LETTER**

Sooty mangabey genome sequence provides insight into AIDS resistance in a natural SIV host

David Palesch1,8, Steven E. Bosinger1,2,8, Gregory K. Tharp1, Thomas H. Vanderford1, Mirko Pajardini1,2, Ann Chahroud1,3, Zachary D. Johnson1, Frank Kirchhoff4, Beatrice H. Hahn5, Robert B. Norgren Jr6, Nirav B. Patel1, Donald L. Sodora7, Reem A. Dawoud1, Caro–Beth Stewart8, Sara M. Seepo8, R. Alan Harris9,10, Yue Liu9, Mutthuswamy Raveendran9,10, Yi Han8, Adam English9, Gregg W. C. Thomas11, Matthew W. Hahn11, Lenore Pipes12, Christopher E. Mason12, Donna M. Muzny9,10, Richard A. Gibbs9,10, Daniel Sauter4, Kim Worley9,10, Jeffrey Rogers9,10 & Guido Silvestri1,2

In contrast to infections with human immunodeficiency virus (HIV) in humans and simian immunodeficiency virus (SIV) in macaques, SIV infection of a natural host, sooty mangabeys (*Cercopithecus atys*), is non-pathogenic despite high viraemia1. Here we sequenced and assembled the genome of a captive sooty mangabey. We conducted genome-wide comparative analyses of transcript assemblies from *C. atys* and AIDS-susceptible species, such as humans and macaques, to identify candidates for host genetic factors that influence susceptibility. We identified several immune-related genes in the genome of *C. atys* that show substantial sequence divergence from macaques or humans. One of these sequence divergences, a C-terminal frameshift in the toll-like receptor-4 (TLR4) gene of *C. atys*, is associated with a blunted *in vitro* response to TLR-4 ligands. In addition, we found a major structural change in exons 3–4 of the immune-regulatory protein intercellular adhesion molecule 2 (ICAM-2); expression of this variant leads to reduced cell surface expression of ICAM-2. These data provide a resource for comparative genomic studies of HIV and/or SIV pathogenesis and may help to elucidate the mechanisms by which SIV-infected sooty mangabeys avoid AIDS.

SIV infection of natural hosts, such as sooty mangabeys, is typically non-pathogenic despite high viraemia. This is in stark contrast to HIV infection in humans and experimental SIV infection in rhesus macaques (*Macaca mulatta*) that progress to AIDS unless treated with antiretroviral therapy. The main virological and immunological features of natural SIV infection in sooty mangabeys have been described over the past 15 years in studies that compared and contrasted this infection with the pathogenic infections of HIV and SIV in humans and rhesus macaques1. SIV-infected sooty mangabeys show several features that have been observed in pathogenic infections, including high viraemia, short *in vivo* lifespan of productively infected cells, depletion of mucosal CD4+ T cells, strong type-I interferon response in the acute infection, and cellular immune responses that fail to control virus replication. However, in contrast to pathogenic infections, SIV-infected sooty mangabeys (i) have healthy CD4+ T cell levels; (ii) do not experience mucosal immune dysfunction, avoiding depletion of T helper 17 (T\(_{\text{H}17}\)) cells, intestinal epithelial damage and microbial translocation; (iii) maintain low levels of immune activation during the chronic infection; and (iv) achieve compartmentalization of virus replication that preserves central-memory and stem-cell memory CD4+ T cells as well as follicular T\(_{\text{H}17}\) cells1,2. An additional notable feature of SIV infection in natural hosts is the low rate of mother-to-infant transmission that is related to low expression of CCR5 on circulating and mucosal CD4+ T cells3. Although many aspects of the natural course of SIV infection in sooty mangabeys have now been described, the key molecular mechanisms by which these animals avoid AIDS remain poorly understood.

In this study, we sequenced the genome of a captive sooty mangabey and compared this genome to the genomes of AIDS-susceptible primates to look for candidate genes that may influence susceptibility to AIDS in SIV-infected hosts. We sequenced genomic DNA to a whole-genome coverage of about 180× using the Illumina HiSeq 2000 platform, and produced an initial assembly using ALLPATHS-LG, Atlas-Link and Atlas-GapFill (see Methods for details). The total size of the assembled *C. atys* genome (Caty_1.0; NCBI accession number GCA_000955945.1) is around 2.85 Gb, with a contig N50 size of 112.9 kb and scaffold N50 size of 12.85 Mb (Table 1). Genome annotation identified 20,829 protein-coding genes and 4,464 non-coding genes in the *C. atys* assembly, which is comparable to other available draft quality genomes of nonhuman primates (Table 1). These analyses demonstrate that the Caty_1.0 reference genome is of sufficient quality to facilitate population-scale whole-genome and transcriptome sequencing studies.

To identify novel immunogenetic factors specific to *C. atys* that may be involved in the ability of this species to avoid progression to AIDS, we established a bioinformatic pipeline for a comparative protein analysis (Fig. 1 and Extended Data Fig. 1, see Methods for details). Using this approach, we found 34 candidate immune-related genes with sequences that diverged between *C. atys* and *M. mulatta* (Table 1 and Extended Data Table 1). Although we cannot exclude a role of immune genes with minor differences in *C. atys* and *M. mulatta*, the highly divergent genes listed in Table 1 and Extended Data Table 1 constitute candidate genes involved in the outcomes of SIV infection in these two species.

Our screen identified sequence divergence in a number of proteins that are important during HIV infection, such as APOBEC3C (91.6%) and BST2 (also known as tetherin, 95.1%), as well as pattern-recognition receptors (MBL2, CLEC4A, CLEC4D and CLEC6A), the antiviral sensor cyclic GMP–AMP synthase (cGAS (also known as MB21D1)) and other immune mediators (Extended Data Table 1). Because CD4 and CCR5 are important for AIDS pathogenesis, we aligned the sequences of *CaCD4* and *CaCCR5* to *MmCD4* and *MmCCR5*, respectively4,5. Neither gene showed any major structural changes in the wild-type variants, although CD4 was slightly below the 97%
threshold of identity (Extended Data Fig. 1b, c). In addition, we found specific gene families in *C. atys* that are expanded relative to *M. mulatta*, humans and other primates (Extended Data Table 2a). Notably, we detected localized regions of increased substitution, defined by a clustered difference of three or more amino acids, in 10 genes. The most marked variations in the amino acid sequence of *C. atys* compared to *M. mulatta* were observed in ICAM-2 and TLR-4 (Table 1).

ICAM-2 is an approximately 60-kDa transmembrane glycoprotein of the immunoglobulin superfamily, which is expressed on various immune cells and implicated in lymphocyte homing and recirculation. ICAM-2 ligands are lymphocyte function-associated antigen-1, which is expressed on various immune cells and thus codes for a markedly different final gene product (Extended Data Figs 2, 3). Splice-junction sequence analysis showed intact splicing for all four exons in *M. mulatta*, but no splice junctions were found between exons 3 and 4 in *C. atys*, indicating severe splicing defects due to the deletion (Extended Data Fig. 4).

To test whether the observed genetic difference in ICAM2 has functional consequences, we measured ICAM-2 surface expression on immune cells from humans, *M. mulatta* and *C. atys* with an antibody that recognizes a conserved epitope between these species. ICAM-2 was readily detected on T cells and B cells from humans and *M. mulatta*, but not from *C. atys* (Fig. 2b, c), suggesting that ICAM-2 is not functional in lymphocytes of *C. atys*. However, a truncated, lower expressed ICAM-2 protein was detected in all individuals (data not shown). The ICAM-2 deletion may be specific to *C. atys*, as it is not present in any other known primate sequences, including other natural SIV hosts, such as the African green monkey, drill and colobus monkey. Transcript models generated from de novo assembled *C. atys* RNA-sequencing (RNA-seq) data from 14 different tissues showed that the mature mRNA sequence of CaICAM2 retains substantial portions of what is part of the intronic sequence in other nonhuman primates, and thus codes for a markedly different final gene product (Extended Data Figs 2, 3). Splice-junction sequence analysis showed intact splicing for all four exons in *M. mulatta*, but no splice junctions were found between exons 3 and 4 in *C. atys*, indicating severe splicing defects due to the deletion (Extended Data Fig. 4).

