Learning from the past: Searching for novel *TRIM*, *CypA*, and *TRIMCyp* antiviral factors in primates

Raymond Malfavon-Borja

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Harmit S. Malik, Chair

Michael Emerman

Adam P. Geballe

Program Authorized to Offer Degree:

Department of Genome Sciences

University of Washington

Abstract

Learning from the past: Searching for novel *TRIM*, *CypA*, and *TRIMCyp* antiviral factors in primates

Raymond Malfavon-Borja

Chair of the Supervisory Committee:
Affiliate Assistant Professor Harmit S. Malik
Department of Genome Sciences

The evolutionary history and genetic composition of mammals has been strongly influenced by viruses. This is reflected by evolved mechanisms of host defense mediated by restriction factors that are in an arms race to win over recurrent viral pressure. Restriction factors demonstrate genetic innovation, observed in forms such as positive selection and recurrent births of novel antiviral genes, that serves as a beacon to signify and study this arms race. In this dissertation, I explore these signals of genetic innovation to identify new restriction factors and extrapolate from them insights into the history of host-virus interactions. I first describe an ancient antiviral *TRIM5-CyclophilinA* gene fusion, termed *TRIMCypA3,* which likely protected primate ancestors 43 million years ago, but has since decayed. I then present an analysis of the primate *TRIM* multigene family and highlight members displaying signatures of positive selection, which represent novel restriction factor candidates. Amongst these, I focus on *TRIM52* that demonstrates a unique genetic innovation in the RING domain, suggesting a novel recognition domain. Finally, I present an additional analysis of primate genomes designed to catalogue *CypA* retrogenes and explore their evolutionary history, given that a pilot exploration of *CypA* retrogenes led to the discovery of *TRIMCypA3*. A systematic examination has highlighted several other retrogene copies with diverse evolutionary histories suggesting both preservation and innovation. The genetic innovation explored in this dissertation has highlighted several restriction factor candidates amongst the *TRIM* gene family and numerous *CypA* retrogenes encoded within primate genomes that has set a foundation to discover novel antiviral genes.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to humbly thank my PhD advisor, Harmit Malik. The opportunity to learn and develop in his laboratory and under his supervision has been an honor and privilege that is without parallels. The personal and professional admiration that Harmit has emphatically earned amongst his peers is an honest reflection of his character. I am grateful for his approach to mentoring and for teaching me the value of being deliberate. I am also grateful for the assembly of brilliant individuals that Harmit attracted to his lab. The storied and continued success of the Malik lab has been and is still proportional to the quality and intellect of the individuals that make up the lab. While I am fortunate to have been surrounded by many extraordinary lab mates, I would to particularly acknowledge the following people: Eric Smith, Nels Elde, Nitin Phadnis, Matt Daugherty, Maulik Patel, SaraH Zanders, Mia Levine, Rick McLaughlin, Janet Young, Kevin Roach, Ben Ross, Pat Mitchell, Emily Baker, Aimee Littleton, Michael Eickbush, and Aida de la Cruz. I would like to separately acknowledge and thank Adriana Ludwig.

I am also tremendously thankful for a supportive co-advisor, Michael Emerman. Thank you for welcoming me into your lab. I am grateful and appreciative of your unique mentoring and communication style that has taught me to be a more considerate and direct researcher. I would also like to thank the past and present members of the Emerman lab, in particular: Tsai-Yu Lin, Semih Tareen, Melody Li, Nisha Duggal, Efrem Lim, and Lily Wu.

I would like to acknowledge and thank the members of my committee for their guidance and support: Josh Akey, Willie Swanson, Evan Eichler, and Adam Geballe. I am appreciative of the useful and insightful discussions with Adam Geballe, Evan Eichler, and Willie Swanson. I would also like to thank Josh Akey for providing me with resources to contribute to my computational toolbox, as well as amusing and engaging conversations.

My graduate career was filled with many friends that entered my life for a reason, a season, or a lifetime. I thank them for making Seattle a home. I would like to particularly acknowledge Kat Claw, Jackie Martinez, and Ceci Martinez-Vasquez. I would like to separately acknowledge Dan Skelly and Cailyn Spurrell. Both are amazing friends and colleagues. I would also like to separately thank Vicky Yan and her family. I am grateful for being welcomed into their family and home.

Finally, I would like to thank my family. It is because of their support and guidance that I pursued a graduate degree. It is because of their continued support, guidance, and gift packages that I was able to complete my graduate degree.

My research was supported by the Graduate Opportunities and Minority Achievement Program (GO-MAP) Fellowship, Genomics Outreach for Minorities (GenOM) Project Fellowship, Ruth L. Kirschstein National Research Service Awards for Individual Predoctoral Fellowship to Promote Diversity in Health-Related Research, and NIH Interdisciplinary Training Grant Fellowship.

Chapter 1

Introduction

Viruses are an unavoidable burden. This is evident by the rich documented history of viral infections that has been recorded on modern-day infections, as well as records dating back 100s and 1000s of years (Barquet and Domingo 1997, Riedel 2005). Along with this written record, host genomes contain evidence of ancient viral infections that can be detected via evolutionary analysis of host-encoded genes. This process is termed indirect paleovirology and has focused on detecting genetic innovation of host antiviral genes to infer selective pressure from ancient viruses (paleoviruses) (Patel, Emerman et al. 2011). In this dissertation (Chapters 2-4), I describe several forms of genetic innovation in host genes that indicate ancient viral infections and explore the implications of these signals.

Although records of paleoviruses are not likely to be found in fossil layers of rock or captured in amber, their existence is occasionally recorded in the genomes of their animal hosts, revealed by direct paleovirology. Remnants of viral genomes found in host genomes are referred to as Endogenous Viral Elements (EVEs) (Katzourakis 2010). The first EVEs to be characterized were of tumor-associated retroviruses in the 1960s (Weiss 2006). Endogenization of these viral genomes (proviruses) is due to an obligatory integration stage of the viral replication lifecycle involving virus-encoded reverse transcriptase and integrase genes (Figure 1.1). Retroviral-EVEs can be found in all mammalian genomes and comprise ~8% of the human genome (Griffiths 2001). Since the discovery of these first retroviral-EVEs and the development of genome sequencing technologies, the endogenized genomes of both retrovirus and non-retroviruses have surfaced in animal genomes (Holmes 2011, Horie and Tomonaga 2011).

Finding EVEs of non-retroviruses (e.g. DNA viruses) in animal genomes is unique given that integration is not obligatory (Figure 1.2). Their presence in host genomes is likely due to interactions with processes such as LINE-mediated retrotransposition or non-homologous end joining. If integration has occurred in the host germline, the EVE is able to propagate into subsequent generations as a heritable unit (Holmes 2011, Horie and Tomonaga 2011). The age of an EVE can be calculated based on the presence or absence of orthologous EVEs between closely and distantly related species (Figure 1.2) (Holmes 2011). In specialized cases (i.e. retroviral-EVEs), the homology of the long terminal repeats (LTRs) can be used

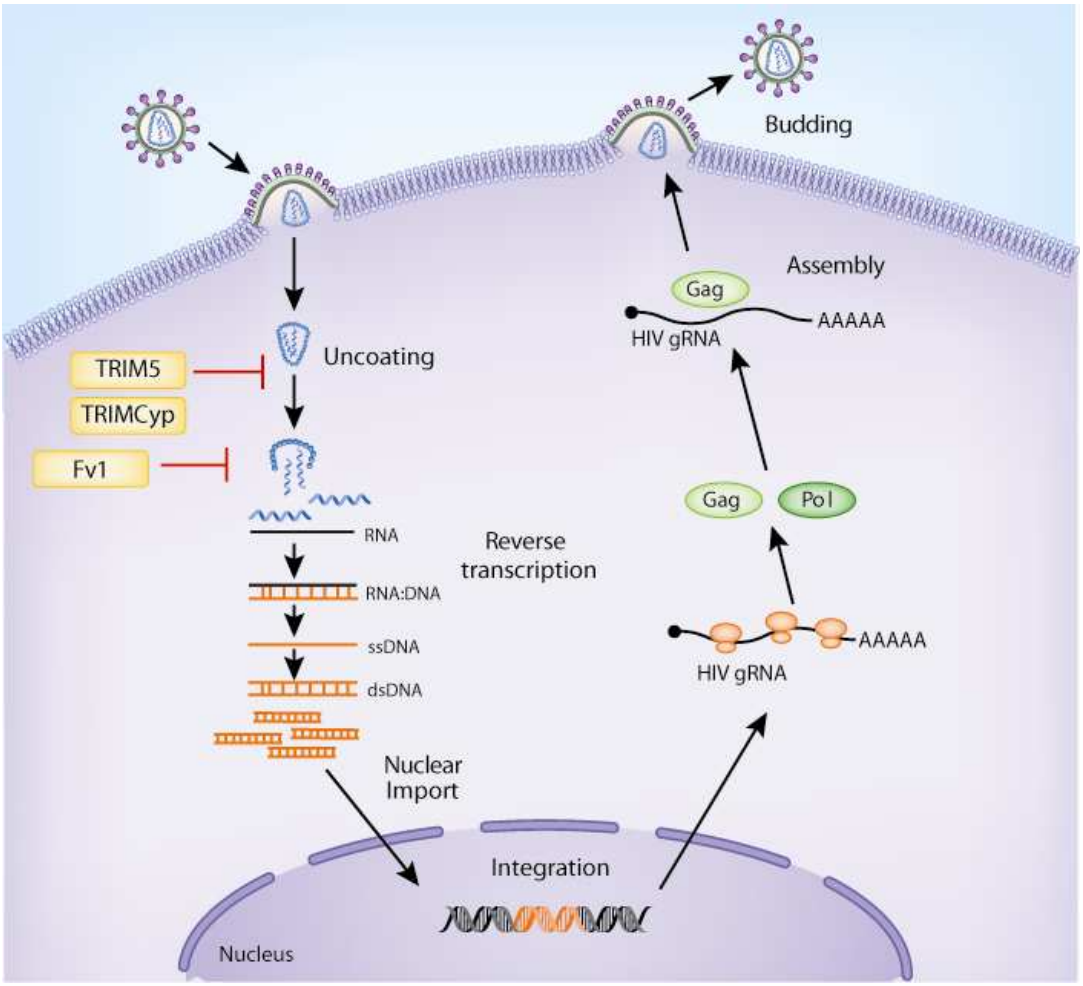


Figure 1.1 Postentry restrictions within the retroviral lifecycle. Primate TRIM5 and related TRIMCyp restrict lentiviruses postentry at the stage of uncoating (Diaz-Griffero, Vandegraaff et al. 2006, Stremlau, Perron et al. 2006). Mus Fv1 similarly restricts postentry, though further downstream to inhibit nuclear import (Jolicoeur and Rassart 1980). This figures is from (Yan and Chen 2012) with permission (License Number: 3080370479841).

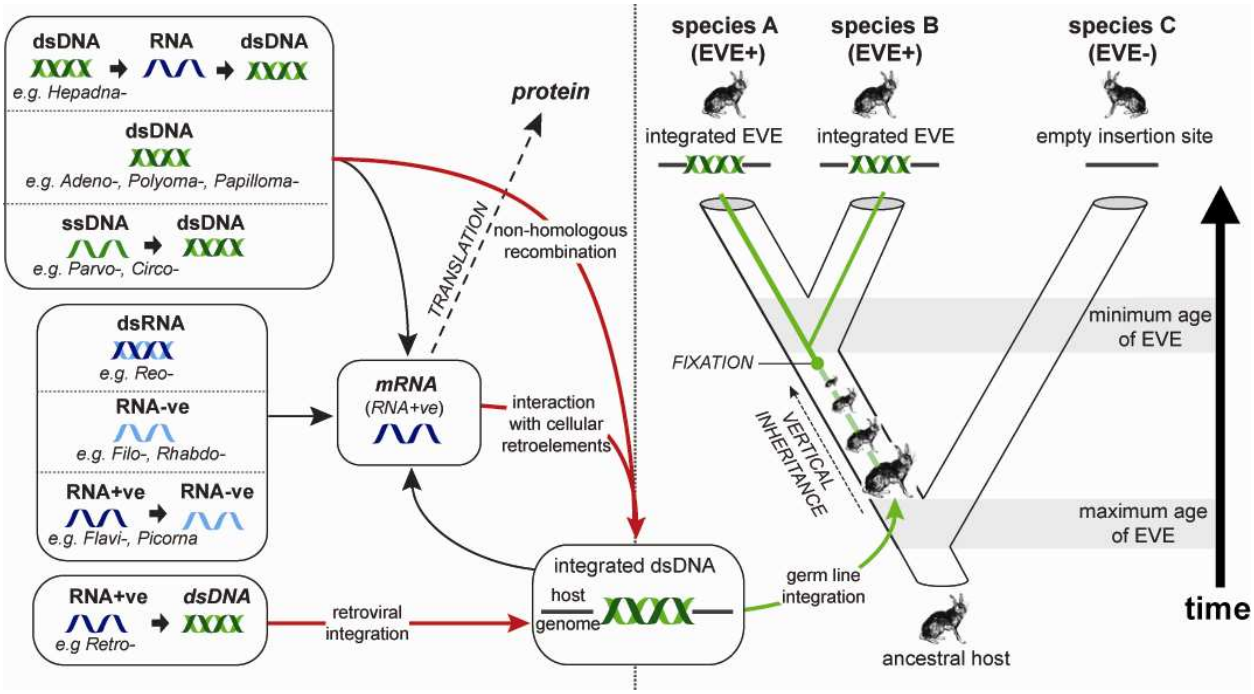


Figure 1.2 History of diverse viral infections documented by endogenous viral elements (EVEs) in host genomes. Despite distinct replication strategies, viral fossils (EVEs) of DNA and RNA have been found within the genomes of mammalian hosts. The date of fixation serves as a proxy for dating the ancient infections (host-virus interactions) and can be decoded by evaluating orthologous EVEs in closely and distantly related species. This figure is from (Katzourakis 2010).

to determine the date of integration based on the host neutral substitution rate (Katzourakis, Tristem et al. 2007). Based on these methods, the age of a virus fossil in an animal genome can be reliably determined and has revealed a deep and rich history of viruses (e.g. Bornaviruses, Hepadnaviruses, and Lentiviruses) (Holmes 2011, Horie and Tomonaga 2011, Gifford 2012).

1.1 History of lentiviruses

The human immunodeficiency virus (HIV), causative of acquired Immunodeficiency syndrome (AIDS), is a relatively new pathogen that entered the public awareness during the early/mid 1980s (www.aids.gov). However, sampling and phylogenetic analysis of tissue samples preserved in 1960 from Kinshasa, Democratic Republic of the Congo support HIV circulating in the human population as early as the late 19th century/early 20th century (Worobey, Gemmel et al. 2008). HIV belongs to the family of retroviruses, identified on the genus level as *Lentiviridae*. Retroviruses have a high mutation rate and sequence diversity that confounds reliable dating of viruses (Sharp, Bailes et al. 2000). While the true date that HIV entered the human population is contentious, mounting evidence suggest lentiviruses have infected other mammalian lineages for millions of years.

Mammals harboring lentiviruses can be placed into two major groups: Laurasiatheria and Supraprimates (Gifford 2012). Lentiviruses were initially recovered from ungulates, with the first being recovered from horses (equine), now known as equine infectious anemia virus (EIAV) (Clements and Zink 1996, Leroux, Cadoré et al. 2004). However, the agent infecting horses was not initially characterized as a retrovirus due to tissue culturing-related complications. Instead, the first retrovirus to be identified as “lentivirus” was characterized in sheep and identified as ovine maedi-visna virus (OMVV) (Sigurdsson 1954, Straub 2004). Goats and cattle (bovine) were later found to harbor lentiviruses (Clements and Zink 1996). Also amongst Laurasiatheria, several species of the Felidae family are infected by lentiviruses (Clements and Zink 1996, O'Brien, Troyer et al. 2012). Unique amongst the Laurasiatheria group, evidence of lentiviral infections amongst species of the weasel family derive from the discovery of lentivirus fossils within their genomes (Cui and Holmes 2012, Han and Worobey 2012). The lentivirus-EVEs were initially found due to bioinformatic queries of the ferret (*Mustela putorius furo*) genome and labeled *Mustelidae* endogenous lentivirus of *Mustela putorius furo* (MELVmpf). Closely related species within the weasel family were investigated for the presence or absence of MELVmpf to determine the date of the ancient lentivirus infections by taxonomic distribution and concluded the acquisition occurred 8.8-11.8 million

years ago (Mya). Thus, some lentiviral infections within Laurasiatheria are predicted to be ancient, occurring millions of years ago.

The first lentivirus-EVE was discovered in the European rabbit (*Oryctolagus cuniculus*) genome and labeled rabbit endogenous lentivirus type K (RELIK) (Katzourakis, Tristem et al. 2007). This finding dramatically altered the comprehension of lentivirus, as this was the first time lentiviruses demonstrated the capacity for germline infection and an infectious history in the millions of years range. RELIK orthologs were later found in the European hare (*Lepus europaeus*) and determined to only be found within a subset of species within lagomorphs (Keckesova, Ylinen et al. 2009). Based on the phylogenetic relationship of rabbits and hares, the acquisition of RELIK in lagomorph genomes is predicted to have occurred at least 12 Mya. Supported by the more recent discovery of MELVmpf, lentiviruses appear to have been circulating amongst mammals for at least 12 million years and were capable of more diverse infections than is demonstrated by modern-day exogenous lentiviruses.

Primates are unique as certain lineages are infected by modern-day lentiviruses, while other lineages show evidence of ancient lentiviral infections. Evidence of ancient infections derive from finding two distinct lentivirus-EVEs in Madagascar lemurs, labeled grey mouse lemur and fat-tailed dwarf lemur prosimian immunodeficiency virus (pSIVgml and pSIVfdl) (Gifford, Katzourakis et al. 2008, Gilbert, Maxfield et al. 2009). Remarkably, it was estimated that the two distinct lentivirus-EVEs were acquired around roughly the same time, ~4.2 Mya. Modern, exogenous lentiviruses are not naturally found in prosimian primates. Thus, pSIVgml and pSIVfdl demonstrate that prosimians were at one time infected with lentiviruses, but that the pathogen was overcome and abolished within this lineage of primates. No other extant primate has been found to harbor lentivirus-EVEs within their genome despite the fact that more than 40 different exogenous simian immunodeficiency viruses (SIVs) are currently circulating amongst simian primates (Keele, Jones et al. 2009, Sharp and Hahn 2010, Sharp and Hahn 2011).

Modern-day lentiviral infections of primates demonstrate a host range that is not reflected by lentiviral-EVEs. To date, 4 distinct HIV-1 groups have been identified: M (Major or Main), N (non-M, non-O), O (outlier), and P (Plantier, Leoz et al. 2009, Sharp and Hahn 2010, Sharp and Hahn 2011), which derive from cross-species transmissions of SIV from chimpanzee (*Pan troglodytes*) and gorilla (*Gorilla gorilla gorilla*) (SIVcpz*Ptt* and SIVgor). HIV-2 was characterized several years after HIV-1 and cases to date have been split into groups A-H (Sharp and Hahn 2011). All cases of HIV-2 are thought to derive from cross-

species transmissions from the SIV of sooty mangabey (Cercocebus atys atys) (SIVsmm) (Sharp, Robertson et al. 1995, Santiago, Range et al. 2005, Sharp and Hahn 2011). More than 40 SIVs with natural non-human primate hosts have been discovered since the initial discovery of HIV-1 and HIV-2 (Clements and Zink 1996, Liégeois, Lafay et al. 2009, Worobey, Telfer et al. 2010). This has revealed that cross-species transmissions have also occurred between non-human primates (Charleston and Robertson 2002, Sharp and Hahn 2011). Discoveries of SIVs from primates isolated for ~10,000 years on Bioko Island from mainland Africa has set the current age estimate of SIVs in primates to ~32,000 years based on phylogenetic analysis (Worobey, Telfer et al. 2010). This estimate derives from the comparison of SIVs from island primates to their mainland counterparts. This supports the conclusion that lentivirus infections of primates are not a recent occurrence.

1.2 Host restriction factors

Animals have not been helpless during this long-lived exposure to viral pathogens and have evolved defense mechanisms in response to continued and recurrent viral threats. Complemented by the adaptive immune system, the innate immune system is the front line defense against microbial pathogens. This system is primed to detect non-host, viral ligands (referred to as pathogen-associated molecular pattern or PAMP) via pathogen recognition receptors (PRRs). When triggered, PRRs will initiate a signaling cascade activating interferon and interferon stimulated genes (ISGs) that function to defend an infected cell and directly inhibit viral replication (reviewed in (Wilkins and Gale 2010, Yan and Chen 2012)). Here I will focus on an arm of the innate immune response that encodes for potent antiviral genes, termed here as restriction factors.

Host-encoded restriction factors have several distinctive characteristics. For example, we can distinguish restriction factors from other components of the innate immune response as restriction factors directly interact with viral proteins, with interactions being driven by either the host or virus. In some instances, this direct interaction can lead to successive adaptation by both the virus and host, establishing an evolutionary arms race in which one of the two players (host or virus) adaptively/rapidly evolves to gain an upper hand that then places evolutionary pressure on the "losing" player to subsequently adapt (Figure 1.3A). This scenario was formally described by Leigh Van Valen (Van Valen 1973) via the "Red Queen hypothesis", inspired by Lewis Carroll's 1871 "Through the Looking-Glass, and What Alice Found There." The Red Queen hypothesis describes an evolutionary scenario where two genetic species antagonize each other in a cyclical fashion (Daugherty and Malik 2012). Intriguingly, restriction factors

are themselves almost always targeted and antagonized by a viral protein (viral antagonist) (Duggal and Emerman 2012), leading to an arms race in which the viral antagonist and host-encoded restriction factor successively adapt to antagonize and evade, respectively (Figure 1.3B). We can detect signals of a restriction factor engaged in a host-virus arms race when it exhibits signatures of positive selection (Sawyer, Emerman et al. 2004, Sawyer, Emerman et al. 2007), an enrichment for non-synonymous (amino acid-altering mutations) changes compared to synonymous (silent mutations) in the coding sequence (Yang 1997, Nielsen and Yang 1998, Yang 1998, Suzuki and Gojobori 1999, Yang and Bielawski 2000) (Figure 1.3C). In the context of a host-virus arms race, rapid evolution of the restriction factor is interpreted as the consequence of repeated selective pressure imposed by the virus. This provides a unique opportunity to date selected changes via phylogenetics, thus resolving the date of the selective pressure.

Species-specificity underlies the discovery of a majority of restriction factors. Often, identifying differences in permissiveness versus non-permissiveness of a cell type to viral infection leads to the discovery of a novel restriction factor or extends the functionality of those that are presently established. Differences between orthologous restriction factors of related species can be explained as the consequence of adaptive evolution driven by the host-virus arms race. Therefore, while a bona-fide restriction factor in one species functions to restrict a particular virus in one species, an orthologous restriction factor may not function equally in another species.

Similar to the direct paleovirology approach of investigating animal genomes for the fossil records of viruses, an alternative, indirect methodology focused on restriction factors can also be informative of age and nature of ancient viruses (Patel, Emerman et al. 2011). In this dissertation, I will describe my approaches to paleovirology using a restriction factor-centric approach that focuses on the evolutionary innovations exhibited by presently established restriction factors (e.g., positive selection). I will demonstrate that this approach is able to identify bona-fide restriction factors and generate a source of restriction factor candidates. I will further demonstrate that the characterization of these restriction factors also serves to enrich the history present-day viral pathogens. This serves to cement the utility of evolutionary biology in the understanding of present-day and historical host-virus interactions.

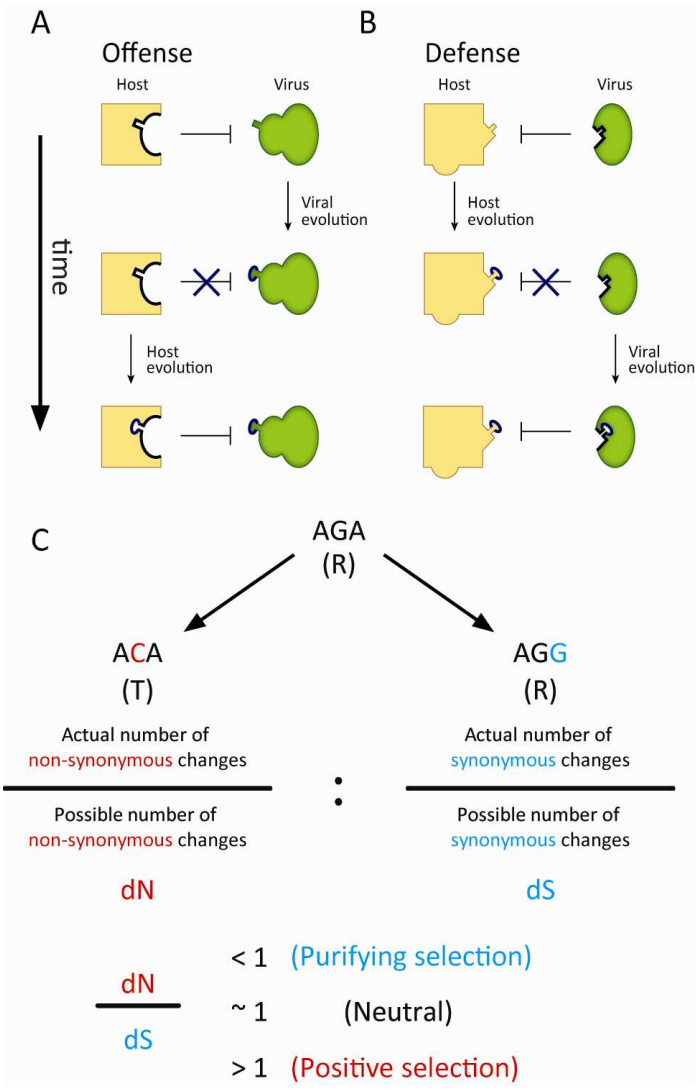


Figure 1.3 Host-virus arms race. A host-virus evolutionary arms race is posed from the interaction of host and virus proteins. (A) A scenario in which the host-encoded protein targets a viral protein for restriction places selective pressure on the virus to evade recognition. If the virus evolves to evade the "offensive" host protein, the virus is acknowledged as "winning" in this snapshot of the arms race and places selective pressure on the host protein to evolve and regain recognition. (B) A scenario in which the virus instigates an interaction places the targeted host protein on the "defensive" and under selective pressure to abrogate this interaction. If this takes place, the host and virus are thus acknowledged as "winning" and "losing", respectively. Selective pressure on the virus directs the protein to regain the interaction and antagonism of the host. (C) The sites of interaction between the host and virus proteins engaged in an arms race may exhibit positive selection, reflected by an enrichment of non-synonymous changes and residue changes compared to synonymous (silent) mutations. The rate of non-synonymous changes (dN) compared to the rate of synonymous changes (dS) generates a ratio that describes the tempo of evolution acting on a gene or domain. When this ratio (dN/dS) is greater than 1, this reflects positive selection. A ratio less than 1 reflects purifying selection (suppression of residue-altering codon changes). When the ratios are statistically indistinct, dN/dS is equivalent to 1 and reflects the gene to be evolving at a neutral rate. This figure is adapted from (Daugherty and Malik 2012).

1.2.1 FV1

The first restriction factor to be characterized was Friend virus susceptibility factor 1 (Fv1) from studies on resistance to Friend murine leukemia virus (MLV) in mice (Lilly 1967). Prototypical Fv1 alleles recovered from National Institutes of Health (NIH) Swiss ($Fv1^N$) and BALB/c ($FV1^B$) mice were soon identified for their differential restriction of MLV strains (Hartley, Rowe et al. 1970). Homozygous $Fv1^N$ mice were susceptible to N-tropic MLV strains, while homozygous $FV1^B$ mice restricted N-tropic MLV infection. Conversely, homozygous $FV1^B$ mice were susceptible to B-tropic MLV strains, while homozygous $Fv1^N$ mice restricted B-tropic MLV infection. Fv1 was later cloned, making it the first restriction factor to be cloned. Unexpectedly, Fv1 was found to have viral origins; it encodes the gag region of a retroviral-EVE (MERV-L) (Best, Le Tissier et al. 1996). Thus, *Mus* genomes utilize a domesticated retroviral gag as a restriction factor. An investigation of Fv1 evolution across the *Mus* genus determined the gene to be rapidly evolving for ~7 million years (Qi, Bonhomme et al. 1998, Yan, Buckler-White et al. 2009). Sites of positive selection were found to overlap with regions and specific residues previously highlighted for impacting or being critical for restriction, and supported a model in which Fv1 directly binds to the capsid (CA) of viruses (Kozak and Chakraborti 1996, Bishop, Bock et al. 2001, Stevens, Bock et al. 2004, Yan, Buckler-White et al. 2009). While restriction by Fv1 was known to occur postentry and pre-integration (Figure 1.1), details to the mechanism have remained elusive. However, it was soon after demonstrated that FV1 directly interacts with the CA protein of the assembled viral core (Hilditch, Matadeen et al. 2011) reinforcing the long-standing model and the conclusions of the evolutionary analysis. Thus, while Fv1 derives from viral origins, the restriction factor has been adaptively evolving in the direct defense of *Mus* species for millions of years.

1.2.2 TRIM5

Non-murine Mammals also exhibit evidence of a FV1-like restriction factor despite not encoding a FV1 gene (Shibata, Sakai et al. 1995, Best, Le Tissier et al. 1996, Himathongkham and Luciw 1996, Hofmann, Schubert et al. 1999, Besnier, Ylinen et al. 2003, Towers, Hatziioannou et al. 2003). The human Fv1-like restriction factor was termed Restriction Factor 1 (REF1) (Towers, Bock et al. 2000) and the factor identified in non-human primates was termed Lentivirus Factor 1 (Lv1) (Cowan, Hatziioannou et al. 2002). REF1 and Lv1 shared many features: restriction occurred postentry and prior to reverse transcription, their viral target was CA, and restriction activity was saturable (Shibata, Sakai et al. 1995, Himathongkham and Luciw 1996, Towers, Bock et al. 2000, Besnier, Takeuchi et al. 2002, Cowan, Hatziioannou et al. 2002, Hatziioannou, Cowan et al. 2003). While the range of restriction differed

between REF1 and Lv1, it was later concluded that the human and non-human primate factors were in fact variants of the same antiviral gene based on competition (cross-abrogation) assays (Besnier, Takeuchi et al. 2002, Cowan, Hatziioannou et al. 2002, Hatziioannou, Cowan et al. 2003). For example, an African green monkey (Agm) cell line was sequentially treated with two viruses known to be restricted by that cell line (e.g., SIVmac and HIV-1). Upon treatment of the second virus, investigators found that Agm cells were incapable of additional restriction activity. Based on the known restriction limitations imposed by saturation kinetics, it was concluded that Lv1 was the sole factor responsible for restriction of the two viruses. Similar cross-abrogation experiments and conclusions were made for human REF1, supporting REF1 and Lv1 being the same factor. Shortly after, a screen of a rhesus macaque cDNA library identified *TRIM5* as the gene encoding Lv1 activity (Stremlau, Owens et al. 2004). This was immediately followed by work to confirm that human REF1 and other primate Lv1 were also in fact *TRIM5* (Hatziioannou, Perez-Caballero et al. 2004 Cowan, and Bieniasz 2004, Keckesova, Ylinen et al. 2004, Perron, Stremlau et al. 2004, Yap, Nisole et al. 2004, Song 2005 Park, Stremlau, Sodroski 2005). To date, the antiviral activity of *TRIM5* has further been extended and documented amongst closely related homologs belonging to glires (Schaller, Hue et al. 2007, Tareen, Sawyer et al. 2009, Fletcher, Hué et al. 2010) and cows (Si, Vandegraaff et al. 2006, Ylinen, Keckesova et al. 2006). Later work showed that *TRIM5* was well conserved in mammalian genomes, with the exception of cat and dog genomes, where the *TRIM5* gene had undergone independent pseudogenization/loss events (Sawyer, Emerman et al. 2007, McEwan, Schaller et al. 2009)

TRIM5 is a member of the *TRIM* multigene family, which encodes more than 70 genes in humans and is similarly expansive throughout primates (Reymond, Meroni et al. 2001, Han, Lou et al. 2011). Members of the *TRIM* multigene family are characterized by a tripartite motif consisting of a RING (Really Interesting New Gene) domain, one or two B-Boxes, and a Coiled-Coil motif, the order and spacing of which are generally conserved (Meroni and Diez-Roux 2005, Nisole, Stoye et al. 2005). The RING domain encodes a zinc finger motif and is associated with E3 ubiquitin ligase activity (Ikeda 2000, Meroni and Diez-Roux 2005, Pertel, Hausmann et al. 2011). The B-Box domain is another zinc binding motif that is unique to *TRIM* genes and is responsible for higher-order assembly of TRIM5 dimers (Li and Sodroski 2008, Li, Yeung et al. 2011). Dimerization (lower-order oligomerization) of TRIM5 is facilitated by the Coiled-Coil domain (Kar, Diaz-Griffero et al. 2008, Langelier, Sandrin et al. 2008). An additional C-terminal domain can be found on most *TRIM* genes, which is used to categorize *TRIM* genes into

classifications denoted C-I to C-XI (reviewed in (Ozato, Shin et al. 2008, McNab, Rajsbaum et al. 2011)). *TRIM5* and the majority of *TRIM* genes are classified as C-IV and encode a C-terminal B30.2 (PRYSPRY) domain (Sardiello, Cairo et al. 2008). *TRIM5* encodes several isoforms, all of which contain the RING, B-Box, and Coiled-Coil domains (Reymond, Meroni et al. 2001, Battivelli, Migraine et al. 2011). Only the longest of the isoforms, TRIM5α, contains the B30.2 and has antiviral activity (Stremlau, Owens et al. 2004).

Considerable variation in both the Coiled-Coil and B30.2 domains amongst primate *TRIM5* orthologs led to the discovery of ancient positive selection (Sawyer, Wu et al. 2005, Maillard, Ecco et al. 2010). Positive selection in *TRIM5* served to rationalize the species-specific restriction variation observed amongst primates and cemented the gene as involved in genetic conflict. Following the identification of *TRIM5* as a host-encoded factor targeting viral CA, the B30.2 domain gained recognition for regulating restriction (Sayah, Sokolskaja et al. 2004, Javanbakht, Yuan et al. 2006 Sodroski 2006, Li, Li et al. 2006 Lee, Sodroski 2006, Perron, Stremlau et al. 2006, Stremlau, Perron et al. 2006 Sodroski 2006). Significant work went into resolving the physical details of how TRIM5α interacts with CA. A tremendous step was the finding that hexagonal TRIM5α assembles as a lattice directly to the surface of the retroviral core (Ganser-Pornillos, Chandrasekaran et al. 2011). The B30.2 domain was found to be necessary for the higher-ordered TRIM5α structure to associate with CA. The consequence of this interaction is thought to accelerate the uncoating of the CA from the viral core structure and abrogate subsequent replication stages (Stremlau, Perron et al. 2006 Sodroski 2006). As would be predicted from the detection of positive selection, restriction is mediated by variation in the B30.2 domain and viral CA (Sawyer, Wu et al. 2005, Sebastian and Luban 2005, Ohkura, Yap et al. 2006, Kirmaier, Wu et al. 2010, Maillard, Ecco et al. 2010). Based on where functional variation occurred on a phylogeny, we can determine when in evolutionary time selective pressure, from involvement in an arms race with a virus, resulted in those changes (Patel, Emerman et al. 2011, Daugherty and Malik 2012). However, we cannot be sure what the true identity of the viral pressure was for *TRIM5*, although it was most certainly retroviral based on the presently recognized restriction range (Hatziioannou, Perez-Caballero et al. 2004, Keckesova, Ylinen et al. 2004, Perron, Stremlau et al. 2004, Yap, Nisole et al. 2004, Song, Javanbakht et al. 2005, Yap 2008). Thus, as complemented by the viral fossil record contained within primate genomes, *TRIM5* evolved under recurrent and episodic positive selection as a consequence of a host-retrovirus arms race for millions of years of primate history (Sawyer, Wu et al. 2005, Sawyer, Emerman et al. 2007). In Chapter 3,

I present an evolutionary based analysis that identifies some novel examples of putative restriction factors within the *TRIM* multigene family, based on their signatures of positive selection.

The mechanistic details of TRIM5α restriction remain elusive and an active of research. Some confusion arises because TRIM5α antiviral activity manifests in several seemingly distinct routes. Restriction prior to reverse transcription occurs in a proteasome-dependent manner (Figure 1.1) (Anderson, Campbell et al. 2006, Wu, Anderson et al. 2006). Intriguingly, the proteasome serves to degrade TRIM5α, but only when the restriction factor is in the presence of a restriction-sensitive virus (Rold and Aiken 2008). Inhibition of the proteasome by MG132 treatment reveals a second restriction pathway that permits reverse transcription of the viral genome, but inhibits nuclear entry (Anderson, Campbell et al. 2006, Wu, Anderson et al. 2006). Restriction by TRIM5α also occurs via an indirect route through the activation of TGF-activated kinase 1 (TAK1) that leads to the activation of the NF-κB and AP-1, promoting downstream innate immunity signaling (Pertel, Hausmann et al. 2011, Tareen and Emerman 2011). Due to the E3 ubiquitin ligase activity of the RING domain, TRIM5α generates unattached, K63-linked ubiquitin chains that activate TAK1. This activity is amplified when TRIM5α interacts with the CA of the assembled viral core. Therefore, TRIM5α is able to initiate a signaling cascade that leads to restrictive conditions in an infected cell. Thus, TRIM5α restriction occurs via direct and indirect pathways.

1.2.3 CyclophilinA, an unconventional restriction factor

While the interaction of TRIM5α and CA results in restriction, the interaction of host-encoded CyclophilinA (*CypA*) with CA has the potential for both viral restriction and enhancement. *CypA* encodes peptidyl-prolyl isomerase activity catalyzing the *cis-trans* isomerization of proline residues within peptides (Takahashi, Hayano et al. 1989). The role of *CypA* in the lentiviral lifecycle was initially recognized as a positive factor for HIV-1, resulting in an increase in infectivity (Thali, Bukovsky et al. 1994). *CypA* was found to target the proline at position 90 (P90) of HIV-1 CA protein, facilitating a conformational change of the proline and disassembly of the viral core (Franke, Yuan et al. 1994, Braaten, Aberham et al. 1996, Gamble, Vajdos et al. 1996, Gitti, Lee et al. 1996). However, this CypA-facilitated activity is not necessary for the fulfillment of the lentivirus lifecycle (Thali, Bukovsky et al. 1994, Wiegers, Rutter et al. 1999). In some primate lineages, a retrotransposed copy of *CypA* is linked to the tripartite motif of *TRIM5* (RING, B-Box, and Coiled-Coil domains) (Sayah, Sokolskaja et al. 2004, Ribeiro, Menezes et al. 2005, Brennan, Kozyrev et al. 2008, Newman, Hall et al. 2008, Virgen, Kratovac et al. 2008, Wilson, Webb et al. 2008). The fusion of *TRIM5* and *CypA*, termed *TRIMCyp*, combines the CA

binding ability of CypA with the antiviral effector domains of TRIM5. In these instances, the *CypA* domain structurally and functionally replaces the B30.2 domain (See Chapter 2 for additional details). This fusion generates a potent restriction factor that functions at the same early, postentry stage before reverse transcription as TRIM5α (Figure 1.1) (Hatziioannou, Perez-Caballero et al. 2004, Perron, Stremlau et al. 2004, Stremlau, Owens et al. 2004, Sebastian and Luban 2005, Stremlau, Perron et al. 2006). Remarkably, the discovery of an antiviral *TRIMCyp* gene fusion occurred at nearly the same time that *TRIM5* was identified as the gene encoding REF1 and Lv1 activity (Sayah, Sokolskaja et al. 2004, Stremlau, Owens et al. 2004). Thus, while lentiviruses hijack host-encoded *CypA* to improve fitness, this interaction can also betray lentiviruses and result in robust restriction.

CypA is the prototypic representative of a family of *Cyclophilins*. There are as many as 9 *Cyclophilin* gene family members identified in human that are all characterized by their peptidyl-prolyl isomerase activity (Fischer, Bang et al. 1984, Takahashi, Hayano et al. 1989, Wang and Heitman 2005, Schaller, Ocwieja et al. 2011). *CypA* was initially characterized for its high affinity to cyclosporineA (CsA), an immunosuppressant drug (Handschumacher, Harding et al. 1984). The natural biological role of *CypA* is contentious, and is thought to have a role in several biological processes, such as protein folding, and apoptosis (Stamnes 1992; Matouschek 1995; Ou 2001; Min 2005; Uittenbogaard 1998; Nahreini 2001; Decker 2003; Colgan 2004; Grimim 2007; Helekar 1994; Wang 2005). *CypA* is conserved throughout eukaryotes and demonstrates evidence of evolving under strong purifying selection amongst primates, suggesting a conserved biological role (Ortiz, Bleiber et al. 2006).

The human genome contains copious numbers of *CypA* retrogenes (Haendler and Hofer 1990, Willenbrink, Halaschek et al. 1995, Zhang, Harrison et al. 2003). Retrotransposed genes are viewed as evolutionary dead ends, as they are not expected to transpose with the necessary regulatory elements. However, it is suspected that 20% of retrogenes in the human genome are transcriptionally active (Marques 2005; Vickenbosch 2006). Indeed, several of the *CypA* retrogenes functionally express mRNA (Harrison, Zheng et al. 2005) and the multiple *TRIMCyp* cases (Chapter 2) demonstrate that the *CypA* retrogenes are capable of evolutionary routes other than dead ends. Expanding outside the *TRIMCyp* cases, in Chapter 4, I use the conservation and evolution of *CypA* retrogenes to infer their impact on primate and human evolution.

Chapter 2

Birth, decay, and reconstruction of an ancient *TRIMCyp* gene fusion in primate genome

TRIM5 is a host antiviral gene with an evolutionary history of genetic conflict with retroviruses. The *TRIMCyp* gene encodes a protein fusion of TRIM5 effector domains with the capsid-binding ability of a retrotransposed *CyclophilinA* (*CypA*), resulting in novel antiviral specificity against lentiviruses. Previous studies have identified two independent primate *TRIMCyp* fusions that evolved within the past 6 million years (My). Here, we describe an ancient primate *TRIMCyp* gene (that we call *TRIMCypA3*), which evolved in the common ancestor of simian primates 43 million years ago (Mya). Gene reconstruction shows that *CypA3* encoded an intact, likely active, *TRIMCyp* antiviral gene, which was subject to selective constraints for at least 10 My, followed by pseudogenization or loss in all extant primates. Despite its decayed status, we found *TRIMCypA3* gene fusion transcripts in several primates. We found that the reconstructed "newly born" *TrimCypA3* encoded robust and broad retroviral restriction activity but that this broad activity was lost via eight amino acid changes over the course of the next 10 My. We propose that *TRIMCypA3* arose in response to a viral pathogen encountered by ancestral primates but was subsequently pseudogenized or lost due to a lack of selective pressure. Much like imprints of ancient viruses, fossils of decayed genes, such as *TRIMCypA3*, provide unique and specific insight into paleoviral infections that plagued primates deep in their evolutionary history.

2.1 Introduction

Ancient viruses have selected for changes in host antiviral genes throughout primate evolution (Emerman and Malik 2010, Patel, Emerman et al. 2011). Understanding when these adaptive changes occurred, together with how they altered the antiviral specificities of these genes, can lead to strong inferences about the existence of ancient viruses and their consequences on the modern function and specificity of the primate innate immune system. For example, although the TRIM5α protein encodes a retroviral restriction factor that blocks the viral life cycle of several retroviruses (Stremlau, Owens et al. 2004, Song 2005, Zhang 2006, Yap 2008), retroviral specificity varies among primates as a result of ancient selection for changes in antiviral specificity (Sawyer, Wu et al. 2005, Li, Li et al. 2006, Kaiser, Malik et al. 2007, Kirmaier, Wu et al. 2010). These species-specific differences in TRIM5α are due to dramatic variation in both the Coiled-Coil and B30.2 domains, which are responsible for the interaction with the viral capsid protein of a variety of retroviruses (Sebastian and Luban 2005, Maillard, Ecco et al. 2010). Thus, innovation for capsid-binding specificity has directly resulted in rapid changes in TRIM5α. An additional form of genetic innovation in the *TRIM5* locus involves novel gene fusions. Such a gene fusion was first identified in owl monkeys (*Aotus trivirgatus*), which encode a fusion protein between the *TRIM5* gene and a retrotransposed *CyclophilinA* (*CypA1*) gene, called the *TRIMCyp* gene fusion (Sayah, Sokolskaja et al. 2004). The retrotransposition of *CypA* between exons 7 and 8 of owl monkey *TRIM5* (Sayah, Sokolskaja et al. 2004) occurred 4.5–6 Mya (Ribeiro, Menezes et al. 2005, Perelman, Johnson et al. 2011). Like TRIM5α, the resulting TRIMCyp protein contains RING, B-Box 2, and Coiled-Coil domains. However, a CypA domain has structurally and functionally replaced the B30.2 domain as the capsid-binding determinant (Thali, Bukovsky et al. 1994, Lin and Emerman 2006). The resulting fusion of TRIM5 effector domains with the capsid-binding ability of CypA in owl monkeys generated a protein with novel antiviral defense activity against HIV-1 (Nisole, Lynch et al. 2004, Sayah, Sokolskaja et al. 2004). This restriction occurs at the same early, postentry stage before reverse transcription as TRIM5α (Hatziioannou, Perez-Caballero et al. 2004, Perron, Stremlau et al. 2004, Stremlau, Owens et al. 2004, Sebastian and Luban 2005, Stremlau, Perron et al. 2006).

