UNIL | Université de Lausanne
Faculté de biologie et de médecine

Institut de Microbiologie (IMUL)

Analysis of susceptibility to Human Immunodeficiency Virus 1 (HIV-1) by comparative genetics

THESE DE DOCTORAT

Présentée à la Faculté de biologie et de médecine de l'Université de Lausanne

Par

Millán ORTIZ SERRANO

Diplômé en Biologie
Universidad Autonoma de Madrid

Jury
Prof. Thierry Pedrazzini, Président
Prof. Amalio Telenti, Directeur de thèse
Prof. Alicia Sanchez Mazas, Experte
Prof. Henrik Kaessmann, Expert

LAUSANNE
Mars 2009

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président	Monsieur Prof.	Thierry **Pedrazzini**
Directeur de thèse	Monsieur Prof.	Amalio **Telenti**
Experts	Monsieur Prof.	Henirk **Kaessmann**
	Madame Prof.	Alicia **Sanchez Mazas**

le Conseil de Faculté autorise l'impression de la thèse de

Monsieur Millan Ortiz Serrano

Biologiste diplômé de l'Université de Madrid

intitulée

Analysis of susceptibility to Human Immunodeficiency Virus 1 (HIV-1) by comparative genetics

Lausanne, le 6 mars 2009

pour Le Doyen
de la Faculté de Biologie et de Médecin

Prof. Thierry Pedrazzini

Table of Contents

“Nothing in biology makes sense except in the light of evolution”

Theodosius Dobzhansky

Summary

From the beginning of the 20th century the world population has been confronted with the human immune deficiency virus 1 (HIV-1). This virus has the particularity to mutate fast, and could thus evade and adapt to the human host. Our closest evolutionary related organisms, the non-human primates, are less susceptible to HIV-1. In a broader sense, primates are differentially susceptible to various retrovirus. Species specificity may be due to genetic differences among primates. In the present study we applied evolutionary and comparative genetic techniques to characterize the evolutionary pattern of host cellular determinants of HIV-1 pathogenesis. The study of the evolution of genes coding for proteins participating to the restriction or pathogenesis of HIV-1 may help understanding the genetic basis of modern human susceptibility to infection.

To perform comparative genetics analysis, we constituted a collection of primate DNA and RNA to allow generation of de novo sequence of gene orthologs. More recently, release to the public domain of two new primate complete genomes (bornean orang-utan and common marmoset) in addition of the three previously available genomes (human, chimpanzee and Rhesus monkey) help scaling up the evolutionary and comparative genome analysis. Sequence analysis used phylogenetic and statistical methods for detecting molecular adaptation.

We identified different selective pressures acting on host proteins involved in HIV-1 pathogenesis. Proteins with HIV-1 restriction properties in non-human primates were under strong positive selection, in particular in regions of interaction with viral proteins. These regions carried key residues for the antiviral activity. Proteins of the innate immunity presented an evolutionary pattern of conservation (purifying

selection) but with signals of relaxed constrain if we compared them to the average profile of purifying selection of the primate genomes. Large scale analysis resulted in patterns of evolutionary pressures according to molecular function, biological process and cellular distribution.

The data generated by various analyses served to guide the ancestral reconstruction of TRIM5α a potent antiviral host factor. The resurrected TRIM5α from the common ancestor of Old world monkeys was effective against HIV-1 and the recent resurrected hominoid variants were more effective against other retrovirus. Thus, as the result of trade-offs in the ability to restrict different retrovirus, human might have been exposed to HIV-1 at a time when TRIM5α lacked the appropriate specific restriction activity.

The application of evolutionary and comparative genetic tools should be considered for the systematical assessment of host proteins relevant in viral pathogenesis, and to guide biological and functional studies.

Résumé

La population mondiale est confrontée depuis le début du vingtième siècle au virus de l'immunodéficience humaine 1 (VIH-1). Ce virus a un taux de mutation particulièrement élevé, il peut donc s'évader et s'adapter très efficacement à son hôte. Les organismes évolutivement le plus proches de l'homme les primates non-humains sont moins susceptibles au VIH-1. De façon générale, les primates répondent différemment aux rétrovirus. Cette spécificité entre espèces doit résider dans les différences génétiques entre primates. Dans cette étude nous avons appliqué des techniques d'évolution et de génétique comparative pour caractériser le modèle évolutif des déterminants cellulaires impliqués dans la pathogenèse du VIH-1. L'étude de l'évolution des gènes, codant pour des protéines impliquées dans la restriction ou la pathogenèse du VIH-1, aidera à la compréhension des bases génétiques ayant récemment rendu l'homme susceptible.

Pour les analyses de génétique comparative, nous avons constitué une collection d'ADN et d'ARN de primates dans le but d'obtenir des nouvelles séquences de gènes orthologues. Récemment deux nouveaux génomes complets ont été publiés (l'orang-outan du Bornéo et Marmoset commun) en plus des trois génomes déjà disponibles (humain, chimpanzé, macaque rhésus). Ceci a permis d'améliorer considérablement étendue de l'analyse. Pour détecter l'adaptation moléculaire nous avons analysé les séquences à l'aide de méthodes phylogénétiques et statistiques.

Nous avons identifié différentes pressions de sélection agissant sur les protéines impliquées dans la pathogenèse du VIH-1. Des protéines avec des propriétés de restriction du VIH-1 dans les primates non-humains présentent un taux particulièrement haut de remplacement d'acides aminés (sélection positive). En

particulier dans les régions d'interaction avec les protéines virales. Ces régions incluent des acides aminés clé pour l'activité de restriction. Les protéines appartenant à l'immunité inné présentent un modèle d'évolution de conservation (sélection purifiante) mais avec des traces de "relaxation" comparé au profil général de sélection purifiante du génome des primates. Une analyse à grande échelle a permis de classifier les modèles de pression évolutive selon leur fonction moléculaire, processus biologique et distribution cellulaire.

Les données générées par les différentes analyses ont permis la reconstruction ancestrale de TRIM5α, un puissant facteur antiretroviral. Le TRIM5α ressuscité, correspondant à l'ancêtre commun entre les grands singes et les groupe des catarrhiniens, est efficace contre le VIH-1 moderne. Les TRIM5α ressuscités plus récents, correspondant aux ancêtres des grands singes, sont plus efficaces contre d'autres rétrovirus. Ainsi, trouver un compromis dans la capacité de restreindre différents rétrovirus, l'homme aurait été exposé au VIH-1 à une période où TRIM5α manquait d'activité de restriction spécifique contre celui-ci.

L'application de techniques d'évolution et de génétique comparative devraient être considérées pour l'évaluation systématique de protéines impliquées dans la pathogenèse virale, ainsi que pour guider des études biologiques et fonctionnelles.

Abbreviations

A	Adenine
AIDS	Acquired immunodeficiency syndrome
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
C	Cytosine
CA	Capsid
CD4	Cluster of differentiation antigen 4
DC	Dendritic cell
DNA	Deoxyribonucleic acid
EMBOSS	European Molecular Biology Open Software Suite
Env	Envelope
ER	Endoplasmic reticulum
G	Guanine
Gag	Group-specific antigen
HFV	Human foamy virus
HIV-1	Human Immunodeficiency Virus 1
HIV-2	Human Immunodeficiency Virus 2
ICAM	Intercellular adhesion molecule
IN	Integrase
LPS	Lipopolysaccharides
LTR	Likelihood ratio test
MHC II	Major histocompatibility complex class II
ML	Maximum likelihood
MLV	Murine leukaemia virus
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
nef	Negative effector
NF-kB	Nuclear factor-kB
OMK	Owl monkey
PAML	Phylogenetic Analysis by Maximum Likelihood
PAMP	Pathogen-associated molecular pattern
PCR	Polymerase Chain Reaction
PIC	Preintegration complexes
PRR	Pattern recognition receptor
Rbx1	Ring-box 1

Rev	Regulator of viral gene expression
RNA	Ribonucleic acide
SFV	Simian foamy virus
siRNA	Small interfering RNA
SIV	Simian immunodeficiency virus
SRV	Simian type D retrovirus
STLV	Simian T-cell lymphotropic virus
T	Thymine
Tat	Transactivator of the transcription
TLR	Toll like receptor
UTR	Untranslated Region
Vif	Viral infectivity factor
Vpr	Viral protein R
Vpu	Viral protein U
VSV	Vesicular stomatitis virus

Acknowledgements

I would not have realised this work without the help of many people. First I would like to thank my scientific mentor and thesis director, Prof. Amalio Telenti, for having given me the opportunity to work on a such a very exciting subject. I really appreciate his knowledge and the enthusiasm he has shown during the time of my research project.

It is also a pleasure to thank the many people who made this thesis possible. The people who supervised me as PhD student, Dr. Gabriela Bleiber and Dr. Valérie Goldschmidt.

I express my special appreciation to Prof. Henrik Kaessmann, for his fruitful collaboration and guidance in the field of molecular evolution.

I also thank Prof. Thierry Pedrazzini and Prof. Alicia Sanchez Mazas, who kindly accepted to be part of my jury. They gave interesting and helpful inputs to my work.

Many thanks to my colleagues and friends of the Institute of Microbiology for their constant help and the good moments we have shared.

I owe my deepest gratitude to my parents Fernando and Maria del Carmen for their love, support and encouragement during my whole life and especially during this work. You have always been there for me and believed in my projects.

My gratitude goes out also to my brother Alvaro and grandmother Matilde for their encouragements sent from the distant Madrid

Chapter 1. General introduction

1.1 Introduction

Human and non human primates share genomes with high degree of similarity. However, conspicuous differences exist in how these species respond to pathogens, including retroviruses. The study of the evolution of genes coding for primate proteins participating to the restriction or pathogenesis of retroviruses may help understanding the genetic basis of human susceptibility to infection. Furthermore, this information may be used as a tool to elucidate aspects of Virus-Host co-evolution.

1.2 Virus-Host Co-evolution, the Red queen effect

Viruses are dependent of the host cellular machinery for replication. During this cohabitation, the virus may improve and adapt to the host environment, and at the same time the host may enhance its defences against the invading pathogen. Co-evolution is the result of mutual influence between virus and host. This is commonly referred as the “Red queen effect” [1]. The term is taken from Lewis Carroll's “Through the Looking Glass” book [2]. The Red Queen says, "*It takes all the running you can do, to keep in the same place*". In scientific terms, this means that in an evolutionary system, continuing development is needed just to maintain its fitness relative to the system it is co-evolving with. At molecular level we observe changes in the genetic composition of the host in response to a genetic adaptation of the virus.

1.2.1 The host: primate classification

During millions of years primates have been natural host and reservoirs of a number of different retroviruses, including simian immunodeficiency virus (SIV), simian type D

retrovirus (SRV), simian T-cell lymphotropic virus (STLV), and simian foamy virus (SFV). SIV infections of more than twenty primates species have thus far been identified and confirmed by molecular analysis [3]. The close phylogenetic relationship between humans and non-human primates increases the potential for cross-species transmission of these retroviral agents. This is the case of Human immunodeficiency viruses types 1 and 2 (HIV-1, HIV-2) which originated from cross-species transmission of closely related viruses of chimpanzees and sooty mangabeys, respectively [3]. Thus, primates have undergone several episodes of retroviral infection, adaptation and cross-species transmission. This leads to footprints at the molecular level in the form of genetic variation among primate orthologous genes due to the differences on susceptibility to retrovirus infection and diseases. Primates provide a genetic evolution time window of more than 40 mys [4], sufficient to observe important genetic modifications. Analysis of a large number of primate sequences, relatively close in evolution time, allows a detailed analysis of genetic adaptation.

Primates can be classified in four main groups: great apes (including humans), gibbons, monkeys and prosimians (Figure 1). Informally, monkeys are divided in two groups, the Old World and New World monkeys. This denomination reflects their geographical distribution. Old World species are from Asia and Africa, whereas New World species are from Central and South America.

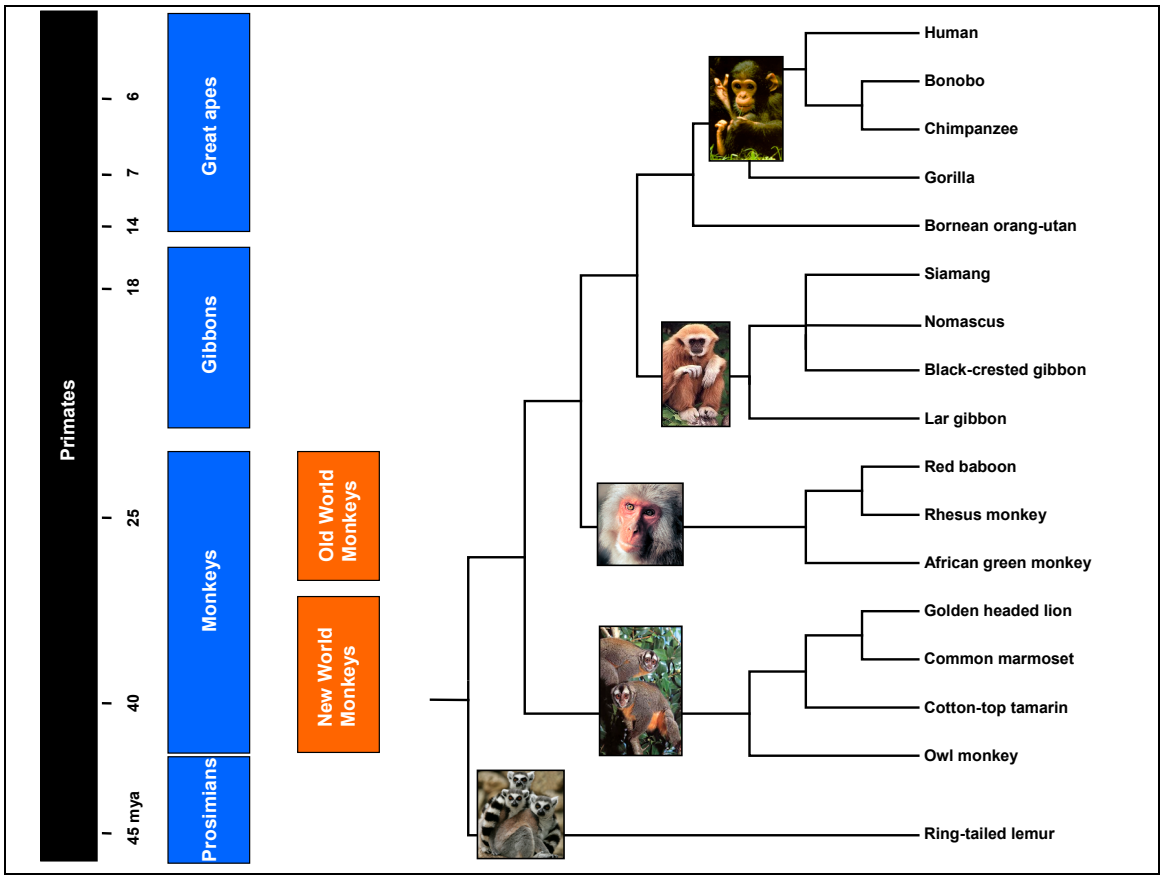


Figure 1: Primate classification. Phylogenetic tree of primates used in this work. Approximate divergence times in millions of years (mya) are shown.

1.2.2 The virus: HIV-1

The Human Immunodeficiency Virus 1 (HIV-1) [5] is a lentivirus belonging to the retrovirus family. It is an enveloped virus possessing a RNA genome that replicates via a DNA intermediate. The following steps occur when the retrovirus encounters the cell: first, the virus enters the cell by fusing with the cellular membrane, taking advantage of receptor and co-receptor host proteins, that otherwise play important roles in immunity and inflammation. Then, the viral genetic material is delivered into the cytoplasm in the form of a nucleoprotein core. The viral RNA genome is copied into DNA by the enzyme reverse transcriptase, transported to the cell nucleus, and integrated in the host genome via the activity of an integrase. The proviral HIV-1 DNA

is transcribed into viral mRNAs, which are processed and exported to the cytoplasm. Finally, viral products are transported to budding sites where virions are assembled together with viral RNA producing new virions.

Different retroviruses may have exerted selective pressures on host genes for millions of years. HIV-1 is a recent human pathogen, dating around 1931 [6], with a new report proposing a more recent date, between 1902 and 1921 [7]. The extensive genetic variation observed is this virus makes HIV-1 one of the fastest evolving of all organisms. The mutation rate is 3×10^{-5} par base per replication cycle in a genome of 10^4 base pairs [8]. It has a generation time of ~2.6 days and produces ~10^{11} new virions each day [9]. Moreover the frequent recombination and natural selection further elevate its rate of evolutionary change.

1.3 Evolutionary and comparative genetics

The majority of the work presented in this thesis is based in the field of **Evolutionary genetics.** This is the broad field of studies that resulted from the integration of genetics and evolutionary processes. The aim is to describe how the evolutionary forces influence genetic variation. To identify genetic variation we use **comparative genetics** (comparing the same gene in different primate species to define the selection pressures exerted on it).

1.4 Evolution: The nature of adaptation and selection

The two main evolutionary forces exerted on DNA sequences are **natural selection** and **neutral evolution**.

Natural selection is the process by which favorable traits that are heritable become more common in successive generations of a population of reproducing organisms, and unfavorable traits that are heritable become less common. Thus natural selection operates by different ways:

- **Positive or directional selection** also know as positive Darwinian selection, increases the frequency of a beneficial mutation. In evolutionary terms, there is fixation of amino acid replacements at protein level.

- **Purifying or stabilizing selection** also know as negative selection maintains a common trait in the population by decreasing the frequency of harmful mutations and weeding them out of the population. There is maintenance of conserved amino acids over long periods.

- **Balancing selection** maintains genetic polymorphisms (or multiple alleles) within a population. This is the situation in which natural selection within a population is able to maintain stable frequencies of two or more phenotypic forms.

In **neutral evolution**, mutations that do not affect functionality are called neutral substitutions and their accumulation is not affected by natural selection. A large number of evolutionary changes are the result of the fixation of neutral mutations that

do not have effects on the fitness of an organism. An independent process from natural selection that produces random changes in the frequency of traits in a population is the **Genetic drift**. This phenomenon results from the role that chance plays in determining whether a given trait will be passed on as individuals survive and reproduce. A **population bottleneck** is an evolutionary event in which a significant percentage of a population or species is killed or otherwise prevented from reproducing. Population bottlenecks increase genetic drift, as the rate of drift is inversely proportional to the population size.

In this work we focus on two forms of natural selection driving primate genome evolution. The degree of **positive** and **purifying selection** acting on primate genes could indicate their evolutionary pattern. Constrains imposed by protein folding and function result in domains under purifying selection. Regions under positive selection could reveal regions of protein-virus interaction due to the red queen effect. In this particular case these replacements could confer an advantage to the host by allowing escape from viral pressure.

1.5 Methods for detecting selection: the K_A/K_S ratio

To analyse the nature of adaptation and selection (positive, purifying or neutral) in different categories of host genes, Goldman and Yang developed a method to estimate the nonsynonymous and synonymous substitutions rates under realistic evolutionary models [10]. Comparison of orthologous genes sequences among primate species identifies nucleic acid changes in codons. Changes are synonymous (silent) when the encoded amino acid remains the same due to redundancy in the genetic code. A nonsynonymous change leads to amino acid replacement.

Two other major features of DNA sequence evolution are implemented in this method. First, the transition/transversion rate of the nucleic acid changes, where transition refers to a mutation changing a purine to another purine (A ↔ G) or a pyrimidine to another pyrimidine (T ↔C) and transversion refers to the substitution of a purine for a pyrimidine or vice versa. Secondly, it accounts for codon-using bias, because different organisms show particular preferences for one of the several codons that encode a given amino acid.

The above information allows estimation of the K_A/K_S ratio (also called dN/dS or ω), which is the number of non-synonymous substitutions per nonsynonymous site (K_A) divided by the number of synonymous substitutions per synonymous site (K_S). Simplistically, a value of K_A/K_S <1 reflects purifying selection (deleterious mutations are eliminated), a value of K_A/K_S =1 reflects neutral evolution (the protein is not under selection), and a value of K_A/K_S >1 reflects positive selection (there is fixation of amino acid replacements). This method provides an overview of coding sequence evolution, estimating the number of nonsynonymous (K_A) over synonymous (K_S) substitutions per site (averaged over the entire sequence) for each branch of the primates tree based on the accepted primate phylogeny [4].

The maximum likelihood (ML) method calculates K_A and K_S values among different sequences using explicit models of codon substitution [11] (Box 1). Parameters in the model (sequence divergence *t*, transition/transversion rate κ and the K_A/K_S ratio) are estimated from the data by ML, and are used to calculate K_A and K_S.

Box 1. A model of codon substitution

The codon is considered the unit of evolution. The substitution rate from codons i to j ($i \neq j$) is given as:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \boldsymbol{k}\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\boldsymbol{k}\pi_j, & \text{for nonsynonymous transition.} \end{cases}$$

Parameter $\boldsymbol{k}$ is the transition/transversion rate ratio, π_j is the equilibrium frequency of codon j and ω (= K_A/K_S) measures the selective pressure on the protein. Given the rate matrix $Q=\{q_{ij}\}$, the transition probability matrix over time t is calculated as:

$$P(t)= \{p_{ij}(t)\} = e^{Qt}$$

Where $p_{ij}(t)$ is the probability that codon i becomes codon j after t. Likelihood calculation on a phylogeny involves summing over all possible codons in extinct ancestors (internal nodes of the tree)

1.6 Methods for detecting amino acid sites under positive selection

Analysis of a gene averages the K_A/K_S ratio across all sites and positive selection is detected only if that average is >1. There is the potential for a loss of information because sites under strong purifying selection may bring the K_A/K_S average value below one, even if a limited number of sites are under strong positive selection. To detect such sites in a likelihood model, the standard approach is to use a statistical distribution to describe the variation of K_A/K_S among sites. This assumes several classes of sites in the protein with different K_A/K_S (>1 positive selection; 1< purifying selection; =1 neutral evolution). Testing of sites under positive selection involves two major steps. First, a likelihood-ratio test compares a **model** that does not allow for sites where K_A/K_S>1 with a more general model that does. Secondly, **empirical**

Bayes methods identify positively selected sites when they exist. Sites having a posterior probability (Post prob. >0.95) are estimated to be under positive selection.

The models are: The null model, M1a (neutral), assumes a class of conserved sites with $K_A/K_S=0$ (purifying selection) and another class of neutral sites with $K_A/K_S=1$. The alternative model, M2a (positive selection), adds a third class of sites with K_A/K_S estimated from the data. If M2a fits the data significantly better than M1a and the estimated K_A/K_S ratio for the third class in M2a is >1, then some sites are under positive selection. Estimating whether the model M2a fits the data significantly better than M1a uses the **likelihood ratio test**.

Empirical Bayes methods uses empirical data to evaluate the conditional probability distributions that arise from Bayes' theorem. This method estimates the probability of certain site, given the data at the site, to be under neutral, purifying or positive selection.

Likelihood ratio test (LTR) evaluates whether the increase in likelihood obtained by adding parameters (=degrees of freedom) to a model is significant. M1a and M2a are two nested models: M1a (*p* parameters) is a special instance of M2a (*p*+*n* parameters), and L_1 and L_2 the maximum likelihoods under M1a and M2a, respectively. Twice the log-likelihood ratio is asymptotically χ^2 (Chi square) distributed (n degrees of freedom) under M1a. The present work uses two (n=2) degrees of freedom as suggested by Yang *et al.* (2005) [12] for the M1a-M2a model.

The formula: $\mathbf{2\,[\ln(L_1) - \ln(L_2)] \sim \chi^2}$ (2 degrees of freedom)

1.7 Molecular adaptation, consequences of the red queen effect

While different retroviruses naturally infect primates, these hosts rarely develop an immunodeficiency syndrome (AIDS). AIDS is characterized by decreasing CD4+ T cells, increased levels of T-cell immune activation, and activation-induced cell death. Thus, the infected primate species represents a natural reservoir of SIVs. Primates may however experience severe infection after exposure to a virus whose natural host is another primate species. They differ in their handling of related retroviruses. These differences include resistance to infection (for example, differences in the susceptibility of primate cells to various retrovirus [13,14]), the satisfactory control of viral replication when infected (for example, in most experimentally infected chimpanzees, HIV-1 replicates poorly) [15], and the occurrence of infections that are characterized by high-level replication without the hallmarks of disease progression (for example, in sooty mangabey and African green monkeys) [16].Thus, there are different host defense mechanisms responsible for differences on susceptibility to retrovirus infections. These mechanisms are present in all primates but with genetic differences. Differences responsible for the various virus restriction capacities might be explained by primate lineage-specific pandemics that shape and redirect the different antiviral defense mechanisms. This is the consequence of the red queen effect, where the host adapts its antiviral defense mechanisms against a retroviral element which evolves at the same time to avoid host defense. Assuring the control of a specific retrovirus may increase susceptibility to a second retrovirus. The constant selective pressures exerted on host proteins implicated in viral control may lead to signatures of rapid evolution, as well as regions containing sites under positive selection due to protein-virus interaction.

1.8 Factors encounter by HIV-1 in the host cell

1.8.1 Description of the host proteins chosen for the detailed analysis

A number of proteins will be assessed in detail in this thesis. Two proteins have a specific anti-HIV restriction activity, TRIM5α and APOBEC3G. One protein PML was also reported to participate in HIV-1 restriction, and Cyclophilin A participates in HIV-1 pathogenesis. The DC-SIGN family of proteins include surface receptors that plays an important role in the recognition of a vast panel of pathogens, including HIV-1. Toll like receptors are a superfamily of receptors responsible of the recognition of molecular motifs from pathogens.

1.8.1.1 TRIM5α and PML

Tripartite motif-containing protein 5 alpha (TRIM5α) and Promyelocytic leukemia (PML) also called TRIM19 belong to the large TRIM protein family [17]. They contain a RING domain, one (TRIM5α) or two (TRIM19) B-boxes domains and a predicted coiled-coil region followed by a C-terminal domain. In TRIM5α, the C-terminal domain it is a B30.2 domain [18] whereas in TRIM19 it's a EXOIII domain (exonuclease domain in DNA-polymerase alpha and epsilon chain, ribonuclease T and other exonucleases).

TRIM5α it's a cytoplasmic restriction factor implicated in the early steps of retroviral replication. This protein was identified from a rhesus macaque library screened for simian factors restricting HIV-1 replication upon transfer into permissive human cells

[13]. Rhesus TRIM5α restricts HIV-1 but the human TRIM5α does not. The mechanism by which TRIM5α restricts HIV-1 is not yet well understood but it is mediated by the B30.2 domain through a direct interaction with the viral capsid [19-22]. A recent report indicates that TRIM5α is autoubiquitinylated within cells, and rapidly processed by the proteasome in a RING domain-dependent way [23]. If TRIM5α encounters incoming sensitive retroviral cores, the complex is recruited to the proteasome and destroyed before initiation of the reverse transcription (Figure 2.1).

PML is implicated in different cell functions, including apoptosis, transcriptional regulation, senescence, cell proliferation and signal transduction. PML localizes in nucleoplasm and in discrete subnuclear matrix associated compartments known as nuclear bodies, where it represents a main constituent. Recently PML has been also associated with antiviral activity in cooperation with type I Interferons. Overexpression of PML confers resistance to infection by vesicular stomatitis virus (VSV) and influenza A virus [24]. PML has also been reported to be active against retroviruses (HIV, murine leukaemia virus (MLV) and human foamy virus (HFV)). One proposed mechanism implicates transient cytoplasmatic export of PML and its subsequent recruitment by the incoming retroviral preintegration complexes (PICs), that inhibits HIV replication (Figure 2.2) [25]. The second proposed mechanism implicates the viral protein *Tat* (transactivator of the transcription) responsible of a correct viral mRNAs transcription of HFV; PML prevents its binding to the viral mRNA (Figure 2.3) [26]. Interestingly, some viruses for which the replicative cycle is inhibited by PML have developed strategies to alter, to various extents, the integrity and localization of nuclear bodies-associated PML. However, how PML interferes with so many viruses with widely varying replication strategies is still unclear.

1.8.1.2 APOBEC3G

The Apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like-3G (APOBEC3G) is an endogenous inhibitor of HIV-1 replication [27,28]. It belongs to the APOBEC superfamily proteins which plays an important role in innate anti-viral immunity. APOBEC3G antiretroviral activity is associated with the hypermutation of viral DNA through cytidine deamination. APOBEC3G is incorporated into HIV-1 particles in the producer cell, and during reverse transcription deaminates cytosine bases to uracil in the negative-sense DNA strand, resulting in G to A hypermutations in the complementary positive sense DNA strand. This hypermutation renders the viral cDNA vulnerable to degradation by nucleases. Those cDNA that manages to integrate into the host genome carries multiple mutations that likely result in aberrant viral products [29] (Figures 2.4, 2.5, 2.6). HIV-1 and other retroviruses possess the viral infectivity factor (Vif) protein [30] (Figure 2.7). To induce APOBEC3G degradation, Vif binds the cellular proteins Cul5, elonginB, elonginC, and Rbx1, to form a cullin5-based E3 ubiquitin ligase complex that leads to polyubiquitination and ultimately to proteasomal degradation of APOBEC3G [31]. The human APOBEC3G fails to restrict HIV-1 due to the degradation imposed by the HIV-1 Vif. In contrast, a number of primates APOBEC3G orthologs display activity against HIV-1 [14,32,33].