Table 1 | *C. atys* assembly statistics and proteins with major structural variations in the *C. atys* genome

| Assembly | Annotation |
|----------|-----------|
| Average coverage per base | Protein-coding genes |
| Total sequence length | Non-coding genes |
| Total assembly gap length | Pseudogenes |
| Number of scaffolds | mRNA transcripts |
| Scaffold N50 | IncRNA transcripts |
| Scaffold L50 | Exons in coding transcripts |
| Number of contigs | Exons in non-coding transcripts |
| Contig N50 | 18,754 |
| Contig L50 | 65,920 |
| GC content | 20,829 |

| Gene | Function | Variation type | Length variation (amino acids) |
|------|----------|---------------|-------------------------------|
| ICAM2 | Lymphocyte extravasation and recirculation | indel, fs | 107 |
| TLR4 | LPS sensing | indel, fs | 17 |
| BPIFA1 | Antimicrobial function in airways | indel | 8 |
| INOS2 | Prolinflammatory messenger | pm, early stop | 8 |
| MBL2 | Pattern recognition receptor for microbial products | pm, early start | 7 |
| TREM2 | Chronic proinflammatory signaling in myeloid cells | indel, fs | 6 |
| PSCLR1 | Enhancement of the interferon response | indel | 5 |
| LST1 | Inhibition of lymphocyte proliferation | indel, fs, pm, indel | 5 |
| CRTAM | T and natural killer cell activation | pm, indel | 4 |

Structural variations were identified by the immunogenomic comparison pipeline. N50, 50% of the genome is in fragments of this length or longer; L50, smallest number of fragments needed to cover more than 50% of the genome; lncRNA, long non-coding RNA; indel, insertion/deletion; fs, frameshift; pm, point mutation.

Figure 1 | Bioinformatic pipeline for the identification of divergent *C. atys* proteins. (1) Sooty mangabey (SM) orthologues were selected by BLAST alignment of *C. atys* NCBI protein predictions (blue) to curated rhesus macaque (RM) protein models (green) and alignment scores were calculated. (2) NCBI transcript predictions with RNA-seq support were identified by BLAT alignment of *de novo* assembled *C. atys* RNA-seq transcripts (orange) to *C. atys* NCBI coding sequence (CDS) predictions (red). (3) Subsequently, corresponding RNA-seq-supported *C. atys* NCBI protein predictions were selected. (4) *C. atys* proteins with high similarity (>97% identity) to *M. mulatta* proteins were filtered out. (5) Immune genes according to Gene Ontology (GO) term classification (immune response) were chosen for further analysis and (6) confirmed by manual inspection.
molecular weight form of ICAM-2 could be detected intracellularly by western blot in C. atys cells (Fig. 2d), thus demonstrating the presence of the predicted truncated ICAM-2 protein. Overall, these data indicate that the presence of a species-specific gene sequence difference in CaICAM2 results in the abrogation of surface expression of this protein in C. atys. Further studies are needed to elucidate potential links between this truncated form of ICAM-2 and the remarkable immunological features of SIV infection in this species.

TLR-4 is a pattern recognition receptor that senses lipopolysaccharides (LPS) on gram-negative bacteria and initiates pro-inflammatory cytokine induction, maturation and activation in macrophages, dendritic cells and other immune cells. During pathogenic HIV or SIV infections, exacerbated TLR-4 stimulation and concomitant pro-inflammatory signalling elicited by microbial translocation is considered a primary mechanism that underlies HIV-induced chronic inflammatory signalling elicited by microbial translocation is considered a primary mechanism that underlies HIV-induced chronic illness in these species (Extended Data Fig. 6b). Although a naive analysis of this pattern would suggest two independent mutational changes in TLR4, the short internal branch of the species tree implies that incomplete lineage sorting of an ancestral polymorphism could also generate this pattern12 (Fig. 3b). To test this hypothesis, we examined the TLR4 gene tree among 17 primate species. While generally supporting the relationships among these species (Fig. 3b), the analysis also found a number of nucleotide positions—spaced throughout the gene—consistent with incomplete lineage sorting between C. atys, baboon and M. mulatta (Extended Data Fig. 7). The incomplete lineage sorting hypothesis is also more likely, given that balancing selection is often found to be acting on immune-related genes. Therefore, even though baboons are believed to be more closely related to sooty mangabeys and drills than to rhesus macaques, the phylogeny of Old World monkeys is compatible with the possibility of a single G-to-A mutation creating the truncated form of the protein in the common ancestor of baboons, rhesus macaques and sooty mangabeys12,16 (Fig. 3b).

We next investigated potential differences in TLR-4 function between M. mulatta and C. atys. Our previous work has shown that macropores from C. atys exhibit higher expression of tetherin, APOBEC and TRIM5α in response to LPS compared to M. mulatta15. This is consistent with the relative resistance of C. atys macropores to in vivo SIV infection after experimental CD4+ T cell depletion compared to SIV-infected M. mulatta macropores.16 Here we analysed cytokine gene expression and protein production after LPS stimulation, and found reduced mRNA expression and secretion of TNF (also known as TNF-α) and IL-6 in cells from C. atys compared to M. mulatta (Fig. 3c, d). Because some commercial LPS preparations contain lipoprotein contaminants that can induce TLR-2 signalling, we confirmed the TLR-4 specificity of the reduced LPS response using the selective TLR-4 agonist lipid-A (Extended Data Fig. 8a, b). Next, we found that the species-specific differences between C. atys and M. mulatta in LPS-induced TNF and IL-6 production were maintained in acute and chronic infection (Fig. 3e and Extended Data Fig. 8c). Additionally, we did not observe any difference in the mRNA levels

![Figure 2: Genomic deletion in CaICAM2 results in a truncated and dysfunctional protein.](image-url)
of TLR4 in cells from C. atys and M. mulatta, nor did the expression of any factors in the TLR-4–MyD88–TRIF signalling axis contribute to the observed difference.

Moreover, using gene set enrichment analysis (GSEA), we observed a significantly attenuated NF-κB signalling in C. atys (Fig. 3f, g and Extended Data Fig. 9). This indicates that induction of pro-inflammatory genes was broadly and significantly reduced in cells from C. atys (Fig. 3f, g and Extended Data Fig. 9). Overall, these results indicate that LPS stimulation of blood cells from C. atys results in a blunted production of pro-inflammatory cytokines. To establish a link between the C-terminal TLR4 sequence difference and the responsiveness to LPS, we analysed the TLR-4 orthologues of humans, C. atys and M. mulatta in an NF-κB reporter assay. We observed a significantly attenuated NF-κB response to LPS of C. atys TLR-4 (CaTLR-4) compared to M. mulatta TLR-4 (MmTLR-4).

Using chimaeric constructs encoding MmTLR4 with the C terminus of CaTLR4 or CaTLR4 with the C terminus of MmTLR4, we confirmed that the TLR4 C terminus is responsible for this phenotypic difference (Fig. 3h). This demonstrates a sequence–function relationship of the TLR4 C terminus and suggests a novel mechanism contributing to the lower immune activation of SIV-infected sooty mangabeys.

Over the past decade the genomes of more than 25 nonhuman primate species have been sequenced, assembled and annotated. This knowledge has improved our understanding of primate evolution, biology and general physiology, which has informed human biology and medicine. Here, we report a high-coverage, high-contiguity whole-genome sequence for CaTLR4, which has informed human biology and medicine.

Figure 3 | The TLR-4 C terminus is distinctive in natural SIV hosts.

a, Alignment of C-terminal TLR-4 protein sequences from different primate species (starting at human TLR-4 amino acid position 821). b, Primate phylogenetic tree with colour-coding according to the TLR-4 C terminus as indicated in a. Phylogeny appears as in ref. 14. c, Cytokine release from blood of rhesus macaques (n = 9 biologically independent samples) and sooty mangabeys (n = 8 biologically independent samples) after LPS stimulation as measured by cytometric bead array. d, mRNA expression in whole blood after LPS stimulation quantified by quantitative PCR (qPCR). n = 4 biologically independent samples for each species. e, TNF and IL-6 cytokine release from blood of rhesus macaques and sooty mangabeys over the course of SIV infection. n = 5 biologically independent samples for each species. Data are mean ± s.d. (c–e), unpaired two-sided Student’s t-test, P values are indicated (c, d, f).
characterize the host response to SIV infection of C. atys and African green monkeys. Here, we examined the mechanisms of AIDS resistance of a natural SIV host genome-wide using genome sequencing. We identified candidate genes that show sequence changes that are specific to C. atys and two gene products (ICAM-2 and TLR-4), which show structural differences between C. atys and M. mulatta that may influence cell-surface expression (ICAM-2) and downstream signalling (TLR-4) of these proteins. Our findings may also explain prior results showing that not all natural SIV hosts respond to infection in the same way, suggesting that in each primate species, multiple distinct mechanisms may contribute to the phenotype, rather than mutations in single genes, as has been purported, and eventually refuted, in other studies. Further comparative studies with additional natural SIV host species may identify additional similarities (or differences) in the genes involved in the evolutionary pathways that led to AIDS resistance in different species of African nonhuman primates.

In this study, we used whole-genome sequencing and comparative genomic analysis to identify candidate genes regulating host resistance to AIDS. Future studies in which these candidate genes are manipulated in vivo during SIV infection are needed to characterize to what extent these genes may influence the non-pathogenic nature of SIV infection in sooty mangabeys.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 January; accepted 16 November 2017.

1. Chahroudi, A., Bosinger, S. E., Vanderford, T. H., Paiardini, M. & Silvestri, G. Reduced Simian immunodeficiency virus replication in natural nonhuman primate hosts. Science 335, 1188–1193 (2012).