Several macaque species also encode *TRIMCyp*, which is the consequence of another, independent *CypA* retrotransposition (*CypA2* retrogene) downstream of the *TRIM5* gene (Brennan, Kozyrev et al. 2008, Newman, Hall et al. 2008, Virgen, Kratovac et al. 2008, Wilson, Webb et al. 2008). This event is also estimated to have occurred 5–6 Mya (Dietrich, Jones-Engel et al. 2010). Unlike *CypA1* in owl monkeys, the *CypA2*-encoding *TRIM5* allele is found at varying frequencies across macaque species (Brennan,

Kozyrev et al. 2008, Newman, Hall et al. 2008, Wilson, Webb et al. 2008, Dietrich, Jones-Engel et al. 2010). These two *TRIMCyp* gene fusions thus represent a remarkable case of convergent evolution in the generation of novel antiviral specificity in the *TRIM5* locus.

Here, by reconstructing a detailed evolutionary history of *CypA* retrogenes proximal to *TRIM5* across primates, we find two additional currently pseudogenized *CypA* retrogenes that inserted downstream of the *TRIM5* gene 18–43 Mya in primate evolution. One of these (which we refer to as *CypA3* in keeping with prior nomenclature) is still expressed as a novel *TRIMCyp* gene fusion transcript in several Old World monkeys. Our phylogenetic analyses date the origin of *CypA3* to 43 Mya and find that *TRIMCypA3* was maintained as an intact gene for at least 10 My, making it the most ancient *TRIMCyp* yet identified in primates. Although *CypA3* is decayed in all extant primates and the resulting *TRIMCyp* gene fusion is defective, our evolutionary reconstruction and virological assays suggest that *TRIMCypA3* encoded broad and potent restriction activity following its birth. Our findings reveal that convergent evolution has led to at least four independent *CypA* retrogene insertions proximal to *TRIM5* and, consequently, the formation of *TRIMCyp* at least three independent times in primate history. This further reflects the intense, recurrent pressure imposed by ancient viruses. We posit that the currently inactive *TRIMCypA3* gene fusion represents the fossil remnants of an ancient antiviral innovation that points to a retroviral challenge before the common ancestor of all simian primates. Our study highlights the utility of antiviral gene evolution for the study of paleovirology (Emerman and Malik 2010, Patel, Emerman et al. 2011).

2.2 Results

2.2.1 CypA retrogenes proximal to TRIM5

The *TRIM5* locus in primates consists of four intact *TRIM* genes: *TRIM22*, *TRIM5*, *TRIM34*, and *TRIM6* (Figure 2.1A), as well as a TRI*M* pseudogene, *TRIMP1* (Figure 2.1A, dotted outline). Our analyses revealed the presence of three *CypA* retrogenes proximal to and downstream of *TRIM5* (Figure 2.1A). The most proximal of these, located ~1 kb downstream of *TRIM5* (Figure 2.1A and Figure A.1), is the polymorphic *CypA2* identified in previous studies (Brennan, Kozyrev et al. 2008, Newman, Hall et al. 2008, Stoye 2008, Virgen, Kratovac et al. 2008, Wilson, Webb et al. 2008) but missing from the reference rhesus macaque genome, as it is not fixed within this macaque lineage. We discovered another *CypA*, located ~14 kb downstream of *TRIM5*, which we labeled as *CypA3* (Figure 2.1A and Figure A.1). *CypA3* lies within *TRIMP1*, which is ~40 kb long in the rhesus macaque genome but ~20 kb shorter in the human and chimpanzee genomes. Finally, we discovered *CypA4*, located ~99 kb downstream of *TRIM5*

(Figure 2.1A and Figure A.1) in the rhesus macaque genome. Neither *CypA3* nor *CypA4* was found in the human and chimpanzee genomes (Figure 2.1B and Figure A.1).

To determine if the recurrence of *CypA* retrotranspositions into the *TRIM5* locus was greater than what we would expect from random insertions into the genome, we calculated the probability of finding three independent *CypA* retrogenes within 100 kb of rhesus macaque *TRIM5*. We queried available primate genomes for all *CypA* retrogenes and found over 100 *CypA* retrogenes distributed in the human, chimpanzee, and rhesus macaque genomes, consistent with previous analyses of the human genome (Zhang 2003}. Based on the number of *CypA* retrogenes and their distribution, we found the probability of the three retrogenes in such close proximity to be highly non-random ($P < 0.0233$; Methods). We therefore conclude that some recurrently acting selective pressure must have preserved *CypA* retrogenes within the *TRIM5* locus.

2.2.2 Estimating the age of CypA2, CypA3, and CypA4

We sought to understand the temporal distribution of the *CypA* retrogenes proximal to the *TRIM5* locus among primate species. Using PCR and primers to flanking regions of each retrogene, we genotyped the panel of primate genomes for the presence or absence of *CypA2*, *CypA3*, and *CypA4* (Figure A.1). All *CypA* retrogenes recovered in this analysis were subsequently sequenced to determine their potential to encode a full-length ORF and for phylogenetic analysis. Consistent with previous reports (Newman, Hall et al. 2008, Virgen, Kratovac et al. 2008), we found *CypA2* to be present in lion-tailed macaque (*Macaca silenus*) and pig-tailed macaque (*Macaca nemestrina*) genomes, each generating an ~2.3-kb band (Figure A.1). We found no evidence of *CypA2* outside of macaques. Our results are consistent with a previous study that showed this retrogene is present only within the macaque lineage that arose 5–6 Mya (Newman, Hall et al. 2008).

In contrast to *CypA2*, we found *CypA3* to be present throughout Old World monkeys as well as in gibbons (Figure 2.1B and Figure A.1). *CypA3* primers were not expected to generate a PCR product from human and chimpanzee genomes due to an ~20-kb region deletion in *TRIMP1* that corresponds to the genomic region containing *CypA3* (Figure 2.1A). Results from other hominoids (gorilla and orangutan) suggest that this ~20-kb deletion occurred before the branching of humans and orangutans. We did not observe the presence of *CypA3* in any New World monkey genomes. Investigations of the assembled marmoset genome (WUGSC3.2/calJac3 and GenBank accession no. AC148555) revealed no evidence of

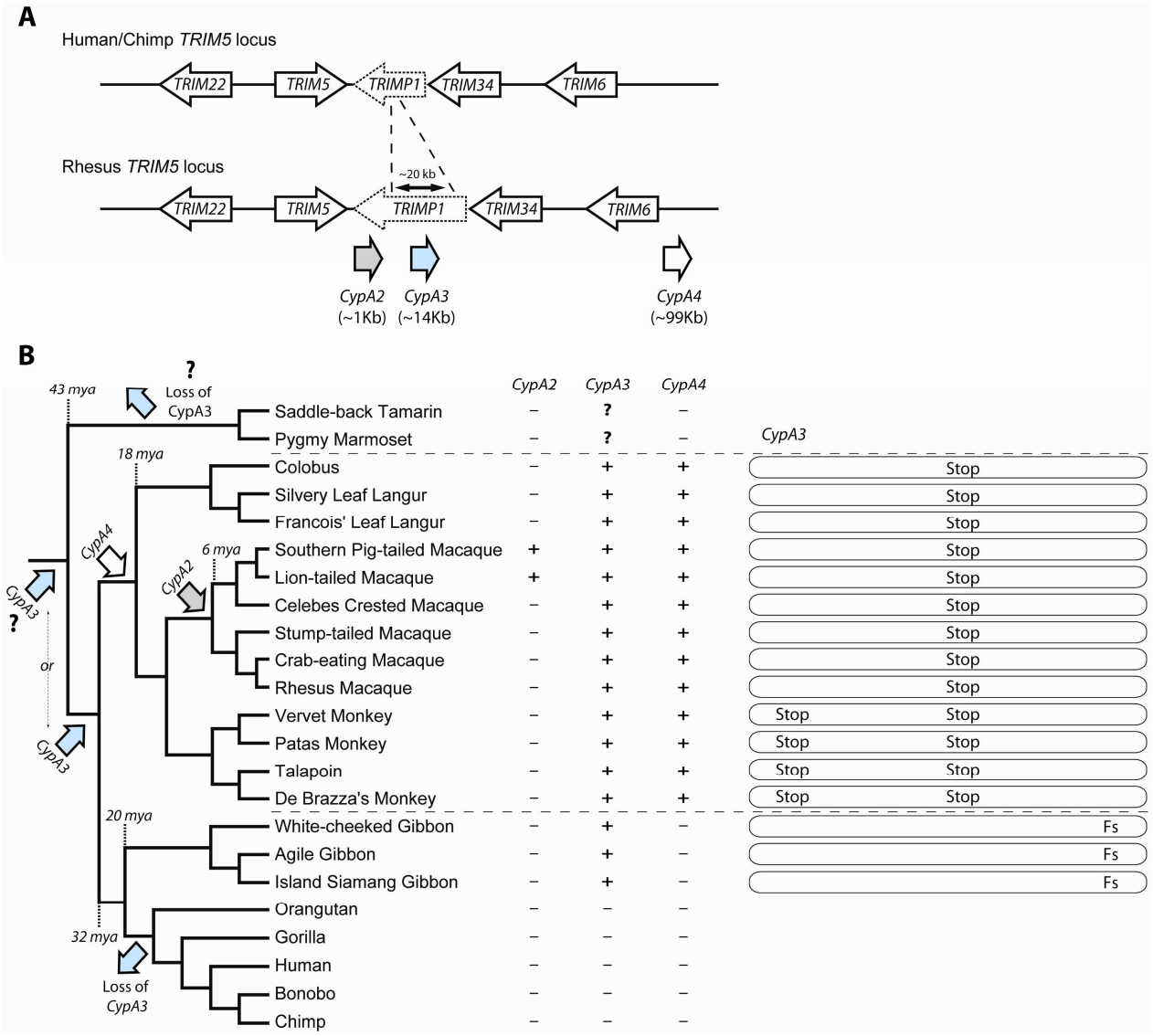


Figure 2.1 Summary of *CypA* retrogenes proximal to *TRIM5*. (A) Relative locations of the *CypA* retrogenes are displayed below the representations of the *TRIM5* locus. *CypA2* (light gray arrow), *CypA3* (light blue arrow), and *CypA4* (white arrow) retrogenes are ~1 kb, ~14 kb, and ~99 kb downstream of *TRIM5*, respectively. The rhesus macaque *TRIMP1* region contains an additional ~20 kb not present in the human and chimpanzee *TRIMP1*. (B) Panel of primates investigated by PCR for *CypA2*, *CypA3*, and *CypA4* is shown in the phylogeny with the notation of retrogene presence or absence indicated to the right. The plus (+) symbol indicates the presence of the *CypA* retrogene. The minus (−) symbol indicates the absence of the *CypA* retrogene. Arrows with labels (Cyp*A*2, *CypA3*, Loss of *CypA3*, and *CypA4*) indicate the point at which the retrogene was acquired or lost in primate evolution. In our analysis, we recovered *CypA2* from southern pig-tailed and lion-tailed macaques. However, given previous reports of the origin and spread of *CypA2* (28), we could place the date of its acquisition at the root of the macaque lineage. *CypA3* sequences, along with pseudogenizing mutations, are represented for those primates found to encode the retrogene. Stop and Fs denote a stop codon and a frameshift mutation in the *CypA3* sequence, respectively.

CypA3 within the ~20-kb stretch between *TRIM5* and *TRIM34* (Figure A.2). Additional searches of another New World monkey, Nancy Ma's night monkey (*Aotus nancymaae*; Genbank accession no. AC183999), also did not reveal the presence of *CypA3* in *TRIMP1*. Therefore, based on orthologous *CypA3* retrogenes in gibbons and Old World monkeys (see below), we can estimate that the ancestral *CypA3* retrogene was acquired in primates at least before the Old World monkey/hominoid split (Figure 2.1B), which occurred 32 Mya (Perelman, Johnson et al. 2011).

Similar assays revealed that *CypA4* is present in all Old World monkeys assayed but not outside this clade (Figure 2.1B and Figure A.1). This suggests that *CypA4* retrotransposed before the common ancestor of Old World monkeys, at least 18 Mya (Perelman, Johnson et al. 2011). Thus, both *CypA3* and *CypA4* considerably predate the macaque-specific *CypA2* retrogene.

2.2.3 Extant transcriptional expression of a pseudogenized TRIMCypA3 gene fusion

To determine whether the retrotransposition of *CypA3* or *CypA4* into the *TRIM5* locus led to the formation of novel *TRIMCyp* gene fusions, we probed total mRNA from fibroblasts from 16 primate species by RT-PCR, with the forward primer located in the RING domain of *TRIM5* and the reverse primer designed to either *CypA3* or *CypA4*. We identified four Old World monkeys (vervet monkey, De Brazza's monkey, patas monkey, and talapoin) that expressed *TRIMCyp* transcripts, which included the *CypA3* retrogene on their 3′-end (Figure 2.2). We found three distinct isoforms of *TRIM5-CypA3* (*TRIMCypA3*) transcripts. Only one of these, isoform-1, has its *CypA3* in-frame with *TRIM5* exons, where it would be translated as a *TRIMCyp* gene fusion. Isoform-1 encodes *TRIM5* exons 2–7, a short stretch of the upstream region of *CypA3*, and the *CypA3* coding region. The other two isoforms would not result in an in-frame *TRIMCyp* gene fusion (Figure 2.2).

Intriguingly, a shared feature of the gene fusions, including those found in the owl monkey and macaque is the inclusion of a short segment corresponding to the region immediately upstream of the *CypA* retrogene coding region (Figure 2.2, labeled *CypA* upstream region, and Figure A.3A). This short DNA segment, which appears to originate from the 5′-untranslated region of the parental *CypA* gene, encodes a cryptic splice acceptor site that appears conserved throughout mammals (Figure A.3B). At least among primates, this region provides the splice acceptor site and sequences necessary for an in-frame fusion of the *CypA* retrogene with the *TRIM5* effector domains, thereby facilitating formation of the *TRIMCyp* fusion transcripts.

Sequencing the *CypA3* retrogenes from our PCR survey (Figure A.1) revealed signs of pseudogenization in each case (Figure 2.1B and Figure A.3), with either a nonsense mutation or a frameshift in their ORF (Figure 2.1B). A premature stop codon was identified at the 19th codon of *CypA3* in a subset of Old World monkeys (talapoin, patas monkey, and De Brazza's monkey) of the family Cercopithecidae. In addition, all Old World monkey *CypA3* sequences shared a premature stop codon at the site corresponding to the 90th codon of the *CypA* coding sequence. However, neither stop codon is found within the three orthologous, syntenic gibbon *CypA3* sequences from the agile gibbon, island siamang, and white-cheeked gibbon. Instead, all three gibbons encode a frameshift mutation that is predicted to truncate the 3□-end of the *CypA3* coding sequence by 69 nt. Thus, gibbons maintain a large portion of their *CypA3* coding sequence and do not encode the pseudogenizing mutations found in Old World monkeys. Despite encoding a longer intact ORF than Old World monkeys, we did not detect any evidence of a *TRIMCypA3* transcript in gibbons based on RT-PCR. The *CypA3* found in Old World monkeys was likely pseudogenized in their common ancestor, suggesting that the retrogene has existed as a pseudogene in that lineage of primates for 18 My (Perelman, Johnson et al. 2011). Likewise, we estimate that gibbon *CypA3* sequences acquired their frame shift mutation 9–20 Mya in either the gibbon or the hominoid common ancestor (Perelman, Johnson et al. 2011). Thus, although modern *CypA3* sequences are expressed as a *TRIMCyp* gene fusion, the product is likely defective in all extant primates. However, because the pseudogenizing nonsense mutations found in Old World monkey *CypA3* sequences are completely distinct from the frameshift mutation found in gibbon sequences, our analysis strongly implies that *CypA3* in the Old World monkey/hominoid ancestor (*32myoCypA3*) encoded an active ORF (Perelman, Johnson et al. 2011).

Several attempts to identify a *TRIMCyp* transcript that includes *CypA4* yielded no such product from fibroblast mRNA. It is possible that such a product could be expressed in different tissues. However, it is also likely that *CypA4* never contributed to the formation of a *TRIMCyp* gene fusion, because sequencing of *CypA4retrogenes* revealed a shared indel (2-bp deletion) at the position corresponding to the seventh codon, resulting in a pseudogenizing frameshift. Because the *CypA4* from all Old World monkeys shares this common pseudogenizing mutation, *CypA4* may have become pseudogenized at or shortly after birth.

2.2.4 Evolutionary analysis of CypA3

Despite being decayed in all extant primates, the disparate pattern of pseudogenizing mutations in gibbons vs. Old World monkeys suggests that *TRIMCypA3* might have encoded an active antiviral gene at one time. To test this, we built a phylogeny (Figure 2.3A) composed of intact functional owl monkey *CypA1* and macaque *CypA2* retrogenes, pseudogenized retrogenes *CypA3* and *CypA4*, as well as primate *CypA* (parental) genes to calibrate the ages of the retrogenes. The phylogeny shows that all four *CypA* retrogenes split into four distinct monophyletic groups and that the tree topology and branch lengths are consistent with the estimated evolutionary origins of the retrogenes. For instance, the closest outgroup to the *CypA* retrogenes that gave rise to *Aotus TRIMCyp* (*CypA1*) is the *Aotus CypA* gene, confirming that *Aotus CypA1* retrogenes are derived from a *CypA* gene within the *Aotus* genus (Ribeiro, Menezes et al. 2005). We are unable to gain high resolution within the hominoid and Old World monkey *CypA* (parental) genes due to the very high identity of these sequences, likely the consequence of extremely strong purifying selection (Ortiz, Bleiber et al. 2006). However, our phylogenetic analysis also places *CypA2* retrogenes close to the macaque genus, albeit with poor resolution due to the phylogenetic proximity of the Old World monkey and hominoid *CypA* genes.

Our PCR genotyping for the presence or absence of *CypA* retrogenes in primate genomes allowed a tentative dating of their age on a primate phylogenetic tree (Figure 2.1). Consistent with our genotyping results, we find that a phylogenetic analysis of the *CypA* sequences themselves shows that *CypA3* and *CypA4* retrogenes are older than *CypA2* retrogenes, with *CypA4* appearing to branch slightly before the common ancestor of the Old World monkeys and hominoids (Figure 2.3A). Surprisingly, based on the phylogeny, with 99% bootstrap support, we find that *CypA3* was acquired in the simian common ancestor (Figure 2.3A) at least 43 Mya (Perelman, Johnson et al. 2011), which is even earlier than the 32 Mya that we had inferred from PCR genotyping (Figure A.1). However, our attempts to detect *CypA3* in New World monkeys were unsuccessful (Figure A.2). This discordance between our genotyping and phylogenetic analysis could be a result of discordance in mutation rates between retrogenes compared with the parental *CypA* genes. However, because we assume that all these retrogenes were the product of a single cycle of retrotransposition (i.e., retrogenes did not give rise to other retrogenes), we do not believe this difference in mutation rates is sufficient to skew our phylogenetic analysis. We also note the consistency between the genotyped and phylogenetic "age" inferences for *CypA1* (Ribeiro, Menezes et al. 2005), despite *CypA1* having also gone through a retrotransposition event. Instead, we conclude that the *CypA3* retrogene was independently lost in the lineage of New World monkeys, similar to its loss in

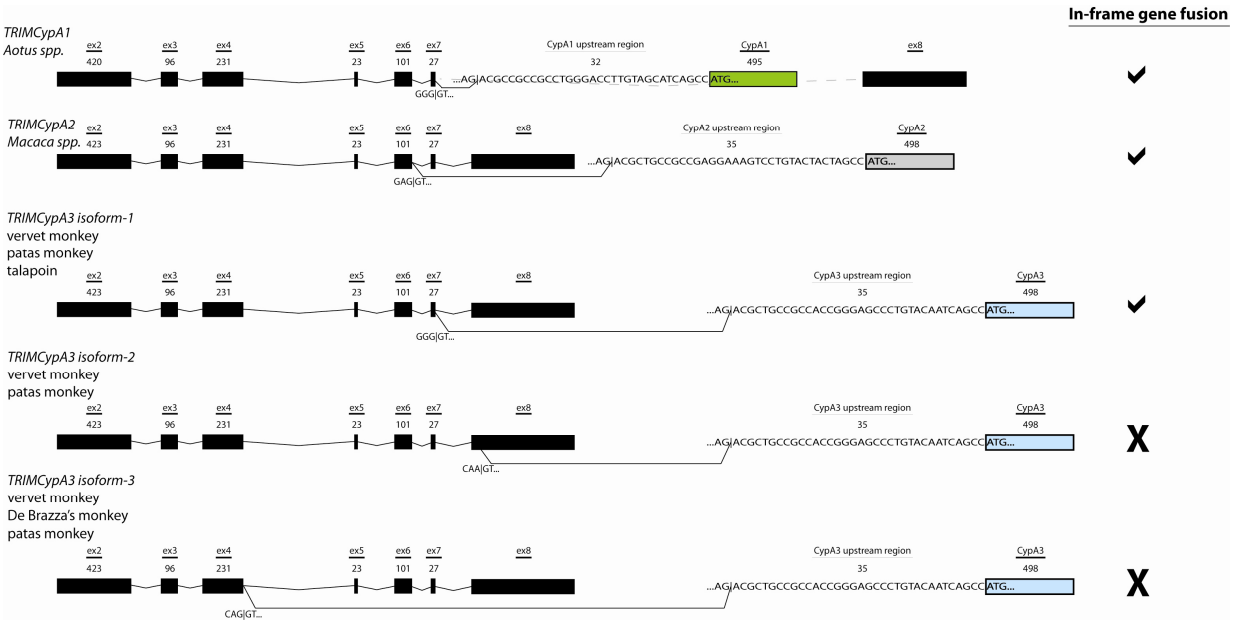


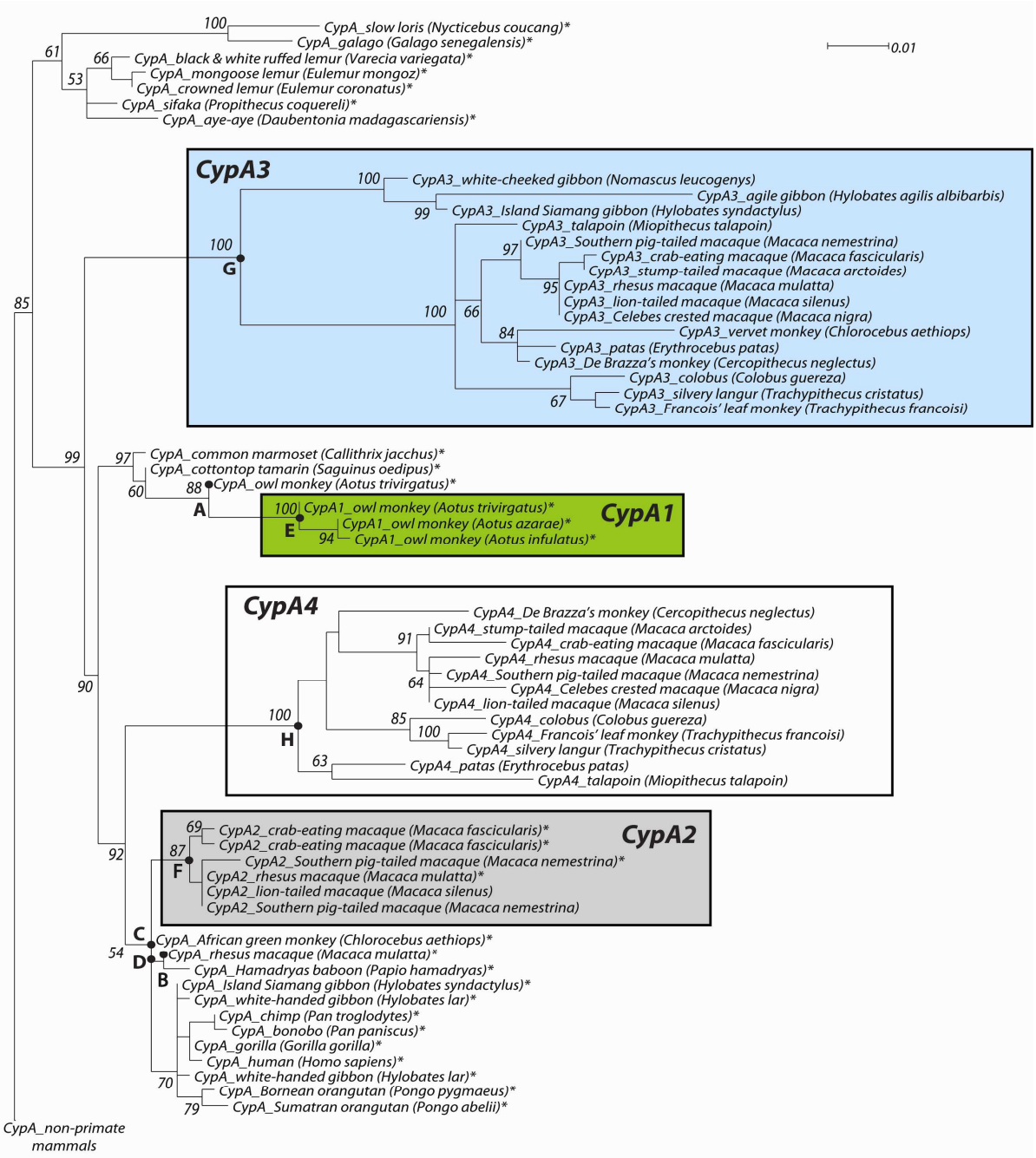
Fig. 2.2 Structure of *TRIMCyp* transcripts. From a subset of Old World monkeys, we found three *TRIMCypA3* isoforms transcribed. *TRIM5* exons (black blocks) are joined to a stretch of the upstream region of *CypA* and the subsequent *CypA* coding sequence (*CypA1*, green; *CypA2*, light gray; *CypA3*, light blue). We included the sequence of the intron boundaries and the approximate size of each exon. The structures of owl monkey *TRIMCypA1* and macaque *TRIMCypA2* are also presented for comparison (15, 24, 27). The splice acceptor sequence "AG|AC," also present in the *CypA* upstream region, is shown for each retrogene. Isoform-1 encodes *TRIM5* exons 2–7 fused in-frame to the *CypA3* upstream region and *CypA*. In contrast, isoform-2 encodes TRIM5 exons 2–8 (62 nt of exon 8) to the *CypA3* upstream region and *CypA3* coding region, whereas isoform-3 encodes *TRIM5* exons 2–4 to the *CypA3* upstream region and *CypA3* coding region. To the right, we indicated whether the gene fusion produces a product, where TRIM5 effector domains are in-frame with *CypA*, with a "check mark," indicating an in-frame product or an "X," indicating that the product would not be in-frame. Both isoform-2 and isoform-3 would result in an "out-of frame" gene fusion with *CypA*.

some hominoids. Thus, our phylogenetic results clearly reveal *CypA3* as being at least 43 My old, which means that it is, by far, the oldest of the four primate *CypA* retrogenes.

Using both parsimony and likelihood criteria, we were able to reconstruct the sequence of this "intact" Old World monkey/hominoid ancestral *CypA3* (hereafter referred to as *32myoCypA3*) based on extant *CypA3* sequences. Only a single site, residue 144 of CypA3, could not be resolved (Figure A.3), and it could encode a proline (P), arginine (R), or histidine (H). To determine whether selective constraints (either diversifying or purifying selection) acted on *CypA3*, we compared the dN/dS ratio of *32myoCypA3* with that of the ancestral parental *CypA* gene from which it likely derived (Figure 2.3B). Finding a high dN/dS ratio would indicate *CypA3* was under pressure to evolve adaptively, a low dN/dS ratio would be evidence of selective pressure for protein constraint, and a dN/dS ratio ∼1 would indicate an absence of selective pressure. *CypA3* showed evidence suggestive of purifying selection (probability of [dN < dS] = 0.0470–0.0520; Figure 2.3B). Applying the same analysis to evaluate the modern, functional owl monkey and macaque *TRIMCyps* (Figure 2.3B), we also find evidence for purifying selection in owl monkey *CypA1* (probability of [dN < dS] = 0.1010) and in macaque *CypA2* (probability of [dN < dS] = 0.0010). In contrast, an analysis of the ancestral version of *CypA4* shows no evidence of selective constraint (probability of [dN < dS] = 0.429). Thus, during the period that it encoded an intact ORF, *CypA3* seems to have evolved under similar selective pressures as both owl monkey and macaque *CypA* retrogenes (Figure 2.3B).

The dS values of these retrogene-parental gene comparisons are also informative as a rough proxy for their age of divergence (assuming roughly equal rates of evolution at silent sites). Both the owl monkey and macaque *TRIMCyp* gene comparisons have a dS of 0.02 (Figure 2.3B), consistent with their birth ∼4.5–6 Mya (Ribeiro, Menezes et al. 2005, Dietrich, Jones-Engel et al. 2010). In contrast, the *CypA3* comparison between the parental gene and *32myoCypA3* reveals a dS of 0.04 (Figure 2.3B), which is twice that of the value estimated for the owl monkey or macaque, suggesting that *CypA3* was preserved as an ORF for at least twice as long as the currently intact *TRIMCyp* gene fusions. This suggests that *CypA3* was preserved as an intact retrogene from the time it was acquired 43 to 32 Mya, when we begin to observe evidence of independent pseudogenization (or loss) events across the primate lineages. The signature of purifying selection (Figure 2.3B) further suggests the *TRIMCypA3* gene fusion was functional during this period of ∼10 My. Because no extant *CypA3* sequences could be isolated in New World monkeys, the "oldest" version of ancestral *CypA3* that we could faithfully reconstruct represents the

A



B

Pairwise Comparison	dN	dS	Prob. [dN<dS]	Age of retrogene (estimated)
(A) owl monkey CypA gene vs. (E) 'anc' owl monkey CypA1 retrogene	0.011	0.02	0.101	6 million years
(B) macaque CypA gene vs. (F) 'anc' macaque CypA2 retrogene	0.003	0.02	0.001	6 million years
(C) 'anc' OWM-Hominid CypA gene vs. (G) OWM-Hominid 32myo CypA3 retrogene	0.021-0.024	0.04	0.047-0.052	>32 million years
(D) 'anc' OWM CypA gene vs. (H) 'anc' OWM CypA4 retrogene	0.032	0.035	0.429	18 million years

Figure 2.3 Phylogeny of *CypA* retrogenes. (A) We built a phylogeny of parental *CypA* genes and retrogenes using maximum likelihood methodologies (57). *CypA* gene sequences were collected from rodents (outgroup), prosimians, New World monkeys, Old World monkeys, and hominoids. The *CypA* retrogenes that we included were *CypA1* (green-filled box) from owl monkeys, *CypA2* (light gray-filled box) from macaques, *CypA3* (light blue-filled box), and *CypA4* (white-filled box). Bootstrap support values are shown at the nodes. The phylogeny has been rooted to the rodent parental *CypA* genes: mouse (*Mus musculus*), rat (*Rattus norvegicus*), and squirrel (*Ictidomys tridecemlineatus*). (B) We used the K-Estimator program (58) to evaluate the rates of dN and dS for *CypA* retrogenes and computed the probability that dN is significantly different from dS by confidence interval tests.

version that existed in the last common ancestor of hominoids and Old World monkeys (*32myoCypA3*) 10 My following the birth of *CypA3*.

2.2.5 Testing ancient and de novo TRIMCyp proteins for restriction of modern or ancient lentiviruses

Previous studies have explored the interactions between *CypA* genes and retrogenes with lentiviral capsids (Towers, Hatziioannou et al. 2003, Sayah, Sokolskaja et al. 2004, Brennan, Kozyrev et al. 2008, Newman, Hall et al. 2008, Virgen, Kratovac et al. 2008, Wilson, Webb et al. 2008, Price 2009, Dietrich, Brennan et al. 2011). We were therefore interested in assessing whether ancient, potentially active versions of *TRIMCyA3* might have interacted with lentiviral capsids. Both naturally occurring and artificial *TRIMCyp* genes have highlighted the modularity of the *TRIM5-CypA* gene arrangement (Yap, Mortuza et al. 2007). We therefore designed two synthetic *TRIMCypA3* versions. We elected to use the *TRIM5* effector domains from the owl monkey because the exon structure of *TRIMCypA1* closely resembles *TRIMCypA3*. Thus, the first version (*TRIM5-32myoCypA3*) consisted of owl monkey *TRIM5* effector domains fused to *32myoCypA3* (Figure 2.4A). Because we were not able to resolve the identity of residue 144 unambiguously from our evolutionary reconstruction, we constructed three separate TRIM5-32myoCypA3 versions, which encoded a P, R, or H at this site. We also designed a chimera composed of owl monkey *TRIM5* effector domains and the inferred parental *CypA* gene (*TRIM5-parentalCypA*), representing the *CypA* gene from which *CypA3* derived at the time of its birth. Because *CypA* genes have been evolving under strong purifying selection throughout primate history, this approach allows us to evaluate the lentiviral specificity of a de novo *CypA* retrogene unambiguously, representing *TRIMCypA3* immediately following its birth (Neagu, Ziegler et al. 2009).

Consistent with previous results, we observed that owl monkey TRIMCypA1 was not able to restrict HIV-2 (Zhang 2006) but was able to restrict HIV-1 (Sayah, Sokolskaja et al. 2004), simian immunodeficiency virus from African green monkeys (SIVagm) (Lin and Emerman 2006), and feline immunodeficiency virus

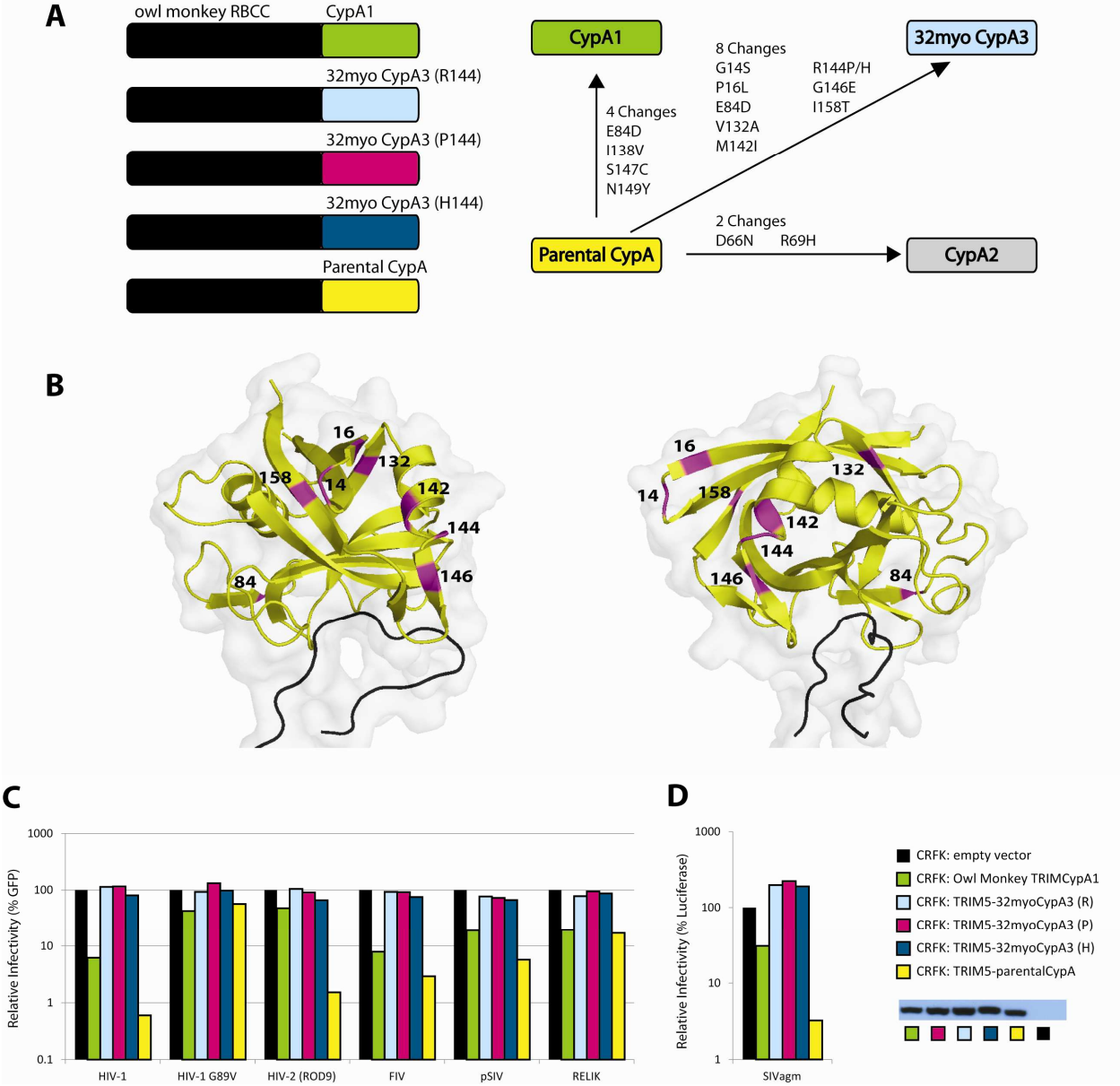


Figure 2.4 Reconstructed ancestral CypA3 vs. modern and extinct (reconstructed) lentiviruses. (A) Cartoon representations of the owl monkey TRIMCypA1 (CypA1, green), owl monkey TRIM5-32myoCypA3 [32myoCypA3 (R144), light blue; 32myoCypA3 (P144), magenta; 32myoCypA3 (H144), teal], and owl monkey TRIM5-parental CypA (yellow) gene fusions, with the owl monkey TRIM5 effector domains (RING, B-box, coiled-coil) represented as a black block. Differences in residues encoded by CypA1, parental CypA, CypA2 (light gray), and 32myoCypA3 have been identified and listed according to the direction of the arrow. (B) Eight residues unique to 32myoCypA3 (P144) (magenta) can be mapped onto a structure of parental CypA (yellow) interacting with capsid (black) using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC). Two different orientations of the structure are presented. (C) Stable CRFK cell lines encoding an empty vector (black box), owl monkey TRIMCypA1 (green box), TRIM5-32myoCypA3 (R/P/H144) (light blue, magenta, and teal boxes, respectively), and owl monkey TRIM5-parentalCypA (yellow box) were assayed against chimeric EIAV

encoding the ancient (resurrected) capsid of paleolentiviruses RELIK and pSIV (38) and the modern lentiviruses HIV-1 (LAI strain), HIV-1 G89V0, HIV-2 (ROD9 strain), and FIV (9, 56). (D) SIVagm. 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 against CRFK cells encoding an empty vector. The virus inoculums were standardized to give the absolute percentage of GFP between 15% and 30%. In the case of SIVagm, this system used a luciferase reporter. Shown is a representative experiment that was repeated three times. We confirmed the stable expression of TRIMCyp proteins by Western blot analysis, using 30 µg of protein extract for each sample (lane 1, owl monkey TRIMCypA1; lane 2, TRIM5-32myoCypA3 (R144); lane 3, P144; lane 4, H144; lane 5, TRIM5-parental CypA; lane 6, CRFK with an empty LPCX vector).

(FIV) (Diaz-Griffero, Kar et al. 2007), as well as chimeric viruses encoding the reconstructed capsid of the "paleoviruses" RELIK and pSIV (Goldstone 2010) (Figure 2.4 C and D). Remarkably, we found that the TRIM5-parentalCypA encodes broad and potent antiviral activity, because it restricts all these lentiviruses tested except a mutant that disrupts the CypA binding site on capsid (HIV-1 G89V) (Figure 2.4 C and D). Thus, *TRIM5-parentalCypA* could restrict all representatives tested from the modern-day lentiviruses and the paleolentiviruses (Figure 2.4 C and D). The variation between the slightly narrowed binding specificity of CypA1 and the broad specificity of parentalCypA is attributed to four amino acid differences that occurred during CypA1 evolution (Figure 2.4A). On the other hand, we also found that *TRIM5-32myoCypA* could not restrict any of the paleolentiviruses or modern-day lentiviruses (Figure 2.4 C and D). Depending on the ambiguous residue 144, parentalCypA and 32myoCypA3 differ at seven (or eight) residues (Figure 2.4 A and B), one of which has also independently occurred during CypA1 evolution. We attempted to explore the loss of restrictive ability by evaluating residues unique to 32myoCypA3 within the parentalCypA backbone using both 32myoCypA3 (P144)/parentalCypA and parentalCypA/32myoCypA3 (P144) chimeras (Figure A.4A). We found that the loss of restriction activity in 32myoCypA3 could not be reversed by replacing N- of C-terminal regions with parentalCypA (Figure A.4B), suggesting that this observed loss of restrictive ability in 32myoCypA3 is attributable to a combination of multiple residues among the seven (or eight) residues specific to the 32myoCypA3-encoded protein (Figure 2.4 A and B). These results indicate that the most ancient version of the *Trim5-CypA* fusion gene had the broadest specificity for restriction of retroviruses but that subsequent evolution either narrowed its specificity to (ancient) retroviral capsids that we were not able to test in our assays or destroyed this activity.

2.3 Discussion

2.3.1 Recurrent TRIM5-CypA gene fusions across the primate phylogeny

Retrotransposition of *CypA* retrogenes proximal to the *TRIM5* locus has the instantaneous effect of creating a new restriction factor, potentially expanding the restrictive range of primate genomes (Brennan, Kozyrev et al. 2008, Newman, Hall et al. 2008, Virgen, Kratovac et al. 2008, Wilson, Webb et al. 2008). Including the present study, at least three such instances of *TRIMCyp* gene fusion are now documented in primate genomes. In addition to the still active *CypA1* and *CypA2* retrogenes that were born in the owl monkey and macaque species 4.5–6 Mya, we have identified a third, much more ancient retrogene that is still present as a fusion transcript and likely encoded a putative restriction factor 43 Mya in primate history. This remarkably convergent retrotransposition proximal to *TRIM5*, in contrast to the frequent but otherwise random pattern of *CypA* retrogene insertions elsewhere in primate genomes, strongly suggests that the *CypA* retrogene bearing haplotype must have had a strong enough selective advantage to sweep through populations and species. Based on the potent antiviral activity of TRIMCyp fusion proteins, we posit that it is most likely that this selective advantage was conferred by protection against an ancient viral infection.

Previous studies have suggested that *CypA* fusions function in concert with a variety of *TRIM* genes (Javanbakht, Diaz-Griffero et al. 2007, Yap, Mortuza et al. 2007); however, only *TRIM5* has recurrently been revealed to accommodate a functional gene fusion with *CypA* naturally. It may be that the expression patterns of other *TRIM* genes could not accommodate a functional antiviral gene fusion without compromising endogenous function. Alternatively, given that *TRIM* genes homomultimerize via their B-Box and Coiled-Coil domains, homomultimerization of a *TRIM* gene with *CypA* might have had deleterious consequences that would only be tolerated when this involved a canonical restriction factor like *TRIM5* but perhaps not a *TRIM* gene that plays an essential housekeeping function in the cell. Consistent with this hypothesis, our survey of rhesus macaque, chimpanzee, and human genomes has not revealed any other *TRIMCyp* candidates in which a *CypA* retrogene was found within 20 kb of a *TRIM* gene.

The retention pattern of *TRIMCyp* genes could pose a cost to antiviral defense and the cell. In the case of *CypA1* and *CypA2*, it precludes the production of a B30.2-containing TRIM5α from that allele, providing a tradeoff in terms of restrictive potential. This would explain why *CypA2* has variably swept through to fixation in macaque species (Ylinen 2010), likely as a result of balancing selection, as seen previously in the *TRIM5* locus of macaque populations (Newman, Hall et al. 2006). An additional explanation could be that the fusion of the TRIM5 E3 ubiquitin ligase domain to a CypA protein that may bind numerous client

proteins in the primate proteome increases the toxic burden of such gene fusions, or leads to aberrant cell signaling (Pertel, Hausmann et al. 2011). In light of this "cost," if the restriction activity of the evolved *TRIMCyp* is obviated, either because the restricted viral capsid is eliminated or evolves away from *TRIMCyp* recognition, the advantage to retain the *TRIMCyp* gene fusion diminishes greatly. It is also possible that rapid evolution of the B30.2 domain of TRIM5α to recognize the target retroviral capsid might obviate the need to maintain *TRIMCyp*. Thus, it is not surprising that the evolution of these gene fusions is recurrent and dynamic, both remarkably convergent but also relatively short-lived in evolutionary time. This would also help explain why *TRIMCypA3*, which may have encoded an active antiviral protein in primate history, is now an extinct gene. Such "extinct" *TRIMCyp* gene fusions have also been identified outside primates. Indeed, such a pseudogenized *TRIMCyp* gene fusion (ftr52) was recently identified in fish genomes (Boudinot, van der Aa et al. 2011). Finally, even though the *CypA* domain of *TRIMCypA3* has decayed, it is formally possible that the *TRIMCypA3* fusion transcript in some primates still serves the same function as some of the alternate *TRIM5* isoforms, to attenuate TRIM5α function (Berthoux, Sebastian et al. 2005).

