Retrocopying expands the functional repertoire of APOBEC3 antiviral proteins in primates

Lei Yang[1], Michael Emerman[2], Harmit S. Malik[3,4*], & Richard N. McLaughlin Jr.[1,3*],

[1]Pacific Northwest Research Institute, Seattle, WA, USA; [2]Division of Human Biology, [3]Division of Basic Sciences, [4]Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Running head: Expansion of APOBEC3 genes in primates

*Correspondence to:

Richard N. McLaughlin Jr., 720 Broadway, Seattle WA 98122 USA Tel: +001 (206) 568-1457; rmclaughlin@pnri.org

Harmit S. Malik, 1100 Fairview Avenue N A2-205, Seattle WA 98109 USA Tel: +001 (206) 667-5204; hsmalik@fhcrc.org

Abstract

Host-virus arms races are inherently asymmetric; viruses evolve much more rapidly than host genomes. Thus, there is high interest in discovering mechanisms by which host genomes keep pace with rapidly evolving viruses. One family of restriction factors, the *APOBEC3 (A3)* cytidine deaminases, has undergone positive selection and expansion via segmental gene duplication and recombination. Here, we show that new copies of *A3* genes have also been created in primates by reverse transcriptase-encoding elements like LINE-1 or endogenous retroviruses via a process termed retrocopying. First, we discovered that all simian primate genomes retain the remnants of an ancient *A3* retrocopy: *A3I*. Furthermore, we found that some New World monkeys encode up to ten additional *APOBEC3G* (*A3G*) retrocopies. Some of these *A3G* retrocopies are transcribed in a variety of tissues and able to restrict retroviruses. Our findings suggest that host genomes co-opt retroelement activity in the germline to create new host restriction factors as another means to keep pace with the rapid evolution of viruses. *(163)*

Introduction

Host genomes have an ancient history of coevolution with selfish genetic elements. One type of these selfish elements, called endogenous retroelements, created a substantial fraction of most animal genomes (Canapa et al., 2015; de Koning et al., 2011; Lander et al., 2001; Smit et al., 2015; Sotero-Caio et al., 2017). Endogenous retroelements such as endogenous retroviruses (ERVs) and Long Interspersed Element-1s (LINE-1s) reside in host genomes where they 'copy-and-paste' themselves via the action of their reverse transcriptase. These retroelements can negatively impact host fitness by disrupting genes or regulatory regions, and by increasing the likelihood of ectopic recombination (Boissinot et al., 2001; Hancks and Kazazian, 2012; Kaer and Speek, 2013; Petrov, 2003; Song and Boissinot, 2007).

In addition to acting on their own RNA to ensure duplication, the reverse transcription/ integration functions encoded by LINE-1s and ERVs also occasionally act on host mRNAs. This 'off-target' activity, termed retrocopying, entails the duplication of a host gene via the reverse transcription and integration of an mRNA. These 'retrocopies' are intronless and removed from the chromosomal location of the parental intron-containing gene. Previous studies estimated that 3,700-18,000 retrocopies are present in the human genome (Casola and Betrán, 2017; Navarro and Galante, 2015; Potrzebowski et al., 2008).

Two features distinguish retrocopies from other types of gene duplications. First, 'DNA-based' mechanisms of duplication (*e.g.,* segmental gene duplications) result in a new copy of the gene including its promoter and distal regulatory elements. In contrast, retrocopying typically duplicates only the exons, leading to the moniker 'processed pseudogene'. Thus, transcription of a new retrocopy depends on the genomic neighborhood into which it integrates (Carelli et al., 2016). Second, retrocopying relies on the machinery encoded by endogenous retroelements like LINE-1, which are highly active in germline and early embryo (Friedli et al., 2014; Garcia-Perez et al., 2007; Klawitter et al., 2016; Muotri, 2016; Wissing et al., 2012). Therefore, unlike DNA-based duplications, retrocopying is almost exclusively limited to RNAs expressed in germline or early embryonic tissues. It follows that the level of germline expression of host mRNAs should be highly correlated to their probability of generating retrocopies. For example, ribosomal proteins that are highly expressed in germline tissues represent the most abundant class of processed pseudogenes in the human genome (Balasubramanian et al., 2009). While germline expression of host mRNAs predicates the generation of retrocopies, the vast majority of these retrocopies show characteristic signatures of pseudogenization (Casola and Betrán, 2017; Navarro and Galante, 2015; Potrzebowski et al., 2008).

While most retrocopies do not increase the genic capacity of the host due to inactivating mutations, a subset of retrocopies escaped mutational abrasion, presumably because they provide a selective advantage to the host. Indeed, evidence of functional retention in retrocopied sequences has been found in diverse organisms and includes functions such as novel subcellular localization of proteins (Rosso et al., 2008), neurotransmitter metabolism (Burki and Kaessmann, 2004), courtship (Wang et al., 2002), fertility (Kalamegham et al., 2006), and pathogen restriction (Malfavon-Borja et al., 2013; Sayah et al., 2004). Such functional retention may be particularly beneficial in the case of host defense genes, whose functional diversification is necessary for host genomes to keep pace with pathogens. For example, retrocopying of the *CypA* gene between coding exons of the *TRIM5* gene has created novel *TRIMCyp* fusion genes that can potently restrict retroviruses including HIV-1 (Malfavon-Borja et al., 2013; Newman et al., 2008; Nisole et al., 2004; Sayah et al., 2004; Virgen et al., 2008; Wilson et al., 2008). In a remarkable case of convergent evolution, retrocopying has created *TRIMCyp* fusion genes multiple times during primate evolution, further expanding and diversifying the *TRIM* gene family for retroviral defense (Brennan et al., 2008; Virgen et al., 2008). In other examples, mobile element and viral genes themselves have been retrocopied and domesticated for various functions including antiviral defense (Best et al., 1996; Fujino et al., 2014; Malik and Henikoff, 2005; McLaughlin et al., 2014; Nishigaki et al., 2013; Yan et al., 2009).

Here, we investigated whether retrocopying may have similarly diversified another family of host defense genes: the *APOBEC3* (*A3*) cytidine deaminases. Although the common ancestor of placental mammals likely encoded three *A3* genes (Münk et al., 2012), this locus has recurrently expanded and contracted throughout mammalian evolution (Ito et al., 2020), including a dramatic expansion to seven paralogous genes in catarrhine primates (Old World monkeys and hominoids) followed by recurrent positive selection of this expanded gene set (Bulliard et al., 2009; Compton et al., 2012; Duggal et al., 2011; Henry et al., 2012; McLaughlin et al., 2016;

OhAinle et al., 2006). We found that ancient and recent retrocopying has further diversified the already expansive *A3* gene repertoire in primates, adding as many as ten new *A3*s outside the well-studied '*A3* locus'. Our work uncovered an ancient *A3* born via retrocopying in the common ancestor of simian primates and a dramatic, ongoing history of *A3* retrocopying in New World monkeys (NWMs). Many of these NWM-specific *A3* retrocopies are expressed and some retain the capability to restrict retroviruses. Thus, retrocopied *A3*s have continually expanded host defense repertoires in primate genomes.

Results

A3I: an ancient *A3* retrocopy in simian primates

The *A3* locus of human, other hominoids, and Old World monkeys comprises seven clustered genes, *A3A-A3H* (Jarmuz et al., 2002; OhAinle et al., 2006; Silvas and Schiffer, 2019) (Figure 1A). We undertook an analysis to search for any variation in this gene structure in other primate genomes. We used BLASTn with each of the seven human *A3* nucleotide sequence to query all sequenced primate genomes on NCBI. As expected, each genome contained a series of proximal hits, presumably comprising the *A3* locus. However, in every simian primate genome examined, we also found exactly one shared syntenic region, distinct from the *A3* locus, with high similarity to *A3G*. This sequence match spanned the exonic sequences of *A3G* in a single contiguous region of around 1,100 bp (Figure 1B). Based on the absence of introns, we concluded that this sequence represents a retrocopy of an *A3* gene. We found no evidence of the syntenic copy of this *A3* retrocopy in the genomes of prosimians including the tarsier, bushbaby, and mouse lemur. We, therefore, conclude that this retrocopy was born in the common ancestor of simian primates, and hence propose the name *A3I* for this retrocopy, extending the nomenclature scheme for other human *A3* genes.

To understand the relationship between *A3I* and other *A3* genes, we created a maximum likelihood tree using the nucleotide sequences from simian primate *A3B*, *A3G*, and *A3I* genes. *A3I* could be aligned to primate *A3G*s and *A3B*s, since *A3I* shares the deaminase domain organization of these *A3*s (*A3Z2-A3Z1*; Figure 1A). With high bootstrap support, we found that all *A3I*s share a common ancestor to the exclusion of *A3B* and *A3G* genes. This pattern held with a tree of individual deaminase domains from all human *A3*s (Figure S1A). Our analyses found the *A3G* genes to be the closest phylogenetic neighbors of the *A3I*s, suggesting a common ancestry of these two genes (Figure 1C). Since *A3G* is predicted to have been born soon after the simian-prosimian split (Münk et al., 2012), we propose that *A3I* arose via retrocopying of *A3G* in the common ancestor of simian primates, approximately 43 million years ago (MYA) (Perelman et al., 2011).

In all species analyzed, *A3I* has acquired potentially inactivating mutations relative to *A3G*. We found that all *A3I* retrocopies share a nonsense mutation at codon position 261 (File S1), and thus, in most species, *A3I* encodes only a short putative open reading frame (ORF) of 153 codons (compared to the 384 codon ORF of *A3G*) which spans the N-terminal deaminase domain. Following this initial truncation, there were additional lineage-specific disruptions of the *A3I* ORF during the diversification of simian primates. These results suggest that *A3I* was born in the simian ancestor either as a truncated retrocopy or acquired a truncating mutation shortly following birth. Nonetheless, it is possible that the ancient *A3I*s encoded functional A3 proteins before becoming disabled by mutation.