2. Cartwright, E. K. et al. Divergent CD4+ T memory stem cells in dynamic and pathogenic nonhuman simian immunodeficiency virus infections. J. Immunol. 192, 4666–4673 (2014).

3. Pandrea, I. et al. Paucity of CD4+CCR5+ T cells may prevent transmission of simian immunodeficiency virus in natural nonhuman primate hosts by breast-feeding. J. Virol. 82, 5501–5509 (2008).

4. Paiardini, M. et al. Low levels of SIV infection in sooty mangabey central memory CD4+ T cells are associated with limited CCR5 expression. Nat. Med. 17, 830–836 (2011).

5. Beaumier, C. M. et al. CD4 downregulation by memory CD4+ T cells in vivo renders African green monkeys resistant to progressive SIVagm infection. Nat. Med. 15, 879–885 (2009).

6. Halai, K., Whiteford, J., Ma, B., Nourshargh, S. & Woodfin, A. ICAM-2 facilitates luminal interactions between neutrophils and endothelial cells in vivo. J. Cell Sci. 127, 620–629 (2014).

7. Staunton, D. E., Dustin, M. L. & Springer, T. A. Functional channelling of ICAM-2, a cell adhesion ligand for LFA-1 homologous to ICAM-1. Nature 339, 61–64 (1989).

8. Xue, C. et al. The population genomics of rhesus macaques (Macaca mulatta) based on whole-genome sequences. Genome Res. 26, 1651–1662 (2016).

9. Casasnovas, J. M., Pironen, C. & Springer, T. A. Lymphocyte function-associated antigen-1 binding residues in intercellular adhesion molecule-2 (ICAM-2) and the integrin binding surface in the ICAM subfamily. Proc. Natl Acad. Sci. USA 96, 2017–2022 (1999).

10. Brenchley, J. M. et al. Microbial translocation is a cause of systemic immune activation in chronic HIV infection. Nat. Med. 12, 1365–1371 (2006).

11. Brenchley, J. M. & Douek, D. C. HIV infection and the gastrointestinal immune system. Mucosal Immunol. 1, 23–30 (2008).

12. Perreault, P. et al. A molecular phylogeny of living primates. PLoS Genet. 7, e1001342 (2011).

13. Mendes, F. K. & Hahn, M. W. Gene tree discordance causes apparent substitution rate variation. Syst. Biol. 65, 711–721 (2016).

14. Finstermeier, K. et al. A mitogenomic phylogeny of living primates. PLoS ONE 8, e69504 (2013).

15. Mir, K. D. et al. Reduced Simian immunodeficiency virus replication in macaques of sooty mangabeys is associated with increased expression of host restriction factors. J. Virol. 89, 10136–10144 (2015).

16. Klatt, N. R. et al. Availability of activated CD4+ T cells dictates the level of viremia in naturally SIV-infected sooty mangabeys. J. Clin. Invest. 118, 2039–2049 (2008).

17. Raetz, C. R. & Whitfield, C. Lipopolysaccharide endotoxins. Annu. Rev. Biochem. 71, 635–700 (2002).

18. Rogers, J. & Gibbs, R. A. Comparative primate genomics: emerging patterns of genome content and dynamics. Nat. Rev. Genet. 15, 347–359 (2014).

19. Bosinger, S. E. et al. Global analysis reveals a dynamic and robust innate response in SIV-infected sooty mangabeys. J. Clin. Invest. 119, 3536–3572 (2009).

20. Jacqueline, B. et al. Nonpathogenic SIV infection of African green monkeys induces a strong but rapidly controlled type I IFN response. J. Clin. Invest. 119, 3544–3555 (2009).

21. Bosinger, S. E. et al. Intact type I interferon production and IRF7 function in sooty mangabeys. PLoS Pathog. 9, e1003597 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Ehrt and all animal care and veterinary staff at the Yerkes National Primate Research Center (YNPRC). B. Cervasia and K. Gill at the Emory University Flow Cytometry Core; M. T. Nega at the Emory CFAR Virology & Drug Discovery Core; R. Linsenmeyer for excellent technical assistance; and O. Laur and the Emory Custom Cloning Core Division. We are grateful for the sequence production and related activities at the Human Genome Sequencing Center at Baylor College of Medicine. The production teams involved were: sample tracking (S. Jhangiani, C. Kovar), laboratory production (H. Doddapaneni, H. Chao, S. L. Lee, G. Weissenberger and M. Wang), illumina sequencing (H. Dinh, G. Okwuonu and J. Santibanez), PacBio sequencing (K. Perelman, P. Zimin, A. V. Rutovitz, J. Santibanez, P. Dongen, D. Delong, D. Warner, Y. Li, and S. N. Markulis) and production informatics (M. Dahdouli, Z. Khan, J. G. Reid and D. Sexton). D. Rio Deros and S. C. Murali also contributed to the assembly of the genome sequence. This work was funded by HHS/National Institutes of Health (NIH) R37 AI66998. Research reported in this publication was also supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number P51 OD011132 (to the YNPRC). D. M. M., R. A. G., K. W. and J. R. were supported by NIH grant U54-HG006484-01. D. S. was supported by the Prolynity Program ‘Innate Sensing and Regulation of Retroviruses’ (SPP 1923) of the German Research Foundation (DFG). F. K. was funded by the Advanced ERC grant ‘Anti-Virome’ and the DFG-funded SFB 1279, G.W.C.T. and M.W.H. were supported by the Precision Health Initiative of Indiana University. B.H.H. was funded by NIH grant R37 AI050529. Comparative transcriptomics research was funded by NIH grant R24 OD010445.

Author Contributions D.P. and S.E.B. designed and performed experiments and analysed data. S.E.B. and G.K.T. designed and performed bioinformatics analyses. F.K. and B.H.H. designed experiments. T.H.V., M.P. and A.C. contributed to the study design and data interpretation. R.B.N. performed custom annotation of macaque and mangabey genomes and Sanger sequencing. Z.P.I. collected samples and analysed data. Y.H. contributed to sequencing. T.A. contributed to genome assembly. M.R., D.M.M., and R.A.G. supervised and/or managed the sequencing of the C. atys genome. R.A.H. and Y.L. performed genome assembly tasks. R.A.D. performed RNA-seq sample processing and analysis. D.L.S. analysed TLR-4 functional data. K.W. and J.R. supervised the assembly and analysis of genome. C.B.S. and S.M.S. sequenced and interpreted genetic data of ICAM-2. G.W.C.T. and M.W.H. analysed gene family evolution. N.B.P. collected samples and conducted RNA-seq experiments. L.P. and C.E.M. sequenced and assembled RNA-seq transcripts. D.S. designed and analysed TLR-4 experiments. J.R. conceived the study, designed experiments and analysed genomic data. G.S. conceived, designed and led the study. D.P., S.E.B., J.R. and G.S. wrote the manuscript with input from all authors.

Author Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to G.S. (gsilves@emory.edu).

Reviewer Information Nature thanks M. Martin and the other anonymous reviewer(s) for their contribution to the peer review of this work.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.
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 by using the guidelines from the NIH guidelines. All studies were approved by the Emory University Institutional Animal Care and Use 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 300 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 assayed 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 1500146 Rev. A Pilot Release). In brief, 5 µg (for 2- and 3-kb size libraries) or 10 µg (for 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 (2kb), 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/). 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 (v14.9.9) 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. We used 12.3× coverage of long Pacific Biosciences RS and RS II sequences, along with the gap-filled Illumina sequencing read, 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 kbp 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 non-human primate species (around 40 kbp for chimpanzee and macaque, 80 kbp for gorilla, and 110 kbp for rhesus macaque). To assess genome completeness, we mapped 21,772 human protein-coding canonical transcripts to Caty1.0 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 2032 Benchmarking Universal Single-Copy Orthologues (BUSCO) genes and found that over 95% are present in Caty1.0 (88.8% complete single copy and the others present but duplicated or fragmented).

