Ancient hybridization and strong adaptation to viruses across African vervet monkey populations

Hannes Sv达尔1,17, Anna Jasinska2,3, Cristian Apetrei4,5, Giovanni Coppola6,7, Yu Huang7, Christopher A Schmitt8, Beatrice Jacquelin9, Vasily Ramensky2,10, Michaela Müller-Trutwin9, Martin Antonio11, George Weinstock12, J Paul Grobler13, Ken Dewar14, Richard K Wilson12,17, Trudy R Turner13,15, Wesley C Warren16, Nelson B Freimer2,16, Magnus Nordborg1

Vervet monkeys are among the most widely distributed nonhuman primates, show considerable phenotypic diversity, and have long been an important biomedical model for a variety of human diseases and in vaccine research. Using whole-genome sequencing data from 163 vervets sampled from across Africa and the Caribbean, we find high diversity within and between taxa and clear evidence that taxonomic divergence was reticulate rather than following a simple branching pattern. A scan for diversifying selection across taxa identifies strong and highly polygenic selection signals affecting viral processes. Furthermore, selection scores are elevated in genes whose human orthologs interact with HIV and in genes that show a response to experimental simian immunodeficiency virus (SIV) infection in vervet monkeys but not in rhesus macaques, suggesting that part of the signal reflects taxon-specific adaptation to SIV.

Vervet monkeys (genus Chlorocebus; also known as African green monkeys) are highly abundant in savannahs and riverine forests throughout sub-Saharan Africa, as well as on several Caribbean islands where they were introduced during the colonial era. There is a long history of research on vervet monkeys, ranging from studies of their social behavior1–3 to their use as an important model for a variety of human diseases4–6. Vervet research colonies have been established, one of which is currently being genetically characterized2,8. Vervets are particularly interesting for HIV/AIDS research, as they are the most abundant natural hosts of SIV, a close relative of HIV. SIV is highly prevalent across African vervets, but infected individuals typically avoid progression to immunodeficiency, despite high viral loads6,10. Elucidating the genetic mechanisms for host defense against this virus in vervets may identify new targets for preventive and therapeutic interventions for HIV/AIDS.

Here we follow the publication of the vervet reference genome11 by presenting a genus-wide survey of polymorphism. The genus Chlorocebus has alternatively been viewed as a single species (Chlorocebus aethiops) with several subspecies or as five or six species with additional subspecies12. Our sampling strategy was intended to capture this diversity by including a total of 163 individuals from nine countries in Africa and two countries in the Caribbean, which harbor sizable feral populations (Fig. 1a and Supplementary Data 1). To our knowledge, no previous study has conducted genome-wide resequencing in a nonhuman primate in such a large sample and over such a geographically extensive area.

Vervets harbor extensive polymorphism, both within and between taxa, and we found clear evidence that taxonomic divergence involved gradual divergence and gene flow rather than following a simple branching pattern. A scan for diversifying selection across vervet taxa yielded gene enrichments much stronger than in similar studies on humans13. In particular, we report strong and highly polygenic selection signals affecting viral processes—in line with recent evidence that proposes a driving role for viruses in protein evolution in mammals14. These signals are furthermore enriched in genes with known relevance to SIV: either their human orthologs interact with HIV or they show a vervet-specific transcriptional response to SIV infection. Interestingly, rather than affecting genes with antiviral and inflammation-related functions15, selection in vervets appears to have primarily targeted

1Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter (VBC), Vienna, Austria. 2Center for Neurobehavioral Genetics, University of California, Los Angeles, Los Angeles, California, USA. 3Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland. 4Center for Vaccine Research, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. 5Department of Microbiology and Molecular Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. 6Department of Neurology, University of California, Los Angeles, Los Angeles, California, USA. 7State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, China. 8Department of Anthropology, Boston University, Boston, Massachusetts, USA. 9Institut Pasteur, Unité HIPER, Paris, France. 10Moscow Institute of Physics and Technology, Dolgoprudy, Russian Federation. 11Medical Research Council (MRC), The Gambia Unit, Fajara, The Gambia. 12Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA. 13Department of Genetics, University of the Free State, Bloemfontein, South Africa. 14Department of Human Genetics, McGill University, Montreal, Quebec, Canada. 15Department of Anthropology, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin, USA. 16McDonnell Genome Institute, Washington University in St. Louis, St. Louis, Missouri, USA. 17Present addresses: Department of Genetics, University of Cambridge, Cambridge, UK (H.S.) and Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, Ohio, USA (R.K.W.). Correspondence should be addressed to M.N. (magnus.nordborg@gmi.oeaw.ac.at).

Received 10 November 2016; accepted 27 September 2017; published online 30 October 2017; doi:10.1038/ng.3980
genes involved in the transcriptional regulation of viruses and, in particular, those that are harmful only under immunodeficiency, suggesting evolved tolerance to SIV rather than resistance against infection.

RESULTS
Sequencing and polymorphism detection
100-bp paired-end Illumina data were generated for all samples. Coverage was relatively low (median 4.4×), but at least one member of each taxon was sequenced to 10× coverage or higher (Supplementary Data 1). Employing a standard pipeline for alignment to the reference *Chlorocebus_sabaeus* 1.1 (ref. 11), derived from a St. Kitts–origin monkey, and joint variant detection across all samples, we discovered a total of over 97 million SNPs, 61 million of which passed our quality filters (Online Methods and Supplementary Figs. 1–4 and Supplementary Table 1).

Genetic relationships among vervet groups and SIV strains
Clustering of individuals based on pairwise genetic distance (Fig. 1b,c) and principal-component analysis (Supplementary Fig. 5) generally agreed with prior morphological and geographic classification and led us to define six African and two Caribbean taxonomic groups: *sabaeus* (West Africa), *aethiops*, *tantalus*, *hilgerti*, *cynosuros*, *pygerythrus*; and *sabaeus* (St. Kitts and Nevis) and *sabaeus* (Barbados). The genetic relatedness pattern also clearly confirmed the status of *sabaeus* as an outgroup to other vervet taxa (Supplementary Figs. 6 and 7, and Supplementary Note) and suggested isolation by distance within groups (Fig. 1b). Both geographic location and group identity contributed significantly to explaining the overall pattern of polymorphism (likelihood-ratio test, \( P < 2 \times 10^{-3} \) and \( P < 1 \times 10^{-52} \), respectively; Supplementary Table 2). Our data agree with the morphology-based taxonomy\(^{12} \) in that *sabaeus*, *aethiops* and *tantalus*...
Figure 2 Evidence for gene flow across taxa. (a) Admixture clustering of individuals. Each pie chart represents an individual, and colors represent contributions from five assumed admixture clusters. The choice of five clusters is discussed in the legend of Supplementary Figure 13. Full results are shown in Supplementary Figure 14. Colored lines mark comparisons in b and c. (b,c) MSMC plots of cross-coalescence rate, a measure of gene flow, across time (on a log scale). Shaded areas correspond to ±3 block jackknifing s.d. (d) UPGMA tree of pairwise distance matrix summarized by country. Arrows point to evidence of cross-taxon gene flow. (e) D statistic (ABBA-BABA test) for instances of gene flow shown in d. For full results, see Supplementary Figure 14 and Supplementary Data 2.