2.3.2 CypA3 as a Potential Paleoviral Marker in Primate Evolution

Based on a relatively abundant record of endogenization revealed by sequencing and bioinformatic efforts, retroviral lineages have been shown to date back many millions of years (Katzourakis, Gifford et al. 2009, Han and Worobey 2012). Indeed, lentiviruses have been estimated to be at least 4 My old in primates (Gilbert, Maxfield et al. 2009) and ∼12 My old in other mammals based on the presence of endogenous copies within the host genome, although these dates are likely vast underestimations of the true age of lentiviruses (Keckesova, Ylinen et al. 2009, Cui and Holmes 2012, Gifford 2012, Han and Worobey 2012). In response to these retroviral challenges faced throughout their evolution, primates encode a number of intrinsic mechanisms with the capability of inhibiting viral replication. Positive selection of such restriction factors is a potent mechanism for primate genomes to respond to novel or adapted viral pathogens (Sawyer, Wu et al. 2005), but it is not the only mode of adaptation. Primate genomes also use other mechanisms, such as gene duplications (Han, Lou et al. 2011), and in the case of *TRIMCyp*, recurrent gene fusions, to respond to new viral challenges.

TRIMCyp evolution not only serves to belie the traditional view that retrotransposed genes are evolutionary dead ends but suggests that *CypA* retrogenes are highly labile modules that can be gained and lost throughout primate history. Although whole-gene dN/dS analyses strongly suggest that *CypA3*

evolved under purifying selection for 10 My following its birth, we identified seven (or eight) residues within 32myoCypA3 that differentiate a loss of capsid-binding from broad-range capsid-binding (as exhibited by parentalCypA). Similarly, only four residues separate the broad binding of parentalCypA to the narrowed binding specificity documented from CypA1. Macaque CypA2 further demonstrates this trajectory of narrowed binding specificity (Ylinen 2010). In the cases of *CypA1* and *CypA2*, deviation from broad capsid-binding evolved within 6 My. Therefore, although broad capsid binding appears as an innate feature of *CypA*, the specificity that each *CypA* retrogene evolves is determined by minor changes that have a great impact on the capsid-binding trajectory (Price 2009, Dietrich, Jones-Engel et al. 2010, Ylinen 2010). Based on our results with *TRIM-parentalCypA*, we predict that *TRIMCypA3* was also capable of interacting with a broad range of lentiviral capsids on birth. Similar to the specificity-narrowing changes that occurred during *TRIMCypA1* and *TRIMCypA2* evolution, we posit that in the 10 My after its birth, *TRIMCypA3* narrowed its specificity to restrict only ancient retroviruses rather than any of the retroviruses we tested. Finally, after the utility of *TRIMCypA3* as a retroviral restriction factor was exhausted ~32 Mya, the *TRIMCypA3* gene decayed in all extant primates.

From a paleovirology perspective, even currently inactive or pseudogenized *CypA* retrogenes may represent remnants of antiviral genes that were active at an earlier time in primate evolution. We propose that ancient *TRIMCypA3* arose in response to a pathogen encountered by evolutionarily successful ancestors. Although it is formally possible that the true target of *TRIMCypA3* was a non-lentiviral or even a non-retroviral pathogen, there is little precedent for this conjecture. It is also unlikely that *TRIMCypA3* was a genomic innovation due to some other "housekeeping" adaptation, based both on the intrinsic costs of *TRIM5-CypA* gene fusions and the recurrent pseudogenization/loss of *TRIMCypA3* in extant primates. Instead, we propose that the birth and demise of *TRIMCypA3* are more consistent with the model wherein it helped protect host genomes against viral invasions for as long as 10 My of primate history. Thus, "fossil" antiviral genes like *CypA3* provide unique paleoviral insight into viral challenges encountered by primate ancestors 43 Mya and complement the incomplete fossil record of retroviral imprints in animal genomes.

2.4 Methods

2.4.1 Identifying CypA Retrogenes Proximal to TRIM5

The human *CypA* gene (NC_000007) and mRNA (NM_021130) sequences were used as query sequences in a BLAST-like alignment tool (BLAT) analysis to identify *CypA* homologs (Kent 2002). BLAT searches

were performed from the University of California, Santa Cruz Genome Browser on the human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), and rhesus macaque (*Macaca mulatta*) genomes (Kent, Sugnet et al. 2002). For each of the primate genomes, the BLAT search results from the two query sequences were combined to assemble a comprehensive list of *CypA* homologs that was evaluated and compiled into a catalog of *CypA* retrogenes. Cyp*A* retrogenes were mapped back to their respective primate genome, and *CypA* retrogenes proximal to *TRIM5* were identified (Figure 2.1). *CypA* retrogenes were named according to their distance from *TRIM5* and based on previously established nomenclature. These *CypA* retrogenes were then evaluated for an ORF, indels, and premature stop codons. We also catalogued the distribution of *CypA* retrogenes and organized these based on the number of *CypA* retrogenes found in a random stretch of 100 kb of the evaluated primate genomes. To calculate the probability of multiple *CypA* insertions within a given distance, 100 kb in this case, we counted the number of 100-kb stretches that contained 0–10 *CypA* retrogenes. We focused on rhesus macaques because the largest number of events in which multiple *CypA* retrogenes could be found in any 100-kb stretch of its genome was reported in this species. We identified 116 cases of only finding 1 *CypA* retrogene within 100 kb. In addition, we identified four cases of finding 2 *CypA* retrogenes and one case of finding 3 *CypA* retrogenes within 100 kb of each other in the rhesus macaque genome. Thus, of 129 total *CypA* retrogenes in the rhesus macaque genome, only 3 *CypA* retrogenes could be found within 100 kb of each other (proximal to the *TRIM5* locus), which we can calculate as a probability ($P = 0.02325$).

2.4.2 Determining the Presence or Absence of Proximal CypA Retrogenes

Genomic DNA (gDNA) was isolated from primate fibroblast cells purchased from Coriell Cell Repositories. The primate panel was composed of human, chimpanzee (ID no. 3448), bonobo (*Pan paniscus*, ID no. 5253), gorilla (*Gorilla gorilla*, ID no. 5251), orangutan (*Pongo pygmaeus*, ID no. 5252), island siamang gibbon (*Hylobates syndactylus*, PR00722), agile gibbon (*Hylobates agilis albibarbis*, PR00773), rhesus macaque (ID no. 7098), crab-eating macaque (*Macaca fascicularis*, ID no. 3446), celebes-crested macaque (*Macaca nigra*, ID no. 7101), pig-tailed macaque (*M. nemestrina*, ID no. 8452), stump-tail macaque (*Macaca arctoides*, ID no. 3443), lion-tailed macaque (*M. silenus*, OR1890), silvery leaf langur (*Trachypithecus cristatus*, bl.4381), Francois' leaf langur (*Trachypithecus francoisi*, PR01099), colobus (*Colobus guereza*, PR00980), talapoin (*Miopithecus talapoin*, PR00716), patas monkey (*Erythrocebus patas*, ID no. 6254), De Brazza's monkey (*Cercopithecus neglectus*, PR01144), pygmy marmoset (*Callithrix pygmaea*, OR690), and saddle-back tamarin (*Saguinus fuscicollis nigrifrons*, OR621)

species. gDNA from this diverse primate panel, representing New World monkeys, Old World monkeys, and hominoids, was used to determine the presence or absence of *CypA2*, *CypA3*, and *CypA4* throughout primates in a PCR survey. All PCR reactions were performed using 25-µL reaction volumes and the PCR SuperMix High Fidelity (Invitrogen) reagent. The thermocycler parameters were 94 °C for 3 min; 39 cycles at 94 °C for 15 s, 60 ° for 15 s, and 72 °C for 2 min; and a final extension step at 72 °C for 10 min. All products were directly sequenced using BigDye sequencing (Applied Biosystems). *CypA2* reactions were performed using primers 105 (forward: 5′-CTGTGCTCACCAAGCTCTTGAAC-3′) and 103 (reverse: 5′-TCCCACATAATTCAGTTTGTTTGATAAA-3′), and *CypA4* reactions were performed using primers 108 (forward: 5′-AATCTGCTGGCACCTTGTTTTGTAC-3′) and 110 (reverse: 5′-TAGCTTTTGGGCAGCTAGGAGG-3′). We used nested PCR analysis to amplify *CypA3* from primates, with the first-round primers being 87 (forward: 5′-GAACTACTTGAATCCAGGAGGCAGA-3′) and 101 (reverse: 5′-TATCCTCTTTTTGAATCAATTCCTTTGTCA-3′) and the second round primers being 100 (forward:5′-GCAGGAGTAAGTCCTCACCTATC-3′) and 84 (reverse: 5′-TTATTCGAGTTGTCCACAGTCAGCAG-3′).

2.4.3 Detecting TRIM5-CypA3 and TRIM5-CypA4 Transcripts

A two-step RT-PCR/semi-nested PCR-based method was used to amplify *TRIMCyp* from primate RNA. The primates used were human, chimpanzee, island siamang gibbon, agile gibbon, talapoin, patas monkey, De Brazza's monkey, vervet monkey (*Cercopithecus aethiops*, PR01190), Francois' leaf monkey, colobus, rhesus macaque, woolly monkey (*Lagothrix lagotricha*, ID no. 5356), spider monkey (*Ateles belzebuth*, KB6701), titi monkey (*Callicebus donacophilus*, OR1522), and owl monkey (*Aotus trivirgatus*, CRL-1556). Total RNA was isolated from fibroblast cells purchased from Coriell Cell Repositories. The initial RT-PCR step was performed using a primer designed to the start of the coding region of the *TRIM5* gene (primer 80) and an oligo-dT reverse primer. This primer combination was used to amplify all products encoded by the *TRIM5* gene. Next, we used either a *CypA3*- or *CypA4*-specific reverse primer in combination with primer 80 to confirm the transcription of a *TRIMCyp* gene fusion. RT-PCR reactions were performed using SuperScript III Reverse Transcriptase with Platinum Taq (Invitrogen) in 12.5-µL volume reactions. The RT-PCR parameters were an initial RT step at 50 °C for 30 min; followed by 34 cycles at 94 °C for 15 s, 60 °C for 15 s, and 68 °C for 3 min; and a final extension at 72 °C for 10 min. A 1:300 dilution of the RT-PCR product was then prepared for the subsequent semi-nested PCR step. This was performed using primers 80 and 73 (reverse: 5′-TTATTMGAGTTGTCCACAGTCAGCARTRGTGA-3′) to amplify *TRIMCyp* without targeting a specific *CypA* retrogene. The PCR parameters were kept

unchanged. All products were TOPO TA (Invitrogen) cloned and BigDye sequenced using M13 universal primers.

2.4.4 Construction of CypA Phylogeny, Alignment, and 32myoCypA3

The nucleotide sequences of modern *CypA* genes and *CypA1–4* retrogenes were used to build an alignment of all *CypA* sequences using Clustal W2 (Larkin, Blackshields et al. 2007). This was done for *CypA* sequences at the nucleotide and protein levels. The nucleotide alignment was used in reconstructing the 32 million year old form of *CypA3* (*32myoCypA3*). We were able to use a parsimony-based approach to reconstruct the sequence of *32myoCypA*, which was in agreement with a maximum likelihood reconstruction.

Phylogenetic trees were generated using Mr. Bayes (version 3.1) in the construction of the *CypA* phylogeny. We performed 1 million Markov chain Monte Carlo generations with a sampling every 1,000th generation and discarded the first 250 samplings as run-in.

2.4.5 Assessing Ancient and de Novo TRIMCyp Proteins for Antiviral Activity

To test 32myoCypA3 for the ability to interact with viral capsid protein, we used "stitch-PCR" to join the TRIM5 effector domains (RING, B-Box, and Coiled-Coil) from owl monkey TRIMCypA1 to 32myoCypA. All PCR parameters were as previously mentioned. The first-round set of stitch-PCR used primers 144 (forward: 5′-GCGCTTCTCGAGGCCACCAT-3′) and 134 (reverse: 5′-GGGGTTGACCATGGCTGATGCTAC-3′) to amplify the *TRIM5* region of owl monkey *TRIMCypA1* and primers 133 (forward: 5′-GTAGCATCAGCCATGGTCAACCCC-3′) and 177 (reverse: 5′-GCGCGCTTATCGATGAATTCTTATTC-3′) to amplify *32myoCypA*. Dilutions of the first-round products were combined and stitched together by PCR using primers 144 and 177. This PCR product was sequenced to verify the successful construction of the *TRIM5-32myoCypA3* gene fusion. Owl monkey *TRIMCypA1* and *TRIM5-32myoCypA3* were cloned into the expression vector pLPCX and then transduced into Crandell-Rees feline kidney (CRFK) cell lines to establish the following stable cell lines: CRFK (owl monkey *TRIMCypA1*) and CRFK (*TRIM5-32myoCypA*). *TRIM5-32myo-CypA3 P144R* and *P144H* were generated using a QuikChange II Site-Directed Mutagenesis Kit (Agilent). A CRFK cell line containing pLPCX without an insert was also established to serve as a negative control. The gene fusion of owl monkey *TRIM5* effector domains and *parentalCypA* (*TRIM5-parentalCypA*) was built by first amplifying the mRNA from the rhesus macaque *CypA* gene with primers: 262 (forward: 5′-CTGGGACCTTGTAGCATCAGCCATGGTCAACCCCACCGTGTTCTTC-3′) and 264

(reverse: 5′-GCGCGCTTATCGATGAATTATTCGAGTTGTCCACAGTCAGCAATG-3′). Next, the owl monkey *TRIM5* amplicon was combined with the rhesus macaque *CypA* gene using primers 177 and 264. We confirmed *TRIMCyp* protein expression in the appropriate cell lines by Western blot analysis. We used the following viruses in assessing our cell lines: HIV-1 (LAI strain), HIV-2 (ROD9), FIV, RELIK, and pSIV, and we used no virus as a control. RELIK and pSIV were prepared by cotransfection of 293T cells with pL-vesicular stomatitis virus- G, pCMV-tat, equine infectious anemia virus (EIAV) GFP 6.1 (encoding the genome of EIAV with a GFP expression cassette), and either pEIAV-RELIK or pEIAVpSIV (Goldstone 2010) (a kind gift from Melvyn Yap, MRC National Institute for Medical Research, London). Viruses were harvested by collecting supernatant and titered on CRFK cells to determine the dose of virus that would infect between 15% and 30% of the cells in a 12-well plate based on flow cytometry for GFP expression. Other viruses were similarly constructed and assayed (Yamashita and Emerman 2004, Sawyer, Wu et al. 2005). For infection assays, cell lines were seeded onto 12-well plates and subsequently infected with the aforementioned viruses using the predetermined viral titers. Three days post-infection, cells were collected from the 12-well plates and suspended in fixing agent for immediate analysis by flow cytometry. SIVagm infections were performed using 96-well plates. We did not need to fix RELIK or pSIV-infected samples before flow cytometry. In all experiments, mock infected cells were used to set the GFP gate.

Chapter 3

An evolutionary screen highlights candidate antiviral genes within the primate *TRIM* gene family

Recurrent viral pressure has acted on host-encoded antiviral genes during primate and mammalian evolution. This selective pressure has resulted in dramatic episodes of adaptation in host antivirals, often detected via positive selection. These evolutionary signatures of adaptation have the potential to highlight previously unrecognized antiviral genes and to expand the repertoire of known restriction factors. While the *TRIM* multigene family is recognized for encoding several bona fide restriction factors (e.g. TRIM5alpha), most members of this expansive gene family remain uncharacterized. Here, we investigated the *TRIM* multigene family for signatures of positive selection in order to identify novel candidate antiviral genes. Our analysis reveals previously undocumented signatures of positive selection in 14 *TRIM* genes, 10 of which represent novel candidate restriction factors. These include the unusual *TRIM52* gene, which has evolved under strong positive selection despite its encoded protein lacking a putative viral recognition (B30.2) domain. We show that *TRIM52* arose via gene duplication from the *TRIM41* gene. Both *TRIM52* and *TRIM41* have dramatically expanded RING domains compared to the rest of the *TRIM* multigene family, yet this domain has evolved under positive selection only in primate *TRIM52*, suggesting that it represents a novel host-virus interaction interface. Our evolutionary-based screen not only documents positive selection in known *TRIM* restriction factors but also highlights candidate novel restriction factors, providing insight into the interfaces of host-pathogen interactions mediated by the *TRIM* multigene family.

3.1 Introduction

Host encoded restriction factors confer an intrinsic line of defense that inhibits viruses at various stages of the viral life cycle (Goff 2004, Duggal and Emerman 2012, Yan and Chen 2012). One example of this type of antiviral defense gene is *TRIM5*, which was identified as the block to HIV-1 infection in rhesus macaques (Stremlau, Owens et al. 2004). The potent restriction by *TRIM5* is conserved in other mammals, including primates (Yap, Nisole et al. 2004, Song, Javanbakht et al. 2005, Zhang 2006, Kratovac, Virgen et al. 2008, Yap 2008, Rahm 2011) and closely related paralogs belonging to glires (Schaller, Hue et al. 2007, Tareen, Sawyer et al. 2009, Fletcher, Hué et al. 2010) and cows (Si, Vandegraaff et al. 2006, Ylinen, Keckesova et al. 2006). Restriction activity is attributed to the assembly of a *TRIM5* lattice directly to the surface of the retroviral core (Ganser-Pornillos, Chandrasekaran et al. 2011) that is thought to mediate premature capsid disassembly (Stremlau, Perron et al. 2006). Antiviral activity of *TRIM5* has also been attributed to the induction of an inflammatory response (Pertel, Hausmann et al. 2011, Tareen and Emerman 2011). Retroviral specificity of *TRIM5* dramatically differs amongst primate orthologs due to ancient and on-going selective pressures reflected by variation in the Coiled-Coil and B30.2 domains, which influence the interaction with viral proteins (Sawyer, Wu et al. 2005, Sebastian and Luban 2005, Kirmaier, Wu et al. 2010, Maillard, Ecco et al. 2010).

TRIM5 is a member of the *TRIM* multigene family, which encodes more than 70 genes in humans and is similarly expansive throughout primates (Han, Lou et al. 2011). Proteins encoded by the *TRIM* multigene family are characterized by a tripartite motif consisting of a RING domain, one or two B-boxes, and a Coiled-Coil motif, the order and spacing of which are generally conserved (Reymond, Meroni et al. 2001, Meroni and Diez-Roux 2005, Nisole, Stoye et al. 2005). Like *TRIM5*, several other *TRIM* genes have been implicated in innate immunity and antiviral defense (reviewed in (Nisole, Stoye et al. 2005, Ozato, Shin et al. 2008, Johnson and Sawyer 2009, Kawai and Akira 2011, McNab, Rajsbaum et al. 2011)). However, the majority of *TRIM* genes remain largely uncharacterized, along with their potential for encoding antiviral activities.

Previous studies have used functional characterizations to identify *TRIM* gene family members that encode antiviral activity. For example, a screen of a subset of human and mouse *TRIM* genes highlighted hitherto unidentified members that positively or negatively impacted retroviral fitness (Uchil, Quinlan et al. 2008). Other functional characterizations have focused on hallmarks of restriction factors, including induction on interferon treatment (Carthagena, Bergamaschi et al. 2009, Uchil, Hinz et al. 2012). While

candidate restriction factors were identified from each of these approaches, functional identification of novel restriction factors in the *TRIM* gene family is complicated due to a number of reasons. First, multiple alternatively-spliced transcripts are produced from each *TRIM* gene. PML, for instance, is only one of eleven TRIM19 protein isoforms while TRIM5alpha is the longest of at least nine reported transcripts of the *TRIM5* gene (Reymond, Meroni et al. 2001, Brennan, Kozyrev et al. 2007, Battivelli, Migraine et al. 2011), but the only protein isoform with antiviral activity. Homodimerization of TRIM5alpha with other TRIM5 isoforms (gamma, delta, and iota) causes dominant negative suppression of the antiviral activity of TRIM5alpha (Stremlau, Owens et al. 2004, Passerini, Keckesova et al. 2006, Battivelli, Migraine et al. 2011), so antiviral activity requires that the correct isoform or combination of isoforms be appropriately expressed in the cells being assayed. Second, viral restriction specificity may further impede identification of antiviral function especially for those restriction factors that act directly at the host-virus interface (like TRIM5alpha) compared to those that may indirectly affect the immune response (like PML); for the former case, detection of antiviral activity would depend on the right combination of *TRIM* genes and viruses. For instance, whereas rhesus macaque *TRIM5* has potent antiviral activity against HIV-1, the human ortholog only has relatively modest effects (Stremlau, Owens et al. 2004).

In order to bypass these difficulties associated with a functional screening approach, here we have taken a complementary, evolutionary approach to identify candidate antiviral restriction factors in this family. This approach exploits a common feature of restriction genes: the unique selective pressures they are subjected to by virtue of their antagonistic relationship with viral pathogens (Meyerson and Sawyer 2011, Daugherty and Malik 2012). Any mutation that improves the ability of an antiviral gene to recognize the virus is advantageous to the host. In contrast, the virus selectively favors mutations that weaken or destroy this interaction. Repeated rounds of mutation in which one party increases affinity while the other party decreases affinity can lead to rapid evolution at the protein-protein binding interface. Specifically, such interactions will result in the rapid accumulation of changes at non-synonymous (amino acid-altering) positions in coding DNA compared to the relatively benign mutations at synonymous sites, a selective regime referred to as positive selection. Such positive selection analysis was successfully used to precisely identify the region of TRIM5alpha that determines its specificity for different retroviral capsids (Sawyer, Wu et al. 2005). Importantly, positive selection has also been detected in nearly all other known restriction factors that directly interact with viral proteins (reviewed

in (Duggal and Emerman 2012)). Indeed, signals of adaptive evolution are often a hallmark amongst restriction factors with roles at the direct interface of host-pathogen interactions.

Here, using reference genomes, we analyzed members from the *TRIM* gene family for positive selection in primates. Via our evolutionary screen, we recovered both *TRIM* genes previously identified to be under positive selection due to their antiviral role (Sawyer, Wu et al. 2005, Sawyer, Emerman et al. 2007), four antiviral genes whose evolutionary signatures were previously unknown (e.g., *TRIM25* (Gack, Shin et al. 2007)), and as many as ten novel antiviral genes that have not been previously identified in any analyses. We also present a more detailed analysis of the most intriguing restriction factor candidate revealed by our screen, *TRIM52*. *TRIM52* lacks a C-terminal B30.2 domain, but encodes a massively expanded RING domain that we find has been subject to intense positive selection. Our analysis of *TRIM52* evolution reveals its age and birth via a partial duplication of the *TRIM41* gene, followed by independent loss or pseudogenization of *TRIM52* in multiple mammalian and primate lineages. Based both on the strong signatures of adaptive evolution, and the recurrent losses, we propose that *TRIM52* represents a novel, non-canonical antiviral *TRIM* gene in primate genomes with unique specificity determined by the rapidly evolving RING domain. Our evolutionary screen to identify novel restriction factors reveals several intriguing candidates that warrant further study to fully elucidate the role played by *TRIM* genes either directly or indirectly in mediating antiviral defense.

3.2 Results

3.2.1 Positive selection has acted on several TRIM genes in primates

To screen the *TRIM* gene family for signatures of having participated in an evolutionary arms race, we evaluated *TRIM* orthologs from primates for recurrent positive selection via maximum likelihood analyses using the CODEML program from the PAML package (Yang 2007). We compared *TRIM* orthologs from human, chimpanzee, orangutan, rhesus, and marmoset genomes and identified genes where specific residues have recurrently evolved under positive selection throughout primate history (Table 3.1; Table B.1). In some instances, we were able to identify additional orthologs from other primate genome sequencing projects that are underway via Ensembl (Vilella, Severin et al. 2009) or from previous gene-directed sequencing efforts. For a few TRIM genes, we were unable to identify the full complement of orthologs, as the genes are absent or not intact in the available genome assemblies. Our screen identified 16 out of 65 genes as having evolved under positive selection using a p-value cutoff of 0.05: *TRIM2*, *TRIM5, TRIM7, TRIM10, TRIM15, TRIM21, TRIM22, TRIM25, TRIM31, TRIM38, TRIM52,*

TRIM58, TRIM60, TRIM69, TRIM75, and *TRIM76* (Table 3.1; Table B.1). Among these recovered members, *TRIM5* and *TRIM22* represent a bona-fide and suspected restriction factor, respectively, that were previously reported to show strong evidence of positive selection (Stremlau, Owens et al. 2004, Sawyer, Wu et al. 2005, Sawyer, Emerman et al. 2007, Barr, Smiley et al. 2008).

Our screen recovered known restriction factors *TRIM15, TRIM21, TRIM25* and *TRIM38*. *TRIM25* activates RIG-I signaling via ubiquitination (Gack, Shin et al. 2007), and *TRIM25*-mediated signal transduction is known to be inhibited by the direct interaction of Influenza A protein NS1 to the Coiled-Coil domain of TRIM25 (Gack, Albrecht et al. 2009). Intriguingly, we find that *TRIM25* exhibits a number of sites under positive selection with notable clusters in between the Coiled-Coil and B30.2 domains (Figure 3.1). Based on this, we speculate that the positive selection exhibited by *TRIM25* indicate sites influencing its interactions with viral antagonist proteins. *TRIM15* was discovered in a knockdown screen to inhibit the release of retroviruses (Uchil, Quinlan et al. 2008), and later found to have a role in the RIG-I sensing pathway (Uchil, Hinz et al. 2013). We find sites of positive selection within the RING, Coiled-Coil, and B30.2 domains (Figure 3.1). *TRIM21* is able to degrade viruses via an intracellular antibody-mediated (Mallery, McEwan et al. 2010) and positive selection was detected within the B-Box and B30.2 domains (Figure 3.1). *TRIM38* is known to negatively regulate innate immunity by targeting TRIF (Xue, Zhou et al. 2012), NAP1 (Zhao, Wang et al. 2012), and TRAF6 (Zhao, Wang et al. 2012) for ubiquitination and degradation. *TRIM38* is also suspected to improve the fitness of HIV-1 during entry by an unknown mechanism (Uchil, Quinlan et al. 2008). Only a single site of positive selection residing between the Coiled-Coil and B30.2 domains was detected (Figure 3.1). Based on the location of positive selection outside of the notable domains, little can be inferred. Similar to *TRIM25*, *TRIM15* may be targeted by a viral antagonist based on its role in RIG-I signaling, so positive selection may be a reflection of an evolutionary trajectory to evade recognition. *TRIM21* is likely to be rapidly evolving to maintain recognition of viral proteins for proteasomal degradation. Thus, we are able to apply our evolutionary screen to further characterize known restriction factors, offering insight into specific domains of host-virus interaction interfaces.

The majority of the known restriction factors that we identified via our analysis belong to the C-IV family of *TRIM* genes based on their domain structure (Ozato, Shin et al. 2008). Amongst the candidates that also belong to the C-IV family, we recovered *TRIM7*, *TRIM10*, *TRIM58*, *TRIM60, TRIM69,* and *TRIM75*. A single site was found to be under positive selection in *TRIM7* between the Coiled-Coil and B30.2

domains (Figure 3.1); however, this site of positive selection is uninformative as it does not reside in any of the defined domains. *TRIM75* has a single informative site of positive selection in its RING domain associated with E3 ubiquitin ligase activity. Positive selection was found in the Coiled-Coil and B30.2 domains of *TRIM10*. We can only speculate that positive selection within these domains of *TRIM10* may reflect changes to target recognition (Sawyer, Wu et al. 2005, Maillard, Ecco et al. 2010). *TRIM58* and *TRIM60* presented the most sites of positive selection amongst the candidate C-IV family members (Figure 3.1). *TRIM58* presented a cluster of sites in the B30.2 domain. *TRIM60* had sites throughout its gene and sites in each of the notable domains: RING, B-Box, Coiled-Coil, and B30.2. Moreover, we found that many of these positively selected sites formed clusters in the Coiled-Coil and B30.2 domains. While classification to any particular family of *TRIM* genes is not indicative of a restriction factor, there is precedent for identifying restriction factors within the C-IV family that warrants further investigation into these.

Outside of the C-IV family, the candidate restriction factors highlighted by our screen were *TRIM2* (C-VII), *TRIM31* (C-V), *TRIM52* (C-V), and *TRIM76* (UC). *TRIM2* is the only recovered *TRIM* gene containing a Filamin-type immunoglobulin domain and array of NHL repeats. Two sites of positive selection were found in *TRIM2*, with one of these residing within the beginning of the NHL repeats (Figure 3.1). *TRIM31* is implicated in inhibiting HIV-1 entry and the release of MLV, though the details of restriction by *TRIM31* is a suspected retroviral restriction factor that acts at the stages of entry and release (Uchil, Quinlan et al. 2008). We identified three sites exhibiting signatures of positive selection within regions not associated within the defined domains of *TRIM31* (Figure 3.1). As the mechanism of restriction by *TRIM31* has yet to be elucidated, it is not clear whether these sites of positive selection are the consequence of viral pressure or some other selective force. *TRIM76* encodes for a large protein that contains B-Box, Coiled-Coil, Fibronectin III, and B30.2 domain in the C-terminus region of the protein. The bulk of the protein does not contain homology to any known domain; however, it is within this region of ~3,500 amino acids that we identified six sites of positive selection. *TRIM52* is unique amongst the restriction factor candidates as it only encodes the RING and B-Box domains. We identified the majority of positive selection within the RING domain and a single site immediately upstream of the B-Box domain. Intriguingly, the RING domain of *TRIM52* has expanded and is the largest amongst the genes recovered by our screen (Figure 3.1). We find that within this expanded region of the RING domain resides the rapidly evolving site.

Table 3.1 Primate *TRIM* genes recovered via PAML screen

TRIM gene	M7vsM8	p-value	% of positively selected sites	Average dN/dS for selected sites	Positively selected sites	# of taxa
TRIM2	7.306736	0.025903738	0.645	1.54007	98, **497**	12
TRIM5	73.47338	<0.005	20.464	3.29159	**7**, 139, **175, 182, 213, 215, 228, 257, 258, 310, 311**, 317, **324, 379**, 381, **382, 418, 421, 423, 471, 483**	22
TRIM7	9.924556	0.006996971	0.364	11.03782	**258**	10
TRIM10	6.060552	0.048302305	1.999	3.61613	152, **329**	11
TRIM15	10.736928	<0.005	5.835	2.24534	**18, 42, 150, 460**	11
TRIM21	7.01864	0.029917251	3.378	4.81422	**124, 407**	10
TRIM22	10.195488	<0.005	4.887	6.16845	**99, 171, 220, 308**	13
TRIM25	19.270786	<0.005	9.368	2.40224	58, 259, **297, 338, 377,** 415, **418**, 420, 435	10
TRIM31	15.056244	<0.005	5.265	8.68452	**72, 227, 250**	7
TRIM38	6.511954	0.038543146	3.655	2.94482	**215**	12
TRIM52	16.514108	<0.005	6.635	5.8355	**75**, 111, 149, 153, 221	7
*TRIM52**	11.500758	<0.005	3.615	6.88558	**75, 149**, 221	14
TRIM58	17.753758	<0.005	4.24	2.33532	**223, 443, 472, 475, 480**	10
TRIM60	8.00201	0.018297241	20.431	2.14558	8, **82, 96, 134**, 200, 251, 252, **255**, 264, 271, **302**, 322, 370, 405, **459**	11
TRIM69	6.650254	0.035967951	19.156	2.46053	14, 158, **192, 226, 246, 261**, 285, 353, 371, 473	10
TRIM75	10.614008	<0.005	0.665	12.23692	**45, 227**	10
TRIM76	42.288758	<0.005	1.711	7.94347	**306, 651, 1507, 2727, 2797, 3314**	10

Sites marked in bold and underline were also found to be under positive selection by Datamonkey (Delport, Poon et al. 2010)

*PAML was executed using a combination of orthologs retrieved from online databases and sequencing.

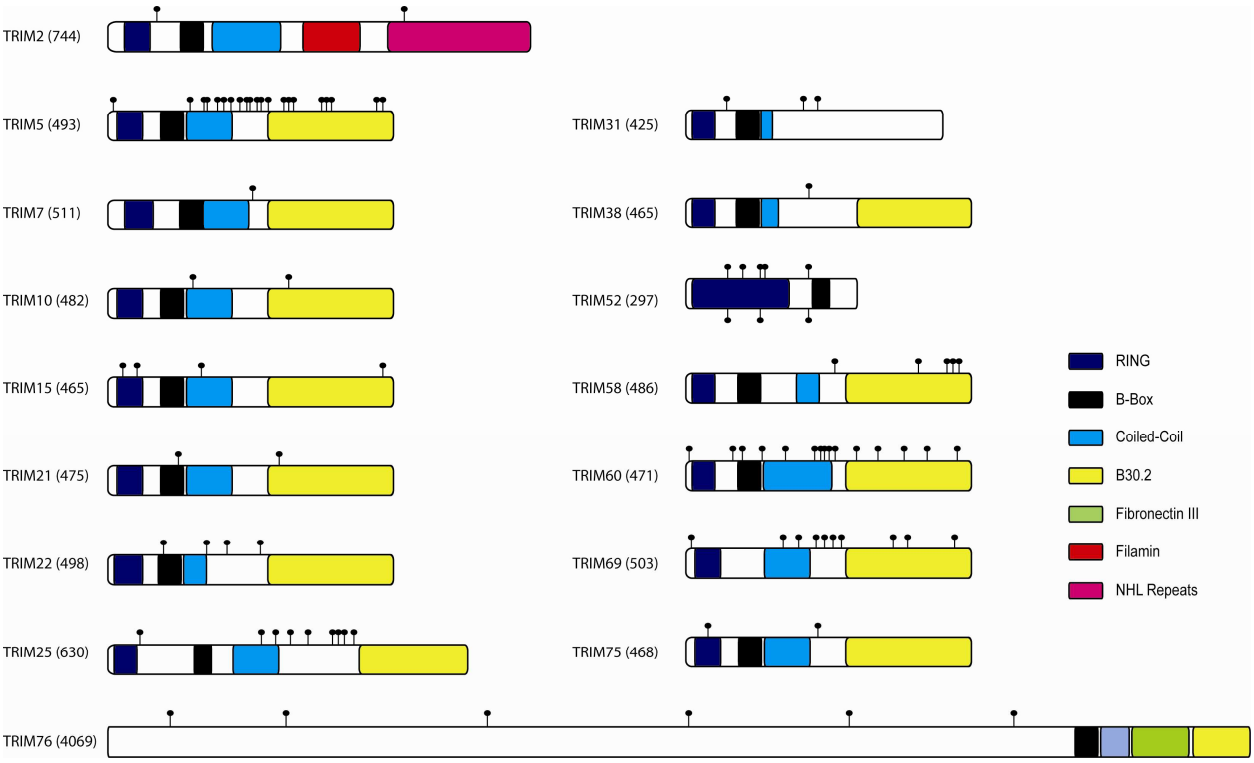


Figure 3.1 Architecture of *TRIM* family members exhibiting positive selection. Locations, lengths, and nomenclature of protein domains are based on GENBANK and ENSEMBL reports. Sites of positive selection are marked with lollipops. The sites identified by the in-depth analysis of *TRIM52* are lollipops on the underside of the protein representation.

Thus, our evolutionary screen for novel restriction factors amongst the *TRIM* gene family identified 14 members not previously known to be under positive selection and as many as 10 novel candidates. Though not all of these *TRIM* genes are expected to reveal host-pathogen interactions, the extent of positive selection found and the recovery of known restriction factors support the hypothesis that some of these candidates will have a role in host-pathogen interactions.

3.2.2 Rapid evolution of the TRIM52 RING domain in primates

We selected to evaluate *TRIM52* in more detail since it structurally deviated the most from the canonical *TRIM* restriction factors (i.e. *TRIM5* and *TRIM22*). For example, *TRIM52* lacks the viral recognition (B30.2) domain and displays signatures of rapid evolution within the RING domain. Moreover, *TRIM52* appears to lack an intact Coiled-Coil domain within its coding region (Figure 3.1). Thus, *TRIM52* is comprised solely of the RING and B-Box2 domains, making it a highly unusual member of the TRIM multigene family. Even the RING domain of *TRIM52* is highly unusual. RING domains of the *TRIM* family are generally defined by the consensus sequence Cx2Cx9-45Cx1-3Hx2-3Cx2Cx4-48Cx2C where eight cysteine, histidine, or aspartic acid residues coordinate two zinc atoms (Meroni and Diez-Roux 2005). The region between the sixth and the seventh coordinating cysteine residues is referred to as the "loop 2" region of the RING tertiary structure using the precedent of the human c-cbl RING-containing E3 ubiquitin ligase (Zheng, Wang et al. 2000). The majority of *TRIM* genes encode between 4-48 amino acids in their "loop2" region, with the mode being 13 amino acids (Figure 3.2). However, several *TRIM* genes were found to deviate from the consensus range. Most notably, *TRIM52* encodes 139 amino acids in its "loop 2" region (Figure 3.2). Thus, *TRIM52* encodes the largest RING domain of any *TRIM* gene encoded in the human genome. BLAST (Altschul, Gish et al. 1990) analysis of this region reveals similarity only to mammalian *TRIM52* and *TRIM41* genes, both of which have exceptionally large RING domain expansions.

In order to elucidate the evolutionary relationship between *TRIM52* and *TRIM41*, and to deduce when this large "loop 2" RING expansion occurred, we carried out phylogenetic analyses of *TRIM52* and *TRIM41* sequences that were obtained from BLAST (Altschul, Gish et al. 1990) searches of vertebrate genomes. Our analyses revealed that *TRIM52* and *TRIM41* are close paralogs in mammalian genomes (Figure 3.3). We found that the reptile (anole lizard), avian (chicken and wild turkey), and marsupial (Tasmanian devil and opossum) genomes have only single *TRIM41*-like genes, which are phylogenetic outgroups to both the *TRIM52* and *TRIM41* clades from eutherian mammals (Figure 3.3). This suggests

that *TRIM52* was born in eutherian mammals ~190 million years ago via a partial duplication of *TRIM41*, having lost both the Coiled-Coil and B30.2 domains at birth (Meredith, Janečka et al. 2011).

Despite their evolutionary relationship, our screen for positive selection in primates highlighted *TRIM52*, but not *TRIM41*. To further evaluate the evolutionary history of *TRIM52* we repeated our analysis of recurrent, site-based positive selection via maximum likelihood analyses using primate *TRIM52* orthologs obtained by our own sequencing efforts. From this in-depth analysis, we refined the sites of recurrent, codon-based positive selection (Figure 3.1; Table 3.2). The sites of positive selection reside primarily within the expanded "loop 2" region of *TRIM52*. This rapid evolution of the RING domain is especially evident in an evolutionary comparison focused on the "loop 2" expansion unique to *TRIM41* and *TRIM52*, which highlights the dramatic acceleration of amino acid replacements in *TRIM52* (Figure 3.4A; 3.4B). However, we found no evidence of positive selection having acted on the *TRIM41* using available primate sequences from databases (Figure 3.4A; Table 3.2). Thus, in stark contrast to *TRIM41* and its RING domain that has been evolving under constraint, we find that *TRIM52* has been rapidly evolving throughout primate history, with much of that selection acting on the RING domain.

3.2.3 Repeated loss/pseudogenization of TRIM52 in mammals

Our sequencing survey also revealed at least two instances of *TRIM52* loss or pseudogenization over the course of primate evolution (Figure 3.5). For instance, within marmoset and other New World monkey genomes, we were only able to identify exon 2 using a combination of BLAST (Altschul, Gish et al. 1990) and BLAT (Kent 2002) searches (Figure B.1), and our own PCR analyses, suggesting that *TRIM52* is present but pseudogenized throughout the New World Monkey lineage. We were also unable to detect *TRIM52* from gibbon genomes (*Nomascus leucogenys*, *Hylobates agilis*, and *Symphalangus syndactylus*) via PCR with gDNA, despite using PCR primers that amplified *TRIM52* from all other Hominoids and Old World monkeys. BLAST (Altschul, Gish et al. 1990) and BLAT (Kent 2002) analyses support the absence of *TRIM52* from publicly available gibbon genomes.

This pattern of stochastic *TRIM52* loss was also evident in other mammalian orders. We identified *TRIM52* pseudogenization in African elephant (*Loxodonta africana*), which contains premature nonsense mutations (Figure B.1). We were also unable to identify *TRIM52* throughout the Glires (Rodentia and

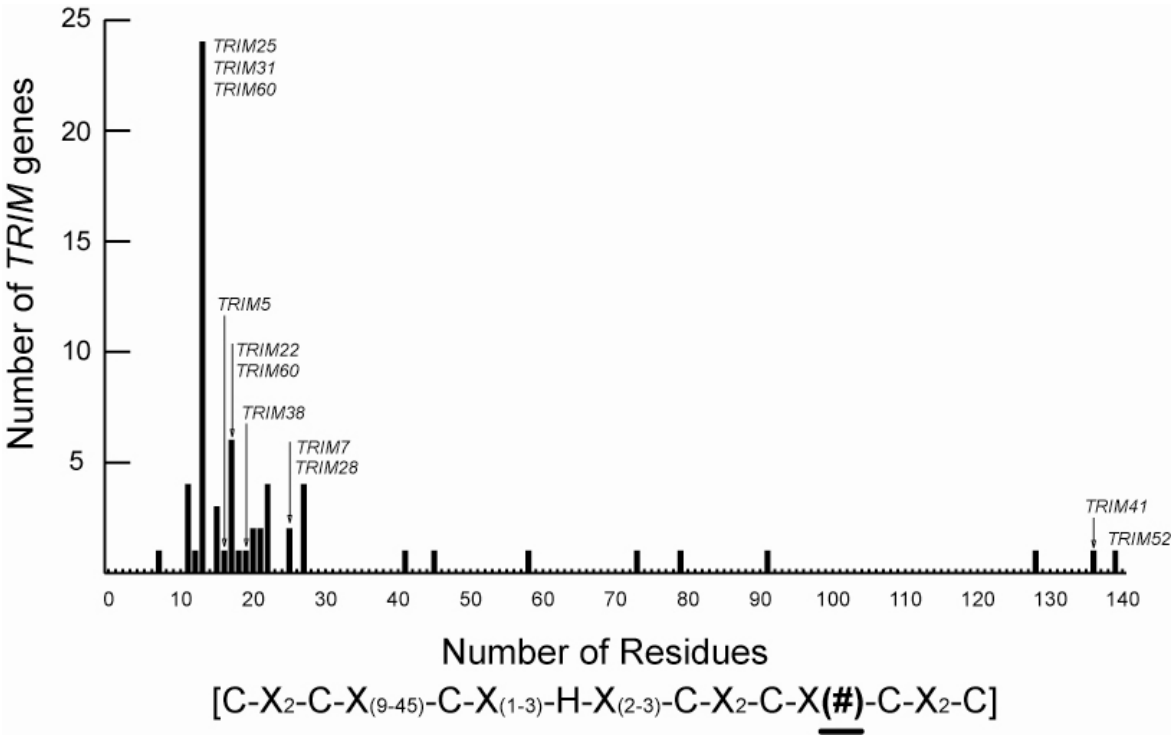


Figure 3.2 Variability in the length of the RING domain. The RING domains from 73 annotated human TRIM genes were collected from Ensembl (Flicek, Amode et al. 2012) and GenBank, and evaluated to determine the length of a variable region located within the domain. Alignments of homologous regions were built using ClustalX (Larkin, Blackshields et al. 2007) and the number of residues residing in the variable region was counted by hand. The predicted length of this variable region ranges from 7-74 amino acids. *TRIM52* and *TRIM41* have the largest expansion of their RING domains.

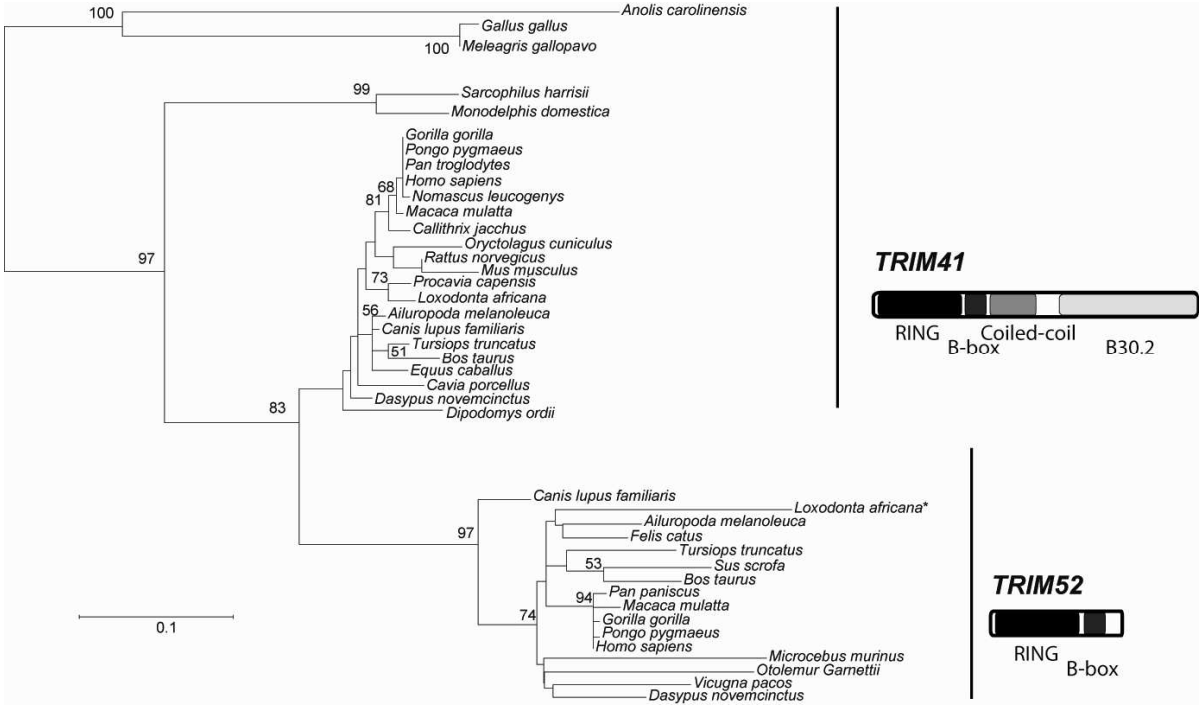


Figure 3.3 Phylogenetic relationship of *TRIM52* and *TRIM41.* A phylogram of homologous regions of the RING and B-Box2 domains from *TRIM52* and *TRIM41* orthologs was built using a maximum likelihood based approach via PhyML (Guindon, Dufayard et al. 2010). Statistical support is represented by bootstrap values, collected from 100 iterations. The "*" symbol denotes the presence of non-sense mutations that result in pseudogenization.