1.8.1.3 Cyclophilin A

The role of peptidylprolyl isomerase A (Cyclophilin A, CypA) in HIV replication has been investigated in detail after discovering in a yeast two-hybrid screen that the CypA binds the HIV-1 capsid [34]. CypA performs cis/trans isomerisation of proline

peptide bonds in sensitive proteins in the cell cytoplasm. CypA interacts with viral protein gag in infected cells leading to its recruitment into nascent HIV-1 virions [35] and also with incoming HIV-1 cores in newly infected cells (Figure 2). In humans, CypA promotes HIV-1 infectivity in target cells by a mechanism that does not require CypA incorporation into virions [36]. However the exact mechanism of action remains unclear. CypA appears to prevent the action of restriction factors by altering HIV-1 capsid conformation in a manner that makes it less sensitive to their inhibitory effect. In Old World monkey cells, CypA decreases HIV-1 infectivity, but only in the presence of TRIM5α [37]. In the owl monkey, a New World monkey, a CypA pseudogene has been inserted into the TRIM5α coding region, replacing the B30.2 domain with CypA and leading to a molecule called TRIMCyp [38,39]. This restriction factor strongly restricts HIV-1 by the recruitment of the incoming viral capsid to the TRIM5α domain facilitated by interaction between the CypA domain and the viral capsid [40]. Recently studies in three different old world monkeys (macaques) identified TRIMCyp chimeras that had arisen independently from that found in owl monkeys. None restricts HIV-1 [41],[42]. Thus, CypA might have a important role in the host immunity, that differs among different primate species [43].

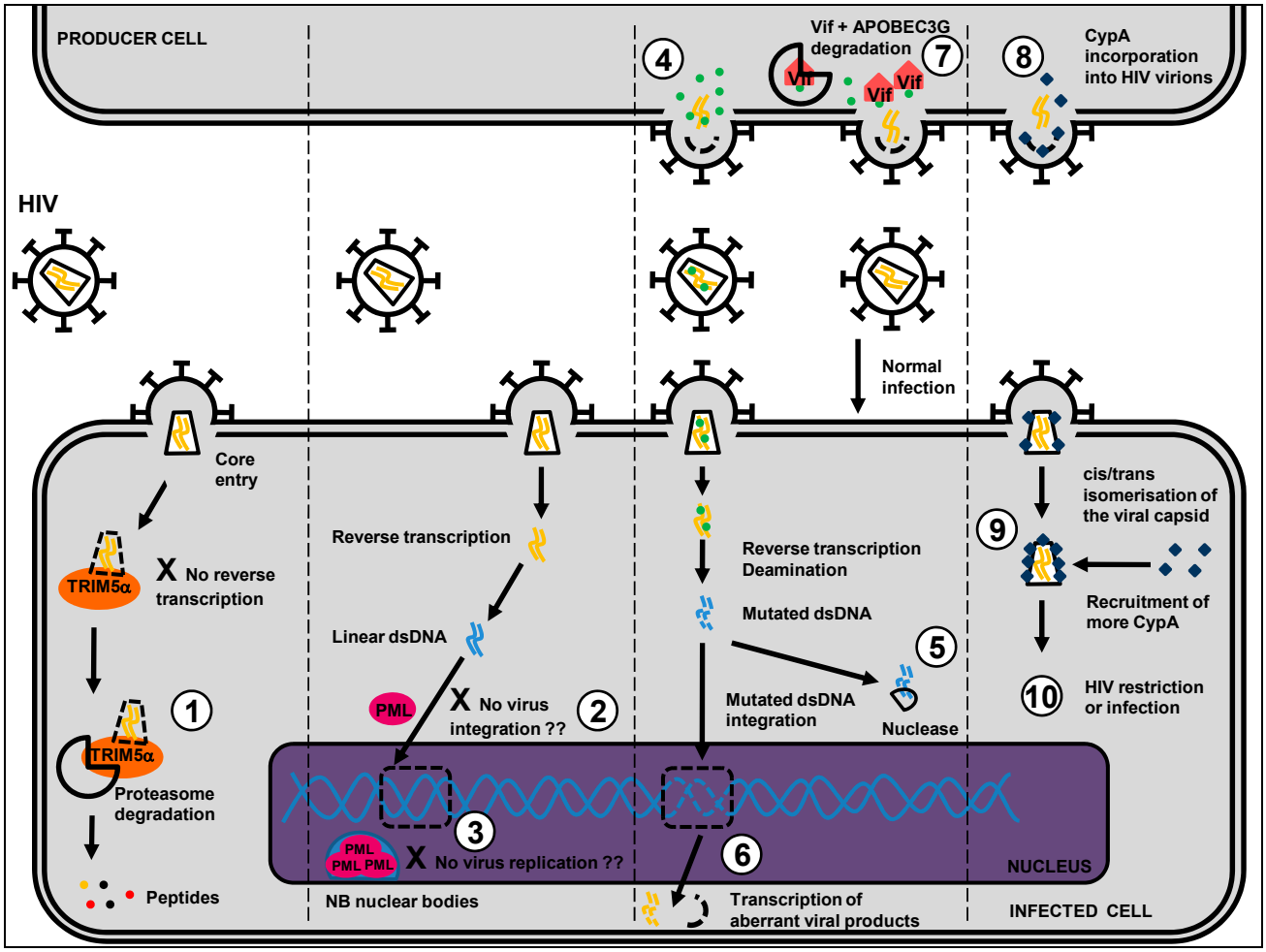


Figure 2. Host cell proteins related with HIV restriction or pathogenesis. (1) After HIV-1 infection and before virus uncoating TRIM5α binds the viral capsid, and then the complex TRIM5α–HIV-1 core is degraded by the proteasome. (2-3) PML viral restriction activity remains unclear. (4) APOBEC3G is incorporated into HIV-1 particles in the producer cell, (5) due to cytidine deaminase activity, the viral mutated DNA it's degradated by nucleases or (6) integrated in the host cell genome, resulting in aberrant viral products after transcription. (7) To escape APOBEC3G hypermutation activity, HIV-1 uses Vif to prevent its incorporation in the producer cell. (8) CypA interacts with gag in the producer cells leading to its recruitment into nascent HIV-1 virions. (9) More CypA is recruited by the viral core in the infected cell. (10) In human cells the interaction between HIV-1 and CypA is important for maximal infection. In Old World monkey cells CypA decreases HIV-1 infectivity.

1.8.1.4 DC-SIGN family

The DC-SIGN family of proteins include the dendritic-cell specific ICAM-grabbing non-integrin (DC-SIGN) and the related proteins liver/lymph node “L”-SIGN (L-SIGN also called DC-SIGNR), and CD209 antigen like protein 2 (CD209L2). These proteins are C-type lectin receptors with roles as cell adhesion receptors and in innate immunity as pathogen receptors [44,45]. DC-SIGN and L-SIGN recognize a vast range of bacteria, mycobacteria, parasites and viruses, including HIV (Figure 3. Panel 1). Dendritric Cells (DCs) are thought to be among the first cells infected by HIV-1 on the genital mucosa. Infected DCs migrate to lymph nodes where they transfer viruses to T cells. HIV-1 utilizes DCs as Trojan horses to spread the virus to the lymph nodes (Figure 3. Panel 2). DC-SIGN and L-SIGN function depends on a carbohydrate-recognition domain separated from a particular transmembrane region by a neck region made up of multiple 23 amino acid repeats. These proteins are expressed in a tissue-specific manner. DC-SIGN is expressed on phagocytic cells such as dendritic cells and macrophages, while L-SIGN expression is restricted to lymph node sinus endothelia and hepatic sinusoidal endothelium. The third homologue, CD209L2 is absent in humans but present in other primates. In the Rhesus macaque, CD209L2 has been identified in liver, spleen, lymph node, heart, and skin [46].

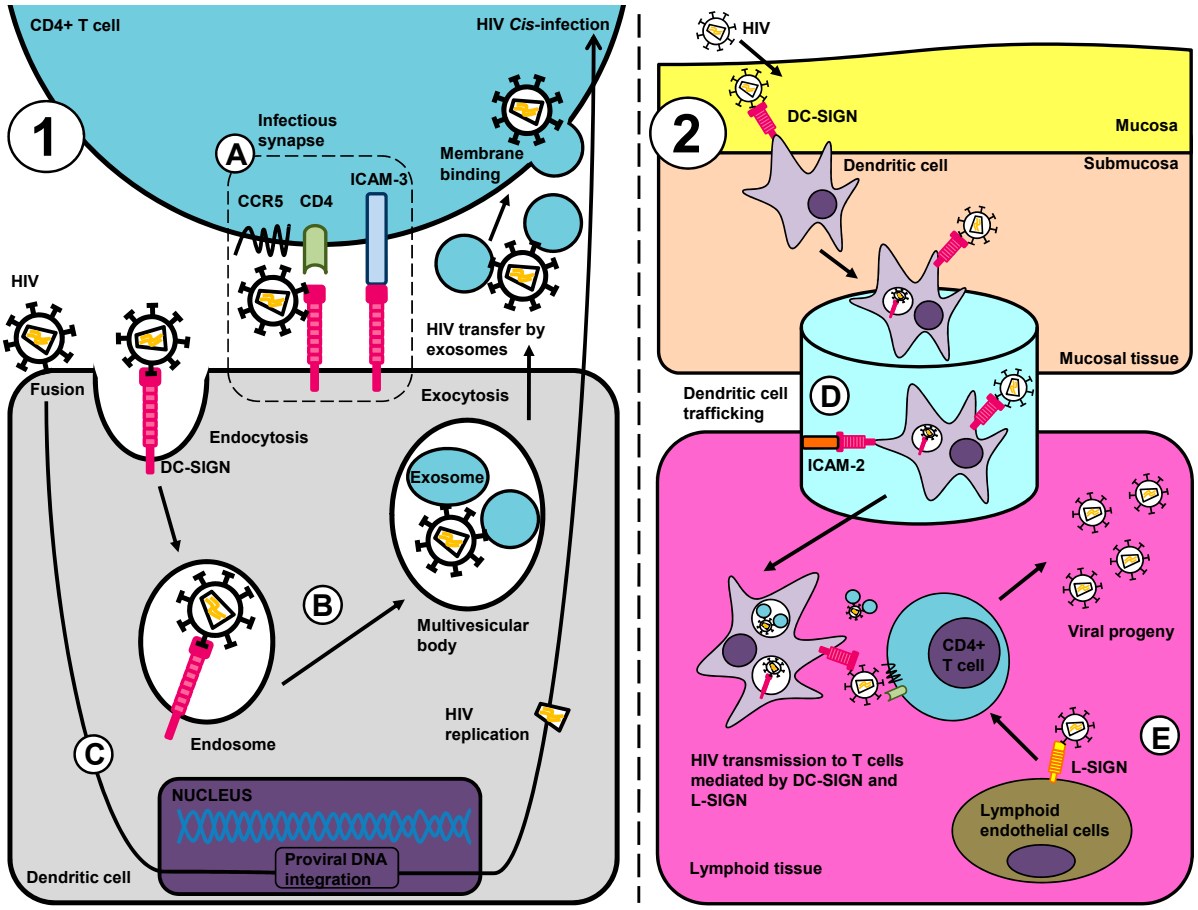


Figure 3. Mechanism of dendritic cell mediated HIV transmission and migration to lymphoid tissues. Panel 1. There are two types of the dendritic cell (DC)-mediated HIV-1 transmission to CD4+ cells in *trans*-infection. (A) First, by the infectious synapse, where the DC-specific intercellular adhesion molecule 3 (ICAM-3) participates in the formation of the infectious synapse. (B) Secondly, by HIV-1-associated exosomes. Endocytosed HIV-1 gain access to endosomal multivesicular bodies, enabling the release of HIV-1 associated with exosomes. Then, exosome-associated HIV-1 virions enter the CD4+ cells through membrane binding and fusion. (C) Direct HIV-1 infection of DCs by *cis*-infection. Panel 2. DCs capture HIV-1 that enter peripheral tissue as the mucosa, and migrate to lymphoid tissues facilitating HIV-1 particle transport, where the infection of T cells occurs in *trans*. (D) Interaction between DC-SIGN and ICAM-2 supports DCs tethering and rolling in endothelium. At the same time the HIV-1-infected DCs migrates to the lymphoid tissue. (E) L-SIGN is able to capture HIV-1 and bind ICAM-3 for in-*trans* infection of T cells with similar efficiency as DC-SIGN. Thus L-SIGN enhance the HIV-1 presentation to CD4+ T cells in the lymphoid tissues.

1.8.1.5 Toll like receptors

Toll-like receptors (TLRs) are a protein superfamily participating in the innate immune response against pathogens [47]. They are pattern recognition receptors (PRRs) that recognize specific pathogen-associated molecular patterns (PAMPs, conserved molecular patterns of pathogen structures). TLRs are located at the cell surface (TLR 1, 2, 4, 5, 6 and 10) or intracellularly (TLR 3, 7, 8, 9). Signalling through TLRs results in a type-1 interferon response and/or the production of pro-inflammatory cytokines. The first piece of evidence suggesting a role for TLRs in HIV-1 pathogenesis dates back to 1990, when it was shown that bacterial LPS activates the viral LTR, a process later found to be mediated by TLR4 [48]. To date, at least five TLR members (TLR2, TLR4, TLR7, TLR8, and TLR9) have been implicated in induction of HIV-1 expression during opportunistic co-infections [49]. TLR2 and TLR9 have shown to induce HIV replication [50]. Recently, ssRNAs from HIV were shown to be recognized by TLR7 and 8 [51].

1.8.2 Description of the host proteins chosen for the large scale analysis

The recent publication of two additional primate genomes (bornean orang-utan and common marmoset) allows the larger scale analysis of evolutionary pressures in the primate lineages. In a large scale analysis, we studied the evolution pattern of the majority of the cellular host factors implicated in the viral cell cycle and in pathogenesis (137 genes). HIV-1 usurps the host cellular machinery at multiple steps to infect and complete a productive cycle. Host factors modulate the viral entry, the post-entry events and late steps of viral replication. An increasing number of genes

and proteins have been identified over the years, mostly through functional and cell biology studies [52,53], eg. the receptors CD4 and CCR5. New genes have now been identified through large scale siRNA screens as ”host dependency factors” [54,55]. Progress in understanding of innate immunity and intrinsic cellular defence against retroviruses have also generated a series of host factors for analysis [56]. Many of these have been evaluated for genetic polymosphism that could modulate the individual susceptibility to disease. Gene name, Molecular function, cellular location, possible interaction between the host factors and HIV-1 proteins are listed in the Supplementary table S1 in chapter 4.3, page 75.

Chapter 2. Hypothesis and aims

Hypothesis:

- Primates and retrovirus have co-evolved during millions of years, each exerting mutual selective pressures.
- Selective pressures may be identified through use of comparative genetics. Signatures of positive selection could reveal regions of protein-virus interaction, identify relevant functional amino acids in host proteins implicated in HIV-1 restriction or pathogenesis and in the various steps of the viral life cycle.

The aims of the present study are:

- Analyse and classify in detail the evolution pattern of fifteen proteins related to HIV-1 restriction or pathogenesis. For this, we need to:
    - Build a DNA-RNA-Cell collection that includes material from sixteen primate species representative of the four main primate groups. (Chapter 3.1)
    - Create a dataset of new primate sequences.
- Perform a large scale evolutionary analysis of the majority of known host proteins involved in the HIV-1 life cycle and pathogenesis. For this, we need to:

- Use the available completed genomes of human, chimpanzee, bornean orang-utan, Rhesus macaque and common marmoset to identify the orthologous proteins (Chapter 3.3)

- Classify the evolutionary pattern for each protein. (Chapter 4.1-2-3)

- Identify “patches” of amino acids or individual amino acids under positive selection. (Chapter 4.1-2-3)

- Localized which steps of the retroviral cell cycle have been under selective pressures due to ancient genetic conflicts with ancestral retrovirus. (Chapter 4.3)

- Use this information to direct functional analyses. (Chapter 4.4)

Chapter 3. Materials and methods

Materials and methods

3.1 Primate collection

DNA samples of bonobo (*Pan paniscus*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), bornean orang-utan (*Pongo pygmaeus*), nomascus (*Hylobates leucogenys*), siamang (*Hylobates syndactylus*) and red baboon (*Papio hamadryas*) were kindly provided by Dr. Henrik Kaessmann from the Center of Integrative Genomics and Dr. Robert Hammond from the Department of Ecology and evolution, UNIL Lausanne. In addition, RNA-DNA extraction was performed with RNeasy Mini Kit (Cat No.74104) and QIAamp DNA Blood Mini Kit (Cat No.51104) on cells from: lar gibbon (*Hylobates lar*) lymphoblast (MLA144 cell line); African green monkey (*Cercopithecus* [*chlorocebus*] *aethiops*) kidney fibroblast (COS-7 cell line); rhesus monkey (*Macaca mulatta*) skin fibroblast (AG07128 cell line); owl monkey (*Aotus trivirgatus*) kidney (OMK cell line); cotton-top tamarin (*Saguinus Oedipus*) lymphoblast (B95-8 cell line); common marmoset (*Callithrix jacchus jacchus*) cells from a piece of frozen liver kindly provided by Keith Mansfield and Kuei-Chin from the New England primate research center; golden headed lion (*Leontopithecus rosalia chrysomelas*) cells from peripheral blood kindly provided by Charles Buillard and Eugène Chabloz from the Zoo of Servion; ring-tailed lemur (*Lemur catta*) skin fibroblast (AG07099 cell line).

3.2 Molecular analysis

To obtain the coding region of the genes analysed herein, we amplified and sequenced the exons with primers designed for the flanking intron regions on

genomic DNA and the 3' and 5' UTR regions on cDNA, when available. The primers were designed based on the alignment of the three complete available genomes (human-chimpanzee-rhesus monkey) at this period, from the UCSC (University of California Santa Cruz) Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html). HotStarTaq Master Mix (QIAGEN) was used for PCR (Polymerase Chain Reaction) amplification of DNA fragments smaller than 1 kb, and PrimeSTART DNA polymerase (TAKARA) or PfuTurbo DNA Polymerase (Stratagene) for fragments larger than 1 kb to avoid errors in DNA amplification. PCR conditions of annealing, step cycling and extension were optimised for each experiment. For sequencing, BigDye Terminator V1.1 Cycle sequencing Kit (Applied Biosystems) was used.

3.3 Orthologous genes identification in non-human primates complete genomes

To obtain the orthologous sequences of genes in four non human primates, the last genome assembly (hg18) of human (*Homo sapiens*), (panTro2) of chimpanzee (*Pan troglodytes*), the available genome assembly of (ponAbe2, WUSTL Pongo_albelii-2.0.2) Sumatran orang-utan (*Pongo pygmaeus abeli*), the genome assembly (rheMac2) of rhesus monkey and the the available genome assembly (calJac1, WUSTL Callithrix_jacchus-2.0.2) of common marmoset (*Callithrix jacchus*) were downloaded from UCSC genome browser. Human complete genome was release in April 2003, chimpanzee in March 2005, rhesus monkey in January 2006. Sumatran orang-utan and common marmoset draft assemblies were release in March and April 2008 respectively. BlatSuite.34 version of BLAT (BLAST – Like Alignment Tool) [57] was used with the human coding region sequence as template to retrieve homology

sequences from each genome with parameters "-t dnax –q dnax" (i.e. translated DNA). BLAT of the querying sequences in the five primates genomes was performed chromosome by chromosome using positive strand chain, then the homologous sequence with the maximum match value was taken.

3.4 Evolutionary analysis

Sequences were aligned using the sequence analysis tool MUSCLE (MUltiple Sequence Comparison by Log-Expectation) [58]. Coding regions were aligned according to their corresponding amino acid sequences using the tranalign application of the EMBOSS package (European Molecular Biology Open Software Suite) [59].

3.4.1 Branch Models

To trace the evolutionary history of the genes, we analyzed their substitutional patterns in the framework of the accepted primate phylogeny [4] using several codon-based maximum likelihood procedures as implemented in the codeml tool of the Phylogenetic Analysis by Maximum Likelihood (PAML) program package [60]. To obtain an overview of the coding sequence evolution, we estimated the number of nonsynonymous (K_A) over synonymous (K_S) substitutions per site (averaged over the entire sequence) for each branch of the trees using the free-ratio model of codeml.

3.4.2 Site specific models

In a more detailed analysis, to identify regions containing sites under positive selection, we utilized models that allow for different K_A/K_S rates at different sites of the sequence, because adaptive evolution often occurs at a limited number of sites

[11]. We first compared a null model ("M1a", [12,61]), which assumes two site classes (sites under purifying selection and neutrally evolving sites), to an alternative model ("M2a", [12,61]), which adds a third site class that allows for sites with $K_A/K_S > 1$, using likelihood ratio tests [62].

For additional information about materials and methods, refer to the detailed "Materials and methods" in the original articles (Chapter 4. Results).

Chapter 4. Results

4.1 Original article

Patterns of evolution of host proteins involved in retroviral Pathogenesis

Millan Ortiz[1], Gabriela Bleiber[1], Raquel Martinez[1], Henrik Kaessmann*[2] and Amalio Telenti*[1]

Address: [1]Institute of Microbiology and University Hospital, University of Lausanne, Switzerland and [2]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

Retrovirology 2006 Feb 7;3:11[63].

Comments on the article:

This study analyzed patterns of coding sequence evolution of four genes with known (*TRIM5*α and *APOBEC3G*) or suspected (*PML*) role in virus restriction, or in viral pathogenesis (*PPIA*, encoding Cyclophilin A), in the same set of human and non-human primate species. Detailed analysis of the four model proteins confirmed the previously described pattern of (i) strong positive selection on *TRIM5*α and multiple regions of conflict, and identified (ii) strong positive selection on *APOBEC3G* with better defined regions of host gene-virus co-evolution than previously reported, (iii) strong purifying selection for *PML* with absence of residues under positive selection, and (iv) full conservation of *PPIA* among primates. This suggests that *PPIA* and *PML* are not direct effectors of antiviral response. Together, the results presented here further support that an evolutionary genomics approach may be very useful for systematically assessing functional roles of primate host proteins potentially relevant in viral pathogenesis.

Retrovirology

BioMed Central

Short report

Open Access

Patterns of evolution of host proteins involved in retroviral pathogenesis

Millan Ortiz[1], Gabriela Bleiber[1], Raquel Martinez[1], Henrik Kaessmann*[2] and Amalio Telenti*[1]

Address: [1]Institute of Microbiology and University Hospital, University of Lausanne, Switzerland and [2]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

Email: Millan Ortiz - millan.Ortiz-serrano@chuv.ch; Gabriela Bleiber - Gabriela.x.bleiber@gsk.com; Raquel Martinez - Raquel.martinez@chuv.ch; Henrik Kaessmann* - Henrik.Kaessmann@unil.ch; Amalio Telenti* - amalio.telenti@chuv.ch

* Corresponding authors

Published: 07 February 2006

*Retrovirology*2006, 3:11 doi:10.1186/1742-4690-3-11

Received: 23 December 2005

Accepted: 07 February 2006

This article is available from: http://www.retrovirology.com/content/3/1/11

Abstract

Background: Evolutionary analysis may serve as a useful approach to identify and characterize host defense and viral proteins involved in genetic conflicts. We analyzed patterns of coding sequence evolution of genes with known (*TRIM5α* and *APOBEC3G*) or suspected (*TRIM19/PML*) roles in virus restriction, or in viral pathogenesis (*PPIA*, encoding Cyclophilin A), in the same set of human and non-human primate species.

Results and conclusion: This analysis revealed previously unidentified clusters of positively selected sites in *APOBEC3G* and *TRIM5α* that may delineate new virus-interaction domains. In contrast, our evolutionary analyses suggest that *PPIA* is not under diversifying selection in primates, consistent with the interaction of Cyclophilin A being limited to the HIV-1M/SIVcpz lineage. The strong sequence conservation of the *TRIM19/PML* sequences among primates suggests that this gene does not play a role in antiretroviral defense.

Background

Evolutionary genomics approaches have been proposed as powerful tools to identify protein regions relevant for host-pathogen interactions [1]. Identifying signatures of genetic conflict can open the way to biological testing of hypotheses regarding the function of host proteins. In retrovirology, the utility of this approach was recently demonstrated in evolutionary analyses of the antiretroviral defense genes *TRIM5α*, encoding a retrovirus restriction factor targeting the viral capsid [2,3], and *APOBEC3G*, coding for a cytidine deaminase that hypermutates viral DNA in primates [4-6]. Both genes were shown to have been shaped by positive selection, which led to the rapid fixation of adaptive amino acid replacement substitutions. The two genes revealed two different patterns of positive selection: a localized region of rapid change in *TRIM5α* [3], and a pattern where positively selected residues are scattered throughout the sequence in *APOBEC3G* [5].

To assess the potential of an evolutionary approach to identify further primate genes/proteins involved in virus defense, we analyzed coding sequence evolution of two additional genes, *TRIM19* (*PML*) and *PPIA*, and reassessed the selective signatures of *TRIM5α* and *APOBEC3G* in a common set of primates, representing 40 million years of evolution [7]. TRIM19 (PML) was proposed to possess anti(retro)viral activity [8,9], while Cyclophilin A,

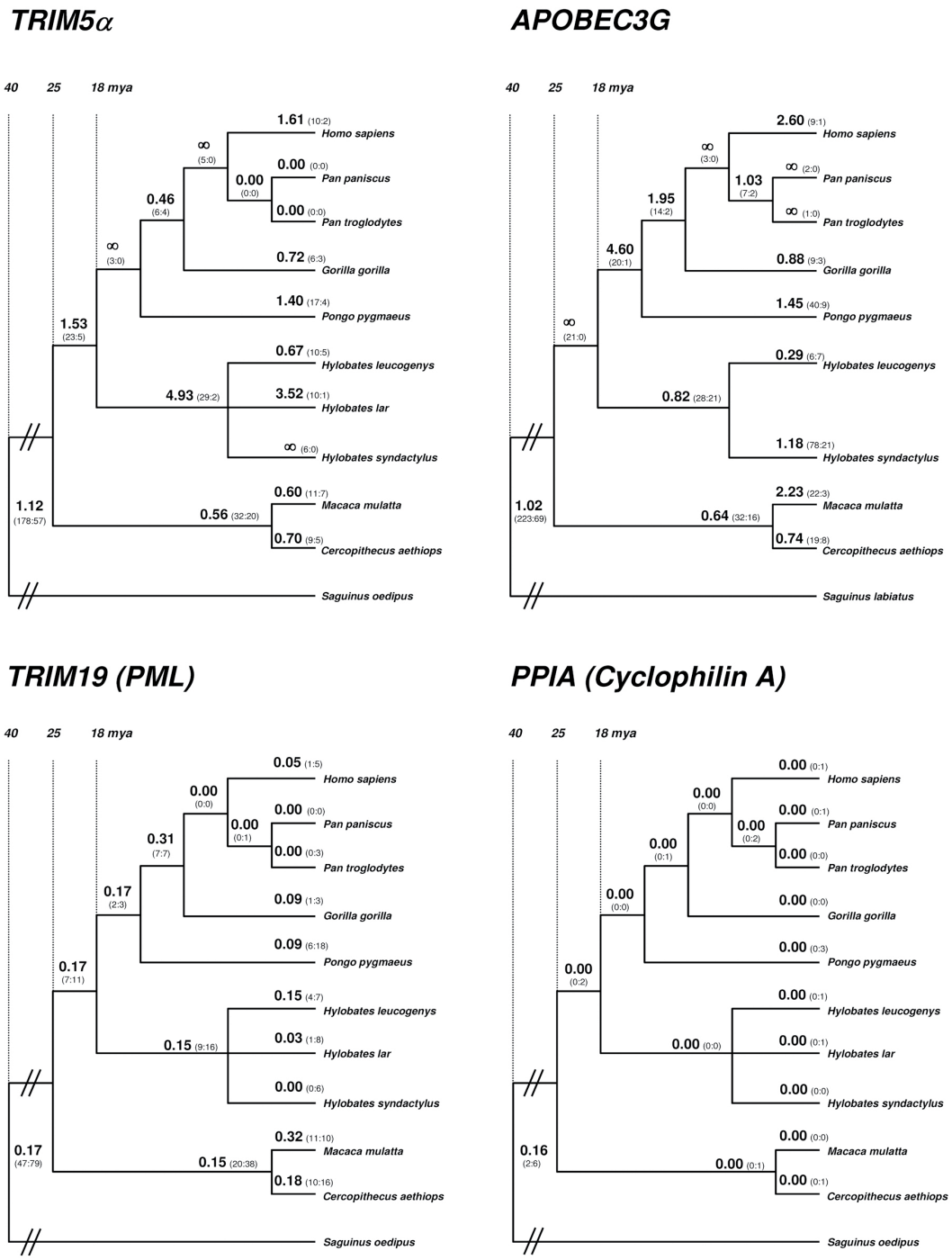


Figure 1

Phylogenetic trees of candidate antiviral defense genes. K_A/K_S values and the estimated number of nonsynonymous and synonymous substitutions (in parentheses) for each branch are indicated. Approximate divergence times in millions of years (mya) are shown [7].