appeared to be well-defined taxa, whereas hilgerti, cynosuros and pygerythrus were comparatively closer to each other, exhibiting substantial amounts of shared variation and strongly correlated allele frequencies (Fig. 1c and Supplementary Fig. 8). However, while morphological evidence groups hilgerti and pygerythrus as a single species (Chlorocebus pygerythrus) distinct from cynosuros (Chlorocebus cynosuros), our data showed that pygerythrus and cynosuros are closer to each other than either is to hilgerti. Indeed, two pygerythrus individuals from Botswana were more closely related to cynosuros than to other pygerythrus individuals. This is probably due to admixture: as we note below, there was abundant evidence for admixture between these groups. Finally, the pattern of relatedness among SIV strains mirrored the pattern in vervets (Fig. 1d), suggesting that SIV existed in vervets before their initial divergence more than a half a million years ago and coevolved with the taxa.

Our data also confirmed that, as surmised from the historical record, Caribbean vervets are derived from Western African sabaeus. Additionally, we are able to clarify the relationship between the vervet populations on different Caribbean islands. The fact that vervets from Barbados were nearly as different from vervets from St. Kitts and Nevis as they were from Gambian sabaeus individuals (Fig. 1c) suggests that these two Caribbean populations were founded independently and experienced two independent bottlenecks (28% and 17% reduction in diversity relative to Gambia, respectively). The vervet population from Nevis, on the other hand, was genetically a subset of the St. Kitts population (17% reduction in diversity relative to St. Kitts) and likely was founded by individuals from this island, which is less than 4 km away. In human genetics, there is currently great interest in sequencing studies of recently expanded bottlenecked populations, both for elucidating population genetic processes and for identifying deleterious variants with a strong impact on phenotypes that have reached high frequency through drift. However, while the site-frequency spectrum in Caribbean vervets was generally biased toward higher allele frequencies, there was no evidence in our data that this effect was relatively stronger for putatively deleterious alleles (Supplementary Fig. 9).

Returning to Africa, variation within vervet taxa was much larger than in humans and other great apes, but was typical for that seen in other primates (Fig. 1c and Supplementary Fig. 10)—which is perhaps surprising given the ubiquity of vervets. Divergence between vervet taxa was generally higher than between subspecies of other primates, with average pairwise sequence divergence between taxa of ~0.4%, as compared to 0.2% to 0.32% across great ape subspecies, and relative genetic differentiation ($F_{ST}$) ranging from 25% to 71% (Fig. 1c and Supplementary Fig. 11), as compared to <15% across human populations or macaque subspecies. However, maximum sequence divergence was substantially lower than between human and chimpanzee (~1.24%). This intermediate status was supported by the presence of substantial amounts of both shared variation and fixed differences between vervet taxa (Supplementary Fig. 8).

Evidence for genetic admixture

The process that gave rise to the current taxa was much more complex than a series of population splits. We used Admixture to cluster individuals into groups. This analysis generally resolved the
The alternative explanation, ancestral population structure, seems less parsimonious, as the structure would have had to persist through the number of genes in a category (capped at 474). Terms are grouped using Cytoscape clustermaker.

Figure 3 Enrichment map network of GO categories enriched for high average gene selection scores. Edges represent overlap in genes. Colors represent $P$ values on a log scale (with red corresponding to the most highly significant, TopGO Kolmogorov–Smirnov weight $P < 1 \times 10^{-300}$). Node size represents the number of genes in a category (capped at 474). Terms are grouped using Cytoscape clustermaker.

Turning to the eastern and southern African cynosuros–hilgerti pygerythrus complex, we found that, while simple clustering suggests that hilgerti is an outgroup to cynosuros and pygerythrus (Fig. 1b,c), Admixture represents cynosuros individuals as a mixture of hilgerti and pygerythrus with a larger contribution of the former (Fig. 2a). MSMC suggested a complex history of varying gene flow between the three groups (Supplementary Fig. 16). We also investigated the status of pygerythrus from Botswana, which appeared as a sister group to Zambian cynosuros in the clustering tree (Fig. 2d): the $D$ statistic indicated that they have an additional genetic contribution from South African pygerythrus ($D = 7.6\%$, jackknifing $P < 1 \times 10^{-300}$, Fig. 2e), and MSMC confirmed an intermediate status of Botswanian pygerythrus with comparable levels of genetic exchange with both South African pygerythrus and cynosuros until total separation from both groups 10,000 years ago (Fig. 2c), again compatible with isolation by distance. In summary, while inferred genetic relationships are generally consistent with current taxonomy, strong signals of excess allele sharing along geographic axes suggest that the evolutionary history of these taxa involved processes of gradual divergence, isolation and secondary contact.
Strong signals of selection

Our data provide an opportunity to look for signals of adaptation on a continent-wide scale, across multiple taxa. To identify footprints of selection, we used an approach that incorporates information on both the distortion of allele frequency spectra within groups and the increase in differentiation among pairs of groups at loci close to a group-specific selective sweep (XP-CLR)\(^{26}\). To summarize the 30 XP-CLR comparisons between African taxa (Supplementary Figs. 17–22), we calculated ‘selection scores’—the root-mean-square XP-CLR scores (across taxon comparisons)—on a 1,000-bp grid along the genome.

These scores clearly captured strong signals of selection, as they were significantly higher in genic than in intergenic regions (one-sided Mann–Whitney U test, \(P < 1 \times 10^{-30}\), Supplementary Fig. 23). To gain further insight, we compared the distribution of average selection scores for genes (Supplementary Data 3) across Gene Ontology (GO) terms with the R package TopGO. Testing for enrichment using the relative rank of all scores yielded stronger signals than testing genes with the highest scores against the background, suggesting that weaker, polygenic effects contribute strongly to the signal of selection (Supplementary Fig. 24). We found 157 significantly enriched GO terms, many of which are related to RNA transcription and cell signaling (Fig. 3, Supplementary Fig. 25 and Supplementary Data 4).

These GO enrichments showed partial overlap with similar scores comparing human populations (Supplementary Fig. 26)\(^{13}\). However, there were many more significantly enriched GO terms for vervets than for humans and shared enrichments were generally much more significant in the current data set, suggesting that vervet taxa provide a powerful model to study diversifying selection across closely related primate taxa. The strongest selection scores are consistent with a dominant role of viral pathogens as selective agents in vervets. In particular, we note viral process (\(P = 5 \times 10^{-9}\)) and positive and negative regulation of transcription from the polymerase II promoter (\(P = 5 \times 10^{-17}\) and \(5 \times 10^{-14}\)), which is known to interact with viral proteins (for example the HIV Tat protein during transcription elongation of HIV-1 LTR)\(^{27}\). These virus-related categories were not only enriched in the root-mean-square summary but also in many two-taxon XP-CLR comparisons (Supplementary Note). Furthermore, these categories did not show particularly large neutrality indices or significant enrichment for conserved elements\(^{28}\) (Supplementary Figs. 27–30 and Supplementary Note), providing evidence that these signals are not predominantly driven by purifying selection (background selection), which can lead to confounding signals\(^{29}\).