Lagomorpha) lineage of mammals, suggesting it has been deleted early within this lineage. However, utilizing UCSC (Kent, Sugnet et al. 2002) and Ensembl (Flicek, Amode et al. 2012) predictions, we were able to recover *TRIM52* from the genomes of the mouse and rat. Sequence analysis of these predicted mouse and rat *TRIM52* revealed that they do not encode a B-Box domain. Therefore, the annotated mouse and rat *TRIM52* are only comprised of a RING domain. Furthermore, the *TRIM52* orthologs we did identify in mouse and rat genomes were not located proximal to *TRIM41* and are therefore the only non-syntenic *TRIM52* orthologs in mammals (Figure B.2). When we included mouse and rat *TRIM52* in our phylogenetic analysis (Figure B.3), branch support at the node separating the *TRIM41* and *TRIM52* clades was lowered (even though mouse and rat *TRIM52* genes localized within the *TRIM52* clade). Due to the apparent loss of *TRIM52* throughout glires, the truncated structure of mouse and rat *TRIM52*, and their ambiguous phylogenetic placement, we therefore cannot confidently assign these mouse and rat *TRIM* genes as bona fide *TRIM52* orthologs, labeling them *TRIM52*-like instead (Figure B.1). Accordingly, we omitted the mouse and rat *TRIM52*-like sequences from our phylogenetic analysis (Figure 3.3). Additional genome sequencing within eutherian mammals may reveal still additional instances of *TRIM52* loss or pseudogenization, suggestive of episodes of relaxed selective pressure amongst individual lineages.

3.2.4 Preservation of TRIM52 in humans

Given the pattern of recurrent pseudogenization and the sub-telomeric position of *TRIM52* in the human genome (Figure B.2), we explored whether there is pseudogenization of *TRIM52* in human populations. We identified 27 reported indels (insertions, deletions, or single nucleotide polymorphisms) from NCBI's database of short genetic variations (dbSNP) throughout the 2 exons that encode *TRIM52* (Table B.2). Of the 27 reported variations, 21 represent missense mutations or codon insertions/deletions. We identified a single case of the start codon being reported as a missense mutation resulting in the methionine being converted to an isoleucine. A non-sense mutation corresponding to the 83rd codon of *TRIM52* was reported and validated by the 1000 Genomes Project, which would result in the truncation of over two-thirds of the open reading frame. The remaining 6 reported variations were synonymous changes and would cause no functional change to *TRIM52*. Thus, we identified only 2 possible loss-of-function variants from dbSNP. In addition, our own sequencing efforts of a human diversity panel (24 African American; 24 European Caucasian; 24 Han Chinese) only recovered SNPs consistent with several of the synonymous change or codon deletion variants. These data suggest that *TRIM52* has been largely preserved in humans and its role is still actively maintained.

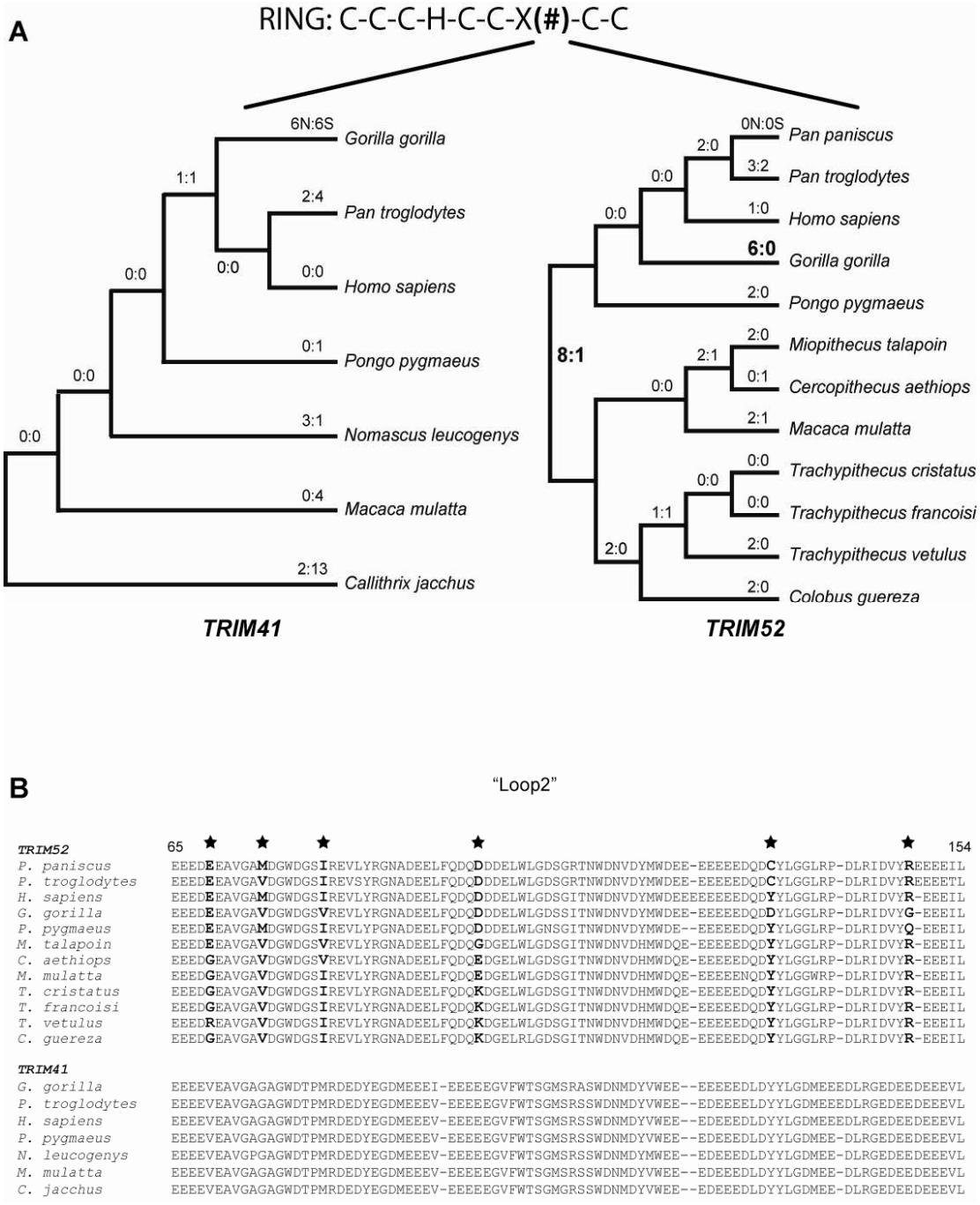


Figure 3.4 Positive selection within the RING domain of *TRIM52.* *(*A) More than half the sites predicted to be evolving under positive selection (Table 3.2) are located within the RING domain of *TRIM52*. To further highlight this, we identified the number of synonymous (S) and non-synonymous (N) substitutions that have occurred in the expanded "loop 2" region of *TRIM52* in primate evolution (the equivalent domain of *TRIM41* is shown for comparison). Examples of dramatic episodes of positive selection in *TRIM52's* RING domain are highlighted in bold. (B) The differences in evolutionary signals are further demonstrated by an alignment of a 90 amino acid long stretch of the "loop 2" region from primate *TRIM41* and *TRIM52*. Sites of positive selection (Table 3.2) have indicated with a star and bolded.

Table 3.2 Maximum likelihood analyses of *TRIM41* and *TRIM52* genes in primates

TRIM gene	M7 vs M8 (2lnλ; p value)	% sites with dN/dS>1 (avg. dN/dS for those sites)	Positively selected sites (posterior probability)	Primate sequences utilized for comparison
TRIM41	0; p=1.00	0 (n.a)	none	**Hominoid**: *Homo sapiens; Pan paniscus; Pan troglodytes; Gorilla gorilla; Pongo pygmaeus; Nomascus leucogenys* **Old World monkeys**: *Macaca mulatta; Papio anubis* **New World monkeys**: *Callithrix jacchus; Saimiri boliviensis* **Prosimian:** *Otolemur garnettii*
TRIM52 *(RING domain expansion only)*	34.75; p<0.005	18.93 (12.51)	**75(1.000)***; 82(0.965); **100(0.998)***; **111(0.997)***; 134(0.966); **149(0.997)***; **153(0.993)***	**Hominoid**: *Homo sapiens; Pan troglodytes; Pan paniscus; Gorilla gorilla; Pongo pygmaeus* **Old World monkeys:** *Cercopithecus aethiops; Miopithecus talapoin; Macaca mulatta; Papio anubis; Trachypithecus vetulus; Trachypithecus cristatus; Trachypithecus francoisi; Colobus guereza* **New World monkeys**: none (all pseudogenized) **Prosimian:** *Microcebus murinus*

3.2.5 Human and rhesus TRIM52 do not restrict lentiviruses

The history of positive selection uncovered amongst primate *TRIM52* orthologs indicates that its function has been adaptively evolving. While many members of the *TRIM* family positively and negatively impact retroviruses (Uchil, Quinlan et al. 2008), it is unclear whether *TRIM52* reflects this. Indeed, given the degree of adaptive evolution within the RING domain of *TRIM52*, it appears that the role of viral recognition has shifted from the absent B30.2 domain to the RING domain and that the target is likely not retroviral. To assess this, we evaluated human and rhesus *TRIM52* orthologs for antiviral activity against a limited panel of lentiviruses (Figure B.4). Although many of these viruses are restricted by *TRIM* proteins, we found no evidence of restriction by either human or rhesus *TRIM52*. Thus, although the evolutionary patterns of positive selection and episodic loss are consistent with the function of *TRIM52* in some form of genome defense, the target(s) of this activity is still unknown.

3.3 Discussion

3.3.1 TRIM52, a novel antiviral gene highlighted by unique genetic innovation

From a screen of positive selection within the *TRIM* gene family, we recovered 14 members not previously known to be evolving under positive selection. Of these, *TRIM52* exhibited signatures of novel rapid evolution amongst *TRIM* genes. In the absence of these evolutionary analyses, *TRIM52* might not draw attention as a candidate antiviral factor because of the lack of a canonical virus-interaction domain. While *TRIM52* lacks B30.2 and Coiled-Coil domains, the gene bears an ancient expansion of the RING domain that exhibits positive selection. We previously described an expanded set of rodent *TRIM5* paralogs, and highlighted mouse (*Mus musculus*) *TRIM12* that only encodes RING, B-Box2, and Coiled-Coil domains (Tareen, Sawyer et al. 2009). Similar to *TRIM52*, mouse *TRIM12* exhibits signatures of positive selection despite the absence of a recognized interaction interface (i.e. B30.2 domain). Indeed, our finding of positive selection within the RING domain leads to the intriguing model whereby the antiviral interaction interface of *TRIM52* may have now shifted to within its RING domain. This is an unusual exception to the highly modular arrangement of the mammalian and fish *TRIM* gene family in which the target interaction interface is usually restricted to the Coiled-Coil or B30.2 domains, which are also the hotspots for positive selection (Reymond, Meroni et al. 2001, Meroni and Diez-Roux 2005, Nisole, Stoye et al. 2005, Song, Gold et al. 2005, Yap, Nisole et al. 2005, Sawyer, Emerman et al. 2007, van der Aa, Levraud et al. 2009).

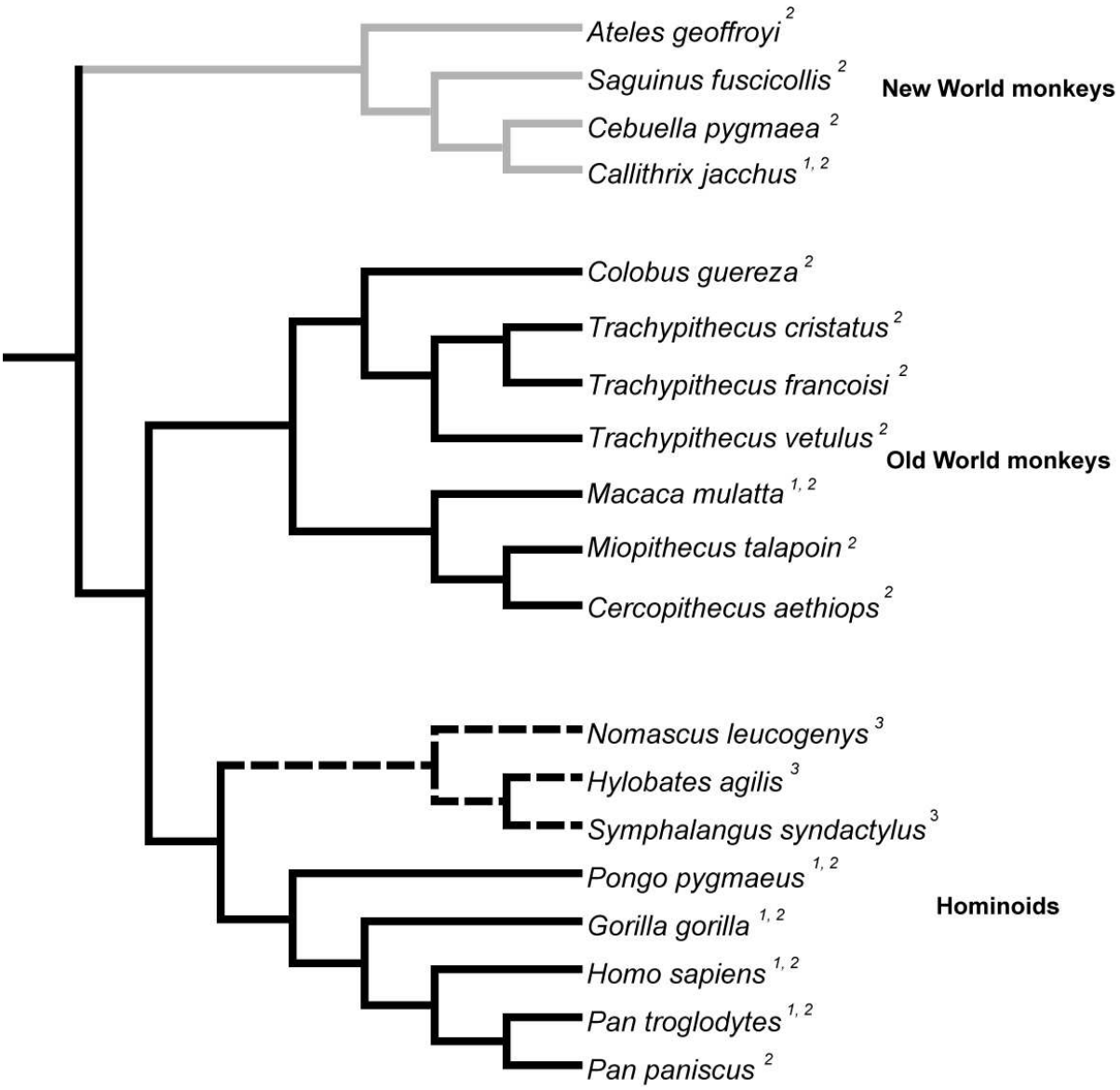


Figure 3.5 Presence/Absence of *TRIM52* in primates. We evaluated *TRIM52* from a range of Hominoids, Old World monkeys, and New World monkeys, using sequences collected from (1) Ensembl (Flicek, Amode et al. 2012) and Genbank and via (2) PCR. Primates surveyed by our analysis are presented in a guide tree of the well-accepted primate phylogeny (Perelman, Johnson et al. 2011). (3) We were unable to amplify *TRIM52* from the gibbon lineage of Hominoids (represented by dotted branches), despite the use of primers that we used to amplify orthologs from other Hominoids and Old World monkeys. Greyed branches represent lineages where we observed *TRIM52* to be pseudogenized.

Despite the strong signature of positive selection, we identified at least four independent losses of *TRIM52* within mammals, with half of these events occurring in primates. The absence of *TRIM52* from gibbon genomes may reflect its genomic position, proximal to the telomeric region in Hominoids and Old World monkeys. However, this genomic positioning is not shared in other mammals (Figure B.2) and therefore cannot account for the multiple loss events we have observed. Furthermore, we found no evidence for either loss or pseudogenization of the proximally located *TRIM41* gene. This suggests that the parental gene is under strong functional constraint, while the episodes of *TRIM52* loss strongly suggest that this *TRIM* gene does not carry out a conserved, housekeeping function in mammalian genomes. Intriguingly, the recurrent loss of *TRIM52* demonstrates the dynamic evolutionary history observed by other *TRIM* genes with antiviral function. For instance, the dog *TRIM5* ortholog is pseudogenized (Sawyer, Emerman et al. 2007), while cats encode a truncated form of *TRIM5* with a disrupted B30.2 domain; both lineages are unable to express TRIM5alpha (McEwan, Schaller et al. 2009). Both rodent (mouse and rat) and cow genomes lack *TRIM22* orthologs, but contain expanded sets of *TRIM5* paralogs (Sawyer, Emerman et al. 2007, Tareen, Sawyer et al. 2009). Expansions are not unique to *TRIM5*. Han et al (Han, Lou et al. 2011) identified several *TRIM* genes that are copy number variable in human genomes. Similar dynamics have also been observed in several teleost species, where unique *TRIM* genes (fintrims) have expanded and diversified in each lineage (van der Aa, Levraud et al. 2009). Thus, considering the dynamic history of *TRIM52* and our evidence of positive selection, we posit that the unusual *TRIM* gene is involved in genome defense, but which can be lost either due to relaxed selection or because of the high costs borne by encoding such a defense (Sawyer, Wu et al. 2006). We have previously suggested that positive selection and the expansion of *TRIM* genes is driven by new or continuous selective pressure, likely provided by viral pathogens (Sawyer, Emerman et al. 2007, Tareen, Sawyer et al. 2009, Han, Lou et al. 2011). Similarly, the loss or relaxation of such a selective pressure could result in the loss of a *TRIM* gene (Sawyer, Wu et al. 2006, Sawyer, Emerman et al. 2007).

It is also formally possible that *TRIM52* is under positive selection not because of antiviral activity but instead to maintain its interaction with a 'host' target substrate that is also adaptively evolving. However, we find this co-evolutionary scenario unlikely because such host-host interaction surfaces are not typically found to evolve under positive selection unless they are challenged by a pathogenic influence (Koyanagi, Kerns et al. 2010, Daugherty and Malik 2012). Furthermore, this scenario would posit that the many incidences of *TRIM52* loss we have documented would have to coincide with the simultaneous loss of the target substrate or the need to maintain the interaction.

3.3.2 Positive selection within the TRIM gene family

Based on the unbiased approach of our screen, we predicted the recovery of several known restriction factors. In particular, there was an expectation of identifying *TRIM5* and *TRIM22*, both previously highlighted for their positive selection (Sawyer, Wu et al. 2005, Sawyer, Emerman et al. 2007). In addition to these, we recovered other known or suspected restriction factors: *TRIM15, TRIM21*, *TRIM25*, and *TRIM38*. We detected positive selection occurring all along *TRIM25*, in particular within the Coiled-Coil and B30.2 domain (Figure 3.1). *TRIM25* plays a role in influenza infection, where its activity is critical for the activation of the RIG-I dependent signaling cascade (Gack, Shin et al. 2007). Specifically, Influenza A encodes protein NS1 that directly interacts and inhibits *TRIM25* at the Coiled-Coil domain inhibiting the ubiquitination and activation of RIG-I. This is reminiscent of adaptive evolution in other known restriction factors, such as in the case of MAVS to evade protease cleavage by Hepatitis C virus (Patel, Loo et al. 2012) or in Tetherin to evade lentivirus Nef or Vpu (Lim, Malik et al. 2010). In both cases, positive selection highlights regions of the host-encoded protein targeted by viral antagonists and provided insight into mechanisms of host evasion. Thus, it is likely that the sites of positive selection exhibited by *TRIM25* reveal adaptation during primate history to evade NS1 or NS1-like antagonist. *TRIM15* similarly plays a role regulating innate immune signaling. As we highlighted *TRIM15* in our screen for positive selection, we posit that the rapid evolution is the consequence of selective pressure from the direct interaction with viral proteins, either from targeting or antagonism. While we cannot differentiate solely on the profile of positive selection, additional functional work will be required to shed further light on the antiviral role of *TRIM15*. *TRIM21* is able to target cytosolic antibodies bound to viruses and autoubiqutinate, leading to the proteasomal degradation of the *TRIM21* bound complex (Mallery, McEwan et al. 2010). As this complex forms from *TRIM21* binding to the invariant region of antibodies (James, Keeble et al. 2007), it is unlikely that the interaction between host-encoded products is responsible for the positive selection we detected. Instead, it is much more likely that *TRIM21* is targeted by a viral antagonist and that positive selection is also reflective of evasion from viral antagonist. Unique among these recovered *TRIM* genes, *TRIM38* has been found to assist HIV-1 during entry (Uchil, Quinlan et al. 2008). This may be explained as *TRIM38* has recently been recognized for having a role in negatively regulating innate immunity by targeting components of innate immunity for degradation (Xue, Zhou et al. 2012, Zhao, Wang et al. 2012, Zhao, Wang et al. 2012). It is difficult to speculate. In the cases of these known restriction factors, our analysis of positive selection provides insight into the interface and nature of host-pathogen interactions, as well as clues regarding the evolution of present day function.

One *TRIM* gene that did not show a signature of positive selection that could be statistically supported is *TRIM19*/*PML*. This is in agreement with an extended sequencing of primate *TRIM19*/*PML* orthologs which concluded that there was no evidence for positive selection of this gene (Ortiz, Bleiber et al. 2006). This may be surprising in light of the evidence that PML functions as an antiviral (reviewed in (Nisole, Stoye et al. 2005)). However, positive selection would only be expected to act on genes encoding proteins which directly interact with viral proteins, and so any up-stream or down-stream effector may not present such a signal. It is interesting that *TRIM1* (MID2) also shows no adaptive signature, given that the human *TRIM1* protein has been shown to have moderate anti-MLV activity (Yap, Nisole et al. 2004). This may reflect a retroviral restriction that is not currently being utilized by humans or chimpanzees, since it was detected against a virus that does not infect these primates naturally. Nonetheless, it is important to point out that the absence of positive selection does not preclude *TRIM* genes from being candidate restriction factors, but those *TRIM* genes that have evolved under positive selection represent the most likely candidates for having an antiviral role.

As many of the *TRIM* genes remain largely uncharacterized, our evolutionary screen is able to highlight candidate restriction factors based on exhibition of positive selection, a hallmark of antiviral genes at the direct interface of the host-viral pathogen arms race (Daugherty and Malik 2012). In addition to *TRIM52*, *TRIM58* and *TRIM60* were recovered with intriguing profiles of positive selection. We found that *TRIM58* had a cluster of sites within the viral recognition (B30.2) domain, and *TRIM60* presented positively selected sites in each of the notable domains: RING, B-Box, Coiled-Coil, and B30.2 (Figure 3.1). Based on the extent of rapid evolution observed amongst these *TRIM* genes, these represent the most likely restriction factor candidates and should therefore be included in subsequent antiviral surveys of the gene family

Based on previous studies with APOBEC and *TRIM5* restriction genes, it is informative to identify antiviral restriction factors even if they are not currently active against modern viral pathogens. Restriction factors honed against evolutionarily “recent” viral infections might protect us against future viruses or viral variants, or might be artificially enhanced to be active against current forms. Genes with partial activity might vary in potency within the human population. Furthermore, such genes serve as barriers to animal models of viral infection (Hatziioannou, Princiotta et al. 2006, Kirmaier, Wu et al. 2010). To this end, our evolutionary approach to identify potential restriction factors in the *TRIM* family has revealed ten members that bear previously unrecognized signatures of recent positive selection.

These primate *TRIM* genes are therefore primate candidates to be investigated as novel restriction factors against viruses.

3.4 Methods

3.4.1 Collecting TRIM orthologs

Human (*Homo sapiens*) *TRIM* gene sequences were obtained from Ensembl (Flicek, Amode et al. 2012) and GenBank. Chimpanzee (*Pan troglodytes*), bonobo (*Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo abelii*), white-cheeked gibbon (*Nomascus leucogenys*), rhesus macaque (*Macaca mulatta*), baboon (*Papio anubis*), squirrel monkey (*Saimiri boliviensis*), marmoset (*Callithrix jacchus*), tarsier (*Tarsius syrichta*), mouse lemur (*Microcebus murinus*), and bushbaby (*Otolemur garnettii*) orthologs were obtained when reported from NCBI by BLASTing (Altschul, Gish et al. 1990) the "RefSeq RNA" databases with the human *TRIM* sequence as the query and from Ensembl gene orthology/paralogy predictions (Vilella, Severin et al. 2009). Additional primate orthologs were collected when available (e.g., African green monkey (*Chlorocebus aethiops*)). Subsequent collection of *TRIM* sequences, specifically *TRIM52* and TRIM41, via publically available databases were carried out utilizing Ensembl's genome databases to recover annotated sequences from available animals, including Reptilia, Avian, and Mammalian species.

3.4.2 Sequencing TRIM52

To expand our collection our collection of primate *TRIM52* sequences to improve the power of downstream evolutionary analysis, we amplified *TRIM52* using genomic DNA from the following primates: human, chimpanzee, bonobo, gorilla, orangutan, rhesus macaque, African green monkey, talapoin monkey (*Miopithecus talapoin*), colobus monkey (*Colobus guereza*), Francois' leaf monkey (*Trachypithecus francoisi*), purple-faced langur (*Trachypithecus vetulus*), and silvery langur (*Trachypithecus cristatus*). Exon 1 was amplified and sequenced using the following primer pair: Forward – CCACCGATCCCAGAGAGAGG & Reverse – CCTCTGGGGAAGCCAATCTGC. We amplified exon2 by nested PCR with the following primer pairs: Initial primer pair: Forward – GTYGCATGATTTAGAAYTTTACTGACCAA & Reverse – GACAATCCAGGCATCCAGTTATGC. Second, nested primer pair: Forward – ATWATGGTTTATTTAATAYARTATACATTATC & Reverse – GAACTCTAACTCATGGGATGGACAAA. The second, nested primer pair was used to sequence exon2. We used PCR Supermix (Invitrogen, 10790-020) for amplification reactions and performed 40 PCR cycles.

Sequencing reactions were carried out using BigDye. *TRIM52* sequences are being deposited in the GenBank database (accession numbers forthcoming).

3.4.3 Phylogenetic Analysis

Non-primate *TRIM52* and *TRIM41* sequences were obtained by BLAST (Altschul, Gish et al. 1990) analysis with the human *TRIM52* protein as query, and psi-blast (Altschul, Madden et al. 1997) analysis with the human *TRIM52* RING expansion as query. Psi-blast of the RING expansion recovers only *TRIM52* and TRIM41 orthologs, suggesting that these are the only *TRIMs* with this expansion. We found no evidence of a protein domain downstream of the B-Box2 domain, with homology to TRIM41, in any of the *TRIM52* orthologs. For instance, there is no identifiable Coiled-Coil domain or B30.2 domain downstream of the human *TRIM52* gene in the human genome assembly. All of the *TRIM41* sequences are predicted to encode a Coiled-Coil and B30.2 domain. Non-primate and primate *TRIM* sequences (*TRIM52* and *TRIM41*) that we recovered from BLAST (Altschul, Gish et al. 1990) and Ensembl (Flicek, Amode et al. 2012) were aligned using ClustalX (Larkin, Blackshields et al. 2007). We only included the RING (omitting the region containing the RING expansion) and B-Box domain. Using this alignment we constructed a tree using maximum likelihood methodology (Guindon, Dufayard et al. 2010) and used the program Dendroscope (Huson, Richter et al. 2007) to present a phylogram.

3.4.4 Delineation of TRIM protein domain boundaries and secondary structure

RING, B-box1, and B-box2 domains were identified based on the consensus sequences (Meroni and Diez-Roux 2005). Coiled-Coil domain boundaries were identified by predicting secondary protein structure with PSIPRED (McGuffin, Bryson et al. 2000) and identifying the long alpha-helix that is associated with this motif (Lupas 1996). B30.2 or other C-terminal domains were identified by using the CDD (Marchler-Bauer, Anderson et al. 2005) and SMART (Schultz, Copley et al. 2000) domain databases, and the N-terminal boundary of B30.2 domains was aided by secondary structure prediction, as the B30.2 domain consists entirely of sequential tandem beta-strands (Seto, Liu et al. 1999, Masters, Yao et al. 2006).

3.4.5 Computational analysis of positive selection

DNA sequences were aligned using ClustalX (Larkin, Blackshields et al. 2007). dN and dS for pairwise comparisons, as well as their confidence values, were calculated using the K-estimator software package (Comeron 1999). Sliding window analysis for human - chimpanzee comparisons were performed with a

window size of 600bp, and a slide size of 10bp. For chimpanzee - rhesus and human - rhesus comparisons, the window was adjusted to 300bp to account for the greater divergence between these species. A p-value was obtained for the window with the highest value of dN/dS. Detection of recurrent positive selection by multiple alignment comparisons was carried out using the CODEML program from the PAML package (Yang 1997). Codon-based modeling and dN/dS calculations for multiple alignment comparisons were executed by CODEML from the PAML package (Yang 1997). Constrained model M7 was tested against unconstrained model M8 under the following parameters: f61 (codon frequencies of 61 non-stop codons are calculated), starting omega: 0.4 and 1.5. All simulations converged and results are consistent between both codon models (2ln□; p=values were calculated assuming two degrees of freedom). We present the percentage of sites estimated to evolve under positive selection and the average dN/dS for those sites. Posterior probabilities were calculated according to the Naive Empirical Bayes model (Yang 1997). Codons under positive selection with a posterior probability of >95% have additionally been listed (Table 3.2). The in-depth PAML analysis of *TRIM52* was complemented by Fast Unbiased Bayesian AppRoximation (FUBAR), implemented through Datamonkey suite of phylogenetic analysis tools (Delport, Poon et al. 2010). Sites of positive selection identified via FUBAR are denoted with a "*" (Table 3.2).

3.4.6 TRIM52 restriction assays

We generated CRFK cell lines that stably express HA-tagged human and rhesus *TRIM52* by transduction of a retrovirus vector (LPCX) encoding human and rhesus *TRIM52* as described (Sawyer, Wu et al. 2005). Stable cell lines, including a negative control empty vector CRFK cell line, were plated on 12-well plates (0.8x10^5 cells/well). These were allowed to incubate overnight and then infected with the following GFP encoding retroviruses: HIV-1, HIV-2 (ROD9), and FIV. We used a virus titer determined to give us at least 15% infection. Three days after infection, cells were fixed with paraformaldehyde and GFP expression was measured by flow cytometry.

Chapter 4

A catalogue of *CyclophilinA* retrogenes in primates

Elucidation of the biological role of *CyclophilinA* (*CypA*) has been complicated due to its involvement in numerous biological processes, such as protein folding, cell cycle regulation, and apoptosis. Furthermore, *CypA* has a role in affecting the lifecycle of several viruses, including HIV-1 and HCV. Many retrogene copies of *CypA* have been reported within primate genomes, and these have occasionally led to the birth of novel *TRIMCyp* antiviral factors. Yet, there has been little characterization of *CypA* retrogenes. We surveyed 6 primate reference genomes from hominoids, Old World monkeys, and New World monkeys to catalogue *CypA* retrogenes and describe their copy number and putative functional state. We find that primate genomes encode well over 100 *CypA* retrogenes. While the majority of these are pseudogenes, we find that on average ~18% of the copies encode a putatively functional open reading frame and demonstrate diverse evolutionary histories that implicate them in important functional roles.

4.1 Introduction

CyclophilinA (*CypA*) is the prototypic representative of the *Cyclophilin* gene family. At least 9 *Cyclophilin* genes have been identified in human. These are all characterized by their predicted peptidyl-prolyl isomerase activity, the catalytic *cis-trans* isomerization of proline residues (Fischer, Bang et al. 1984, Takahashi, Hayano et al. 1989, Wang and Heitman 2005, Schaller, Ocwieja et al. 2011). *CypA* was initially identified as the target for the immunosuppressant drug cyclosporineA (CsA) (Handschumacher, Harding et al. 1984). The CypA-CsA complex inhibits calcineurin protein phosphatase activity and consequently downregulates T-cell activation (Liu, Farmer et al. 1991, Colgan, Asmal et al. 2004, Roehrl, Kang et al. 2004).

CypA is believed to play a role in a number of biological processes including protein folding, cell cycle regulation, and apoptosis (Stamnes 1992; Matouschek 1995; Ou 2001; Min 2005; Uittenbogaard 1998; Nahreini 2001; Decker 2003; Colgan 2004; Grimim 2007; Helekar 1994; Wang 2005). However, none of these roles are firmly established. Indeed, *CypA* interactions demonstrate a promiscuity that extends to interactions with "foreign" targets. *CypA* interacts with exposed proline residues of lentiviral CA protein to facilitate the uncoating process to enhance viral fitness (Franke, Yuan et al. 1994, Thali, Bukovsky et al. 1994, Braaten, Aberham et al. 1996, Gamble, Vajdos et al. 1996, Gitti, Lee et al. 1996, Lin and Emerman 2006). This function of *CypA* is not necessary for the fulfillment of this stage of the lentivirus lifecycle (Thali, Bukovsky et al. 1994, Wiegers, Rutter et al. 1999). *CypA* also improves the fitness of Hepatitis Virus C (HCV) by direct interactions with phosphoprotein nonstructural protein 5A (NS5A) and RNA-dependent RNA polymerase nonstructural protein 5B (NS5B) (Fernandes, Poole et al. 2007, Yang, Robotham et al. 2008, Chatterji, Bobardt et al. 2009, Hanoulle, Badillo et al. 2009, Kaul, Stauffer et al. 2009, Fernandes, Ansari et al. 2010). Additional interactions with diverse viruses continue to emerge (reviewed in (Baugh and Gallay 2012, Zhou, Mei et al. 2012)) and add further complexity to the role and activity of *CypA* in the cell.

CypA is conserved throughout eukaryotes, and demonstrates evidence of evolving under strong purifying selection amongst primates (Ortiz, Bleiber et al. 2006). In addition to the intron-containing *CypA* parental gene, the human genome possesses many *CypA* retrogenes (Haendler and Hofer 1990, Willenbrink, Halaschek et al. 1995, Zhang, Harrison et al. 2003) that result from the reverse transcription of the abundant *CypA* transcript and subsequent genomic integration into a new location by the LINE-1 (L1) retrotransposon machinery (Kaessmann, Vinckenbosch et al. 2009); these retrogenes are also

referred to as processed pseudogenes or RNA-based duplicates. Retrotransposed genes are typically viewed as evolutionary dead ends, as they are not expected to transpose with the necessary regulatory elements. However, the lack of transposition with regulatory elements does not pose a significant barrier to all retrogenes; it is suspected that 20% of retrogenes in the human genome are transcriptionally active (Brosius 1991; Marques 2005; Vickenbosch 2006). Indeed, retrogenes provide a unique source for novel, unconstrained genetic material (Brosius 1991, Bieniasz 2003). Relevant to this chapter, several *CypA* retrogenes exhibit transcriptionally activity (Harrison, Zheng et al. 2005). Furthermore, functionality has been demonstrated as at least 3 *CypA* retrogenes form gene fusions with the restriction factor *TRIM5*, generating a novel restriction factor termed *TRIMCyp* (Sayah, Sokolskaja et al. 2004, Brennan, Kozyrev et al. 2008, Newman, Hall et al. 2008, Virgen, Kratovac et al. 2008, Wilson, Webb et al. 2008, Malfavon-Borja, Wu et al. 2013) (also see Chapter 2). In these instances, *CypA* has structurally and functionally replaced the CA targeting motif of *TRIM5*. Such examples demonstrate the functional and diverse utility of *CypA* retrogenes in the genome.

We wished to further elucidate the evolutionary contribution of *CypA* retrogenes to the human and other primate genomes. We therefore surveyed several assembled primate genomes (where the chromosomal positions are available) for *CypA* retrogene copy number and putative functional state (whether it encoded an intact open reading frame). This was carried out by performing BLAT searches (Kent 2002) on the UCSC Genome browser (Karolchik, Hinrichs et al. 2012). We found that primate genomes contain more than 100 *CypA* retrogene copies with an average of ~18% of copies encoding a putatively functional open reading frame (ORF), which we term "*intact*". From species comparisons, we identified 24 sets of orthologous *CypA* retrogenes, where at least 2 primate species in the set were *intact*. Amongst these we found evidence of evolution under diverse pressures, including positive and purifying selection. Rhesus macaque and marmoset revealed many examples of recently acquired (young) *CypA* retrogenes, whiles apes collectively contained fewer young lineage-specific *CypA* retrogenes. We found a single case of an ancient *CypA* retrogene that pseudogenized in the ape common ancestor, but has reverted to an *intact* ORF in the human lineage, regaining protein-coding capacity. Our findings reveal an abundance of *CypA* retrogenes within primate genomes. While many of these are presumed to be pseudogenes (e.g., not functional or transcriptionally inactive), several have been highlighted by their unique evolutionary and molecular history.

4.2 Results

4.4.1 Identifying CypA Retrogenes

CypA retrogenes were identified within the human genome since initial molecular investigations of the parental gene (Haendler and Hofer 1990, Willenbrink, Halaschek et al. 1995). Despite the recognition of copious numbers contained in the human genome (Zhang, Harrison et al. 2003), the functional state and the evolutionary context of *CypA* retrogenes in humans has yet to be explored. Towards this, we screened 6 primate reference genomes (4 hominoid, 1 Old World monkey, and 1 New World monkey) to catalogue *CypA* retrogenes using the UCSC genome browser (Kent, Sugnet et al. 2002). We found 109 *CypA* retrogenes to be encoded within the human reference genome (Figure 4.1). From these, we identified 24 that encoded putatively functional open reading frames (ORFs), termed "*intact*" (see methods). Chimpanzee, gorilla, and orangutan were found to encode 19 *intact* retrogenes of 114, 17 of 104, and 21 of 118, respectively. Rhesus macaque and marmoset exhibited similar numbers of *CypA* retrogenes, 24 *intact* retrogenes of 143, and 28 of 128, respectively. We find that primate genomes universally contain over 100 *CypA* retrogene copies with an average of ~18% maintaining a putatively functional, *intact* ORF.

To determine *intact CypA* retrogene orthology and estimate the date of acquisition we relied on synteny. While it is possible that translocations could relocate *CypA* retrogenes to alternative (non-syntenic) locations, we conservatively identified cases of unidentified syntenic orthologs, when its presence was expected, as pseudogenes. We found 24 sets of orthologs that had *intact* copies in at least 2 primate lineages and termed these the "*intact* ortholog" set (Figure 4.1; Table C.1). To improve the resolution of our dating approach we screened an additional Old World monkey (baboon) and New World monkey (squirrel monkey) reference assembly. Baboon and squirrel monkey genomes were not screened further as the quality of their assemblies was not to the same extent as the 6 initially screened primates at the time of my analysis. 15 ortholog sets were encoded in all major primate lineages and acquired at least 43 Mya. 7 ortholog sets were only found in hominoids and Old World monkeys, placing their acquisition date at 32 Mya. The youngest case of an "*intact* ortholog" was found in only a subset of the great apes and estimated to be 8 Myo.

An additional 10 *intact* retrogenes had pseudogenized orthologs amongst the original primate genomes assayed. We termed this set of retrogenes the "Single *intact* ortholog" (Figure 4.1; Table C.2). It is possible that membership in the "*intact* ortholog" versus "single *intact* ortholog" categories is arbitrary

since it was dependent of which genomes were available for our analysis. Five of the "single *intact* ortholog" cases were acquired 43 mya. 2 of these are present in rhesus macaque, while there is one each in the human, orangutan, and marmoset genomes. 2 cases of 32Myo "Single *intact* orthologs" were found in rhesus macaque and orangutan genomes respectively. The orangutan genome also encode 2 "Single *intact* orthologs" acquired 17 Mya whereas human genomes encode a "Single *intact* ortholog" estimated to be 20 Myo. The 20 Myo human (Hsa1_2264934) and 32 Myo orangutan (Pab9_55196143) "Single *intact* orthologs" are unique amongst the other cases as the mutations leading to pseudogenization in the majority of the orthologs have been reverted or compensated, respectively (Figure 4.2A; Figure C.1A). In the other 8 "Single *intact* ortholog" sets, we are not able to distinguish whether the *intact* state of the *CypA* retrogene is a function of preservation prior to divergence or reversion/compensation after divergence. Orthologs of Hsa1_2264934 do not encode a canonical start codon, as this suffered a missense mutation shortly after its acquisition 20 Mya. Only in the human lineage has the canonical start codon been "restored". Similarly, Pab9_55196143 and its orthologs have lost the canonical start due to a missense mutation (Figure 4.2B; Figure C.1B), but a 1nt deletion in the orangutan lineage places an upstream ATG in-frame with the *CypA* coding sequence that may putatively function as a novel start codon for the Pab9_55196143 ORF.

Based on synteny, 46 *intact* retrogenes did not have identifiable orthologs and represent recently acquired or young acquisitions, which we termed "Lineage specific *intact* retrogenes" (Figure 4.1; Table C.3). The majority of these were found outside of the ape lineage, with 16 encoded by rhesus macaque and 20 encoded by marmoset. Amongst apes, human contain 7 "Lineage specific *intact* retrogenes", while orangutan contain 2 and gorilla contains a single "Lineage specific *intact* retrogenes". From our analysis, it appears that apes possess fewer "Lineage specific" *CypA* retrogenes compared to other primates.

4.4.2 Identifying selection amongst "intact ortholog" sets

We sought to investigate the evolutionary history of the "intact ortholog" sets to identify the selective pressures acting on these preserved *CypA* retrogenes. In our limited analysis, we identified 5 "*Intact* ortholog" retrogenes that displayed signatures suggestive of positive selection: sets 6, 9, 20, 21, and 24 (Table 4.1; light grey fill). To evaluate these in greater detail, we performed additional sequencing and PAML analysis ('asterisks' in Table 4.1). Based on this expanded survey of each of the 5 ortholog sets, we find that only one of them (Set 6) has evolved under strong positive selection. Additional dN/dS

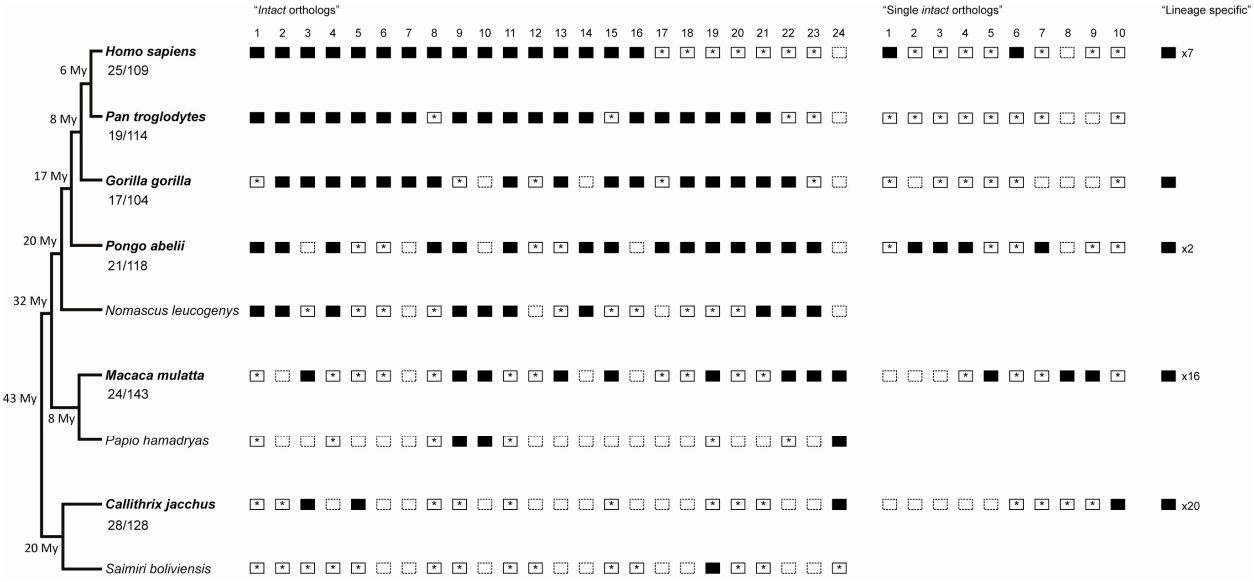


Figure 4.1 "*Intact" CypA* retrogenes. *CypA* retrogenes were recovered from 6 primate reference genomes and evaluated for a putatively functional open reading frame (ORF) that were subsequently termed "*intact*" (black-filled box). Below the name of each primate genome systematically (**BOLD**) evaluated is the number of "*Intact*" over the total number of *CypA* retrogenes identified. "*Intact*" were placed into 1 of 3 categories depending on the presence and state of orthologous *CypA* retrogenes, and orthology was determined by synteny. When orthologs "*intact*" were identified, these were binned into the "*intact* ortholog" category. "*Intact*" that only had pseudogenized orthologs (white-filled box with asterisk) were placed in the "Single *intact* ortholog" category. Orthologs that could not be found were denoted with a broken white-filled box. An ortholog may not be found due to an assembly gap or because the *CypA* retrogene was acquired after the divergence of the lineage not containing the ortholog of interest. For example, orthologs of Set 6 could not be found in New World monkeys, but were recovered in Old World monkeys and hominoids. Thus, this *CypA* retrogene was likely acquired in the common ancestor of Old World monkeys and hominoids, the lineage of primates that diverged from New World monkeys.