Table 1: Codeml analyses using site-specific models.

			TRIM5α		
Site-specific Models[a]	ω_0[b]	ω_1[c]	ω_2[d]	LogL	Sites with $\omega > 1$ [e]
C: M1a	0.00 (34.91%)	1.00 (65.09%)		-4117.12	
D: M2a	0.00 (26.04%)	1.00 (61.67%)	6.37* (12.29%)	-4087.97	11 sites
			APOBEC3G		
Site-specific Models	ω_0	ω_1	ω_2	LogL	Sites with $\omega > 1$
C: M1a	0.03 (37.56%)	1.00 (62.44%)		-4187.55	
D: M2a	0.00 (28.28%)	1.00 (48.60%)	4.40* (23.11%)	-4148.85	24 sites
			TRIM19 (PML)		
Site-specific Models	ω_0	ω_1	ω_2	LogL	Sites with $\omega > 1$
C: M1a	0.09 (91.47%)	1.00 (8.53%)		-5215.40	
D: M2a	0.11 (97.25%)	1.00 (0.00%)	2.5 (2.75%)	-5214.46	n/a [f]
			PPIA (Cyclophilin A)		
Site-specific Models	ω_0	ω_1	ω_2	LogL	Sites with $\omega > 1$
C: M1a	0.05 (100%)	1.00 (0%)		-751.04	
D: M2a	0.05 (100%)	1.00 (0.00%)	1.00 (0.00%)	-751.04	n/a [f]

[a] the likelihood models used are described in the text
[b] class of sites under purifying selection
[c] class of sites evolving neutrally
[d] class of sites that may show $K_A/K_S > 1$
[e] sites pinpointed to be under positive selection by Bayes Empirical Bayes analysis
[f] test not applicable (M1a and M2a not significantly different)

encoded by *PPIA* (peptidyl-prolyl cis-trans isomerase), is incorporated into HIV-1 particles through an interaction with the viral capsid [10]. Cyclophilin A is incorporated only into viral particles of viruses of the HIV-1M/SIV_{CPZ} lineage, where it is required for viral replication [11].

To trace the evolutionary history of these genes, we first sequenced their coding regions from eleven primate species [see Additional files 1 and 2]. We then analyzed their substitutional patterns in the framework of the accepted primate phylogeny [7] using several codon-based maximum likelihood procedures as implemented in the codeml tool of the PAML program package [12] (Figure 1).

To obtain an overview of the coding sequence evolution, we estimated the number of nonsynonymous (KA) over synonymous (KS) substitutions per site (averaged over the entire sequence) for each branch of the trees using the free-ratio model of codeml [12]. Similarly to previous reports [3,5,6], this analysis revealed generally high KA/KS values on the different branches of the TRIM5α and APOBEC3G trees (average KA/KS ~1.1 for both genes), indicating that these genes show accelerated amino acid replacement rates due to the action of positive selection [13]. In contrast, PPIA and TRIM19 (PML) show low KA/KS values (0.05 and 0.15, respectively, when averaged over the entire tree), suggesting that their protein sequences have been strongly preserved by purifying selection (Figure 1).

In more detailed analyses, we then utilized models that allow for different K_A/K_S rates at different sites of the sequences, because adaptive evolution often occurs at a limited number of sites [14]. We first compared a null model ("M1a", [15,16]), which assumes two site classes (sites under purifying selection and neutrally evolving sites), to an alternative model ("M2a", [15,16]), which adds a third site class that allows for sites with $K_A/K_S > 1$, using likelihood ratio tests [17]. This comparison revealed that the alternative model provides a significantly better fit ($P < 10^{-30}$) for the *TRIM5α* and *APOBEC3G* genes than the null model, whereas the null model could not be rejected for *TRIM19* and *PPIA* (Table 1). The K_A/K_S for the additional site class is larger than 1 for both *TRIM5α* ($K_A/K_S \sim 6.4$) and *APOBEC3G* ($K_A/K_S \sim 4.4$), strongly suggesting adaptive protein evolution driven by positive selection at a subset of sites. Thus, this analysis supports the hypothesis that *TRIM5α* and *APOBEC3G* evolved under positive selection. Contrary to this, nearly all sites of *TRIM19* and *PPIA* (91.5% and 100%, respectively) are under purifying selection (Table 1).

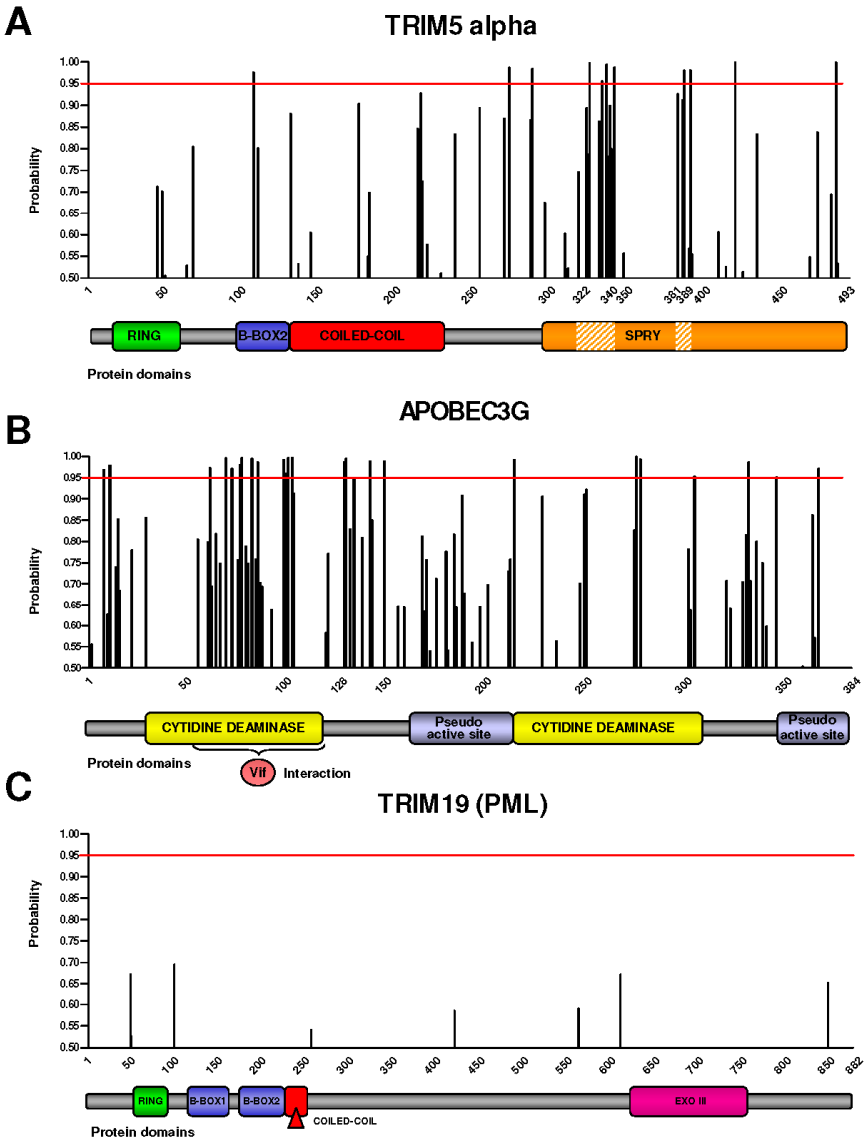


Figure 2
Codons under positive selection in *TRIM5*α and *APOBEC3G*. Y-axis: Probabilities of positively selected codons (see text). X-axis: amino acid numbering and functional domains. *TRIM19* is shown for comparison.

Using a recently developed Bayesian approach [16], we analyzed the site class under positive selection in *TRIM5α* and *APOBEC3G* in more detail. For *TRIM5α*, 11 of 493 (2%) codon sites can be predicted to be positively selected with high confidence ($P > 0.95$, Figure 2A). Two clusters of positive selection are found in the SPRY domain. The first cluster resides between amino acids 322 to 340 in the variable region 1 (v1, [18]), a region previously described as a "patch" of positive selection [3]. Replacement of the v1 region, or of specific amino acids within v1, modifies the restriction pattern of TRIM5α [19,20]. The second cluster, localized between amino acids 381 to 389, corresponds to the previously described variable region v2 of the SPRY domain [18]. Substitution of the human v2 region by a Rhesus monkey v2 exhibits no inhibitory activity against HIV-1 or a N-MLV$_{L117H}$ chimera [19,20]. However, the role of v2 in species-specific lentiviral restriction has not yet been extensively tested.

The analysis also predicts a large number (24 of 384, 6%) of positively selected sites in the *APOBEC3G* (Figure 2B) sequence. This result is consistent with previous reports by Sawyer et al. [5]. However, the inclusion of several new species from an additional hominoid lineage, Hylobatidae (gibbons and siamangs), points to the existence of a cluster of residues under positive selection between amino acids 62 and 103, the region that defines the Vif-interaction domain [21]. The protein Vif, which counteracts the activity of APOBEC3G, is encoded by nearly all lentiviruses [22]. Within the Vif-interaction domain of APOBEC3G, 10 residues can be pinpointed to have evolved under strong positive selection. Interestingly, the APOBEC3G amino acid position 128, which controls the ability of the HIV-1 Vif protein to bind and inactivate this host defense factor [23,24], is correctly identified as being positively selected ($P > 0.987$).

The parallel assessment of multiple genes in the same set of primates allows for several considerations and conclusions. First, by including additional primate lineages, we modify and complement previously observed patterns for two antiviral defense genes/proteins. For TRIM5α, our analysis confirms previous results by Sawyer et al [3], but underscores the potential interest of the second variable region of the SPRY domain that may be of functional relevance and merits further experimental analysis. With respect to APOBEC3G, our analysis extends previous reports that showed protein-wide distribution of positively selected residues. It suggests that this protein potentially carries a functionally relevant cluster of selected residues that coincides with the region of HIV-1-Vif interaction [23,24]. Positive selected sites by Bayes Empirical Bayes Inference with probabilities $P > 0.95$ for the two proteins are listed in Additional file 3.

Second, the failure to identify signatures of positive selection in the *TRIM19* (*PML*) gene suggests that its encoded protein does not have antiviral activity, or that the protein acts as an intermediary, lacking a physical protein-protein interaction with the pathogen. TRIM19 (PML) has been implicated in many functions, for example, in apoptosis and cell proliferation [9]. In addition, TRIM19 (PML) expression may act as an effector of the antiviral state induced by type I interferons [9]. Overexpression of TRIM19 (PML) is reported to confer resistance to infection by vesicular stomatitis virus and influenza A virus. Rabies, Lassa virus and lymphocytic choriomeningitis virus replicate to higher levels in PML-negative cells, whereas overexpression of the protein has no significant effect. Various roles have been proposed for TRIM19 (PML) in retroviral replication [8,25], although these findings remain controversial [26]. Many other viruses, including herpes simplex type 1 disturb the nuclear bodies that contain, among other proteins, TRIM19 (PML). However, it is unclear whether these effects are a consequence of the viral infection or a sign of its participation in antiviral defense. Thus, the effect of TRIM19 (PML) might be indirect. Failure to identify a signature of positive selection militates against a direct role of this protein in antiviral defense, because it would be expected that a prolonged contact with multiple pathogens over long evolutionary time periods would have resulted in signatures of positive selection indicative of a genetic conflict.

Finally, the absence of a signature of positive Darwinian selection in Cyclophilin A provides a complement to the understanding of the role of this protein in retroviral pathogenesis. Cyclophilin A interacts directly with the HIV-1 capsid, an interaction that may protect HIV-1 from antiviral restriction activity [27]. Although required by members of the HIV-1M/SIV$_{CPZ}$ lineage for replication, it is not needed by other primate immunodeficiency viruses [11]. Owl monkeys exhibit post-entry restriction of HIV-1 mediated by a TRIM5-Cyclophilin A fusion protein generated by retroposition [28]. Evolutionary analysis of *PPIA* indicates that Cyclophilin A has been preserved by strong purifying selection, leaving its protein sequence virtually unchanged. This is consistent with the interaction of Cyclophilin A and the viral capsid being limited to the HIV-1M/SIVcpz lineage.

Together, the results presented here further support that an evolutionary genomics approach may be very useful for systematically assessing functional roles of primate host proteins potentially relevant in viral pathogenesis [29]. Candidates for this approach may include other members of the TRIM or APOBEC families [30,31] as well as proteins involved in pathogen recognition and life cycle. Signatures of positive selection, but also the absence of signs of a genetic conflict, constitute relevant informa-

tion for understanding the nature of virus-host protein interactions.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

MO carried out the molecular genetic studies, performed sequence and phylogenetic analysis and contributed to drafting of the manuscript. GB and RM carried out molecular genetic studies. HK conceived the study, performed the evolutionary genomic analyses and drafted the manuscript. AT conceived the study, supervised the molecular genetic analysis, assured funding, and drafted the manuscript.

Additional material

Additional file 1

GenBank accession numbers.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4690-3-11-S1.doc]

Additional file 2

Primers for amplification and sequence analysis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4690-3-11-S2.doc]

Additional file 3

Positive selected sites by Bayes Empirical Bayes Inference with probabilities P > 0.95.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4690-3-11-S3.doc]

Acknowledgements

Supported by Swiss National Science Foundation grant no. 310000-110012/1 (to A.T.) and 3100A0-104181 (to H.K.), research awards of the Cloëtta and Leenaards Foundations (to A.T.), and a grant for interdisciplinary research from the Faculty of Biology and Medicine of the University of Lausanne (to A.T. and H.K.).

References

1. Yang Z: **The power of phylogenetic comparison in revealing protein function.** *Proc Natl Acad Sci U S A* 2005, **102:**3179-3180.
2. Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, Sodroski J: **The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys.** *Nature* 2004, **427:**848-853.
3. Sawyer SL, Wu LI, Emerman M, Malik HS: **Positive selection of primate TRIM5{alpha} identifies a critical species-specific retroviral restriction domain.** *Proc Natl Acad Sci U S A* 2005, **102:**2832-2837.
4. Sheehy AM, Gaddis NC, Choi JD, Malim MH: **Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein.** *Nature* 2002, **418:**646-650.
5. Sawyer SL, Emerman M, Malik HS: **Ancient Adaptive Evolution of the Primate Antiviral DNA-Editing Enzyme APOBEC3G.** *PLoS Biol* 2004, **2:**E275.
6. Zhang J, Webb DM: **Rapid evolution of primate antiviral enzyme APOBEC3G.** *Hum Mol Genet* 2004, **13:**1785-1791.
7. Goodman M: **The genomic record of Humankind's evolutionary roots.** *Am J Hum Genet* 1999, **64:**31-39.
8. Turelli P, Doucas V, Craig E, Mangeat B, Klages N, Evans R, Kalpana G, Trono D: **Cytoplasmic recruitment of INI1 and PML on incoming HIV preintegration complexes: interference with early steps of viral replication.** *Mol Cell* 2001, **7:**1245-1254.
9. Nisole S, Stoye JP, Saib A: **TRIM family proteins: retroviral restriction and antiviral defence.** *Nat Rev Microbiol* 2005, **3:**799-808.
10. Franke EK, Luban J: **Inhibition of HIV-1 replication by cyclosporine A or related compounds correlates with the ability to disrupt the Gag-cyclophilin A interaction.** *Virology* 1996, **222:**279-282.
11. Braaten D, Franke EK, Luban J: **Cyclophilin A is required for the replication of group M human immunodeficiency virus type 1 (HIV-1) and simian immunodeficiency virus SIV(CPZ)GAB but not group O HIV-1 or other primate immunodeficiency viruses.** *J Virol* 1996, **70:**4220-4227.
12. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555-556.
13. Li WH: *Molecular evolution* Sunderland MA, Sinauer Associates; 1997.
14. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation.** *Trends Ecol Evo* 2000, **15:**496-503.
15. Yang Z, Nielsen R, Goldman N, Pedersen AM: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155:**431-449.
16. Yang Z, Wong WS, Nielsen R: **Bayes empirical bayes inference of amino acid sites under positive selection.** *Mol Biol Evol* 2005, **22:**1107-1118.
17. Yang Z: **Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution.** *Mol Biol Evol* 1998, **15:**568-573.
18. Song B, Gold B, O'Huigin C, Javanbakht H, Li X, Stremlau M, Winkler C, Dean M, Sodroski J: **The B30.2(SPRY) domain of the retroviral restriction factor TRIM5alpha exhibits lineage-specific length and sequence variation in primates.** *J Virol* 2005, **79:**6111-6121.
19. Stremlau M, Perron M, Welikala S, Sodroski J: **Species-Specific Variation in the B30.2(SPRY) Domain of TRIM5{alpha} Determines the Potency of Human Immunodeficiency Virus Restriction.** *J Virol* 2005, **79:**3139-3145.
20. Yap MW, Nisole S, Stoye JP: **A Single Amino Acid Change in the SPRY Domain of Human Trim5alpha Leads to HIV-1 Restriction.** *Curr Biol* 2005, **15:**73-78.
21. Conticello SG, Harris RS, Neuberger MS: **The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G.** *Curr Biol* 2003, **13:**2009-2013.
22. Gaddis NC, Sheehy AM, Ahmad KM, Swanson CM, Bishop KN, Beer BE, Marx PA, Gao F, Bibollet-Ruche F, Hahn BH, Malim MH: **Further investigation of simian immunodeficiency virus Vif function in human cells.** *J Virol* 2004, **78:**12041-12046.
23. Schrofelbauer B, Chen D, Landau NR: **A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif).** *Proc Natl Acad Sci U S A* 2004, **101:**3927-3932.
24. Mangeat B, Turelli P, Liao S, Trono D: **A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action.** *J Biol Chem* 2004, **279:**14481-14483.
25. Regad T, Saib A, Lallemand-Breitenbach V, Pandolfi PP, de TH, Chelbi-Alix MK: **PML mediates the interferon-induced antiviral state against a complex retrovirus via its association with the viral transactivator.** *EMBO J* 2001, **20:**3495-3505.
26. Berthoux L, Towers GJ, Gurer C, Salomoni P, Pandolfi PP, Luban J: **As(2)O(3) enhances retroviral reverse transcription and counteracts Ref1 antiviral activity.** *J Virol* 2003, **77:**3167-3180.
27. Franke EK, Yuan HE, Luban J: **Specific incorporation of cyclophilin A into HIV-1 virions.** *Nature* 1994, **372:**359-362.
28. Sayah DM, Sokolskaja E, Berthoux L, Luban J: **Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1.** *Nature* 2004, **430:**569-573.
29. Telenti A: **Adaptation, co-evolution, and human susceptibility to HIV-1 infection.** *Infect Genet Evol* 2005, **5:**327-334.
30. Reymond A, Meroni G, Fantozzi A, Merla G, Cairo S, Luzi L, Riganelli D, Zanaria E, Messali S, Cainarca S, Guffanti A, Minucci S, Pelicci PG, Ballabio A: **The tripartite motif family identifies cell compartments.** *EMBO J* 2001, **20:**2140-2151.
31. Bogerd HP, Wiegand HL, Doehle BP, Lueders KK, Cullen BR: **APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells.** *Nucleic Acids Res* 2006, **34:**89-95.

4.2 Original article

The evolutionary history of the CD209 (DC-SIGN) family in human and non-human primates

Millan Ortiz[1], Henrik Kaessmann[2], Mary Carrington[3], Lluis Quintana-Murci[4], Amalio Telenti[1*]

[1]Institute of Microbiology, University of Lausanne, Lausanne, Switzerland, [2] Center of Integrative Genomics, University of Lausanne, Lausanne, Switzerland, [3]Laboratory of Molecular Immunology, US Army Medical Research Institute of Infectious Diseases, Frederick, MD, USA; [4]Cancer and Inflammation Program, Laboratory of Experimental Immunology, SAIC-Frederich Inc., NCI-Frederick, Frederick, MD, USA and [5] Institute Pasteur, Human Evolutionary Genetics, CNRS, URA3012, Paris, France

Genes & immunity 2008 Sep;9(6):483-92 [64]

Comments of the article:

CD209 family genes encode for C-lectin receptors that recognize a vast range of pathogens. It includes three members, DC-SIGN, L-SIGN and CD209L2. DC-SIGN and L-SIGN function depends on a carbohydrate-recognition domain separated from a transmembrane region by a neck region made up of several repeats. The importance of these receptors in pathogen recognition, as well as the importance of dendritic cells in pathogenesis of HIV-1 infection make these genes an excellent model system to investigate their distribution in primates in order to (i) thoroughly assess their evolutionary history (ii) identify amino acids under positive selection (iii) date the origin of gene duplication (iv) understand the evolutionary process giving raise to the neck repeat region. The phylogenetic tree supports a scenario where *CD209L* results from a duplication of *CD209*. Overall, *CD209* and *CD209L2* genes present a evolution pattern consistent with purifying selection. However, detailed analysis shows that L-SIGN presents a small subset of sites class under positive

selection, where we identifiedy three residues under positive selection. A second family of major innate immunity microbial sensors, namely the Toll-like receptor (TLR) gene family that identifies specific pathogen-associated molecular patterns (PAMPs) was investigated in parallel. The observed degree of purifiying selection of the Toll-like receptor family may be expected given the need to faithfully recognize various pathogens motifs and the inability of pathogens to modify the specific molecular patterns.

Genes and Immunity (2008) 9, 483–492

www.nature.com/gene

ORIGINAL ARTICLE

The evolutionary history of the CD209 (DC-SIGN) family in humans and non-human primates

M Ortiz[1], H Kaessmann[2], K Zhang[1], A Bashirova[3], M Carrington[4], L Quintana-Murci[5] and A Telenti[1]
[1]Institute of Microbiology, University of Lausanne, Lausanne, Switzerland; [2]Center of Integrative Genomics, University of Lausanne, Lausanne, Switzerland; [3]Laboratory of Molecular Immunology, US Army Medical Research Institute of Infectious Diseases, Frederick, MD, USA; [4]Cancer and Inflammation Program, Laboratory of Experimental Immunology, SAIC-Frederick Inc., NCI–Frederick, Frederick, MD, USA and [5]Institut Pasteur, Human Evolutionary Genetics, CNRS, URA3012, Paris, France

The CD209 gene family that encodes C-type lectins in primates includes CD209 (DC-SIGN), CD209L (L-SIGN) and CD209L2. Understanding the evolution of these genes can help understand the duplication events generating this family, the process leading to the repeated neck region and identify protein domains under selective pressure. We compiled sequences from 14 primates representing 40 million years of evolution and from three non-primate mammal species. Phylogenetic analyses used Bayesian inference, and nucleotide substitutional patterns were assessed by codon-based maximum likelihood. Analyses suggest that CD209 genes emerged from a first duplication event in the common ancestor of anthropoids, yielding CD209L2 and an ancestral CD209 gene, which, in turn, duplicated in the common Old World primate ancestor, giving rise to CD209L and CD209. K_A/K_S values averaged over the entire tree were 0.43 (CD209), 0.52 (CD209L) and 0.35 (CD209L2), consistent with overall signatures of purifying selection. We also assessed the Toll-like receptor (TLR) gene family, which shares with CD209 genes a common profile of evolutionary constraint. The general feature of purifying selection of CD209 genes, despite an apparent redundancy (gene absence and gene loss), may reflect the need to faithfully recognize a multiplicity of pathogen motifs, commensals and a number of self-antigens.
Genes and Immunity (2008) **9,** 483–492; doi:10.1038/gene.2008.40; published online 5 June 2008

Keywords: *C-type lectins; HIV; Ebola; Mycobacteria; innate immunity; DC-SIGN*

Introduction

The *CD209* family of genes codes for DC-SIGN (dendritic cell-specific ICAM-grabbing non-integrin) and related proteins L-SIGN (for liver/lymph node 'L'-SIGN, encoded by *CD209L*) and CD209L2. These homologous genes cluster on chromosome 19p13.3.

DC-SIGN is expressed on phagocytic cells such as dendritic cells and macrophages, whereas L-SIGN expression is restricted to lymph node sinus endothelia and hepatic sinusoidal endothelium. The third homologue, *CD209L2*, is absent in humans but present in other primates. In the *Rhesus macaque*, *CD209L2* is expressed in liver, spleen, lymph node, heart and skin.[1] One or more DC-SIGN-like proteins have been reported in mice, rats and dogs. These proteins are type-II C-type lectin receptors with roles as cell-adhesion receptors and in innate immunity as pathogen receptors.[2–4] DC-SIGN and L-SIGN recognize a vast range of bacteria, mycobacteria, viruses and protozoa (reviewed in Koppel *et al*.[2]). Their function depends on a carbohydrate-recognition domain separated from a transmembrane region by a neck region made up of several 23-amino acid repeats. DC-SIGN has affinity for mannose oligosaccharides and fucose-containing moieties whereas L-SIGN binds only to mannose oligosaccharides (reviewed in Koppel *et al*.[2]). The neck region plays a role in the orientation and flexibility of the carbohydrate-recognition domain. The type and number of the neck-region repeats are important in dimer formation and stabilization of protein tetramers.[5] Thus, neck-length variation could influence the pathogen-binding properties of these lectins. In previous analyses in primates,[1] *CD209* and *CD209L* have variable numbers of repeats. This contrasts with the single partial repeat that characterizes primate *CD209L2*, and that is a general feature of the neck region of other mammal *CD209*-like genes.

The importance of DC-SIGN family members in pathogen recognition as well as the importance of dendritic cells in pathogenesis of human immunodeficiency virus, Ebola and *Mycobacterium tuberculosis* infection, together with their particular genomic organization make these genes an excellent model system to investigate the mode and intensity of selective pressures that may act on pathogen defense genes.[6] As a category, immunity- and defense-related genes have experienced by far the most positive selection in humans and other organisms.[7]

The aim of the present study was to extend the analysis of the distribution of the three recognized *CD209*

Correspondence: Professor A Telenti, Institute of Microbiology, Centre Hospitalier Univeritaire Vaudois, Bugnon 48, CHUV, Lausanne 1011, Switzerland.
E-mail: amalio.telenti@chuv.ch
Received 2 April 2008; revised 24 April 2008; accepted 25 April 2008; published online 5 June 2008

family genes in primates to thoroughly assess the evolutionary history of this gene family, to identify possible domains and amino acids under selective pressure, to date the gene-duplication events resulting in this family and to understand the evolutionary process giving rise to the neck repeat region. To trace the evolutionary history of these genes, gene-coding sequences are determined for a representative number of primates (hominoids, Old World and New World monkeys), followed by the analysis of amino-acid substitutional patterns in the framework of the accepted primate phylogeny.[8] The analyses result in global estimates of the patterns of evolution (purifying, neutral, positive selective pressure) as well as shed light on episodes of adaptive evolution at specific sites.[6]

To help build a more contextual interpretation of the CD209 family data, we investigate in parallel the evolution of a second family of major innate immunity microbial sensors, namely the Toll-like receptor (TLR) gene family. TLRs are pattern-recognition receptors that identify specific pathogen-associated molecular patterns (PAMPs, conserved molecular patterns of pathogen structures). TLRs are located at the cell surface (TLR1, -2, -4, -5, -6 and -10) or intracellularly (TLR3, -7 and -8). Signalling through TLRs results in a type-1 interferon response and/or the production of pro-inflammatory cytokines.

Results

Homology and species distribution of CD209 *family members*
Phylogenetic analysis groups the various primate genes with high bootstrap values (Figure 1). Within a species, the percentage of identity at the nucleotide level (excluding the length-variable neck region) is 79.0 (range 78.3–79.6) for comparisons of *CD209* and *CD209L* sequences, 78.8 (range 74.4–80.2) for comparisons of *CD209* and *CD209L2*, and 72.1 (range 71.2–72.5) for comparisons of *CD209L* and *CD209L2*.