Multiple *A3G* retrocopies in New World monkeys

A3I was not the only hit uncovered by our search for *A3* retrocopies. Our analyses also revealed *A3F* retrocopies in three Old World monkeys within the Colobinae subfamily, *A3H* retrocopies in two New World monkeys, and a single *A3* retrocopy that cannot be assigned to a specific *A3* parent in the greater galago (*Otolemur garnettii*) (Figure 1D). However, our most striking finding was that every sequenced NWM genome contained numerous *A3G* retrocopies. This abundance of *A3* retrocopies motivated a deeper investigation of the evolution and function of these NWM *A3G* retrogenes.

Initially focusing on the common marmoset (*Callithrix jacchus*) genome, we found three intron-containing *A3* genes on chromosome 1 (likely orthologous to human *A3A*, *A3G*, and *A3H*). We also found nine loci outside of the *A3* locus with high sequence similarity to human and marmoset *A3G* genes (Figure 2A). In contrast to the marmoset *A3G* gene, each of these additional hits lacked introns suggesting they were retrocopies. Seven *A3G* retrocopies spanned more than 1,100 bp and most of the coding exons of the marmoset *A3G*. In addition, one retrocopy spanned 811 bp, and one shorter retrocopy spanned 260 bp (Figure 2A). These shorter retrocopies showed a marked 3' bias, consistent with the 5'-truncation-prone, target-primed reverse transcription mechanism of LINE-1 and related retrotransposons (Cost, 2002; Luan et al., 1993). Seven *A3G* retrocopies possessed premature stop codons, deletions, or early truncations (Figure 2A). Two of these retrocopies encode ORFs that span a single cytidine deaminase domain ('*C1*' and '*C2*'). However, two *A3G* retrocopies ('*SS1'* and '*SS2'*, Figure 2A) were predicted to encode a 382 amino acid protein (comparable to the full-length 384 amino acid protein encoded by the intron-containing *A3G* gene). Thus, the marmoset genome contains nine *A3G* retrocopies, which may encode 2-4 additional A3G-like proteins.

Next, we expanded our search for *A3G* retrocopies to the other assembled NWM genomes on NCBI: the Bolivian squirrel monkey (*Saimiri boliviensis*), the white-faced capuchin (*Cebus*

capucinus), and Ma's night monkey (*Aotus nancymaae*). Like marmoset, we found multiple *A3G* retrocopies in each of these genomes – eight in capuchin, seven in night monkey, and ten in squirrel monkey (Figure 2A). All but one of the NWM *A3G* retrocopies aligned (without large gaps) to each other and intron-containing *A3G*s, and a phylogenetic tree confirmed these retrocopies cluster with NWM *A3G*s (Figure S1B). The exception, one squirrel monkey *A3G* retrocopy (GenBank: JH378161), contains a 286 bp insertion of another shorter *A3G* retrocopy, most likely the result of a nested insertion of one retrocopy into another.

A maximum likelihood phylogenetic tree (Figure 2B) using the alignable NWM *A3Gs* and *A3G* retrocopies revealed six bootstrap-supported 'clusters' of retrocopies with representatives from all four analyzed NWM species (Figure 2B and Table S1 clusters *C1-C5* and *A3I*). One of these clusters contains the NWM *A3I* sequences which date back to at least the last common ancestor of simian primates. Our findings suggest that the other five clusters represent orthologs of *A3G* retrocopies born in or before the last common ancestor of these four NWM species. This orthology was further supported by shared synteny in two NWM genomes (marmoset and squirrel monkey) for all clusters (Figure S2). We, therefore, conclude that five orthologous NWM *A3G* retrocopies were likely born via independent retrotransposition events in or prior to the most recent common ancestor of these four species analyzed.

To more precisely date the origins and species distribution of these *A3G* retrocopies, we investigated their presence in additional NWM species lacking publicly available genome sequences. For each retrocopy, we used UCSC MultiZ alignments (Blanchette, 2004) to find flanking sequence conservation in shared syntenic locations of marmoset, human, and mouse genomes to design oligos specific to a single retrogene-containing locus in marmoset. We were able to do so for seven of nine retrocopies (all but *SS3* and *C4*, Figure 2A). Confirming the specificity of these oligos, six of these seven oligo pairs reproducibly amplified a single locus

from marmoset genomic DNA with touchdown PCR whereas only *A3I* was amplified from human genomic DNA (Figure S3). Using these oligos and genomic DNA from other species, we observed that many retrocopies were present in the shared syntenic loci in other NWMs. Specifically, retrocopies *C1*, *C2*, *C3*, and *C5* of marmoset were also present in titi (*Plecturocebus moloch*) and saki monkeys (*Pithecia pithecia*), two species in the basal family of the NWM phylogeny, suggesting these retrocopies were born in or prior to the most recent common ancestor of all NWMs (~25 MYA).

In addition to the six orthologous 'clusters' of retrocopies found in all or many NWMs, our phylogenetic analysis also reveals 'species-specific' retrocopies (*SS*) with no apparent ortholog in the other three species with genome assemblies (Figure 2B). These *A3G* retrocopies instead share a recent common ancestor with the intron-containing *A3G* gene from the same species, suggesting that they were born recently. Our PCR analysis revealed that some 'marmoset-specific' retrocopies were also present in the closely related tamarin (*Saguinus oedipus*). Thus, *A3G* retrocopies vary in age, from being found in only one species, in a few closely related species, in all NWM species, or in all simian primates. The different branch lengths leading to each of the NWM *A3G* retrocopies or retrocopy ortholog clusters also reflect their variable ages (Figure 2B). Our findings suggest that rather than a single burst, *A3G* retrocopies have been continually born throughout the evolutionary history of NWMs.

Retention of putatively functional NWM *A3G* retrogenes

Retrocopies are often assumed to be nonfunctional at birth since they usually consist of only the sequence within the mRNA of the parent gene and therefore lack promoters, and enhancers. However, there are well-documented examples of retrogenes that have been retained for their functionality (Casola and Betrán, 2017). To investigate whether any of the *A3* retrocopies might

be functional, we used several criteria to eliminate retrocopies likely to be non-functional. We assumed that functional retrocopies should be transcribed and should have evolved under selective constraint on an intact open reading frame with an intact cytidine deaminase motif. To be conservative, we narrowed our focus to the *A3G* retrocopies detectable in the four sequenced, publicly available New World monkey genomes – nine copies in common marmoset, eight copies in white-faced capuchin, ten copies in Bolivian squirrel monkey, and seven copies in Nancy Ma's night monkey.

Many *A3* genes are expressed in the germline and early development (Friedli et al., 2014; Marchetto et al., 2013; Refsland et al., 2010) where they protect against a diverse range of infectious and endogenous elements including retroviruses and LINE-1 (Arias et al., 2012; Harris and Dudley, 2015). In order to similarly assess expression of *A3G* retrocopies *in vivo*, we queried publicly available NWM RNA-seq datasets (Table S2) using all reference *A3G* retrocopy sequences. We organized all perfect-matching, uniquely-mapping read counts by species (Figure 3, x-axis) and specific retrocopy (Figure 3, y-axis). The intron-containing *A3G* gene itself showed detectable expression in most datasets in all four species (organized by tissue source of each dataset, colored bars down columns denote tissues). We also found that several *A3G* retrocopies showed expression in each species (dark bars within each species). For example, *C2* retrocopies (Figure 3) are expressed in three of the four analyzed NWM species. In marmoset, the *C2* retrocopy is expressed in stem cells and induced pluripotent cells, similar to the intron-containing *A3G* gene. We also found that each species expressed at least one species-specific retrocopy. In several species, these younger, species-specific *A3G* retrocopies were expressed in ovaries and testes just like the intron-containing *A3G*. Overall, our data suggest that a large subset of NWM *A3G* retrocopies are expressed *in vivo*, including in tissues relevant for defense against pathogens.

Second, we evaluated the *A3G* retrocopies for their predicted ORFs. Full-length NWM *A3G* genes encode a ~384 amino acid protein. In contrast, most *A3G* retrocopies encode only short putative ORFs, typically less than 100 codons (Figure 2A, yellow boxes). However, a subset of *A3G* retrocopies have retained a predicted ORF of at least 250 codons (Figure 2A, yellow boxes; Figure 2B, empty circles on branches) which encompasses one or two of the core deaminase domains. Most of these retrocopies conserve the core amino acids within these domains required for antiviral or anti-retroelement activities of *A3G* (Figure 4) (Huthoff and Malim, 2007; Navarro et al., 2005).

Third, we evaluated the *A3G* retrocopies for evidence of selective retention. Although *A3G* retrocopies are expected to lack stop codons upon their birth from an intact *A3G* gene, absence of stop codons in older *A3G* retrocopies could indicate functional retention. We adapted a previously published approach (Young et al., 2018) to simulate the rate of decay of ORFs in the absence of selection based on ORF length and conservative and liberal bounds on NWM background mutation frequency and generation time (Campbell and Eichler, 2013; Tacutu et al., 2018; Thomas et al., 2018). We found that that less than 5% of *A3G* ORFs were expected to remain intact after 20 million years (less than 1% after 40 million years) (Figure 5). In contrast to this expectation, we found two *A3G* retrocopies have remained intact despite being at least 20 million years old (Bininda-Emonds et al., 2007). These include one *C1 A3G* retrocopy (with a preserved ORF in capuchin monkey) and a *C2* retrocopy (with a preserved ORF in marmoset and night monkey). Based on these findings, we hypothesize that some NWM *A3G* retrocopies have been retained for their function.

