Methods.

Assembly. To sequence and assemble the vervet reference genome we therefore utilized genomic DNA from a male vervet within the VRC. Our overall strategy differed from that used in previous NHP assemblies, in that we leveraged a physical map consisting of genome-wide clone end sequences from a vervet BAC library with 201,000 BACs (CHORI-252), and the use of recently developed methods for deriving assembled genomes from multiple sequencing platforms (Bradnam et al. 2013); for this purpose we constructed and then merged two independent assemblies (Yao et al. 2012).

To create a chromosomal version of this assembly the assembled vervet genome was aligned against the rhesus macaque genome rheMac7 (Zimin et al. 2014) and the human genome (GRCh38 at UCSC Genome Browser (http://genome.ucsc.edu/) utilizing BLASTZ to align and score non-repetitive vervet regions against repeat-masked rhesus macaque and human sequence, respectively. Alignment chains differentiated between orthologous and paralogous alignments and only "reciprocal best" alignments were retained in the alignment set. The assembled vervet genome was broken into 1kb segments and then aligned against the rhesus macaque and human genomes using BLAT (Kent 2002) to identify uniquely aligning segments of the vervet genome to aid in identifying breakpoints. A vervet genetic linkage map was evaluated but insufficient marker numbers (<500) prevented the collation of most scaffolds. To further ensure accuracy and improve long range scaffold connectivity we aligned all BAC end sequences, using a best hit versus all other hits threshold (best hit >20 BLAT score above all else), to evaluate concordance between paired sequences of know length. When sufficient discordant evidence presented within scaffolds breaks were made manually and all assembly scaffolds and contigs were subsequently renumbered.

Repeats. To compute the proportion of the genome that such elements represent, the size of the genome was determined, excluding undefined nucleotide characters. The vervet RepeatMasker report was used to extract all repeat coordinates and grouped elements corresponding to RepeatMasker-assigned families and classes; calculations of the amount of full-length elements allowed 10-nucleotide non-matching flanks. Artificial entries, such as elements with overlapping coordinates, false annotations, or contaminations from taxonomically distant groups (e.g., crocodilian L2 elements), were removed. The fragmentation report of RepeatMasker (rank) was used to merge interrupted elements. For long terminal repeat elements (LTRs), full-length retroviral elements were distinguished from singular LTRs corresponding to the giri annotation.

These sequences were extracted by calculating a 2-way genome alignment of the soft masked (http://www.repeatmasker.org/RMDownload.html) genomes of vervet (settings: -species "Macaca") and rhesus macaque (rheMac7, http://hgdownload.soe.ucsc.edu/goldenPath/rheMac3/bigZips/rheMac3.fa.masked.gz) using LASTZ (Harris 2007) to generate and score alignments. Subsequent chaining and netting of the obtained alignments used various external scripts (http://hgdownload.cse.ucsc.edu/admin/jksrc.zip), producing a 2-way alignment in axt-format. From this alignment, "Unique" vervet sequences (in the range of 30-2,500 bp) were selected with a corresponding gap in rhesus macaque, allowing a maximum of 10 bp of missing flanks.

Structural variant detection.

Deletions. For deletions variants generated with LUMPY (Layer et al. 2014) the breakpoints for each deletion were defined as a pair of probability distributions derived from split-pair and split-read alignment discordance. Overlapping breakpoints were clustered and when probabilities were supported by both independent read-type sources a predicted SV is returned. When determining which breakpoints delineating a deletion that were shared among samples we required the multi-read type support evidence to be contained within a given breakpoint and only returned shared deletions that matched exactly the deletion sequence coordinate boundaries defined in each vervet at a minimum of any three independent vervet samples. Only these LUMPY deletions denoting a CNVnator read depth of <1.5 were retained for further analysis. The LUMPY deletions for each vervet were filtered to extract deletions ranging from 500bp to 1Mb in size and then genotyped using CNVnator for each of the LUMPY calls. A gene was considered altered by a deletion event if the collective breakpoints of each deletion intersected with any vervet exon sequence.