To test more specifically for virus-related selection signals, we looked for enrichment of signals among the orthologs of human HIV-interacting genes. Indeed, 43 of 166 gene sets in the NCBI HIV-1
human interaction database showed significant enrichment for high selection scores (Supplementary Fig. 31 and Supplementary Data 5). However, we note that 21 and 71 partly overlapping gene categories from this database also showed enrichment for human selection scores and conserved elements, respectively (Supplementary Figs. 31 and 32, and Supplementary Note), suggesting that these gene sets are not very specific.

Selection signals linked to vervet-specific response to SIV

SIVAgm is prevalent in African vervets and has diverged into taxon-specific strains (Fig. 1d). Furthermore, while SIVAgm is highly pathogenic when used experimentally to infect pigtailed macaques that are not natural SIV hosts, infected vervets generally do not progress to AIDS, suggesting coevolution of virus and host. We hypothesized that coevolution between taxon-specific SIV strains and vervet taxa could lead to an ongoing evolutionary arms race that would manifest itself as diversifying selection across taxa, specifically on genes involved in host defense (whereas adaptations shared across the genus would be very difficult to detect). To test this hypothesis, we reanalyzed microarray data comparing the transcriptional response of vervets and macaques to infection with SIV. If some of the selection signals reflect adaptation to SIV in vervets, we would expect selection scores to be elevated in genes that are differentially expressed in vervets—but not in macaques—as a response to infection. Indeed, selection scores were much higher in genes that showed a significant difference in expression before and after infection in vervets only, as compared to genes showing a difference in expression in both species (one-sided Mann–Whitney U test, $P = 5 \times 10^{-9}$) or in macaque only ($P = 1 \times 10^{-4}$) (Supplementary Fig. 33). Conversely, vervet-specific (but not shared or macaque-specific) differentially expressed genes were significantly enriched in high selection scores ($P = 0.003$, $P > 0.99$ and $P > 0.99$, respectively).

To further investigate the underlying mechanisms, we grouped differentially expressed genes by coexpression patterns using weighted gene coexpression network analysis (WGCNA; Supplementary Figs. 34 and 35). Five of the 33 gene coexpression modules showed a significant enrichment for genes with high selection scores (FWER $< 0.05$; Fig. 4a, Supplementary Fig. 35 and Supplementary Data 6). The significant modules had similar expression patterns with strong changes in vervets post-infection and weak, inconsistent signals in macaques. In particular, all the modules that were enriched for diversifying selection showed changes in gene expression in vervets 6 d post-infection, which is around the time that the virus becomes detectable and activates early immune responses. Two modules also showed...
expression differences in the chronic stage (115 d post-infection), which is most relevant for progression to immunodeficiency. We ran GO enrichment analysis separately on the genes in the enriched WGCNA modules showing early (‘acute’) and late (‘chronic’) expression changes (Fig. 4). We found 30 and 20 significantly enriched GO categories, respectively, many of which are involved in response to HIV in humans (Supplementary Data 7 and 8). For example, for early expression response, enriched GO categories included clathrin-mediated endocytosis35, autophagosome assembly, positive regulation of type I interferon (IFN-I) production35,36, and innate immune response (marginally significant at $P = 0.01004$). This is consistent with recent findings that in macaques the type I IFN response is delayed following SIV infection and inhibited during the first week of SIV infection39, while natural hosts mount a very early and transient type I IFN response35,36. Conversely, the three most highly enriched GO categories for genes in modules with late expression changes were positive regulation of natural killer (NK) cell activation, regulation of cellular response to heat40 and somatic hypermutation of immunoglobulin genes41, consistent with differences in NK cell responses during SIV infection in natural hosts as compared to non-natural hosts, the lack of viral replication in B cell follicles (T follicular helper ($T_{FH}$) cells) and preservation of lymph node immune function in natural hosts in contrast to macaques, as well as better adaptation to the stress induced by chronic infection42-45.

Candidate targets of selection

While enrichment analysis identifies categories of genes under selection and is likely driven by large numbers of genes with moderate effects, the highest selection scores identify candidate regions for strong selection (Fig. 5). The highest score was for an uncharacterized gene on chromosome 6 (Fig. 5b) with 97% sequence identity to the human gene encoding RAN-binding protein 3 (RANBP3), a protein connected to influenza A virus replication42 that is involved in nucleocyttoplasmic export of RNA from human T cell leukemia virus type 1 (HTLV-1)47 and HIV48,49. Another gene that displayed among the highest selection scores was NFIX (nuclear factor I/X; Fig. 5d), which encodes a transcription factor that binds the palindromic sequence $5′$-TTGGCNNNNGCCAA-3′ in viral and cellular promoters. Nuclear factor I proteins can serve as a transcription selectivity factor for RNA polymerase II and have a critical role in transcription and regulation of JC virus in humans50 and simian virus 40 in vervet cells51. These closely related viruses are usually harmless but cause disease under immunodeficiency, specifically in SIV infection in macaques52 and HIV infection in humans53. However, the lack of common genetic variants in the coding sequence of this gene suggests that selection is more likely to have targeted regulatory variants.

There has been considerable debate about vervet taxonomy, both concerning the taxonomic levels of different groups (species or subspecies) and relationships between groups. While taxonomic assignment can reflect a variety of morphological, genetic and behavioral information, our results suggest that—despite evidence for substantial genetic exchange—Chlorocebus includes both genetically well-separated taxa (sabaeus, aethiops, tantalus and pygerythus) and more closely related groups (pygerythus, cynosurus and hilgerti). The latter groups would naturally fall at a taxonomic level below the former.

Different phylogenetic relationships between vervet taxa have been proposed. Using two different clustering algorithms, we find that West African sabaeus split off the common ancestor of other vervets first, followed by aethiops, with the last split separating tantalus from pygerythus (with the latter including cynosurus and hilgerti). This result is consistent with the findings of Warren et al.,11 who inferred the same branching pattern using whole-genome sequencing data from a single representative per taxonomic group. It contradicts the findings of Pfeifer54, who suggested that aethiops (rather than sabaeus) constitutes the outgroup of the vervet taxa. Pfeifer attributed this difference to Warren et al. having neglected to distinguish fixed from shared polymorphism. A potential reason for this discrepancy is that Pfeifer apparently used fixed differences only to infer taxon relatedness. This procedure can lead to erroneous conclusions because inferred branch lengths for taxa with comparatively high rates of drift (low effective population size), such as aethiops, will be biased upward and branch lengths for taxa that have been exchanging genes, such as sabaeus and tantalus, will be biased downward. By aligning a subset of the data against the macaque reference genome and computing genome-wide summaries of the coalescent histories of the samples, we strongly confirm the outgroup status of sabaeus (Supplementary Fig. 7 and Supplementary Note).