To further evaluate the selective constraint acting on *A3G* retrocopies, we used computational models to test whether their evolution more closely resembles a functional gene or a pseudogene. We first used the RELAX method (Wertheim et al., 2015) to test whether the *A3G*

retrocopies show relaxed selection relative to intron-containing *A3G* genes. Significant relaxed selection was not detected in the putatively intact retrocopies relative to the intron-containing *A3G*s (Figure S4). Instead, RELAX suggests that the *A3G* retrocopies have evolved more rapidly than the intron-containing *A3G* genes in the same set of species (Figure S4). Next, using a branch model of PAML (Yang, 2007), we observed that two retrocopies (capuchin-*C1* and marmoset-*SS1*) had elevated dN/dS (2.6 and 2.9 respectively, significantly greater than the neutral expectation of 1), while the rest of the branches were suggestive of neutral evolution or purifying selection (Figure S4). These analyses suggested that, overall, the retrocopies evolved at a similar or accelerated rate compared to intron-containing *A3G*s. Further, capuchin-*C1* and marmoset-*SS1* show evidence of accelerated evolution. Overall, our three lines of evidence suggest that at least a subset of the *A3G* retrocopies are likely to have been retained for their function.

Antiviral activity of NWM *A3G* retrocopies

We reasoned that *A3G* retrocopies could have a role in innate immunity/genome defense similar to intron-containing *A3* genes. To test this possibility, we cloned and assayed intron-containing *A3Gs* and each *A3G* retrocopy encoding an intact near-full-length ORF for its ability to restrict the endogenous retroelement LINE-1 using established *in vitro* retrotransposition assays (Dewannieux et al., 2003; Moran et al., 1996). These assays require that a LINE-1 sequence be transcribed, spliced, and reverse transcribed back into the genome. As controls, we tested the anti-LINE-1 restriction of human A3A and human A3G. Consistent with previous reports (Bogerd et al., 2006; Chen et al., 2006; Muckenfuss et al., 2006; Niewiadomska et al., 2007), we observed potent restriction of LINE-1 by human A3A, and no restriction by human A3G. In contrast to human A3G, we found that the intron-containing A3Gs from marmoset and squirrel monkey restricted LINE-1 more than 10-fold, comparable to A3A. However, we observed no

appreciable restriction of LINE-1 by any of the *A3G* retrocopies (Figure 6). Thus, despite potent anti-LINE-1 restriction by NWM A3Gs, it appears that this activity is not retained by any of the retrocopies tested.

Next, we investigated the antiviral restriction by NWM *A3G* genes and retrocopies. Using single-cycle infectivity assays, we measured the ability of NWM *A3G* genes and retrocopies to block infectivity of HIV-1ΔVif, which lacks Vif, a known antagonist of APOBEC3 proteins. Consistent with previous results, we found that human A3G potently restricts HIV-1ΔVif but human A3A is a poor restrictor; this restriction pattern is the opposite to that observed for LINE-1 restriction in here and in previous findings (Bogerd et al., 2006; Chen et al., 2006; Turelli et al., 2004) (Figure 6). We also observed 100-fold or greater restriction of HIV-1ΔVif infectivity by intron-containing *A3G* genes from marmoset and squirrel monkey (Figure 6), consistent with a previous report of restriction by NWM A3Gs (Wong et al., 2009). Finally, we observed that two retrocopies – marmoset-*SS1* and capuchin-*C1* – restrict HIV-1ΔVif at least 10-fold, suggesting that these two *A3G* retrocopies encode *bona fide* A3G-like anti-retroviral activity. Thus, retrocopying has expanded the functional repertoire of *A3* antiviral genes in NWMs. At least 2 of these genes are expressed at the RNA level in at least some tissues and encode a functional protein with antiviral activity.

Discussion

Replicating retrotransposons inflict deleterious consequences on host genomes via insertional mutagenesis, ectopic recombination, and dysregulation of proximal genes (Beck et al., 2011). Despite these negative consequences, retrotransposons can bring about innovation in host genomes via the birth of new exons or genes (Mi et al., 2000; Schmitz and Brosius, 2011), or novel regulatory mechanisms and gene-regulatory networks (Chuong et al., 2016; Kunarso et

al., 2010; Wang et al., 2007). In this work, we show that retrotransposon-mediated gene birth can lead to continual evolution of new innate immune genes. We show that all simian primate genomes contain the remnants of *A3I*, an ancient *A3* retrocopy. We further find that NWM genomes have continually acquired *A3G*-derived retrocopies, a subset of which are transcribed, retain intact ORFs and functional motifs, and are capable of restricting retroviruses.

This history of ancient and young retrocopies provides a valuable resource in understanding how antiviral genes coevolve with pathogens, including changes in Vif-interacting residues or viral restriction profile (Krupp et al., 2013). Although numerous methods exist for reconstruction of ancestral sequences, rapidly evolving genes like the *A3s* violate assumptions and often limit the utility of these methods, thereby preventing reliable reconstruction of ancestral sequences. However, retrocopies are molecular fossils, an evolutionary snapshot of the ancient parental gene sequence which presumably evolved neutrally after inserting into the genome. *A3I* provides such a record of an *A3G*-like gene from 40 MYA, which was present in the common ancestor of simian primates. Given the rapid gene turnover of the *A3* locus in mammals, it is possible that the parent of *A3I* no longer exists in modern primates. In this scenario, the *A3I* retrocopy may be all that remains of this ancient *A3* gene which predates simian primate diversification.

Recent computational analysis corroborates the presence of *A3* retrocopies in two of the genomes we analyzed (Hayward et al., 2018; Ito et al., 2020) and adds to a growing literature suggesting the *A3* content of mammalian genomes may be even more variable and dynamic than previously appreciated (Hayward et al., 2018; Ito et al., 2020). Our data suggests that *A3* retrocopying is more prevalent in NWM genomes compared even to other simian primates. This abundance is consistent with a previous study that reported an increased number of retrocopies of all genes in marmoset and squirrel monkey genomes, correlated with an increase in the

activity of two LINE-1 subfamilies L1PA7 and L1PA3 (Navarro and Galante, 2015). It is unclear whether increased LINE-1 activity is sufficient to explain our observations since some NWMs like the *Ateles* lineage may have low or no retroelement activity (Boissinot et al., 2004). Even if NWM LINE-1 activity is high, it would not necessarily explain why *A3G* rather than the other NWM *A3* genes are subject to recurrent retrocopying. Although duplication of some nuclear A3 proteins like human A3A or A3B are likely to be more toxic due to increased genomic mutation (Hultquist et al., 2011; McLaughlin et al., 2016), we favor the alternate hypothesis that *A3G* expression in the germline/early embryos of NWMs is unusually high, rendering it a more likely substrate for retrocopying relative to other NWM *A3* genes. Following their insertion into a new genomic location, these retrocopies could be expressed by exaptation of a neighboring transposable element, promoter piggybacking, or recruitment of a novel promoter (Carelli et al., 2016). Recent work suggests that most of the mouse genome is transcribed over relatively short evolutionary timescales (Neme and Tautz, 2016). Such 'genome-wide' transcription could be the first step in exposing an advantageous function of a retrogene (Jaganathan et al., 2019).

We showed that intron-containing NWM *A3G* genes restricted both LINE-1 and HIV-1. Thus, it is likely that *A3G* retrocopies retained both of these functions immediately following birth. Yet, over time, all retrocopies that restrict HIV-1 have lost the ability to restrict LINE-1 (Figure 6). Although this could reflect idiosyncratic events, our finding that anti-LINE-1 activity, but not anti-retroviral activity, was repeatedly lost, suggests otherwise. It is possible that the anti-LINE-1 function is simply more sensitive to random mutation, such that mutations are more likely to result in loss of LINE-1 restriction; we also cannot rule out the possibility that certain *A3G* retrocopies retain the capacity to restrict NWM-specific LINE-1 lineages. Alternatively, *A3G* retrocopies may have been absolved of selection for LINE-1 restriction, perhaps due to sufficient silencing by *A3G* and other restriction factors. Nevertheless, the retrocopies present a natural 'separation of function'

event that can delineate the requirements for *A3G* proteins to restrict LINE-1 versus retroviruses.

Although we used the lentivirus HIV-1ΔVif to measure the anti-retroviral activity of NWM *A3G* retrocopies, lentiviruses have not yet been found in NWMs. Even apart from lentiviruses, few active retroviruses in general have been found in NWMs; those that have been found likely represent the tip of an understudied aspect of monkey and virus biology (Colcher et al., 1977; Muniz et al., 2013). Thus, HIV-1ΔVif only serves as a proxy for the activity of *A3G* retrocopies towards some relevant viral pathogen in the natural environment of these monkeys.

While the NWM *A3G* retrocopies did not restrict LINE-1, such a mechanism of gene duplication could, in theory, function as a feedback mechanism on excess retroelement activity in the germline/early embryo. Retroelement restriction factors expressed in these tissues could be retrocopied and increase dosage or diversity of anti-retroelement restriction factors (Kondrashov et al., 2002). In this way, the retrocopies may represent a 'revolving door' of new gene substrates for neo- or sub-functionalization; the needs of the genome would dictate which functions persist.

In conclusion, our findings suggest retrocopied gene sequences represent a prevalent, recurrent, and rapid mechanism in primates and other organisms to evolve new genome defense functions including restriction of viruses. Although the presence of endogenous retroelements is probably net deleterious to the host, retrogene birth represents a mechanism whereby host genomes could nevertheless take advantage of the activities of these genomic pathogens to protect themselves against endogenous and infectious pathogens.

