Supporting Information

Malfavon-Borja et al. 10.1073/pnas.1216542110

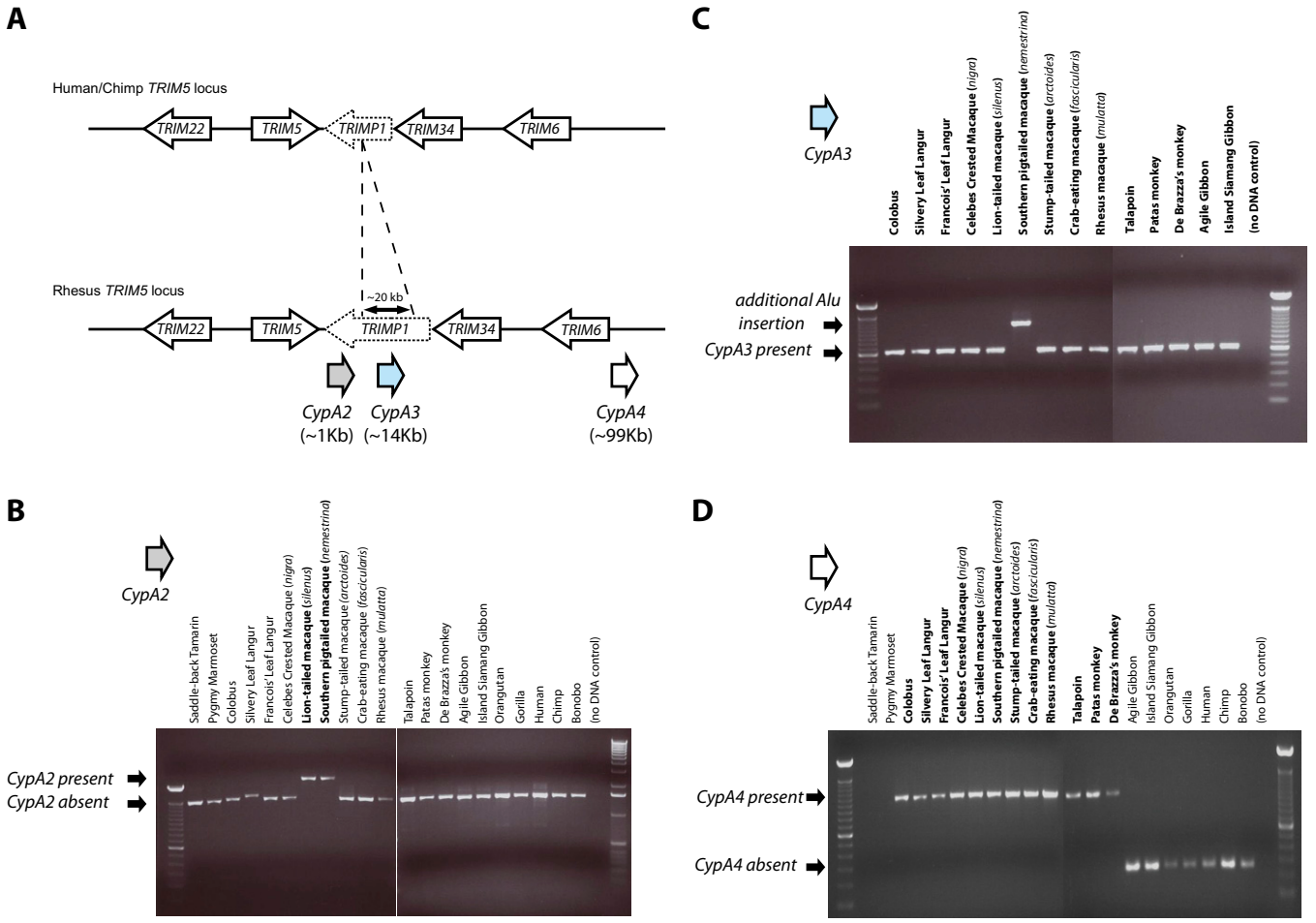

Fig. S1. Summary of *CypA* retrogenes proximal to *TRIM5*. Comparison of the human, chimpanzee, and rhesus macaque *TRIM5* locus. (*A*) Relative locations of *CypA* are illustrated below the representation of the *TRIM5* locus. *CypA2* (light gray arrow), *CypA3* (light blue arrow), and *CypA4* (white arrow) are ~1 kb, ~14 kb, and ~99 kb downstream of *TRIM5*, respectively. The rhesus macaque *TRIMP1* region contains an additional ~20 kb not present in the human and chimpanzee *TRIMP1*. (*B–D*) Determining the presence or absence of *CypA* retrogenes was done using the following primates: saddle-back tamarin (*Saguinus fuscicollis nigrifrons*), pygmy marmoset (*Callithrix pygmaea*), colobus (*Colobus guereza*), silvery leaf langur (*Trachypithecus cristatus*), Francois' leaf langur (*Trachypithecus francoisi*), celebes crested macaque (*Macaca nigra*), lion-tailed macaque (*Macaca silenus*), Southern pig-tailed macaque (*Macaca nemestrina*), stump-tailed macaque (*Macaca arctoides*), crab-eating macaque (*Macaca fasicularis*), rhesus macaque (*Macaca mulatta*), talapoin (*Miopithecus talapoin*), patas monkey (*Erythrocebus patas*), De Brazza's monkey (*Cercopithecus neglectus*), agile gibbon (*Hylobates agilis albibarbis*), Island Siamang gibbon (*Hylobates syndactylus*), orangutan (*Pongo pygmaeus*), gorilla (*Gorilla gorilla*), human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), and bonobo (*Pan paniscus*). Amplifying the site