These results notwithstanding, our analyses also clearly show that a simple phylogenetic tree cannot fully capture the pattern of relatedness across vervets. As in several recent studies,56,57 the process of speciation must have been gradual and involved gene flow, leading to a fundamentally reticulate pattern of relatedness. Although we see no evidence for current gene flow between vervet taxa, definitive conclusions will require more appropriate sampling (especially in putative hybrid zones, as these taxa are expected to readily hybridize58,59) and sequencing strategies that are better able to resolve haplotypes.

Finally, we carry out a screen for diversifying selection across vervet taxa, our primary goal being to look for signs of adaption to SIV. To this end, we use a method that tests whether the change in allele frequency between taxa at a locus occurred too quickly to be due to random drift. This approach is expected to be sensitive to both recent and relatively older selection events and to pick up signals of recurrent adaptation56.

Gene ontology analysis yields strong enrichment of selection scores in multiple biological processes, generally driven by polygenic signals. Our data have the potential to yield insights into taxon-specific adaptations (for example, altitude adaptation in aethiops). In the present study, we focus on loci that show signals of repeated (but differential) adaptations across multiple taxa, consistent with host–pathogen coevolution. As hypothesized, our screen identified a strong excess of signal in genes that interact with viruses, consistent with findings in other organisms14,26. While there is potential for coevolution with different types of viruses in vervets, integration of our selection study with gene expression analysis of SIV-infected monkeys provides evidence that part of the signal results from vervet coevolution with SIV. Interestingly, the genes identified do not include the virus (co)receptor genes involved in the virus docking mechanism, but rather are genes...
involved in cell signaling and transcriptional regulation, consistent with recent results suggesting that natural selection has shaped pri-
mate CD4 T cell transcription\(^6\) and suggesting adaptation to living with the virus rather than avoidance of infection. Indeed, one of the highest scoring genes controls the expression of a virus known to cause disease under SIV-induced immunodeficiency.

While we are confident that our analysis is picking up real signals of selection in virus-related genes, it is difficult to determine the mode of selection conclusively. For example, we carried out several tests to confirm that our results were not primarily driven by purifying (rather than diversifying) selection, but further orthogonal approaches and functional validation will be necessary to ultimately understand the evolutionary dynamics of verties and their pathogens. The data and results presented here should aid this endeavor and may prove useful in the quest for antiviral vaccinations and therapies.

**URLs.** Vervet reference genome, [https://www.ncbi.nlm.nih.gov/assembly/GCF_000409795.2](https://www.ncbi.nlm.nih.gov/assembly/GCF_000409795.2); Los Alamos National Laboratory HIV sequence database, [https://www.hiv.lanl.gov/content/index; vervet reference gene annotation, [https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Chlorocebus_sabaeus/100/](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Chlorocebus_sabaeus/100/); MSMC2, [https://github.com/stschiff/msmc2; HIV-1 Human Interaction Database](https://github.com/stschiff/msmc2; HIV-1 Human Interaction Database) (last accessed January 2016), [https://www.ncbi.nlm.nih.gov/refsseq/HIVInteractions; IUCN Red List of Threatened Species (accessed September 2017), [http://www.iucnredlist.org](http://www.iucnredlist.org)/].

**METHODS**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

**ACKNOWLEDGMENTS**

Samples were collected through the UCLA Systems Biology Sample Repository funded by US National Institutes of Health grants R01RR016300 and R01OD010980 to N.F. For permits allowing us to collect samples, we thank the Gambia Department of Parks & Wildlife Management; the Botswana Ministry of Environment & Wildlife and Tourism; the Ghana Wildlife Division, Forestry Commission; the Zambia Wildlife Authority; the Ethiopian Wildlife Conservation Authority; the Ministry of Forestry & the Environment, Department of Environmental Affairs, South Africa; the Department of Economic Development and Environmental Affairs, Eastern Cape; the Department of Tourism, Environmental and Economic Affairs, Free State; Ezemvelo KZN Wildlife, KwaZulu-Natal; and the Department of Economic Development, Environment and Tourism, Limpopo. We also thank G. Redmond and the St. Kitts Biomedical Research Foundation for facilitating sample collection in St. Kitts and Nevis. We thank J. Brenchley, K. Reimann (R24OD010976), and J. Baulu and the Barbados Primate Research Center and Wildlife Reserve for providing samples of Tanzanian origin and Barbadian vervets. For help with sample collection and processing, we thank J. Danzy-Cramer, T. Jung, O. Morton and J. Freimer. We thank J. Kamon for discussion, U. Seren, J. Wasserscheid and N. Juretic for IT support, and R. Halai for help with figure design. H.S. has been supported by a travel grant from J. Kamm for discussion, Ü. Seren, J. Wasserscheid and N. Juretic for IT support, and R. Halai for help with figure design. H.S. has been supported by a travel grant from J. Kamm for discussion, Ü. Seren, J. Wasserscheid and N. Juretic for IT support, and R. Halai for help with figure design. H.S. has been supported by a travel grant from J. Kamm for discussion, Ü. Seren, J. Wasserscheid and N. Juretic for IT support, and R. Halai for help with figure design. H.S. has been supported by a travel grant from J. Kamm for discussion, Ü. Seren, J. Wasserscheid and N. Juretic for IT support, and R. Halai for help with figure design.