Materials and Methods

Identification of *A3* retrocopies

A3G retrocopies were identified using BLAT of UCSC genome databases for marmoset (*Callithrix jacchus* draft assembly, WUGSC 3.2, GCA_000004665.1) and squirrel monkey (*Saimiri boliviensis*, saiBol1, GCA_00023585.1) with marmoset *A3G* (NM_001267742) as a query sequence. Additional copies were identified using BLASTn of the NCBI genome assemblies of Ma's night monkey (*Aotus nancymaae*, Anan_2.0) and capuchin (*Cebus capucinus imitator*, Cebus_imitator-1.0). The spider monkey retrocopy was identified using BLAST to query the NCBI HTGS database for reads from New World monkeys. See Table S1 for detailed coordinates of each sequence.

Mapping inactivating mutations in retrocopies

A3I sequences were queried using the codon-based and indel-sensitive alignment program LAST (http://last.cbrc.jp). The translated *A3G* sequence of *Callithrix jacchus* (NC_013914.1) was used as the reference sequence and indexed using the setting of 'lastdb -p -cR01' of the LAST aligner, and then the *A3I* sequences were queried using the setting of 'lastal -F15' to output in 'maf' format. The longest indel-sensitive translation of each *A3I* was then manually extracted from the maf output and aligned with mafft (https://mafft.cbrc.jp/alignment/software/) using the setting of '--anysymbol' to allow stop codons and frame shifting changes to be shown.

Analysis of syntenic *A3G* retrocopies

Synteny of *A3G* retrocopies in marmoset and squirrel monkey was analyzed using UCSC table browser to download gene names within 1Mbp of either side of the retrocopy. Synteny was confirmed if the same gene was adjacent next to the retrocopy in both species. For five pairs of sequences that the tree suggested should be orthologous, we found shared genes on both

sides of the retrocopies. For one retrocopy (*C5*), we found a shared gene on only one side of the retrocopies (Figure S2).

Construction of *A3* phylogeny

A3G and *A3G* retrocopies were aligned using MAFFT (Katoh and Standley, 2013) with auto algorithm parameters within Geneious version 11.1.4 (Kearse et al., 2012). All retrocopies (both ORF-containing and retropseudogenes) were aligned using the complete alignable region defined by BLASTn. Trees were constructed using PHYML (Guindon et al., 2010) with NNIs topology search, BioNJ initial tree, HKY85 nucleotide substitution model, and 100 bootstraps.

ORF retention simulation

To simulated the decay of retro *A3G* ORFs, we used the 'mutator' and 'orf_scanner' scripts developed by Young *et al.* (Young et al., 2018). The ORF of *Callithrix jacchus A3G* (identified from GenBank accession NC_013914.1, 1,150 bp) was used as the starting ORF. Combinations of several substitution, insertion, deletion rate and sexual maturation time were used for the simulation. We used substitution, insertion and deletion rate of 1.16×10^{-8}, 2×10^{-10} and 5.5×10^{-10} per site per generation for human (Campbell and Eichler, 2013), substitution, insertion and deletion rate of 5.4×10^{-9}, 1.55×10^{-10} and 1.55×10^{-10} per site per generation for mouse (Uchimura et al., 2015), and substitution rate of 8.1×10^{-9} for Night monkey (*Aotus nancymaae*) (Thomas et al., 2018). Sexual maturation time of human, mouse and New World monkeys were estimated to be 25, 0.3 and 1-9 years (http://genomics.senescence.info; https://animaldiversity.org). Each run simulates the mutation of the starting ORF for 50 million years, and the simulations with each set of parameters were repeated 10,000 times. The number of ORFs that were still open and at the same length of the starting ORF were counted at every 50,000 years of each simulation.

Analysis of selective constraints in *A3G* retrocopies

RELAX (Wertheim et al., 2015) was carried out using the Datamonkey webserver (Weaver et al., 2018) and a PhyML (Guindon et al., 2010) tree of the MAFFT (Katoh and Standley, 2013) aligned nucleotide sequences of the subset of retrocopies that encode an ORF longer than 250 amino acids in addition to the New World monkey *A3Gs* with or without human *A3G*. We defined the branches leading to the *A3Gs* as reference branches and all of the other branches as test branches. The above nucleotide alignment and PhyML tree were input into the CODEML NSsites model of PAML (Yang, 2007). To test for selection along branches, these same input files were input into the branch model of PAML. To test for significance of branches with apparent dN/dS <1, we fixed that branch at dN/dS=1 and calculated the likelihood of this tree.

RNA-seq analysis for retrocopy and *A3G* expression

We searched the NCBI GEO and SRA databases (October 2018) with the keywords 'Callithrix', 'Aotus', 'Saimiri' and 'Cebus' to find existing RNA-seq datasets from these species. *Callithrix jacchus*, *Aotus nancymaae*, and *Cebus capucinus* are used, matching the available species where retrocopies of *A3G* were identified. For *Saimiri*, *Saimiri sciureus* RNA-seq was used, for which no genome sequence has been published, and the retrocopy analysis in the rest of the text analyzes *Saimiri boliviensis*. All RNA-seq datasets (Table S2) were queried using the default parameters of the 'blastn_vdb' tool of SRA toolkit (Leinonen et al., 2011) and the identified *A3Gs* and *A3G* retrocopies in this work as query sequences. RNA-seq reads hit by blastn_vdb were then processed with a custom perl (https://www.perl.org) script to only keep the reads that match the query sequence at 100% identity across the entire RNA-seq read and maps uniquely to only one of the queried retrocopies or *A3G*. Read that passed these filters were tallied and organized by species, tissue type, and the retrocopy or *A3G* copy they match.

LINE-1 retrotransposition assays

LINE-1 retrotransposition assays were carried out as previously described (Xie et al., 2011). For the mouse ORFeus luciferase assays 25,000 293T cells were seeded into each well of a 96-well clear bottom, white-wall plate. 24 hours later, each well was transfected with 200ng pYX016 (CAG promoter driving mouse ORFeus LINE-1 with globin intron and luciferase reporter) or pYX015 (Xie et al., 2011) (JM111 inactive human LINE-1 construct which contains loss-of-function mutations in ORF1p of LINE-1) and pCMV-HA-A3 or pCMV-HA-empty. 24 hours post-transfection, transfected cells were selected with 2.5 µg/ml puromycin for 72 hours. Cells were lysed and luciferase substrates provided using the Dual-Glo Luciferase Assay System (Promega E2920). Renilla and firefly luciferase activity were measured using the LUMIstar Omega luminometer. Retrotransposition is reported as firefly/renilla activity to control for toxicity.

Virus infectivity assays

Single-round HIV-1 infectivity assays were performed as described previously (OhAinle et al., 2006; Yamashita and Emerman, 2004). To produce VSV-G-pseudotyped HIV-1, 50,000 293T cells were plated in a 24-well plate, and 24 hours later, co-transfected with 0.3 µg lentiviral vector encoding luciferase in the place of the nef gene (pLai3ΔenvLuc2 (Yamashita and Emerman, 2004), pLai3ΔenvLuc2ΔVif (OhAinle et al., 2006), 50 ng L-VSV-G, and 300 ng pCMV-HA-A3G or pCMV-HA-empty. All viruses were harvested 48 hours after transfection and filtered through a 0.2-µm filter. p24 *gag* in viral supernatants was quantified using an HIV-1 p24 Antigen Capture Assay (ABL Inc.). Virus equivalents to two nanograms of p24 *gag* were used to infect 50,000 SupT1 cells per well in a 96-well plate in the presence of 20 µg/ml DEAE-dextran. Forty-eight hours after infection, cells from triplicate infections were lysed in 100 µl Bright-Glo luciferase assay reagent (Promega) and read on a LUMIstar Omega luminometer (BMG Labtech). A3A Western blots were carried out using Covance mouse HA.11 Clone 16B12 anti-HA monoclonal antibody.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

We thank members of the Malik, Emerman, and McLaughlin labs for valuable discussions. We especially thank Janet Young for comments and suggestions critical for the generation of this manuscript, and Lily Wu for technical training and consultation on virus restriction assays. This work was supported by a Howard Hughes Medical Institute postdoctoral fellowship of the Helen Hay Whitney Foundation, a National Institute of General Medical Sciences (NIGMS) at the National Institutes of Health (NIH) K99/R00 Pathway to Independence Award (grant number GM112941) to RNM; National Institute of Allergy and Infectious Disease at the NIH R01 to ME (grant number AI3092); grants from the Mathers Foundation and an NIGMS at the NIH R01 (grant number GM074108). HSM is an Investigator of the Howard Hughes Medical Institute.

Figure legends

Figure 1

Identification and phylogenetic distribution of *A3I*

A) *A3I* is located away from the *A3* locus at a distant but highly conserved syntenic locus in all simian primates. The human genome is shown as an example. B) ORF structure of *A3I* in various primate species. Purple boxes represent sequences that can be aligned to the intron-containing *A3* copies, whereas yellow boxes represent the longest ORF of the *A3I* in corresponding species. Stars (*) indicate the position of stop codons. C) Maximum likelihood

phylogeny of *A3Is* and the intron-containing *A3Bs* and *A3Gs*. Clusters of *A3Is*, *A3Gs*, and *A3Bs* are highlighted by their respective color, and bootstrap values leading to these clusters are shown on the nodes. D) Expansion of *A3* retrocopies along the primate phylogeny. The number of retrocopies of each A3 is shown in color boxes at the inferred point of retrocopy birth in the primate phylogeny. The white '*A3*' box represents a sequence that could not be assigned to a particular ortholog.