Table 4.1 PAML screen of "*intact* ortholog" sets

Intact CypA retrogenes	M7vsM8	dN/dS	P-value	% of sites within dN/dS category
Set 1	0.072	7.282	0.965	1.360
Set 2	0.139	1.179	0.933	100.000
Set 3	0.023	1.000	0.989	100.000
Set 4	1.701	3.180	0.427	41.186
Set 5	5.508	34.236	0.064	2.485
Set 6	**6.460**	**49.375**	**0.040**	**14.478**
Set 6*	**8.367**	**33.983**	**0.015**	**15.098**
Set 7	<0.005	28.169	1.000	0.000
Set 8	<0.005	27.107	1.000	0.000
Set 9	6.06	7.167	0.048	8.188
Set 9*	3.861	8.433	0.145	2.454
Set 10	0.410	1.271	0.815	100.000
Set 11	<0.005	1.000	1.000	0.000
Set 12	<0.005	1.000	1.000	0.000
Set 13	3.251	3.055	0.197	37.087
Set 14	0.366	2.043	0.833	59.987
Set 15	1.481	2.547	0.477	44.190
Set 16	<0.005	2.175	1.000	0.000
Set 17	0.306	4.662	0.858	32.904
Set 18	0.995	1.666	0.608	100.000
Set 19	2.430	3.853	0.297	18.422
Set 20	7.778	23.077	0.020	8.598
Set 20*	5.74913	17.505	0.0564	9.154
Set 21	3183.674	98.019	0.000	2.103
Set 21*	5.795	26.375	0.0552	1.885
Set 22	<0.005	1.000	1.000	0.000
Set 23	0.504	1.421	0.777	90.515
Set 24	9.226	39.485	0.010	2.274
Set 24*	4.229	4.687	0.121	6.501

analysis in each ortholog set revealed a significant fraction of the "*Intact* ortholog" retrogenes have evolved under purifying selection (Table 4.2) suggesting that these genes have been subject to selective constraint since their birth. Therefore, we find that *CypA* retrogenes have been evolving under diverse selective pressures throughout human and primate history.

4.4.3 Conservation of transcriptionally active "Single intact orthologs"

Given the unique observation of a "restored" ORF by "Single *intact* orthologs" Hsa1_2264934 and Pab9_55196143, we investigated these further for transcriptional evidence of activity. Using RT-PCR, we find that Hsa1_2264934 is transcriptionally active in human embryonic kidney (293T) cell lines (Figure 4.2C). We were unable to detect Pab9_55196143 expression in orangutan fibroblast cell lines; however, this does not rule out the possibility of expression in other tissues.

As we detected transcriptional evidence of Hsa1_2264934 by RT-PCR, we further investigated the *CypA* retrogene for evidence of conservation via SNP analysis. We identified 13 reported indels from NCBI's database of short genetic variations (dbSNP) (Table 4.3). The majority of the SNPs reported (12 of 13) result in residue alterations. Of these, rs35460778 reflects a non-synonymous change and a frameshift mutation (1nt deletion). However, there is no population diversity reported and this SNP contains no validation status by NCBI. A single reported SNP results in a synonymous change, thus there is no alternation to the predicted protein sequence. Based on these findings, Hsa1_2264934 appears well preserved in humans and represents a *CypA* retrogene that has "restored" an ORF from a pseudogenized ancestral state.

4.3 Discussion

To better understand the depth of *CypA* retrogenes in the human genome and their evolutionary dynamics, we screened primate genomes to evaluate copy number and putative functional state of *CypA* retrogenes. From an in-depth screen of 6 primate reference genomes, we find that each encode more than 100 *CypA* retrogenes of which an average of ~18% of the retrogenes have a putatively functional ORF. These *intact CypA* retrogenes are comprised of copies that are related by orthology, born in a common ancestor and inherited by extant primates, along with younger, species-specific copies. Amongst human *intact retrogenes*, we find that 18 have ancient origins ranging from 8-43 Mya. The remaining *intact* retrogenes appear to have been acquired within the last 6 million years. The human genome contains more "Lineage specific" *CypA* retrogenes than other hominoids. However, rhesus

macaque and marmoset genomes contain the most "Lineage specific" retrogenes, 16 and 20 respectively. This may be an overestimate in rhesus macaque and marmoset as we lack additional primate genomes to contextualize the Old World monkey and New World monkey "Lineage specific" *intact sets.* Thus, along hominoid evolution, there was a dearth of *intact CypA* gene acquisitions, but a recent burst of activity occurred along the human lineage. This is consistent with evidence of L1 activity found in recent human history (Boissinot, Chevret et al. 2000) and reflects a continued accumulation of *CypA* retrogenes to be utilized and functionalized.

The parental *CypA* gene represents a dramatic example of purifying selection amongst primates (Ortiz, Bleiber et al. 2006, Perelman, Johnson et al. 2011). We found many *CypA* retrogenes have also been subject to strong purifying selection, although these tend to be less constrained than parental *CypA*. Due to this preservation, several orthologs could not be distinguished from each other or the parental gene (e.g., Set 16). The cases of purifying selection may represent cases in which the *CypA* retrogenes support or expand the *CypA* parental gene function, via either dosage or altered expression patterns. These could also be substrates for new gene fusions of the *CypA* domain with other genes just like *TRIMCyp*, although EST evidence did not find any such examples at least within humans.

One of the *CypA* retrogenes (Set 6) demonstrated a robust signature of positive selection. Set 6 comprises a collection of 32 Myo *CypA* retrogenes, but only a subset of the great apes retained *intact* copies. The parallels between this *CypA* retrogenes and the *TRIM52* putative antiviral factor (in Chapter 4) are striking. Both are genes evolving under positive selection and yet both have been subject to idiosyncratic losses suggesting their function is not essential. The Set 6 *CypA* retrogenes may represent a viable candidate for a new antiviral based on its potential property of binding retroviral capsids or other proteins, thus warranting further functional evaluation.

We find that *CypA* retrogenes are more prevalent within the human genome than previously suspected. Intriguingly, the depth of total and putatively functional *CypA* retrogenes is comparable amongst primates. While the function of these retrogene copies was not explored in this study, the evolutionary histories of several "intact ortholog" sets warrant follow up on the basis of their strong evolutionary signatures. Piotukh et al (Piotukh, Gu et al. 2005) explored the targeting preference of human parental *CypA* using a library of linear peptide sequences. A similar approach could be adapted to a yeast-2 hybrid system to explore the targeting ability of several of the sets. Among the "intact ortholog" sets

found to be slowly evolving, we'd expect little deviation from the preferences of parental *CypA*. In the case of rapidly evolving Set 6, identifying sequences homologous to the target preference will likely shed light on the nature of the target (host-encoded or virus-encoded). Also warranting similar functional evaluation is the case of "Single *intact* ortholog" Hsa1_2264934. This *CypA* retrogene is reminiscent of IRGM that pseudogenized in the primate common ancestor, but recently "resurrected" in the lineage of hominoids that gave rise to human, chimpanzee, and gorilla (Bekpen, Marques-Bonet et al. 2009). Despite its derivation from a pseudogenized ORF 6 Mya, we recovered transcriptional activity and population SNP data consistent with a functional Hsa1_2264934. *CypA* has been associated with a wide range of biological pathways, such as protein folding, cell cycle regulation, and apoptosis (Stamnes 1992; Matouschek 1995; Ou 2001; Min 2005; Uittenbogaard 1998; Nahreini 2001; Decker 2003; Colgan 2004; Grimim 2007; Helekar 1994; Wang 2005). Given the number of *intact CypA* retrogenes eligible for replacing the parental gene, it is possible that several of these biological roles are in fact carried out by the products encoded from a *CypA* retrogene. Indeed, there are several cases of "orphan" retrogenes replacing the parental gene throughout mammalian evolution (Ciomborowska, Rosikiewicz et al. 2013). As there is no evidence of the parental *CypA* gene being loss in any of the primate lineages we investigated, we posit that the *CypA* retrogenes have instead evolved to compensate or functionally replace select endogenous roles associated with the *CypA* parental gene.

4.4 Methods

4.4.1 Identifying CypA Retrogenes

The human *CypA* mRNA (NM_021130) sequence was used as a query sequence to submit to BLAST-like alignment tool (BLAT) (Kent 2002). BLAT searches were performed on the human (*Homo sapiens*; Feb. 2009 (GRCh37/hg19)), chimpanzee (*Pan troglodytes*; Feb. 2011 (CSAC 2.1.4/panTro4)), gorilla (*Gorilla gorilla gorilla*; May 2011 (gorGor3.1/gorGor3)), orangutan (*Pongo pygmaeus abelii*; July 2007 (WUGSC 2.0.2/ponAbe2)), rhesus macaque (*Macaca mulatta*; Oct. 2010 (BGI CR_1.0/rheMac3)), and marmoset (*Callithrix jacchus*; March 2009 (WUGSC 3.2/calJac3)) assemblies. Retrogenes were aligned to NM_021130 to identify homology and pseudogenizing mutations (e.g., nonsense mutations; frameshift mutations). In cases where the canonical start was not present, the upstream region was evaluated until a start codon or stop codon was identified in the same frame as the ORF. If an ORF contained a premature stop codon, it was still considered an "*intact"* retrogene if the stop was not within the first

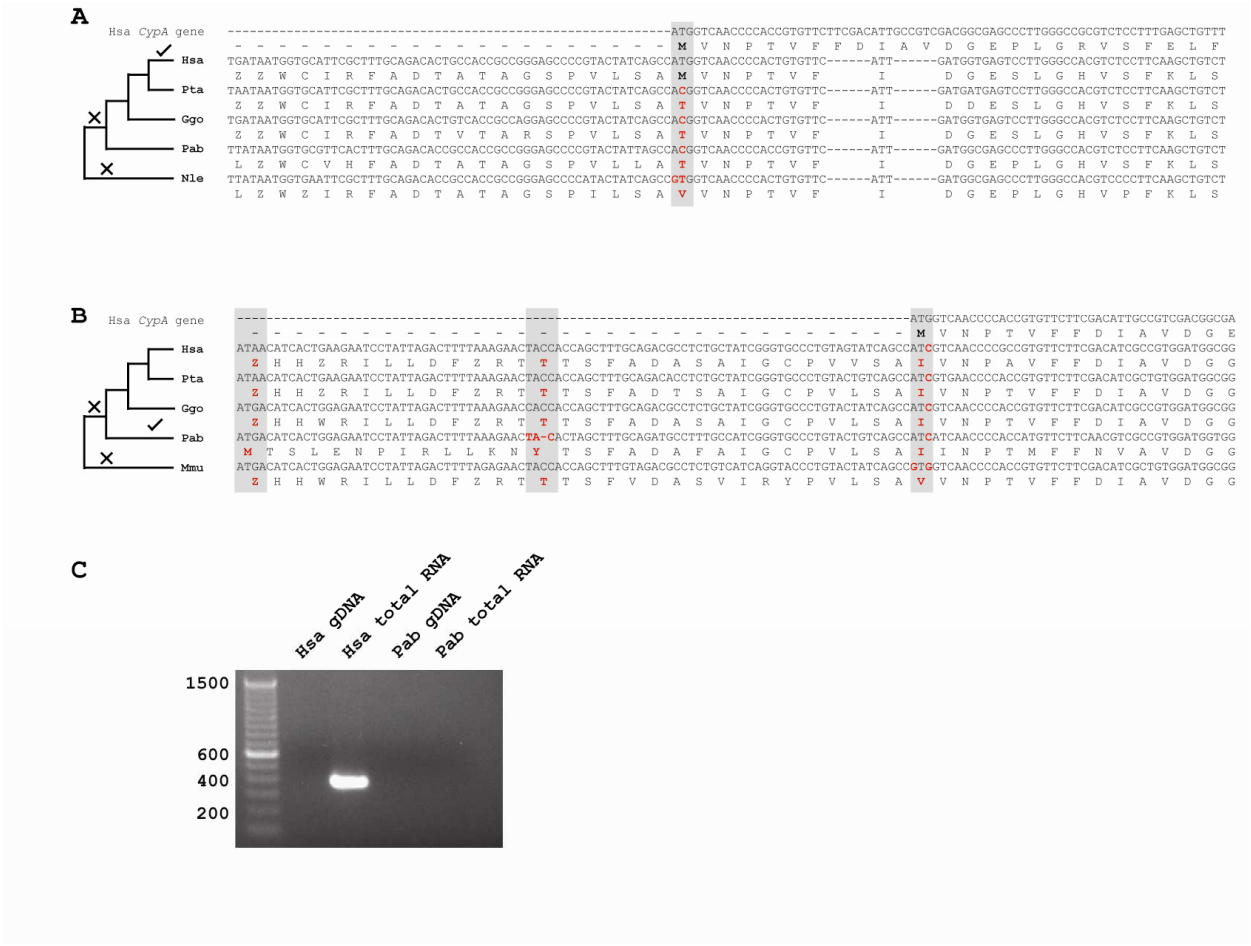


Figure 4.2 Restoration of "Single *intact* ortholog" *CypA* retrogenes. (A-B) Hsa1_2264934 and Pab9_55196143 were highlighted amongst the sets of "Single *intact* orthologs" as they appeared to derive from a pseudogenized common ancestor. Thus, the ORFs of these *CypA* retrogenes have "restored" the pseudogenizing mutation to a state that is putatively functional. (C) The expression of Hsa1_2264934 and Pab9_55196143 was explored by RT-PCR using primers specifically designed for each retrogene. Lane1: Has gDNA; Lane2: Has total RNA; Lane3: Pab gDNA; Lane4: Pab total RNA.

Table 4.2 Whole gene dN/dS analysis

Set1	Hsa:Ptr	Hsa:Pab	Hsa:Pab			
	0.3089	1.206	0.9175			
Set2	Hsa:Ptr	Hsa:Ggo	Hsa:Pab	Ptr:Ggo	Ptr:Pab	Ggo:Pab
	0.0114/0	1.712	1.937	2.552	2.359	1.297
Set3	Hsa:Ggo	Hsa:Ptr	Hsa:Mmu	Hsa:Cja	Ggo:Ptr	Ggo:Mmu
	0.8475	0.4175	0.7135	0.8602	0.3098	0.6547
	Ggo:Cja	Ptr:Mmu	Ptr:Cja	Mmu:Cja		
	0.8048	0.5974	0.6989	0.8744		
Set4	Hsa:Ptr	Hsa:Ggo	Hsa:Pab	Ptr:Ggo	Ptr:Pab	Ggo:Pab
	0.2188	0.01453/0	2.239	1.126	1.246	3.195
Set5	Hsa:Ptr	Hsa:Ggo	Hsa:Cja	Ptr:Ggo	Ptr:Cja	Ggo:Cja
	0.4519	0.3413	0.5803	1.124	0.6773	0.6339
Set6	Hsa:Ptr	Hsa:Ggo	Ptr:Ggo			
	0.00374/0	1.797	2.205			
Set7	Hsa:Ptr	Hsa:Ggo	Ptr:Ggo			
	0.6677	0.2558	0.2777			
Set8	Hsa:Ggo	Hsa:Pab	Ggo:Pab			
	0.2332	0.5832	0.5636			
Set9	Hsa:Ptr	Hsa:Pab	Hsa:Mmu	Ptr:Pab	Ptr:Mmu	Pab:Mmu
	0/0.0203	0.6965	0.7969	0.5986	0.7371	0.8923
Set10	Hsa:Ptr	Hsa:Mmu	Ptr:Mmu			
	0.0049/0	2.008	2.0103			
Set11	Hsa:Ptr	Hsa:Ggo	Hsa:Pab	Ptr:Ggo	Ptr:Pab	Ggo:Pab
	0.7646	1.747	0.6403	1.075	0.6831	0.8109
Set12	Hsa:Ptr					
	0.4789					
Set13	Hsa:Ptr	Hsa:Ggo	Hsa:Mmu	Ptr:Ggo	Ptr:Mmu	Ggo:Mmu
	2.5899	0.6373	0.9521	1.477	1.126	1.231
Set14	Hsa:Ptr	Hsa:Pab	Ptr:Pab			
	0.8195	1.342	1.864			
Set15	Hsa:Ggo	Hsa:Pab	Hsa:Mmu	Ggo:Pab	Ggo:Mmu	Pab:Mmu
	1.865	1.049	0.6546	1.084	0.6319	0.6441
Set16	HsaGgo	Hsa:Ptr	Ggo:Ptr			
	0/0	0.4347	0.4347			
Set17	Ptr:Pab					
	0.9578					
Set18	Ggo:Ptr	Ggo:Pab	Ptr:Pab			
	1.803	1.361	1.422			
Set19	Pab:Ptr	Pab:Ggo	Pab:Mmu	Ptr:Ggo	Ptr:Mmu	Ggo:Mmu

	0.8827	0.5676	2.079	2.249	6.346	3.409
Set20	Ggo:Ptr 0.6995	Ggo:Pab 0.5323	Ptr:Pab 0.8579			
Set21	Ggo:Ptr 6.105	Ggo:Pab 1.527	Ptr:Pab 0.7452			
Set22	Pab:Ggo 0.3121	Pab:Mmu 0.3949	Ggo:Mmu 0.4936			
Set23	Pab:Mmu 0.8501					
Set24	Mmu:Cja 0.4914					

Highlighted "*intact* ortholog" sets exhibited whole gene dN/dS < 1 unanimously in pairwise comparisons.

Table 4.3 Human *CypA* retrogene SNPs

dbSNP ID	Nucleotide Change	Synonymous Change	Non-Synonymous Change
rs189779579	G to A		G146S
rs181156822	G to A		D119N
rs142883803	A to T		T115S
rs11580218	C to T; C to A		A113V; A133E
rs115153819	C to T		T96M
rs186178723	C to A		T89K
rs146993086	A to C		H88P
rs148028730	G to A	T64T	
rs9803657	G to A		G60D
rs9803658	G to A		R59Q
rs150540801	G to A		R33H
rs190597208	C to T		R33C
rs35460778	A to C; - to C		T5P

145 (of 165) codons. The neighboring genes of "*intact"* retrogenes were catalogued and used to determine orthology based on synteny. 24 sets of orthologs, containing" *intact" CypA* retrogenes from at least 2 species, were collected and collectively termed "*intact* ortholog" sets. There were a separate 10 sets of orthologs that only a single "*intact" CypA* retrogene could be found amongst the orthologs and these were termed "Single *intact* ortholog" sets. *CypA* retrogenes that did not have any orthologs were identified as recently acquired (e.g., young, lineage specific) and termed "Lineage specific".

All *intact" CypA* retrogenes were used as queries for a second screen to recover additional *CypA* retrogenes not previously identified with only NM_021130. These queries were also used to identify orthologs in additional primate genomes: white-cheeked gibbon (*Nomascus leucogenys*; Jun. 2011 (GGSC Nleu1.1/nomLeu2)), baboon (*Papio hamadryas*; Nov. 2008 (Baylor 1.0/papHam1)), and squirrel monkey (*Saimiri boliviensis*; Oct. 2011 (Broad/saiBol1)). As these assemblies are not as well developed as those used in the first screen, these only served to complement "*intact* ortholog" sets and "Single *intact* ortholog" sets, but not to identify "Lineage specific" retrogenes.

CypA retrogenes were labeled based on their position (e.g., location). For example, human *CypA* retrogene found at position "chr1:22649345-22649824" was labeled "Hsa1_22649345". Primate abbreviations were as follows- human: Hsa, chimpanzee: Ptr, gorilla: Ggo, orangutan: Pab, white-cheeked gibbon: Nle, rhesus macaque: Mmu, baboon: Pha, marmoset: Cja, squirrel monkey: Sbo.

4.4.2 Evolutionary Analysis

We investigated "*intact* ortholog" sets for recurrent positive selection using the CODEML program from the PAML package (Yang 2007). Codon-based modeling and dN/dS calculations for multiple alignment comparisons were executed by CODEML from the PAML package. Constrained model M7 was tested against unconstrained model M8 under the following parameters: f61 (codon frequencies of 61 non-stop codons are calculated), starting omega: 0.4 and 1.5. All simulations converged and results are consistent between both codon models. For each set we present: M7vsM8 (2lnλ), p=values (calculated assuming two degrees of freedom), the percentage of sites estimated to evolve under positive selection, and the average dN/dS for those sites.

4.4.3 Additional Collection and Analysis

To refine our evolutionary analysis, we collected additional primate sequences by BLAST queries (Altschul, Gish et al. 1990) and PCR amplification. BLAST queries were used to collect orthologs from bonobo (*Pan paniscus*), and Hamadryas baboon (*Papio hamadryas*). PCR was used to obtain Set21 ortholog from Island Siamang gibbon (*Symphalangus syndactylus*; PR00722) siamang gibbon using forward: 5'-TATCAGCCATGGTCAACCCCAC-3'; reverse: 5'-GTTATCCACAGTCAACAATGGTGATC-3'. PCR was used to obtain Set24 orthologs from Woolly monkey (*Lagothrix lagotricha*; 5356), Titi monkey (*Callicebus donacophilus*; OR1522), Talapoin (*Miopithecus talapoin*; PR00716), and Colobus (*Colobus guereza*; PR00980) using forward: 5'-CAGCSATGGTCAACCCCACC-3'; reverse: 5'-TCCTGAGCTGCAGAAGGAATGG-3'.

All PCR reactions were performed using 25-µL reaction volumes and the PCR SuperMix High Fidelity (Invitrogen) reagent. The thermocycler parameters were 94°C for 3min; 39 cycles at 94°C for 15sec, 60°C for 15sec, and 72°C for 2min; and a final extension step at 72°C for 10min. All products were TOPO TA (Invitrogen) cloned and BigDye sequenced using M13 universal primers.

4.4.4 Transcriptional Expression

RT-PCR was used to investigate transcriptional expression of Hsa1_22649345 from "Single *intact* ortholog" sets. Reactions were performed SuperScript III Reverse Transcriptase with Platinum Taq (Invitrogen) in 12.5-µL volume reactions. The RT-PCR parameters were an initial RT step at 50 °C for 30 min; followed by 34 cycles at 94 °C for 15 s, 60 °C for 15 s, and 68 °C for 3 min; and a final extension at 72 °C for 10 min. Primers were forward: 5'- CAACCCCACTGTGTTCATTGATGG-3';
reverse: 5'- CATATTGCCAAAGACCACGTGCTG-3'. gDNA and total RNA was isolated from fibroblast cells purchased from Coriell Cell Repositories. All products were TOPO TA (Invitrogen) cloned and BigDye sequenced using M13 universal primers.

Chapter 5

Concluding remarks and future directions

5.1 Summary

In this dissertation, I described a variety of analyses, delving into primate genomes, to identify putative antiviral genes and extrapolate the history of viral interactions embedded within these genes.

First, I detailed the discovery of an ancient antiviral gene fusion, termed *TRIMCypA3* (Malfavon-Borja, Wu et al. 2013). *TRIMCyp* gene fusions, *TRIMCypA1* and *TRIMCypA2*, were discovered previously in distinct primate lineages and estimated to have evolved quasi-simultaneously 6 Mya. *TRIMCyp* antiviral activity has focused on lentiviruses and has been described as a lentivirus-specific restriction factor. I demonstrated that *TRIMCypA3* evolved in the common ancestor of simian primates 43 Mya and was able to target and restrict lentiviruses. I found that *TRIMCyp* gene fusions uniformly are born with a broad restriction spectrum, and that this narrows and becomes specific over the course of its evolution. Thus, I posit that the antiviral *TRIMCypA3* evolved and swept to fixation in primates due to selective pressure from a lentivirus-like virus 43 Mya.

Next, I presented an analysis of the primate *TRIM* multigene family to describe the evolutionary history of *TRIM* genes and identify signatures of positive selection, interpreted as a hallmark of involvement in a host-virus arms race. This form of selective pressure has resulted in dramatic episodes of adaptation in host antivirals that manifests as positive selection. This approach identified previously undocumented signatures of positive selection in 14 *TRIM* genes, 10 of which represent novel candidate restriction factors that may not be otherwise recognized using more traditional approaches. I focused on the *TRIM52* gene that demonstrates unique genetic innovation: (I) loss of the canonical viral recognition domain (B30.2), (II) expansion of the RING domain, and (III) positive selection, most notably in the expansion of the RING domain. Taken together, this suggests that the genetic innovation identified within *TRIM52* represents the evolution of a novel host-virus interaction interface. This gene family analysis serves to highlight candidate novel restriction factors by positive selection and provides insight into the interfaces of host-pathogen interactions mediated by the *TRIM* multigene family.

Finally, I presented an additional analysis of primate genomes designed to catalogue *CypA* retrogenes and explore their evolutionary history. I previously established the utility of *CypA* retrogenes as a mobile, modular unit in the form of the *TRIMCyp* gene fusion. I took a systematic approach to describe their copy number and putative functional state. We find that primate genomes encode well over 100 *CypA* retrogenes. While the majority of these are pseudogenes, we find that on average ~18% of the copies encode a putatively functional open reading frame and demonstrate diverse evolutionary histories; at least one of these *CypA* retrogenes may have become co-opted in an antiviral role.

5.2 Future directions

5.2.1 TRIMCyp

Unmistakably, *CypA* is the target of a wide range of viruses and serves to enhance infectivity and replication within infected cells (Franke, Yuan et al. 1994, Thali, Bukovsky et al. 1994, Braaten, Aberham et al. 1996, Gamble, Vajdos et al. 1996, Gitti, Lee et al. 1996, Lin and Emerman 2006, Fernandes, Poole et al. 2007, Yang, Robotham et al. 2008, Chatterji, Bobardt et al. 2009, Hanoulle, Badillo et al. 2009, Kaul, Stauffer et al. 2009, Fernandes, Ansari et al. 2010) (reviewed in (Baugh and Gallay 2012, Zhou, Mei et al. 2012)). This is intriguing given the interaction made by lentiviruses has been exploited, where *TRIM5* structurally and functionally replaces the viral recognition (B30.2) domain with *CypA* to re-target and restrict lentiviruses. Remarkably, the range of *TRIMCyp* restriction has been poorly explored. This may be a function of the similarly limited exploration of the *TRIM5* restriction range, only reported to restrict retroviruses. It has been previously demonstrated that *CypA* binding and related *TRIMCyp* restriction is difficult to predict (Goldstone 2010). Thus, preconceptions should be discounted and a diverse range of viruses should be explored for sensitivity to *TRIMCyp* restriction. Based on the current understanding of *CypA* interactions with viral proteins, assessing the restriction potential of *TRIMCyp* against viruses like HCV is particularly interesting. Several other virus-encoded proteins known or thought to interact with *CypA*, including influenza M1 protein and Hepatitis B virus (HBV) small surface proteins (reviewed in (Baugh and Gallay 2012, Zhou, Mei et al. 2012)).

5.2.2 TRIM gene family

The screen to systematically evaluate the *TRIM* gene family for positive selection was of high interest in the on-going struggles of identifying novel restriction factors. The *TRIM* gene family has routinely demonstrated itself to be a family of restriction factors, both direct and indirect (proteins regulating viral fitness via indirect interactions) (reviewed in (Ozato, Shin et al. 2008, Kajaste-Rudnitski, Pultrone et

al. 2010, McNab, Rajsbaum et al. 2011). Therefore, investigating this multigene family was expected to highlight candidates exhibiting signatures of genetic conflict in the form of positive selection. Indeed, we identified positive selection in several members with known involvement in viral, but not previously known to be rapidly evolving fitness (*TRIM15, TRIM21, TRIM25,* and *TRIM38*). Furthermore, we identified 10 other *TRIM* genes (*TRIM2, TRIM7*, *TRIM10, TRIM31*, *TRIM52*, *TRIM58*, *TRIM60*, *TRIM69*, *TRIM75,* and *TRIM76*) with no previously known involvement in viral fitness or detailed evolutionary history. We focused on *TRIM52* as this demonstrated the most diverse genetic innovation amongst the recovered restriction factor candidates. However, the signals of positive selection exhibited by the other candidates necessitate follow up to evaluate their potential range of restriction. In particular, attention should be given to *TRIM58* and *TRIM60* as these exhibited the most diverse signals of positive selection. Along with *TRIM5*, *TRIM58* and *TRIM60* are classified as C-IV. Therefore, any exploration for viral targets should begin with retroviruses. Uchil et al (Uchil, Quinlan et al. 2008) established a framework for systematically evaluating human and mouse *TRIM* genes for early and late restriction against representative gammaretroviruses (N, B, and NB-tropic MLV) and a representative lentivirus (HIV-1). Remarkably, the majority of the *TRIM* genes highlighted in this dissertation were not evaluated by Uchil et al (Uchil, Quinlan et al. 2008). Extending their study to include *TRIM7*, *TRIM10*, *TRIM52*, *TRIM58*, *TRIM60*, *TRIM69*, *TRIM75*, and *TRIM76* orthologs from human and other primates would give an early indication to the validity of some of these *TRIM* genes as bona-fide restriction factors. This should then be complemented to include additional viral target candidates (e.g., additional lentiviruses, spumaretroviruses, etc…) to explore a wider range of restriction.

5.2.3 CypA retrogenes

Several sets of *CypA* retrogenes were highlighted due to their distinct evolutionary signatures. Rapidly evolving Set 6 represents an intriguing collection of retrogenes to investigate for neo-function and to determine the binding impact of the changes to their coding sequence. On the opposite spectrum, are those that we found with evolutionary signatures of purifying selection. This group demonstrated the most diversity in regards to age of acquisition and number of "*intact* orthologs" sets. To evaluate the preservation or divergence of fast and slow evolving *CypA* retrogenes from canonical *CypA* activity, a 2-hybrid system approach has been demonstrated to be an effective system (Luban, Bossolt et al. 1993, Piotukh, Gu et al. 2005). Indeed, this allows a range of targets to be evaluated for interactions with *CypA* retrogenes and can be made high-throughput. Indeed, the interaction between *CypA* and HIV-1 CA was first demonstrated via a 2-hyrbid system (Luban, Bossolt et al. 1993). While the systematic analysis of

primate *CypA* retrogenes has provided a useful catalogue and highlighted several evolutionarily intriguing individuals and ortholog sets, a complementary functional assay would provide valuable information towards the utility of these mobile modules (See Chapter 4.3).

5.3 Paleovirology Insight and Conclusions

The exploratory research described by this dissertation was pursued to better understand the evolutionary history of restriction factors, and to better understand the history of viruses encountered by primate hosts via indirect paleovirology. This methodology relies on identifying traceable signals in host genomes, the birth of novel restriction factors or positive selection driven by recurrent viral pressure. These signals require two elements in order to be truly informative. The first element is a time stamp, placing an age on the signal. The second element is information about the target of restriction. This is the more difficult question to answer – What virus or pathogen does my putative antiviral factor restrict? While there are struggles to identify an informative host-encoded gene and viral antagonist, together these two elements provide immense historical insight into ancient viral infections. Indeed, the discovery of an ancient gene fusion (Chapter 2) combines both elements to implicate the age of lentiviruses to be millennia older than present estimates suggest. In addition, I have provided a collection of restriction factor candidates from the *TRIM* multigene family and a catalogue of *CypA* retrogenes encoded by primate genomes. Discovering candidates solely from these two sources suggests the existence of many other presently unknown restriction factors residing within host genomes. By expanding this evolutionary based approach to other genes and gene families additional restriction factors may indeed by revealed. Alternatively, the investigation of *TRIM* genes and *CypA* retrogenes should be expanded to other animal lineages not already explored. By expanding the library of bona-fide restriction factors, the repertoire of antiviral genes that has defended us from recurrent viral infection is improved.

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Anderson, J. L., E. M. Campbell, X. Wu, N. Vandegraaff, A. Engelman and T. J. Hope (2006). "Proteasome inhibition reveals that a functional preintegration complex intermediate can be generated during restriction by diverse TRIM5 proteins." J Virol **80**(19): 9754-9760.

Barquet, N. and P. Domingo (1997). "Smallpox: the triumph over the most terrible of the ministers of death." Ann Intern Med **127**(8 Pt 1): 635-642.

Barr, S. D., J. R. Smiley and F. D. Bushman (2008). "The interferon response inhibits HIV particle production by induction of TRIM22." PLoS Pathog **4**(2): e1000007.

Battivelli, E., J. Migraine, D. Lecossier, S. Matsuoka, D. Perez-Bercoff, S. Saragosti, F. Clavel and A. J. Hance (2011). "Modulation of TRIM5alpha activity in human cells by alternatively spliced TRIM5 isoforms." J Virol **85**(15): 7828-7835.

Baugh, J. and P. Gallay (2012). "Cyclophilin involvement in the replication of hepatitis C virus and other viruses." Biol Chem **393**(7): 579-587.

Bekpen, C., T. Marques-Bonet, C. Alkan, F. Antonacci, M. B. Leogrande, M. Ventura, J. M. Kidd, P. Siswara, J. C. Howard and E. E. Eichler (2009). "Death and resurrection of the human IRGM gene." PLoS Genet **5**(3): e1000403.

Berthoux, L., S. Sebastian, D. M. Sayah and J. Luban (2005). "Disruption of human TRIM5alpha antiviral activity by nonhuman primate orthologues." J Virol **79**(12): 7883-7888.

Besnier, C., Y. Takeuchi and G. Towers (2002). "Restriction of lentivirus in monkeys." Proc Natl Acad Sci U S A **99**(18): 11920-11925.

Besnier, C., L. Ylinen, B. Strange, A. Lister, Y. Takeuchi, S. P. Goff and G. J. Towers (2003). "Characterization of murine leukemia virus restriction in mammals." J Virol **77**(24): 13403-13406.

Best, S., P. Le Tissier, G. Towers and J. P. Stoye (1996). "Positional cloning of the mouse retrovirus restriction gene Fv1." Nature **382**(6594): 826-829.

Bieniasz, P. D. (2003). "Restriction factors: a defense against retroviral infection." Trends Microbiol **11**(6): 286-291.

Bishop, K. N., M. Bock, G. Towers and J. P. Stoye (2001). "Identification of the regions of Fv1 necessary for murine leukemia virus restriction." J Virol **75**(11): 5182-5188.

Boissinot, S., P. Chevret and A. V. Furano (2000). "L1 (LINE-1) retrotransposon evolution and amplification in recent human history." Mol Biol Evol **17**(6): 915-928.

Boudinot, P., L. M. van der Aa, L. Jouneau, L. Du Pasquier, P. Pontarotti, V. Briolat, A. Benmansour and J. P. Levraud (2011). "Origin and evolution of TRIM proteins: new insights from the complete TRIM repertoire of zebrafish and pufferfish." PLoS One **6**(7): e22022.

Braaten, D., C. Aberham, E. K. Franke, L. Yin, W. Phares and J. Luban (1996). "Cyclosporine A-resistant human immunodeficiency virus type 1 mutants demonstrate that Gag encodes the functional target of cyclophilin A." J Virol **70**(8): 5170-5176.

Brennan, G., Y. Kozyrev and S. L. Hu (2008). "TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis." Proc Natl Acad Sci U S A **105**(9): 3569-3574.

Brennan, G., Y. Kozyrev, T. Kodama and S. L. Hu (2007). "Novel TRIM5 isoforms expressed by Macaca nemestrina." J Virol **81**(22): 12210-12217.

Brosius, J. (1991). "Retroposons--seeds of evolution." Science **251**(4995): 753.

Carthagena, L., A. Bergamaschi, J. M. Luna, A. David, P. D. Uchil, F. Margottin-Goguet, W. Mothes, U. Hazan, C. Transy, G. Pancino and S. Nisole (2009). "Human TRIM gene expression in response to interferons." PLoS One **4**(3): e4894.

Charleston, M. A. and D. L. Robertson (2002). "Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny." Syst Biol **51**(3): 528-535.
Chatterji, U., M. Bobardt, S. Selvarajah, F. Yang, H. Tang, N. Sakamoto, G. Vuagniaux, T. Parkinson and P. Gallay (2009). "The isomerase active site of cyclophilin A is critical for hepatitis C virus replication." J Biol Chem **284**(25): 16998-17005.
Ciomborowska, J., W. Rosikiewicz, D. Szklarczyk, W. Makalowski and I. Makalowska (2013). ""Orphan" retrogenes in the human genome." Mol Biol Evol **30**(2): 384-396.
Clements, J. E. and M. C. Zink (1996). "Molecular biology and pathogenesis of animal lentivirus infections." Clin Microbiol Rev **9**(1): 100-117.
Colgan, J., M. Asmal, M. Neagu, B. Yu, J. Schneidkraut, Y. Lee, E. Sokolskaja, A. Andreotti and J. Luban (2004). "Cyclophilin A regulates TCR signal strength in CD4+ T cells via a proline-directed conformational switch in Itk." Immunity **Aug;21**(2): 189-201.
Cowan, S., T. Hatziioannou, T. Cunningham, M. A. Muesing, H. G. Gottlinger and P. D. Bieniasz (2002). "Cellular inhibitors with Fv1-like activity restrict human and simian immunodeficiency virus tropism." Proc Natl Acad Sci U S A **99**(18): 11914-11919.
Cui, J. and E. C. Holmes (2012). "Endogenous lentiviruses in the ferret genome." J Virol **86**(6): 3383-3385.
Daugherty, M. and H. Malik (2012). Rules of engagement: molecular insights from host-virus arms races. Annual Review of Genetics. **46:** 675-698.
Delport, W., A. F. Poon, S. D. Frost and S. L. Kosakovsky Pond (2010). "Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology." Bioinformatics **26**(19): 2455-2457.
Diaz-Griffero, F., A. Kar, M. Lee, M. Stremlau, E. Poeschla and J. Sodroski (2007). "Comparative requirements for the restriction of retrovirus infection by TRIM5alpha and TRIMCyp." Virology **369**(2): 400-410.
Diaz-Griffero, F., N. Vandegraaff, Y. Li, K. McGee-Estrada, M. Stremlau, S. Welikala, i. Z. S, A. Engelman and J. Sodroski (2006). "Requirements for capsid-binding and an effector function in TRIMCyp-mediated restriction of HIV-1." Virology **Aug 1;351**(2): 404-419.
Dietrich, E. A., G. Brennan, B. Ferguson, R. W. Wiseman, D. O'Connor and S. L. Hu (2011). "Variable prevalence and functional diversity of the antiretroviral restriction factor TRIMCyp in Macaca fascicularis." J Virol **85**(19): 9956-9963.
Dietrich, E. A., L. Jones-Engel and S. L. Hu (2010). "Evolution of the antiretroviral restriction factor TRIMCyp in Old World primates." PLoS One **5**(11): e14019.
Duggal, N. K. and M. Emerman (2012). "Evolutionary conflicts between viruses and restriction factors shape immunity." Nat Rev Immunol **12**(10): 687-695.
Emerman, M. and H. S. Malik (2010). "Paleovirology--modern consequences of ancient viruses." PLoS Biol **8**(2): e1000301.
Fernandes, F., I. U. Ansari and R. Striker (2010). "Cyclosporine inhibits a direct interaction between cyclophilins and hepatitis C NS5A." PLoS One **5**(3): e9815.
Fernandes, F., D. S. Poole, S. Hoover, R. Middleton, A. C. Andrei, J. Gerstner and R. Striker (2007). "Sensitivity of hepatitis C virus to cyclosporine A depends on nonstructural proteins NS5A and NS5B." Hepatology **46**(4): 1026-1033.
Fischer, G., H. Bang and C. Mech (1984). "[Determination of enzymatic catalysis for the cis-trans-isomerization of peptide binding in proline-containing peptides]." Biomed Biochim Acta **43**(10): 1101-1111.
Fletcher, A. J., S. Hué, T. Schaller, D. Pillay and G. J. Towers (2010). "Hare TRIM5α restricts divergent retroviruses and exhibits significant sequence variation from closely related lagomorpha TRIM5 genes." J Virol **84**(23): 12463-12468.
Flicek, P., M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kähäri, D. Keefe, S. Keenan, R.

Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Harrow, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa and S. M. Searle (2012). "Ensembl 2012." Nucleic Acids Res **40**(Database issue): D84-90.
Franke, E. K., H. E. Yuan and J. Luban (1994). "Specific incorporation of cyclophilin A into HIV-1 virions." Nature **372**(6504): 359-362.
Gack, M. U., R. A. Albrecht, T. Urano, K. S. Inn, I. C. Huang, E. Carnero, M. Farzan, S. Inoue, J. U. Jung and A. García-Sastre (2009). "Influenza A virus NS1 targets the ubiquitin ligase TRIM25 to evade recognition by the host viral RNA sensor RIG-I." Cell Host Microbe **5**(5): 439-449.
Gack, M. U., Y. C. Shin, C. H. Joo, T. Urano, C. Liang, L. Sun, O. Takeuchi, S. Akira, Z. Chen, S. Inoue and J. U. Jung (2007). "TRIM25 RING-finger E3 ubiquitin ligase is essential for RIG-I-mediated antiviral activity." Nature **446**(7138): 916-920.
Gamble, T. R., F. F. Vajdos, S. Yoo, D. K. Worthylake, M. Houseweart, W. I. Sundquist and C. P. Hill (1996). "Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid." Cell **87**(7): 1285-1294.
Ganser-Pornillos, B. K., V. Chandrasekaran, O. Pornillos, J. G. Sodroski, W. I. Sundquist and M. Yeager (2011). "Hexagonal assembly of a restricting TRIM5alpha protein." Proc Natl Acad Sci U S A **108**(2): 534-539.
Gifford, R., A. Katzourakis, M. Tristem, O. Pybus, M. Winters and R. Shafer (2008). "A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution." Proc Natl Acad Sci U S A **Dec 23;105**(51): 20362-20367.
Gifford, R. J. (2012). "Viral evolution in deep time: lentiviruses and mammals." Trends Genet **28**(2): 89-100.
Gilbert, C., D. G. Maxfield, S. M. Goodman and C. Feschotte (2009). "Parallel germline infiltration of a lentivirus in two Malagasy lemurs." PLoS Genet **5**(3): e1000425.
Gitti, R. K., B. M. Lee, J. Walker, M. F. Summers, S. Yoo and W. I. Sundquist (1996). "Structure of the amino-terminal core domain of the HIV-1 capsid protein." Science **273**(5272): 231-235.
Goff, S. P. (2004). "Retrovirus restriction factors." Mol Cell **16**(6): 849-859.
Goldstone, D. C., Yap, M.W., Robertson, L.E., Haire, L.F., Taylor, W.R., Katzourakis, A., Stoye, J.P., Taylor, I.A. (2010). "Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface." Cell Host Microbe **8**(3): 248-259.
Griffiths, D. J. (2001). "Endogenous retroviruses in the human genome sequence." Genome Biol **2**(6): REVIEWS1017.
Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk and O. Gascuel (2010). "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0." Syst Biol **59**(3): 307-321.
Haendler, B. and E. Hofer (1990). "Characterization of the human cyclophilin gene and of related processed pseudogenes." Eur J Biochem **190**(3): 477-482.
Han, G. Z. and M. Worobey (2012). "An Endogenous Foamy-like Viral Element in the Coelacanth Genome." PLoS Pathog **8**(6): e1002790.
Han, G. Z. and M. Worobey (2012). "Endogenous Lentiviral Elements in the Weasel Family (Mustelidae)." Mol Biol Evol.
Han, K., D. I. Lou and S. L. Sawyer (2011). "Identification of a genomic reservoir for new TRIM genes in primate genomes." PLoS Genet **7**(12): e1002388.
Handschumacher, R. E., M. W. Harding, J. Rice, R. J. Drugge and D. W. Speicher (1984). "Cyclophilin: a specific cytosolic binding protein for cyclosporin A." Science **226**(4674): 544-547.