**AUTHOR CONTRIBUTIONS**

N.B.F., T.R.T., M.N., A.J.I., K.D., W.C.W. and R.K.W. conceived the study. M.N. and H.S. designed the analysis strategy. H.S. analyzed the data and prepared tables and figures. C.A. contributed the SV\textsubscript{yne} sequence analysis. G.C. contributed the WGCNA analysis. B.I. and M.M.-T. provided expertise on the expression data analysis and SIV. Y.H. and V.R. provided bioinformatic support. C.A.S., J.P.G., M.A. and T.R.T. collected samples and obtained permits. N.B.E., G.W., R.K.W., K.D. and W.C.W. oversaw sequencing. M.N., H.S., N.F. and A.J.J. wrote the manuscript. All authors read and approved the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Reprints and permissions information is available online at [http://www.nature.com/reprints/index.html](http://www.nature.com/reprints/index.html). Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Cheney, D.L. & Seyfarth, R.M. The recognition of social alliances by vervet monkeys. Anim. Behav. 34, 1722–1731 (1986).
2. Seyfarth, R.M., Cheney, D.L. & Marler, P. Vervet monkey alarm calls: semantic communication in a free-ranging primate. Anim. Behav. 28, 1070–1094 (1980).
3. Janssens, A.J. et al. Systems biology of the vervet monkey. ILAR J. 54, 122–143 (2013).
4. Briggs, C.M. et al. Live attenuated tetravalent dengue virus host range vaccine is immunogenic in African green monkeys following a single vaccination. J. Virol. 88, 7279–7284 (2014).
5. Matsumura, Y. et al. African green monkeys recapitulate the clinical experience with replication of live attenuated pandemic influenza virus vaccine candidates. J. Virol. 88, 8139–8152 (2014).
6. Pirzufu, N.S. et al. Exploring of primate models of tick-borne flaviviruses infection for evaluation of vaccines and drugs efficacy. PLoS One 8, e61094 (2013).
7. Huang, Y.S. et al. Sequencing strategies and characterization of 721 vervet monkey genomes for future genetic analyses of medically relevant traits. BMC Biol. 13, 41 (2015).
8. Jasinska, A.J. et al. Genetic variation and gene expression across multiple tissues and developmental stages in a nonhuman primate. Nat. Genet. http://dx.doi.org/10.1038/ng.3959 (2017).
9. Ma, D. et al. SHV\textsubscript{yne} infection in wild African green monkeys from South Africa: epidemiology, natural history, and evolutionary considerations. PLoS Pathog. 9, e1003011 (2013).
10. Ma, D. et al. Factors associated with simian immunodeficiency virus transmission in a natural African nonhuman primate host in the wild. J. Virol. 88, 5687–5705 (2014).
11. Warren, W.C. et al. The genome of the vervet (Chlorocebusaapellaabesaeus). Genome Res. 25, 1921–1933 (2015).
12. Enstam, K.L. & Isbel, L.A. In Primates in Perspective (ed. Campbell, C.J.) 252–274 (Oxford University Press, 2007).
13. Daub, J.T. et al. Evidence for polygenic adaptation to pathogens in the human genome. Mol. Biol. Evol. 30, 1544–1558 (2013).
14. Enard, D., Cai, L., Gwennap, C. & Petrov, D.A. Viruses are a dominant driver of protein adaptation in mammals. eLife 5, e12469 (2016).
15. Quach, H. et al. Genetic adaptation and Neandertal admixture shaped the immune system of modern humans. Cell 167, 643–656 (2016).
16. Jeffreys, E.M. et al. Revisiting an old niddle: what determines genetic diversity levels within species? PLoS Biol. 10, e1001388 (2012).
17. Prado-Martinez, J. et al. Great ape genetic diversity and population history. Nature 499, 471–475 (2013).
18. Perry, G.H. et al. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. Genome Res. 23, 602-610 (2012).
19. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).
20. Hernandez, R.D. et al. Demographic histories and patterns of linkage disequilibrium in Chinese and Indian rhesus macaques. Science 316, 240–243 (2007).
21. Ebersberger, I., Metzler, D., Schwarz, C. & Pääbo, S. Genomewide comparison of DNA sequences between humans and chimpanzees. Am. J. Hum. Genet. 70, 1490–1497 (2002).
22. Alexander, D.H., Nye, J. et al. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664 (2009).
23. Durand, E.Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28, 2239–2252 (2011).
24. Patterson, N. et al. Ancient admixture in human history. Genetics 192, 1065–1093 (2012).
25. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46, 919–925 (2014).
26. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. Genome Res. 20, 393–402 (2010).
27. Zhou, C. & Rana, T.M. A bimolecular mechanism of HIV-1 Tat protein interaction with RNA polymerase II transcription elongation complexes. J. Mol. Biol. 320, 925–942 (2002).
28. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050 (2005).
29. Zeng, K. & Corcoran, P. The effects of background and interference selection on patterns of genetic variation in subdivided populations. Genetics 201, 1539–1554 (2015).
30. Ako-Adjie, D. et al. HIV-1, human interaction database: current status and new features. Nucleic Acids Res. 43, D566–D570 (2015).
31. Kaposiánsky, B. et al. Local virus extinction following a host population bottleneck. J. Virol. 89, 8152–8161 (2015).
32. Müller, M.C. et al. Simian immunodeficiency viruses from central and western Africa: evidence for a new species-specific lentivirus in tantalus monkeys. J. Virol. 67, 1227–1235 (1993).
33. Goldstein, S. et al. Plateau levels of viremia correlate with the degree of CD4+-T-cell loss in simian immunodeficiency virus SIVagm-infected pigtailed macaques: variable pathogenicity of natural SIVagm isolates. J. Virol. 79, 5153–5162 (2005).
34. Mandell, D.T. et al. Pathogenic features associated with increased virulence upon simian immunodeficiency virus cross-species transmission from natural hosts. J. Virol. 88, 6778–6792 (2014).
35. 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).
36. Jacquelin, B. et al. Innate immune responses and rapid control of inflammation in African green monkeys treated or not with interferon-γ during primary SIVagm infection. PLoS Pathog. 10, e1004241 (2014).
37. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).
38. Mlyauchi, K., Kim, Y., Latinovic, O., Morozov, V. & Melikyan, G.B. HIV enters cells via endocytosis and dynamin-dependent fusion with endosomes. Cell 137, 433–444 (2009).
39. Barouch, D.H. et al. Rapid inflammasome activation following mucosal SIV infection of rhesus monkeys. Cell 165, 656–667 (2016).
40. Pan, X.-Y. Recent biomarker analysis on early immune responses and rapid control of inflammation in simian immunodeficiency virus (SIV)-infected rhesus monkeys. J. Clin. Invest. 129, 1077–1082 (2016).
41. Klein, F. et al. Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. Cell 153, 126–138 (2013).
42. Pereira, L.E., Johnson, R.P. & Ansari, A.A. Sooty mangabeys and rhesus macaques exhibit significant divergent natural killer cell responses during both acute and chronic phases of SIV infection. Cell. Immunol. 254, 10–19 (2008).
43. Myfthaler, M. et al. Early induction of polyfunctional simian immunodeficiency virus (SIV)-specific T lymphocytes and rapid disappearance of SIV from lymph nodes of sooty mangabeys during primary infection. J. Immunol. 186, 5151–5161 (2011).
44. Brenchley, J.M. et al. Differential infection patterns of CD4+ T cells and lymphoid tissue viral burden distinguish progressive and nonprogressive lentiviral infections. Blood 120, 4172–4181 (2012).
45. Zhang, R. et al. Envelope-specific B-cell populations in African green monkeys chronically infected with simian immunodeficiency virus. Nat. Commun. 7, 12131 (2016).
46. Predicata, R. & Zhou, Y. The role of Ran-binding protein 3 during influenza A virus replication. J. Gen. Virol. 94, 977–984 (2013).
47. Hakata, Y., Yamada, M. & Shida, H. A multifunctional domain in human CRM1 (exportin 1) mediates RanBP3 binding and multimerization of human T-cell leukemia virus type 1 Rex protein. Mol. Cell. Biol. 23, 8751–8761 (2003).
48. Langer, K., Dian, C., Rybin, V., Müller, C.W. & Petosa, C. Insights into the function of the CRM1 cofactor RanBP3 from the structure of its Ran-binding domain. PLoS One 6, e17011 (2011).
49. Shida, H. Role of nucleocytoplasmic RNA transport across the nuclear envelope. BioEssays 39, 1077–1082 (2016).
50. Novikova, P.Y. et al. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat. Commun. 6, 1077–1082 (2015).
51. Loseva, O. et al. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515, 261–263 (2014).
52. Novikova, P.Y. et al. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat. Commun. 6, 1077–1082 (2015).
53. Ferenczy, M.W. et al. Molecular biology, epidemiology, and pathogenesis of progressive multifocal leukoencephalopathy, the JC virus–induced demyelinating disease of the human brain. Clin. Microbiol. Rev. 25, 471–506 (2012).
54. Pfeifer, S.P. The demographic and adaptive history of the African green monkey. Mol. Biol. Evol. 34, 1055–1065 (2017).
55. Romiguier, J. et al. Comparative population genomics in animals uncovers the determinants of genetic diversity. Nature 515, 261–263 (2014).
56. Novikova, P.Y. et al. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat. Commun. 6, 1077–1082 (2015).
57. Mallet, J., Besansky, N. & Hahn, M.W. How reticulated are species? BioEssays 38, 140–149 (2016).
58. Matsuhashi, K., Hirai, M., Watanabe, T., Ohkura, Y. & Nozawa, K. A case of patas-verte hybrid in captivity. Primates 19, 785–793 (1978).
59. de Jong, Y.A., Butynski, T.M. Three Sykes’s monkey Cercopithecus mitis x vervet monkey Chlorocebus pygerythrus hybrids in Kenya. Primate Conserv. 25, 43–56 (2010).
60. Danko, C.G. et al. Natural selection has shaped coding and non-coding transcription in primate CD4+ T-cells. Preprint at bioRxiv https://doi.org/10.1101/083212 (2016).
61. Haus, T. et al. Mitochondrial diversity and distribution of African green monkeys (Chlorocebus sabe) in Kenya. Am. J. Primatol. 75, 350–360 (2013).
62. Hill, W.C.O. in Primates, Comparative Anatomy and Taxonomy 533–581 (Edinburgh University Press, 1966).
63. Demchak, B. et al. Cytoscape: the network visualization tool for GenomesSpace workflows. F1000Res 3, 151 (2014).
Sample collection and sequencing. All blood samples were collected under approved country-specific permits that met standardized bioprospecting regulations (except for in Botswana, where our work on population genomics was officially approved but bioprospecting would be the subject of a separate application). DNA samples were obtained from blood (PaxGene DNA tubes, ACD tubes or archival blood cell pellet collection) except for sample AG23, which was obtained from a B-lymphoblastoid cell line transformed with herpesvirus papio. Individuals were sequenced at variable coverage (Supplementary Data 1) on an Illumina HiSeq 2000 platform obtaining 100-bp paired-end reads.