Genome annotation was performed through the NCBI Genome Annotation Pipeline, which generated models for genes, transcripts and proteins. 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 (NHPTRR, http://www.nhprtr.org/) 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 breeders 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 Human and Macaque MacaM protein models by alignment with BLASTp (v2.2.28). The C. atys protein model alignment with the lowest e-value or highest bitscore (for equal e-values)
was selected for each MacaM protein model, yielding the set of orthologous C. atys protein predictions most similar to the M. mulatta protein models. The spliced CDS sequence for each Caty_1.0 transcript prediction was extracted with gffread (utility from cufflinks v2.1.1). Caty_1.0 transcript prediction CDS sequences were screened against the de novo RNA-seq assembly transcript models by alignment with BLAT (v.34) and an alignment score was calculated as the number of matching bases minus the number of CDS sequence bases missing in alignment gaps normalized by the CDS sequence length. This score penalizes bases missing from the CDS sequence without penalizing extra sequence that may have been added to the RNA-seq transcript model during the assembly process. Only predicted CDS sequences that had a score >0.99 were retained as supported by RNA-seq data. The MacaM best match selected Caty1.0 protein models were then cross-referenced with the RNA-seq supported Caty_1.0 transcript models to eliminate protein models without RNA-seq evidence. The protein alignments to MacaM for these models were then re-examined to find genes for which the alignment identity was less than 97%, where there were gaps in the alignments or the alignment was not the full length of the protein model. These two species share a common ancestor about 10–11 million years ago, and therefore the expectation is that most proteins will be ~97% identical. This was confirmed by using a maximum likelihood amino acid model (WAG amino acid matrix) to estimate sequence distances between the C. atys and M. mulatta orthologues (Extended Data Fig. 1). Proteins of interest for differential response to lentivirus infection may be more divergent than expected on average. These represent potentially divergent genes and were further screened against the Gene Ontology (GO) term ‘immune response’. This list of divergent immune genes was then further curated by manual inspection of multiple alignments of CDS transcript and genomic sequences of C. atys (Caty_1.0). M. mulatta (MacM) and human (GRCh38.p7). Multiple alignment analysis was performed using Multalin (http://multalin.toulouse.inra.fr/). TLRA4 and ICAM2 sequence alignments were generated using Jalview.

Gene family evolution methods. In order to identify rapidly evolving gene families along the C. atys lineage, we obtained peptides from human, chimpanzee, orangutan, gibbon, macaque, baboon, vervet, marmoset and mouse from ENSEMBL 8332. The C. atys peptides were obtained from NCBI1. To ensure that each gene was counted only once, we used only the longest isoform of each protein in each species. We then performed an all-versus-all BLAST search on these filtered sequences14. The resulting e values from the search were used as the main clustering criterion for the MCL program to group peptides into gene families13. This resulted in 14,889 clusters. We then removed all clusters only present in a single species, resulting in 10,967 gene families. We also obtained an ultrametric tree from a previous study and added sooty mangabey based on its divergence time from baboon (TimeTree)16.37.

With the gene family data and ultrametric phylogeny as input, we estimated gene gain and loss rates (λ) with CAFE v.3.08. This version of CAFE is able to estimate the amount of assembly and annotation error (ε) present in the input data using a distribution across the observed gene family counts and a pseudo-likelihood search. CAFE is then able to correct for this error and obtain a more accurate estimate of λ. We find an ε of about 0.04, which implies that 4% of gene families have observed counts that are not equal to their true counts. After correcting for this error rate, we find λ = 0.0020. These values for ε and λ are on par with those previously reported for mammalian datasets13,39 (Extended Data Table 3b). Using the estimated λ value, CAFE infers ancestral gene counts and calculates P values across the tree for each family and lineage to assess the significance of any gene family changes along a given branch. CAFE uses Monte Carlo re-sampling to assess if a given family is rapidly evolving. For those families found to be rapidly evolving (P < 0.01), it then calculates P values for each lineage within the family using the Viterbi method. Those lineages with low P values (P < 0.01) are said to be rapidly evolving.

We observed 1,561 rapidly evolving families across the 10 species of mammals sampled here. Extended Data Table 3c summarizes the gene family changes for all 10 species. Humans have the highest average expansion rate across all families at 0.20 whereas gibbons have the lowest at ~0.09, meaning that they have the most gene family contractions. C. atys has undergone 535 gene family expansions of which 96 are rapid expansions and 340 gene family contractions of which 48 are rapid contractions.

Genetic distance between C. atys and M. mulatta orthologues. The amino acid sequences of 9,257 C. atys proteins with RNA-seq support (Fig. 1) were aligned to M. mulatta orthologues as described above. We used the same RNA-seq alignment package from PAML (v.6.9a) on each of these alignments with the WAG amino acid rate matrix to calculate maximum likelihood genetic distances between the two sequences. A histogram was generated from these distances with R (Extended Data Fig. 1a).

TLRA4 gene tree. TLRA4 nucleotide sequences for 17 primate species were obtained from the NCBI GenBank resource (human: NM_138554.4; rhesus macaque: XM_015116960.1; sooty mangabey: manually curated XM_012091593.1; bonobo: NM_001279223.1; Nancy Ma’s night monkey: XM_012477256.2; drill: XM_011973281.1; colobus monkey: XM_011950600.1; crab-eating macaque: NM_001319615.1; squirrel monkey: XM_003925187.2; baboon: XM_003911309.4; pig-tailed macaque: NM_013058891.1; gorilla: XM_00408514.2; chimpanzee: NM_001144663.1; orangutan: AB445642.1). These sequences were aligned with PASTA2 and we then constructed a maximum likelihood tree with RAXML3, performing 100 bootstrap replicates41,42 (Extended Data Fig. 7). Finding low bootstrap support amongst nodes ancestral to sooty mangabey, drill and baboon, we counted the number of sites that were discordant with respect to the gene tree topology. That is, the number of sites in which baboon and C. atys share the same state and C. atys and drill share a different state with an outgroup species (one of the two Old World monkeys).

Sample collection and processing. Peripheral blood samples from SIV-negative rhesus macaques and SIV-negative sooty mangabeys were collected by venipuncture according to standard procedures at the Yerkes National Primate Research Center of Emory University and in accordance with US National Institutes of Health guidelines. Human blood samples were obtained from healthy donors at the Yerkes National Primate Research Center in accordance with Institutional Review Board protocol IRB0004582 and all relevant ethical regulations. Informed consent was obtained from all blood donors. Peripheral blood mononuclear cells (PBMCs) were isolated from whole blood using Ficoll density-gradient centrifugation.

In vitro TLRA4 ligand stimulation assay. The assay used in this study is a modified version of the procedure previously described13. Ultrapure LPS (Escherichia coli 0111:B4) and monophosphoryl lipid-A (Salmonella minnesota) were purchased from Invivogen. Whole blood collected in EDTA vacutainers was diluted 1:4 with RPMI 1640 medium and 195 μl aliquots were transferred to 96-well, round-bottom micro-titre plates. Agonists were diluted in RPMI 1640 and 5 μl were applied to the wells at the following final concentrations: LPS, 1,000–10 ng ml−1; lipid-A, 10–1 μg ml−1. Suspensions were then mixed by pipet and incubated at 37 °C, 5% CO2 for 4 h. After incubation, plates were centrifuged at 700 r.p.m. for 10 min, and 120 μl of cell-free supernatant was removed and stored at −80 °C until the assay was carried out. Each TLRA4 ligand at a given concentration was performed in triplicate for each animal.

Cytokine bead array (CBA). Samples were obtained from sooty mangabeys and rhesus macaques housed at the YNPRC. Sooty mangabeys were naturally infected at the YNPRC and rhesus macaques had been infected previously with SIVsmm as previously described39. Supernatant levels of TNF and IL-6 were measured using the human immunflotation CBA kit (BD Biosciences Immunocytometry Systems) according to the manufacturer’s instructions, with the modification that the sample volumes for supernatant, antibody-coupled bead mix and PE-conjugated detection antibody solution were all reduced to 25 μl instead of 50 μl44. After incubation, samples were washed with 2% paraformaldehyde in PBS, resuspended in 150 μl PBS, and analysed using a FACSCalibur flow cytometer (BD Biosciences Immunocytometry Systems). The average of triplicate cytokine measurements was used as the representative value for individual animals, and variations in cytokine levels between species groups were tested for statistical significance using unpaired t-tests in Prism 6.0. To quantify the level of TLRA4 mRNA, and to perform linear regression of TLRA-signalling molecules with TNF and IL6 cytokine levels, in the LPS-stimulated blood samples in the longitudinal SIVsmm-infected samples, we used microarray expression data from matched whole-blood samples; these data are available from the NCBI Geo database (accession GSE16147).

Plasma viral load measurement. Quantification of SIVsmm plasma viral RNA levels were quantified using qPCR as described previously45. RNA-seq analysis of LPS-stimulated orthologues. RNA-seq analysis was conducted at the Yerkes Nonhuman Primate Genomics Core Laboratory (http://www.yerkes.emory.edu/np_genomics_core/). CD14 monoocytes were isolated from Ficoll-purified PBMCs using CD14 MicroBeads according to the manufacturer’s instructions (Miltenyi Biotec). Subsequently, 0.4 × 10⁶ cells were stimulated for 6 h with 10 ng ml⁻¹ LPS and then immediately lysed in 350 μl RLT buffer (Qiagen). RNA was purified using Micro RNAEasy columns (Qiagen) and RNA quality was assessed using Agilent Bioanalyzer. Then, 10 ng of total RNA was used as input for mRNA amplification using 5’ template-switch PCR with the Clontech SMART-Seq v4 Ultra Low Input RNA kit, according to the manufacturer’s instructions. Amplified cDNA was concentrated and appended with dual indexed barcodes using Illumina NexteraXT DNA Library Prep kits. Libraries were validated by capillary electrophoresis on an Agilent 4200 TapeStation, pooled and sequenced on an Illumina HiSeq 3000 using (100 bp paired-end reads) at an average read depth of 18 million. RNA-seq data were analysed by alignment and annotation to either.
the MacaM v7.8.2 assembly of the Indian rhesus macaque genome (available at https://www.unmc.edu/rhesusgenechip/index.htm) or to the Caty_1.0 assembly. Alignment was performed using STAR v2.5.2b using the annotation as a splice junction and abundance estimation reference, and non-unique mappings were removed from downstream analysis. Transcripts were annotated using both the MacaM and Caty_1.0 assemblies and annotation as described in the text. Transcript abundance was estimated internally in STAR using the algorithm of HT-Seq and differential expression analyses were performed using the DESeq2 packages53. To quantitate expression to which LPS treatment induced changes, gene expression between species, we used GSEA50. GSEA was performed using Partek Genomics software, v.6.6.