surrounding *CypA2* was expected to generate an ~2-kb band when the retrogene was present and an ~1.5-kb band when the retrogene was absent. Amplifying the site surrounding *CypA4* was expected to generate an ~250-bp band when the retrogene was present and an ~750-bp band when the retrogene was absent. The gel displaying results for *CypA3* only shows primates for which we were able to amplify products containing the retrogene. The primates that were not included were not predicted to encode a TRIMP1 that contained *CypA3*.

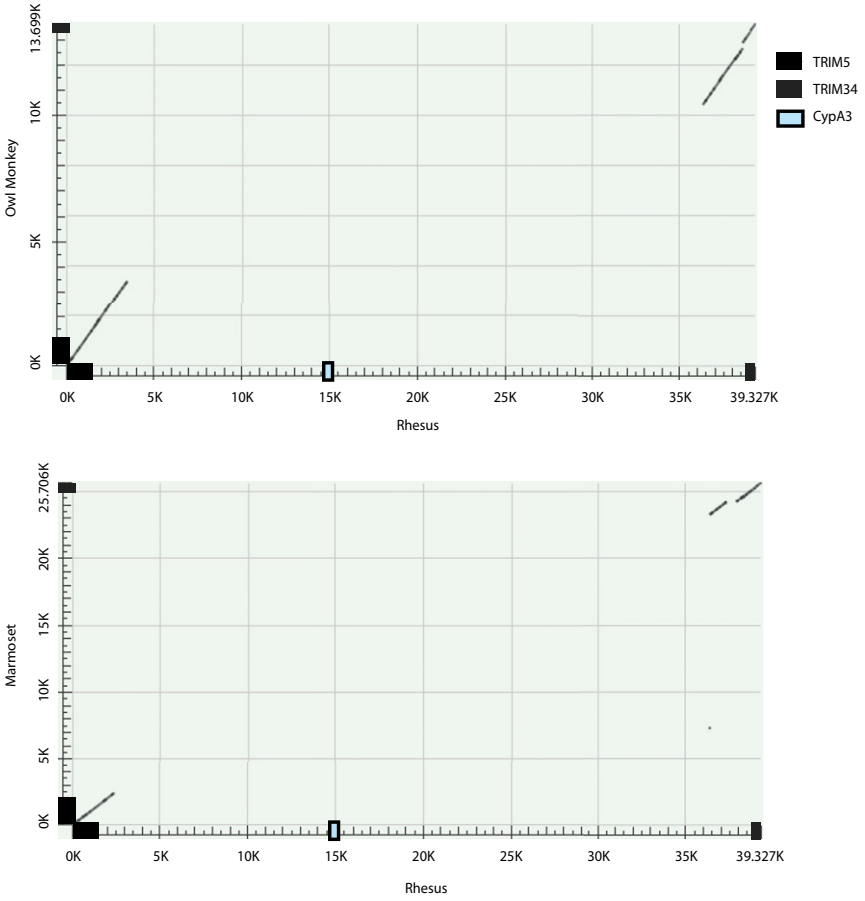

Fig. S2. TRIMP1 dot plot. To evaluate the presence or absence of *CypA3* in New World monkeys *in silico*, we mapped *TRIMP1* between the common marmoset (*Callithrix jacchus*, accession no. AC148555), Nancy Ma's night monkey (*Aotus nancymaae*, accession no. AC183999), and rhesus macaque (University of California, Santa Cruz Genome Browser) at ~25 kb, ~13 kb, and ~40 kb in length, respectively. Pairwise alignments were prepared using the National Center for Biotechnology Information's bl2seq. Rhesus macaque *TRIMP1* was used as the query sequence, whereas marmoset *TRIMP1* and owl monkey *TRIMP1* were used as the subject sequence. We set the filter to mask species-specific repeats for the human species. All other general parameters were kept at default. Output was visualized as a dot matrix plot with rhesus macaque as the *x* axis and either marmoset or owl monkey as the *y* axis. We annotated the location of *TRIM5* (black box), *TRIM34* (dark gray box), and *CypA3* retrogene (light blue box).