Figure 2

Discovery and phylogenetic analysis of *A3G* retrocopies in New World monkeys

A) The common marmoset (*Callithrix jacchus*) genome encodes a single *A3G* and nine retrocopies of *A3G* (orange boxes). *A3G* resides on four coding exons at the *A3* locus on chromosome 1, while the retrocopies are intronless and found throughout the genome. Some retrocopies contain putative protein-coding ORFs (yellow boxes) of varying lengths that retain alignable sequence similarity to the *A3G* protein (gray within yellow boxes, regions of poor alignment caused by frame-shifting mutations). B) PhyML tree of *A3G* and *A3G* retrocopies from the four sequenced and assembled New World monkey genomes suggests that six retrocopies are orthologous and conserved in all four species (clusters *C1-C5* and *A3I*). The genome of each species (colors correspond to species) contains an intron-containing *A3G* (dotted line) as well as retrocopies that are closely related to *A3G*. These more recent copies are found in only one genome, without identifiable orthologs in the other three species. Some retrocopies retain a putative protein coding ORF (indicated by a circle at the branch tip).

Figure 3

A3G retrocopies are transcribed in New World Monkey tissues

A heat map shows the counts of RNA-seq reads ($\log_{10}$ of read count + 1) that map uniquely at 100% identity and coverage. Each pixel represents the average read counts of available data for the corresponding tissue type and *A3G* retrocopy. Tissue types are marked by the colored lines behind the pixels. Green represents germline tissues including iPSC, ESC, testis and ovary; orange represents brain tissues of various regions; blue represents blood samples including whole blood and lymphocytes. Retrocopies which retain a putative protein coding ORF are labeled with 'ORF'.

Figure 4

A3G retrocopies retain core deaminase motifs

An amino acid alignment of the core deaminase motifs shows that *A3Gs* of various primates have conserved HxE-CxxC motifs in both the N- and C-terminal domains. The putative ORF-encoding retrocopies all retain a conserved C-terminal motif, and most retain an N-terminal motif.

Figure 5

Simulation and evolution suggest selection to retain ORFs in *A3G* retrocopies

A) A simulation of ORF retention suggests most ORFs are lost within 10-20 million years of evolution, in the absence of any selection to retain the ORF. Dots indicate the proportion of simulated ORFs (10,000 total) that were still intact after a given time. Colors represent three sets of parameters intended to match New World monkeys (green) or provide liberal (orange, mouse-like) and conservative (blue, human-like) bounds on the parameter sets of indel rate and generation time. The substitution rate of Ma's night monkey was used for all three sets of

simulations. Horizontal red lines indicate the 1st and 5th percentile of intact ORFs. Vertical red lines mark the key time points of last common ancestor (LCA) among New World monkeys.

Figure 6

A3G retrocopies restrict HIV-1 but not LINE-1

Bar charts of measured restriction of LINE-1 (retrotransposition assays) and HIV-1ΔVif (single cycle infectivity assays) show that NWM *A3Gs* and some *A3G* retrocopies restrict retrovirus. Only NWM *A3G*s, but not retrocopies restrict LINE-1.

Figure S1

A phylogeny of primate *A3s* and *A3G* retrocopies

A) A PhyML tree of the domains of *A3*s and *A3I*s. Primate *A3*s with two deaminase domains were split into their constituent N-terminal and C-terminal domains and aligned together with all single deaminase domain *A3*s. Black triangles indicate bootstraps > 90%. B) A PhyML tree of nucleotide sequences of an alignable region of *A3Gs* and *A3Bs* from diverse simian primates in addition to *A3G* retrocopies from New World monkeys. Tree is rooted on the *A3B* group. New World monkey retrocopies group with New World monkey *A3G*s which group in a larger group with *A3G*s from all simian primates. Exceptionally, one retrocopy (*A3I*) is found in all simian primate genomes and forms a monophyletic group that branches before the diversification of simian primates (and their *A3G*s).

Figure S2

Synteny analysis of retrocopies in marmoset and squirrel monkey for inference of orthology

UCSC table browser was used to identify the genes on either side of each retrocopy in marmoset and squirrel monkey. An occurrence of the same genes on both sides of a copy in each species was used as an indicator of synteny and therefore orthology. All inferred orthologous pairs were also supported by phylogenetic groupings.

Figure S3

PCR of genomic DNA of New World monkeys to date retrocopy births

A) PCR amplification of marmoset genomic DNA using oligos designed to designed to amplify each retrocopy locus in marmoset shows filled loci for 6/7 retrocopies and a failed PCR reaction for retrocopy *SS1*. Oligos designed to amplify a squirrel-monkey copy, not found in marmoset, show no band. B) Amplification of marmoset retrocopy-containing loci in a panel of New World monkeys plus human genomic DNA show variable presence and/or retention of retrocopies across species.

Figure S4

Analysis of selection in the evolution of retrocopies

Top, PAML free ratio analysis of selection along branches (omega values for terminal branches shown in orange). Omega values less than one for all terminal branches leading to *A3G*s suggest these genes have evolved under purifying selection. The branches leading to the two retrocopies that restrict retrovirus have elevated omega values (significantly higher than dN/dS=1, p=0.058 for *capuchin-C1*, p=0.025 for *marmoset-SS1*) suggesting these retrocopies have evolved under positive selection. Bottom, RELAX analysis of overall selection when comparing *A3G*s (purple) to *A3G* retrocopies (black) suggests the retrocopies have evolved

under intensified selection (no detection of relaxation of selection) relative to the presumably functional *A3G*s.

Figure S5

Western blot of A3s and A3G retrocopies

For each construct, 50ng plasmid was transfected into 25,000 293T cells in a single well of a 24 well plate. Forty-eight hours later, cells were harvested, lysed, and probed with Covance mouse HA.11 Clone 16B12 anti-HA monoclonal antibody.

Table S1

Sequence coordinates, orthology groups, and ORF retainment for *A3G*s and *A3G* retrocopies

Table S2

Read counts of retrocopies across 98 New World monkey RNAseq datasets

File S1

Codon-based and indel-sensitive alignment of primate *A3I*s

Stop codons and frame shifts were included in the alignment: star (*) represents a stop codon, slash (/) represents a frame shift caused by deletion, and backslash (\) represents a frame shift caused by insertion. Header of the sequences indicate the names of species and the NCBI accession numbers where the sequences are extracted from.

References

Arias JF, Koyama T, Kinomoto M, Tokunaga K. 2012. Retroelements versus APOBEC3 family members: No great escape from the magnificent seven. *Front Microbiol* **3**:275.

doi:10.3389/fmicb.2012.00275

Balasubramanian S, Zheng D, Liu Y-J, Fang G, Frankish A, Carriero N, Robilotto R, Cayting P, Gerstein M. 2009. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol* **10**:R2. doi:10.1186/gb-2009-10-1-r2

Beck CR, Garcia-Perez JL, Badge RM, Moran J V. 2011. LINE-1 Elements in Structural Variation and Disease. *Annu Rev Genomics Hum Genet* **12**:187–215. doi:10.1146/annurev-genom-082509-141802

Best S, Le Tissier P, Towers G, Stoye JP. 1996. Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature* **382**:826–829. doi:10.1038/382826a0

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature* **446**:507–12. doi:10.1038/nature05634

Blanchette M. 2004. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Res* **14**:708–715. doi:10.1101/gr.1933104

Bogerd HP, Wiegand HL, Hulme AE, Garcia-Perez JL, O'Shea KS, Moran J V, Cullen BR. 2006. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. *Proc Natl Acad Sci* **103**:8780–8785. doi:10.1073/pnas.0603313103

Boissinot S, Entezam A, Furano A V. 2001. Selection Against Deleterious LINE-1-Containing Loci in the Human Lineage. *Mol Biol Evol* **18**:926–935. doi:10.1093/oxfordjournals.molbev.a003893

Boissinot S, Roos C, Furano A V. 2004. Different rates of LINE-1 (L1) retrotransposon amplification and evolution in New World monkeys. *J Mol Evol* **58**:122–130. doi:10.1007/s00239-003-2539-x

Brennan G, Kozyrev Y, Hu S-L. 2008. TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis. *Proc Natl Acad Sci* **105**:3569–3574. doi:10.1073/pnas.0709511105

Bulliard Y, Turelli P, Rohrig UF, Zoete V, Mangeat B, Michielin O, Trono D. 2009. Functional Analysis and Structural Modeling of Human APOBEC3G Reveal the Role of Evolutionarily Conserved Elements in the Inhibition of Human Immunodeficiency Virus Type 1 Infection and Alu Transposition. *J Virol*. doi:10.1128/jvi.01491-09

Burki F, Kaessmann H. 2004. Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat Genet* **36**:1061–1063. doi:10.1038/ng1431

Campbell CD, Eichler EE. 2013. Properties and rates of germline mutations in humans. *Trends Genet* **29**:575–84. doi:10.1016/j.tig.2013.04.005

Canapa A, Barucca M, Biscotti MA, Forconi M, Olmo E. 2015. Transposons, Genome Size, and Evolutionary Insights in Animals. *Cytogenet Genome Res* **147**:217–39. doi:10.1159/000444429

Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res* **26**:301–314. doi:10.1101/gr.198473.115

Casola C, Betrán E. 2017. The Genomic Impact of Gene Retrocopies: What Have We Learned from Comparative Genomics, Population Genomics, and Transcriptomic Analyses? *Genome Biol Evol* **9**:1351–1373. doi:10.1093/gbe/evx081

Chen H, Lilley CE, Yu Q, Lee D V, Chou J, Narvaiza I, Landau NR, Weitzman MD. 2006. APOBEC3A Is a Potent Inhibitor of Adeno-Associated Virus and Retrotransposons. *Curr Biol* **16**:480–485. doi:10.1016/j.cub.2006.01.031

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science (80-)* **351**:1083–1087. doi:10.1126/science.aad5497

Colcher D, Heberling RL, Kalter SS, Schlom J. 1977. Squirrel monkey retrovirus: an endogenous virus of a new world primate. *J Virol*.