Alignment, variant detection and filtering. Sequences were aligned against ChlSab1.1 (ref. 11) using bwa-mem64 with a total coverage of 798× and a median coverage of 4.4×. On average, more than 98% of the reads mapped for all taxa, suggesting that reference bias was weak. Following the GATK recommended workflow65,66, alignments against ChlSab1.1 (see URLs) were locally realigned, base quality scores were recalibrated using a first round of variant calling and variants were detected using the GATK UnifiedGenotyper. We also called variants using the GATK HaplotypeCaller but found these calls to have a strong bias toward homozygous reference alleles in the low-coverage samples. We hence only used UnifiedGenotyper variant calls for further analysis. Biallelic SNP calls were hard filtered with a combination of GATK best practices66 and custom filters to yield the data set used for further analysis (Supplementary Fig. 1 and Supplementary Table 1). We used VCFtools diff-indv-discordance to compare the genotype calls from Warren et al.11 and the current data set for the five individuals that were shared between the studies. For non-filtered SNPs, the discordance rate was 0.12% to 0.26%; for all non-filtered sites, it was 0.006% to 0.012%.

Given the large differences in coverage between individuals (2× to 45×), a stringent control on the false positive rate would have led to strong bias toward lower diversity (and especially a lower number of singletons) in low-coverage samples. We suggest that for population genomic analysis it is conservative to reduce bias at the cost of increased noise. Our data set did show correlation between coverage and individual heterozygosity (Pearson’s r = 0.48, P = 1×10⁻⁸⁰; Supplementary Fig. 3), especially for individuals with less than 4× coverage, but cross-individual sequence divergence within and between taxa was not strongly affected by coverage (Pearson’s r = 0.009 and 0.026, respectively; Supplementary Fig. 3). The ancestral state for each SNP was determined by aligning the macaque reference genome, rheMac2, against ChlSab1.1 using nucmer67, only considering one-to-one mappings with a minimum length of 200 bp.

To test whether our results could be affected by coverage bias, we repeated some of the analysis with a subset of the sequencing reads aligned to the rhesus macaque genome, Mmul 8.0.1 (Supplementary Figs. 6 and 36, and Supplementary Note).

Minor in silico contamination of the original read files of 18 South African individuals with RNA-seq reads was detected at a late stage of the project. While the effect of this on variant calls was very small (the median geno- type concordance between the original call set and an updated version with contamination removed was 99.91%), it led to some highly expressed genes being masked by the high-coverage filter (~2% additional PASS SNPs in the recall). We therefore repeated the entire selection analysis (Figs. 3–5 and Supplementary Figs. 15–31) using the updated call set.

Accessible genome size. To compare levels of polymorphism and divergence across individuals and to previous studies, we measured the proportion of the genome accessible to our variant detection process. In particular, we excluded all sites that did not pass our quality filters and, for each individual, all sites for which the UnifiedGenotyper could not make a genotype call (Ns) (Supplementary Fig. 2).

Diversity and divergence. Nucleotide diversity was calculated by computing the number of pairwise differences for each comparison divided by the accessible genome size for each pair as derived above. For each group, nucleotide diversity was estimated as the average of within-group comparisons. We found 16–27 million SNPs segregating within taxa, corresponding to an average number of pairwise differences per site (nucleotide diversity) of 0.17–0.22% (Fig. 1c and Supplementary Fig. 7) and effective population sizes generally above 35,000 (except for aethiops, for which we estimated ~29,000). Site-frequency spectrums within taxa (Supplementary Fig. 4) generally agreed with neutral expectations, except for a general lack of low-frequency variants and an excess of high-frequency derived variants, most likely as a consequence of low power to call low-frequency variants and erroneous inference of the ancestral state, respectively.

Divergence was calculated as the average of pairwise differences across all comparisons of two groups. Two taxon site-frequency spectrums (Supplementary Fig. 9) generally showed fixed differences as well as shared and private variation, except for hilgery–cynosorus–pygerythrus, which showed few fixed differences and highly correlated allele frequencies.

To assess the relative contributions of geography and taxon label to explaining the genetic relatedness among vervets, we calculated principal components from autosomal SNPs using PCAdapt version 05/26/14 in mode fast68 setting K = 6, which gave the best fit to our data. Next, for each of the six principal components, we performed likelihood-ratio tests in R (function anova with option test = “LRT”) to test whether a linear model “PC – latitude + longitude + taxon label” gave a significantly better fit than a model using either only geography or only taxon (Supplementary Table 4). The outgroup branch of the vervet phylogeny was confirmed using genome-wide summaries of the average time to the most recent common ancestor among samples from different taxa (Supplementary Fig. 7 and Supplementary Note).