ICAM2 exon splice junction analysis. RNA-seq alignments from all 24 LPS-stimulated sample monocytes, and alignments derived from deep RNA-seq (over 50 million reads) from two samples derived from flow-sorted, purified, blood stimulated monocyte samples, and alignments derived from deep RNA-seq was tested using the Wald test as part of the DESeq2 workflow. Bars represent data was tested using the Wald test as part of the DESeq2 workflow. Bars represent

Fold change was calculated by dividing the normalized post-treatment sample quantity with the normalized untreated control quantity from the same animal, and calculating the average of fold changes for each species. Flow cytometry of PBMCs. Multicolour flow cytometry staining was performed using the following antibodies and reagents: CD3–APC/Cy7 (SP34-2), CD14–PE/Cy7 (MS2) and CD20–PE/Cy5 (2H7) from BD; CD4–BV650 (OKT4), CD8–BV711 (RPA-T8), ICAM–2–FITC (CR23/2), Mouse IgG2a–(v)–FITC (MOPC-173) isotype control from Biologend; Live/Dead Fixable Aqua from Thermo Fisher Scientific. Cells were stained for flow cytometry and data were acquired on an LSR II cytometer (BD) and analysed by FlowJo 10 software (TreeStar). Further analyses were performed using PRISM (GraphPad) and Excel (Microsoft Office 2011) software.

ICAM-2 western blot. PBMCs were lysed in RIPA buffer and equal amounts of cell lysate were boiled after addition of sample buffer including β-mercaptoethanol, resolved with a 4–15% SDS–PAGE (Bio-Rad), and proteins were transferred to an Immobilon-P PVDF membrane (Millipore). Afterwards membranes were blocked for 1 h in blocking buffer (Bio-Rad) and incubated overnight with polyclonal rabbit ICAM-2–specific antibody (Bethyl). After washing (PBS with 0.05% Tween-20), anti-rabbit HRP-conjugated secondary antibody was incubated for an additional 1 h, washed, and HRP activity was determined using the Super Signal West Pico Kit (Bio-Rad and visualized using the ChemiDoc XRS+ (Bio-Rad). Then the membrane was stripped with buffer (2% SDS, 0.5 M Tris, pH 2.2), blocked again and β-actin was detected using a rabbit anti-β-actin antibody as primary antibody and anti-rabbit HRP antibody as secondary antibody.

Statistical analysis. Statistical significance was determined using an unpaired Student’s t-test with Welch’s correction. P < 0.05 was considered significant. *P < 0.05; **P < 0.01; NS, not significant. Data are mean ± s.d. or s.e.m. as indicated. Significance for comparisons of mRNA levels of individual genes in RNA-seq data was tested using the Wald test as part of the DESeq2 workflow. Bars represent group means, and dots represent read counts for individual samples normalized to library size. P values denoted are adjusted using Benjamini–Hochberg correction.

Code availability. We used a custom script to quantify ICAM-2 splice junctions. This script is available at Github: https://github.com/BosinglerLab/splicing-analysis. Data availability. Raw sequences of the C. aty reference genome have been deposited in NCBI under Bioproject accession number PRJNA157077. The genome assembly is available at NCBI as Caty1.0 (RefSeq accession GCF_000959945.1). The multi-tissue C. aty RNA-seq reads are available from the Nonhuman Primate Reference Transcriptome Resource (NCBI SRA accession numbers SRX270666 and SRX270667). Data from Sanger sequencing of TLR4 and ICAM2 are available at NCBI accession numbers MF648275–MF648286. Microarray data used for TLR-4 measurement and linear regression with TNF and IL-6 are available from the NCBI GEO database (accession GSE16147). The RNA-seq data for LPS-stimulated monocytes was submitted to the GEO database (accession numbers GSM2711028–GSM2711051 and GSE101617).

The non-human primate reference transcriptome resource (NHPTR) for comparative functional genomics. Nucleic Acids Res 41, D906–D914 (2015).

Pipes, L. et al. The non-human primate reference transcriptome resource (NHPTR) for comparative functional genomics. Nucleic Acids Res 41, D906–D914 (2015).

Amplification of CD19 and CD5. A custom script was used to quantify ICAM-2 splice junctions.

Iyengar, A. et al. The non-human primate reference transcriptome resource (NHPTR) for comparative functional genomics. Nucleic Acids Res 41, D906–D914 (2015).

Pipes, L. et al. The non-human primate reference transcriptome resource (NHPTR) for comparative functional genomics. Nucleic Acids Res 41, D906–D914 (2015).

Iyengar, A. et al. The non-human primate reference transcriptome resource (NHPTR) for comparative functional genomics. Nucleic Acids Res 41, D906–D914 (2015).

Pipes, L. et al. The non-human primate reference transcriptome resource (NHPTR) for comparative functional genomics. Nucleic Acids Res 41, D906–D914 (2015).

Iyengar, A. et al. The non-human primate reference transcriptome resource (NHPTR) for comparative functional genomics. Nucleic Acids Res 41, D906–D914 (2015).

Pipes, L. et al. The non-human primate reference transcriptome resource (NHPTR) for comparative functional genomics. Nucleic Acids Res 41, D906–D914 (2015).
38. Han, M. V., Thomas, G. W., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* **30**, 1987–1997 (2013).
39. Carbone, L. et al. Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513**, 195–201 (2014).
40. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
41. Mirarab, S. et al. PASTA: ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. *J. Comput. Biol.* **22**, 377–386 (2015).
42. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
43. Tapping, R. I., Akashi, S., Miyake, K., Godowski, P. J. & Tobias, P. S. Toll-like receptor 4, but not toll-like receptor 2, is a signaling receptor for *Escherichia* and *Salmonella* lipopolysaccharides. *J. Immunol.* **165**, 5780–5787 (2000).
44. Morgan, E. et al. Cytometric bead array: a multiplexed assay platform with applications in various areas of biology. *Clin. Immunol.* **110**, 252–266 (2004).
45. Gordon, S. N. et al. Severe depletion of mucosal CD4+ T cells in AIDS-free simian immunodeficiency virus-infected sooty mangabeys. *J. Immunol.* **179**, 3026–3034 (2007).
46. Sumpter, B. et al. Correlates of preserved CD4+ T cell homeostasis during natural, nonpathogenic simian immunodeficiency virus infection of sooty mangabeys: implications for AIDS pathogenesis. *J. Immunol.* **178**, 1680–1691 (2007).
47. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
49. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
50. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
51. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics* **23**, 3251–3253 (2007).
52. Medvedev, A. E. & Vogel, S. N. Overexpression of CD14, TLR4, and MD-2 in HEK 293T cells does not prevent induction of in vitro endotoxin tolerance. *J. Endotoxin Res.* **9**, 60–64 (2003).
53. Sauter, D. et al. Differential regulation of NF-κB-mediated proviral and antiviral host gene expression by primate lentiviral Nef and Vpu proteins. *Cell Rep.* **10**, 586–599 (2015).
54. Bosinger, S. E. et al. Gene expression profiling of host response in models of acute HIV infection. *J. Immunol.* **173**, 6858–6863 (2004).
Extended Data Figure 1 | Genetic distances of *C. atys* and *M. mulatta* orthologues and protein sequence alignments of CD4 and CCR5.

a, Genetic distances of *C. atys* and *M. mulatta* orthologues. The dotted blue line represents a mean distance of 0.00755 expected substitutions, and the solid red line represents the 97th percentile. This percentile indicates that 8,979 out of 9,257 genes have a distance less than 0.0294.