Compton AA, Hirsch VM, Emerman M. 2012. The host restriction factor APOBEC3G and

retroviral Vif protein coevolve due to ongoing genetic conflict. *Cell Host Microbe*. doi:10.1016/j.chom.2011.11.010

Cost GJ. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**:5899–5910. doi:10.1093/emboj/cdf592

de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet* **7**:e1002384. doi:10.1371/journal.pgen.1002384

Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**:41–48. doi:10.1038/ng1223

Duggal NK, Malik HS, Emerman M. 2011. The Breadth of Antiviral Activity of Apobec3DE in Chimpanzees Has Been Driven by Positive Selection. *J Virol*. doi:10.1128/jvi.05046-11

Friedli M, Turelli P, Kapopoulou A, Rauwel B, Castro-Díaz N, Rowe HM, Ecco G, Unzu C, Planet E, Lombardo A, Mangeat B, Wildhaber BE, Naldini L, Trono D. 2014. Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency. *Genome Res* **24**:1251–1259. doi:10.1101/gr.172809.114

Fujino K, Horie M, Honda T, Merriman DK, Tomonaga K. 2014. Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. *Proc Natl Acad Sci* **111**:13175–13180. doi:10.1073/pnas.1407046111

Garcia-Perez JL, Marchetto MCN, Muotri AR, Coufal NG, Gage FH, O'Shea KS, Moran J V. 2007. LINE-1 retrotransposition in human embryonic stem cells. *Hum Mol Genet* **16**:1569–1577. doi:10.1093/hmg/ddm105

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* **59**:307–321. doi:10.1093/sysbio/syq010

Hancks DC, Kazazian HH. 2012. Active human retrotransposons: variation and disease. *Curr Opin Genet Dev* **22**:191–203. doi:10.1016/j.gde.2012.02.006

Harris RS, Dudley JP. 2015. APOBECs and virus restriction. *Virology* **479–480**:131–145. doi:10.1016/j.virol.2015.03.012

Hayward JA, Tachedjian M, Cui J, Cheng AZ, Johnson A, Baker ML, Harris RS, Wang L-F, Tachedjian G. 2018. Differential Evolution of Antiretroviral Restriction Factors in Pteropid Bats as Revealed by APOBEC3 Gene Complexity. *Mol Biol Evol* **35**:1626–1637. doi:10.1093/molbev/msy048

Henry M, Terzian C, Peeters M, Wain-Hobson S, Vartanian JP. 2012. Evolution of the primate APOBEC3A cytidine deaminase gene and identification of related coding regions. *PLoS One*. doi:10.1371/journal.pone.0030036

Hultquist JF, Lengyel JA, Refsland EW, LaRue RS, Lackey L, Brown WL, Harris RS. 2011. Human and Rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H Demonstrate a Conserved Capacity To Restrict Vif-Deficient HIV-1. *J Virol*. doi:10.1128/jvi.05238-11

Huthoff H, Malim MH. 2007. Identification of Amino Acid Residues in APOBEC3G Required for Regulation by Human Immunodeficiency Virus Type 1 Vif and Virion Encapsidation. *J Virol* **81**:3807–3815. doi:10.1128/JVI.02795-06

Ito J, Gifford RJ, Sato K. 2020. Retroviruses drive the rapid evolution of mammalian APOBEC3 genes. *Proc Natl Acad Sci* **117**:610–618. doi:10.1073/pnas.1914183116

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia A, Batzoglou S, Sanders SJ, Farh KK-H. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**:535-548.e24. doi:10.1016/j.cell.2018.12.015

Jarmuz A, Chester A, Bayliss J, Gisbourne J, Dunham I, Scott J, Navaratnam N. 2002. An Anthropoid-Specific Locus of Orphan C to U RNA-Editing Enzymes on Chromosome 22. *Genomics* **79**:285–296. doi:10.1006/geno.2002.6718

Kaer K, Speek M. 2013. Retroelements in human disease. *Gene* **518**:231–241.

doi:10.1016/j.gene.2013.01.008

Kalamegham R, Sturgill D, Siegfried E, Oliver B. 2006. Drosophila mojoless, a Retroposed GSK-3, Has Functionally Diverged to Acquire an Essential Role in Male Fertility. *Mol Biol Evol* **24**:732–742. doi:10.1093/molbev/msl201

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**:772–80. doi:10.1093/molbev/mst010

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**:1647–1649. doi:10.1093/bioinformatics/bts199

Klawitter S, Fuchs N V., Upton KR, Muñoz-Lopez M, Shukla R, Wang J, Garcia-Cañadas M, Lopez-Ruiz C, Gerhardt DJ, Sebe A, Grabundzija I, Merkert S, Gerdes P, Pulgarin JA, Bock A, Held U, Witthuhn A, Haase A, Sarkadi B, Löwer J, Wolvetang EJ, Martin U, Ivics Z, Izsvák Z, Garcia-Perez JL, Faulkner GJ, Schumann GG. 2016. Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. *Nat Commun* **7**:10286. doi:10.1038/ncomms10286

Kondrashov FA, Rogozin IB, Wolf YI, Koonin E V. 2002. Selection in the evolution of gene duplications. *Genome Biol* **3**:RESEARCH0008. doi:10.1186/gb-2002-3-2-research0008

Krupp A, McCarthy KR, Ooms M, Letko M, Morgan JS, Simon V, Johnson WE. 2013. APOBEC3G Polymorphism as a Selective Barrier to Cross-Species Transmission and Emergence of Pathogenic SIV and AIDS in a Primate Host. *PLoS Pathog* **9**:e1003641. doi:10.1371/journal.ppat.1003641

Kunarso G, Chia N-Y, Jeyakani J, Hwang C, Lu X, Chan Y-S, Ng H-H, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic

stem cells. *Nat Genet* **42**:631–634. doi:10.1038/ng.600

Lander E, Heaford A, Sheridan A, Linton L, Birren B, Subramanian A, Coulson A, Nusbaum C, Zody M, Dunham A, Baldwin J, Hunt A, Devon K, McMurray A, Mungall A, Dewar K, Doyle M, Marra M a, Fulton L a, FitzHugh W, Chinwalla a T, Funke R, Delehaunty A, Gage D, Harris K, Olsen A, Gibbs R a, Fujiyama A, Howland J, Toyoda A, Kann L, Rosenthal A, Lehoczky J, LeVine R, Rump A, McEwan P, Madan A, McKernan K, Federspiel N a, Abola a P, Meldrim J, Mesirov JP, Evans G a, Roe B a, Miranda C, de La Bastide M, Morris W, Bailey J a, Naylor J, Bateman A, Raymond C, Rosetti M, Jones T a, Kaspryzk A, Santos R, McLysaght A, Sougnez C, Smit a F, Williams A, Stange-Thomann N, Stojanovic N, Felsenfeld A, Wetterstrand K a, Wyman D, Patrinos A, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Deadman R, Deloukas P, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Jones M, Lloyd C, Matthews L, Mercer S, Milne S, Mullikin JC, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Mardis ER, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Lucas S, Elkin C, Uberbacher E, Frazier M, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Hattori M, Yada T, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Platzer M, Nyakatura G, Taudien S, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Qin S, Davis RW, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson M V, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Athanasiou M, Schultz R, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala

R, Aravind L, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Kasif S, Kennedy S, Kent WJ, Kitts P, Koonin E V, Korf I, Kulp D, Lancet D, Lowe TM, Mikkelsen T, Moran J V, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, Szustakowki J. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921. doi:10.1038/35057062

Leinonen R, Sugawara H, Shumway M. 2011. The Sequence Read Archive. *Nucleic Acids Res* **39**:D19–D21. doi:10.1093/nar/gkq1019

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell* **72**:595–605. doi:10.1016/0092-8674(93)90078-5

Malfavon-Borja R, Wu LI, Emerman M, Malik HS. 2013. Birth, decay, and reconstruction of an ancient TRIMCyp gene fusion in primate genomes. *Proc Natl Acad Sci* **110**:E583–E592. doi:10.1073/pnas.1216542110

Malik HS, Henikoff S. 2005. Positive selection of iris, a retroviral envelope-derived host gene in Drosophila melonogaster. *PLoS Genet*. doi:10.1371/journal.pgen.0010044

Marchetto MCN, Narvaiza I, Denli AM, Benner C, Lazzarini TA, Nathanson JL, Paquola ACM, Desai KN, Herai RH, Weitzman MD, Yeo GW, Muotri AR, Gage FH. 2013. Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503**:525–529. doi:10.1038/nature12686

McLaughlin RN, Gable JT, Wittkopp CJ, Emerman M, Malik HS. 2016. Conservation and Innovation of APOBEC3A Restriction Functions during Primate Evolution. *Mol Biol Evol*

33:1889–1901. doi:10.1093/molbev/msw070

McLaughlin RN, Young JM, Yang L, Neme R, Wichman HA, Malik HS. 2014. Positive Selection and Multiple Losses of the LINE-1-Derived L1TD1 Gene in Mammals Suggest a Dual Role in Genome Defense and Pluripotency. *PLoS Genet* **10**:e1004531. doi:10.1371/journal.pgen.1004531

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang X-Y, Edouard P, Howes S, Keith JC, McCoy JM. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403**:785–789. doi:10.1038/35001608