CD209L2 resembles the ancestral form, because the observed shorter neck region appears to be reminiscent of the shorter neck regions present in *CD209*-like sequences in other mammals. Analysis of the dog genome identifies a single *CD209* homologue. Similarly, we identified by BLAST analysis of cow sequences the presence of a single *CD209* homologue. The dog and cow *CD209*-like sequences are closely related to mouse *CD209g* and *Signr8* (Figure 1). The rat genome codes for the same set of paralogues as mice (not shown). We and others[1] have failed to amplify sequences of *CD209* genes in prosimians. Our study confirms the previous observation that orangutan has a truncated form of *CD209L* and that *CD209L*2 is a pseudogene in the gorilla.[1]

The phylogenetic tree (Figure 1) supports a scenario where the three extant *CD209* genes emerged from two duplication events; a first duplication event occurred in the common ancestor of anthropoids, yielding *CD209L2* and an ancestral *CD209* gene, which, in turn, duplicated in the common Old World primate ancestor, giving rise to *CD209L* and *CD209*.

Analysis of sequence evolution and selective pressures
Analysis of the substitutional pattern using several codon-based maximum likelihood procedures allowed the estimation of the number of non-synonymous (K_A) over synonymous (K_S) substitutions per site. Overall, all the three genes have values consistent with purifying selection (Figures 2a–c). The K_A/K_S values averaged over the entire tree were 0.43 (*CD209*), 0.52 (*CD209L*) and 0.35 (*CD209L2*). Due to the high variability in the number of neck-region repeats across the species and the difficulty of obtaining a reliable alignment, this region was not included in this analysis (a separate analysis of the neck region is presented below). Detailed analyses of the trees (Figures 2a–c) did, however, identify a number of branches with K_A/K_S values higher than 1.0, in particular in the gibbon lineage. There were no significant differences in the pattern of selective constrain among *CD209* gene family among primates having all three or only two functional homologues (Figure 2d).

In more detailed analyses, we utilized models that allow for different K_A/K_S rates at different sites of the sequences, because adaptive evolution often occurs at a limited number of sites. This comparison revealed that the null model (which includes sites under purifying selection and neutrally evolving) could not be rejected for *CD209* and *C209L2*. However, the alternative model that adds a third site class that allows for sites under positive selection provided a significantly better fit to the data with respect to *CD209L* with a *P*-value of 6.5E−05 (Table 1). The K_A/K_S for the additional site class is larger than one for L-SIGN ($K_A/K_S \sim 5$), suggesting adaptive protein evolution driven by positive selection at a small subset of sites (Table 1). Using a Bayesian approach,[9] we analysed the site class under positive selection in L-SIGN in more detail. Only one residue, alanine in position 88, located in the first neck repeat is pinpointed to be under positive selection (posterior probability, $P = 0.99$). In addition, threonine 319 and alanine 393, in the C-lectin domain, are identified with lower confidence ($P = 0.90$ and $P = 0.93$ respectively). These two L-SIGN residues map at the protein surface away from the region that contains residues involved in carbohydrate interactions (Figure 3). The binding of small carbohydrate compounds by L-SIGN takes place principally at a calcium coordination site in the carbohydrate-recognition domain. The amino-acid residues involved directly in coordinating the calcium ion,[10] Glu359, Asn361, Glu366 and Asp378, are fully conserved across all primates.

Analysis of the neck region
The neck region is characterized by a variable number of a conserved 23-amino acid repeats. In primates included in this work, *CD209* and *CD209L* code five to nine repeats, a number that varies according to the species. The number of repeats may vary also within a species (apes and Old World monkeys).[1] In contrast, the primate *CD209L2* neck region contains a single (partial) repeat element, a general feature of the neck region of other mammal *CD209*-like genes—an exception being the mouse *Cd209b* (four repeats) and *Cd209c* (two repeats).

To better understand the evolutionary process resulting in the extended and variable length of the neck regions of primate *CD209* and *CD209L* genes, we computed a phylogenetic tree considering all repeats in mammal *CD209* family genes (Figure 4a). The C-terminal partial repeat of primate *CD209* and *CD209L*, and the single partial repeat of *CD209L2* were found to be most closely related to those of *CD209*-like repeats in other

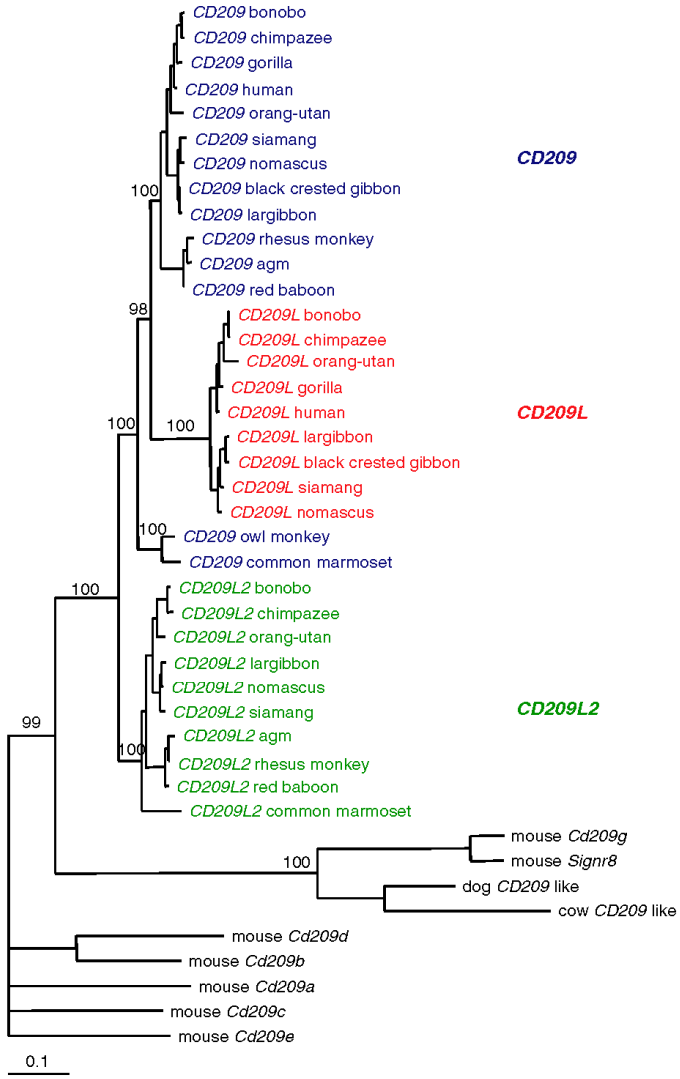


Figure 1 Phylogenetic tree of *CD209* family of genes in mammals. Bayesian estimation of the evolutionary relationships among coding sequences (excluding the neck repeat region) of *CD209* family. Bootstrap values are indicated.

mammals. The first (proximal) repeat of primate *CD209* and *CD209L* appears to result from a duplication of the last (distal) repeat. Although intermediate repeats have an unclear origin (that is the analysis does not differentiate whether they originated from the proximal or distal repeat), they are highly conserved within and across the various *CD209* and *CD209L* genes, the only exception being the penultimate repeat of *CD209L* (Figure 4b).

Comparison of evolutionary pattern of CD209 *and* TLR *gene families*

To better define the significance of the pattern of selection in the *CD209* gene family, we performed an evolutionary genetic analysis of a second family of innate immunity receptors. For this, we selected five primate species, representatives of apes and Old and New World monkeys. We used sequence of *TLR*1 (95.7% of coding sequence), *TLR*2 (97.0%), *TLR*3 (97.1%), *TLR*4 (85.2%),

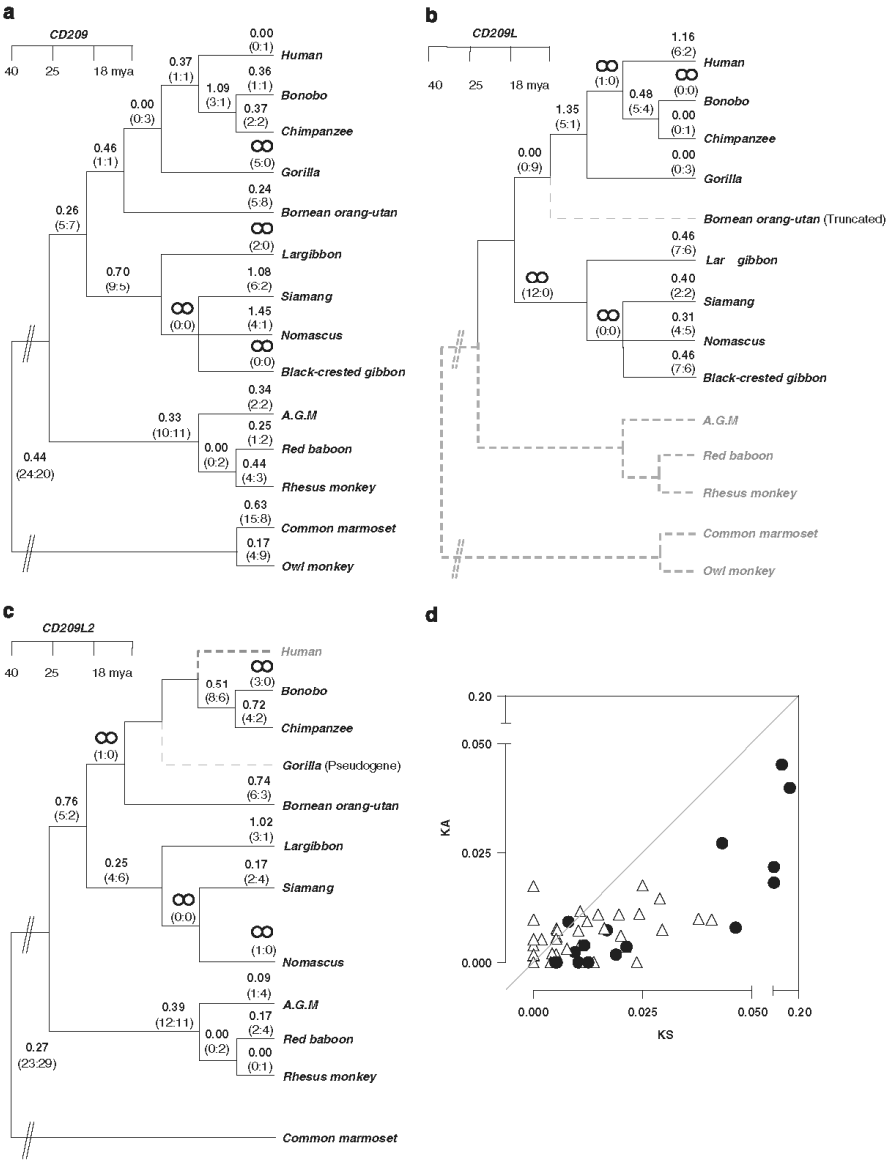


Figure 2 Analysis of sequence evolution and selective pressure acting on *CD209*, *CD209L* and *CD209L2*. (**a**–**c**) K_A/K_S values and the estimated number of non-synonymous and synonymous substitutions (in parentheses) for each branch are indicated. Approximate divergence time in Mya is shown. Species in which a gene is absent or disabled by truncation or pseudogenization are identified by discontinued grey branches. (**d**) Differences in the degree of purifying selection among primates carrying two or three functional members of the *CD209* family. K_A and K_S values obtained with PAML analysis are represented. Triangles represent values for primates carrying all three *CD209* family members (*CD209*, *CD209L* and *CD209L2*). Black circles represent values for primates carrying two *CD209* family members or having a copy disabled by truncation or pseudogenization. The upper left area represents positive selection; lower right area represents purifying selection. PAML, phylogenetic analysis by maximum likelihood.

Table 1 Codeml analyses using site-specific models

Site-specific models[a]	ω_0[b]	ω_1[c]	ω_2[d]	*Log L*	*Sites with* $\omega > 1$[e]
CD209					
M1a	0.00 (56.44%)	1.00 (43.55%)		−1999.16	
M2a	0.07 (68.38%)	1.00 (0.00%)	1.36 (31.61%)	−1998.71	N/A[f]
CD209L2					
M1a	0.23 (80.99%)	1.00 (19.00%)		−1825.16	
M2a	0.23 (80.99%)	1.00 (13.43%)	1.00 (5.57%)	−1825.16	N/A[f]
CD209L					
M1a	0.00 (68.62%)	1.00 (31.37%)		−1786.23	
M2a	0.00 (75.46%)	1.00 (14.69%)	5.28 (9.83%)	−1778.26	1 site

[a]The likelihood models used are described in the text.
[b]Class of sites under purifying selection.
[c]Class of sites evolving neutrally.
[d]Class of sites that may show $K_A/K_S > 1$ positive selection.
[e]Sites pinpointed to be under positive selection by Bayes Empirical Bayes analysis.
[f]Test not applicable (M1 and M2a not significantly different).

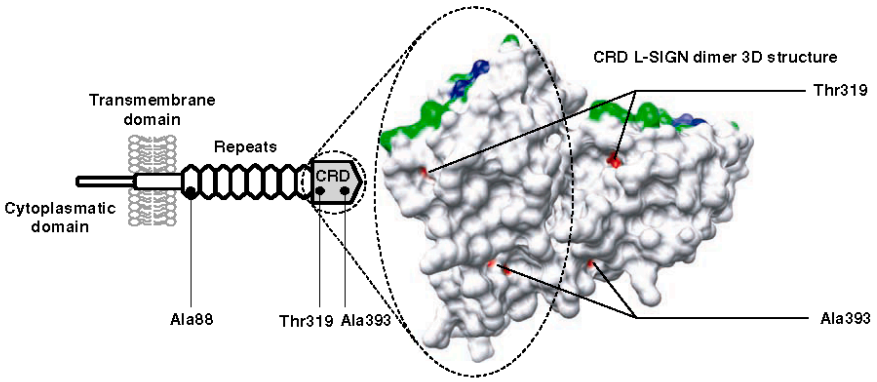


Figure 3 Structure of the carbohydrate-recognition domain of L-SIGN. Three residues were predicted to be under positive selection (in red): alanine 88 (in the neck repeat domain) and threonine 319 and alanine 393. The last two L-SIGN residues map at the protein surface away from recognized domains involved in carbohydrate interaction. Residues involved directly in coordinating the calcium ion are shown in blue, residues important for binding with carbohydrates are shown in green.

TLR5 (65.2%), *TLR6* (98.8%), *TLR8* (36.0%) and the complete sequence of *TLR7* and -10.

Overall, all nine *TLR* gene sequences had values consistent with purifying selection (Figure 5). The K_A/K_S values averaged over the entire tree were 0.50 (*TLR1*), 0.49 (*TLR2*), 0.23 (*TLR3*), 0.52 (*TLR4*), 0.59 (*TLR5*), 0.36 (*TLR6*), 0.34 (*TLR7*), 0.47 (*TLR8*) and 0.44 (*TLR10*).

In the set of more detailed analyses using models that allow different K_A/K_S rates at different sites of the sequences (because adaptive evolution often occurs at a limited number of sites), we found that the null model (which includes sites under purifying selection and neutrally evolving) could not be rejected for *TLR2*, *-3*, *-4*, *-5*, *-6*, *-7*, *-8* and *-10*. The alternative model (which adds a third site class that allows for sites under positive selection) provided a significantly better fit to the data with respect to *TLR1*, with a *P*-value of 2.46E 04. One per cent of TLR1 codons have a $K_A/K_S \sim 16.9$, suggesting adaptive protein evolution driven by positive selection. Using a Bayesian approach,[9] we analysed the site class under selective pressure in TLR1 in more detail. Two residues, tryptophan 61 located in the extracellular domain and serine 748 located in the Toll/interleukin-1 receptor domain, were predicted to be under positive selection (posterior probability, $P > 0.95$).

Discussion

The evolutionary history of the *CD209* gene family in primates is characterized by several episodes of gene duplication, recent gene deletion/truncation and elongation of the neck repeat region. Given the role of DC-SIGN and related proteins in pathogen uptake, it is plausible that these evolutionary changes have occurred in response to temporal selective pressures during primate evolution. However, the evolutionary analysis of various

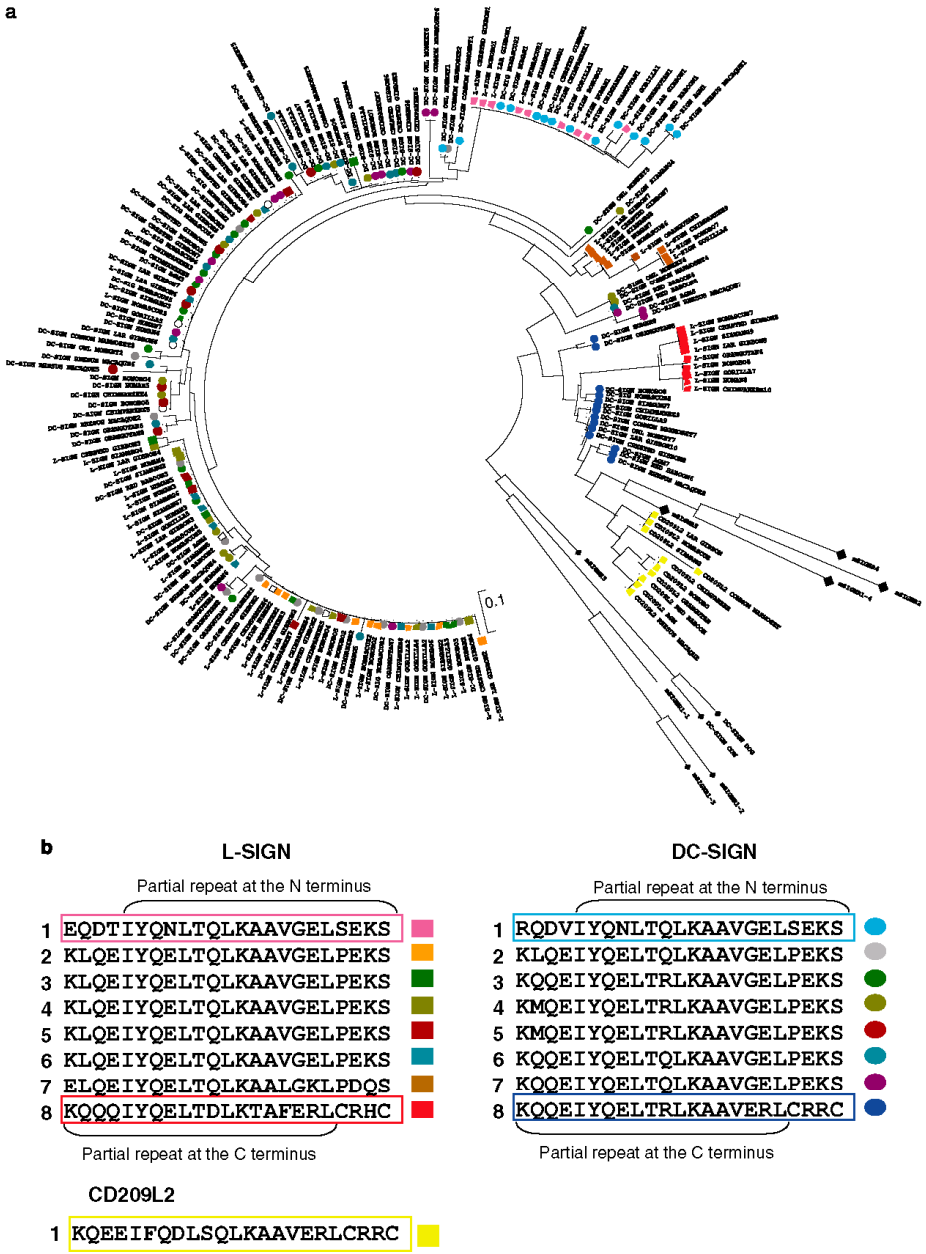


Figure 4 Analysis of the neck repeat region of DC-SIGN family in mammals. (**a**) Neighbor-joining tree of amino-acid sequences of the neck repeat region. Colour coding of circles and squares correspond to those described in (**b**). Black rhombus represents repeats of mouse SIGNR1–5, and dog and cow CD209-like. (**b**) Alignment of the neck region of DC-SIGN, L-SIGN and CD209L2 from bonobo (as representative of other primate repeat sequences). Repeats are numerated 1–8 (N- to C-terminal), colour coded and represented by circles for DC-SIGN and by squares for L-SIGN and CD209L2.

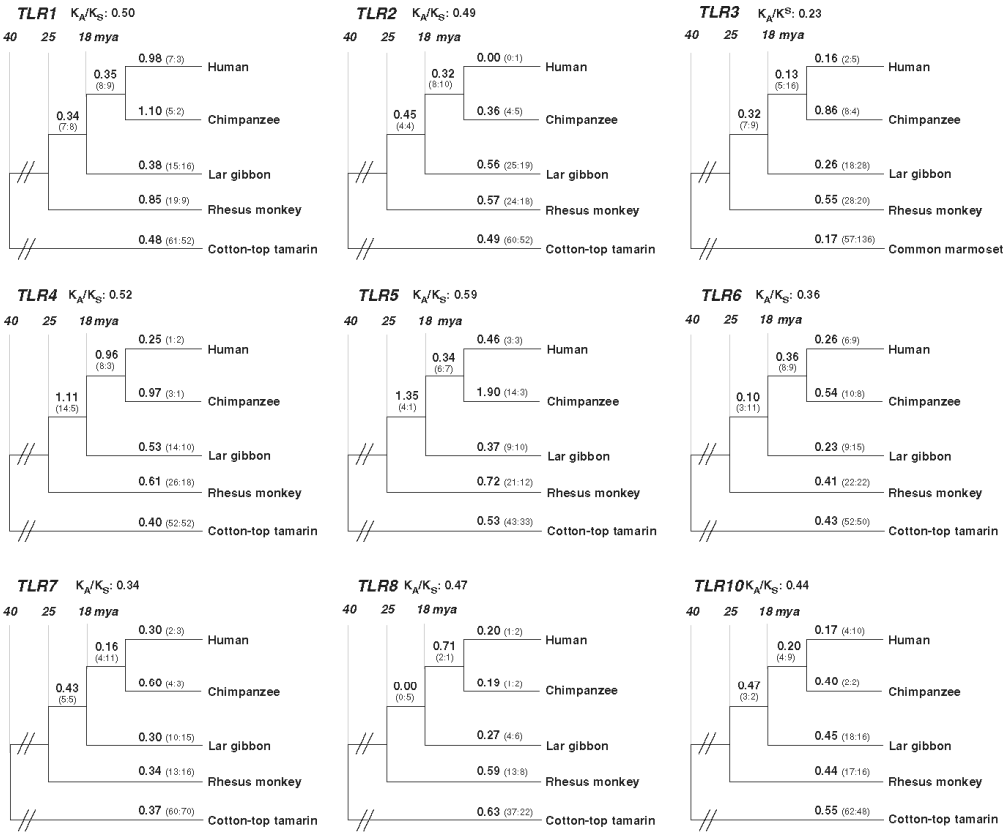


Figure 5 Phylogenetic trees of Toll-like receptor1, -2, -3, -4, -5, -6, -7, -8 and -10. K_A/K_S values and the estimated number of non-synonymous and synonymous substitutions (in parantheses) for each branch are indicated. Approximate divergence time in Mya is shown.

primate homologues and paralogues failed to identify widespread signs of positive selection, as assessed by global K_A/K_S values, by lineage and species-specific K_A/K_S values, or by site-specific analyses. Only more detailed analysis of some lineages suggested episodes of recent evolution (that is species-specific positive selection), in particular among gibbons. L-SIGN (encoded by *CD209L*) in humans may also be a relevant example of species-specific positive selection that has continued to operate during recent human evolution. Indeed, a study screening for the degree of sequence-based diversity in different human populations has shown that *CD209L* exhibits higher diversity in its coding region, as compared to its homologue *CD209*.[11] These observations suggest that there has been an advantage for a higher diversity in *CD209L* with respect to *CD209* within humans, probably driving *CD209L* to accumulate new mutations and eventually new functions, possibly compensating the recent loss of *CD209L2* in humans.

The general pattern of purifying selection of the *CD209* family of genes (K_A/K_S range 0.35–0.52) and *TLR* family of genes (K_A/K_S range 0.23–0.59) appears similar. These values are greater than the genome-wide average (K_A/K_S $\sim$0.2) for human–chimpanzee gene pairs.[12] They are, however, much lower than the K_A/K_S estimates for intrinsic defense genes investigated in the context of human immunodeficiency virus pathogenesis, such as *TRIM5α* and *APOBEC3G* ($K_A/K_S = 1.1$ for both genes).[13] Comparative analysis of the genes that encode APOBEC3G and TRIM5α has revealed the intensity of the selective pressures resulting from the long-standing conflict between retroviruses and their hosts.[14,15] These two genes/proteins constitute a paradigm of pathogen-driven positive selection pressure, not only because of their global elevated K_A/K_S values, but also by the identification of precise residues and domains under strong positive selection. Analysis of APOBEC3G across primate species reveals many residues in the amino-terminal cytidine deaminase domain that are under positive selection, which coincide indeed with the proposed region of interaction with the human immunodeficiency virus-1 Vif protein. Analysis of the TRIM5α

pinpoints a patch of amino acids that is under positive selective pressure at variable regions V1 and V2. The variable regions of TRIM5α have in turn evolved independently to recognize the various retroviral capsids.[16]

A second point of discussion concerns the length of the neck region that characterizes primate *CD209* and *CD209L* among other *CD209*-like genes in mammals. In this study and in previous analyses in primates[1], *CD209* and *CD209L* have variable numbers of repeats. A single partial repeat characterizes primate *CD209L2*; a general feature of the neck region of other mammal *CD209*-like genes. The first and last repeat units of the neck region in *CD209* and *CD209L* originate from a common ancestral motif. We studied only one individual per species, but Barreiro *et al*.[11] studied the degree of polymorphism of the neck region in both *CD209* and *CD209L* by genotyping 1064 individuals from 52 worldwide populations. Striking differences were observed between the two genes. Although minimal variation was observed for *CD209*, *CD209L* exhibited a strong variation in allelic frequencies of different neck lengths. In their population genetics study, the *CD209* genes did not experience selective sweeps (that is directional selection, leading to rapid spread/fixation of an advantageous allele) but rather experienced purifying selection (*CD209*) or balancing selection (*CD209L*).[11] Thus, these duplicated genes have evolved, and might still evolve, under completely different selective pressures.

The strong contrast observed in length variation of the neck region between the various genes may have consequences on function. A number of association studies in human populations have attempted to correlate length variation of the neck region and promoter variants with susceptibility to infectious diseases whose etiological agents are known to interact with one (or both) of these lectins. Results, in particular from the analysis of susceptibility to human immunodeficiency virus-1 or *M. tuberculosis* infection, remain under discussion.[11,17–23]

The usurpation of DC-SIGN and other family members by pathogens such as retroviruses, Ebola and *Mycobacteria* might have led to a pattern of distinctive patches or domains with the characteristic of positive selective pressure as a result of a genetic conflict.[6,13] Against this expectation, the evolutionary analysis of the family suggests both a pattern of apparent redundancy of *CD209* genes (gene absence and gene loss) and of positive selection in specific lineages, in the context of a general pattern of gene conservation. A similar pattern is identified for a second family of proteins of the innate immunity, the TLRs. This pattern is consistent with the concept put forward by Lynch and Conery[24] proposing that most duplicated genes experience a brief period of relaxation of the selective constraint early in their history. Thereafter, most gene duplicates are silenced within a few million years, with a minority of duplicated genes subsequently experiencing strong purifying selection. The observed degree of purifying selection may be expected given the need to faithfully recognize various pathogen motifs, the inability of the pathogen to modify the molecular pattern,[25] and in the case of the DC-SIGN family, the need to recognize the commensal flora, as well as ICAM-2, ICAM-3 adhesion molecules and other self-proteins.