Hanoulle, X., A. Badillo, J. M. Wieruszeski, D. Verdegem, I. Landrieu, R. Bartenschlager, F. Penin and G. Lippens (2009). "Hepatitis C virus NS5A protein is a substrate for the peptidyl-prolyl cis/trans isomerase activity of cyclophilins A and B." J Biol Chem **284**(20): 13589-13601.
Harrison, P. M., D. Zheng, Z. Zhang, N. Carriero and M. Gerstein (2005). "Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability." Nucleic Acids Research **2005 Apr 28;33**(8): 2374-2383.
Hartley, J. W., W. P. Rowe and R. J. Huebner (1970). "Host-range restrictions of murine leukemia viruses in mouse embryo cell cultures." J Virol **5**(2): 221-225.
Hatziioannou, T., S. Cowan, S. P. Goff, P. D. Bieniasz and G. J. Towers (2003). "Restriction of multiple divergent retroviruses by Lv1 and Ref1." EMBO J **22**(3): 385-394.
Hatziioannou, T., D. Perez-Caballero, A. Yang, S. Cowan and P. D. Bieniasz (2004). "Retrovirus resistance factors Ref1 and Lv1 are species-specific variants of TRIM5alpha." Proc Natl Acad Sci U S A **101**(29): 10774-10779.
Hatziioannou, T., M. Princiotta, M. Piatak, F. Yuan, F. Zhang, J. D. Lifson and P. D. Bieniasz (2006). "Generation of simian-tropic HIV-1 by restriction factor evasion." Science **314**(5796): 95.
Hilditch, L., R. Matadeen, D. C. Goldstone, P. B. Rosenthal, I. A. Taylor and J. P. Stoye (2011). "Ordered assembly of murine leukemia virus capsid protein on lipid nanotubes directs specific binding by the restriction factor, Fv1." Proc Natl Acad Sci U S A **108**(14): 5771-5776.
Himathongkham, S. and P. A. Luciw (1996). "Restriction of HIV-1 (subtype B) replication at the entry step in rhesus macaque cells." Virology **219**(2): 485-488.
Hofmann, W., D. Schubert, J. LaBonte, L. Munson, S. Gibson, J. Scammell, P. Ferrigno and J. Sodroski (1999). "Species-specific, postentry barriers to primate immunodeficiency virus infection." J Virol **73**(12): 10020-10028.
Holmes, E. C. (2011). "The evolution of endogenous viral elements." Cell Host Microbe **10**(4): 368-377.
Horie, M. and K. Tomonaga (2011). "Non-retroviral fossils in vertebrate genomes." Viruses **3**(10): 1836-1848.
Ikeda, K. (2000). TRIM PROTEINS AS RING FINGER E3 UBIQUITIN LIGASES. S. Inoue. Madame Curie Bioscience.
James, L. C., A. H. Keeble, Z. Khan, D. A. Rhodes and J. Trowsdale (2007). "Structural basis for PRYSPRY-mediated tripartite motif (TRIM) protein function." Proc Natl Acad Sci U S A **104**(15): 6200-6205.
Javanbakht, H., F. Diaz-Griffero, W. Yuan, D. F. Yeung, X. Li, B. Song and J. Sodroski (2007). "The ability of multimerized cyclophilin A to restrict retrovirus infection." Virology **367**(1): 19-29.
Javanbakht, H., W. Yuan, D. F. Yeung, B. Song, F. Diaz-Griffero, Y. Li, X. Li, M. Stremlau and J. Sodroski (2006). "Characterization of TRIM5alpha trimerization and its contribution to human immunodeficiency virus capsid binding." Virology **353**(1): 234-246.
Johnson, W. E. and S. L. Sawyer (2009). "Molecular evolution of the antiretroviral TRIM5 gene." Immunogenetics **61**(3): 163-176.
Jolicoeur, P. and E. Rassart (1980). "Effect of Fv-1 gene product on synthesis of linear and supercoiled viral DNA in cells infected with murine leukemia virus." J Virol **33**(1): 183-195.
Kaessmann, H., N. Vinckenbosch and M. Long (2009). "RNA-based gene duplication: mechanistic and evolutionary insights." Nat Rev Genet **10**(1): 19-31.
Kaiser, S. M., H. S. Malik and M. Emerman (2007). "Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein." Science **316**(5832): 1756-1758.
Kajaste-Rudnitski, A., C. Pultrone, F. Marzetta, S. Ghezzi, T. Coradin and E. Vicenzi (2010). "Restriction factors of retroviral replication: the example of Tripartite Motif (TRIM) protein 5 alpha and 22." Amino Acids **39**(1): 1-9.
Kar, A. K., F. Diaz-Griffero, Y. Li, X. Li and J. Sodroski (2008). "Biochemical and biophysical characterization of a chimeric TRIM21-TRIM5alpha protein." J Virol **82**(23): 11669-11681.

Karolchik, D., A. S. Hinrichs and W. J. Kent (2012). "The UCSC Genome Browser." Curr Protoc Bioinformatics **Chapter 1**: Unit1.4.
Katzourakis, A., R. J. Gifford, M. Tristem, M. T. Gilbert and O. G. Pybus (2009). "Macroevolution of complex retroviruses." Science **Sep 18;325**(5947): 1512.
Katzourakis, A., Gifford, R.J. (2010). "Endogenous viral elements in animal genomes." PLoS Genetics **6**(11).
Katzourakis, A., M. Tristem, O. G. Pybus and R. J. Gifford (2007). "Discovery and analysis of the first endogenous lentivirus." Proc Natl Acad Sci U S A **104**(15): 6261-6265.
Kaul, A., S. Stauffer, C. Berger, T. Pertel, J. Schmitt, S. Kallis, M. Zayas, M. Z. Lopez, V. Lohmann, J. Luban and R. Bartenschlager (2009). "Essential role of cyclophilin A for hepatitis C virus replication and virus production and possible link to polyprotein cleavage kinetics." PLoS Pathog **5**(8): e1000546.
Kawai, T. and S. Akira (2011). "Regulation of innate immune signalling pathways by the tripartite motif (TRIM) family proteins." EMBO Mol Med **3**(9): 513-527.
Keckesova, Z., L. M. Ylinen and G. J. Towers (2004). "The human and African green monkey TRIM5alpha genes encode Ref1 and Lv1 retroviral restriction factor activities." Proc Natl Acad Sci U S A **101**(29): 10780-10785.
Keckesova, Z., L. M. Ylinen, G. J. Towers, R. J. Gifford and A. Katzourakis (2009). "Identification of a RELIK orthologue in the European hare (Lepus europaeus) reveals a minimum age of 12 million years for the lagomorph lentiviruses." Virology **384**(1): 7-11.
Keele, B. F., J. H. Jones, K. A. Terio, J. D. Estes, R. S. Rudicell, M. L. Wilson, Y. Li, G. H. Learn, T. M. Beasley, J. Schumacher-Stankey, E. Wroblewski, A. Mosser, J. Raphael, S. Kamenya, E. V. Lonsdorf, D. A. Travis, T. Mlengeya, M. J. Kinsel, J. G. Else, G. Silvestri, J. Goodall, P. M. Sharp, G. M. Shaw, A. E. Pusey and B. H. Hahn (2009). "Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz." Nature **460**(7254): 515-519.
Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Research **2002 Apr;12**(4): 656-664.
Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler (2002). "The human genome browser at UCSC." Genome Research **2002 Jun;12**(6): 996-1006.
Kirmaier, A., F. Wu, R. M. Newman, L. R. Hall, J. S. Morgan, S. O'Connor, P. A. Marx, M. Meythaler, S. Goldstein, A. Buckler-White, A. Kaur, V. M. Hirsch and W. E. Johnson (2010). "TRIM5 suppresses cross-species transmission of a primate immunodeficiency virus and selects for emergence of resistant variants in the new species." PLoS Biol **8**(8).
Koyanagi, M., J. A. Kerns, L. Chung, Y. Zhang, S. Brown, T. Moldoveanu, H. S. Malik and M. Bix (2010). "Diversifying selection and functional analysis of interleukin-4 suggests antagonism-driven evolution at receptor-binding interfaces." BMC Evol Biol **10**: 223.
Kozak, C. A. and A. Chakraborti (1996). "Single amino acid changes in the murine leukemia virus capsid protein gene define the target of Fv1 resistance." Virology **225**(2): 300-305.
Kratovac, Z., C. A. Virgen, F. Bibollet-Ruche, B. H. Hahn, P. D. Bieniasz and T. Hatziioannou (2008). "Primate lentivirus capsid sensitivity to TRIM5 proteins." J Virol **82**(13): 6772-6777.
Langelier, C. R., V. Sandrin, D. M. Eckert, D. E. Christensen, V. Chandrasekaran, S. L. Alam, C. Aiken, J. C. Olsen, A. K. Kar, J. G. Sodroski and W. I. Sundquist (2008). "Biochemical characterization of a recombinant TRIM5alpha protein that restricts human immunodeficiency virus type 1 replication." J Virol **82**(23): 11682-11694.
Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.
Leroux, C., J. L. Cadoré and R. C. Montelaro (2004). "Equine Infectious Anemia Virus (EIAV): what has HIV's country cousin got to tell us?" Vet Res **35**(4): 485-512.

Li, X., Y. Li, M. Stremlau, W. Yuan, B. Song, M. Perron and J. Sodroski (2006). "Functional replacement of the RING, B-box 2, and coiled-coil domains of tripartite motif 5alpha (TRIM5alpha) by heterologous TRIM domains." J Virol **80**(13): 6198-6206.
Li, X. and J. Sodroski (2008). "The TRIM5alpha B-box 2 domain promotes cooperative binding to the retroviral capsid by mediating higher-order self-association." J Virol **82**(23): 11495-11502.
Li, X., D. F. Yeung, A. M. Fiegen and J. Sodroski (2011). "Determinants of the higher order association of the restriction factor TRIM5alpha and other tripartite motif (TRIM) proteins." J Biol Chem **286**(32): 27959-27970.
Li, Y., X. Li, M. Stremlau, M. Lee and J. Sodroski (2006). "Removal of arginine 332 allows human TRIM5alpha to bind human immunodeficiency virus capsids and to restrict infection." J Virol **80**(14): 6738-6744.
Lilly, F. (1967). "Susceptibility to two strains of Friend leukemia virus in mice." Science **155**(3761): 461-462.
Lim, E. S., H. S. Malik and M. Emerman (2010). "Ancient adaptive evolution of tetherin shaped the functions of Vpu and Nef in human immunodeficiency virus and primate lentiviruses." J Virol **84**(14): 7124-7134.
Lin, T. Y. and M. Emerman (2006). "Cyclophilin A interacts with diverse lentiviral capsids." Retrovirology **2006 Oct 12**(3): 70.
Liu, J., J. D. Farmer, W. S. Lane, J. Friedman, I. Weissman and S. L. Schreiber (1991). "Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes." Cell **66**(4): 807-815.
Liégeois, F., B. Lafay, P. Formenty, S. Locatelli, V. Courgnaud, E. Delaporte and M. Peeters (2009). "Full-length genome characterization of a novel simian immunodeficiency virus lineage (SIVolc) from olive Colobus (Procolobus verus) and new SIVwrcPbb strains from Western Red Colobus (Piliocolobus badius badius) from the Tai Forest in Ivory Coast." J Virol **83**(1): 428-439.
Luban, J., K. L. Bossolt, E. K. Franke, G. V. Kalpana and S. P. Goff (1993). "Human immunodeficiency virus type 1 Gag protein binds to cyclophilins A and B." Cell **73**(6): 1067-1078.
Maillard, P. V., G. Ecco, M. Ortiz and D. Trono (2010). "The specificity of TRIM5 alpha-mediated restriction is influenced by its coiled-coil domain." J Virol **84**(11): 5790-5801.
Malfavon-Borja, R., L. I. Wu, M. Emerman and H. S. Malik (2013). "Birth, decay, and reconstruction of an ancient TRIMCyp gene fusion in primate genomes." Proc Natl Acad Sci U S A.
Mallery, D. L., W. A. McEwan, S. R. Bidgood, G. J. Towers, C. M. Johnson and L. C. James (2010). "Antibodies mediate intracellular immunity through tripartite motif-containing 21 (TRIM21)." Proc Natl Acad Sci U S A **107**(46): 19985-19990.
McEwan, W. A., T. Schaller, L. M. Ylinen, M. J. Hosie, G. J. Towers and B. J. Willett (2009). "Truncation of TRIM5 in the Feliformia explains the absence of retroviral restriction in cells of the domestic cat." J Virol **83**(16): 8270-8275.
McNab, F. W., R. Rajsbaum, J. P. Stoye and A. O'Garra (2011). "Tripartite-motif proteins and innate immune regulation." Curr Opin Immunol **23**(1): 46-56.
Meredith, R. W., J. E. Janečka, J. Gatesy, O. A. Ryder, C. A. Fisher, E. C. Teeling, A. Goodbla, E. Eizirik, T. L. Simão, T. Stadler, D. L. Rabosky, R. L. Honeycutt, J. J. Flynn, C. M. Ingram, C. Steiner, T. L. Williams, T. J. Robinson, A. Burk-Herrick, M. Westerman, N. A. Ayoub, M. S. Springer and W. J. Murphy (2011). "Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification." Science **334**(6055): 521-524.
Meroni, G. and G. Diez-Roux (2005). "TRIM/RBCC, a novel class of 'single protein RING finger' E3 ubiquitin ligases." Bioessays **27**(11): 1147-1157.
Meyerson, N. R. and S. L. Sawyer (2011). "Two-stepping through time: mammals and viruses." Trends Microbiol **19**(6): 286-294.

Neagu, M. R., P. Ziegler, T. Pertel, C. Strambio-De-Castillia, C. Grutter, G. Martinetti, L. Mazzucchelli, M. Grutter, M. G. Manz and J. Luban (2009). "Potent inhibition of HIV-1 by TRIM5-cyclophilin fusion proteins engineered from human components." J Clin Invest **119**(10): 3035-3047.
Newman, R. M., L. Hall, M. Connole, G. L. Chen, S. Sato, E. Yuste, W. Diehl, E. Hunter, A. Kaur, G. M. Miller and W. E. Johnson (2006). "Balancing selection and the evolution of functional polymorphism in Old World monkey TRIM5alpha." Proc Natl Acad Sci U S A **103**(50): 19134-19139.
Newman, R. M., L. Hall, A. Kirmaier, L. A. Pozzi, E. Pery, M. Farzan, S. P. O'Neil and W. Johnson (2008). "Evolution of a TRIM5-CypA splice isoform in old world monkeys." PLoS Pathog **4**(2): e1000003.
Nielsen, R. and Z. Yang (1998). "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene." Genetics **148**(3): 929-936.
Nisole, S., C. Lynch, J. P. Stoye and M. W. Yap (2004). "A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1." Proc Natl Acad Sci U S A **101**(36): 13324-13328.
Nisole, S., J. P. Stoye and A. Saib (2005). "TRIM family proteins: retroviral restriction and antiviral defence." Nat Rev Microbiol **3**(10): 799-808.
O'Brien, S. J., J. L. Troyer, M. A. Brown, W. E. Johnson, A. Antunes, M. E. Roelke and J. Pecon-Slattery (2012). "Emerging viruses in the Felidae: shifting paradigms." Viruses **4**(2): 236-257.
Ohkura, S., M. W. Yap, T. Sheldon and J. P. Stoye (2006). "All three variable regions of the TRIM5alpha B30.2 domain can contribute to the specificity of retrovirus restriction." J Virol **80**(17): 8554-8565.
Ortiz, M., G. Bleiber, R. Martinez, H. Kaessmann and A. Telenti (2006). "Patterns of evolution of host proteins involved in retroviral pathogenesis." Retrovirology **3**: 11.
Ozato, K., D. M. Shin, T. H. Chang and H. C. Morse (2008). "TRIM family proteins and their emerging roles in innate immunity." Nat Rev Immunol **8**(11): 849-860.
Passerini, L. D., Z. Keckesova and G. J. Towers (2006). "Retroviral restriction factors Fv1 and TRIM5alpha act independently and can compete for incoming virus before reverse transcription." J Virol **80**(5): 2100-2105.
Patel, M. R., M. Emerman and H. S. Malik (2011). "Paleovirology - Ghosts and gifts of viruses past." Curr Opin Virol **1**(4): 304-309.
Patel, M. R., Y. M. Loo, S. M. Horner, M. Gale and H. S. Malik (2012). "Convergent evolution of escape from hepaciviral antagonism in primates." PLoS Biol **10**(3): e1001282.
Perelman, P., W. E. Johnson, C. Roos, H. N. Seuánez, J. E. Horvath, M. A. Moreira, B. Kessing, J. Pontius, M. Roelke, Y. Rumpler, M. P. Schneider, A. Silva, S. J. O'Brien and J. Pecon-Slattery (2011). "A molecular phylogeny of living primates." PLoS Genet **7**(3): e1001342.
Perron, M. J., M. Stremlau and J. Sodroski (2006). "Two surface-exposed elements of the B30.2/SPRY domain as potency determinants of N-tropic murine leukemia virus restriction by human TRIM5alpha." J Virol **80**(11): 5631-5636.
Perron, M. J., M. Stremlau, B. Song, W. Ulm, R. C. Mulligan and J. Sodroski (2004). "TRIM5alpha mediates the postentry block to N-tropic murine leukemia viruses in human cells." Proc Natl Acad Sci U S A **101**(32): 11827-11832.
Pertel, T., S. Hausmann, D. Morger, S. Züger, J. Guerra, J. Lascano, C. Reinhard, F. A. Santoni, P. D. Uchil, L. Chatel, A. Bisiaux, M. L. Albert, C. Strambio-De-Castillia, W. Mothes, M. Pizzato, M. G. Grütter and J. Luban (2011). "TRIM5 is an innate immune sensor for the retrovirus capsid lattice." Nature **472**(7343): 361-365.
Piotukh, K., W. Gu, M. Kofler, D. Labudde, V. Helms and C. Freund (2005). "Cyclophilin A binds to linear peptide motifs containing a consensus that is present in many human proteins." J Biol Chem **280**(25): 23668-23674.
Plantier, J. C., M. Leoz, J. E. Dickerson, F. De Oliveira, F. Cordonnier, V. Lemée, F. Damond, D. L. Robertson and F. Simon (2009). "A new human immunodeficiency virus derived from gorillas." Nat Med **15**(8): 871-872.

Price, A. J., Marzetta, F., Lammers, M., Ylinen, L.M., Schaller, T., Wilson, S.J., Towers, G.J., James, L.C. (2009). "Active site remodeling switches HIV specificity of antiretroviral TRIMCyp." Nat Struct Mol Biol. **16**(10): 1036-1042.
Qi, C. F., F. Bonhomme, A. Buckler-White, C. Buckler, A. Orth, M. R. Lander, S. K. Chattopadhyay and H. C. Morse (1998). "Molecular phylogeny of Fv1." Mamm Genome **9**(12): 1049-1055.
Rahm, N., Yap, M., Snoeck, J., Zoete, V., Muñoz, M., Radespiel, U., Zimmermann, E., Michielin, O., Stoye, J.,P., Ciuffi, A., Telenti, A. (2011). "Unique Spectrum of Activity of Prosimian TRIM5{alpha} against Exogenous and Endogenous Retroviruses." J Virol **May;85(9)**: 4173-4183.
Reymond, A., G. Meroni, A. Fantozzi, G. Merla, S. Cairo, L. Luzi, D. Riganelli, E. Zanaria, S. Messali, S. Cainarca, A. Guffanti, S. Minucci, P. G. Pelicci and A. Ballabio (2001). "The tripartite motif family identifies cell compartments." Embo Journal **20**(9): 2140-2151.
Ribeiro, I. P., A. N. Menezes, M. A. Moreira, C. R. Bonvicino, H. N. Seuanez and M. A. Soares (2005). "Evolution of cyclophilin A and TRIMCyp retrotransposition in New World primates." J Virol **79**(23): 14998-15003.
Riedel, S. (2005). "Edward Jenner and the history of smallpox and vaccination." Proc (Bayl Univ Med Cent) **18**(1): 21-25.
Roehrl, M. H., S. Kang, J. Aramburu, G. Wagner, A. Rao and P. G. Hogan (2004). "Selective inhibition of calcineurin-NFAT signaling by blocking protein-protein interaction with small organic molecules." Proc Natl Acad Sci U S A **101**(20): 7554-7559.
Rold, C. J. and C. Aiken (2008). "Proteasomal degradation of TRIM5alpha during retrovirus restriction." PLoS Pathog **4**(5): e1000074.
Santiago, M. L., F. Range, B. F. Keele, Y. Li, E. Bailes, F. Bibollet-Ruche, C. Fruteau, R. Noë, M. Peeters, J. F. Brookfield, G. M. Shaw, P. M. Sharp and B. H. Hahn (2005). "Simian immunodeficiency virus infection in free-ranging sooty mangabeys (Cercocebus atys atys) from the Taï Forest, Côte d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2." J Virol **79**(19): 12515-12527.
Sardiello, M., S. Cairo, B. Fontanella, A. Ballabio and G. Meroni (2008). "Genomic analysis of the TRIM family reveals two groups of genes with distinct evolutionary properties." BMC Evol Biol **8**: 225.
Sawyer, S. L., M. Emerman and H. S. Malik (2004). "Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G." PLoS Biol **2**(9): E275.
Sawyer, S. L., M. Emerman and H. S. Malik (2007). "Discordant evolution of the adjacent antiretroviral genes TRIM22 and TRIM5 in mammals." PLoS Pathog **3**(12): e197.
Sawyer, S. L., L. I. Wu, J. M. Akey, M. Emerman and H. S. Malik (2006). "High-frequency persistence of an impaired allele of the retroviral defense gene TRIM5alpha in humans." Curr Biol **16**(1): 95-100.
Sawyer, S. L., L. I. Wu, M. Emerman and H. S. Malik (2005). "Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain." Proc Natl Acad Sci U S A **102**(8): 2832-2837.
Sayah, D. M., E. Sokolskaja, L. Berthoux and J. Luban (2004). "Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1." Nature **430**(6999): 569-573.
Schaller, T., S. Hue and G. J. Towers (2007). "An active TRIM5 protein in rabbits indicates a common antiviral ancestor for mammalian TRIM5 proteins." J Virol **81**(21): 11713-11721.
Schaller, T., K. E. Ocwieja, J. Rasaiyaah, A. J. Price, T. L. Brady, S. L. Roth, S. Hué, A. J. Fletcher, K. Lee, V. N. KewalRamani, M. Noursadeghi, R. G. Jenner, L. C. James, F. D. Bushman and G. J. Towers (2011). "HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency." PLoS Pathog **7**(12): e1002439.
Sebastian, S. and J. Luban (2005). "TRIM5alpha selectively binds a restriction-sensitive retroviral capsid." Retrovirology **2**: 40.
Sharp, P. M., E. Bailes, F. Gao, B. E. Beer, V. M. Hirsch and B. H. Hahn (2000). "Origins and evolution of AIDS viruses: estimating the time-scale." Biochem Soc Trans **28**(2): 275-282.

Sharp, P. M. and B. H. Hahn (2010). "The evolution of HIV-1 and the origin of AIDS." Philos Trans R Soc Lond B Biol Sci **365**(1552): 2487-2494.
Sharp, P. M. and B. H. Hahn (2011). "Origins of HIV and the AIDS Pandemic." Cold Spring Harb Perspect Med **1**(1): a006841.
Sharp, P. M., D. L. Robertson and B. H. Hahn (1995). "Cross-species transmission and recombination of 'AIDS' viruses." Philos Trans R Soc Lond B Biol Sci **349**(1327): 41-47.
Shibata, R., H. Sakai, M. Kawamura, K. Tokunaga and A. Adachi (1995). "Early replication block of human immunodeficiency virus type 1 in monkey cells." J Gen Virol **76 (Pt 11)**: 2723-2730.
Si, Z., N. Vandegraaff, C. O'Huigin, B. Song, W. Yuan, C. Xu, M. Perron, X. Li, W. A. Marasco, A. Engelman, M. Dean and J. Sodroski (2006). "Evolution of a cytoplasmic tripartite motif (TRIM) protein in cows that restricts retroviral infection." Proc Natl Acad Sci U S A **103**(19): 7454-7459.
Sigurdsson, B. (1954). Maedi, a slow progressive pneumonia of sheep: an epizootiological and pathologic study. British Veterinary Journal. **110:** 255-270.
Song, B., B. Gold, C. O'Huigin, H. Javanbakht, X. Li, M. Stremlau, C. Winkler, M. Dean and J. Sodroski (2005). "The B30.2(SPRY) domain of the retroviral restriction factor TRIM5alpha exhibits lineage-specific length and sequence variation in primates." J Virol **79**(10): 6111-6121.
Song, B., H. Javanbakht, M. Perron, D. H. Park, M. Stremlau and J. Sodroski (2005). "Retrovirus restriction by TRIM5alpha variants from Old World and New World primates." J Virol **79**(7): 3930-3937.
Song, B., Javanbakht, H., Perron, M., Park, D.,H., Stremlau, M., Sodroski, J. (2005). "Retrovirus restriction by TRIM5alpha variants from Old World and New World primates." J Virol **Apr;79(7)**: 3930-3937.
Stevens, A., M. Bock, S. Ellis, P. LeTissier, K. N. Bishop, M. W. Yap, W. Taylor and J. P. Stoye (2004). "Retroviral capsid determinants of Fv1 NB and NR tropism." J Virol **78**(18): 9592-9598.
Stoye, J. P., Yap, M.Y. (2008). Chance favors a prepared genome. PNAS. **105:** 3177-3178.
Straub, O. C. (2004). "Maedi-Visna virus infection in sheep. History and present knowledge." Comp Immunol Microbiol Infect Dis **27**(1): 1-5.
Stremlau, M., C. M. Owens, M. J. Perron, M. Kiessling, P. Autissier and J. Sodroski (2004). "The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys." Nature **427**(6977): 848-853.
Stremlau, M., M. Perron, M. Lee, Y. Li, B. Song, H. Javanbakht, F. Diaz-Griffero, D. J. Anderson, W. I. Sundquist and J. Sodroski (2006). "Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor." Proc Natl Acad Sci U S A **103**(14): 5514-5519.
Suzuki, Y. and T. Gojobori (1999). "A method for detecting positive selection at single amino acid sites." Mol Biol Evol **16**(10): 1315-1328.
Takahashi, N., T. Hayano and M. Suzuki (1989). "Peptidyl-prolyl cis-trans isomerase is the cyclosporin A-binding protein cyclophilin." Nature **337**(6206): 473-475.
Tareen, S. U. and M. Emerman (2011). "Human Trim5α has additional activities that are uncoupled from retroviral capsid recognition." Virology **409**(1): 113-120.
Tareen, S. U., S. L. Sawyer, H. S. Malik and M. Emerman (2009). "An expanded clade of rodent Trim5 genes." Virology **385**(2): 473-483.
Thali, M., A. Bukovsky, E. Kondo, B. Rosenwirth, C. Walsh, J. Sodroski and H. Göttlinger (1994). "Functional association of cyclophilin A with HIV-1 virions." Nature **1994 Nov 24;372**(6504): 363-365.
Towers, G., M. Bock, S. Martin, Y. Takeuchi, J. P. Stoye and O. Danos (2000). "A conserved mechanism of retrovirus restriction in mammals." Proc Natl Acad Sci U S A **97**(22): 12295-12299.
Towers, G. J., T. Hatziioannou, S. Cowan, S. P. Goff, J. Luban and P. D. Bieniasz (2003). "Cyclophilin A modulates the sensitivity of HIV-1 to host restriction factors." Nat Med **9**(9): 1138-1143.
Uchil, P. D., A. Hinz, S. Siegel, A. Coenen-Stass, T. Pertel, J. Luban and W. Mothes (2012). "TRIM protein mediated regulation of inflammatory and innate immune signaling and its association with antiretroviral activity." J Virol.

Uchil, P. D., A. Hinz, S. Siegel, A. Coenen-Stass, T. Pertel, J. Luban and W. Mothes (2013). "TRIM protein-mediated regulation of inflammatory and innate immune signaling and its association with antiretroviral activity." J Virol **87**(1): 257-272.
Uchil, P. D., B. D. Quinlan, W. T. Chan, J. M. Luna and W. Mothes (2008). "TRIM E3 ligases interfere with early and late stages of the retroviral life cycle." PLoS Pathog **4**(2): e16.
van der Aa, L. M., J. P. Levraud, M. Yahmi, E. Lauret, V. Briolat, P. Herbomel, A. Benmansour and P. Boudinot (2009). "A large new subset of TRIM genes highly diversified by duplication and positive selection in teleost fish." BMC Biol **7**: 7.
Van Valen, L. (1973). A new evolutionary law**:** 1-30.
Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin and E. Birney (2009). "EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates." Genome Res **19**(2): 327-335.
Virgen, C. A., Z. Kratovac, P. D. Bieniasz and T. Hatziioannou (2008). "Independent genesis of chimeric TRIM5-cyclophilin proteins in two primate species." Proc Natl Acad Sci U S A **105**(9): 3563-3568.
Wang, P. and J. Heitman (2005). "The cyclophilins." Genome Biology **6**(7): 226.
Weiss, R. A. (2006). "The discovery of endogenous retroviruses." Retrovirology **3**: 67.
Wiegers, K., G. Rutter, U. Schubert, M. Grättinger and H. G. Kräusslich (1999). "Cyclophilin A incorporation is not required for human immunodeficiency virus type 1 particle maturation and does not destabilize the mature capsid." Virology **257**(1): 261-274.
Wilkins, C. and M. Gale (2010). "Recognition of viruses by cytoplasmic sensors." Curr Opin Immunol **22**(1): 41-47.
Willenbrink, W., J. Halaschek, S. Schuffenhauer, J. Kunz and A. Steinkasserer (1995). "Cyclophilin A, the major intracellular receptor for the immunosuppressant cyclosporin A, maps to chromosome 7p11.2-p13: four pseudogenes map to chromosomes 3, 10, 14, and 18." Genomics **28**(1): 101-104.
Wilson, S. J., B. L. Webb, L. M. Ylinen, E. Verschoor, J. L. Heeney and G. J. Towers (2008). "Independent evolution of an antiviral TRIMCyp in rhesus macaques." Proc Natl Acad Sci U S A **105**(9): 3557-3562.
Worobey, M., M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J. J. Muyembe, J. M. Kabongo, R. M. Kalengayi, E. Van Marck, M. T. Gilbert and S. M. Wolinsky (2008). "Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960." Nature **455**(7213): 661-664.
Worobey, M., P. Telfer, S. Souquière, M. Hunter, C. A. Coleman, M. J. Metzger, P. Reed, M. Makuwa, G. Hearn, S. Honarvar, P. Roques, C. Apetrei, M. Kazanji and P. A. Marx (2010). "Island biogeography reveals the deep history of SIV." Science **329**(5998): 1487.
Wu, X., J. L. Anderson, E. M. Campbell, A. M. Joseph and T. J. Hope (2006). "Proteasome inhibitors uncouple rhesus TRIM5alpha restriction of HIV-1 reverse transcription and infection." Proc Natl Acad Sci U S A **103**(19): 7465-7470.
www.aids.gov. (2013). "U.S. Department of Health & Human Services." from www.aids.gov.
Xue, Q., Z. Zhou, X. Lei, X. Liu, B. He, J. Wang and T. Hung (2012). "TRIM38 Negatively Regulates TLR3-Mediated IFN-β Signaling by Targeting TRIF for Degradation." PLoS One **7**(10): e46825.
Yamashita, M. and M. Emerman (2004). "Capsid is a dominant determinant of retrovirus infectivity in nondividing cells." J Virol **78**(11): 5670-5678.
Yan, N. and Z. J. Chen (2012). "Intrinsic antiviral immunity." Nat Immunol **13**(3): 214-222.
Yan, Y., A. Buckler-White, K. Wollenberg and C. A. Kozak (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**(9): 3259-3263.
Yang, F., J. M. Robotham, H. B. Nelson, A. Irsigler, R. Kenworthy and H. Tang (2008). "Cyclophilin A is an essential cofactor for hepatitis C virus infection and the principal mediator of cyclosporine resistance in vitro." J Virol **82**(11): 5269-5278.

Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Comput Appl Biosci **13**(5): 555-556.
Yang, Z. (1998). "Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution." Mol Biol Evol **15**(5): 568-573.
Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-1591.
Yang, Z. and J. P. Bielawski (2000). "Statistical methods for detecting molecular adaptation." Trends Ecol Evol **15**(12): 496-503.
Yap, M., W., Lindemann, D., Stanke, N., Reh, J., Westphal, D., Hanenberg, H., Ohkura, S., Stoye, J., P. (2008). "Restriction of Foamy Viruses by Primate Trim5alpha." J Virol. **June; 82(11)**: 5429–5439.
Yap, M. W., G. B. Mortuza, I. A. Taylor and J. P. Stoye (2007). "The design of artificial retroviral restriction factors." Virology **365**(2): 302-314.
Yap, M. W., S. Nisole, C. Lynch and J. P. Stoye (2004). "Trim5alpha protein restricts both HIV-1 and murine leukemia virus." Proc Natl Acad Sci U S A **101**(29): 10786-10791.
Yap, M. W., S. Nisole and J. P. Stoye (2005). "A single amino acid change in the SPRY domain of human Trim5alpha leads to HIV-1 restriction." Curr Biol **15**(1): 73-78.
Ylinen, L. M., Z. Keckesova, B. L. Webb, R. J. Gifford, T. P. Smith and G. J. Towers (2006). "Isolation of an active Lv1 gene from cattle indicates that tripartite motif protein-mediated innate immunity to retroviral infection is widespread among mammals." J Virol **80**(15): 7332-7338.
Ylinen, L. M., Price, A.J., Rasaiyaah, J., Hué, S., Rose, N.J., Marzetta, F., James, L.C., Towers, G.J. (2010). "Conformational adaptation of Asian macaque TRIMCyp directs lineage specific antiviral activity." PLoS Pathogen **6**(8).
Zhang, F., Hatziioannou, T., Perez-Caballero, D., Derse, D., Bieniasz, P.,D. (2006). "Antiretroviral potential of human tripartite motif-5 and related proteins." Virology **Sep 30;353(2)**: 396-409.
Zhang, Z., P. Harrison, Y. Liu and M. Gerstein (2003). "Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome." Genome Research **Dec;13**(12): 2541-2558.
Zhao, W., L. Wang, M. Zhang, P. Wang, C. Yuan, J. Qi, H. Meng and C. Gao (2012). "Tripartite motif-containing protein 38 negatively regulates TLR3/4- and RIG-I-mediated IFN-β production and antiviral response by targeting NAP1." J Immunol **188**(11): 5311-5318.
Zhao, W., L. Wang, M. Zhang, C. Yuan and C. Gao (2012). "E3 ubiquitin ligase tripartite motif 38 negatively regulates TLR-mediated immune responses by proteasomal degradation of TNF receptor-associated factor 6 in macrophages." J Immunol **188**(6): 2567-2574.
Zheng, N., P. Wang, P. D. Jeffrey and N. P. Pavletich (2000). "Structure of a c-Cbl-UbcH7 Complex: RING Domain Function in Ubiquitin-Protein Ligases." Cell **102**: 533-539.
Zhou, D., Q. Mei, J. Li and H. He (2012). "Cyclophilin A and viral infections." Biochem Biophys Res Commun **424**(4): 647-650.

Appendix A

Supplemental Information for Chapter 2

C
CypA3
additional Alu insertion
CypA3 present
Colobus
Silvery Leaf Langur
Francois' Leaf Langur
Celebes Crested Macaque (nigra)
Lion-tailed macaque (silenus)
Southern pigtailed macaque (nemestrina)
Stump-tailed macaque (arctoides)
Crab-eating macaque (fasciculatris)
Rhesus macaque (mulatta)
Talapoin
D
CypA4
CypA4 present
CypA4 absent
Saddle-back Tamarin
Pygmy Marmoset
Colobus
Silvery Leaf Langur
Francois' Leaf Langur
Celebes Crested Macaque (nigra)
Lion-tailed macaque (silenus)
Southern pigtailed macaque (nemestrina)
Stump-tailed macaque (arctoides)
Crab-eating macaque (fasciculatris)
Rhesus macaque (mulatta)
Talapoin
Patas monkey
De Brazza's monkey
TRIM6
TRIM34
TRIMP1
TRIM5
22
~20 kb
CypA2 (~1Kb)
CypA3 (~14Kb)
CypA4 (~99Kb)
locus
Saddle-back Tamarin
Pygmy Marmoset
Colobus
Silvery Leaf Langur
Francois' Leaf Langur
Celebes Crested Macaque (nigra)
Lion-tailed macaque (silenus)
Southern pigtailed macaque (nemestrina)
Stump-tailed macaque (arctoides)
Crab-eating macaque (fasciculatris)
Rhesus macaque (mulatta)
Talapoin
Patas monkey
De Brazza's monkey
Agile Gibbon
Island Siamang Gibbon
Orangutan
Gorilla
Human
Chimp
Bonobo
(no DNA control)

Figure A.1 *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*.

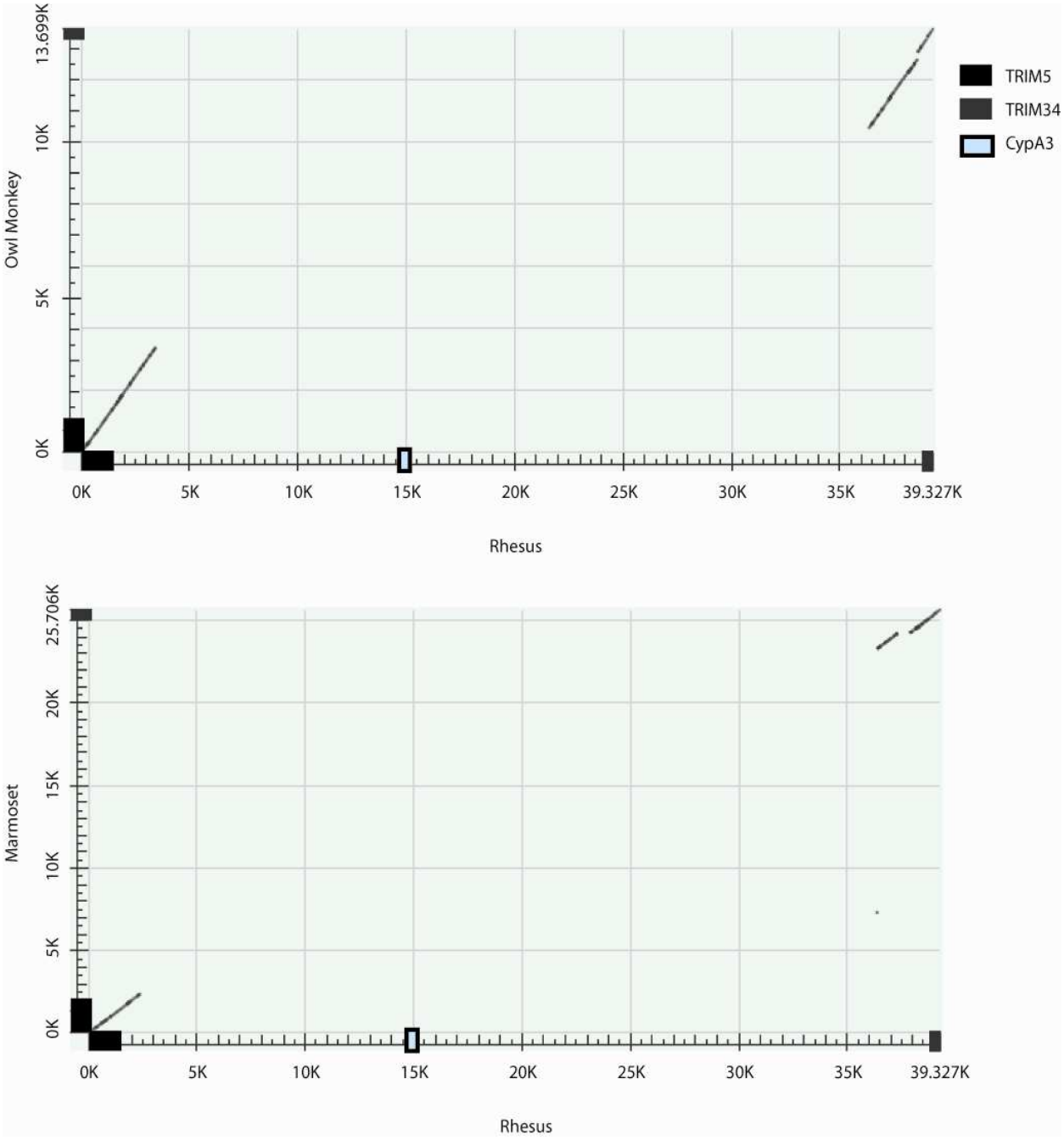

Owl Monkey
13.699K
10K
5K
0K
0K
5K
10K
15K
20K
25K
30K
35K
39.327K
Rhesus
TRIM5
TRIM34
CypA3
Marmoset
25.706K
20K
15K
10K
5K
0K
0K
5K
10K
15K
20K
25K
30K
35K
39.327K
Rhesus

Figure A.2 *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).

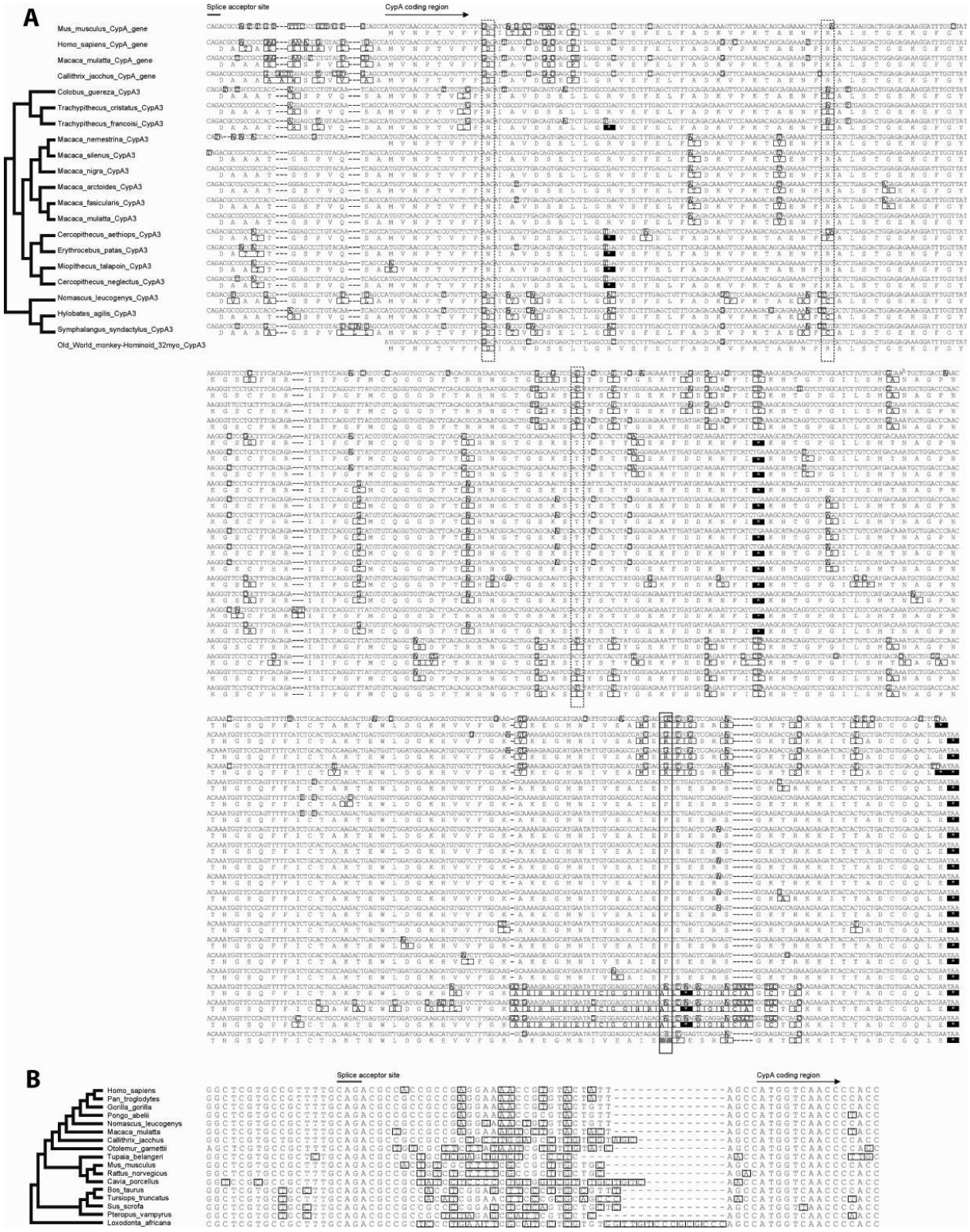
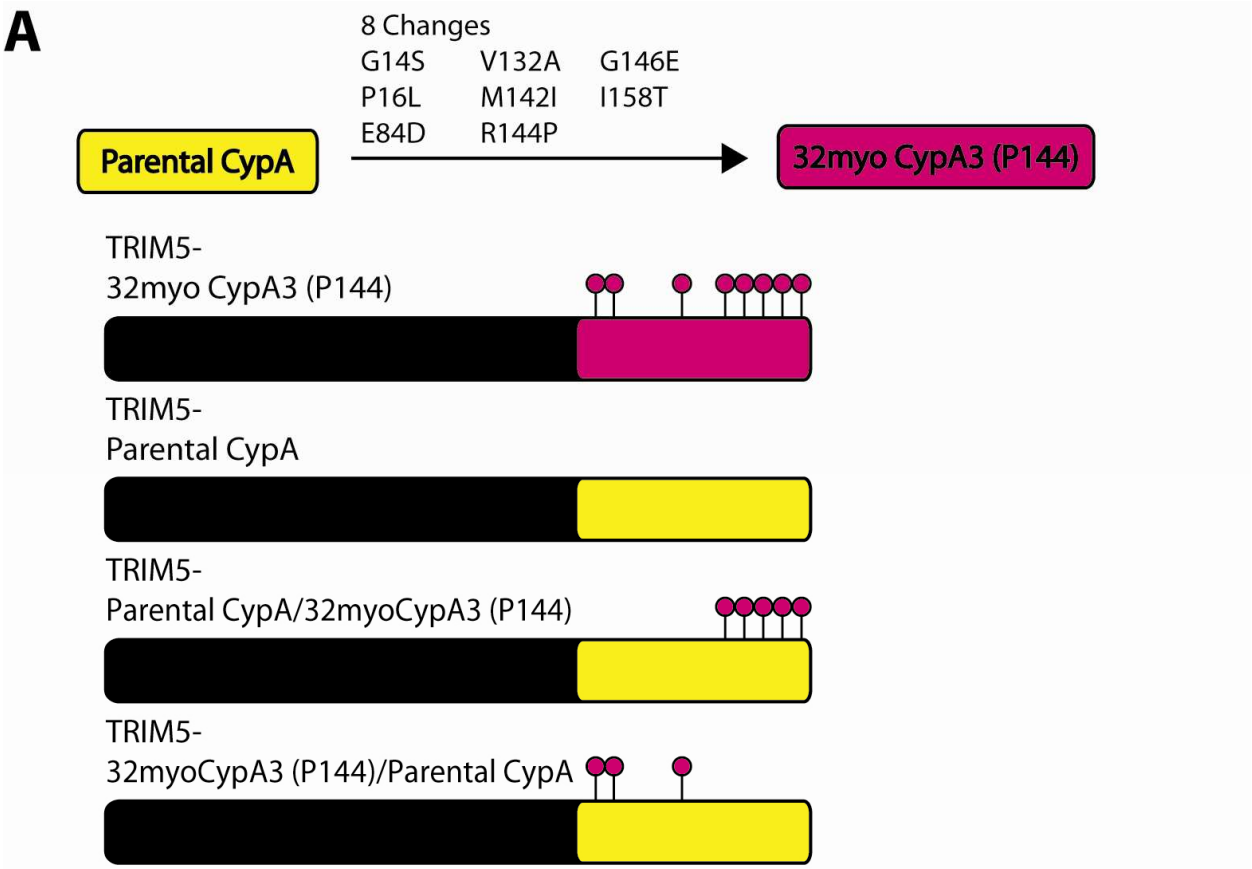

A
Splice acceptor site
CypA coding region
Mus_musculus_CypA_gene
Homo_sapiens_CypA_gene
Macaca_mulatta_CypA_gene
Callithrix_jacchus_CypA_gene
Colobus_guereza_CypA3
Trachypithecus_cristatus_CypA3
Trachypithecus_francoisi_CypA3
Macaca_nemestrina_CypA3
Macaca_silenus_CypA3
Macaca_nigra_CypA3
Macaca_arctoides_CypA3
Macaca_fascicularis_CypA3
Macaca_mulatta_CypA3
Cercopithecus_aethiops_CypA3
Erythrocebus_patas_CypA3
Miopithecus_talapoin_CypA3
Cercopithecus_neglectus_CypA3
Nomascus_leucogenys_CypA3
Hylobates_agilis_CypA3
Symphalangus_syndactylus_CypA3
Old_World_monkey-Hominoid_32myo_CypA3
B
Splice acceptor site
CypA coding region
Homo_sapiens
Pan_troglodytes
Gorilla_gorilla
Pongo_abelii
Nomascus_leucogenys
Macaca_mulatta
Callithrix_jacchus
Otolemur_garnettii
Tupaia_belangeri
Mus_musculus
Rattus_norvegicus
Cavia_porcellus
Bos_taurus
Tursiops_truncatus
Sus_scrofa
Pteropus_vampyrus
Loxodonta_africana

Figure A.3 *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 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 non-synonymous 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 non-primate animals, but we used the upstream region from a spread of non-primate 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 ClustalW2 (Larkin, Blackshields et al. 2007), and the locations of the conserved splice acceptor site and the start of the *CypA* coding sequence have been marked.