Moran J V, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH. 1996. High Frequency Retrotransposition in Cultured Mammalian Cells. *Cell* **87**:917–927. doi:10.1016/S0092-8674(00)81998-4

Muckenfuss H, Hamdorf M, Held U, Perković M, Löwer J, Cichutek K, Flory E, Schumann GG, Münk C. 2006. APOBEC3 Proteins Inhibit Human LINE-1 Retrotransposition. *J Biol Chem* **281**:22161–22172. doi:10.1074/jbc.M601716200

Muniz CP, Troncoso LL, Moreira MA, Soares EA, Pissinatti A, Bonvicino CR, Seuánez HN, Sharma B, Jia H, Shankar A, Switzer WM, Santos AF, Soares MA. 2013. Identification and Characterization of Highly Divergent Simian Foamy Viruses in a Wide Range of New World Primates from Brazil. *PLoS One*. doi:10.1371/journal.pone.0067568

Münk C, Willemsen A, Bravo IG. 2012. An ancient history of gene duplications, fusions and losses in the evolution of APOBEC3 mutators in mammals. *BMC Evol Biol* **12**:71. doi:10.1186/1471-2148-12-71

Muotri AR. 2016. L1 Retrotransposition in Neural Progenitor CellsMethods in Molecular Biology (Clifton, N.J.). pp. 157–163. doi:10.1007/978-1-4939-3372-3_11

Navarro F, Bollman B, Chen H, König R, Yu Q, Chiles K, Landau NR. 2005. Complementary function of the two catalytic domains of APOBEC3G. *Virology* **333**:374–386. doi:10.1016/j.virol.2005.01.011

Navarro FCP, Galante PAF. 2015. A Genome-Wide Landscape of Retrocopies in Primate Genomes. *Genome Biol Evol* **7**:2265–2275. doi:10.1093/gbe/evv142

Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *Elife* **5**:e09977. doi:10.7554/eLife.09977

Newman RM, Hall L, Kirmaier A, Pozzi L-A, Pery E, Farzan M, O'Neil SP, Johnson W. 2008. Evolution of a TRIM5-CypA Splice Isoform in Old World Monkeys. *PLoS Pathog* **4**:e1000003. doi:10.1371/journal.ppat.1000003

Niewiadomska AM, Tian C, Tan L, Wang T, Sarkis PTN, Yu X-F. 2007. Differential Inhibition of Long Interspersed Element 1 by APOBEC3 Does Not Correlate with High-Molecular-Mass-Complex Formation or P-Body Association. *J Virol* **81**:9577–9583. doi:10.1128/JVI.02800-06

Nishigaki K, Ito J, Odahara Y, Kuse K, Watanabe S, Kawamura M, Hiratsuka T, Ochi H. 2013. Refrex-1, a Soluble Restriction Factor against Feline Endogenous and Exogenous Retroviruses. *J Virol*. doi:10.1128/jvi.01267-13

Nisole S, Lynch C, Stoye JP, Yap MW. 2004. A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1. *Proc Natl Acad Sci* **101**:13324–13328. doi:10.1073/pnas.0404640101

OhAinle M, Kerns JA, Malik HS, Emerman M. 2006. Adaptive Evolution and Antiviral Activity of the Conserved Mammalian Cytidine Deaminase APOBEC3H. *J Virol* **80**:3853–3862. doi:10.1128/JVI.80.8.3853-3862.2006

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, Schneider MPC, Silva A, O'Brien SJ, Pecon-Slattery J. 2011. A Molecular Phylogeny of Living Primates. *PLoS Genet* **7**:e1001342. doi:10.1371/journal.pgen.1001342

Petrov DA. 2003. Size Matters: Non-LTR Retrotransposable Elements and Ectopic

Recombination in Drosophila. *Mol Biol Evol* **20**:880–892. doi:10.1093/molbev/msg102

Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, Kaessmann H. 2008. Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol*. doi:10.1371/journal.pbio.0060080

Refsland EW, Stenglein MD, Shindo K, Albin JS, Brown WL, Harris RS. 2010. Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucleic Acids Res* **38**:4274–4284. doi:10.1093/nar/gkq174

Rosso L, Marques AC, Weier M, Lambert N, Lambot M-A, Vanderhaeghen P, Kaessmann H. 2008. Birth and Rapid Subcellular Adaptation of a Hominoid-Specific CDC14 Protein. *PLoS Biol* **6**:e140. doi:10.1371/journal.pbio.0060140

Sayah DM, Sokolskaja E, Berthoux L, Luban J. 2004. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**:569–573. doi:10.1038/nature02777

Schmitz J, Brosius J. 2011. Exonization of transposed elements: A challenge and opportunity for evolution. *Biochimie* **93**:1928–1934. doi:10.1016/j.biochi.2011.07.014

Silvas T V., Schiffer CA. 2019. APOBEC3s: DNA-editing human cytidine deaminases. *Protein Sci* pro.3670. doi:10.1002/pro.3670

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. http://www.repeatmasker.org

Song M, Boissinot S. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. *Gene* **390**:206–213. doi:10.1016/j.gene.2006.09.033

Sotero-Caio CG, Platt RN, Suh A, Ray DA. 2017. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. *Genome Biol Evol* **9**:161–177. doi:10.1093/gbe/evw264

Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, Diana E, Lehmann G, Toren D, Wang J, Fraifeld VE, de Magalhães JP. 2018. Human Ageing Genomic Resources: new and updated databases. *Nucleic Acids Res* **46**:D1083–D1090.

doi:10.1093/nar/gkx1042

Thomas GWC, Wang RJ, Puri A, Harris RA, Raveendran M, Hughes DST, Murali SC, Williams LE, Doddapaneni H, Muzny DM, Gibbs RA, Abee CR, Galinski MR, Worley KC, Rogers J, Radivojac P, Hahn MW. 2018. Reproductive Longevity Predicts Mutation Rates in Primates. *Curr Biol* **28**:3193-3197.e5. doi:10.1016/j.cub.2018.08.050

Turelli P, Vianin S, Trono D. 2004. The Innate Antiretroviral Factor APOBEC3G Does Not Affect Human LINE-1 Retrotransposition in a Cell Culture Assay. *J Biol Chem* **279**:43371–43373. doi:10.1074/jbc.C400334200

Uchimura A, Higuchi M, Minakuchi Y, Ohno M, Toyoda A, Fujiyama A, Miura I, Wakana S, Nishino J, Yagi T. 2015. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Res* **25**:1125–1134. doi:10.1101/gr.186148.114

Virgen CA, Kratovac Z, Bieniasz PD, Hatziioannou T. 2008. Independent genesis of chimeric TRIM5-cyclophilin proteins in two primate species. *Proc Natl Acad Sci* **105**:3563–3568. doi:10.1073/pnas.0709258105

Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. 2007. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci* **104**:18613–18618. doi:10.1073/pnas.0703637104

Wang W, Brunet FG, Nevo E, Long M. 2002. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proc Natl Acad Sci* **99**:4448–4453. doi:10.1073/pnas.072066399

Weaver S, Shank SD, Spielman SJ, Li M, Muse S V, Kosakovsky Pond SL. 2018. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Mol Biol Evol* **35**:773–777. doi:10.1093/molbev/msx335

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. 2015. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. *Mol Biol Evol* **32**:820–832.

doi:10.1093/molbev/msu400

Wilson SJ, Webb BLJ, Ylinen LMJ, Verschoor E, Heeney JL, Towers GJ. 2008. Independent evolution of an antiviral TRIMCyp in rhesus macaques. *Proc Natl Acad Sci* **105**:3557–3562. doi:10.1073/pnas.0709003105

Wissing S, Muñoz-Lopez M, Macia A, Yang Z, Montano M, Collins W, Garcia-Perez JL, Moran J V, Greene WC. 2012. Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility. *Hum Mol Genet* **21**:208–218. doi:10.1093/hmg/ddr455

Wong SK, Connole M, Sullivan JS, Choe H, Carville A, Farzan M. 2009. A New World Primate Deficient in Tetherin-Mediated Restriction of Human Immunodeficiency Virus Type 1. *J Virol* **83**:8771–8780. doi:10.1128/JVI.00112-09

Xie Y, Rosser JM, Thompson TL, Boeke JD, An W. 2011. Characterization of L1 retrotransposition with high-throughput dual-luciferase assays. *Nucleic Acids Res* **39**:e16–e16. doi:10.1093/nar/gkq1076

Yamashita M, Emerman M. 2004. Capsid Is a Dominant Determinant of Retrovirus Infectivity in Nondividing Cells. *J Virol* **78**:5670–5678. doi:10.1128/JVI.78.11.5670-5678.2004

Yan Y, Buckler-White A, Wollenberg K, Kozak CA. 2009. Origin, antiviral function and evidence for positive selection of the gammaretrovirus restriction gene Fv1 in the genus Mus. *Proc Natl Acad Sci U S A* **106**:3259–3263. doi:10.1073/pnas.0900181106

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**:1586–1591. doi:10.1093/molbev/msm088

Young GR, Yap MW, Michaux JR, Steppan SJ, Stoye JP. 2018. Evolutionary journey of the retroviral restriction gene *Fv1*. *Proc Natl Acad Sci* **115**:10130–10135. doi:10.1073/pnas.1808516115