b, Pairwise alignment of CD4 and CCR5 protein sequences for *C. atys* and *M. mulatta*. Sequences were aligned using Jalview v.2.9.0.
Extended Data Figure 2 | Sequence alignment of ICAM-2 protein and exon sequence analysis of ICAM2. a, Pairwise alignment of predicted ICAM-2 protein models for sooty mangabey and rhesus macaque. Exon structure is highlighted based on human ICAM-2. Alignment was performed using Jalview v.2.9.0. b, The sequence of exon 3 of CaICAM2 was confirmed in 10 additional individuals. Sequencing reads were aligned to the C. atys reference genome and visualized using Integrative Genomics Viewer (IGV). The red arrow indicates the position of the 499-bp genomic deletion in C. atys.
Extended Data Figure 3 | Predicted model of the ICAM2 gene structure and ICAM2 genomic sequence alignments. a, Predicted model of ICAM2 gene structure of M. mulatta and C. atys and the location of PCR primers for Sanger sequencing. Light blue, untranslated region; dark blue, CDS; red lines, intronic sequence; dotted line, exonic and intronic sequences present in human ICAM2 and MmICAM2 but not in CaICAM2; red box, the sequence that would be intronic in MmICAM2, but which is included in the exonic sequence of CaICAM2; light-purple box for CaICAM2 exon 4 represents the fact that the exon 4 sequence in MmICAM2 is present in CaICAM2 but is not included in the CaICAM2 CDS due to a stop codon in the CaICAM2 exon 3. Primer positions are indicated by arrows. Predicted PCR products are indicated by thick lines. Primers Ex3_F and Ex3_R were designed to amplify a region spanning a putative genomic deletion which includes the 3' region of CaICAM2 exon 3 and intron 3. b, Alignment of ICAM2 genomic sequences. Sanger sequencing of 2 rhesus macaques and 2 sooty mangabeys (including the Caty_1.0 reference animal) was performed to confirm the ICAM2 genomic deletion specific to C. atys. Starting at MmICAM2 nucleotide position 3166, sequences were aligned using Jalview v.2.9.0. Dashed lines denote the deletion in C. atys. RM, rhesus macaque; SM, sooty mangabey.
Extended Data Figure 4 | ICAM2 splice junction analysis in C. atys and M. mulatta by RNA-seq read alignment. 

a, Quantification of observed splicing. Splice site counts for RNA-seq read alignments were added together and sites with more than 100 total reads were compared to find the proportion of reads supporting each splice variant or intronic retention. 

b, MmICAM2 splicing analysed by RNA-seq read alignment to the reference genome and visualized in IGV. 
c, CaICAM2 splicing analysed by RNA-seq read alignment to the reference genome and visualized in IGV.
Extended Data Figure 5 | Sequence alignment of TLR-4 and the structure of the TLR4 gene. a, Pairwise alignment of TLR-4 protein sequences for C. aty and M. mulatta. The sequence difference at the C terminus is highlighted in red. Sequences were aligned using Jalview v.2.9.0. b, TLR4 gene structure and location of PCR primers. Light blue, untranslated region; dark blue, CDS; red lines, intronic sequence. Primer positions are indicated by arrows. Predicted PCR product is indicated by thick line. Primers TLR4_F and TLR4_R were designed to amplify a region including a putative stop-loss mutation present in CaTLR4 but not in MmTLR4. c, Chromatograms showing stop-loss (indicated by arrows) in the TLR4 gene in C. aty with respect to M. mulatta. The relevant codon is underlined.
Extended Data Figure 6 | TLR-4 C terminus sequence alignments.

**a**. The sequence of the CaTLR-4 C terminus was confirmed in 10 additional individuals. Sequencing reads were aligned to the Caty_1.0 reference genome and visualized in IGV. The red arrow indicates the position of the G-to-A stop codon mutation that can be found in MmTLR4 but not CaTLR4.

**b**. Alignment of genomic sequences encoding the TLR4 C terminus from different primate species. Starting at human TLR4 nucleotide position 2461, sequences were aligned using Jalview v.2.9.0.
Extended Data Figure 7 | Maximum likelihood gene tree of TLR4. This topology corresponds to the accepted species relationships for Old World monkeys. However, low bootstrap support among the nodes ancestral to C. atys, drill and baboon indicate that several sites within the gene do not support that ordering and may be indicative of incomplete lineage sorting. The table on the left shows these sites.
Extended Data Figure 8 | Analysis of cytokine expression and release after activation of TLR-4. 

**a**, TNF release from whole blood upon stimulation with lipid-A. Whole blood was stimulated with lipid-A at the indicated concentrations for 4 h and cytokine secretion was measured by cytometric bead array. n = 5 biologically independent samples for *M. mulatta*; n = 4 biologically independent samples for *C. atys*. 

**b**, IL-6 release from whole blood upon stimulation with lipid-A. Whole blood was stimulated with lipid-A at the indicated concentrations for 4 h and cytokine secretion was measured by cytometric bead array. n = 8 biologically independent samples for *M. mulatta*; n = 9 biologically independent samples for *C. atys*. 

**c**, SIVsmm plasma viral load for *M. mulatta* and *C. atys*. SIVsmm RNA levels in plasma were quantified at the indicated time points after intravenous inoculation with a primary uncloned SIVsmm isolate. n = 5 biologically independent samples for each species. 

**d**, TLR4 mRNA levels in LPS-stimulated blood samples. To test the level of TLR4 expression in the LPS-stimulated blood samples shown in Fig. 3e, we isolated RNA from whole blood from time-point matched replicate samples using PAXgene Blood RNA tubes, and analysed expression using Affymetrix GeneChip Rhesus Macaque Genome Arrays, which contains three independent probesets specific for MmTLR4 (denoted on the x axis). Probeset intensities are displayed along the y axis as RNA normalized values. n = 3 biologically independent samples for *M. mulatta*; n = 4 biologically independent samples for *C. atys*. 

**a**–**d**, Dots represent individual animals, and the bar represents the mean. Unpaired two-sided Student's *t*-test, *P* values are indicated. 

**e**, TNF and IL6 mRNA levels in LPS-stimulated monocytes from *M. mulatta* and *C. atys*. RNA-seq was used to assay global changes in gene expression after LPS stimulation of primary CD14+ monocytes. Significance for comparisons of mRNA levels of individual genes was tested using the Wald test as part of the DESeq2 workflow. Bars represent group means, and dots represent read counts for individual samples normalized to library size. Indicated *P* values are adjusted using the Benjamini–Hochberg correction. n = 6 biologically independent samples for each species.
Extended Data Figure 9 | LPS-mediated induction of TNF and IL-6 inflammatory signalling is globally attenuated in C. atys. a, b. Data shown are the leading-edge genes depicted in Fig. 3f, g (GSEA plots), for TNF-signalling genes (a) and IL-6-signalling genes (b). Values are the log2-transformed difference between LPS-treated and untreated samples for each individual animal. Genes selected are the combination of leading-edge/core-enriched genes for *M. mulatta* and *C. atys* GSEA analyses for each pathway. The gene sets selected for enrichment testing were obtained from the MSIGDB database hallmark collection are denoted at the top of each panel. Genes were organized using hierarchical clustering with Spearman dissimilarity and average linkage to estimate distance between genes and clusters, respectively. The colour scale at the bottom denotes the maximum and minimum on a log2 scale. For animal study source data, see Supplementary Table 2.
Extended Data Table 1 | Amino acid divergence in proteins from C. atys identified by the immunogenomic comparison pipeline

| Gene     | Function                                                                 | RM length (aa) | SM length (aa) | Identity (%) |
|----------|---------------------------------------------------------------------------|----------------|----------------|--------------|
| CD24     | B cell and granulocyte activation/differentiation                          | 78             | 77             | 88.5         |
| APOBEC3C | retroviral restriction factor                                              | 190            | 190            | 91.6         |
| DEFB129  | antimicrobial                                                              | 183            | 183            | 94.5         |
| CLEC2D   | inhibits NK-cell-mediated lysis                                            | 198            | 199            | 94.5         |
| GZMA     | cell lysis mediated by CD8+ T cells and NK cells                          | 262            | 262            | 94.7         |
| PGLYRP1  | Peptidoglycan recognition on gram-positive bacteria                       | 196            | 196            | 94.9         |
| CCL24    | chemoattractant for resting T cells                                       | 119            | 119            | 95.0         |
| C5AR1    | complement receptor                                                        | 350            | 350            | 95.1         |
| BST2     | retroviral restriction factor                                              | 182            | 182            | 95.6         |
| PF4      | coagulation, chemoattractant for neutrophils and monocytes                | 196            | 196            | 96.0         |
| S100A7   | antimicrobial, immunomodulatory                                            | 101            | 101            | 96.0         |
| CLEC6A   | mannose-dependent pathogen recognition, proinflammatory                  | 209            | 209            | 96.2         |
| MB21D1   | antiviral, cytosolic DNA sensor                                           | 522            | 522            | 96.2         |
| BPI      | antimicrobial, LPS-sensing                                                | 487            | 487            | 96.3         |
| PRG3     | cytotoxic and cytostimulatory activities                                  | 225            | 225            | 96.4         |
| GSDMD    | antimicrobial, pyroptosis                                                 | 484            | 484            | 96.5         |
| CLEC4A   | Pattern recognition receptor                                              | 204            | 204            | 96.6         |
| CLEC4D   | inflammation and immune responses                                        | 215            | 215            | 96.7         |
| PPBP     | chemoattractant and activator of neutrophils                              | 128            | 128            | 96.8         |
| CD4      | T cell receptor activation, HIV/SIV receptor                              | 458            | 458            | 96.9         |
| CTSG     | lysosomal antigen processing                                              | 255            | 255            | 96.9         |
| CD33     | adhesion molecule on myeloid cells                                        | 359            | 359            | 96.9         |
| LY96/MD2 | associates with TLR4 for LPS binding                                     | 160            | 160            | 96.9         |
| CCL11    | chemoattractant for eosinophils                                           | 97             | 97             | 96.9         |

aa, amino acids.
Extended Data Table 2 | Analysis of immune gene families across species