Materials and methods

Primates

CD209 gene family coding sequences from primates were generated by amplification and sequencing of genomic DNA: bonobo (*Pan paniscus CD209* EU041926, *CD209L* EU041931, *CD209L2* EU041934), chimpanzee (*Pan troglodytes CD209L2* EU041935), bornean orangutan (*Pongo pygmaeus CD209L* EU041932, *CD209L2* EU041936), lar gibbon (*Hylobates lar CD209L2* EU041937), nomascus (*Hylobates leucogenys CD209* EU041927, *CD209L* EU041933, *CD209L2* EU041938), siamang (*Hylobates syndactylus CD209L2* EU041939), red baboon (*Papio hamadryas CD209L2* EU041940), African green monkey (*Cercopithecus* (*chlorocebus*) *aethiops CD209* EU041928, *CD209L2* EU041941), owl monkey (*Aotus trivirgatus CD209* EU41930). Common marmoset (*Callithrix jacchus jacchus CD209* EU041929, *CD209L2* EU041942) sequences were obtained by amplification and sequencing of cDNA from liver. For cotton-top tamarin (*Saguinus Oedipus)* and golden headed lion (*Leontopithecus rosalia chrysomelas*), partial sequences were obtained for *CD209* and *CD209L2* from cDNA of peripheral blood. Other primate and mammal sequences were downloaded from the NCBI database, human (*Homo sapiens CD209* AF290886, *CD209L* BC038851), chimpanzee (*CD209* AY078913, *CD209L* AH011538), gorilla (*Gorilla gorilla CD209* AY078906, *CD209L* AH011537), bornean orangutan (*CD209* AY078905), lar gibbon (*CD209* AH011540, *CD209L* AH011531), siamang (*CD209* AY078878, *CD209L* AH011532), black crested gibbon (*Hylobates concolor CD209* AY078885, *CD209L* AH011533), rhesus monkey (*Macaca mulatta CD209* NH_001032870, *CD209L2* AY074781), red baboon (*CD209* AY078864), mouse (*Mus musculus Cd209a* AF373408, *Cd209b* AF373409, *Cd209c* AF373410, *Cd209d* AF373411, *Cd209e* AF373412, *Cd209g* XM_284376, *Signr8* XM_284386) and dog (*Canis lupus familiaris CD209*-like XM_542118). A cow (*Bos taurus*) *CD209*-like was identified by BLAST search on is entire genome (UCSC Genome Bioinformatics, www.genome.ucsc.edu).

Toll-like receptor family of gene sequences were generated by amplification and sequencing of genomic DNA of lar gibbon (*Hylobates lar* TLR1 EU488847, TLR2 EU488848, TLR3 EU488849, TLR4 EU488850, TLR5 EU488851, TLR6 EU488852, TLR7 EU488853, TLR8 EU488854, TLR10 EU488855), cotton-top tamarin (*Saguinus Oedipus* TLR1 EU488856, TLR2 EU488857, TLR4 EU488859, TLR5 EU488860, TLR6 EU488861, TLR7 EU488862, TLR8 EU488863, TLR10 EU488864) and cDNA from common marmoset (*Callithrix jacchus jacchus* TLR3 EU488858). Human sequences were downloaded from the NCBI database (TLR1NM_003263, TLR2NM_003264, TLR3NM_003265, TLR4NM_138554, TLR5NM_003268, TLR6NM_006068, TLR7NM_016562, TLR8NM_138636, TLR10NM_030956). Chimpanzee (*P. troglodytes*) and rhesus monkey (*M. mulatta*) sequences were obtained by BLAST search on their entire genome (UCSC Genome Bioinformatics, www.genome.ucsc.edu).

Molecular analysis

Exons were amplified by primers designed for the flanking intron regions (for genomic DNA), and for 3′- and 5′-UTR regions (for cDNA). HotStarTaq Master Mix (Qiagen AG, Hombrechtikon, Switzerland) was used for

PCR amplification of fragments smaller than 1 kb and PrimeSTAR DNA polymerase (TAKARA Bio Inc. Shiga, Japan) for fragments larger than 1 kb. All primer sequences are presented in Supplementary Table S1. Sequences were aligned using MUSCLE.[26] Coding regions were aligned according to their corresponding amino-acid sequences using the European Molecular Biology Open Software Suite package.[27] To perform phylogenetic analysis, we use a Bayesian inference of phylogeny with Mr Bayes 3.[28] Different phylogenetic trees were performed at nucleic acid and amino acid levels, with different lengths of sequence (entire sequences, sequences without the neck repeat region, sequences consisting solely of the carbohydrate-recognition domain and of the neck repeat region).

Evolutionary analyses

To trace the evolutionary history of the *CD209* and *TLR* families, we analysed their substitutional patterns in the framework of the accepted primate phylogeny[8] using several codon-based maximum likelihood procedures as implemented in the codeml tool of the phylogenetic analysis by maximum likelihood program package.[29] To obtain an overview of the coding-sequence evolution, we estimated the number of non-synonymous (K_A) over synonymous (K_S) substitutions per site (averaged over the entire sequence) for each branch of the trees using the free-ratio model of codeml.[29] In more detailed analyses, we utilized models that allow for different K_A/K_S rates at different sites of the sequences, because adaptive evolution often occurs at a limited number of sites.[30] We first compared a null model ('M1a'[9,31]), which assumes two site classes (sites under purifying selection and neutrally evolving sites), to an alternative model ('M2a'[9,31]), which adds a third site class that allows for sites with $K_A/K_S > 1$, using likelihood ratio tests.[32]

Acknowledgements

We thank Keith Mansfield and Kuei-Chin Lin from the New England Primate Center, and Charles Buillard and Eugène Chabloz from the Zoo of Servion for materials. This work was funded by the Swiss National Science Foundation and a grant for interdisciplinary research from the Faculty of Biology and Medicine of the University of Lausanne. This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government. This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

1 Bashirova AA, Wu L, Cheng J, Martin TD, Martin MP, Benveniste RE *et al*. Novel member of the CD209 (DC-SIGN) gene family in primates. *J Virol* 2003; **77**: 217–227.
2 Koppel EA, van Gisbergen KP, Geijtenbeek TB, van Kooyk Y. Distinct functions of DC-SIGN and its homologues L-SIGN (DC-SIGNR) and mSIGNR1 in pathogen recognition and immune regulation. *Cell Microbiol* 2005; **7**: 157–165.
3 Wu L, Kewalramani VN. Dendritic-cell interactions with HIV: infection and viral dissemination. *Nat Rev Immunol* 2006; **6**: 859–868.
4 Figdor CG, van KY, Adema GJ. C-type lectin receptors on dendritic cells and Langerhans cells. *Nat Rev Immunol* 2002; **2**: 77–84.
5 Feinberg H, Guo Y, Mitchell DA, Drickamer K, Weis WI. Extended neck regions stabilize tetramers of the receptors DC-SIGN and DC-SIGNR. *J Biol Chem* 2005; **280**: 1327–1335.
6 Yang Z. The power of phylogenetic comparison in revealing protein function. *Proc Natl Acad Sci USA* 2005; **102**: 3179–3180.
7 Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. *Nat Rev Genet* 2007; **8**: 857–868.
8 Goodman M. The genomic record of Humankind's evolutionary roots. *Am J Hum Genet* 1999; **64**: 31–39.
9 Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 2005; **22**: 1107–1118.
10 Snyder GA, Colonna M, Sun PD. The structure of DC-SIGNR with a portion of its repeat domain lends insights to modeling of the receptor tetramer. *J Mol Biol* 2005; **347**: 979–989.
11 Barreiro LB, Patin E, Neyrolles O, Cann HM, Gicquel B, Quintana-Murci L. The heritage of pathogen pressures and ancient demography in the human innate-immunity CD209/CD209L region. *Am J Hum Genet* 2005; **77**: 869–886.
12 Wagner A. Rapid detection of positive selection in genes and genomes through variation clusters. *Genetics* 2007; **176**: 2451–2463.
13 Ortiz M, Bleiber G, Martinez R, Kaessmann H, Telenti A. Patterns of evolution of host proteins involved in retroviral pathogenesis. *Retrovirology* 2006; **3**: 11.
14 Sawyer SL, Emerman M, Malik HS. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol* 2004; **2**: E275.
15 Sawyer SL, Wu LI, Emerman M, Malik HS. Positive selection of primate TRIM5{alpha} identifies a critical species-specific retroviral restriction domain. *Proc Natl Acad Sci USA* 2005; **102**: 2832–2837.
16 Goldschmidt V, Ciuffi A, Ortiz M, Brawand D, Munoz M, Kaessmann H *et al*. Antiretroviral activity of ancestral TRIM5alpha. *J Virol* 2008; **82**: 2089–2096.
17 Barreiro LB, Quintana-Murci L. DC-SIGNR neck-region polymorphisms and HIV-1 susceptibility: From population stratification to a possible advantage of the 7/5 heterozygous genotype. *J Infect Dis* 2006; **194**: 1184–1185.
18 Wichukchinda N, Kitamura Y, Rojanawiwat A, Nakayama EE, Song H, Pathipvanich P *et al*. The polymorphisms in DC-SIGNR affect susceptibility to HIV type 1 infection. *AIDS Res Hum Retroviruses* 2007; **23**: 686–692.
19 Martin MP, Lederman MM, Hutcheson HB, Goedert JJ, Nelson GW, van KY *et al*. Association of DC-SIGN promoter polymorphism with increased risk for parenteral, but not mucosal, acquisition of human immunodeficiency virus type 1 infection. *J Virol* 2004; **78**: 14053–14056.
20 Liu H, Hwangbo Y, Holte S, Lee J, Wang C, Kaupp N *et al*. Analysis of genetic polymorphisms in CCR5, CCR2, stromal cell-derived factor-1, RANTES, and dendritic cell-specific intercellular adhesion molecule-3-grabbing nonintegrin in seronegative individuals repeatedly exposed to HIV-1. *J Infect Dis* 2004; **190**: 1055–1058.
21 Gramberg T, Zhu T, Chaipan C, Marzi A, Liu H, Wegele A *et al*. Impact of polymorphisms in the DC-SIGNR neck domain on the interaction with pathogens. *Virology* 2006; **347**: 354–363.
22 Olesen R, Wejse C, Velez DR, Bisseye C, Sodemann M, Aaby P *et al*. DC-SIGN (CD209), pentraxin 3 and vitamin D receptor gene variants associate with pulmonary tuberculosis risk in West Africans. *Genes Immun* 2007; **8**: 456–467.

23 Vannberg FO, Chapman SJ, Khor CC, Tosh K, Floyd S, Jackson-Sillah D *et al*. CD209 genetic polymorphism and tuberculosis disease. *PLoS ONE* 2008; **3**: e1388.
24 Lynch M, Conery JS. The evolutionary fate and consequences of duplicate genes. *Science* 2000; **290**: 1151–1155.
25 Medzhitov R. Toll-like receptors and innate immunity. *Nat Rev Immunol* 2001; **1**: 135–145.
26 Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**: 1792–1797.
27 Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000; **16**: 276–277.
28 Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 2003; **19**: 1572–1574.
29 Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 1997; **13**: 555–556.
30 Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 2000; **15**: 496–503.
31 Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 2000; **155**: 431–449.
32 Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 1998; **15**: 568–573.

Supplementary Information accompanies the paper on Genes and Immunity website (http://www.nature.com/gene)

4.3 Original article (in preparation)

Evolutionary Pattern of Host Genes Involved in HIV Pathogenesis

Millán Ortiz[1], Nicolas Guex[2], Olivier Martin[2], Ioannis Xenarios[2], Angela Ciuffi[1], Amalio Telenti[1]*

1Institute of Microbiology, University of Lausanne, 2Vital-IT, Swiss Institute of Bioinformatics, Lausanne, Switzerland

Comments of the article:

The recent availability of the complete genomes from five primates allows the analysis of genetic diversity over the last 40 million years evolution. We established a priority list of human candidate genes involved in the HIV-1 life cycle and pathogenesis based on published reports, and a control set of random genes. The orthologous genes were retrieved from the genome of the four non-human primates and we analyzed the nucleotide substitution patterns. Similar median K_A/K_S (~0.2) values were estimated for the set of candidate genes and for the controls genes. The genes were placed in the cellular context and in relation to the different stages of the HIV-1 life cycle. Genes involved in virus entry, early phase, nuclear import, integration, transcription nuclear export and late phases of the viral replication cycle were under significantly stronger purifying selection than the average across control genes. Genes of the innate immunity presented median values of (~0.4), and among these, genes of intrinsic cellular defense were under positive selection. Detailed analysis of genes in the upper K_A/K_S quintile (>0.4) by using up additional primate sequences allowed the identification of residues under positive selection in nine genes. In ten instances, the identified residues were relevant for HIV-1 biology.

Abstract:

Background: The availability of the complete genomes from five primates allows the analysis of genetic diversity over the last 40 million years evolution. We hypothesized that inter-species differences in susceptibility to HIV-1 would be influenced by the long-range selective pressures on host genes associated with HIV-1 pathogenesis.

Materials and Methods: We established a priority list of human genes (n=140) involved in the HIV-1 life cycle and pathogenesis based on published reports, and a control set 100 random genes. We retrieved the orthologous genes from the genome of the four non-human primates (*Pan troglodytes, Pongo pygmaeus abeli, Macaca mulatta* and *Callithrix jacchus*) downloaded from UCSC. We analyzed the nucleotide substitution patterns using codon-based maximum likelihood procedures as implemented in the PAML software.

Results and conclusion: Similar median K_A/K_S value were estimated for the set of 140 genes involved in HIV-1 pathogenesis and for 100 random controls genes; 0.19 and 0.23, respectively. However, while genes involved in HIV-1 early and late replication phases had values similar to those estimated for random genes, other determinants of HIV-1 pathogenesis presented a different evolutionary history. Genes of the innate immunity had median values of 0.40, and among these, genes of intrinsic cellular defense had values K_A/K_S around or greater than 1.0 (positive selection). Detail assessment of thirty genes in the upper K_A/K_S quintile (>0.4) by using additional primate sequences allowed the identification of residues under positive selection in TRIM5a, APOBEC3H, APOBEC3G, CD4, IL4, PTPRC, GML, DEFB1 and L-SIGN. In ten instances, the identified residues are relevant for HIV-1 biology. Thus, long-acting selective pressures on primate genomes may lead to

variation in genes influencing contemporary susceptibility to HIV-1 infection and disease.

Introduction

Infectious diseases are thought to be a major force driving evolution. This concept has been supported by several genome-wide studies that ranked genes involved in immunity and inflammation among those exhibiting the strongest features of positive selective pressure[1,2,3,4]. These studies used comparative genome data across mammals or analysed features of human population differentiation. Therefore, evolutionary genomic analysis could identify genes that play a role in modern susceptibility to infectious diseases[2,5,6].

Among infectious agents, retroviruses can shape the genome both through their contribution to the large mass of genetic material of retroviral origin, and via repetitive attacks by exogenous infection. By now it is well established that a number of cellular factors specifically target retroelemental activity – of endogenous retroviruses, as well as incoming infection[7,8,9,10]. Modern susceptibility to lentiviruses, and in particular to HIV-1 and SIVs, could thus be modulated by past exposure to ancestral retroviral infections, leading to interspecies specificities, and within a given species, to interindividual differences in susceptibility to infection and disease. On the other hand, genetic adaptation to bottlenecks due to epidemics by non-retroviral pathogens may also modify modern susceptibility to HIV-1 through shared determinants of immune response or polymorphism in key cellular processes.

The evolutionary pattern of members of two families of genes involved in intrinsic cellular defense against retroviruses are the best examples of evolutionary pressures exerted by past infections. TRIM5α, a protein that restricts retroviral infection by targeting decapsidation, and APOBE3G and 3F, proteins that deaminate viral RNA, are among the proteins under the strongest positive selective pressure in primates[5,6,11]. Detailed analyses identified patches of residues under positive selection pressure representing regions of direct interaction between viral proteins and the host antiretroviral protein. Reconstruction and functional testing of ancestral antiretroviral proteins or of ancient retroviruses supports evidence for a dynamic process of evolution of antiretroviral specificity in the primate lineages that results in the pattern of restriction to lentivirus and other retrovirus observed in modern primates[12,13,14].

The recent availability of human and of four non-human primate complete genomes facilitates the large scale analysis of evolutionary pressures on coding regions. It allows testing the hypothesis that genes that are under strong selection in primates may be particularly relevant to human susceptibility to HIV-1 disease. For this purpose, we conducted a systematic assessment of 140 genes involved in HIV-1 pathogenesis and biology to characterize their pattern of evolutionary pressure, and to identify single residues in host proteins that are of relevance in infection.

Materials and methods

Selection of genes candidates for the analysis. We screened the literature for (i) genes associated with the biology of HIV-1 (reviewed in[15,16,17,18] and recent studies[10,19,10]); (ii) HIV-1 dependency factors emerging from genome-wide siRNA

screens [21,22,23]; and (iii) genes considered polymorphic and involved in HIV-1 pathogenesis (compiled in www.hiv-pharmacogenomics.org) (**Figure 1**). For the three large siRNA screens, that included over 600 candidates, we restricted analysis to (i) genes identified in at least two of three screens, or to (ii) genes with single nucleotide polymorphisms that reached a nominal significant p value in a recent genome-wide association study of determinants of susceptibility to HIV-1[24].

Orthologous genes identification in non-human primates. To obtain the orthologous sequences of candidate and control genes in the four non-human primates, the last genome assembly (hg18) of human (*Homo sapiens*), (panTro2) of chimpanzee (*Pan troglodytes*), the available genome assembly of (ponAbe2, WUSTL Pongo_albelii-2.0.2) Sumatran orangutan (*Pongo pygmaeus abeli*), the genome assembly (rheMac2) of rhesus monkey (*Macaca mulatta*) and the available genome assembly (calJac1, WUSTL Callithrix_jacchus-2.0.2) of common marmoset (*Callithrix jacchus*) were downloaded from UCSC genome browser. BlatSuite.34 [25] was used with the human coding region sequence (CDSs) as template to retrieve homology sequences from each genome with parameters "-t dnax –q dnax" (i.e. translated DNA). BLAT of the querying sequences in the five primate genomes was performed chromosome by chromosome using positive strand chain, then the homologous sequence with the maximum match value was taken.

Alignment of the orthologous sequences. The CDSs for the 140 candidate genes of the five primates were aligned using the sequence analysis tool MUSCLE (MUltiple Sequence Comparison by Log-Expectation)[26]. CDSs were aligned according to their corresponding amino acid sequences using the tranalign

application of the EMBOSS package (European Molecular Biology Open Software Suite)[27].

Primate evolutionary analysis. To trace the evolutionary history of the 140 candidate host genes and 100 control genes (randomly selected among 23000 human genes), we analysed their substitutional patterns in the framework of the accepted primate phylogeny[28] using several codon-based maximum likelihood procedures as implemented in the codeml tool of the PAML (Phylogenetic Analysis by Maximum Likelihood) program package[29]. To obtain an overview of the coding sequence evolution, we estimated the number of non-synonymous (K_A) over synonymous (K_S) substitutions per site (averaged over the entire sequence) for each branch of the trees using the free-ratio model of codeml. In more detailed analysis, in certain genes, we utilized models that allow for different K_A/K_S rates at different sites of the sequences, because adaptive evolution often occurs at a limited number of site[30]. We first compared a null model (M1a[31,32),] which assumes two site classes (sites under purifying selection and neutrally evolving sites), to an alternative model (M2a[31,32]), which adds a third site class that allows for sites with $K_A/K_S>1$) using likelihood ratio test. We then use empirical Bayes methods to identify positively selected sites when they exist. Sites having a posterior probability (Post prob. >0.95) are estimated to be under positive selection. For this detailed analysis all the primates complete CDSs available in NCBI database were used.

Cellular localization and classification. Genes/proteins were placed in the cellular context and in relation to the different stages of the HIV-1 life cycle. Molecular function, cellular location and possible interaction between the host factors and HIV-1

proteins (**Supplementary table S1**) were assigned using the available literature and dedicated databases (HIV-1 Human Protein Interaction Database from NCBI, http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions; Ingenuity Pathways Analysis Database (Ingenuity® Systems, http://www.ingenuity.com), Gene Ontology Database, http://www.geneontology.org/).

Results

BLAT and alignment efficiency. CDSs of human candidate genes (n=140) were used as BLAT queries against the four non-human primate genomes. We identified and reconstructed all four non-human primate orthologs (558 sequences) with the exception of *L-SIGN* (also called *CLEC4M)*, known to be absent in old and new world monkeys[33,34]. Of the 558 sequences, 338 included the entire CDS compared to the human sequence template. The remaining sequences contained gaps of different size in different primates. Gaps represented 1% to 30% of the CDS compared to the human reference sequence. The only exception was *CCL18* in common marmoset, which missed 68% of the expected CDS. For the one hundred control random genes we identified all four non-human primates orthologs (n=400). Over the 400 sequences, 284 included the entire CDS compared to the human sequence template. The remaining sequences (n=116), contained gaps representing 1% to 10% of CDS. Gap sequences could be due to the incomplete genome assembly or represent an actual event of deletion/insertion. There is debate about the best handling of short deletions or insertions in the analysis of sequence evolution[35]. In this study, insertions and deletions were defined from the consensus sequence. Positions in gaps were coded as "?" to be interpreted as undetermined nucleotide by the PAML program. Insertions were removed from analysis.

K_A/K_S analysis. The K_A/K_S median value for the one hundred random control genes was 0.19 (purifying selection). The K_A/K_S median value for candidate genes involved in HIV-1 pathogenesis was 0.23. Classification of candidate genes according to molecular function resulted in differences in K_A/K_S median values among functional classes. Immune response genes, including the innate and intrinsic immunity had median values of 0.37 and 1.04; p value=0.0001 and p=0.0005, respectively, compared to the control genes (**Figure 2**). Genes involved in virus life cycle steps (entry, early phase, nuclear import, integration, transcription, nuclear export and late phase) presented median values similar or below that of control genes. These differences were statistically significant for genes involved in transcription (p=0.0125), nuclear export (p=0.0158) and late phases of the viral cycle (p<0.0001). K_A and K_S as well as the K_A/K_S ratio values for all genes analyzed in this work are annotated in **Supplementary table S1**.

We compiled and presented data in their cellular context (**Figure 3**). Among 42 genes/proteins encoding membrane receptors, plasma membrane proteins, and soluble molecules, we identified 16 host factors (38%) with values in the upper K_A/K_S quintile (K_A/K_S>0.4). The gene under strongest positive selection (K_A/K_S>1) is *CCL11,* encoding for eotaxin (chemokine C-C motif ligand 11), a small secreted protein involved in immunoregulatory and inflammatory processes. It has been suggested that eotaxin controls the migration of immune cells to sites of HIV-1 infection[36]. Other genes/proteins in the upper quantile of K_A/K_S included *PTPRC* (K_A/K_S=0.71), encoding CD45, a protein abundantly expressed on the surface of B- and T-cells, that has important functions in cell maturation and activation; *CD4* (K_A/K_S=0.70),

encoding the primary receptor for SIV/HIV; and *CLEC4M* (K_A/K_S=0.61), encoding a C-type lectin receptor that binds HIV-1. We also found three members of the interferon response system (*IFNAR1*, K_A/K_S=0.62; *IFNGR1*, K_A/K_S=0.52; *IFNG*, K_A/K_S=0.44) in addition to interleukins, interleukin receptors and chemokines (*IL1A*, K_A/K_S=0.76; *IL18*, K_A/K_S=0.46; *IL2RA*, K_A/K_S=0.42; *CCL7*, K_A/K_S=0.65; *CCL18*, K_A/K_S=0.52). *DEFB1* (K_A/K_S=0.58), encodes a defensin from a family of microbicidal and cytotoxic peptides, and *MBL2* (K_A/K_S=0.59) encodes a mannose-binding protein that binds HIV-1 gp120. Other genes/proteins located in the plasma membrane in the upper K_A/K_S quantile includes *GML* (K_A/K_S=0.67), encoding a GPI-anchor protein of the LY6 family, recently associated with differences in cellular susceptibility to infection[19], and Tetherin/BST2 (K_A/K_S=0.58), a protein that restricts HIV-1 release. It is downregulated from the cell surface by direct biding of HIV-1 Vpu (Viral protein U)[10].

Among 53 genes/proteins involved in molecular functions in the cytoplasm, we identified five host factors (9%) with values in the upper K_A/K_S quintile. *APOBEC3H (*K_A/K_S*=1.25), APOBEC3G (*K_A/K_S*=1.25)*, *TRIM5*α *(*K_A/K_S*=1.04)*and *APOBEC3F (*K_A/K_S*=0.91)* are known to be under positive selective pressure[37,5,6,11]. *MAP4* (K_A/K_S=0.57) was recently proposed as HIV-1 dependency factor in two siRNA large scale screens[21,23].

Among 43 genes/proteins involved in nuclear functions, most were under strong purifying selection, with the exception of three genes (7%) with values in the upper K_A/K_S quintile. *TRIM22* (K_A/K_S=0.63) encodes an intrinsic immunity factor involved in the regulation of transcription[20,38]. *DDX53* (K_A/K_S=0.50), recently proposed as HIV-1

dependency factor in a siRNA large scale screen brass[21], is associated with viral mRNA processing. *CDC25C* (K_A/K_S=0.42) encodes a protein promoting cell cycle arrest that binds HIV-1 Vpr (Viral protein R)[15,39]. Functional details and association with HIV-1 pathogenesis of all genes/proteins are detailed in **Supplementary table S1.**

Genes/Proteins with regions under positive selection. In a more detailed analysis, we applied different PAML statistical models (M1a and M2a) to identify residues under positive selection in gene products with a K_A/K_S values >0.4. This analysis includes 28 of the candidate genes mentioned before. Because the identification of sites under positive selection could be limited by the number of primate sequences, we compiled additional complete CDSs of non-human primates in the NCBI database (**Table 1**). APOBEC3H, APOBEC3G, APOBEC3F, TRIM5α, TRIM22 and CLEC4M/L-SIGN have been already well characterized in terms of their evolutionary history[37,5,6,11,34]. For APOBEC3H, APOBEC3G and TRIM5α the availability of additional sequences allowed the detection of novel sites predicted to be under positive selection. For five additional genes/proteins (CD4, IL4, *PTPRC*/C45, GML, DEFB1), we identified one to ten residues to be under positive selection (Post probs. ≥0.95), (**Table 1**). Residues N64, N77, and A80 of CD4 are distributed on the molecular surface at the top of domain D1 at the interface between CD4 and the HIV protein gp120, and between CD4 and MHC-II[40]. PTPRC/CD45 L33, N276, A277, T292, Q334, K353, F397, V454, R504 and K533 are located in the extracellular spacer and fibronectin domains. This protein region could be under positive selective pressure due to a genetic conflict between host and a pathogen, but there is no evidence for CD45 to bind directly to a retroviral protein although it

has been implicated in the binding of influenza virus[41]. Altered CD45 expression in C77G carriers influences immune function and outcome of hepatitis C infection[42]. A recent study described the promoter region of *IL4* gene to be under selective pressure due to local adaptation to diverse pathogenic challenges[43]; however, IL4 S152 is located at the C-terminus. No information is available regarding DEFB1 R61, or GML S107.

Control genes. Two genes in the control group, *IFNA8* (K_A/K_S=1.33) and *PHYHD1* (K_A/K_S=1.22) are under strong positive selection (**Supplementary table S2**). *PHYHD1* encodes for hytanoyl-CoA dioxygenase domain containing 1, a molecule with oxidoreductase activity. PHYHD1 was described as interacting with Epstein Barr virus BRRF1 protein in a high-throughput yeast two-hybrid system[44]. BRRF1 is a transcription factor cooperating in the induction of the lytic form of EBV infection in certain cell types[45]. No residues were found to be under positive selection (**Table 1**). *IFNA8* encodes an interferon molecule implicated in anti-viral response, although not reported to be specifically associated with HIV-1 pathogenesis. IFNA8 R45, a surface accessible residue, is predicted to be under positive selection (**Table 1**).

Discussion:

Human susceptibility to HIV-1 occurred at a time in evolution of the species when host capacity to abort, control, or resolve infection was limited. The almost-universal susceptibility to HIV-1 in humans stands in stark contrast to the outcome of infection of SIV in their natural primate hosts. Members of the TRIM and APOBEC families of proteins, and Tetherin/Bst2 account for differences in retroviral restriction specificity among primate species. The differences in pathogenesis that underlie the ability to

tolerate viral replication without immune damage in sooty mangabeys or African green monkeys, have been difficult to map to precise genes. While naturally SIV-infected primates are thought to have gone through a long host-pathogen adaptation process – in contrast with the recent infection of humans with HIV-1. Thus the interest and the goal of the current work to assess the evolutionary profiles of 140 candidate and confirmed genes/proteins involved in HIV-1 biology and pathogenesis. This is now facilitated by the possibility to search for orthologs of human genes in the complete genomes of four evolutionary distant primates.

The choice of genes for study covers confirmed cellular factors needed for viral replication, as well as candidate genes that have emerged from genetic association studies and from three genome-wide siRNA screens. This set of genes allows the following observations: (i) three general groups of genes with different evolutionary history of their coding regions in primates (as captured by the K_AK_S ratio), (ii) the non *a priori* identification of residues that are candidate interaction domains with pathogen(s). Three general patterns of selective pressure in coding regions emerged from the analysis of 40 million years of primate evolution.

First, cellular genes proven (or proposed) involved in early phase of the viral cycle, nuclear import, integration, transcription, nuclear export and late phase are under strong purifying selection – with most values below the genome-wide average of $K_AK_S = 0.2$. Thus, it can be argued that differences in susceptibility to retroviruses among primates are less likely to be coded (and coding) in those genes. It could also be argued that the available mutational space available for mutation and polymorphisms in modern humans may be limited. This consideration may be of

interest when planning re-sequencing of these genes for the purpose of identifying the genetic basis for differential HIV-1 susceptibility and disease evolution in humans.