Fig1

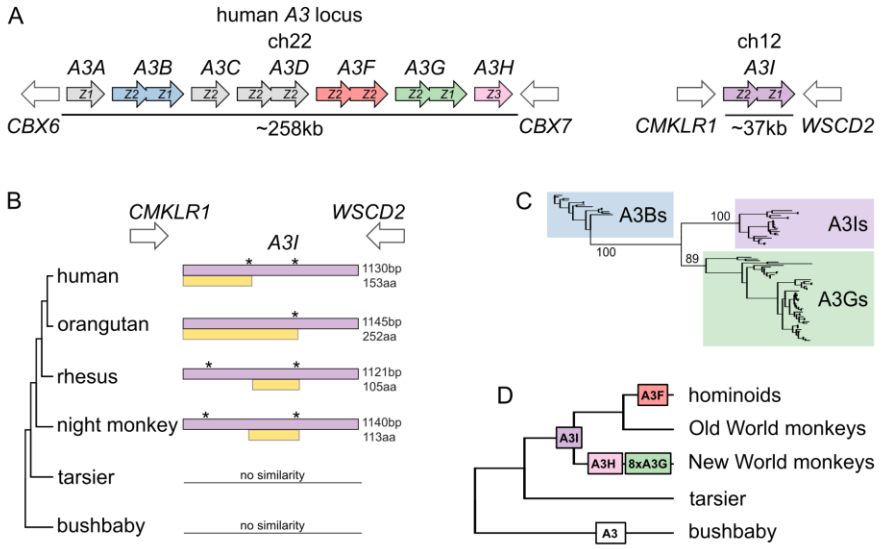

A
human A3 locus
ch22
A3A
A3B
A3C
A3D
A3F
A3G
A3H
CBX6
~258kb
CBX7
ch12
A3I
CMKLR1
~37kb
WSCD2
B
CMKLR1
WSCD2
A3I
human
orangutan
rhesus
night monkey
tarsier
bushbaby
1130bp
153aa
1145bp
252aa
1121bp
105aa
1140bp
113aa
no similarity
no similarity
C
A3Bs
A3Is
A3Gs
100
100
89
D
A3F
A3I
A3H
8xA3G
A3
hominoids
Old World monkeys
New World monkeys
tarsier
bushbaby

Fig2

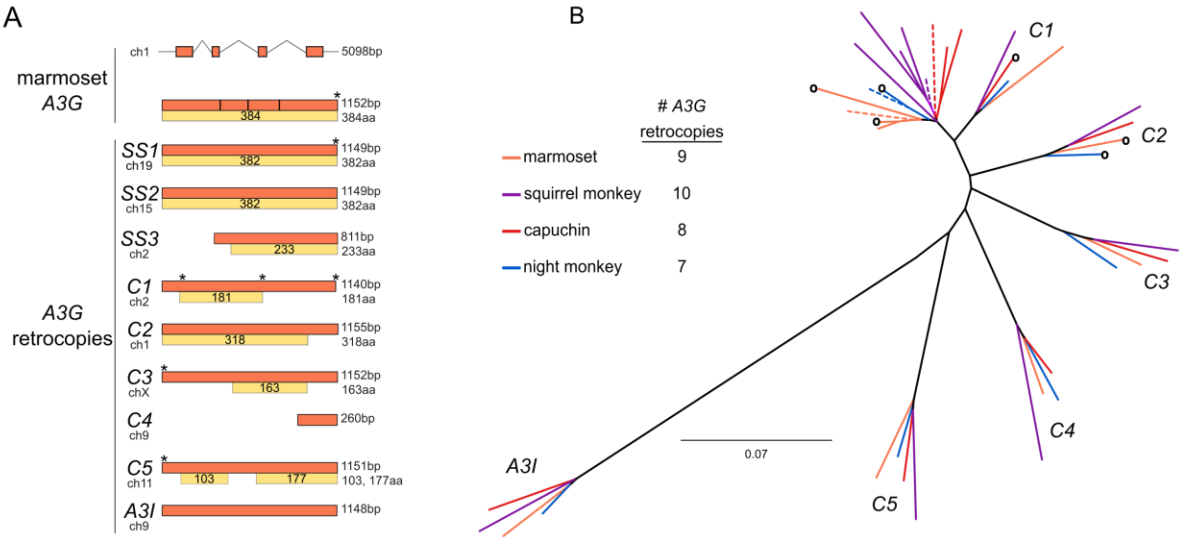

A
ch1
5098bp
marmoset
A3G
1152bp
384
384aa
SS1
ch19
1149bp
382
382aa
SS2
ch15
1149bp
382
382aa
SS3
ch2
811bp
233
233aa
C1
ch2
1140bp
181
181aa
A3G
retrocopies
C2
ch1
1155bp
318
318aa
C3
chX
1152bp
163
163aa
C4
ch9
260bp
C5
ch11
1151bp
103
177
103, 177aa
A3I
ch9
1148bp
B
A3G
retrocopies
marmoset 9
squirrel monkey 10
capuchin 8
night monkey 7
C1
C2
C3
C4
C5
A3I
0.07

Fig3

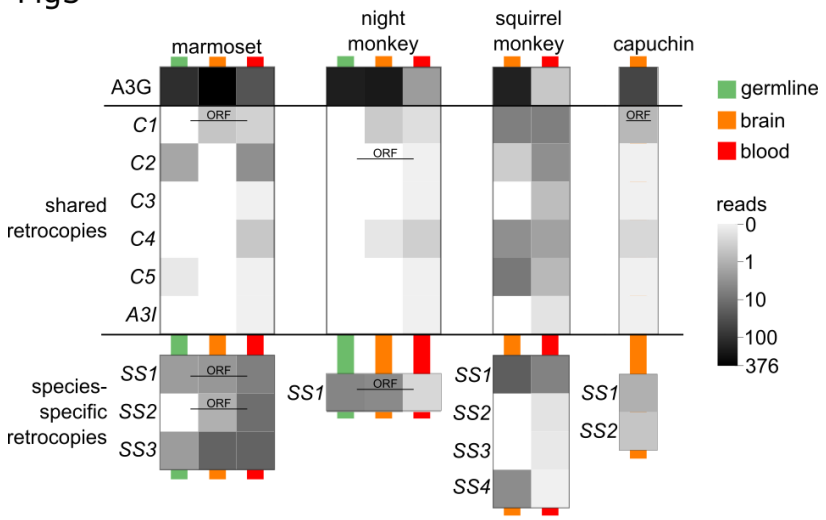

marmoset
night monkey
squirrel monkey
capuchin
A3G
C1
C2
C3
C4
C5
A3I
shared retrocopies
species-specific retrocopies
SS1
SS2
SS3
SS4
ORF
germline
brain
blood
reads
0
1
10
100
376

Fig4

		N-terminal deaminase motif		C-terminal deaminase motif	
		HXE	CXXC	HXE	CXXC
	human A3G	HPE	CTKC	HAE	CFSC
	rhesus A3G	HPE	CTRC	HAE	CFSC
	marmoset A3G	HPE	CPVC	HAE	CFCC
	squirrel monkey A3G	HPE	CPVC	HAE	CFSC
	owl monkey A3G	HPE	CPVC	HAE	CFSC
retrocopies	marmoset-SS1	HPE	CPVC	HAE	CFSC
retrocopies	marmoset-SS2	NPE	CPVC	HAE	CFSC
retrocopies	marmoset-C1	HPE	CPVC	HAE	CFSC
retrocopies	night monkey-C1	HPE	CPVC	HAE	CFSC
retrocopies	night monkey-SS1	HPE	CPVC	HAE	CFSC
retrocopies	capuchin-C1	HPE	SPVC	HAE	CFSC

Fig5

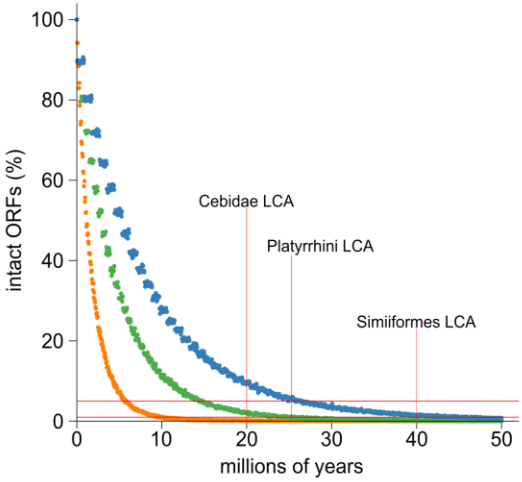
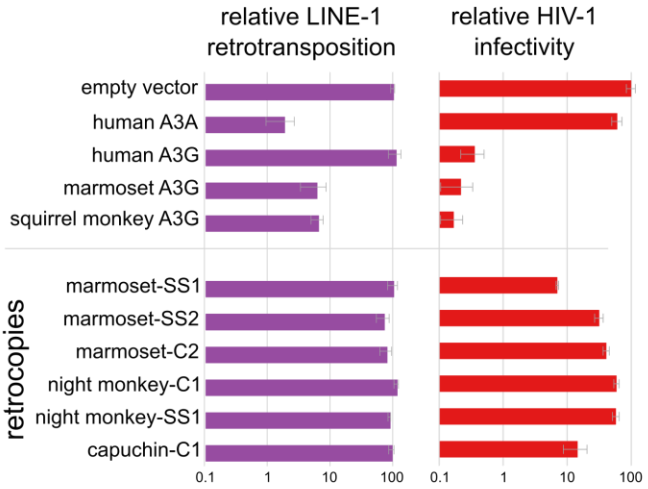

100
80
60
40
20
0
intact ORFs (%)
Cebidae LCA
Platyrrhini LCA
Simiiformes LCA
0
10
20
30
40
50
millions of years

Fig6


relative LINE-1 retrotransposition
relative HIV-1 infectivity
empty vector
human A3A
human A3G
marmoset A3G
squirrel monkey A3G
retrocopies
marmoset-SS1
marmoset-SS2
marmoset-C2
night monkey-C1
night monkey-SS1
capuchin-C1
0.1
1
10
100
0.1
1
10
100