Segmental duplications. We evaluated the overall sequencing performance of the raw reads and we demarcated the regions of the reads that displayed the best qualities. The estimated levels of duplicated reads raw reads were mapped using BWA (Li and Durbin 2010) to the ChlSab1.1 reference and PCR duplicates were removed using PicardTools (http://broadinstitute.github.io/picard/). Subsequently, non-duplicated Illumina reads were clipped into two 36bp length reads (covering positions 10-45 and 46-81, and thus avoiding the inclusion of the lower-quality ends of the reads) to improve the mapping efficiency in the boundaries of repetitive regions. Chromosomes were partitioned into 36bp kmers (with adjacent kmers overlapping 5bp) and these kmers were mapped against the ChlSab1.1 assembly using mrsFast (Hach et al. 2014) to account for multi-mapping. Over-represented K-mers, defined as those K-mers with more than 20 mappings into the assembly, were additionally masked (Supplemental Fig. S9).

To identify SDs (fragments > 5Kbp of duplicated sequences) an in-house python script was used. Sample-specific cut-offs are defined based on the internal copy number distribution. For each sample, the mean and standard deviation copy number of those 1Kb windows on chromosomes CAE1–29 (excluding those windows with copy number in the highest 1% of the distribution) are used to normalize the copy number values across all windows. For a duplication to be called the copy number must exceed the mean copy number by three standard deviations in at least five consecutive windows, allowing one of the internal windows to exceed the mean copy number by two standard deviations. Also, duplications are required to expand a minimum of 10Kb. Regions with copy numbers above 100 in any sample were removed from the analysis. Shared SDs were called by intersecting with bedtools the bed files containing the coordinates of the duplications detected in each sample; shared SDs correspond to those coordinates that were called as duplicated in all six vervets.

*Gene family analysis***.** In order to identify rapidly evolving gene families along the vervet lineage we obtained peptides from human, chimpanzee, gibbon, rhesus macaque, marmoset, tarsier, bushbaby, rat, and mouse from ENSEMBL 76 (Flicek et al. 2014) and peptides from vervet and colugo from NCBI (http://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/). To ensure that each gene was counted only once, we used only the longest isoform of each protein in each species and all other isoforms were filtered out. We then performed an all versus all BLAST (Altschul et al. 1997) search on these filtered sequences. The resulting e-values from the search were used as the main clustering criterion for the MCL program to group peptides into gene families (Enright et al. 2002). This procedure resulted in 15,386 clusters and of these, all single-peptide clusters were filtered out, leaving 11,996 gene families. We constructed an ultrametric tree (Fig. 1) manually for these eleven species based on divergence times obtained from TimeTree (Hedges et al. 2006) to estimate gene family change specific to each species lineage.

With the gene family data and ultrametric phylogeny as input, we estimated gene gain and loss rates (λ) with CAFE v3.0 (Han et al. 2013). 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 0.05, which implies that 5% of gene families have observed counts that are not equal to their true counts. After correcting for this error rate, we find $\lambda = 0.002$. These values for ε and λ are on par with those previously reported for mammalian datasets (Supplemental Table S21) (Han et al. 2013). Using the estimated λ value, CAFE infers ancestral gene counts and calculates p-values across the tree for each family to assess the significance of any gene family changes along a given branch. Those branches with low p-values are said to be rapidly evolving.

References.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**: 3389-3402.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R et al. 2013a. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* **2**: 10.

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R et al. 2013b. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* **2**: 10.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**: 1575-1584.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. 2014. Ensembl 2014. *Nucleic acids research* **42**: D749-755.

Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2014a. mrsFAST-Ultra: A compact, SNP-aware mapper for high performance sequencing applications. *Nucleic Acids Res* **42**: W494-W500.

Hach F, Sarrafi I, Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC. 2014b. mrsFAST-Ultra: a compact, SNP-aware mapper for high performance sequencing applications. *Nucleic acids research* doi:10.1093/nar/gku370.

Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular biology and evolution* **30**: 1987-1997.

Harris RS. 2007. Improved pairwise alignment of genomic DNA. The Pennsylvania University.

Hedges SB, Dudley J, Kumar S. 2006a. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**: 2971-2972.

Hedges SB, Dudley J, Kumar S. 2006b. TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics* **22**: 2971-2972.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome research* **12**: 656-664.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol* **15**: R84.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Li H, Durbin RR. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Research* **13**: 103-107.

Yao G, Ye L, Gao H, Minx P, Warren WC, Weinstock GM. 2012. Graph accordance of next-generation sequence assemblies. *Bioinformatics* **28**: 13-16.

Zimin AV, Cornish AS, Maudhoo MD, Gibbs RM, Zhang X, Pandey S, Meehan DT, Wipfler K, Bosinger SE, Johnson ZP et al. 2014. A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biology direct* **9**: 20.