A
8 Changes
G14S V132A G146E
P16L M142I I158T
E84D R144P
Parental CypA
32myo CypA3 (P144)
TRIM5-
32myo CypA3 (P144)
TRIM5-
Parental CypA
TRIM5-
Parental CypA/32myoCypA3 (P144)
TRIM5-
32myoCypA3 (P144)/Parental CypA

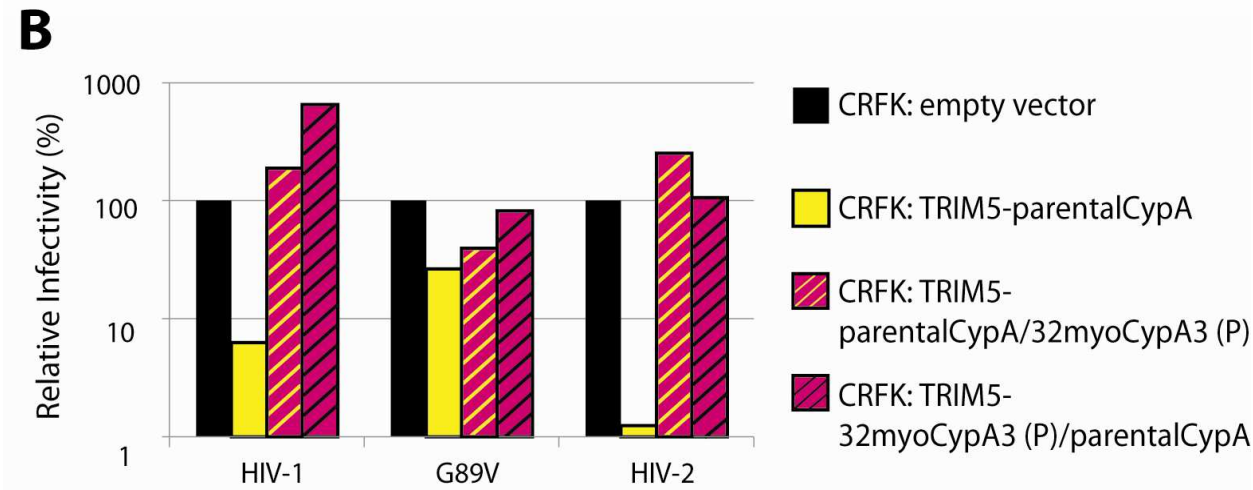

B
Relative Infectivity (%)
1000
100
10
1
HIV-1
G89V
HIV-2
CRFK: empty vector
CRFK: TRIM5-parentalCypA
CRFK: TRIM5-
parentalCypA/32myoCypA3 (P)
CRFK: TRIM5-
32myoCypA3 (P)/parentalCypA

Figure A.4 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%.

Appendix B

Supplemental Information for Chapter 3

cow - *Bos taurus* - cow TRIM52
pig - *Sus scrofa* - pig TRIM52
pda - *Ailuropoda melanoleuca* - panda TRIM52
ele - *Loxodonta africana* - elephant TRIM52
agm - *Cercopithecus aethiops* - African green monkey TRIM52
tal - *Miopithecus talapoin* - talapoin TRIM52
pfl - *Trachypithecus vetulus* - purple faced langur TRIM52
sla - *Trachypithecus cristatus* - silvery langur TRIM52
flm - *Trachypithecus francoisi* - Francois' leaf monkey TRIM52
col - *Colobus guereza* - colobus TRIM52
rhe - *Macaca mulatta* - rhesus TRIM52
chm - *Pan troglodytes* - chimp TRIM52
bon - *Pan paniscus* - bonobo TRIM52
ora - *Pongo pygmaeus* - orangutan TRIM52
mrm - *Callithrix jacchus* - marmoset TRIM52 (exon2)
gor - *Gorilla gorilla* - gorilla TRIM52
hmn - *Homo sapiens* - human TRIM52
bbb - *Otolemur Garnettii* - bushbaby TRIM52
arm - *Dasypus novemcinctus* - armadillo TRIM52
alp - *Vicugna pacos* - alpaca TRIM52
cat - *Felis catus* - cat TRIM52
mle - *Microcebus murinus* - mouse lemur TRIM52
dog - *Canis lupus familiaris* - dog TRIM52
mse - *Mus musculus* - mouse TRIM52 - like
rat - *Rattus norvegicus* - rat TRIM52 – like

```
cow  ATGGCTGGCAGTGCCACTACTCCTAATCCCTTGCAGACACTTCAGGAGGA 50
pig  ATGGCTGGCTGTGCTGCTACTCCCAGTCCTATGCAGACACTTCAGGAGGA 50
pda  ATGGCTGGCTATGCCACTACTCCTAACCCCGTGCAGACCCTTCAGGAGGA 50
ele  ATGGCTAGCTGTGCCACTAC---CAACCCCGCGGAGACACTTCAGGAGGA 47
agm  ATGGCTGGTTATGCCACTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
tal  ATGGCTGGTTATGCCACTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
pfl  ATGGCTGGTTATGCCACTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
sla  ATGGCTGGTTATGCCACTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
flm  ATGGCTGGTTATGCCACTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
col  ATGGCTGGTTATGCCACTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
rhe  ATGGCTGGTTATGCCCCTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
chm  ATGGCTGGTTATGCCACTACTCCCAGCCCCGTGCAGACCCTTCAGGAGGA 50
bon  ATGGCTGGTTATGCCACTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
ora  ATGGCTGGTTATGCCACTACTCCCAGCTCCATGCAGACCCTTCAGGAGGA 50
mrm  --------------------------------------------------
gor  ATGGCTGGTTATGCCACTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
hmn  ATGGCTGGTTATGCCACTACTCCCAGCCCCATGCAGACCCTTCAGGAGGA 50
bbb  ATGGCTAGCTGTGC------------CCGCATGCAGACACTTCAGGAGGA 38
arm  ATGGCAGGCTGTGCCACAACTCCTAGCCCCATGCAGACACTTCAGGAGGA 50
alp  ATGACTGGCTATGCCTCTACTCCCAACCCCACGCAGACACTTCAGCAGGA 50
cat  ATGGCTGGCTGTGCCACTACTCCTAACCCCATGCAGACCCTTCAGGAGGA 50
mle  ATGGCTGGCCACGCCGCCACTCCCAACCCCTTGCAGACGCTGCAGCAGGA 50
dog  ATGGCTGGCTATGCTGCGACTCCTAACCCCATAAAGACCCTTCAGGAGGA 50
mse  ATGGCCACCTCTACACGGCCTCCCAGCCCTATGCAGTCACTTCGGGAAGA 50
rat  ATGGCCGCCCCTACAGGGCCTCCCAGCCCTATGCAGTCACTTCGGGAAGA 50
```

```
cow  AGCCGTGTGTGCCATCTGCCTGGATTACTTCAAGGATCCCGTGTCCATCG 100
pig  AGCGGTGTGTGCCATCTGCCTAGATTACTTCAAGGATCCTGTATCCATCG 100
pda  GGCTGTTTGTGCCATCTGTCTGGATTACTTCAAGGATCCTGTGTCCATAG 100
ele  CGCCGTGTGTGCCATCTGCCTGGATTACTTCAAGGATTCCATAACTATCG 97
agm  AGCGGTGTGTGCCATCTGCTTGGATTACTTCAAGGACCCCGTGTCCATCA 100
tal  AGCGGTGTGTGCCATCTGCTTGGATTACTTCAAGGACCCCGTGTCCATCA 100
pfl  AGCGGTGTGTGCCATCTGCTTGGATTACTTCAAGAACCCCGTGTCCATCA 100
sla  AGCGGTGTGTGCCATCTGCTTGGATTACTTCAAGAACCCCGTGTCCATCA 100
flm  AGCGGTGTGTGCCATCTGCTTGGATTACTTCAAGAACCCCGTGTCCATCA 100
col  AGCGGTGTGTACCATCTGCTTGGATTACTTCAAGGACCCCGTGTCCATCA 100
rhe  AGCGGTGTGTGCCATCTGCTTGGATTACTTCAAGGACCCCGTGTCCATCA 100
chm  AGCGGTGTGTGCCATCTGCTTGGATTACTTCAAGGACCCCGTGTCCATCA 100
bon  AGCGGTGTGTGCCCTCTGCTTGGATTACTTCAAGGACCCCGTGTCCATCA 100
ora  AGCAGTGTGTGCCATCTGCTTGGATTACTTCAAGGACCCCGTGTCCATCA 100
mrm  --------------------------------------------------
gor  AGCGGTGTGTGCCATCTGCTTGGATTACTTCAAGGATCCCGTGTCCATCA 100
hmn  AGCGGTGTGTGCCATCTGCTTGGATTACTTCAAGGACCCCGTGTCCATCA 100
```

```
bbb  AGCCGTGTGCACCATCTACCTGGATTACTTCAAGGATCCGGTGTCCATCG 88
arm  AGCGGTATGCGCTATCTGCCTGGATTACTTCAAGGATCCGGTGTCCATCG 100
alp  AGCGGTGTGCGCCATCTGCCTGGATTACTTGAAGGATCCCTTGTCCATTG 100
cat  AGCGGTGTGTGCCATCTGCCTGGATTACTTCAAGGATCCCGTTTCCATAG 100
mle  GGCGGTGTGCGCCATCTGCCTGGATTACCTCGAGGATCCTGTGTCCATCA 100
dog  GGCGGTGTGCGCCATCTGCCTCGATTACTTCACGGACCCCGTGTCCATCG 100
mse  AGCAGTGTGTGCCATCTGTCTGGATTATTTTAAGGACCCCGTGTCCATTG 100
rat  AGCAGTATGTGCCATCTGTCTGGATTACTTTAAGGACCCCGTGTCCATTG 100

cow  GCTGTGGTCATAACTTTTGCCGTGGGTGTGTGACCCAGCTGTGGGGCAAG 150
pig  GCTGTGGACACAACTTTTGCCGAGGGTGTGTGACTCAGCTGTGGGGCAAG 150
pda  GCTGCGGGCACAACTTCTGCCGAGGGTGTGTGACCCAACTGTGGGGCAAG 150
ele  GCTGTGGGCACAATTTCAGCCGAGGGCGTGTGACCCAGCTGTAGGGCTGG 147
agm  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
tal  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCATCTGTGGGGTAAG 150
pfl  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
sla  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
flm  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
col  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
rhe  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
chm  GCTGCGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
bon  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
ora  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
mrm  --------------------------------------------------
gor  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGTAAG 150
hmn  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGAGTAAG 150
bbb  CCTGCAGGCACAACTTCTGCCACGGGTGTGTGACTCAGCTGTGGGGTAAG 138
arm  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGCAAG 150
alp  GCTGCGGGCACAACTTGTGCAGAGGGTGTGTGACCCAGCTGTGGAGCGAG 150
cat  GCTGTGGGCACAACTTCTGCCGAGGATGTGTGACCCAACTGTGGGGCAAG 150
mle  GCTGCGGGCACAACTTCTGCCGCGGGTGTGCGACCCGGCTGTGGGGTAAG 150
dog  GCTGCGGGCACAACTTCTGCCGAGTGTGTGTAACCCAGCTGTGGGGCAAG 150
mse  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGCAAG 150
rat  GCTGTGGGCACAACTTCTGCCGAGGGTGTGTGACCCAGCTGTGGGGCAAG 150

cow  GAAGATAATGAGCAAGACAGGGAGGAAGAGGAAGATGAATGG---GAGGA 197
pig  GAAGAT---GAGCAAGACAGGGATGAAGAGGAAGATGAATGG-------- 189
pda  GAAGAG---GAGGAAGACAGGGAAGAGGAGGAAGATGAATGG---GAGGA 194
ele  GAGGAT------GAGAACAGGGACGAGAAGGAAGATGAATGG------GA 185
agm  AAGGAC---GAGGAAGACCAGAACGAGGAGGGAGATGAATGG---GAGGA 194
tal  AAGGAC---GAGGAAGACCAGAACGAGGAGGGAGATGAATGG---GAGGA 194
pfl  AAGGAC---AAGGAGGACCAGAACGAAGAGGAAGATGAATGG---GAGGA 194
sla  AAGGAC---AAGGAGGACCAGAACGAAGAGGAAGATGAATGG---GAGGA 194
flm  AAGGAC---AAGGAGGGCCAGAACGAAGAGGAAGATGAATGG---GAGGA 194
col  AAGGAC---GAGGAGGACCAGAACGAAGAGGAAGATGAATGG---GAGGA 194
rhe  AAGGAC---GAGGAGGACCAGAACGAGGAGGAAGATGAATGG---GAGGA 194
chm  GAGGAC------------------GAGGAGGAAGATGAATGGGAGGAGGA 182
bon  GAGGAC------------------GAGGAGGAAGATGAATGGGAGGAGGA 182
ora  GAGGAC---GAGGAGGACCAGAACGAGGAGGAAGATGAATGG------GA 191
mrm  --------------------------------------------------
gor  GAGGAC---GAGGAGGACCAGAACGAGGAGGAAGATGAATGG------GA 191
hmn  GAGGAC---GAGGAGGACCAGAACGAGGAGGAAGATGAATGG---GAGGA 194
bbb  GAAAAT---GAGGAAGATCGG---GAGGAAGATGATGAATGG-------- 174
arm  GAGGAT------GAGGAAAGGGAAGAGGAAATT---GAATGG---GAGGA 188
alp  GAAGATGACGAGGAAGACAGGGACGTAGAGGAAGATGAATGG---GAGAA 197
cat  GAAGAT---GAGGAAGACAGGGACGGGGAAGAAGATGGATGG---GAAGA 194
mle  GATGAG------GAAGACCAGGCCGAGGAAGAGGATGAATCGGAGGAGGG 194
dog  GAAGAC------------------GAGGAGTTGGATGAATGG------GA 176
mse  GAAGAT---------------------------------------------- 156
rat  GAAGAT---------------------------------------------- 156

cow  GGACGAGGACGACGAGGAGGCAGTAGGGGCCATCGGTGGATGGGGCAACT 247
pig  -GAGGAGGACGACGAGGCGGTGG---GGGCCATCGGTGGATGGGACAACT 235
pda  GGACGAGGACGCCGATGTGGAGG---GGGCCATCAGTGGGTGGGACAACT 241
ele  GAAAGAAGAGGACGAGGCCGTGG---GGGCCACTGGTGGACGGGACAACT 232
agm  GGAGGAGGACGGGGAAGCGGTGG---GGGCCGTGGATGGATGGGATGGCT 241
tal  GGAGGAGGACGGGGAAGCGGTGG---GGGCCGTGGATGGATGGGACGGCT 241
pfl  AGAGGAGGACCGGGAAGCGGTGG---GGGCCGTGGATGGATGGGACGGCT 241
sla  AGAGGAGGACGGGGAAGCGGTGG---GGGCCGTGGATGGATGGGACGGCT 241
flm  AGAGGAGGACGGGGAAGCGGTGG---GGGCCGTGGATGGATGGGACGGCT 241
col  GGAGGAGGACGGGGAAGCGGTGG---GGGCCGTGGATGGATGGGACGGCT 241
```

```
rhe  GGAGGAGGACGGCGAAGCGGTGG---GGGCCGTGGATGGATGGGACGGCT 241
chm  GGAGGAGGACGAGGAGGCGGTGG---GGGCCGTGGATGGATGGGACGGCT 229
bon  GGAGGAGGACGAGGAAGCGGTGG---GGGCCATGGATGGATGGGACGGCT 229
ora  GGAGGAGGACGAGGAAGCGGTGG---GGGCCATGGATGGATGGGACGGCT 238
mrm  --------------------------------------------------
gor  GGAGGAGGACGAGGAAGCGGTGG---GGGCCGTGGATGGATGGGACGGCT 238
hmn  GGAGGAGGACGAGGAAGCGGTGG---GGGCCATGGATGGATGGGACGGCT 241
bbb  -AAGGAGGATGAGGAAGCCGTGG---GGGCCACTGGTGGATGGGACAACT 220
arm  AGAAGAGGACGACGAGGTGGTGG---AGGCCATTGGTGGGTGGGACAACT 235
alp  CGAGGAGGACCACGAGGCGGTGG---GGGCTATCGGTGGATGGGACGATT 244
cat  AGACGAGGACGACGATGTGGAGG---AGGCCACCGGTGGATGGGACAACT 241
mle  AGAGGAGGGTGGGGAAGCCGTGG---GGGCCACCGGTGGATGGGACAGCT 241
dog  GAACGAGGACGACGACGTGGAGG---GGGCCATCGGTGGATGGGACAACT 223
mse  -GAGCAGGACCGGGAACAG------------CCTGCTGTGCAGGAGCGCG 193
rat  -GAACAGGACCGGGAACCG------------CCTGCTGTAGGGAACCACG 193

cow  CCATTCGGGAGGTTTTATACCAGGGGAATGCTGACGAGGAGTTGTTCCAG 297
pig  CCATTCGACAGGTTTTATACCAGGGAAATGCTGACGAGGAGTTGTTCCAG 285
pda  CTATTCGAGAGGTTTTGTACCAGGGCAATGCTGATGAGGAGGTGTTCCAG 291
ele  CCATTCGAGAGGTTTTGTACCGGGGCAATGCTGACGAGGGC---TTC--- 276
agm  CCGTTCGAGAGGTGTTGTATCGAGGGAATGCTGACGAAGAGTTGTTCCAA 291
tal  CCGTTCGAGAGGTGTTGTATCGAGGGAATGCTGACGAAGAGTTGTTCCAA 291
pfl  CCATTCGAGAGGTGTTGTATCGAGGGAATGCTGACGAAGAGTTGTTCCAA 291
sla  CCATTCGAGAGGTGTTGTATCGAGGGAATGCTGACGAAGAGTTGTTCCAA 291
flm  CCATTCGAGAGGTGTTGTATCGAGGGAATGCTGACGAAGAGTTGTTCCAA 291
col  CCATTCGAGAGGTGTTGTATCGAGGGAATGCTGACGAAGAGTTGTTCCAA 291
rhe  CCATTCGAGAGGTGTTGTATCGAGGGAATGCTGACGAAGAGTTGTTCCAA 291
chm  CCATTCGAGAGGTGTCGTATCGGGGGAATGCTGACGAAGAGTTGTTCCAA 279
bon  CCATTCGAGAGGTGTTGTATCGGGGGAATGCTGACGAAGAGTTGTTCCAA 279
ora  CCATTCGAGAGGTGTTGTATCCGGGGAATGCTGACGAAGAGTTGTTCCAA 288
mrm  --------------------------------------------------
gor  CCGTTCGAGAGGTGTTGTATCGGGGGAATGCTGACGAAGAGTTGTTCCAA 288
hmn  CCATTCGAGAGGTGTTGTATCGGGGGAATGCTGACGAAGAGTTGTTCCAA 291
bbb  CCATTCTAGAGGTATTGTAT---GGGAATGCCAACGAGGAGCTGTTCCAG 267
arm  CCATCCGAGAGGTTTTGTATCGGGGGAGTGCTGACGAAGAG---TTCCAG 282
alp  CCATTAGAGAGGTTTTATACCAA------GCTGACAAGTTGTTGTTCGAG 288
cat  CTATTCGAGAGGTTTTGTACCAGGGCAGTGCTGACGAGG---TGTTCCAG 288
mle  CCATTCGGGAGGTGTTGTATCGGGGGAACGCTGACGAGGAGCCGTTCCGC 291
dog  CTATTCGAGAGGTTTTGTACCAGCGTAATGGTGATGAGGCAGTGTTCCAG 273
mse  TCATCCGGGAGGTTTTGTTTCCTAGGTACACCGAGCAGGAG------CAG 237
rat  TCATTAGAGAGGTTTTGTTTCATAGGTACACAGAACAGGAGGTGTTTCAG 243

cow  GACCAAGAGGATGATGAACCCTGGGTCGGTGACGGTGGCATAAGG----- 342
pig  GACCAAGAGGATGATGAGCTCTGGGTCGGTGACGCTGGTGTGAGGAATTG 335
pda  GACCAAGAAGATGATGAACTCTGGGTTGGTGACGGTGGCGTCAGGAATTG 341
ele  ---CAAGACGAAGATAAACCCTGGGTTGGTGACGGAGGCATAAGGAATTG 323
agm  GACCAAGAGGACGGTGAACTCTGGCTCGGTGACAGTGGTATAACTAATTG 341
tal  GACCAAGGGGACGGTGAACTCTGGCTCGGTGACAGTGGTATAACTAATTG 341
pfl  GACCAAAAGGACGGTGAACTCTGGCTCGGTGACAGTGGTATAACTAATTG 341
sla  GACCAAAAGGACGGTGAACTCTGGCTCGGTGACAGTGGTATAACTAATTG 341
flm  GACCAAAAGGACGGTGAACTCTGGCTCGGTGACAGTGGTATAACTAATTG 341
col  GACCAAAAGGACGGTGAACTCCGGCTCGGTGACAGTGGTATAACTAATTG 341
rhe  GACCAAGAGGACGGTGAACTCTGGCTCGGTGACAGTGGTATAACTAATTG 341
chm  GACCAAGATGACGATGAACTCTGGCTCGGTGACAGTGGTAGAACTAATTG 329
bon  GACCAAGATGACGATGAACTCTGGCTCGGTGACAGTGGTAGAACTAATTG 329
ora  GACCAAGATGACGATGAACTCTGGCTCGGTAACAGTGGTATAACTAATTG 338
mrm  --------------------------------------------------
gor  GACCAAGATGACGATGAACTCTGGCTCGGTGACAGTAGTATAACTAATTG 338
hmn  GACCAAGATGACGATGAACTCTGGCTCGGTGACAGTGGTATAACTAATTG 341
bbb  GATCCAGAAGATGATGAACTCTGGGTTGGTGACAGTGGTGTAAGTAATTG 317
arm  GACCAAGAGGAT---GAACTCTGGGTCGGTGACGGTGGCGTCAGGAGTTG 329
alp  GACCAAGAGGATGATGAACTCTGGGTCGGTGACGGTGGCATAAGGAATTG 338
cat  GACCAAGAAGAT---GAACTCTGGGTTGGTGACGGTGGCATCAGAAATTG 335
mle  GACCAGGAGGATGACGAATTCTGGGTCGGTTACAGTGGTGCGAGAAATTG 341
dog  GACCAAGAAGATGATGAACTCTGGGTTGGTGATGGTGGGGTCAGGAATCG 323
mse  CACAGAGCACATGCAGGACAGTGGGTCGGTCATAGCCATAGACAGCATCA 287
rat  CACCGAGCACATGCTGGACGCTGGGCTGCTCATAGCCATAGAAGGCATCG 293

cow  -GACAGCATGGATTATGTGTGGGACCAGGAG-----------------G 373
pig  GGACAACATGGATTATGTGTGGGACCAGGAGGAA--------------G 370
pda  GGACAACATGGACTATGCGTGGGACCAGGAGGAAGAA------GAGGAGG 385
```

```
ele  GGACAGTATGGACTATGTGTGGGACCAGGAGG----------------- 355
agm  GGACAACGTAGACCATATGTGGGACCAGGAGGAAGAA------GAAGAGG 385
tal  GGACAACGTAGACCATATGTGGGACCAGGAGGAAGAA------GAAGAGG 385
pfl  GGACAACGTAGACCATATGTGGGACCAGGAGGAAGAA------GAAGAGG 385
sla  GGACAACGTAGACCATATGTGGGACCAGGAGGAAGAA------GAAGAGG 385
flm  GGACAACGTAGACCATATGTGGGACCAGGAGGAAGAA------GAAGAGG 385
col  GGACAACGTAGACCATATGTGGGACCAGGAGGAAGAA------GAAGAGG 385
rhe  GGACAACGTAGACCATATGTGGGACCAGGAGGAAGAA------GAAGAGG 385
chm  GGACAACGTAGACTATATGTGGGACGAGGAGGAAGAA------GAAGAGG 373
bon  GGACAACGTAGACTATATGTGGGACGAGGAGGAAGAA------GAAGAGG 373
ora  GGACAACGTAGACTATATGTGGGACGAGGAGGAAGAA---------GAGG 379
mrm  ------------------------------------------------
gor  GGACAACGTAGACTATATGTGGGACGAGGAGGAAGAA------GAAGAGG 382
hmn  GGACAACGTAGACTATATGTGGGACGAGGAGGAAGAAGAA---GAAGAGG 388
bbb  GAACAATGTGGATTATGTGTGGGACGGGGAAGAAGTG------GACAAGG 361
arm  GGAAGACATGGACTATGTTTGGAACCAGGAGGAA--------------G 364
alp  GGAC---ATGGATTATGTGTGGGACCAGGAGGAAGAA------------G 373
cat  GGACAACATGGACTATGTGTGGGACCAGGAGGAAGAA------GAGGAGG 379
mle  GGGCGACGTGGATGATGGGTGGGACCAGGAGGAAGAAGAGGAGGAGGAGG 391
dog  GGACAATATGGACTATGTGTGGGACCAGGAGGAAGAA---------GAGG 364
mse  GGGCAATGCAAACTCTGAGTGGGATGACGAGGAA--------------G 322
rat  GGGCAATGCAGACTCTGTGTGGGATGATGAGGAA--------------G 328

cow  AAGATACGAACTACTACCTGGGAGGCTTGAGACATGACCTGAGAATTAAC 423
pig  AAGAGAGGGACTACTACTTGGGAGGCTTGAGACAAGACCTGAGAATTGAT 420
pda  AAGACTGGGACTGTTACCTGGGAGGCTTGAGACACGACCTGAGAATTGAC 435
ele  AAGATCGAGACTGTTACCTGGATGGCTTGAGACATGACTTGAGAATTGAC 405
agm  AAGATCAGGACTATTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 435
tal  AAGATCAGGACTATTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 435
pfl  AAGATCAGGACTATTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 435
sla  AAGATCAGGACTATTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 435
flm  AAGATCAGGACTATTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 435
col  AAGATCAGGACTATTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 435
rhe  AAAATCAGGACTATTACCTAGGAGGCTGGAGACCTGACCTGAGAATTGAT 435
chm  AAGATCAGGACTGTTACCTAGGAGGCTTGAGACCTGACCTGAGGATTGAT 423
bon  AAGATCAGGACTGTTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 423
ora  AAGATCAGGACTATTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 429
mrm  ------------------------------------------------
gor  AAGATCAGGACGATTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 432
hmn  AAGATCAGGACTATTACCTAGGAGGCTTGAGACCTGACCTGAGAATTGAT 438
bbb  AAGATCGGGACTATTACCCAGGAGGCTTGAGACTTGACCTGAGAACTGAT 411
arm  ACGTGGAGGAAGATTATCCAGGAGGCTTGAGACATGACCTAAGAATTGAC 414
alp  AGGAGATAGACTGTTACCTGGGAGGCTTGAGACATGACCTGAGA---GAC 420
cat  AAGATCAGGACTATTACCTGGGAAGCTTGAGACATGACCTGAGAATTGAC 429
mle  AAGATCGGGACTATTACCTAGGAGGCTTGAGACGTGACCTGAGGATTGAT 441
dog  AAGACCGGCACTATTACCTGGGAGGCTTGAGACATGACCTGAGAATTGAC 414
mse  AAGACAGGAACAGT---TTACAAGGATTGGTTCATGACCTGAGAATTAGG 369
rat  AAGACTGGAACAGT---TTACAAGGACTGGTACATGACCTGAGAATTAGG 375

cow  GTCTACCTGCAAGAGGAG---GAGATTTTGGAAGAATACGATGAGGACGA 470
pig  GTCTACCTGGAAGAAGAGGAGGAGATAGTGGAAGAATACGATGAAGACGA 470
pda  GTCTACCCAGAAGAG------GAGATACTGGAAGAATACAATGAGGACGA 479
ele  GTCTACCCAGAAAAACAA---GAGATATTTGAAGAATATGATGAGGATGA 452
agm  GTCTACCGAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 479
tal  GTCTACCGAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 479
pfl  GTCTACCGAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 479
sla  GTCTACCGAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 479
flm  GTCTACCGAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 479
col  GTCTACCGAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 479
rhe  GTCTACCGAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 479
chm  GTCTACCGAGAAGAAGAA---GAAACACTGGAAGCATACGATGAGGACGA 470
bon  GTCTACCGAGAAGAAGAA---GAAATACTGGAAGCATACGATGAGGACGA 470
ora  GTCTACCAAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 473
mrm  ------------------------------------------------
gor  GTCTACGGAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 476
hmn  GTCTACCGAGAAGAA------GAAATACTGGAAGCATACGATGAGGACGA 482
bbb  GTCTACCCAGAAGATGAG---GAGGAATTGGAAGCTTACAATGAGAAAGA 458
arm  GTTTACGAAGAG---------GAGATATTGGAAGACTACTATGAGGACGA 455
alp  GTCTACCCCAAAGAAGAT---GGGACATTGGAAGAATACGATGAGGACGA 467
cat  GTCTACCCAGAAGAG------GAGATATCGGAAGAATACGAGGACGAGGA 473
mle  GTTTACCCAGAAGGGGAG---GCGGCACTGGAAGCTTACAGTGAGGGGGA 488
```

```
dog  GTCTACTTAGAA---------GAGATACTGGAAGAATACAACGAAGACGA 455
mse  GTTTTTCCAGAAGAGAGAGATGAACCCCCCCACAATGGCCACCAGTACCA 419
rat  GTTTTTCCAGAAGAAAGAGATGAACACCCCCACGATGGCCACCAATACCA 425

cow  CGAAGAG------------------------------------------ 477
pig  CGAAGAG------------------------------------------ 477
pda  CCAAGAG------------------------------------------ 486
ele  AGAG-------------------------------------------- 456
agm  AGATGAAGAG--------------------------------------- 489
tal  AGATGAAGAG--------------------------------------- 489
pfl  AGATGAAGAG--------------------------------------- 489
sla  AGATGAAGAG--------------------------------------- 489
flm  AGATGAAGAG--------------------------------------- 489
col  AGATGAAGAG--------------------------------------- 489
rhe  AGATGAAGAG--------------------------------------- 489
chm  AGATGAAGAG--------------------------------------- 480
bon  AGATGAAGAG--------------------------------------- 480
ora  AGATGAAGAG--------------------------------------- 483
mrm  ------------------------------------------------
gor  AGAGGAAGAG--------------------------------------- 486
hmn  AGATGAAGAG--------------------------------------- 492
bbb  AGATGAAGAG--------------------------------------- 468
arm  GGATGATGAG--------------------------------------- 465
alp  TGAAGAG----------------------------------------- 474
cat  AGAG-------------------------------------------- 477
mle  GGAG-------------------------------------------- 492
dog  CCAAGAG----------------------------------------- 462
mse  TCGGTTTGGTCGCTACCGCCATCGCCACCGCCACCCTCCAATCTTCCACC 469
rat  TC------------ACTATGGCCGATACCGACACCGTCCAGTTTTCCGCC 463

cow  -------------------------------CTGTACCCTGACACTCAC 495
pig  -------------------------------CTGTATCTTGACAGGCAT 495
pda  -------------------------------CTGTATCCTGAGACCCAC 504
ele  -------------------------------CTGTATTCTGACAGCCAT 474
agm  -------------------------------CTGTATCCTGACATCCAC 507
tal  -------------------------------CTGTATCCTGACATCCAC 507
pfl  -------------------------------CTGTATCCTGACATCCAC 507
sla  -------------------------------CTGTATCCTGACATCCAC 507
flm  -------------------------------CTGTATCCTGACATCCAC 507
col  -------------------------------CTGTATCCTGACATCCAC 507
rhe  -------------------------------CTGTATCCTGACATCCAC 507
chm  -------------------------------CTGTATCCTGACATCCAC 498
bon  -------------------------------CTGTATCCTGACATCCAC 498
ora  -------------------------------CTGTATCCTGACATCCAC 501
mrm  ------------------------------------------------
gor  -------------------------------CTGTATCCTGACATCCAC 504
hmn  -------------------------------CTGTATCCTGACATCCAC 510
bbb  -------------------------------GTGTATCCTGACACCCGC 486
arm  -------------------------------CCGTATGCTGGTACCCAC 483
alp  -------------------------------CTGTA---TGACACCCAC 489
cat  -------------------------------CTGTATCCTGACACCCAC 495
mle  -------------------------------GTGTATCCTGACACCCGC 510
dog  -------------------------------CTGTATCCTGACACCCAT 480
mse  GTGGTCCCCCACATCCGCCTGTGCGTCGGCAGCTCTATCCAGACGCCCGG 519
rat  GTGGTCCCCCACATCCACCTGTGCGTCGGCAGCTCTATCCAGATGCCCGA 513

cow  CTGGCC------------------------------------------ 501
pig  CTGCCT------------------------------------------ 501
pda  CTG--------------------------------------------- 507
ele  CTGCCT------------------------------------------ 480
agm  CCGCCT------------------------------------------ 513
tal  CCGCCT------------------------------------------ 513
pfl  CCGTCT------------------------------------------ 513
sla  CCGTCT------------------------------------------ 513
flm  CCGTCT------------------------------------------ 513
col  CCGTCT------------------------------------------ 513
rhe  CCGCCT------------------------------------------ 513
chm  CCGCCT------------------------------------------ 504
bon  CCGCCT------------------------------------------ 504
ora  CCGCCT------------------------------------------ 507
mrm  ------------------------------------------------
```

```
gor  CCGCCT------------------------------------------- 510
hmn  CCGCCT------------------------------------------- 516
bbb  CCA---------------------------------------------- 489
arm  CTGCCT------------------------------------------- 489
alp  CTGCCT------------------------------------------- 495
cat  CCACCC------------------------------------------- 501
mle  CCACCT------------------------------------------- 516
dog  CTG---------------------------------------------- 483
mse  GTTCCTTCTCCACATGCCCAGGTTCCTCCTCCACATGCC----------- 558
rat  CCACGATCTCCACCTCGAGTACGATCTCCACCTCGAGTACATTCTCCACC 563

cow  -------------CCGCCTCCAGCCCCTCCACGGCAGTTCACCTGCCCCC 538
pig  -------------CCTCCCCTAGCCCCTCCACGGCAGTTCACCTGTCCCC 538
pda  -------------CCTCCTCCGGCCCCTCCACGTCAGTTCACCTGCCCCC 544
ele  -------------CCTCCCCCAACCCCTCAGCATCAGTTCACCAGCCCCC 517
agm  -------------CCTTCCTCGCCCCTTCCAGGGCAGTTCACCTGCCCCC 550
tal  -------------CCTTCCTCGCCCCTTCCAGGGCAGTTCACCTGCCCCC 550
pfl  -------------CCTTCCTCGCCCCTTCTAGGGCAGTTCACCTGCCCCC 550
sla  -------------CCTTCCTCGCCCCTTCCAGGGCAGTTCACCTGCCCCC 550
flm  -------------CCTTCCTCGCCCCTTCCAGGGCAGTTCACCTGCCCCC 550
col  -------------CCTTCCTCGCCCCTTCCAGGGCGGTTCACCTGCCCCC 550
rhe  -------------CCTTCCTCGCCCCTTCCAGGGCAGTTCACCTGCCCCC 550
chm  -------------CCTTCCTTGCCCCTTCCAGGGCAGTTCACCTGCCCCC 541
bon  -------------CCTTCCTTGCCCCTTCCAGGGCAGTTCACCTGCCCCC 541
ora  -------------CCTTCCTTGCCCCTTCCAGGGCAGTTCACCTGCCCCC 544
mrm  ------------------------------------------------- 
gor  -------------CCTTCCTTGCCCCTTCCAGGGCAGTTCACCTGCCCCC 547
hmn  -------------CCTTCCTTGCCCCTTCCAGGGCAGTTCACCTGCCCCC 553
bbb  ----------------------------CCAGGGCAGTTCATTTGCCCCC 511
arm  -------------CTTCCTGCGACCCCTCGGCGGCCGTTCACCTGCCCCC 526
alp  -------------CCTCCCCTGGCCCCTCCATGGCTATTCACCTGCCCCC 532
cat  -------------CCTCCTCCGCCCCCTCCACGTCAGTTCATCTGTCCCC 538
mle  ----------------------------CCGCGGCAGTTCACCTGCCCAC 538
dog  -------------CCTCCACCCACCCGTCCACGTCAGTTCACCTGCCCCC 520
mse  -CAGGTTCCTTCTCTACCTCGGCCCACACCACAGGTCTTCAGCTGCCCGC 607
rat  TCGAGTACGTTCTCCACCTCGTCCCACATCACAGGTCTTCAGCTGCCCAC 613

cow  AATGCCGAAAGAGCTTTAAGCGTCGCAGCTTTCGTCCCAACTTGCAACTG 588
pig  AATGCCGAAAAAGCTTTACACGTCGAAGCTTTCGTCCTAACTTGCAGCTG 588
pda  AGGGCCGAAAGAGCTTTACACGTTGCAGCTTTCATCCCAACTTGCAGCTG 594
ele  AGTGCCTGAAGAACTTTACACGTNCC---TTTCTTCCCAACTTGCAGCTG 564
agm  AGTGCCGAAAGAGCTTTACACGTCGCATCTTTCGTCCCAACTTGCAGCTG 600
tal  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 600
pfl  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 600
sla  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 600
flm  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 600
col  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 600
rhe  AGTGCCGAAAGAGCTTCACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 600
chm  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 591
bon  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 591
ora  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 594
mrm  ------------------------------------------------- 
gor  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 597
hmn  AGTGCCGAAAGAGCTTTACACGTCGCAGCTTTCGTCCCAACTTGCAGCTG 603
bbb  AGTGCCAAAAGAGCTTTATA---TGCAGCTTTCGTTCCAACATGCAGCTG 558
arm  AGTGCCGCAAGAGCTTTACACGTCGCAGCTTTCGCCCCAACTTGCCACTG 576
alp  GGTGCCGCCAGAGCTTTACCCGTCGCAGCTTTCGTCCCAACTTG---CTG 579
cat  AGTGCCGAAAGAGCTTTAAGCGTCGCAGCTTTCGTCCCAACTTGCAGCTA 588
mle  AGTGCCGAAAGGGCTTTACACGTCGCAGCTTCCGCCCCAACCTCCAGCTG 588
dog  AGTGCCGAAAGAGCTTTACCCGTCGCAGCTTTCGTCCCAACTTGCAGCTG 570
mse  AATGCCGAAGGACTTTTCCAAGTCGCAGTTTTCGACCCAATTTGCAGCTG 657
rat  AATGCCGAAGGACATTTCCAAGTCGCAGTTTTCGACCCAATTTGCAGCTG 663

cow  GCGAACATGGTCCAGATAATTCGCCAGATGTGTCCCACTCCTAATCGAGA 638
pig  GCCAACATGGTCCAGATAATTAGCCAGATGTGTCCTACTCCTAATCGAGG 638
pda  GCCAACATGGTCCAGATAATTCGCCAAATGTGCCCCACTCCTTATCGAGG 644
ele  GCCAACATGGTCCAGATAATTCACCAGATGTGCCCCATGCCTTACCAAGG 614
agm  GCCAACATGGTCCAGATAATTCGCCAGATGTGCCCCACTCCTTGTCGGCG 650
tal  GCCAACATGGTCCAGATAATTCGCCAGATGTGCCCCACTCCTTGTCGGGG 650
pfl  GCTAACATGGTCCACATAATTCGCCAGATGTGCCCCACTCCTTGTCGGGG 650
sla  GCTAACATGGTCCACATAATTCGCCAGATGTGCCCCACTCCTTGTCGGGG 650
```

```
flm  GCTAACATGGTCCACATAATTCGCCAGATGTGCCCCACTCCTTGTCGGGG 650
col  GCCAACATGGTCCACATAATTCGCCAGATGTGCCCCACTCCTTGTCGGGG 650
rhe  GCCAACATGGTCCAGATAATTTGCCAGATGTGCCCCACTCCTTGTCGGGG 650
chm  GCCAACATGGTCCAGATAATTCGCCAGATGTGCCCCACTCCTTATCGGGA 641
bon  GCCAACATGGTCCAGATAATTCGCCAGATGTGCCCCACTCCTTATCGGGG 641
ora  GCCAACATGGTCCAGATAATTCGCCAGATGTGCCCCACTCCTTATCGGGG 644
mrm  --------------------------------------------------
gor  GCCAACATGGTCCAGATAATTCGCCAGATGTGCCCCACTCCTTATCGGGG 647
hmn  GCCAACATGGTCCAGATAATTCGCCAGATGTGCCCCACTCCTTATCGGGG 653
bbb  GCTACCATGTTCCAGATAATTCGCCAGATG---CCCACTCGTTATGGGGG 605
arm  GCCAACATGGTCCAGATAATTCGCCAGATGTGCCCCACTCCTTGTCGAGG 626
alp  GCCAACATGGTCCAGATAATTCGCCAGATGTGCCCCACTCCTGATAAAGG 629
cat  GCCAACATGGTCCAGATAATTCGCCAGATGAGCCCCACTCCTTATCGAGG 638
mle  GCCAACATGGTCCAGGTGATCCGGCAGATGCGCCCCACTCCTTACCGAGG 638
dog  GCCAACATGGTGCAGATAATTCGCCAAATGTGCCCCACTCCTTATCAAGG 620
mse  GCCAACATGGTCCATATAATTCGCCAGATTTGCCATACTCCATGA----- 702
rat  GCCAACATGGTCCATATAATTCGCCAGATTTGCCATACGCCATGA----- 708

cow  GAGCCGGGTGAATGATCAGGACATCTGCTCCAAACACCAGGAAGCTCTGA 688
pig  GAGCAGAGAGAATGATCAGGGCATCTGTTCCAAACACCAAGAAGCCCTGA 688
pda  AAGCTGGGGGAATGATGAGGGCATCTGCTCCAAACACCAGGAAGCCCTGA 694
ele  GAGCCGAGGAAATGATCAGGGCATCTGCTCCAAACACCAGGAAGCCCTGA 664
agm  GACCCGGAGTAATGATCAGGGCATGTGCTTCAAACACCAGGAAGCCCTGA 700
tal  GAACCGGAGTAATGATCAGGGCATGTGCTTCAAACACCAGGAAGCCCTGA 700
pfl  GAACCGGAGTAATGATCAGGGCATGTGCTTCAAACACCAGGAATCCCTGA 700
sla  GAACCGGAGTAATGATCAGGGCATGTGCTTCCAACACCAGGGATCCCTGA 700
flm  GAACCGGAGTAATGATCAGGGCATGTGCTTCAAACACCAGGAATCCCTGA 700
col  GAACCGGGGTAATGATCAGGGCATGTGCTTCAAACACCAGGAATCCTTGA 700
rhe  GAACCGGAGTAATGATCAGGGCATGTGCTTCAAACATCAGGAAGCCCTGA 700
chm  AAACCGGAGTAATGATCAGGGCACGTGCTTTAAACACCAGGAAGCCCTGA 691
bon  AAACCGGAGTAATGATCAGGGCATGTGCTTTAAACACCAGGAAGCCCTGA 691
ora  AAACCGGAGAAATGATCAGGGCATGTGCTTTAAACACCAGGAAGCCCTGA 694
mrm  --------------------------------------------------
gor  AAACCGGAGTAATGATCAGGGCATGTGCTTTAAACACCAGGAAGCCCTGA 697
hmn  AAACCGGAGTAATGATCAGGGCATGTGCTTTAAACACCAGGAAGCCCTGA 703
bbb  GAAGGGGAGGACT---GGGGGCATCTGTCCCAAACACCAGGAAGCCCTGA 652
arm  GAGCCGAGGGAACGAGCAGGGAATCTGCTCCAAACACCAAGAAGCCCTGA 676
alp  GAGCCAGGGAAATGATCAGGGCATCTGCTCCAAACACCAAGAAGCCTTCA 679
cat  AAGTCGGGGGAATGATCAGGGCATCTGCTCCAAACACCAGGAAGCCCTGA 688
mle  GAGCCGGGCGAACGAGCAGGGCGTCTGTCCCGAACACCAGGAAGCCCTGA 688
dog  AAGCCGGGGGAATGATCAGGGCATCTGCTCTAAACACCAGGAAGCCCTGA 670
mse  --------------------------------------------------
rat  --------------------------------------------------

cow  AACTTTACTGTGAGGTGGACAAAGAGGCCATCTGTGTGATATGTCGAGAA 738
pig  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGTCGGGAA 738
pda  AGCTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCAGGAA 744
ele  AACTCTTCTGTGAGGTGGACAAAGAGGTTATCTGTGTCGTGTGCCAAGAA 714
agm  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 750
tal  AACTCTTCTGTGAGGTGGACAAGGAGGCCATCTGTGTGGTGTGCCGAGAA 750
pfl  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 750
sla  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 750
flm  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 750
col  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGGGCCGAGAA 750
rhe  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 750
chm  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 741
bon  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 741
ora  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 744
mrm  --------------------------------------------------
gor  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 747
hmn  AACTCTTCTGTGAGGTGGACAAAGAGGCCATCTGTGTGGTGTGCCGAGAA 753
bbb  AACTCTTT------GTCGAC---GAGGCCATCTGTGTGGTGTGC---GAA 690
arm  AGCTCTTCTGTGAGGTGGACGAAGAGGCCATCTGTGTGGTGTGCCAGGAA 726
alp  GGCTCTTCTGTGAGGTGGACAAAGAGGCTATCTGTGTGGTGTGT---GAA 726
cat  AACTTTTTTGTGAAGTGGACAAAGAGGCTATCTGTGTGGTGTGCCGAGAA 738
mle  AACTCTTCTGCGAGGTGGACGAAGAGGCCATCTGGGTGGTGTGCCGGGAA 738
dog  AGCTCTTCTGCGAGGTGGATGAAGAGGCCATCTGTGTGGTGTGCCGAGAA 720
mse  --------------------------------------------------
rat  --------------------------------------------------

cow  TCCAGGAGCCACAAACAGCATAGTGTGGTGCCATTAGACGAAGCGGTGCA 788
```

```
pig  TCCAGGAGCCACAAACAGCACAGTGTGGTACCATTGGAGGAAGTGGCACA 788
pda  TCCAGGAGCCACAAACAGCACAGCGTGGTGCCATTGGAGGAGGTGGTGCA 794
ele  TCCTGGAGCCACAAACAGCACAACGTGGTGCCANTTGAAGAGGTGGTGCA 764
agm  TCCAGGAGCCACAAACACCACAGCGTGTTGACTTTGGAGGAGGTGGTTCA 800
tal  TCCAGGAGCCACAAACACCACAGCGTGGTGCCTTTGGAGGAGGTGGTTCA 800
pfl  TCCAGGAGCCACAAACACCACAGCGTGGTGCCTTTGGAGGAGGTGGTTCA 800
sla  TCCAGGAGCCACAAACACCACAGCGTGGTGCCTTTGGAGGAGGTGGTGCA 800
flm  TCCAGGAGCCACAAACACCACAGCGTGGTGCCTTGGCAGGAGGTTGTTCA 800
col  TCCAGGAGCCACAAACAACACAGCGTGGTGCCTTTGGAGGAGGTGGTTCA 800
rhe  TCCAGGAGCCACAAACAGCACAGCGTGGTGCCTTTGGAGGAGGTGGTTCA 800
chm  TCCAGGAGCCACAAACAGCACAGCGTGCTGCCTTTGGAGGAGGTGGTTCA 791
bon  TCCAGGAGCCACAAACAGCACAGCGTGCTGCCTTTGGAGGAGGTGGTTCA 791
ora  TCCAGGAGCCACAAACAGCACAGCGTGGTGCCTTTGGAGGAGGTGGTTCA 794
mrm  --------------------------------------------------
gor  TCCAGGAGCCACAAACAGCACAGCGTGCTGCCTTTGGAGGAGGTGGTGCA 797
hmn  TCCAGGAGCCACAAACAGCACAGCGTGCTGCCTTTGGAGGAGGTGGTGCA 803
bbb  TCCAGGAGCCACAAACGGCAC---ATGGGGCCTTTGGAGGAAGTGATGCA 737
arm  TCCAGGAGCCACAAACAGCACAGCGAGGTGCCATTGGAGGAGGTGGTGCA 776
alp  TCCAGGAGCCACAAACAGCACAGTGTGGTGCCACTGGAGGAAGTGGTGCA 776
cat  TCCAGGACCCACAAACAACACAGCGTGGTGCCAATGGAGGAGGTGGTGCA 788
mle  TCCGGGAGCCACAGACAGCACAGCGTGGTGCCGCTGGACGAGGTGGTGCA 788
dog  TCCAGGAGCCACAAACAGCACAGCGTGGTGCCATTGGAGAAGGTGGTGCA 770
mse  --------------------------------------------------
rat  --------------------------------------------------

cow  CGAGTACAAGGAGAAAAAAATGGAGAAACTAGTG-------AAACCTTGC 831
pig  TGAGTACAAGGTGAGAGGCACGAGGGAATGTGGG---------GATTTGA 829
pda  GGAGTACAAGGAGAAAAAAATGGAGAAACTTGTG-------AAGCCTTGC 837
ele  GGAGTACACGGGAATAAAGTTGGAAAGAACTCCT-------TAG------ 801
agm  GGACTACCAGGAAATAAAGTTGGAAACA---CTT------GTGGGAATAC 841
tal  GGAGTACCAGGGAATAAAGTTGGAAACA---CTT------GTGGGAATAC 841
pfl  GGAGTACCAGGAAATAAAGTTGGAAACA---CTT------GTGGGAATAC 841
sla  GGAGTACCAGGAAATA---TTG---ACA---CTT------GTGGGAATAC 835
flm  GGAGTACCAGGAAATAAAGGTGGAAACA---CTT------GTGGGAATAC 841
col  GGAGTACCAGGGAATAAAGTTGGAAACA---CTT------GTGGGAATAC 841
rhe  GGAGTACCAGGAAATAAAGTTGGAAACA----TTG-----GTGGGAATAC 841
chm  GGAGTACCAGGAAATAAAGTTGGAAACAACTCTG------GTGGGAATAC 835
bon  GGAGTACCAGGAAATAAAGTTGGAAACAACTCTG------GTGGGAATAC 835
ora  GGAGTACCAGGAAATAAAGTTGGAAACAACTCTG------GTGGGAATAC 838
mrm  ----------GAAATAAAGTTGGAAAGAACTCTT----TGGTGGAAATAC 36
gor  GGAGTACCAGGAAATAAAGTTGGAAACAACTCTG------GTGGGAATAC 841
hmn  GGAGTACCAGGAAATAAAGTTGGAAACAACTCTG------GTGGGAATAC 847
bbb  GGAGTACAAGGAAATAAAGTTGGAAAGAACTCCTC---TGGTAGGAATAC 784
arm  GGAGTACAAGAAAATAAAGTGGGAAAGAACTCTG------GCAGGAATA- 819
alp  GGAGTATAAGAAAATC---TTGGAAAGAATTCCT---TGGCTGGTAGGAA 820
cat  GGAGTACAAGGAGAGAAAGTTGAAAATAACTCCT---TGGCTGGTGGGAA 835
mle  GGAGTACAAGGAAATAAAGTTGGAAAGAACTCCT---CTGGTGGGAATAC 835
dog  GGAGTACAAGGTTCAGCTGTTGACTGCTGTATCA---CAAGTCAGAGTGT 817
mse  --------------------------------------------------
rat  --------------------------------------------------

cow  ATTCCAAGTGCTGCCATAACTTTGAGAGGAAATTGAGGAATTCCAGAATT 881
pig  AGG--GAGTGGAGAAAAAAC----AAGGGGAGCTGGGTACTTGGTAG--- 870
pda  ATTTCACGTGCTGCCACAACTCTGAGAGGAAATTGA-------------- 873
ele  --------------------------------------------------
agm  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAATAA--- 888
tal  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAATAA--- 888
pfl  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAATAA--- 888
sla  TTCAG---AAGCAA---AGCATTCACAGCAAGGCCTATAATCAATAA--- 876
flm  TTCAGATAGAGCAA------ATTCACAGCAAGGCCTATAATCAATAA--- 882
col  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAATAA--- 888
rhe  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAATAA--- 888
chm  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAGTAA--- 882
bon  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAGTAA--- 882
ora  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAATAA--- 885
mrm  TTCAGATAAAGCAAGAGAGCATTCACAGCAAGGCCTGTAAACAATAA--- 83
gor  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAGTAA--- 888
hmn  TTCAGATAGAGCAAGAAAGCATTCACAGCAAGGCCTATAATCAGTAA--- 894
bbb  AAAGGATAGAAGAGGAAAGCACTGAGAGCAAGACCTATAATCAGTGA--- 831
arm  --AAGGTAGTGCAAGGACACATTCAGAGCAAGGCCTAA------------ 855
alp  CACAGAGAAAGAAA---AGCATTCAGAGTAAAGCCTATAATCAATGA--- 864
```

```
cat  TAAAGATAGAACAAGAA---ATTCAGAGTAAGGCCTATAATTGA------ 876
mle  AAAGAAGATAG------------------------------------- 846
dog  TTAAACTTGCTACAAGCTGTTCCTGTGGGTGTACCTGGGACCTCATCTTT 867
mse  ---------------------------------------------------
rat  ---------------------------------------------------

cow  TCTTCATCAACTTCTGCATCTGAATTCATTGACTGAGCCTAGCAAACATC 931
pig  ---------------------------------------------------
pda  ---------------------------------------------------
ele  ---------------------------------------------------
agm  ---------------------------------------------------
tal  ---------------------------------------------------
pfl  ---------------------------------------------------
sla  ---------------------------------------------------
flm  ---------------------------------------------------
col  ---------------------------------------------------
rhe  ---------------------------------------------------
chm  ---------------------------------------------------
bon  ---------------------------------------------------
ora  ---------------------------------------------------
mrm  ---------------------------------------------------
gor  ---------------------------------------------------
hmn  ---------------------------------------------------
bbb  ---------------------------------------------------
arm  ---------------------------------------------------
alp  ---------------------------------------------------
cat  ---------------------------------------------------
mle  ---------------------------------------------------
dog  ACACAGACTTGCTACACATCTGAAATCCTCACCTCCAACTGCACTATGAC 917
mse  ---------------------------------------------------
rat  ---------------------------------------------------

cow  AGGTTTAA---------------------------- 939
pig  ------------------------------------
pda  ------------------------------------
ele  ------------------------------------
agm  ------------------------------------
tal  ------------------------------------
pfl  ------------------------------------
sla  ------------------------------------
flm  ------------------------------------
col  ------------------------------------
rhe  ------------------------------------
chm  ------------------------------------
bon  ------------------------------------
ora  ------------------------------------
mrm  ------------------------------------
gor  ------------------------------------
hmn  ------------------------------------
bbb  ------------------------------------
arm  ------------------------------------
alp  ------------------------------------
cat  ------------------------------------
mle  ------------------------------------
dog  CTTAGCTCTTAGCCTTCTAGTGTAG------------ 942
mse  ------------------------------------
rat  ------------------------------------
```