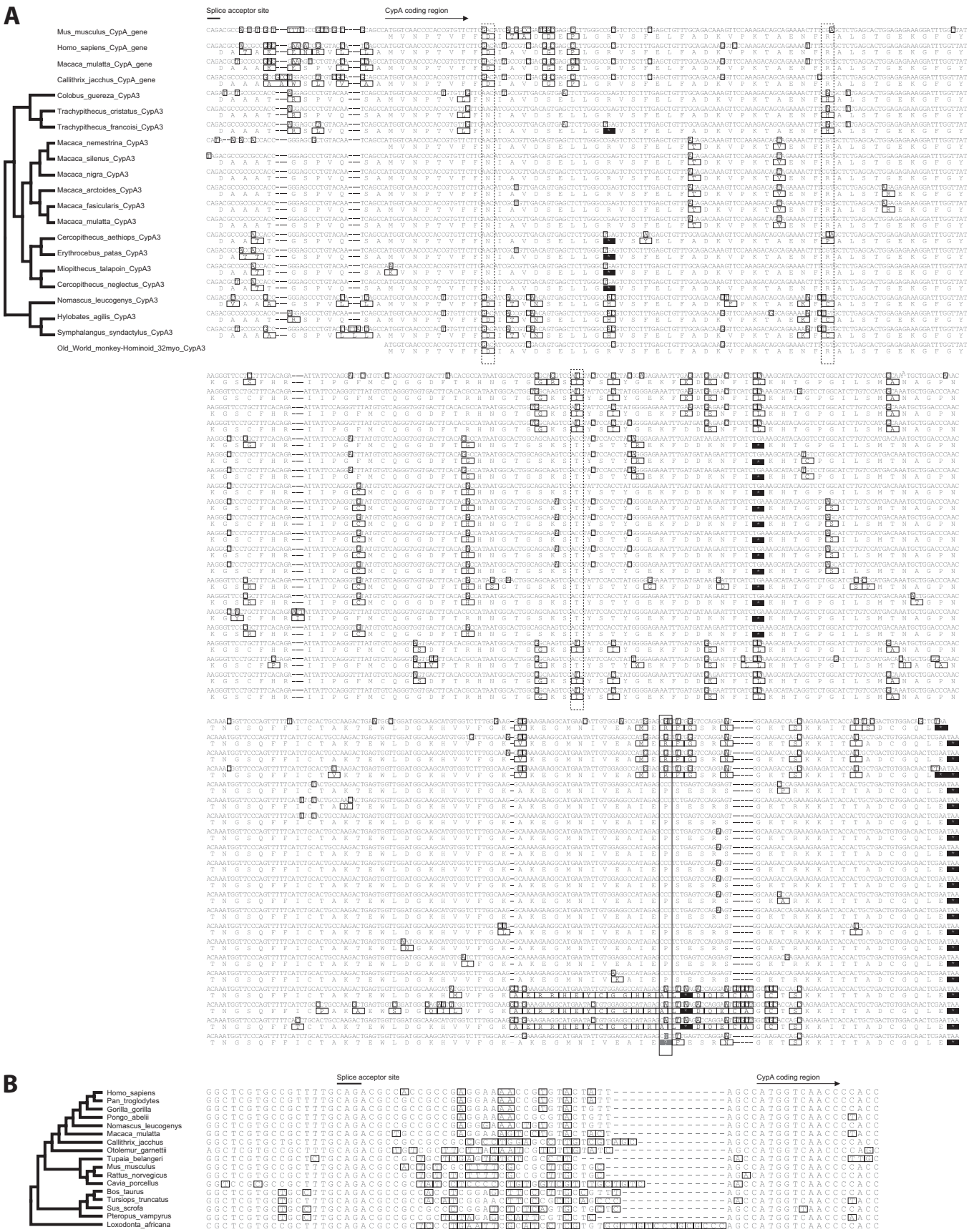

Fig. S3. CypA3 and parental CypA alignments. (*A*) We undertook the reconstruction of a version of *CypA3* that is representative of the Old World monkey/hominoid common ancestor. Reconstruction was carried out by parsimony criteria and supported by a maximum likelihood reconstruction generated by