Second, cytokines, chemokines, and other soluble mediators and receptors of the innate immunity have a distinctive pattern of evolution of coding regions with average K_AK_S values approaching a ratio of 0.4. This values can be referred to as relaxing of purifying selection, balancing selection, or though of as evidence of positive selection acting of domains of these proteins. In contrast with the first set of genes, the available mutational space may be larger, and diversification of these mediators of the innate immunity might have functional consequences – including in the pathogenesis of retroviral infection.

Third, the current study confirms the role of positive selection in shaping intrinsic cellular defense against retroviruses – including the recently described pattern of Tetherin/Bst2. It is apparent from this analysis that a representative set of 5 primate genomes can capture the signal of strong positive selection in those genes, thus highlighting the need to study in detail other uncharacterized genes that present similar K_AK_S ratios for a possible role in anti-retroviral defense.

Next, we used approaches to allow for different K_A/K_S rates at different sites of the sequences, because adaptive evolution often occurs at a limited number of sites[30]. This type of analysis can provide precious information on residues and pathogen-interacting domains. The accuracy of the analysis increases with increasing numbers of primate sequences. Sites predicted under selective pressure could be thus identified in 9 proteins. It confirmed and extended the number of residues under

selective pressure in the intrinsic defense proteins TRIM5α, APOBEC3H and APOBEC3G. In addition, it identified three residues, N64, N77, and A80, of receptor CD4. Residues N64 and N77 appear to be involved directly in binding of gp120, residue A80 may affect gp120 binding indirectly[46]. It is possible that the residues under positive selective pressure identified in IL4, PTPRC, GML, DEFB1, and CLEC4M/L-SIGN are relevant to host-pathogen interactions.

In conclusion, this work represents a large-scale approach to the characterization of the evolutionary history of genes involved in retroviral pathogenesis during primate evolution. This is an effort that needs to be followed by the analysis at human genome level, with the progressive precision provided by additional complete genomes of primates and prosimians. The outcome of such studies should guide functional analysis of candidate genes, and of residues that may identify interaction domains with pathogens.

Acknowledgements

This work was funded by the Swiss National Science Foundation and by a grant from the academic foundation Infectigen.

Figure 1. Number of candidate genes involved in the HIV life cycle and pathogenesis selected for analysis.

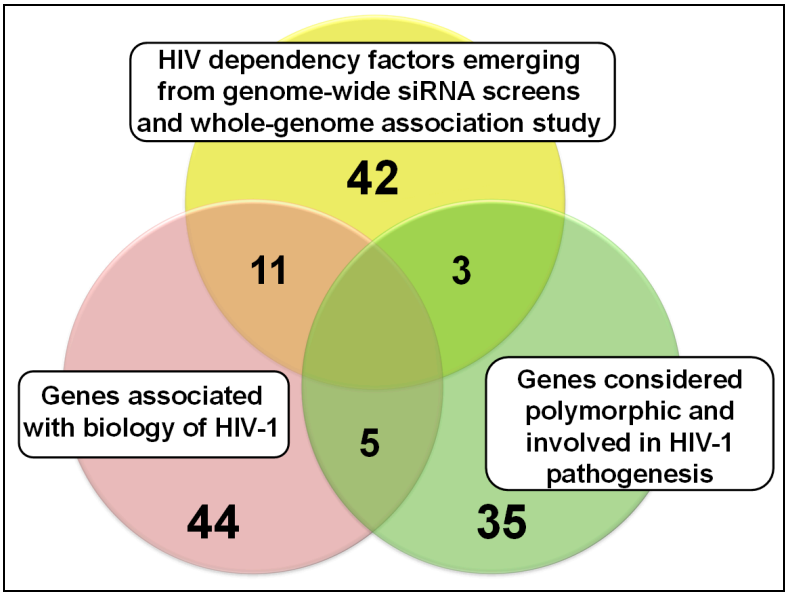


Figure 2. **K_A/K_S values of candidates and control genes.** Each point represents the K_A/K_S ratio value for one gene. Genes with $K_A/K_S>1$ are consider under positive selection. Candidates genes are divided in different categories according to their function in relation with the HIV-1 cell cycle. The bar represents median values of the K_A/K_S ratios of the different genes groups.

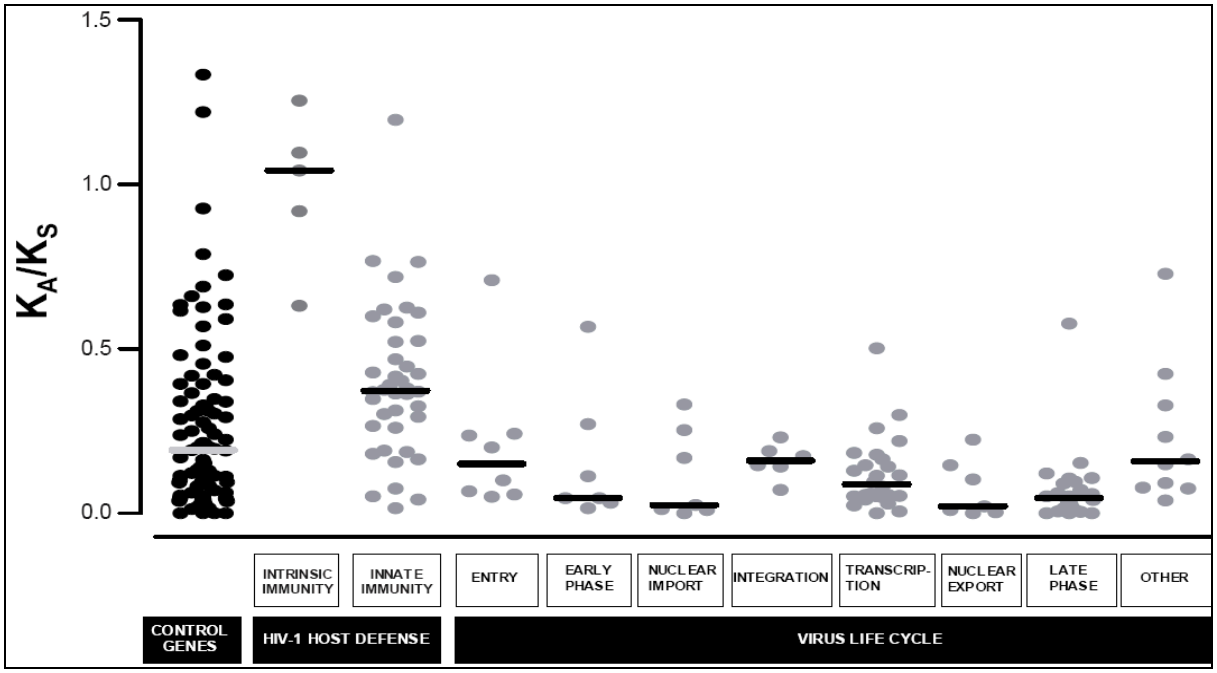


Figure 3. Subcellular location and degree of selective pressure of the candidate genes. Host factors are localized in the cell using the different stages of the HIV-1 life cycle as a framework (blue circles). The subcellular location, function and possible direct interaction with HIV-1 elements was determined using multiple databases and the literature. The degree of molecular evolution is classified in K_A/K_S quantiles: 0 to 0.1151 (grey), 0.1152 to 0.2242 (yellow), 0.2377 to 0.3931 (bisque), 0.4055 to 0.9271 (orange), and higher than one (red). Solid lines correspond to direct protein-protein binding. Dashed lines correspond to ligand-receptor interactions. Selected HIV-1 proteins (Gag, gp120, IN, Nef, Rev, Tat, Vif, Vpr, Vpu) are depicted in black. Endoplasmic reticulum (ER).

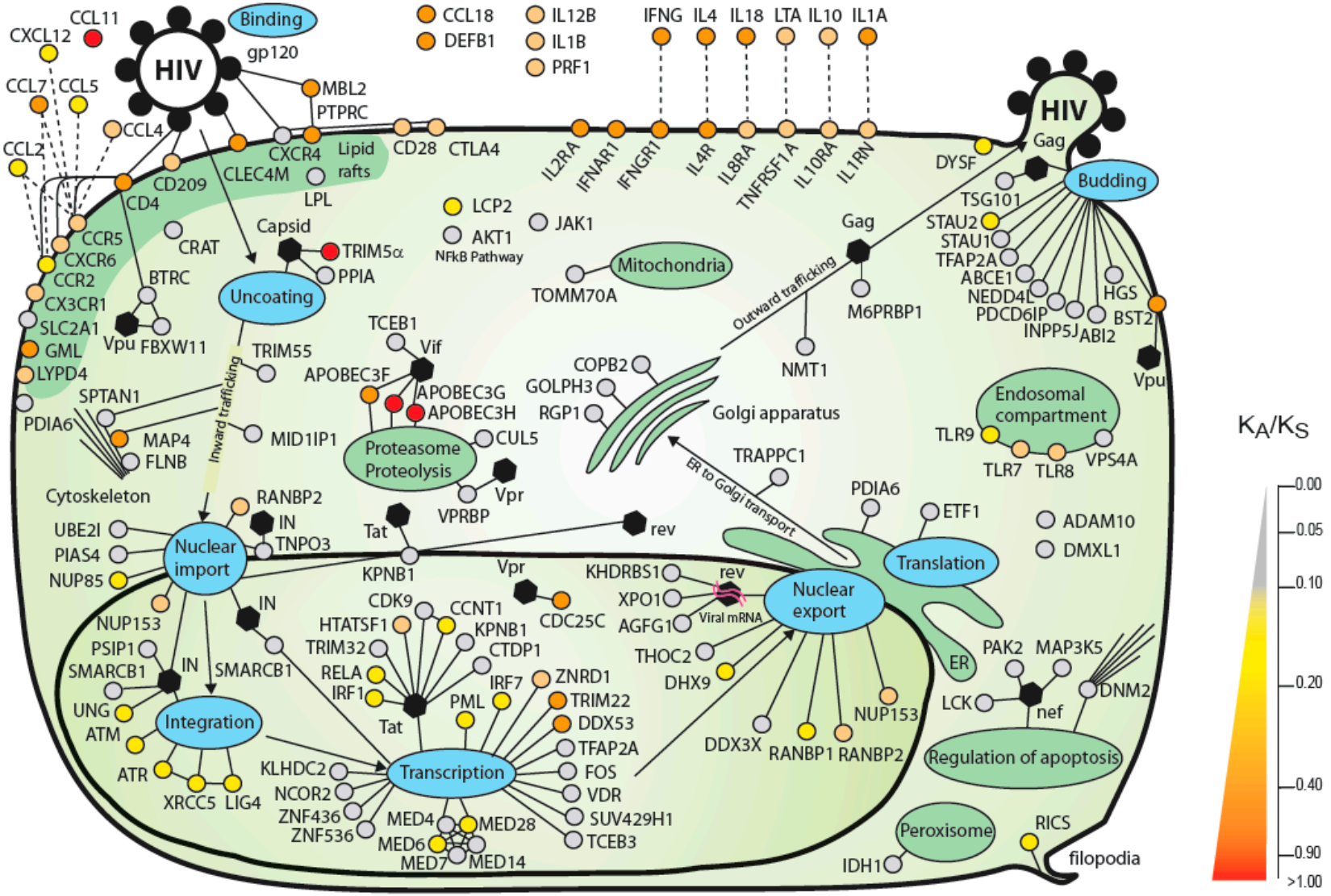

HIV
Binding
gp120
CCL11
CXCL12
CCL7
CCL5
CCL4
CCL2
MBL2
PTPRC
CXCR4
Lipid rafts
CD28
CTLA4
CD209
CD4
CLEC4M
LPL
CCR5
CXCR6
CCR2
CX3CR1
SLC2A1
GML
LYPD4
CRAT
BTRC
Vpu
FBXW11
Capsid
TRIM5α
PPIA
Uncoating
TRIM55
Inward trafficking
SPTAN1
PDIA6
MAP4
FLNB
MID1IP1
Cytoskeleton
CCL18
DEFB1
IL12B
IL1B
PRF1
IFNG
IL4
IL18
LTA
IL10
IL1A
IL2RA
IFNAR1
IFNGR1
IL4R
IL8RA
TNFRSF1A
IL10RA
IL1RN
LCP2
AKT1
NFkB Pathway
JAK1
Mitochondria
TOMM70A
TCEB1
Vif
APOBEC3F
APOBEC3G
APOBEC3H
Proteasome
Proteolysis
CUL5
Vpr
VPRBP
RANBP2
IN
TNPO3
UBE2I
PIAS4
NUP85
Nuclear import
Tat
KPNB1
rev
COPB2
GOLPH3
RGP1
Golgi apparatus
TRAPPC1
ER to Golgi transport
Outward trafficking
Gag
M6PRBP1
NMT1
HIV
Gag
DYSF
Budding
TSG101
STAU2
STAU1
TFAP2A
ABCE1
NEDD4L
PDCD6IP
INPP5J
ABI2
HGS
BST2
Vpu
Endosomal compartment
TLR9
TLR7
TLR8
VPS4A
PDIA6
ETF1
Translation
ADAM10
DMXL1
Vpr
CDC25C
KHDRBS1
rev
XPO1
AGFG1
Viral mRNA
Nuclear export
THOC2
DHX9
DDX3X
RANBP1
RANBP2
NUP153
ER
PAK2
MAP3K5
LCK
nef
DNM2
Regulation of apoptosis
NUP153
PSIP1
SMARCB1
IN
SMARCB1
UNG
ATM
Integration
ATR
XRCC5
LIG4
CDK9
CCNT1
HTATSF1
KPNB1
TRIM32
CTDP1
RELA
IRF1
IRF7
ZNRD1
PML
TRIM22
DDX53
Tat
TFAP2A
FOS
VDR
SUV429H1
TCEB3
Transcription
KLHDC2
NCOR2
ZNF436
ZNF536
MED4
MED28
MED6
MED7
MED14
Peroxisome
IDH1
RICS
filopodia
KA/KS
0.00
0.05
0.10
0.20
0.40
0.90
>1.00

Table 1. Test for positive selection in codons of candidate genes.

Gene name	Gene ID	Primates (1)	K_A/K_S (2)	K_A/K_S>1(3)	*p*-value	Sites with K_A/K_S > 1 (4)
Candidates genes						
TRIM5α	85363	31	**1.209**	4.49 (17.99%)	1.67E-35	**V7, Q107, Q175, L182, T215, L228, Q272, C310, K324, P325, I328, G330, R332, R335, Q337, T338, F339, V340, N343, P381, C385, K389, E405, V408, F418, P421, V423, G483**
APOBEC3H	164668	14	**1.204**	3.79 (31.82%)	8.20E-09	**R20, R21, C53, W90, Y113, K117, P118, Q119, C127, G128, P139, A142, E152, N169, A172**
APOBEC3G	60489	16	**1.144**	3.98 (22.81%)	2.19E-34	**E61, R69, H72, K76, R78, R82, E85, T98, K99, T101, R102, D103, T106, V120, D128, P129, E133, S137, K141, R142, R146, D155, Q168, R169, I187, R213, D274, D276, E330, K344**
CD4	920	12	**0.769**	3.29 (15.97%)	3.34E-05	**N64, N77, A80**
IL4	3565	10	**0.765**	20.22 (1.26%)	1.36E-02	**S152**
PTPRC	5788	5	**0.719**	8.86 (5.54%)	1.60E-17	**L33, N276, A277, T292, Q334, K353, F397, V454, R504, K533**
GML	2765	12	**0.669**	1.94 (36.36%)	3.66E-02	**S107**
DEFB1	1672	5	**0.581**	42.67 (1.5%)	1.14E-03	**R61**
CLEC4M (L-SIGN)	10332	8	**0.538**	5.28 (9.83%)	3.45E-04	**A88**
Control genes						
IFNA8	3445	5	**1.334**	9.36 (6.89%)	6.98E-03	**R45**
PHYHD1	254295	5	**1.220**	10.80 (5.19%)	9.64E-03	**-**

(1) Number of primate sequences available for analysis.
(2) K_A/K_S branch values averaged over the entire tree (free ratio model).
(3) The K_A/K_S value for the additional site class and percentage of residues under positive selection.
(4) Sites under positive selection by Bayes Empirical Bayes analysis. Posterior probability >0.95 (underlined if > 0.99). The numbering is based on the human sequence.

Supplementary table S1. Functional details and association with HIV-1 pathogenesis of all genes.

Gene symbol	Entrez GeneID	K_A	K_S	K_A/K_S	Description (RefSeq)	Location	Molecular function	Biological Process	Reference	Virus partner	Classification for Analysis
ABCE1	6059	0.001	0.159	**0.005**	ATP-binding cassette, sub-family E (OABP), member 1	Cytoplasm	Nucleotide binding	RNA catabolic process	Goff. Nat Rev Microbiology, Swanson and Malim. Cell 2008,	-	LATE PHASE
ABI2	10152	0.002	0.099	**0.020**	abl interactor 2	Cytoplasm	Cytoskeletal adaptor activityDNA binding	Cytoskeleton organization	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	LATE PHASE
ADAM10	102	0.007	0.165	**0.040**	ADAM metallopeptidase domain 10	Cytoplasm	Protein kinase binding	Cell-cell signaling	Brass et al. Science 2008, Fellay et al. Science 2007	-	OTHER
AGFG1	3267	0.002	0.078	**0.022**	HIV-1 Rev binding protein	Nucleus	RNA binding	mRNA nucleur export from nucleus	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	Binds Rev	NUCLEAR EXPORT
AKT1	207	0.013	0.814	**0.016**	v-akt murine thymoma viral oncogene homolog 1	Cytoplasm	Protein import into nucleus	Protein kinase activity	Zhou et al. Cell Host Microbe 2008, Brass et al. Science 2008	-	INNATE IMMUNITY
APOBEC3F	200316	0.140	0.153	**0.919**	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3F	N.A	Cytidine deaminase	Intrinsic immunity	Harris and Liddament. Nat Rev 2004, Malim M. Rev Philos. Trans. R. Soc. Lond., Ser. A 2008	Binds Vif	INTRINSIC IMMUNITY
APOBEC3G	60489	0.346	0.315	**1.096**	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G	Cytoplasm	Cytidine deaminase	Intrinsic immunity	Swanson and Malim. Cell 2008, Brass et al. Science 2008	Binds Vif	INTRINSIC IMMUNITY
APOBEC3H	164668	0.272	0.217	**1.255**	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3H	Cytoplasm	N.A	Cyclic amidines	Harris and Liddament. Nat Rev 2004, OhAinle M et al. Cell Host Microbe 2008	-	INTRINSIC IMMUNITY
ATM	472	0.036	0.155	**0.231**	ataxia telangiectasia mutated	Nucleus	DNA binding	DNA repair	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	INTEGRATION
ATR	545	0.018	0.123	**0.147**	ataxia telangiectasia and Rad3 related	Nucleus	DNA binding	DNA repair	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	INTEGRATION
BST2	684	0.227	0.394	**0.577**	bone marrow stromal cell antigen 2 (tetherin)	Plasma membrane	Signal transducer activity	Immune response	Swanson and Malim. Cell 2008	-	LATE PHASE
BTRC	8945	0.005	0.095	**0.052**	beta-transducin repeat containing	Cytoplasm	Protein binding	Ubiquitin-dependent protein catabolic process	Swanson and Malim. Cell 2008	Binds Vpu	LATE PHASE
CCL11	6356	0.169	0.141	**1.197**	chemokine (C-C motif) ligand 11	Extracellular Space	Chemokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
CCL18	6362	0.152	0.292	**0.522**	chemokine (C-C motif) ligand 18	Extracellular Space	Chemokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
CCL2	6347	0.056	0.295	**0.191**	chemokine (C-C motif) ligand 2	Extracellular Space	Chemokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY

CCL4	6351	0.063	0.194	**0.326**	chemokine (C-C motif) ligand 4	Extracellular Space	Chemokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
CCL5	6352	0.040	0.222	**0.181**	chemokine (C-C motif) ligand 5	Extracellular Space	Chemokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
CCL7	6354	0.106	0.170	**0.625**	chemokine (C-C motif) ligand 7	Extracellular Space	Chemokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
CCNT1	904	0.026	0.139	**0.184**	cyclin T1	Nucleus	DNA binding	Transcription	Goff. Nat Rev Microbiology, Swanson and Malim. Cell 2008, Brass et al. Science 2008, Fellay et al. Science 2007	Binds Tat	TRANSCRIPTION
CCR2	1231	0.045	0.222	**0.201**	chemokine (C-C motif) receptor 2	Plasma membrane	Receptor activity	Immune response	HIV-Pharmacogenomics	Binds gp120	ENTRY
CCR5	1234	0.051	0.211	**0.243**	chemokine (C-C motif) receptor 5	Plasma membrane	Receptor activity	Immune response	Goff. Nat Rev Microbiology, Swanson and Malim. Cell 2008	Binds gp120	ENTRY
CD209	30835	0.140	0.377	**0.370**	C-type lectin receptor CD209 molecule	Plasma membrane	Receptor activity	Immune response	Swanson and Malim. Cell 2008, HIV-Pharmacogenomics	Binds gp120	INNATE IMMUNITY
CD28	940	0.080	0.211	**0.380**	CD28 molecule	Plasma membrane	Coreceptor activity	Immune response	Swanson and Malim. Cell 2008	-	INNATE IMMUNITY
CD4	920	0.174	0.245	**0.709**	CD4 molecule	Plasma membrane	Receptor activity	Immune response	Goff. Nat Rev Microbiology, Swanson and Malim. Cell 2008, Fellay et al. Science 2007, Brass et al. Science 2008.	Binds gp120	ENTRY
CDC25C	995	0.065	0.153	**0.424**	cell division cycle 25 homolog C (S. pombe)	Nucleus	Protein binding	Regulation of mitosis	Swanson and Malim. Cell 2008	Binds Vpr	OTHER
CDK9	1025	0.008	0.329	**0.024**	cyclin-dependent kinase 9	Nucleus	DNA binding	Transcription	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	Binds Tat	TRANSCRIPTION
CLEC4M	10332	0.032	0.052	**0.610**	C-type lectin domain family 4, member M	Plasma membrane	Receptor activity	Immune response	HIV-Pharmacogenomics	Binds gp120	INNATE IMMUNITY
COPB2	9276	0.008	0.116	**0.072**	coatomer protein complex, subunit beta 2 (beta prime)	Cytoplasm. Associated with Golgi	Protein binding	Vesicle-mediated transport	Swanson and Malim. Cell 2008	-	LATE PHASE
CRAT	1384	0.028	0.417	**0.067**	carnitine acetyltransferase	Cytoplasm	Acyltransferase activity	Lipid metabolic process	Goff. Nat Rev Microbiology.	-	ENTRY
CTDP1	9150	0.051	0.450	**0.114**	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) phosphatase, subunit 1	Nucleus	DNA-directed RNA polymerase activity	Transcription	Brass et al. Science 2008, König et al. Cell 2008	Binds Tat	TRANSCRIPTION
CTLA4	1493	0.039	0.151	**0.261**	cytotoxic T-lymphocyte-associated protein 4	Plasma membrane	N.A	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
CUL5	8065	0.001	0.095	**0.006**	cullin 5 ion channel	Cytoplasm	Ubiquitin protein ligase binding	Ubiquitin-dependent protein catabolic process	Swanson and Malim. Cell 2008	-	LATE PHASE
CX3CR1	1524	0.072	0.306	**0.237**	chemokine (C-X3-C motif) receptor 1	Plasma membrane	Chemokine receptor activity	Immune response	HIV-Pharmacogenomics	-	ENTRY
CXCL12	6387	0.016	0.101	**0.156**	chemokine (C-X-C motif) ligand 12 (stromal cell-derived factor 1)	Extracellular Space	Chemokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY

CXCR4	7852	0.011	0.196	**0.058**	chemokine (C-X-C motif) receptor 4	Plasma membrane	Chemokine receptor activity	Immune response	Goff. Nat Rev Microbiology, Swanson and Malim. Cell 2008, Fellay et al. Science 2007	Binds gp120	ENTRY
CXCR6	10663	0.067	0.214	**0.313**	chemokine (C-X-C motif) receptor 6	Plasma membrane	Receptor activity	Immune response	HIV-Pharmacogenomics	Binds gp120	INNATE IMMUNITY
DDX3X	1654	0.001	0.102	**0.010**	DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked	Nucleus	RNA binding ATP-dependent RNA helicase activity	N.A	Zhou et al. Cell Host Microbe 2008, Brass et al. Science 2008	-	NUCLEAR EXPORT
DDX53	168400	0.161	0.320	**0.502**	DEAD (Asp-Glu-Ala-Asp) box polypeptide 53	Nucleus	RNA binding	N.A	Brass et al. Science 2008, Fellay et al. Science 2007	-	TRANSCRIPTION
DEFB1	1672	0.163	0.281	**0.581**	defensin, beta 1	Extracellular Space	N.A	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
DHX9	1660	0.036	0.162	**0.225**	DEAH (Asp-Glu-Ala-His) box polypeptide 9	Nucleus	RNA binding	Nuclear mRNA splicing	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	NUCLEAR EXPORT
DMXL1	1657	0.022	0.148	**0.149**	Dmx-like 1	N.A	Protein binding	N.A	Brass et al. Science 2008, König et al. Cell 2008	-	OTHER
DNM2	1785	0.012	0.370	**0.033**	dynamin 2	Cytoplasm	Microtubule binding	Positive regulation of apoptosis	Swanson and Malim. Cell 2008, König et al. Cell 2008	-	EARLY PHASE
DYSF	8291	0.029	0.236	**0.121**	dysferlin, limb girdle muscular dystrophy 2B (autosomal recessive)	Plasma membrane	N.A	Muscle contraction	Brass et al. Science 2008, Fellay et al. Science 2007	-	LATE PHASE
ETF1	2107	0.000	0.042	**0.000**	eukaryotic translation termination factor 1	Cytoplasm	RNA binding	Regulation of translational termination	Goff. Nat Rev Microbiology, Brass et al. Science 2008	-	LATE PHASE
FBXW11	23291	0.002	0.123	**0.015**	F-box and WD repeat domain containing 11	Cytoplasm	Protein binding	Protein ubiquitination	Brass et al. Science 2008, Fellay et al. Science 2007	Binds Vpu	LATE PHASE
FLNB	2317	0.010	0.211	**0.047**	filamin B, beta (actin binding protein 278)	Cytoplasm	Actin binding	Cytoskeletal anchoring	Goff. Nat Rev Microbiology,	-	EARLY PHASE
FOS	2353	0.012	0.214	**0.056**	v-fos FBJ murine osteosarcoma viral oncogene homolog	Nucleus	Transcription factor activity	Regulation of transcription	Swanson and Malim. Cell 2008	-	TRANSCRIPTION
GML	2765	0.209	0.287	**0.728**	GPI anchored molecule like protein	Plasma membrane	N.A	Apoptosis	Loeuillet et al. PLoS biol. 2008, Brass et al. Science 2008	-	OTHER
GOLPH3	64083	0.005	0.094	**0.052**	golgi phosphoprotein 3 (coat-protein)	Cytoplasm	N.A	N.A	Brass et al. Science 2008, Fellay et al. Science 2007	-	LATE PHASE
HGS	9146	0.016	0.595	**0.027**	hepatocyte growth factor-regulated tyrosine kinase substrate	Cytoplasm	Intracellular protein transport	Protein domain specific binding	Goff. Nat Rev Microbiology, Fellay et al. Science 2007, Brass et al. Science 2008	-	LATE PHASE
HTATSF1	27336	0.048	0.159	**0.299**	HIV-1 Tat specific factor 1	Nucleus	RNA binding	Transcription	Brass et al. Science 2008, Fellay et al. Science 2007	-	TRANSCRIPTION
IDH1	3417	0.013	0.140	**0.092**	isocitrate dehydrogenase 1 (NADP+), soluble	Cytoplasm	Oxidoreductase activity	Metabolic process	Brass et al. Science 2008, König et al. Cell 2008	-	OTHER
IFNAR1	3454	0.114	0.184	**0.620**	interferon (alpha, beta and omega) receptor 1	Plasma membrane	Type I interferon receptor activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IFNG	3458	0.074	0.165	**0.447**	interferon, gamma	Extracellular Space	Cytokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY

IFNGR1	3459	0.109	0.208	**0.524**	interferon gamma receptor 1	Plasma membrane	Interferon-gamma receptor activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL10	3586	0.070	0.168	**0.416**	interleukin 10	Extracellular Space	Cytokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL10RA	3587	0.098	0.243	**0.403**	interleukin 10 receptor, alpha	Plasma membrane	Interleukin-10 receptor activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL12B	3593	0.058	0.198	**0.294**	interleukin 12B (natural killer cell stimulatory factor 2, cytotoxic lymphocyte maturation factor 2, p40)	Extracellular Space	Cytokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL18	3606	0.125	0.267	**0.469**	interleukin 18 (interferon-gamma-inducing factor)	Extracellular Space	Interleukin-1 receptor binding	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL1A	3552	0.154	0.200	**0.767**	interleukin 1, alpha	Extracellular Space	Interleukin-1 receptor binding	Immune response	HIV-Pharmacogenomics, Zhou et al. Cell Host Microbe	-	INNATE IMMUNITY
IL1B	3553	0.088	0.240	**0.365**	interleukin 1, beta	Extracellular Space	interleukin-1 receptor binding	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL1RN	3557	0.091	0.234	**0.391**	interleukin 1 receptor antagonist	Plasma membrane	Interleukin-1 receptor	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL2RA	3559	0.128	0.300	**0.428**	interleukin 2 receptor, alpha	Plasma membrane	Interleukin-2 receptor activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL4	3565	0.103	0.135	**0.765**	interleukin 4	Extracellular Space	interleukin-4 receptor binding	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL4R	3566	0.121	0.285	**0.424**	interleukin 4 receptor	Plasma membrane	Interleukin-4 receptor activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
IL8RA	3577	0.099	0.326	**0.302**	interleukin 8 receptor, alpha	Plasma membrane	Interleukin-8 receptor activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
INPP5J	27124	0.024	0.226	**0.106**	phosphatidylinositol (4,5) bisphosphate 5-phosphatase, A	Cytoplasm	Inositol-polyphosphate 5-phosphatase activity	N.A	Swanson and Malim. Cell 2008	-	LATE PHASE
IRF1	3659	0.019	0.131	**0.143**	interferon regulatory factor 1	Nucleus	Transcription factor activity	Transcription	HIV-Pharmacogenomics	Binds Tat	TRANSCRIPTION
IRF7	3665	0.092	0.419	**0.220**	interferon regulatory factor 7	Nucleus	Transcription factor activity	Regulation of transcription	Mandl et al. Nat Medecine 2008	-	TRANSCRIPTION
JAK1	3716	0.009	0.169	**0.052**	Janus kinase 1 (a protein tyrosine kinase)	Cytoplasm	Receptor binding	Cytokine and chemokine mediated signaling pathway	Zhou et al. Cell Host Microbe 2008, Brass et al. Science 2008	-	INNATE IMMUNITY
KHDRBS1	10657	0.000	0.118	**0.000**	KH domain containing, RNA binding, signal transduction associated 1	Nucleus	RNA binding	regulation of RNA NUCLEAR EXPORT from nucleus	Goff. Nat Rev Microbiology, König et al. Cell 2008, Fellay et al. Science 2007	Binds Rev	NUCLEAR EXPORT
KLHDC2	23588	0.006	0.106	**0.052**	kelch domain containing 2	Nucleus	Protein binding	N.A	Brass et al. Science 2008, Fellay et al. Science 2007	-	TRANSCRIPTION
KPNB1	3837	0.002	0.128	**0.013**	karyopherin (importin) beta 1	Nucleus	Nuclear localization sequence binding	Protein import into nucleus	Swanson and Malim. Cell 2008, König et al. Cell 2008	Binds, Tat, IN, and Rev	NUCLEAR IMPORT
LCK	3932	0.014	0.192	**0.075**	lymphocyte-specific protein tyrosine kinase	Cytoplasm	Protein binding	Induction of apoptosis	Swanson and Malim. Cell 2008	Binds Nef	INNATE IMMUNITY

LCP2	3937	0.036	0.215	**0.165**	lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)	Cytoplasm	Protein binding	Cytokine secretion	Brass et al. Science 2008, Fellay et al. Science 2007	-	OTHER
LIG4	3981	0.033	0.173	**0.189**	ligase IV, DNA, ATP-dependent	Nucleus	DNA binding	DNA repair	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	INTEGRATION
LPL	4023	0.017	0.221	**0.075**	lipoprotein lipase	Cytoplasm	Lipoprotein lipase activity	Lipid catabolic process	Brass et al. Science 2008, Fellay et al. Science 2007	-	OTHER
LTA	4049	0.046	0.123	**0.376**	lymphotoxin alpha (TNF superfamily, member 1)	Extracellular Space	cytokine activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
LYPD4	147719	0.068	0.208	**0.328**	LY6/PLAUR domain containing 4	Plasma membrane	N.A	N.A	Brass et al. Science 2008, Fellay et al. Science 2007	-	OTHER
M6PRBP1	10226	0.043	0.478	**0.091**	mannose-6-phosphate receptor binding protein 1	Cytoplasm	N.A	Vesicle-mediated transport	Goff. Nat Rev Microbiology, Swanson and Malim. Cell 2008, Fellay et al. Science 2007	Binds Gag-p55	LATE PHASE
MAP3K5	4217	0.007	0.167	**0.042**	mitogen-activated protein kinase kinase kinase 5	Cytoplasm	Protein binding	Induction of apoptosis	Swanson and Malim. Cell 2008	Binds Nef	INNATE IMMUNITY
MAP4	4134	0.074	0.130	**0.568**	microtubule-associated protein 4	Cytoplasm	Structural molecule activity	Negative regulation of microtubule depolymerization	Brass et al. Science 2008, König et al. Cell 2008	-	EARLY PHASE
MBL2	4153	0.151	0.253	**0.599**	mannose-binding lectin (protein C) 2	Extracellular Space	Mannose binding	Immune response	HIV-Pharmacogenomics	Binds gp120	INNATE IMMUNITY
MED14	9282	0.006	0.108	**0.052**	mediator complex subunit 14	Nucleus	Transcription	Regulation of transcription	Brass et al. Science 2008, König et al. Cell 2008, Zhou et al. Cell Host Microbe 2008	-	TRANSCRIPTION
MED28	80306	0.025	0.194	**0.130**	mediator complex subunit 28	Nucleus	Protein binding	Regulation of transcription	Zhou et al. Cell Host Microbe 2008, Brass et al. Science 2008	-	TRANSCRIPTION
MED4	29079	0.005	0.167	**0.030**	mediator complex subunit 4	Nucleus	Transcription activator activity	Regulation of transcription	Zhou et al. Cell Host Microbe 2008, Brass et al. Science 2008	-	TRANSCRIPTION
MED6	10001	0.014	0.083	**0.165**	mediator complex subunit 6	Nucleus	Transcription	Regulation of transcription	Brass et al. Science 2008, König et al. Cell 2008	-	TRANSCRIPTION
MED7	9443	0.008	0.141	**0.056**	mediator complex subunit 7	Nucleus	Transcription	Regulation of transcription	Brass et al. Science 2008, König et al. Cell 2008, Zhou et al. Cell Host Microbe 2008	-	TRANSCRIPTION
MID1IP1	58526	0.026	0.234	**0.113**	MID1 interacting protein 1 (gastrulation specific G12 homolog (zebrafish))	Cytoplasm	Protein binding	Negative regulation of microtubule depolymerization	Brass et al. Science 2008, König et al. Cell 2008	-	EARLY PHASE
NCOR2	9612	0.027	0.440	**0.060**	nuclear receptor co-repressor 2	Nucleus	DNA binding	Transcription	Brass et al. Science 2008, Fellay et al. Science 2007	-	TRANSCRIPTION
NEDD4L	23327	0.006	0.179	**0.036**	neural precursor cell expressed, developmentally down-regulated 4-like	Cytoplasm	Protein binding	Excretion	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	LATE PHASE
NMT1	4836	0.008	0.133	**0.062**	N-myristoyltransferase 1	Cytoplasm	Transferase activity	Protein lipoylation	Brass et al. Science 2008, Fellay et al. Science 2007	-	LATE PHASE
NUP153	9972	0.043	0.129	**0.331**	nucleoporin 153kDa,	Nucleus	Transporter activity	mRNA transport	Brass et al. Science 2008, König et al. Cell 2008	-	NUCLEAR IMPORT
NUP85	79902	0.031	0.181	**0.169**	nucleoporin 85kDa	Cytoplasm	N.A	mRNA transport	Brass et al. Science 2008, Fellay et al. Science 2007	-	NUCLEAR IMPORT

PAK2	5062	0.037	0.224	**0.165**	p21 protein (Cdc42/Rac)-activated kinase 2	Cytoplasm	Protein binding	Regulation of cell growth	Swanson and Malim. Cell 2008	Binds Nef	INNATE IMMUNITY
PDCD6IP	10015	0.011	0.102	**0.108**	programmed cell death 6 interacting protein	Cytoplasm	Receptor activity	Protein transport	Swanson and Malim. Cell 2008	Binds Gag-p6	LATE PHASE
PDIA6	10130	0.021	0.206	**0.100**	protein disulfide isomerase family A, member 6	Plasma membrane	Protein binding	Protein folding	Brass et al. Science 2008, Fellay et al. Science 2007	-	ENTRY
PIAS4	51588	0.013	0.528	**0.026**	protein inhibitor of activated STAT, 4	Cytoplasm	Transcription corepressor activity	Regulation of transcription	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	NUCLEAR IMPORT
PML	5371	0.050	0.342	**0.147**	promyelocytic leukemia	Nucleus	Transcription repressor activity	Transcription	Nisole et al. Nat Rev Microbiol 2005, Towers GJ. Retrovirology Rev 2007	-	TRANSCRIPTION
PPIA	5478	0.006	0.126	**0.047**	peptidylprolyl isomerase A (cyclophilin A)	Cytoplasm, Uncoating	Peptidyl-prolyl cis-trans isomerase activity	Protein folding	Goff. Nat Rev Microbiology, Swanson and Malim. Cell 2008, Fellay et al. Science 2007	Binds CA and Gag-p55	EARLY PHASE
PRF1	5551	0.084	0.317	**0.266**	perforin 1 (pore forming protein)	Cytoplasm	Calcium ion binding	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
PSIP1	11168	0.010	0.141	**0.072**	PC4 and SFRS1 interacting protein 1	Nucleus	DNA bindingR	Provirus integration	Swanson and Malim. Cell 2008, Fellay et al.Science 2007	Binds IN	INTEGRATION
PTPRC	5788	0.141	0.197	**0.719**	protein tyrosine phosphatase, receptor type, C	Plasma membrane	Receptor activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
RANBP1	5902	0.035	0.238	**0.147**	RAN binding protein 1	Nucleus	RNA binding	mRNA NUCLEAR EXPORT from nucleus	Goff. Nat Rev Microbiology, Brass et al. Science 2008	-	NUCLEAR EXPORT
RANBP2	5903	0.041	0.162	**0.254**	RNA binding protein 2	Nucleus	Protein binding	Protein import into nucleus	Brass et al. Science 2008, König et al. Cell 2008	-	NUCLEAR IMPORT
RELA	5970	0.019	0.161	**0.116**	v-rel reticuloendotheliosis viral oncogene homolog A (avian)	Nucleus	Transcription factor activity	Regulation of transcription	Brass et al. Science 2008, König et al. Cell 2008, Zhou et al. Cell Host Microbe 2008	Binds Tat	TRANSCRIPTION
RGP1	9827	0.003	0.148	**0.017**	RGP1 retrograde golgi transport homolog (S. cerevisiae)	Cytoplasm	N.A	N.A	Brass et al. Science 2008, Fellay et al. Science 2007	-	LATE PHASE
RICS	9743	0.032	0.137	**0.232**	Rho GTPase-activating protein	Cytoplasm	Protein binding	Signal transduction	Brass et al. Science 2008, Fellay et al. Science 2007	-	OTHER
SLC2A1	6513	0.012	0.234	**0.050**	solute carrier family 2 (facilitated glucose transporter), member 1	Plasma membrane	Transmembrane transporter activity	Transmembrane transport	Goff. Nat Rev Microbiology.	-	ENTRY
SMARCB1	6598	0.001	0.244	**0.006**	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1	Nucleus	protein binding	Transcription	Goff. Nat Rev Microbiology, Swanson and Malim. Cell 2008, Fellay et al. Science 2007	Binds IN	TRANSCRIPTION
SPTAN1	6709	0.003	0.177	**0.016**	spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)	Plasma membrane	Structural constituent of cytoskeleton	Barbed-end actin filament capping	Brass et al. Science 2008, Fellay et al. Science 2007	-	EARLY PHASE
STAU1	6780	0.011	0.112	**0.096**	staufen, RNA binding protein, homolog 1 (Drosophila)	Cytoplasm	RNA binding	RNA localization	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	LATE PHASE
STAU2	27067	0.013	0.087	**0.154**	staufen, RNA binding protein, homolog 2 (Drosophila)	Cytoplasm	RNA binding	Transport	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	LATE PHASE
SUV420H1	51111	0.012	0.124	**0.096**	suppressor of variegation 4-20 homolog 1 (Drosophila)	Nucleus	Transcription	Regulation of transcription	Brass et al. Science 2008, Fellay et al. Science 2007	-	TRANSCRIPTION

TCEB1	6921	0.000	0.107	**0.000**	transcription elongation factor B (SIII), polypeptide 1 (15kDa, elongin C)	Cytoplasm	Protein binding	Regulation of transcription from RNA polymerase II promoter	Swanson and Malim. Cell 2008, König et al. Cell 2008	Binds Vif	LATE PHASE
TCEB3	6924	0.028	0.159	**0.179**	transcription elongation factor B (SIII), polypeptide 3 (110kDa, elongin A),	Nucleus	Transcription elongation factor activity	Regulation of transcription from RNA polymerase II promoter	Zhou et al. Cell Host Microbe 2008, Brass et al. Science 2008	-	TRANSCRIPTION
TFAP2A	7020	0.000	0.139	**0.000**	transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)	Nucleus	Transcription factor activity	Regulation of transcription	Goff. Nat Rev Microbiology, Swanson and Malim. Cell 2008, Fellay et al. Science 2007	-	TRANSCRIPTION
THOC2	57187	0.011	0.110	**0.104**	THO complex 2	Nucleus	RNA binding	mRNA processing	Brass et al. Science 2008, Fellay et al. Science 2007	-	NUCLEAR EXPORT
TLR7	51284	0.042	0.114	**0.369**	toll-like receptor 7	Plasma membrane.	Receptor activity	Immune response	Schlaepfer et al. J. Immunol 2006	-	INNATE IMMUNITY
TLR8	51311	0.054	0.149	**0.363**	toll-like receptor 8	Plasma membrane.	Receptor activity	Immune response	Schlaepfer et al. J. Immunol 2006	-	INNATE IMMUNITY
TLR9	54106	0.065	0.350	**0.187**	toll-like receptor 9	Endosomal compartment	Receptor activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
TNFRSF1A	7132	0.099	0.285	**0.348**	tumor necrosis factor receptor superfamily, member 1A	Plasma membrane	Receptor activity	Immune response	HIV-Pharmacogenomics	-	INNATE IMMUNITY
TNPO3	23534	0.001	0.104	**0.010**	transportin 3	Cytoplasm	Receptor activity	Protein transport	Brass et al. Science 2008, König et al. Cell 2008	Binds IN	NUCLEAR IMPORT
TOMM70A	9868	0.010	0.131	**0.078**	translocase of outer mitochondrial membrane 70 homolog A (S. cerevisiae)	Mitochondria	Protein binding	N.A	Zhou et al. Cell Host Microbe 2008, Brass et al. Science 2008	-	OTHER
TRAPPC1	58485	0.000	0.177	**0.000**	trafficking protein particle complex 1	Cytoplasm	N.A	vesicle-mediated transport	Brass et al. Science 2008, Fellay et al. Science 2007	-	LATE PHASE
TRIM22	10346	0.105	0.166	**0.631**	tripartite motif-containing 22	Cytoplasm	Transcription factor activity	Regulation of transcription	Swanson and Malim. Cell 2008	-	INTRINSIC IMMUNITY
TRIM32	22954	0.007	0.134	**0.049**	tripartite motif-containing 32	Nucleus	Transcription coactivator activity	n.a	Nisole et al. Nat Rev Microbiol 2005	Binds Tat	TRANSCRIPTION
TRIM5	85363	0.234	0.225	**1.042**	tripartite motif-containing 5	Cytoplasm	Protein binding	Intrinsic immunity	Swanson and Malim. Cell 2008, Towers GJ. Retrovirology Rev 2007	Binds CA	INTRINSIC IMMUNITY
TRIM55	84675	0.026	0.097	**0.271**	tripartite motif-containing 55	Cytoplasm	Protein binding	Signal transduction	Brass et al. Science 2008, König et al. Cell 2008	-	EARLY PHASE
TSG101	7251	0.005	0.117	**0.041**	tumor susceptibility gene 101	Cytoplasm	Vesicular transport	Protein transport	Swanson and Malim. Cell 2008	Binds Gag- p6	LATE PHASE
UBE2I	7329	0.000	0.178	**0.000**	ubiquitin-conjugating enzyme E2I (UBC9 homolog, yeast)	Cytoplasm	Protein binding	Ubiquitin-dependent protein catabolic process	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	NUCLEAR IMPORT
UNG	7374	0.040	0.280	**0.142**	uracil-DNA glycosylase	Nucleus	Uracil DNA N-glycosylase activity	DNA repair	Swanson and Malim. Cell 2008	Binds IN	INTEGRATION
VDR	7421	0.029	0.357	**0.081**	vitamin D (1,25- dihydroxyvitamin D3) receptor	Nucleus	Transcription factor activity	Regulation of transcription	HIV-Pharmacogenomics, Zhou et al. Cell Host Microbe	-	TRANSCRIPTION

VPRBP	9730	0.008	0.133	**0.059**	Vpr (HIV-1) binding protein	Cytoplasm	Binding	Interspecies interaction between organisms	Swanson and Malim. Cell 2008, Brass et al. Science 2008	Binds Vpr	LATE PHASE
VPS4A	27183	0.003	0.298	**0.010**	vacuolar protein sorting 4 homolog A (S. cerevisiae)	Cytoplasm	Nucleotide binding	Endosome transport	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	LATE PHASE
XPO1	7514	0.000	0.121	**0.004**	NUCLEAR EXPORTin 1 (CRM1 homolog, yeast)	Nucleus	RNA binding	mRNA NUCLEAR EXPORT from nucleus	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	Binds Rev	NUCLEAR EXPORT
XRCC5	7520	0.022	0.129	**0.174**	X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining)	Nucleus	DNA binding	DNA repair	Goff. Nat Rev Microbiology, Fellay et al. Science 2007	-	INTEGRATION
ZNF436	80818	0.009	0.166	**0.054**	zinc finger protein 436	Nucleus	DNA binding	Transcription	Brass et al. Science 2008, Fellay et al. Science 2007	-	TRANSCRIPTION
ZNF536	9745	0.013	0.310	**0.042**	zinc finger protein 536	Nucleus	DNA binding	Transcription	Brass et al. Science 2008, Fellay et al. Science 2007	-	TRANSCRIPTION
ZNRD1	30834	0.040	0.154	**0.259**	zinc ribbon domain containing 1	Nucleus	Transcription	Regulation of transcription	Fellay et al. Science 2007, Brass et al. Science 2008, HIV-Pharmacogenomics	-	TRANSCRIPTION

4.4 Applications of evolutionary genetic results to other studies

The data generated by these evolutionary and comparative genetics studies was used in additional work.

1. Calculation of nucleotide substitutional pattern of the viral restriction factor TRIM5α allows estimation of the ancestral states of the coding sequence. It has been possible to reconstruct these ancestral states for TRIM5α and test their specificity in vitro against (modern) HIV-1 and five other retroviruses. This, to better understand the evolution of antiretroviral specificity patterns in primates, in particular along the lineage leading to humans. The results are presented in the article "*Antiretroviral Activity of Ancestral TRIM5α*" [65]. The study proposes a statistically significant reduction of HIV-1 restriction by the reconstructed TRIM5α variants (representing 25 million years of evolution) from the common ancestor of old world monkeys and hominoids toward the human variant. The ancestral TRIM5α showed various patterns of restriction capacities against other (modern) retrovirus tested. This study also reveals novel functionally relevant amino acid variants for virus restriction.

2. Evolutionary data allowed a more comprehensive evaluation of TRIM5α polymorphic residues in humans. These were tested for their effect on HIV-1 replication. Common variants of TRIM5α have no effect or modest effect on HIV-1 disease progression. These variants are remote from clusters of positive selection in the primate lineage. The results are presented in the article "*Role of common human TRIM5α variants in HIV-1 disease progression*" [66].

3. Evolutionary information was included in the structural analysis of APOBEC3G. Clusters of positive selection were maped to a 3 D structural model of APOBEC3G and assessed in the context of protein domains important for packaging of APOBEC3G into virions. This model proposes two main clusters with residues under positive selection exposed at the surface. One of the main clusters is at a distinct location from that of the functionally relevant amino acids. The second cluster overlaps with a cluster that includes the important residues D128 and P129, related with the disruption of APOBEC3G packaging into HIV-1. This results are presented in the article "*Model Structure of Human APOBEC3G*" [67].

4.4.1 Original article

Antiretroviral Activity of Ancestral TRIM5α

Valérie Goldschmidt,[1†] Angela Ciuffi,[1†] **Millan Ortiz**,[1] David Brawand,[2] Miguel Muñoz,[1]Henrik Kaessmann,[2*] and Amalio Telenti[1*]

Institute of Microbiology, University Hospital, 1011 Lausanne,[1] and Center for Integrative Genomics, University of Lausanne, 1015 Lausanne,[2] Switzerland
† V.G. and A.C. contributed equally to this study.

Journal of Virology 2008 Mar;82(5):2089-96 [65]

JOURNAL OF VIROLOGY, Mar. 2008, p. 2089–2096 Vol. 82, No. 5
0022-538X/08/$08.00+0 doi:10.1128/JVI.01828-07

Antiretroviral Activity of Ancestral TRIM5α▽

Valérie Goldschmidt,[1]† Angela Ciuffi,[1]† Millan Ortiz,[1] David Brawand,[2] Miguel Muñoz,[1] Henrik Kaessmann,[2]* and Amalio Telenti[1]*

Institute of Microbiology, University Hospital, 1011 Lausanne,[1] and Center for Integrative Genomics, University of Lausanne, 1015 Lausanne,[2] Switzerland

Received 20 August 2007/Accepted 27 November 2007

The antiretroviral protein TRIM5α is known to have evolved different restriction capacities against various retroviruses, driven by positive Darwinian selection. However, how these different specificities have evolved in the primate lineages is not fully understood. Here we used ancestral protein resurrection to estimate the evolution of antiviral restriction specificities of TRIM5α on the primate lineage leading to humans. We used TRIM5α coding sequences from 24 primates for the reconstruction of ancestral *TRIM5*α sequences using maximum-likelihood and Bayesian approaches. Ancestral sequences were transduced into HeLa and CRFK cells. Stable cell lines were generated and used to test restriction of a panel of extant retroviruses (human immunodeficiency virus type 1 [HIV-1] and HIV-2, simian immunodeficiency virus [SIV] variants SIV_{mac} and SIV_{agm}, and murine leukemia virus [MLV] variants N-MLV and B-MLV). The resurrected TRIM5α variant from the common ancestor of Old World primates (Old World monkeys and apes, ~25 million years before present) was effective against present day HIV-1. In contrast to the HIV-1 restriction pattern, we show that the restriction efficacy against other retroviruses, such as a murine oncoretrovirus (N-MLV), is higher for more recent resurrected hominoid variants. Ancestral TRIM5α variants have generally limited efficacy against HIV-2, SIV_{agm}, and SIV_{mac}. Our study sheds new light on the evolution of the intrinsic antiviral defense machinery and illustrates the utility of functional evolutionary reconstruction for characterizing recently emerged protein differences.

A newly described form of innate immunity, coined "intrinsic immunity," provides a constitutive line of defense, which relies on intracellular obstacles to hinder the replication of pathogens (1). This component of the immune system has gained much attention as a cornerstone of the resistance of mammals against several classes of retroelements and retroviruses (43).

Representative components of this cellular defense system include members of the tripartite motif (TRIM) family (21). The best-studied family member, TRIM5α (31), restricts retroviral infection by specifically recognizing the viral capsid and promoting its premature disassembly (3, 20, 32), and, as recently reported, by blocking viral production at a posttranslational stage (23). Human TRIM5α has limited efficacy against human immunodeficiency virus type 1 (HIV-1), while proteins encoded by some primate *TRIM5*α orthologs can potently restrict this particular lentivirus (18, 28, 29, 33). Longstanding selective pressures exerted by retroviruses and retroelements may have contributed to the generation of diverse patterns of antiretroviral specificity of *TRIM5*α and other host defense genes (18, 28).

To better understand the evolution of antiretroviral specificity patterns in primates, in particular along the lineage leading to humans, we utilized a functional evolutionary genomics approach (34). In the present study, we reconstructed ancestral primate TRIM5α sequences and tested their specificity in vitro against HIV-1 and five other retroviruses. We used these six present day viruses as extant markers to evaluate the functional differences over evolutionary time.

MATERIALS AND METHODS

Determination of ancestral sequences. TRIM5α coding sequences from primates were obtained by amplification and sequencing of genomic DNA or cDNA or downloaded from the National Center for Biotechnology Information database (6, 12, 19, 25, 31, 38, 40): human (*Homo sapiens* AY625000), bonobo (*Pan paniscus* DQ229282), chimpanzee (*Pan troglodytes* AY923177), gorilla (*Gorilla gorilla* AY923178), Bornean orang-utan (*Pongo pygmaeus* AY923179), lar gibbon (*Hylobates lar* AY923180), nomascus (*Hylobates leucogenys* DQ229283), siamang (*Hylobates syndactylus* DQ229284), rhesus monkey (*Macaca mulatta* AY625001), olive baboon (*Papio anubis* AY843505), red guenon (*Erythrocebus patas* AY843514), African green monkey (*Cercopithecus* [*chlorocebus*] *aethiops* AY669399, *Cercopithecus* [*chlorocebus*] *tantalus* AY593973), eastern black-and-white colobus (*Colobus guereza* AY843507), douc langur (*Pygathrix nemaeus* AY843508), Bolivian titi (*Callicebus donacophilus* AY843519), Bolivian squirrel monkey (*Saimiri boliviensis boliviensis* AY928202), pygmy marmoset (*Callithrix pygmaea* AY843512), red-chested mustached tamarin (*Saguinus labiatus* AY843518), cotton-top tamarin (*Saguinus Oedipus* DQ229285), white-faced saki (*Pithecia pithecia* AY843515), Bolivian red howler monkey (*Alouatta sara* AY843511), common woolly monkey (*Lagothrix lagotricha* AY843520), and black-handed spider monkey (*Ateles geoffroyi* AY843516).

*TRIM5*α sequences were aligned by using CLUSTAL W. Coding regions were aligned according to their corresponding amino acid sequences by using the EMBOSS package (22). For the reconstruction of *TRIM5*α ancestral sequences (and for the calculation of posterior probabilities of reconstructed amino acids), we used a maximum-likelihood approach as implemented in the codeml tool (parameter settings according to model M0, i.e., model = 0 and NSsites = 0) of the PAML program package (37), in the framework of the accepted primate phylogeny (5). An alternative model (the free-ratio model, where each branch of the phylogeny may have a different K_A/K_S values) reconstructs the same amino acid variants at all sites and nodes, except for the single site 9 (I instead of V). However, since this site is located in the N terminus, which we demonstrate not

* Corresponding authors. Mailing address for A. Telenti: Institute of Microbiology, CHUV 1011 Lausanne, Switzerland. Phone: 41 21 314 05 50. Fax: 41 21 314 40 95. E-mail: amalio.telenti@chuv.ch. Mailing address for H. Kaessmann: University of Lausanne, Bâtiment Génopode, 1015 Lausanne, Switzerland. Phone: 41 21 692 39 07. Fax: 41 21 692 39 65. E-mail: Henrik.Kaessmann@unil.ch.
† V.G. and A.C. contributed equally to this study.
▽ Published ahead of print on 12 December 2007.

to affect anti-HIV-1 specificities of TRIM5α (see below), this ambiguity is not relevant for our conclusions. To validate the predictions from the maximum-likelihood approach, we used a Bayesian approach as implemented in the MrBayes program (7) using the Jones amino acid model and the known species topology as a prior (the topology around node 3 was constrained, i.e., it was fixed to reflect the divergence of the orangutan from the other great apes, as the Bayesian *TRIM5*α gene tree groups the orangutan sequence with gibbons—against the well-established primate phylogeny). Posterior probabilities were calculated on 100 samples representing the last 10,000 generations (sampling frequency 100) of a total of 100,000 generations. We verified that the log likelihoods of the ancestral reconstruction had converged before the last 10,000 generations.