FST was calculated using the Weir–Cockerham estimator69 using vcftools70 for all autosomal SNPs. For each pairwise comparison, we summarized FST values in minor allele frequency (MAF) bins (Supplementary Fig. 8; the maximum across MAFs is shown in Fig. 1c).

SIVagm phylogenetic analyses. A 602-bp pol integrase fragment of SIVagm obtained as described previously9, was used for phylogenetic analyses of a large sample of SIVagm strains from the different subtaxa of vervet with different origin. pol nucleotide sequence alignments were obtained from the Los Alamos National Laboratory HIV Sequence Database (see URLs). Newly derived SIV sequences were aligned using MUSCLE71, and alignments were edited manually where necessary. Regions of ambiguous alignment and all gap-containing sites were excluded.

Phylogenetic trees were inferred from the nucleotide sequence alignments by the neighbor-joining method using the HKY85 model of nucleotide substitution72,73 implemented using PAUP* (ref. 73). The reliability of branching order was assessed by performing 1,000 bootstrap replicates, again using neighbor joining and the HKY85 model. Phylogenetic trees were also inferred by maximum likelihood using PAUP* with models inferred from the alignment using Modeltest74. The neighbor-joining tree topology was used as the starting tree in a heuristic search using TBR branch swapping.

Admixture analysis. The software Admixture 1.23 (ref. 22) was run on autosomal SNPs filtered for MAF >5%, converted to binary .bed format using GATK VariantToBinaryPed and LD pruned using the plink flag --indep 50 10 2 (Fig. 3a and Supplementary Fig. 11).

Cross-taxon gene flow across time. Missing genotypes in our autosomal SNP calls were imputed using Beagle 4 (ref. 75), version 03Oct15.284. java -jar beagle.jar gl = biallelic_pass_snps.vcf out = beagle_out.vcf ibd = false Samples were phased using shapeit.v2.r837.GLIBCv2.12.Linux66. shapeit --phase-input-vcf beagle_out.vcf-window 0.1 -o phased.tmp shapeit --output-vcf-input-haps phased.tmp -o phased.vcf For each geographic sample group of interest, we chose the individual within the highest coverage for further analysis. For pairs of individuals from different groups, we extracted the alleles that segregated within or between the two individuals and their phase as needed as input to MSMC. The number of informative sites between two segregating variants was determined for each pair of individuals separately from the all-sites VCF (the whole genome including non-variant sites) by counting the number of non-filtered sites for which both individuals had genotype calls.
Three runs of MSMC2 (see URLs) were produced for each pair, two inferring coalescent rates across time within each of the two samples

\[ \text{msmc2} -I \, 0,1 -o \, \text{within}_1 \, \text{input}_1 \, \text{chrom} \, 1 \, \text{output} \, \text{chrom} \, 29 \;
\]

\[ \text{msmc2} -I \, 2,3 -o \, \text{within}_2 \, \text{input}_1 \, \text{chrom} \, 1 \, \text{output} \, \text{chrom} \, 29 \;
\]

and one run for inferring coalescent rates across time between the two samples.

\[ \text{msmc2} -P \, 0,0,1,1 -o \, \text{between} \, \text{input}_1 \, \text{chrom} \, 1 \, \text{output} \, \text{chrom} \, 29 \;
\]

The outputs of these runs were combined by interpolating the midpoint of each time interval in the former two on the midpoints in the latter run. Cross-coalescent rate was calculated as \(2 \times \text{between} / (\text{within}_1 + \text{within}_2)\) (Fig. 2bc, Supplementary Figs. 12 and 13, and Supplementary Note). Evolutionary time was scaled to years using a mutation rate of \(1.5 \times 10^{-8}\) mutations per generation and a generation time of 8.5 years1. Each MSMC2 analysis was rerun 29 times leaving one chromosomes out at a time, and block-jackknifing variance was calculated.

**D statistic.** The \(D\) statistic was calculated from autosomal SNPs using Admixtools 3.0 (ref. 24), treating samples from each country as a population and performing all tests that were consistent with the country UPGMA tree shown in Figure 2d. Macaque was used as an outgroup, and the analysis was restricted to sites where the macaque allele could be inferred. Because of limitations in Admixtools, the analysis was restricted to vervet chromosomes 1–24. Block-jackknifing was performed with Admixtools standard settings.

**Diversifying selection scan.** Autosomal biallelic PASS SNP genotypes were converted to XP-CLR input format. For each comparison of two groups, we excluded SNPs if they were not segregating within or between the groups or if they had more than 20% missing genotypes across the two groups. The genetic map from Huang et al.7 was interpolated to our SNP positions to obtain genetic distance in Morgans. A handful of extremely flat regions in the genetic map led to numeric errors in XP-CLR. The problematic markers were removed, and map distance was instead interpolated from the adjacent markers to the left and right.

XP-CLR was run on all 30 possible comparisons of the six African taxa (for each comparison, using each taxon once as the objective and once as the reference population). The parameters supplied to XP-CLR were \(-w\ 0.001 500 1000 -p\ 0.0\), meaning that a set of grid points as the putative selected allele positions are placed along the chromosome with a spacing of 1 kb, the sliding window size was 0.1 cm around the grid points, and, if the number of SNPs within a grid was beyond 500, some SNPs were randomly dropped to control for SNP number. Alleles were assumed unphased (-p 0), and SNPs in high LD were not down-weighted (final 0).

To find loci that were repeatedly under diversifying selection across several group comparisons, we calculated for each grid point the root mean-square selection score across all 30 comparisons (Fig. 5a). To test whether these scores capture biological signal, we confirmed that scores were significantly higher in genic (introns + exons) than in intergenic regions (one-sided Mann–Whitney U test, \(P < 10^{-300}\); Supplementary Fig. 21). Because the Mann–Whitney U test assumes independence of scores, a condition that was not totally met in our analysis owing to LD, we also calculated the average of the mean selection score across each gene and compared the resulting value to a background distribution. The background distribution was obtained by first concatenating all chromosomes in a circle and randomly shifting (rotating) the scores against their genomic positions, then calculating mean gene scores from these rotated data. We again found that genic scores were significantly larger (\(P < 10^{-5}\)).

**Gene set enrichment analysis.** \(z\)-score-transformed selection scores across genes (exons and introns) were used for gene enrichment analysis. Gene locations were extracted from NCBI *Chlorocebus sabaeus* Annotation Release 100 (see URLs). To test for enrichment in GO terms, we first used the R package TopGO with the weight01 algorithm, which allows accounting for the hierarchical structure (and thus overlap) of GO terms when testing significance and thereby implicitly corrects for multiple testing. GO annotations were obtained from the R package org.Hs.eg.db (Bioconductor 3.2). We restricted the analysis to 5,777 GO terms with more than ten genes in our annotation.