Legend continued on following page

codeml (Phylogentic Analysis by Maximum Likelihood package). Visualization of the reconstructed *CypA3* sequence, modern-day CypA3, and parental *CypA* genes was prepared using Geneious 5.3.6 (Biomatters). Human, rhesus macaque, and marmoset *CypA* genes were used as reference sequences (outgroup). The reconstructed version of *CypA3* is listed at the bottom of the alignment. We identified several instances in which the reference sequences served to discern the ancestral codon/residue present between Old World monkey and gibbon *CypA3* sequences (dotted-line boxes). We identified a single instance in which the ancestral codon/residue could not be resolved, residue 144 (gray box with "?" symbol). Synonymous and nonsynonymous changes were highlighted by Geneious (5.3.6) using black-outlined boxes. Nonsense mutations or stop codons have been marked as black-filled boxes containing an asterisk (*) symbol. Reconstruction of the ancestral sequence was limited to the coding region; however, we included the upstream sequence of the *CypA* genes and retrogenes to the alignment to demonstrate the presence of the cryptic splice site used in the formation of *TRIMCyp* gene fusions. We used the mouse parental *CypA* gene to represent nonprimate animals, but we used the upstream region from a spread of nonprimate mammalian species. (*B*) To demonstrate the conservation of the splice acceptor site upstream of the *CypA* coding sequence, we collected the parental *CypA* gene sequence and a portion of the corresponding upstream region from publically available mammalian genomes. These were aligned using Clustal W2 (1), and the locations of the conserved splice acceptor site and the start of the *CypA* coding sequence have been marked.

1. Larkin MA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947–2948.

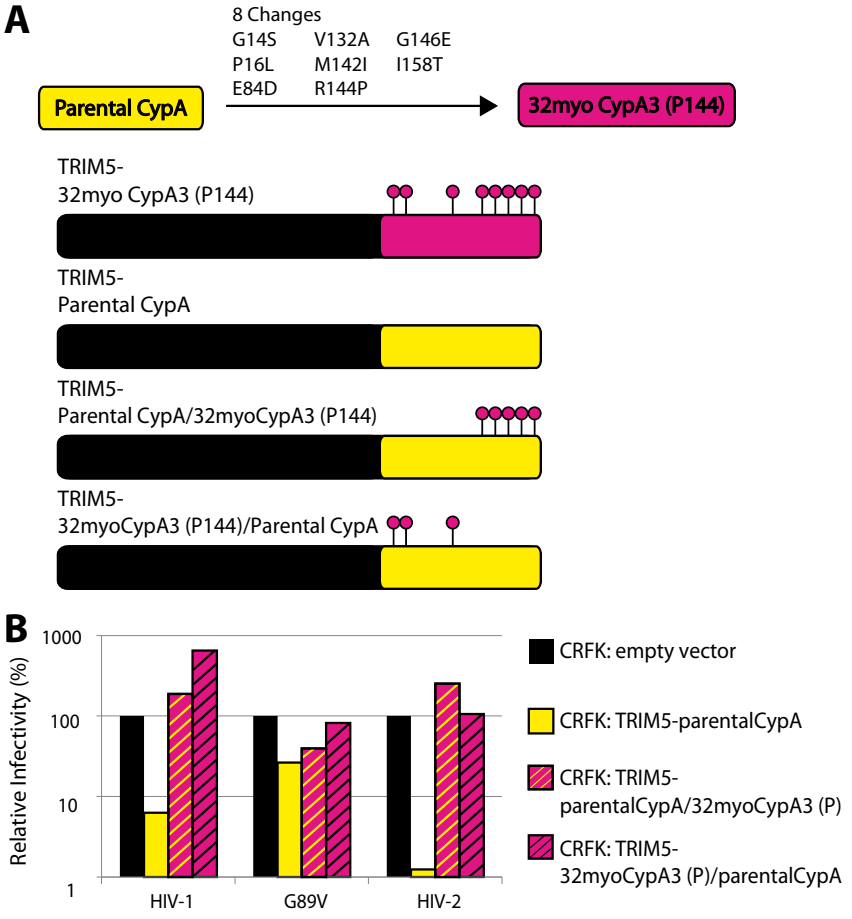

Fig. S4. Evaluation of 32myoCypA3 unique residues by the formation of chimeric *TRIMCyp* gene fusions. (*A*) We illustrated the trajectory of parental CypA (yellow) to 32myoCypA3 (P144) (magenta). The eight residues acquired in the evolution of 32myoCypA3 are highlighted by magenta-filled pegs. (*B*) Stable Crandell-Rees feline kidney (CRFK) cell lines encoding owl monkey *TRIM5-parentalCypA* (positive control), CRFK cell lines with an empty vector (negative control), parental CypA/32myoCypA3 (P144) chimera (magenta with yellow diagonal stripe), and 32myoCypA3 (P144)/parental CypA chimera (magenta with black diagonal stripe) were assayed against lentiviruses: HIV-1 (LAI), HIV-1 G89V, and HIV-2 (ROD9). Viruses are listed along the *x* axis. The *y* axis reflects virus infectivity, determined by the percentage of cells infected with GFP-expressing virus, normalized to 100% for infections of CRFK cells with empty vector. The virus inoculums were standardized to give the absolute percentage of GFP between 15% and 30%.