Panel A

| Change Type | gene family | function | SM  | AGM | RM  | Human | Chimp | Baboon |
|-------------|-------------|----------|-----|-----|-----|-------|-------|--------|
| Expansion (+5) | ADAM metalloproteinas | cytokine regulation | 30  | 20  | 27  | 22    | 18    | 24     |
| Expansion (+6) | scavenger receptors | LDL binding | 17  | 9   | 11  | 9     | 10    | 10     |
| Expansion (+6) | butyrophilin | lymphocyte deactivation | 16  | 10  | 9   | 9     | 7     | 10     |
| Expansion (+3) | TNFRSF10/TRAIL | apoptosis induction | 6   | 4   | 3   | 4     | 5     | 3      |
| Expansion (+2) | CD300 | lipid-binding, immunomodulation | 5   | 3   | 2   | 3     | 3     | 3      |
| Contraction (-3) | C-C-motif chemokines | chemoattractant for immune cells | 6   | 9   | 10  | 20    | 9     | 10     |

Panel B

| | $\lambda$ (No Error Model) | $\epsilon$ (Estimated error) | $\lambda$ (Error Model $= \frac{\lambda}{\epsilon}$) |
|------------------|-----------------------------|---------------------------------|---------------------------------|
| 10 species in this study | 0.00268 | 0.04268 | 0.00204 |
| 11 species Gibbon Genome Project | 0.00258 | 0.04101 | 0.00141 |
| 10 mammal dataset | 0.00238 | 0.07324 | 0.00186 |

Panel C

|                      | Expansions | Contractions | No Change | Avg. Expansion |
|----------------------|------------|--------------|-----------|----------------|
|                      | Families   | Genes gained | genes/ expansion | Families | Genes lost | genes/ contraction |         |          |
| Sooty                | 535 (96)   | 1153         | 2.16      | 340 (48)   | 494        | 1.45          | 10106   | 0.024528 |
| Human                | 1042 (276) | 3471         | 3.33      | 192 (10)   | 210        | 1.09          | 9747    | 0.200967 |
| Marmoset             | 1027 (122) | 2213         | 2.15      | 668 (23)   | 841        | 1.26          | 9286    | 0.107504 |
| Chimp                | 161 (23)   | 384          | 2.39      | 874 (69)   | 1137       | 1.3           | 9946    | -0.081244|
| Gibbon               | 354 (13)   | 552          | 1.56      | 1089 (92)  | 1466       | 1.35          | 9538    | -0.085529|
| Baboon               | 290 (61)   | 660          | 2.28      | 624 (41)   | 737        | 1.18          | 10067   | -0.028084|
| Orang                | 548 (65)   | 1032         | 1.88      | 749 (14)   | 820        | 1.09          | 9684    | -0.003921|
| Macaque              | 1101 (203) | 2904         | 2.64      | 783 (22)   | 835        | 1.07          | 9097    | 0.100666 |
| Mouse                | 631 (38)   | 2719         | 4.31      | 855 (9)    | 1027       | 1.2           | 9495    | 0.013404 |
| Vervet               | 294 (19)   | 658          | 2.24      | 674 (59)   | 921        | 1.37          | 10013   | -0.039209|

a. Expansion and contraction of immune gene families across six primate species. b. Assembly and annotation error estimations and gene gain and loss rates in a single $\lambda$ model in 13 mammals. c. Summary of gene gain and loss events inferred after correcting for annotation and assembly errors across all 13 species. The number of rapidly evolving families is shown in parentheses for each type of change. AGM, African green monkey.
**Extended Data Table 3 | Correlation analysis between TLR-signalling molecules and gene expression**

Pearson’s correlation coefficients ($r$) were calculated separately for cytokines from *C. alyi* and *M. mulatta* (TNF or IL-6) protein measurements versus mRNA levels of TLR-4-signalling genes measured in matched blood samples using Affymetrix GeneChips. $P$ values denote the significance of the Pearson’s correlation coefficient. CI, confidence interval.

| Gene Name | Gene Symbol | Affymetrix Probeset ID | r – SM TNF | p value – SM TNF | Lower CI TNF | Upper CI TNF | r – RM TNF | p value – RM TNF | Lower CI RM | Upper CI RM |
|-----------|-------------|------------------------|------------|----------------|-------------|-------------|------------|----------------|-------------|-------------|
| AP1       | JUN         | MmuG00012219.1_S1_at    | -0.52      | 0.07           | -0.83       | 0.04        | 0.2         | 0.7            | -0.54       | 0.71        |
| CD14      | CD14        | MmuG00012219.1_S1_at    | -0.28      | 0.35           | -0.72       | 0.32        | 0.4         | 0.41           | -0.41       | 0.78        |
| IKKα      | CHUK        | MmuG00012219.1_S1_at    | 0.25       | 0.41           | -0.35       | 0.71        | -0.31       | 0.34           | -0.79       | 0.40        |
| IKKβ      | IKBKB       | MmuG00012219.1_S1_at    | -0.51      | 0.07           | -0.83       | 0.05        | 0.05        | 0.88           | -0.60       | 0.66        |
| IKKγ      | IKBKG       | MmuG00012219.1_S1_at    | -0.26      | 0.4            | -0.71       | 0.34        | 0.14        | 0.7            | -0.54       | 0.71        |
| IRAK1     | IRAK1       | MmuG00012219.1_S1_at    | -0.33      | 0.28           | -0.74       | 0.27        | 0.3         | 0.39           | -0.40       | 0.78        |
| IRF7      | IRF7        | MmuG00012219.1_S1_at    | -0.6       | 0.03           | -0.86       | -0.07       | -0.16       | 0.66           | -0.72       | 0.52        |
| JNK       | MAPK8       | MmuG00012219.1_S1_at    | 0.48       | 0.1            | -0.10       | 0.81        | -0.27       | 0.45           | -0.77       | 0.43        |
| MYD88     | MYD88       | MmuG00012219.1_S1_at    | -0.41      | 0.17           | -0.73       | 0.18        | -0.08       | 0.83           | -0.67       | 0.58        |
| NFκB1     | NFκB1       | MmuG00012219.1_S1_at    | -0.09      | 0.34           | -0.73       | 0.31        | 0.2         | 0.58           | -0.19       | 0.74        |
| NFκB2     | NFκB2       | MmuG00012219.1_S1_at    | 0.04       | 0.08           | -0.82       | 0.07        | 0.06        | 0.86           | -0.59       | 0.67        |
| P38       | MAPK1       | MmuG00012219.1_S1_at    | 0.43       | 0.15           | -0.16       | 0.79        | 0.09        | 0.8            | -0.57       | 0.68        |
| RIP1      | RIP1        | MmuG00012219.1_S1_at    | -0.36      | 0.23           | -0.76       | 0.24        | 0.21        | 0.57           | -0.49       | 0.74        |
| TAB1      | TAB1        | MmuG00012219.1_S1_at    | -0.15      | 0.63           | -0.65       | 0.44        | 0.42        | 0.23           | -0.29       | 0.83        |
| TAK1      | TAK1        | MmuG00012219.1_S1_at    | 0.87       | 0.0             | 0.62        | 0.96        | -0.3        | 0.43           | -0.78       | 0.41        |
| TBK1      | TBK1        | MmuG00012219.1_S1_at    | -0.09      | 0.78           | -0.61       | 0.49        | -0.5        | 0.14           | -0.86       | 0.19        |
| TIRAP      | TIRAP       | MmuG00012219.1_S1_at    | -0.48      | 0.1            | -0.91       | 0.1         | 0.79        | -0.57           | 0.69        | 0.29        |
| TLR4      | TLR4        | MmuG00012219.1_S1_at    | -0.19      | 0.52           | -0.67       | 0.4          | -0.69       | 0.03           | -0.92       | 0.10        |
| TRAF6     | TRAF6       | MmuG00012219.1_S1_at    | 0.48       | 0.09           | -0.09       | 0.82        | 0.04        | 0.91           | -0.60       | 0.65        |
| TRAM      | TRAM        | MmuG00012219.1_S1_at    | -0.24      | 0.43           | -0.7          | 0.36       | -0.6       | 0.07           | -0.89       | 0.05        |
| TRAF5     | TRAF5       | MmuG00012219.1_S1_at    | -0.35      | 0.23           | -0.76       | 0.24        | 0.58        | 0.08           | -0.08       | 0.88        |

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   The primary data reported in this study is the DNA genome sequence of a monkey species, the Sooty Mangabey. 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 most representative of any single animal. To verify sequencing findings and estimate population penetrance of alleles, we sequenced the genome of 10 additional animals - these animals were selected as they collectively represented 76% of the offspring of the Yerkes National Primate Research colony.