Figure B.1 Alignment of *TRIM52* sequences. (A) *TRIM52* sequences from animals collected from BLAST (Altschul, et al. 1990) and BLAT (Kent 2002) searches, as well as our own sequencing of primate orthologs were aligned using Clustal X (Larkin, et al. 2007). (**B)** Sequences were translated and resulting peptide sequences were also aligned via Clustal X.

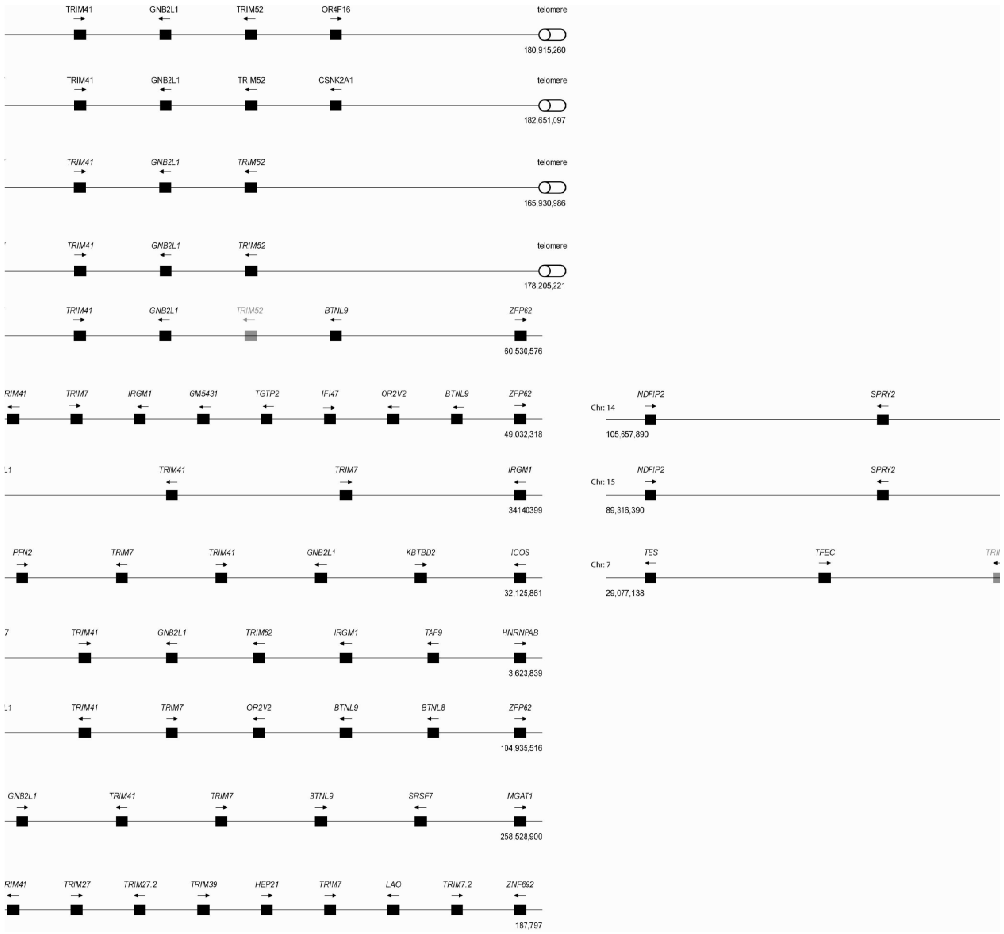

TRIM41
GNB2L1
TRIM52
ORF16
telomere
180,915,260
TRIM41
GNB2L1
TRIM52
CSNK2A1
telomere
182,651,097
TRIM41
GNB2L1
TRIM52
telomere
165,930,586
TRIM41
GNB2L1
TRIM52
telomere
178,205,221
TRIM41
GNB2L1
TRIM52
BTNL9
ZFP62
80,534,576
TRIM41
TRIM7
IRGM1
GNS431
TGTP2
IFI47
OP2V2
BTNL9
ZFP62
49,032,218
TRIM41
TRIM7
IRGM1
34,140,369
IPCN2
TRIM7
TRIM41
GNE2L1
KBTB02
ICOS
32,125,851
TRIM41
GNB2L1
TRIM52
IRGM1
TAF9
3,623,839
TRIM41
TRIM7
OP2V2
BTNL9
BTNL8
ZFP62
104,935,516
GNB2L1
TRIM41
TRIM7
BTNL9
SRSF7
MGAT1
258,528,600
TRIM41
TRIM27
TRIM27.2
TRIM39
HEP21
TRIM7
LAO
TRIM7.2
ZNF852
187,797
MDFIP2
SPRY2
Chr 14
105,357,890
MDFIP2
SPRY2
Chr 15
86,315,390
TES
TPEC
Chr 7
26,077,133

Figure B.2 Genomic localization of *TRIM41* and *TRIM52* across mammals. Based on the human genome assembly (March 2006, UCSC genome browser), *TRIM52* and *OR4F16* are the last known genes before the chromosome 5q telomere. *TRIM41* and *TRIM7* are almost immediate neighbors, with only the *GNB2L1* gene encoded between. This subtelomeric localization is at least conserved throughout Hominoids and Old World monkeys that we were able to evaluate via the UCSC genome browser (Kent, Sugnet et al. 2002). In cases where *TRIM52* (or *TRIM41*) was predicted to be pseudogenized, we represented their location with grayed font. We were unable to identify a rabbit *TRIM52* by prediction programs or BLAST (Altschul, Gish et al. 1990) and BLAT (Kent 2002) searches. Rabbit encodes a second copy of *TRIM41* on chromosome 7; although, sequence analysis predicts that is a pseudogene. We were unable to locate a syntenic mouse and rat *TRIM52*. According to Ensembl's prediction algorithm, mouse and rat *TRIM52-like* are localized to chromosome 14 and 15, respectively.

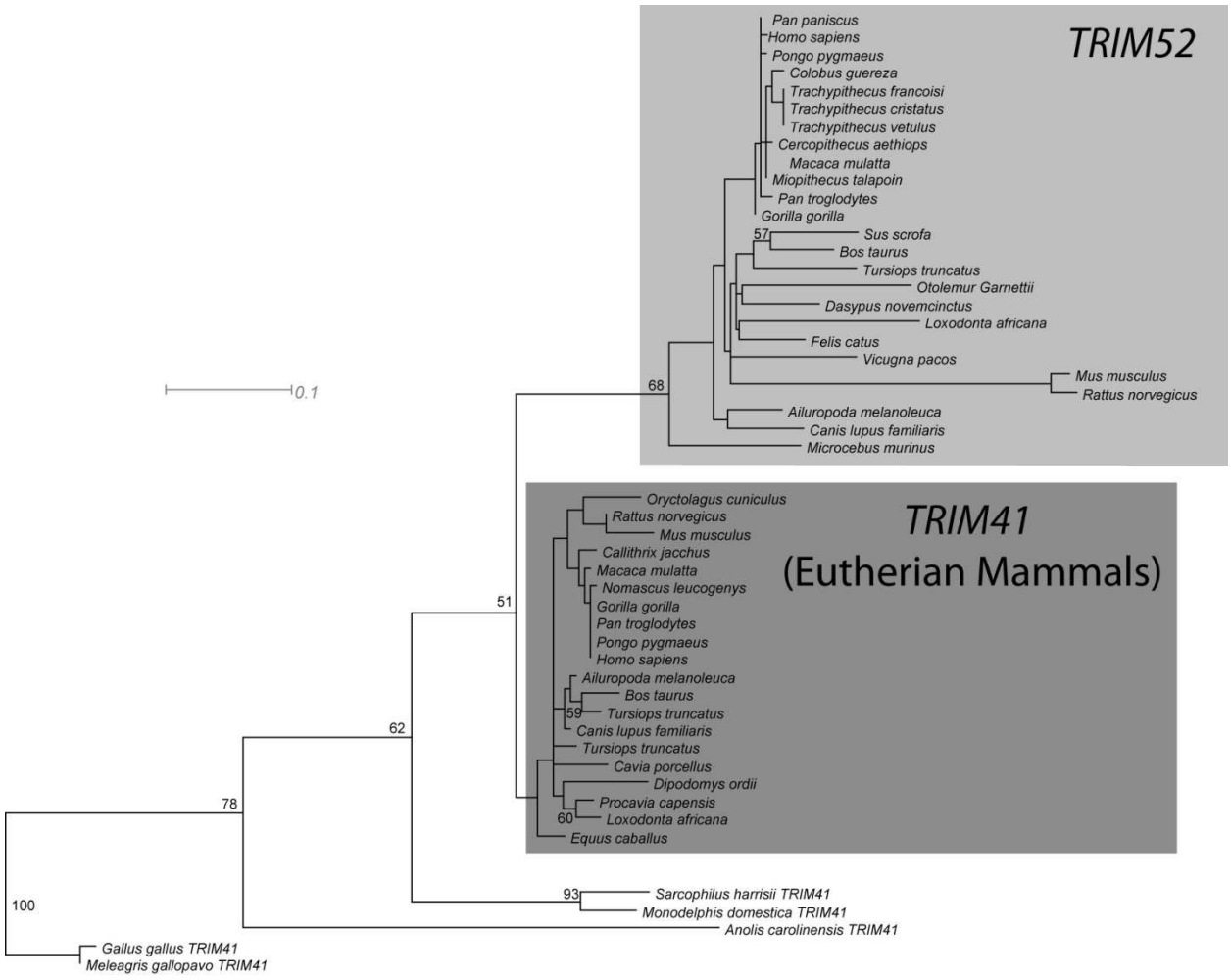


Figure B.3 Phylogenetic relationship of *TRIM52*, *TRIM41,* and *TRIM52-like* genes. A phylogram of *TRIM52* and *TRIM41* that were recovered from BLAST (Altschul, Gish et al. 1990) searches and annotated by Ensembl (Flicek, Amode et al. 2012) was built using a maximum likelihood based approach via PhyML (Guindon, Dufayard et al. 2010). We included mouse and rat TRIM52-*like* sequences. Statistical support is represented by Bootstrap values, collected from 100 iterations.

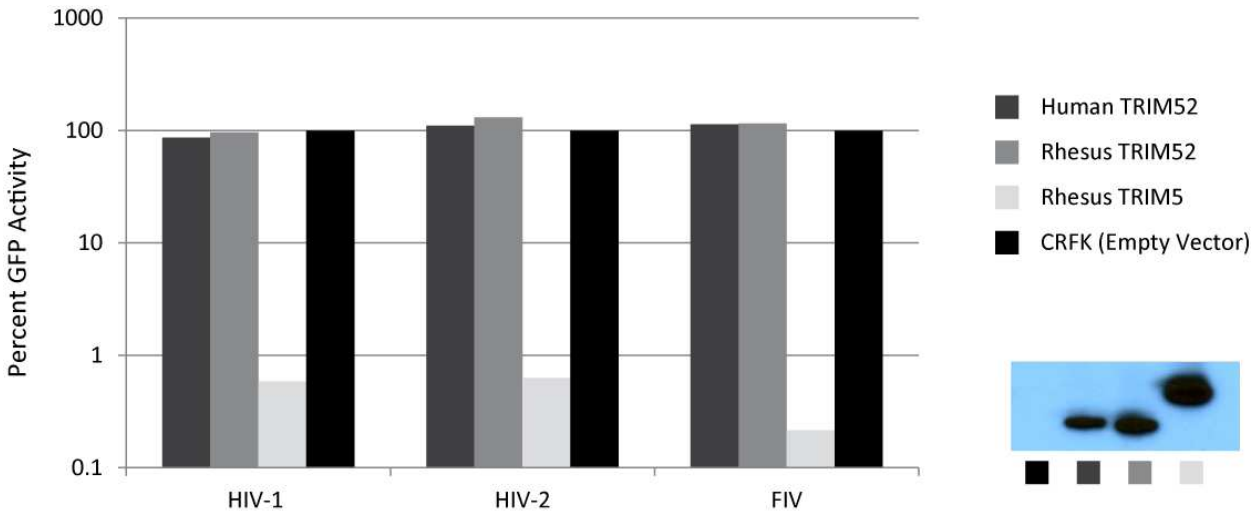


Figure B.4 Testing antiviral activity of *TRIM52* against candidate retroviruses. CRFK cell lines stably expressing HA-tagged rhesus TRIM5alpha (positive control), HA-tagged rhesus TRIM52, HA-tagged human TRIM52, and an empty vector (negative control) were assayed for antiviral activity against HIV-1 (BRU strain), HIV-2 (ROD9 strain), and FIV. The Y-axis reflects infectivity, determined by the percent of GFP reported while the X-axis lists the viruses that were used. Virus titers were set to recover ~15 percent infection. Values of reported GFP were normalized to CRFK, setting CRFK to 100 percent for each virus. We confirmed the stable expression of TRIM5 and TRIM52 proteins by Western blot analysis, using 40ug of protein extract for each sample (Lane1: Negative control; Lane 2: Human TRIM52, Lane 3: Rhesus TRIM52, Lane 4: Rhesus TRIM5). This assay was repeated once to demonstrate reproducibility.

Table B.1 PAML screen of primate TRIM genes

TRIM gene	M7vsM8	dN/dS	P-value	% of positively selected sites	Primates
TRIM1	0.000866	1.28887	0.999567094	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Cae, Marm, Aotus, Tarsier, MLemur, BBaby
TRIM2	5.682542	1.80541	0.058351454	0.885	Human, Chimp, Gor, WCG, Rhesus, Marm, MLemur
TRIM3	0.000294	1	0.999853011	4.25	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm, MLemur
TRIM4	0.083832	1.16824	0.958950329	32.785	Human, Chimp, Orang, WCG, Rhesus, Marm, MLemur
TRIM5	73.47338	3.29159	1.11035E-16	20.464	Human, Chimp, Gor, Orang, Sia, CAe, CTa, Patas, Rhesus, Bab, Douc, Colobus, Woolly, Spider, Howler, Saki, Pygmy, Squirrel, CTTam, Dusky, RBTam, BGTiti
TRIM6	1.191148	2.34464	0.55124606	6.64	Human, Chimp, Orang, WCG, Bab, Rhesus, Marm
TRIM7	6.401934	26.63851	0.040722806	0.317	Human, Chimp, WCG, Marm
TRIM8	2.847932	1.80249	0.240757278	1.711	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM9	0.000018	1	0.999991	0	Human, Chimp, Gor, Orang, WCG, Rhesus
TRIM10	0.969942	10.00483	0.615715052	0.322	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM11	0.000416	64.62517	0.999792022	0	Human, Chimp, Orang, WCG, Rhesus, Marm
TRIM13	0.000012	1	0.999994	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm, Tar, BBaby, MLemur
TRIM14	0.000000	1	1	0	Human, Chimp, Orang, Rhesus, Marm
TRIM15	4.528130	3.09674	0.103927161	5.677	Human, Chimp, Orang, WCG, Rhesus, Marm
TRIM16	1.600080	4.00164	0.449310991	1.024	Human, Chimp, Orang, WCG, Rhesus, Marm
TRIM17	0.513238	3.25894	0.773662923	0.228	Human, Chimp, Orang, WCG, Rhesus, Marm, MLemur
TRIM18	0.000180	1	0.999910004	0	Human, Orang, WCG, Rhesus, Marm, Tar, MLemur
TRIM19	0.290412	2.570260	0.864844117	0.801	Human, Chimp, Bon, Gor, Orang, WCG, WHG, Sia, Rhesus, Cae, CHTam, Marm
TRIM20	29.013012	5.09617	5.01077E-07	2.821	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm, Tar, Mlemur
TRIM21	0.519926	6.03724	0.771080115	0.787	Human, Chimp, Gor, Orang, WCG,

					Rhesus, Marm
TRIM22	10.195488	6.16845	0.006110516	4.887	Human, Chimp, Gor, Orang, Sia, WCG, CAe Rhesus, Bab, Sooty, Patas, Fran, Colobus
TRIM23	0.000036	1	0.999982	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM24	0.797534	1.0725	0.67114706	5.731	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM25	18.713110	2.105	0.0000864	12.06	Human, Chimp, Orang, WCG, Rhesus, Marm, BBaby, MLemur
TRIM26	1.326344	1	0.515214479	2.899	Human, Chimp, Orang, WCG, Rhesus, Marm
TRIM27	0.003346	1	0.998328399	0.509	Human, Chimp, Orang, WCG, Rhesus, Marm
TRIM28	8.600874	12.62902	0.013562631	0.589	Human, Chimp, Gor, Orang, Rhesus, Marm
TRIM29	0.000006	1	0.999997	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm, MLemur
TRIM31	9.639178	8.54013	0.008070103	4.525	Human, Chimp, Gor, Orang, WCG, Rhesus
TRIM32	0.001100	7.27754	0.999450151	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm, Tar, MLemur
TRIM33	1.167274	14.90349	0.557865715	0.299	Human, Chimp, Orang, Rhesus, Marm
TRIM34	3.574824	1.82794	0.167392822	40.24	Human, Chimp, WCG, Bab, Rhesus, Marm
TRIM35	0.000008	1	0.999996	0	Human, Chimp, Orang, WCG, Rhesus, Marm
TRIM36	0.489916	1.95617	0.782737413	1.05	Human, Chimp, Gor, WCG, Rhesus, Marm, Tar
TRIM37	0.003194	1	0.998404275	12.705	Human, Chimp, Orang, Rhesus, Marm
TRIM38	8.238808	4.64912	0.016254199	1.951	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm, Tar, BBaby
TRIM39	0.000054	1	0.999973	0	Human, Chimp, Orang, WCG, Rhesus, Marm
TRIM40	1.746434	1.55113	0.417605948	41.519	Human, Chimp, Gor, Orang, Rhesus, Tar
TRIM41	0.000076	1	0.999962001	0	Human, Chimp, Gor, Orang, Rhesus, Marm
TRIM42	0.000020	1	0.99999	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm, MLemur
TRIM44	0.000058	1	0.999971	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm, BBaby
TRIM45	0.375360	4.27924	0.828879906	0.466	Human, Chimp, Orang, WCG, Rhesus, Marm, MLemur
TRIM46	0.001490	1	0.999255277	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm

TRIM47	0.183104	1	0.912513864	5.57	Human, Chimp, Gor, Orang, Rhesus, Marm
TRIM50	0.000054	1	0.999973	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM52	6.202140	9.82912	0.045	8.24	Human, Chimp, Gor, Orang, Rhesus
TRIM52*	11.31774	4.16531	0.0034864	27.97	Human, Chimp, Bon, Gor, Orang, Rhesus, Cae, Tala, PFL, Sla, FLM, Colobus
TRIM54	0.000098	1	0.999951001	0	Human, Chimp, Orang, WCG, Rhesus, Marm, MLemur
TRIM55	0.000040	1	0.99998	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm, Tar, MLemur
TRIM56	0.000034	1	0.999983	0	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM57	4.409038	15.14907	0.110303569	0.473	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM58	13.327242	1.6149	0.001276516	10.381	Human, Chimp, Orang, WCG, Rhesus, Marm, Tar, Mlemur
TRIM60	17.097654	3.15749	0.000193772	20.828	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM61	0.000016	1	0.999992	0	Human, Chimp, Orang, WCG, Rhesus, Marm
TRIM62	0.001392	1	0.999304242	0	Human, WCG, Rhesus, Marm
TRIM63	1.696708	2.29094	0.428119036	3.641	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM65	0.442102	1	0.801675794	7.574	Human, Chimp, Gor, Orang, Rhesus, MLemur
TRIM66	3.600342	1.77533	0.165270625	26.916	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM67	0.230288	1.72453	0.891237796	0.536	Human, Chimp, WCG, Rhesus, Marm
TRIM68	0.000000	1	1	0	Human, Chimp, Gor, WCG, Rhesus, Marm, BBaby
TRIM69	3.857840	2.6469	0.145305043	19.785	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM71	0.000060	2.17763	0.99997	0	Human, Chimp, Gor, Orang, WCG, Rhesus
TRIM72	0.000028	1	0.999986	0	Human, Chimp, Gor, Orang, WCG, Rhesus, BBaby
TRIM75	0.800806	2.31885	0.670049961	7.432	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm
TRIM76	41.193456	9.01996	1.13489E-09	1.776	Human, Chimp, Gor, Orang, WCG, Rhesus, Marm

*PAML was ran using a combination of orthologs retrieved from online databases and sequencing.

Highlighted *TRIM* genes were found to be under positive selection. Positive selection was based on TRIM genes fulfilling the following criteria: (I) *TRIM* genes met the statistical criteria of PAML (Yang 2007) and (II) were reported by Datamonkey (Delport, Poon et al. 2010) to have at least one site of positive selection that overlapped with sites reported by PAML.

Abbreviations for species used were as follows:
Human (Homo sapiens), Chimp (Pan troglodytes), Bon, (Pan Paniscus), Orang (Pongo abelii), Gor (Gorilla gorilla), WCG (Nomascus leucogenys), WHG (Hylobates lar), Siamang (Symphalangus syndactylus), CAe (Chlorocebus aethiops), CTa (Chlorocebus tantalus), Baboon (Papio anubis), Rhesus (Macaca mulatta), Patas (Erythrocebus patas), Tala (Miopithecus talapoin), SLa (Trachypithecus cristatus), PFL (Trachypithecus vetulus), FLM (Trachypithecus francoisi) Douc (Pygathrix nemaeus), Colobus (Colobus guereza), Woolly (Lagothrix lagotricha), Spider (Ateles geoffroyi), Howler (Alouatta sara), Saki (Pithecia pithecia pithecia), Pygmy (Callithrix pygmaea), Marm (Callithrix jacchus), Squirrel (Saimiri sciureus sciureus), CTTam (Saguinus oedipus), Dusky (Callicebus moloch), RBTam (Saguinus labiatus), BGTiti (Callicebus donacophilus donacophilus), Tar (Tarsius syrichta), MLemur (Microcebus murinus), BBaby (Otolemur garnettii)

Table B.2 Human TRIM52 SNPs

Exon	dbSNP rs# cluster ID	Function	Codon	Validation
Exon1	rs144718973	missense	1	No listing
Exon1	rs149302292	missense	19	3
Exon1	rs138637336	synonymous	35	No listing
Exon1	rs146030758	missense	67	No listing
Exon1	rs182640735	missense	71	1
Exon1	rs191265268	nonsense	83	1
Exon1	rs142657741	missense	100	3
Exon1	rs56956877	frame shift	125	2
Exon1	rs80005177	synonymous	125	2, 3
Exon1	rs80196452	frame shift	125	No listing
Exon1	rs71707263	frame shift	125	3
Exon1	rs3073543	frame shift	128	2
Exon1	rs78075294	synonymous	128	No listing
Exon1	rs33972170	frame shift	128	1, 2, 3
Exon1	rs186360757	missense	134	1
Exon1	rs150292982	missense	136	No listing
Exon1	rs140222786	missense	139	No listing
Exon1	180687568	synonymous	159	No listing
Exon1	rs148225750	missense	168	No listing
Exon1	rs143060535	synonymous	171	No listing
Exon1	rs148982091	missense	187	No listing
Exon1	rs918388	synonymous	201	1, 2, 3, 4, 5

Exon1	rs144966268	missense	223	No listing
Exon1	rs142626341	missense	231	No listing
Exon1	180687760	missense	268	No listing
Exon2	rs149989700	missense	285	No listing
Exon2	rs139236429	missense	290	No listing

Validation code: (1) SNP has been sequenced in 1000Genome project; (2) Validated by multiple, independent submissions to the refSNP cluster; (3) Validated by frequency or genotype data: minor alleles observed in at least two chromosomes; (4) Genotyped by HapMap project; (5) All alleles have been observed in at least two chromosomes apiece.

Appendix C

Supplemental Information for Chapter 4

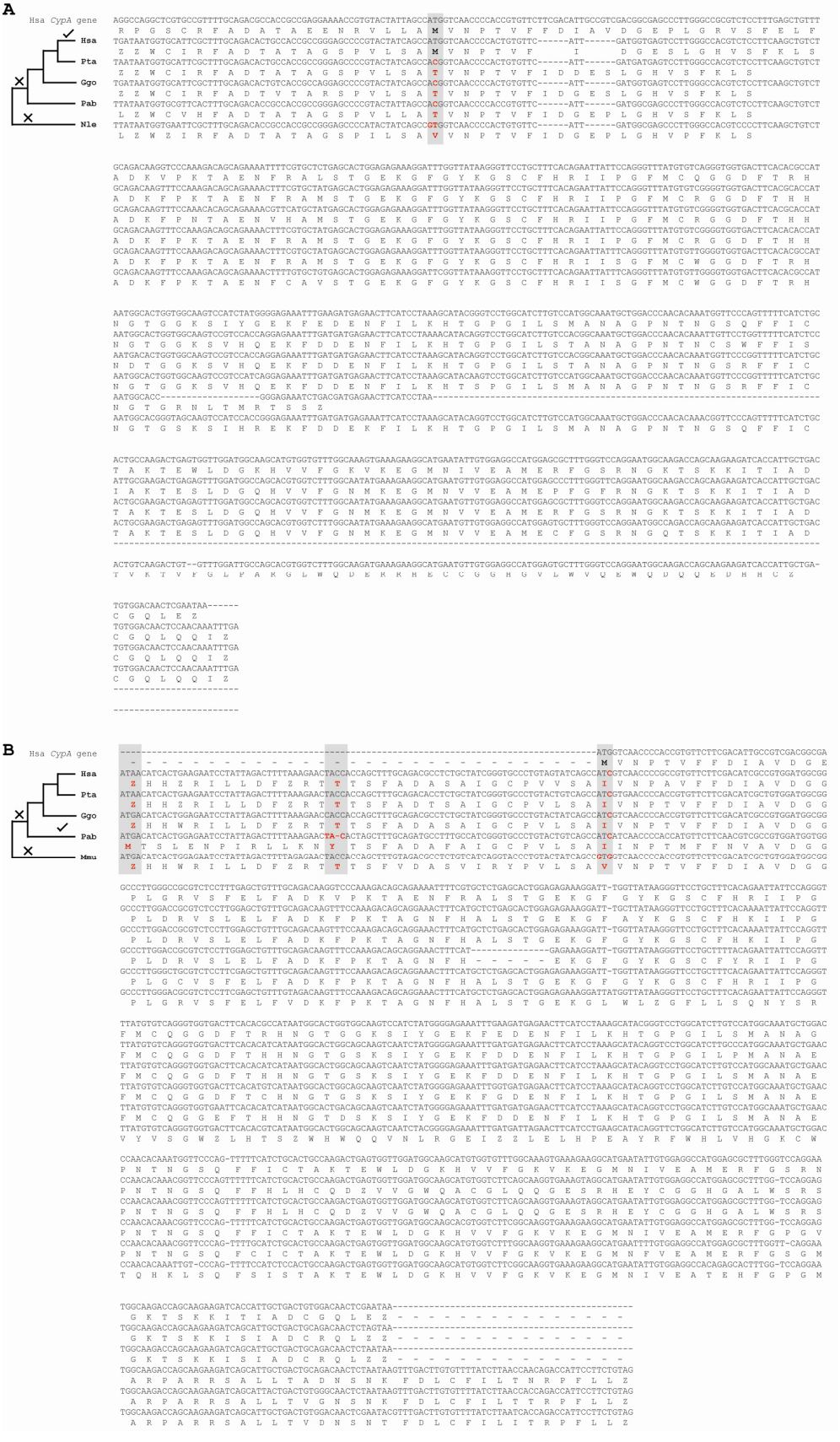

A
Hsa CypA gene
Hsa
Pta
Ggo
Pab
Nle
B
Hsa CypA gene
Hsa
Pta
Ggo
Pab
Mmu

Figure C.1 Alignment of "restored" *CypA* retrogenes. (A) Hsa1_2264934 and (B) Pab9_55196143 were highlighted amongst the sets of "Single *intact* orthologs" as they appeared to derive from a pseudogenized common ancestor. Thus, the ORFs of these *CypA* retrogenes have "restored" the pseudogenizing mutation to a state that is putatively functional.

Table C.1 "*Intact* ortholog" sets labels

Set1	Hsa1_45453909	Ptr1_45336245	Pseudo	Pab1_184930647	Pseudo	Pseudo
Set2	Hsa2_11491753	Ptr2A_11462495	Ggo2a_11610990	Pab2a_101318816	-	Pseudo
Set3	Hsa2_174350593	Ptr2B_177939266	Ggo2b_61367556	-	Mmu12_37107402	Cja6_44991432
Set4	Hsa3_138362721	Ptr3_142161614	Ggo3_138842232	Pab3_141133561	Pseudo	-
Set5	Hsa3_60675965	Ptr3_61610912	Ggo3_62242963	Pseudo	Pseudo	Cja15_36843754
Set6	Hsa5_65867896	Ptr5_48717140	Ggo17_30298988	Pab5_68226297	Pseudo	-
Set7	Hsa5_81305415	Ptr5_33414852	Ggo5_64620917	-	-	-
Set8	Hsa6_31487264	Pseudo	Ggo6_32399063	Pab6_32016739	Pseudo	Pseudo
Set9	Hsa7_28318833	Ptr7_26864040	Pseudo	Pab7_56121476	Mmu3_97943252	Pseudo
Set10	Hsa10_15196795	Ptr10_15593867	-	-	Mmu9_15454144	-
Set11	Hsa11_43488071	Ptr11_43560752	Ggo11_44202248	Pab11_25302073	Pseudo	Pseudo
Set12	Hsa12_98984498	Ptr12_98863945	Pseudo	Pseudo	Pseudo	-
Set13	Hsa13_107516174	Ptr13_107220868	Ggo13_90050086	Pseudo	Mmu17_87054968	-
Set14	Hsa19_30412089	Ptr19_34999596	-	Pab19_30367088	-	-
Set15	Hsa20_41859400	Pseudo	Ggo20_40939463	Pab20_40583506	Mmu10_21205240	Pseudo
Set16	Hsa21_20230097	Ptr21_5306253	Ggo21_7189307	-	-	-
Set17	Pseudo	Ptr5_64924840	Pseudo	Pab5_48800056	Pseudo	-
Set18	Pseudo	Ptr11_98500320	Ggo11_98132935	Pab11_96937733	Pseudo	-
Set19	Pseudo	Ptr11_GL391837_random_3692787	Ggo11_54437615	Pab11_19162975	Mmu14_15694170	Pseudo
Set20	Pseudo	Ptr20_34402675	Ggo20_35014063	Pab20_34835680	Pseudo	Pseudo
Set21	Pseudo	Ptr13_52470131	Ggo13_35112399	Pab13_53122108	Pseudo	Pseudo

Set22	Pseudo	Pseudo	Ggo11_3835817	Pab11_66848521	Mmu14_69667675	-
Set23	Pseudo	Pseudo	Pseudo	Pab7_7264592	Mmu3_45643683	-
Set24	-	-	-	-	Mmu12_33483752	Cja6_48736048

Table C.2 "Single *intact* ortholog" labels

< 20 Myo	32 Myo	43 Myo
Hsa1_22649345 Pab19_52530082 Pab14_69413187	Pab9_55196143 Mmu6_81234515	Hsa21_22200443 Pab16_57523436 Mmu13_25110885 Mmu16_15232421 Cja9_20990773

Table C.3 "Lineage specific" *CypA* retrogene labels

Human	Gorilla	Orangutan	Rhesus macaque	Marmoset
Hsa1_148201973	Ggo10_128627072	Pab1_102216616	Mmu4_7334034	Cja12_4471780
Hsa1_148806133		Pab5_147860419	Mmu2_4955405	Cja1_186916923
Hsa1_148644129			Mmu17_76211448	Cja1_14806534
Hsa1_143767360			Mmu3_162124806	Cja7_8584730
Hsa1_147954856			Mmu20_73474826	Cja14_22627679
Hsa1_149553109			Mmu7_128121419	Cja20_10565575
Hsa1_144363683			MmuX_23188508	CjaX_125394295
			Mmu7_135296871	Cja18_27696897
			Mmu9_127957348	Cja10_79918262
			Mmu14_71153664	Cja10_71670801
			Mmu2_129674877	Cja1_95801495
			Mmu16_53888818	Cja10_91638992
			Mmu10_56523662	Cja16_15599986
			Mmu6_40742170	Cja2_61874166
			Mmu18_32399855	Cja6_108234931
			Mmu6_62939241	Cja15_61262442
				CjaX_2186992
				Cja3_146993719
				Cja13_31386253
				Cja13_36431095

VITA

Raymond Malfavon-Borja was born in Escondido, California. He began his journey for higher education at Palomar Community College that continued at the California State University of San Marcos (CSUSM), where he earned a Bachelor of Science degree. While at CSUSM, Ray received formal lab training in the lab of Dr. Denise Garcia. Ray received additional training from the 2006 Stanford Summer Research Program in the lab of Dr. Dmitri Petrov. He joined the University of Washington, Genome Sciences graduate program in 2007 and completed his dissertation in the laboratories of Dr. Harmit S. Malik and co-advisor Michael Emerman.