Note that our gene scores were not biased by gene length because we were calculating the average score across genes rather than taking the maximum score. However, enrichment results were qualitatively similar if the maximum was taken. Results were also similar if only exons were used (rather than exons + introns). We also note that the most significantly enriched categories contained many genes (Supplementary Data 4) and did not show strong clustering in particular genomic regions.

**HIV-1 human interaction categories.** The NCBI HIV-1 Human Interaction Database50 was downloaded (see URLs). We only kept categories from the database that had ten or more genes in our annotation. We implemented the sunstat statistic13 to compare observed and expected gene-averaged selection scores in HIV-1 human interaction categories to random sets of known genes.

**Gene expression analysis.** We conducted weighted gene coexpression network analysis (WGCNA) on gene expression data from vervets and macaques assayed at different time points before and after infection with SIVagm and SIVmac, respectively, using the WGCNA R package as previously described57. We used as a starting point the list of genes with differentially expressed transcripts in CD4+ cells before and after SIV infection in either vervet or rhesus35. Correlation coefficients were constructed between expression levels of genes, and a connectivity measure (topological overlap, TO) was calculated for each gene by summing the connection strength with other genes. Genes were then clustered on the basis of their TO, and groups of coexpressed genes (modules) were identified. Each module was assigned a color, and the first principal component (eigengene) of a module was extracted from the module and considered to be representative of the gene expression profiles in a module. We identified 36 modules (Supplementary Figs. 34 and 35, and Supplementary Data 6), 33 of which contained more than ten genes in our annotation and were used for enrichment testing.

**Data availability.** All genomic data for the vervets sequenced in this study are available through the Sequence Read Archive (SRA) public repository under BioProject numbers PRJNA168521, PRJNA168472, PRJNA168520, PRJNA168527 and PRJNA168522. Variant call format (VCF) files are available from the European Variation Archive (EVA) under accessions PRJEB22988 (imputed phased whole-genome SNP variants) and PRJEB22989 (effect-annotated whole-genome SNPs, indels and macaque fixed differences). A Life Sciences Reporting Summary is available.
Publisher Correction: Ancient hybridization and strong adaptation to viruses across African vervet monkey populations

Hannes Svardal, Anna J Jasinska, Cristian Apetrei, Giovanni Coppola, Yu Huang, Christopher A Schmitt, Beatrice Jacquelin, Vasily Ramensky, Michaela Müller-Trutwin, Martin Antonio, George Weinstock, J Paul Grobler, Ken Dewar, Richard K Wilson, Trudy R Turner, Wesley C Warren, Nelson B Freimer and Magnus Nordborg

Correction to: Nature Genetics https://doi.org/10.1038/ng.3980, published online 30 October 2017.

In the version of this article published, in the Online Methods eight citations to supplementary material refer to the wrong supplementary items.

Published text: We therefore repeated the entire selection analysis (Figs. 3–5 and Supplementary Figs. 15–31) using the updated call set.
Correct supplementary citation: Supplementary Figs. 17–33.

Published text: We found 16–27 million SNPs segregating within taxa, corresponding to an average number of pairwise differences per site (nucleotide diversity) of 0.17–0.22% (Fig. 1c and Supplementary Fig. 7) and effective population sizes generally above 35,000 (except for aethiops, for which we estimated ~29,000).
Correct supplementary citation: Supplementary Fig. 10

Published text: Two taxon site-frequency spectra (Supplementary Fig. 9) generally showed fixed differences as well as shared and private variation, except for hilgerty–cynosorus–pygerythrus, which showed few fixed differences and highly correlated allele frequencies.
Correct supplementary citation: Supplementary Fig. 8

Published text: Next, for each of the six principal components we performed likelihood-ratio tests in R (function anova with option test = 'LRT') to test whether a linear model “PC ~ latitude + longitude + taxon label” gave a significantly better fit than a model using either only geography or only taxon (Supplementary Table 4).
Correct supplementary citation: Supplementary Table 2

Published text: For each pairwise comparison, we summarized $F_{st}$ values in minor allele frequency (MAF) bins (Supplementary Fig. 8; the maximum across MAFs is shown in Fig. 1c).
Correct supplementary citation: Supplementary Fig. 11

Published text: The software Admixture 1.23 (ref. 22) was run on autosomal SNPs filtered for MAF >5%, converted to binary .bed format using GATK VariantToBinaryPed and LD pruned using the plink flag --indep 50 10 2 (Fig. 3a and Supplementary Fig. 11).
Correct supplementary citation: Supplementary Fig. 13

Published text: Cross-coalescent rate was calculated as $2 \times \text{between}/(\text{within}_1 + \text{within}_2)$ (Fig. 2b,c, Supplementary Figs. 12 and 13, and Supplementary Note).
Correct supplementary citation: Supplementary Figs. 15 and 16, and Supplementary Note

Published text: To test whether these scores capture biological signal, we confirmed that scores were significantly higher in genic (introns + exons) than in intergenic regions (one-sided Mann–Whitney $U$ test, $P < 10^{-300}$; Supplementary Fig. 21).
Correct supplementary citation: Supplementary Fig. 23

Published online: 16 October 2018
https://doi.org/10.1038/s41588-018-0124-x
Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, 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.
   - This paper describes a population sample which was not generated for experimental purposes.

2. **Data exclusions**
   - Describe any data exclusions.
   - Not relevant in this sense. There was of course massive amount of SNP filtering to eliminate unreliable data.

3. **Replication**
   - Describe whether the experimental findings were reliably reproduced.
   - Not relevant.

4. **Randomization**
   - Describe how samples/organisms/participants were allocated into experimental groups.
   - Not relevant.

5. **Blinding**
   - Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   - Not relevant.

   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 the Methods section if additional space is needed).

   
   | n/a | Confirmed |
   |-----|-----------|
   | ✗   | ☑         |
   | ☑   |            |
   | ☑   |            |
   | ☑   |            |
   | ☑   |            |
   | ☑   |            |
   | ☑   |            |
   | ☑   |            |
   | ☑   |            |
   | ☑   |            |

   - 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 summary of the descriptive 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.

### Software

Policy information about availability of computer code

7. **Software**
   - Describe the software used to analyze the data in this study.
   - As described in the paper, only publicly available software was used.
For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The Nature Methods guidance for providing algorithms and software for publication may be useful for any submission.

### Materials and reagents

**Policy information about availability of materials**

8. **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 restriction on any data.

9. **Antibodies**
   
   Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).
   
   Not relevant.

10. **Eukaryotic cell lines**
    
    a. State the source of each eukaryotic cell line used.
    
    Not relevant.
    
    b. Describe the method of cell line authentication used.
    
    Not relevant.
    
    c. Report whether the cell lines were tested for mycoplasma contamination.
    
    Not relevant.
    
    d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.
    
    Not relevant.

### 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.
    
    This paper is based on wild animals.

**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.
    
    Not relevant.