   For RNA-Seq, and microarray we used a power calculation performed on a model set of rhesus macaques in which RNA expression at different effect sizes (fold-changes) gave us an estimate of our sensitivity to detect changes to approximately 1.5 fold. For LPS stimulation studies, as the effect sizes (species differences) were much larger than most gene expression differences based on pilot stimulations, we also used this power calculation dataset.

2. Data exclusions

   Describe any data exclusions.

   To identify novel immunogenetic factors specific for SMs that may be involved in their ability to avoid progression to AIDS, we established a bioinformatic pipeline for a comparative protein analysis. During this process we excluded proteins based on certain criteria: The SM protein model alignment with the lowest e-value or highest bitscore (for equal e-values) was selected for each MacAm protein model, yielding the set of orthologous SM protein predictions most like the rhesus protein models. SM CDS sequences were screened against the de novo RNA-seq assembly transcript models by alignment with blat and an alignment score was calculated as the number of matching bases reduced by the number of CDS sequence bases missing in alignment gaps normalized by the CDS sequence length. This score penalizes bases missing from the CDS sequence without penalizing extra sequence that may have been added to the RNA-seq transcript model during the assembly process. We then excluded SM protein sequences that are not supported by the RNA-seq data (<99 % matching sequence), thereby eliminating those that had been incorrectly identified as divergent from rhesus macaque due to errors in transcript or protein models. We also excluded 6,906 SM protein models with more than 97% identity with rhesus orthologs from further analysis. Proteins of interest for differential response to lentivirus infection may be more divergent than expected on average. We thus filtered the remaining divergent SM proteins (i.e. those with less than 97% identity) that have high sequence confidence using the GO term “immune response,” which after manual inspection resulted in a list of 34 candidate immune-related genes that are sequence divergent between SMs and rhesus macaques (Table 2, Extended data table 1).

   Sequencing and de novo assembly of RNA-seq transcripts required 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.
3. Replication

Describe whether the experimental findings were reliably reproduced.

We reproduced experimental findings in the following manner:
1. To validate our findings of sequence divergence in the sooty mangabey genome, we performed full genome sequencing of an additional 10 animals. We also conducted Sanger sequencing of genomic DNA of genes of interest, and used available RNA-Seq data to verify the sequence data.
2. For the LPS stimulation data, we performed LPS stimulation of 9 RMs and 8 SMs -over three separate experiment days, using multiple dosages. Each data point shown represents the average of three technical replicates per animal. We repeated the LPS experiment in SIV-infected SMs (n = 5) and RMs (n = 4). The finding of reduced TNFa and IL6 cytokine release was measured by protein-based quantiation (CBA), by real-time PCR for RNA levels, and by RNA-Seq - with the observation of reduced production consistent across the different assays. We did not observe a failure to reliably produce our findings.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Where possible samples were run in single batches using equal numbers of samples from species compared (i.e. rhesus macaque and sooty mangabey), thus they were balanced. For the in vivo SIV infection experiment, groups were divided 3 SM + 2 RMs for processing group 1 and 2 SMs and 3 RMs for processing group 2. For the cytokine bead array experiments, ICAM2 surface expression measurements, multiple batches were measured over several days - to minimize batch effects, we would run 3-5 individuals from each species, with an equal number of RM and SM individuals run in a single batch, all samples for a given individual animal (i.e. Time points or stimuli doses) were assayed together. For RNA-Seq processing, samples were run in a single batch for all processing steps, and for sequencing, bar-coded cDNA fragments from all samples were pooled and split evenly across multiple lanes to reduce lane bias.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Blinding was incorporated for RNA-seq and microarray analysis by assigning samples de-identified inventory codes prior to submission for processing in which no information about individual, species or treatment condition was provided.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

7. Software

Describe the software used to analyze the data in this study.

ALLPATHS-LG, Atlas-Link (v.1.0), Atlas GapFil, Phrap, Newbler, Velvet, GraphPad Prism, PBjelly (v14.9.9), GATK, Trinity, blastp v2.2.28+, FlowJo 10, cufflinks v2.1.1,
For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

### Materials and reagents

**Policy information about availability of materials**

**Materials availability**

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

| No restrictions |
|------------------|

**Antibodies**

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

- CD3-APC/Cy7 (SP34-2), CD14-PE/Cy7 (M5E2) and CD20-PE/Cy5 (2H7) from BD; CD4-BV650 (OKT4), CD8-BV711 (RPA-T8), ICAM-2-FITC (CBR-IC2/2), Mouse IgG2α(κ)-FITC (MOPC-173) isotype control from Biolegend; Live/Dead Fixable Aqua from Thermo Fisher Scientific. Antibodies were validated using matching isotype control.

**Eukaryotic cell lines**

- **HEK293T from ATCC.**
  - Cells were purchased from ATCC who provided an authentication certificate.
  - Cells are routinely tested for mycoplasma.
  - The 293T variant of HEK cells is not listed in the ICLAC database. 293T cells are widely used to study protein variants in a well-defined cellular environment by transfection.
Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

Blood draws were obtained from sooty mangabeys and rhesus macaques housed at the Yerkes National Primate Research Center, which is accredited by American Association of Accreditation of Laboratory Animal Care. This study was performed in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health, a national set of guidelines in the U.S. and also to international recommendations detailed in the Weatherall Report (2006). This work received prior approval by the Institutional Animal Care and Use Committees (IACUC) of Emory University (IACUC protocol #2000793, entitled "Comparative AIDS Program").

Sooty mangabey (RefSeq animal):
GenBank Bioproject, species, specimen, sex, age (years) PRJNA279144, sooty mangabey, blood draw, Female, 23

Sooty mangabey diversity sequencing panel:
GenBank BioSample, species, specimen, sex, age (years) SAMN02900445, sooty mangabey, blood draw, Female, 23
SAMN02900446, sooty mangabey, blood draw, Female, 23
SAMN02900447, sooty mangabey, blood draw, Male, 31
SAMN02900448, sooty mangabey, blood draw, Male, 20
SAMN02900449, sooty mangabey, blood draw, Male, 20
SAMN02900450, sooty mangabey, blood draw, Female, 19
SAMN02900451, sooty mangabey, blood draw, Male, 17
SAMN02900452, sooty mangabey, blood draw, Female, 16
SAMN02900453, sooty mangabey, blood draw, Female, 16
SAMN02900454, sooty mangabey, blood draw, Male, 15

Rhesus macaque SIV infection study:
Animal study ID, species, specimen, sex, age (years) RM1, rhesus macaque, blood draw, Male, 12
RM2, rhesus macaque, blood draw, Male, 12
RM3, rhesus macaque, blood draw, Male, 9
RM4, rhesus macaque, blood draw, Male, 9
RM5, rhesus macaque, blood draw, Male, 12
RM6, rhesus macaque, blood draw, Male, 5
RM7, rhesus macaque, blood draw, Male, 5
RM8, rhesus macaque, blood draw, Male, 5
RM9, rhesus macaque, blood draw, Male, 5
RM10, rhesus macaque, blood draw, Male, 5

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Human biospecimens (blood) were obtained from healthy donors by blood draw into EDTA tubes at the Yerkes National Primate Research Center in accordance with internal review board protocol IRB0004582 and all relevant ethical regulations. Informed consent was obtained from all blood donors.
Flow Cytometry Reporting Summary

Form fields will expand as needed. Please do not leave fields blank.

Data presentation

For all flow cytometry data, confirm that:

- 1. The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- 2. The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- 3. All plots are contour plots with outliers or pseudocolor plots.
- 4. A numerical value for number of cells or percentage (with statistics) is provided.

Methodological details

5. Describe the sample preparation. Peripheral blood samples from SIV negative rhesus macaques and SIV negative sooty mangabeys were collected by venipuncture according to standard procedures at the Yerkes National Primate Research Center of Emory University and in accordance with U.S. National Institutes of Health guidelines. PBMCs were isolated by Ficoll density gradient centrifugation.

6. Identify the instrument used for data collection. BD LSRII

7. Describe the software used to collect and analyze the flow cytometry data. Cells were collected using BD FACSDIVA. Data was analyzed using FlowJo 10.

8. Describe the abundance of the relevant cell populations within post-sort fractions. Purity after sorting was >95% as confirmed by flow cytometry.

9. Describe the gating strategy used. Singlets were excluded using FSC-H and FSC-A. Lymphocytes were gated using FSC-A and SSC-A. Dead cells were excluded using LIVE/DEAD™ Fixable Aqua Dead Cell Stain Kit. Then CD3+CD8+ Cytotoxic T lymphocytes and CD3+CD4+ T helper cells and CD3-CD14-CD20+ B cells were defined. We present primary flow cytometry data in histograms since histograms are typically used to compare the staining of the isotype control antibody to the epitope-specific antibody in the same graph.

Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.