Generation of cells stably expressing TRIM5α variants. pLPCX-TRIM5αhu-HA (NIH AIDS Reagent Program, donated by J. Sodroski), an oncoretroviral vector encoding the human TRIM5α (31), was used to generate the ancestral genes of *TRIM5*α by consecutive rounds of site-directed mutagenesis using the QuikChange protocol (Stratagene). *TRIM5*α from African green monkey was cloned from COS-7 (European Collection of Cell Cultures, ECACC no. 87021302), and *TRIM5*α tamarin was cloned from cells derived from cotton-top tamarin (ECACC no. 85011419). pLPCX-TRIM5αrh-HA (J. Sodroski) encodes the rhesus monkey TRIM5α. Expression of hemagglutinin-tagged TRIM5α proteins was determined by Western blotting.

Oncoretroviral vectors were packaged in 293T cells by cotransfecting the various pLPCX-TRIM5α constructs with the pNB-tropic murine leukemia virus (MLV) Gag-Pol and pVSV-G packaging plasmids (a gift from D. Trono) using the calcium phosphate method. As controls, we also used pLPCX and pLPCX-GFP instead of pLPCX-TRIM5α. Supernatants were concentrated and used to transduce 10^5 HeLa and CRFK (feline renal fibroblasts) cells in the presence of 5 μg of Polybrene/ml. At 72 h after transduction, cells were selected in the presence of 0.5 μg (HeLa cells) and 3 μg (CRFK cells) of puromycin/ml for at least 12 days before testing.

Recombinant virus infections. To produce HIV-1-based reporter vector particles, 293T cells (3×10^6 cells) were cotransfected with four plasmids by the calcium phosphate method (16). Plasmids encoded the vesicular stomatitis virus G protein pantropic envelope (pMD.G), the Gag and Pol proteins (pCMVΔR8.92), and Rev (pRSV-Rev), and the fourth plasmid encoded the HIV vector segment carrying green fluorescent protein (GFP) as the reporter transgene (pSIN.cPPT.EF1.GFP.WPRE). Additional constructs represent N-MLV and B-MLV (a gift from D. Trono), HIV-2 (a gift from A. Lever), and the simian immunodeficiency virus (SIV) variants SIV_{mac} (a gift from F.L. Cosset) and SIV_{agm} (a gift from J. Luban).

The infectivity of recombinant viruses was determined by titration on HeLa and CRFK cells. Single-round infectivity assays with the different recombinant viruses in HeLa and CRFK cells were performed at various multiplicities of infection (MOIs). At 48 h after transduction, cells were analyzed by using a fluorescence-activated cell sorter.

RESULTS AND DISCUSSION

Ancestral gene resurrection. To trace the evolutionary history of *TRIM5*α, we aligned the *TRIM5*α coding region from 24 primate species representing 25 million years of primate evolution. Evolutionary analysis allowed the estimation of the most likely sequence at each ancestral node in the framework of the accepted primate phylogeny (5) (Fig. 1A). We then introduced, by site-directed mutagenesis, all predicted amino acid changes at each ancestral node in a stepwise fashion starting from a cloned human *TRIM5*α (Fig. 1B). The oldest reconstructed sequence represents *TRIM5*α from the last common ancestor of Old World monkeys and apes, ~25 million years ago (mya), and differs by 29 amino acids from human TRIM5α. In addition, we tested *TRIM5*α from selected terminal taxa: Old World (rhesus monkey and African green monkey) and New World (cotton-top tamarin) monkeys. Attempts by our laboratory and other groups (28) to identify a *TRIM5*α ortholog in prosimians (e.g., lemurs and galagos) have been unsuccessful. Thus, a suitable outgroup that would permit us to infer the sequence of the common simian ancestor (i.e., Old World primate-catarrhine/New World monkey-platyrhine, 40 mya, Fig. 1A) is not available.

HIV-1 restriction by ancestral TRIM5α. To test the restriction capacity of the different gene variants, we first stably transduced HeLa cells (a human epithelial carcinoma cell line) with oncoretroviral vectors expressing the various ancestral and modern *TRIM5*α sequences. Multiple independently transduced cell lines were examined for each sequence in order to identify those showing comparable levels of transgene expression (Fig. 2A). We then infected the different cell lines with recombinant viruses—HIV-1, HIV-2, SIV_{agm}, SIV_{mac}, and the N- and B-tropic variants of the murine leukemia virus (N-MLV and B-MLV)—expressing the green fluorescent protein (GFP). The proportion of cells expressing GFP (relative infectivity) was then used as a proxy for the capacity of a particular *TRIM5*α sequence to restrict infection of cells by the respective recombinant virus.

We observed a statistically significant reduction of HIV-1 restriction by reconstructed TRIM5α variants (representing 25 million years of evolution) from the common catarrhine ancestral sequence toward the human variant. Ancestral TRIM5α variants are generally poorly restrictive of HIV-2, SIV_{agm} and SIV_{mac} (Fig. 2B). The pattern of HIV-1 restriction by successive ancestral variants was unique among the various retroviruses tested (Fig. 2C to G). Similar to the reconstructed node 1 sequence, TRIM5α from Old World monkeys displays a high capacity to restrict HIV-1. Some Old World monkey lineages appear to have acquired restriction capacity against other SIVs since the common catarrhine ancestor.

We carefully assessed the amino acid variants included in the reconstruction of each ancestral protein in the context of published experimental data of single and complex mutagenesis of TRIM5α (9, 14, 18, 31, 33, 39). The high HIV-1 restriction capacity of the 25 mya ancestral TRIM5α could not be deduced from the reconstructed sequence based on previous evidence from the literature. In particular, no ancestral construct carries the critical residue proline 332, associated with potent restriction of HIV-1 in previous studies (14, 39). The present study reconstructs position 332 with a glutamine (Q) in nodes 1 to 4 with high probability ($P > 0.95\%$; a substitution from glutamine to arginine occurred in the human-chimpanzee ancestor, between nodes 4 and 5, Fig. 1B). A glutamine at position 332 plays a role in the restriction activity of the gorilla B30.2 domain (18), and the change in restriction of HIV-1 between nodes 4 and 5 may be explained by a substitution from Q to R at this site (10). In general, the ability of human TRIM5α to bind the HIV-1 capsid is modulated by the presence of any charged residue at position 332 (14). However, despite the maintained presence of 332Q between nodes 1 and 4, HIV-1 restriction diminishes between these nodes. It should also be emphasized that the change in restriction occurred without any increase in length of the variable regions of the B30.2 domain (28, 33), since these regions were kept stable in size for the reconstruction of ancestral variants. Residue K389 was built as Q in nodes 4 to 1; however, recent analysis of extant sequences indicates that only nodes 3 to 1 should carry 389Q. Functional analysis indicates that this error is unlikely to modify restriction capacity (see below).

Reconstructed ancestral sequences also reveal amino acid substitutions in the other protein domains of TRIM5α. How-

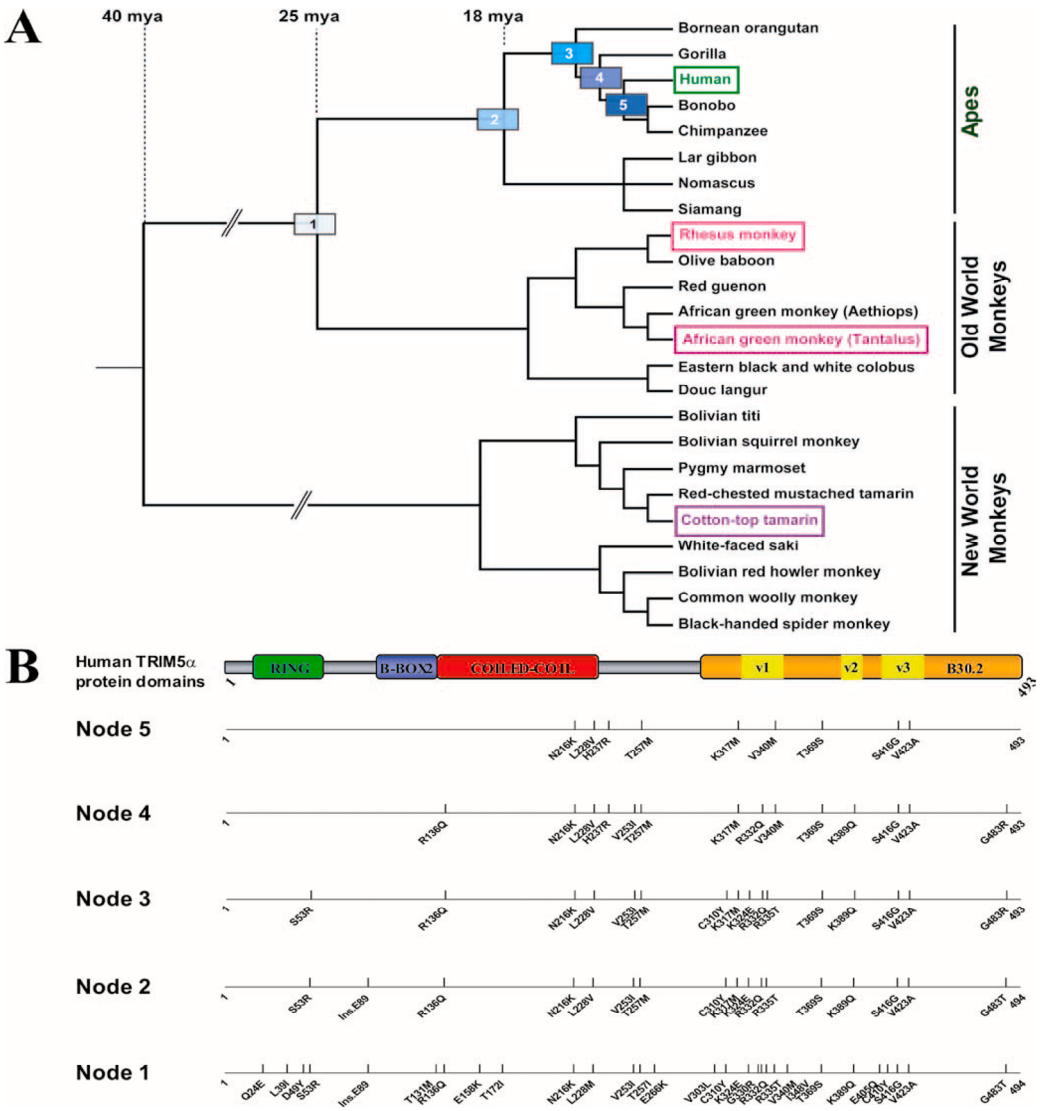


FIG. 1. Reconstruction of ancestral TRIM5α. (A) primate phylogenetic tree. Investigated nodes and species are color-coded. Approximate divergence times in millions of years (mya) are shown. (B) Identity and location of amino acid variants in ancestral nodes. Substitutions are shown relative to the human sequence. Although residue K389 is built as Q in nodes 4 to 1, recent analysis of extant sequences indicates that only nodes 3 to 1 should carry 389Q.

ever, single-amino-acid mutagenesis data from published functional studies (9) do not include any of the positions mutated here, with the exception of the R136Q substitution in the coiled-coil domain. This variant was shown not to modify TRIM5α restriction of HIV-1 (4, 24, 30), while in a study by Javanbahkt et al. the 136Q variant of human TRIM5α was associated with protection from HIV-1 infection in African Americans and with a weak restriction phenotype in vitro (8).

Restriction of other retroviruses. In contrast to the pattern seen for HIV-1, restriction of the murine oncoretrovirus N-

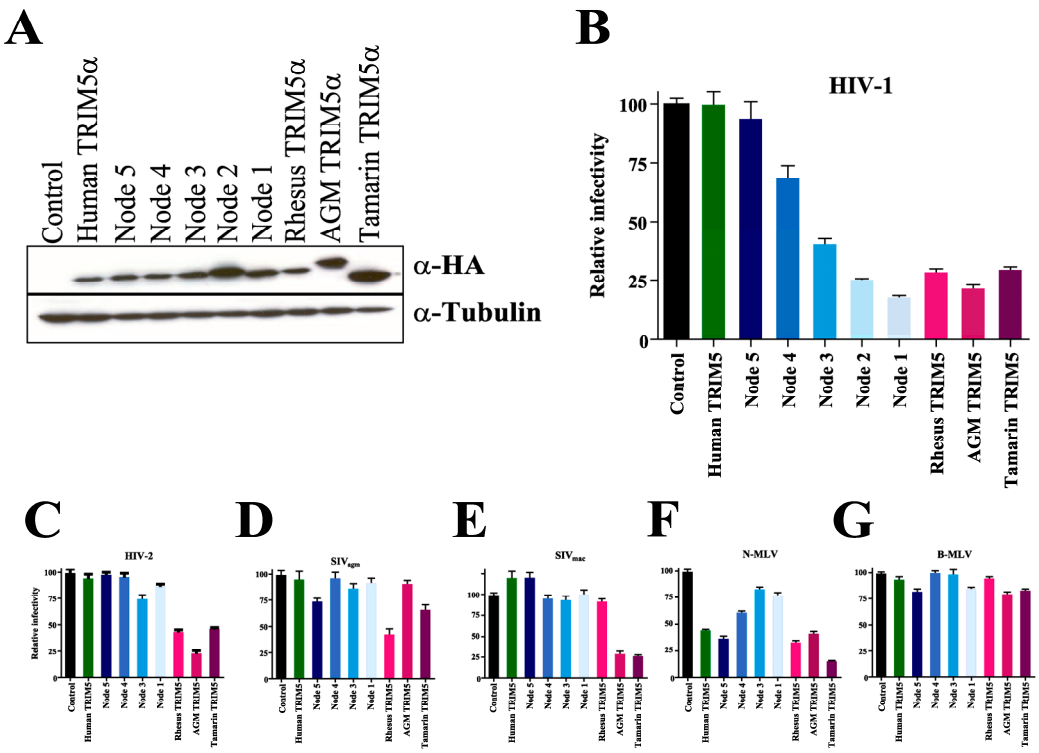


FIG. 2. Functional assessment of antiretroviral activity of ancestral and modern TRIM5α in HeLa cells. (A) Western blot assessment of stable expression of various TRIM5α in HeLa cells. (B) Restriction of HIV-1 recombinant virus in HeLa cells expressing the various reconstructed ancestral *TRIM5*α, as well as *TRIM5*α from humans, Old World monkeys (rhesus monkey, African green monkey [AGM]), and a New World monkey (cotton-top tamarin). Statistical analysis reveals significant restriction differences (one-way analysis of variance, $P < 10^{-3}$). Pairwise comparisons of nodes show significant restriction differences between nodes 1 to 5 ($P > 0.05$, Tukey's post hoc test), except for the node 1 and 2 comparison. Note that the restriction difference between node 1 and macaque, AGM, and tamarin TRIM5α variants is not statistically significant. (C to G) Restriction of five additional retroviruses in HeLa cells expressing reconstructed and extant *TRIM5*α. Node 2, which differs by two amino acids from node 3, is not shown. Shown are representative normalized results from infection at the optimal MOI for each virus. The error bars represent the standard error of the mean on three replicates.

MLV was stronger for reconstructed TRIM5α of extant taxa than for more ancestral TRIM5α constructs (Fig. 2F). The functional change was observed without changes in the RING domain, previously reported to be relevant for N-MLV restriction (24) (Fig. 1B). Specifically, ancestral nodes do not carry the substitution H43Y, a frequent human TRIM5α allele that may negatively affect its putative E3 ubiquitin ligase activity and is associated with impaired restriction of N-MLV (4, 24). The detrimental allele dates back to before the emergence of the African diaspora and is found at a frequency of 43% in indigenous Central and South Americans (24).

Ancestral TRIM5α showed various patterns of restriction capacities against other retroviruses tested (HIV-2, SIV_{agm}, SIV_{mac}, B-tropic variants of the MLV, B-MLV) (Fig. 2C to E and G). Overall, the 25-mya ancestral TRIM5α (node 1) exhibited limited efficacy against HIV-2, SIV_{agm}, and SIV_{mac}. Extant TRIM5α variants from Old World monkeys display higher restriction capacity against HIV-2 than the reconstructed TRIM5α from the common catarrhine ancestor (suggesting a gain in restriction capacity), whereas ancestral hominoid TRIM5αs all show little restriction efficiency with respect to this virus.

To rule out cell line-specific effects, we assessed the antiretroviral activities of selected ancestral and modern TRIM5α in another cell line (Fig. 3), CRFK (feline renal fibroblasts, no intrinsic retroviral restriction [6]). The CRFK series was less complete than that for HeLa cells, because we failed to generate stable cell lines carrying node 3 or cotton-top tamarin TRIM5α variants, whereas all tested clones carrying node 2 TRIM5α were massively overexpressing the protein. However, the overall pattern was confirmed: (i) decreasing restriction of modern HIV-1 from nodes 1 to 5, (ii) more restriction of N-MLV in humans and the human-chimpanzee ancestral TRIM5α, and (iii) greater restriction of HIV-2 by TRIM5α

TABLE 1. Identification of potentially ambiguous reconstructions and of residues responsible for restriction by ancestral TRIM5a[a]

Residue	Human TRIMα amino acid	Node 1 amino acid (probability)[b]		Positive selection	Node 1 amino acid (alternative amino acid selected for functional assays)
		ML	Bayes		
24	Q	E (0.93)	E (1.00)	–	None
39	L	I (0.94)	I (1.00)	–	None
49	D	Y (0.59)	H (1.00)	–	H
53	S	R (1.00)	R (1.00)	–	None
89		Insertion		–	Not assessed
131	T	M (0.84)	V (1.00)	–	V
136	R	Q (1.00)	Q (1.00)	–	None
158	E	K (0.79)	K (1.00)	–	None
172	T	I (0.83)	I (1.00)	–	None
216	N	K (1.00)	Q (0.99)	–	None
228	L	M (0.82)	M (0.54)	–	None
253	V	I (0.83)	V (0.99)	–	V (failure to build)
257	T	I (0.94)	I (1.00)	–	None
266	E	K (0.83)	K (0.99)	–	None
303	V	L (0.95)	L (1.00)	–	None
310	C	Y (0.94)	Y (0.76)	–	None
324	K	E (0.77)	E (0.98)	+	K
330	G	R (0.95)	P (0.85)	–	None
332	R	Q (0.98)	V (0.99)	+	R (failure to build)
335	R	T (0.96)	T (0.91)	+	R
340	V	M (0.42)	V (0.74)	+	V
348	I	V (0.85)	V (0.99)	–	None
369	T	S (1.00)	S (0.98)	–	None
389	K	Q (0.94)	G (0.90)	+	K
405	E	Q (0.95)	R (0.91)	–	None
410	C	Y (0.85)	Y (1.00)	–	None
416	S	G (0.73)	G (1.00)	–	None
423	V	A (1.00)	A (1.00)	–	None
483	G	T (0.72)	E (1.00)	+	E and G

[a] Among the 29 residues that differ between node 1 and human TRIM5α, we identified amino acid positions for which the maximum-likelihood procedure and an alternative Bayesian approach for TRIM5α reconstruction yielded potentially ambiguous reconstructions (i.e., relatively low posterior probabilities for reconstructed variants). Five residues showed low codeml probabilities (<95%), and an alternative amino acid was supported by Bayesian analysis. In addition, a number of residues that differed from human TRIM5α and were under positive selective pressure were identified as possibly responsible for the HIV-1 restriction capacity of the node1 ancestral TRIM5α.

[b] The maximum-likelihood (ML) posterior and Bayesian probabilities are given in parentheses.

based on the maximum-likelihood procedure, 5 showed low codeml probabilities (<95%), and an alternative amino acid was supported by Bayesian analysis (Table 1). We successfully tested four of the alternative amino acids (49Y/H, 131 M/V, 340 M/V, and 483T/E) by introducing them into the original node 1 sequence and failed to build 253V. None of the ambiguous-position variants modified the restriction of HIV-1 (Fig. 4), which ensures that potentially erroneously reconstructed residues likely do not account for the efficient restriction of HIV-1 by the node 1 TRIM5α sequence.

Another issue relates to the recent finding of long-term persistence of multiple TRIM5α alleles at critical positions due to balancing selection (17). Thus, reconstruction of specific nodes may be complicated by the simultaneous presence of key alleles in multiple lineages. Humans present features of balancing selection at position R136Q (4), while sooty mangabeys and rhesus monkeys (17) present multiple alleles at positions 182, 194, 213, 331, 332, indel 337-338, and 339 (human numbering). However, we analyzed the accuracy of reconstruction for these positions (except for indel 337-338, which is not present in all species and is thus not included in the analysis) and found that the ancestral inference at these sites appears to generally be unambiguous (with high posterior probabilities, >98%, for the reconstructed amino acid variant) and not affected by the putative trans-species polymorphisms. In addition, we searched novel TRIM5α submissions to GenBank, aligned available entries to the sequences from our study, and screened critical residues under positive selective pressure for the presence of polymorphism within the species. We re-estimated node 1 under the premise of balancing selection and found the original prediction unchallenged on the basis of currently available data on intraspecies polymorphism. Only residue 253 could have been built as 253V in addition to 253I at node 1, due to the fact that gorilla, orangutan, and nomascus primates are now known to be polymorphic at this position.

Identifying residues responsible for restriction by ancestral TRIM5α. After confirming the reliability of reconstruction of node 1, we sought to describe domains and residues that could contribute to node 1 restriction. For this, we generated reciprocal chimeric constructs of the N^{1-284}- and $C^{285-493}$-terminal portions of the node 1 and node 4 sequences. We then assessed their restriction capacities using CRFK cells (Fig. 4). This analysis demonstrates that only substitutions in the C-terminal domain appear to have affected restriction of HIV-1 by reconstructed TRIM5α on the human lineage (Fig. 4), since it is the hybrid construct with the node 1 C terminus that displays high restriction capacity (as high as that of the intact node 1 sequence).

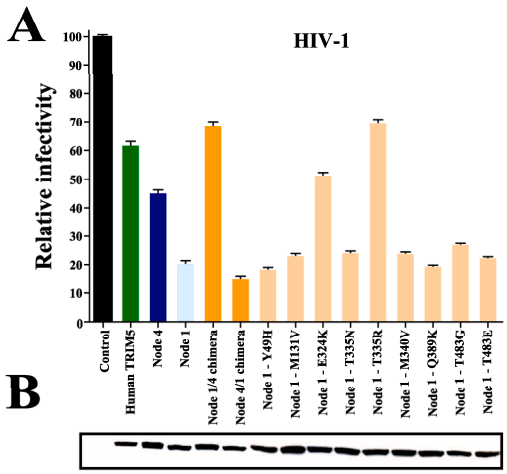


FIG. 4. Functional assessment of ambiguous ancestral variants and identification of residues responsible for restriction by ancestral TRIM5 α. (A) Restriction of HIV-1 recombinant virus in CRFK cells expressing *TRIM5*α from humans, selected reconstructed ancestral *TRIM5*α, N-terminal and C-terminal chimeras of nodes 4 and 1, and key mutants of node 1. Mutants 49H, 131V, 340V, and 483E test alternative reconstructions derived from Bayesian analysis. Mutants 324K, 335R, 340V, 389K, and 483G test the role of critical residues under positive selection by reintroducing the human TRIM5α amino acid variant into the node 1 sequence. Mutant 335N tests an alternative amino acid seen in African green monkeys at this position. Shown are normalized, representative results from infections at an MOI of 5. The error bars represent the standard error of the mean on three replicates. (B) Anti-HA immunoblot of the different TRIM5α-CRFK cell lines.

We then selected residues in the C-terminal region of node 1 that (i) differed from human TRIM5α and (ii) were under positive selective pressure (Table 1). Using site-directed mutagenesis experiments, we demonstrate that changing the ancestral 324E to the more recent 324K (a substitution occurring on the lineage leading to the last common African ape ancestor, node 4) or changing the ancestral 335T residue to 335R (originating in the common human/chimpanzee ancestor on the lineage leading to node 5) strongly reduces the ability of the 25-mya ancestral TRIM5α to restrict HIV-1. These positions have been previously investigated as 324N and 335L and found to be associated with restriction of HIV-1 (14, 33). In addition, a sequence context that includes 324E/332Q/335T in variable region v1 and 389Q in the v2 region was associated with effective restriction in a recent study (18). For amino acid residue 335, we tested another amino acid (335N) found in Old World monkeys. The resulting sequence variant maintains the high restriction capacity seen for the original node 1 sequence (Fig. 4), which underscores that the loss of restriction capacity is specific to reversion to 335R. Thus, the present study reveals novel functionally relevant amino acid variants at positions 324 and 335 and generally confirms the importance of these sites for virus restriction by TRIM5α. It is noteworthy, however, that although this sequence context is maintained from nodes 1 to node 3, restriction efficacy is gradually diminishing between these nodes. This suggests that other substitutions and/or sites (potentially in combination) explain the diminishing restriction capacity of reconstructed TRIM5α compared to reconstructed TRIM5α of the last common Old World primate ancestor (node 1). We individually tested three additional residues (340V, 389K, and 483G) that were previously shown to have evolved under positive selection (Table 1) in this domain, but none of these substitutions appear to reduce the restriction of the original node 1 TRIM5α protein when changed individually (Fig. 4). We did not assess the ambiguous position C385, which corresponds to the V2 region of length polymorphism. Substitution of tyrosine with cysteine at position 385 does not modify the restriction of HIV-1 (18).

The novelty in the present study resides in the experimental approach used to characterize recently emerged protein differences (reconstruction of ancestral host sequences). The strategy was able to identify and experimentally test new variants despite the fact that this protein has been very thoroughly investigated over the past 3 years. Thus, evolutionary genetics and ancestral reconstruction could be of interest in future investigation of less-well-characterized proteins. Interestingly, as for TRIM5α, evolutionary analysis of a second antiretroviral protein, APOBEC3G (27), would also predict shifts in antiretroviral specificity on the human lineage since the catarrhine ancestor, because the ancestral APOBEC3G likely carried the amino acid variant K128 that governs sensitivity of this protein to modern HIV-1 Vif-mediated inhibition (2, 15, 26, 36). There is clearly a great level of complexity in imputing an evolutionary direction of restriction of modern viruses by reconstructed TRIM5α variants. However, ancestral proteins can in the future be tested against resurrected infectious agents (10, 13).

Such various restriction capacities might be explained by lineage-specific pandemics that shape and redirect the intrinsic defense mechanisms. For example, the recent retroviral infection of two great ape lineages—chimpanzees and gorillas—by horizontal transmission from an exogenous source of PtERV1 has not affected the human genome (41). Recent data indicate that the R332 mutation in human and chimpanzee TRIM5α improves the ability of this protein to restrict PtERV1 while resulting in increased susceptibility to HIV-1 (10). Thus, as a result of trade-offs in the ability to restrict different retroviruses, humans might have been exposed to SIV_{cpz} and/or HIV-1 at a time when one or several intrinsic defense proteins lacked the appropriate specificity to avert its transmission from chimpanzees.

ACKNOWLEDGMENTS

We thank D. Trono, A. Lever, F. L. Cosset, J. Luban, and G. Towers for reagents.

A.T. and H.K. conceived the study. A.T., V.G., and A.C. designed the experiments. V.G., A.C., M.O., and M.M. performed the experiments. MO., D.B., and H.K. performed the evolutionary analysis. A.T., H.K., V.G., and A.C. wrote the report.

This study was funded by the Swiss National Science Foundation and a grant for interdisciplinary research from the Faculty of Biology and Medicine of the University of Lausanne.

REFERENCES

1. **Bieniasz, P. D.** 2004. Intrinsic immunity: a front-line defense against viral attack. Nat. Immunol. **5:**1109–1115.
2. **Bogerd, H. P., B. P. Doehle, H. L. Wiegand, and B. R. Cullen.** 2004. A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor. Proc. Natl. Acad. Sci. USA **101:**3770–3774.
3. **Chatterji, U., M. D. Bobardt, P. Gaskill, D. Sheeter, H. Fox, and P. A. Gallay.** 2006. Trim5α accelerates degradation of cytosolic capsid associated with productive HIV-1 entry. J. Biol. Chem. **281:**37025–37033.
4. **Goldschmidt, V., G. Bleiber, M. T. May, R. Martinez, M. Ortiz, and A. Telenti.** 2006. Role of common human TRIM5alpha variants in HIV-1 disease progression. Retrovirology **3:**54.
5. **Goodman, M.** 1999. The genomic record of Humankind's evolutionary roots. Am. J. Hum. Genet. **64:**31–39.
6. **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 TRIM5α. Proc. Natl. Acad. Sci. USA **101:**10774–10779.
7. **Huelsenbeck, J. P., and F. Ronquist.** 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17:**754–755.
8. **Javanbakht, H., P. An, B. Gold, D. C. Petersen, C. O'Huigin, G. W. Nelson, S. J. O'Brien, G. D. Kirk, R. Detels, S. Buchbinder, S. Donfield, S. Shulenin, B. Song, M. J. Perron, M. Stremlau, J. Sodroski, M. Dean, and C. Winkler.** 2006. Effects of human TRIM5α polymorphisms on antiretroviral function and susceptibility to human immunodeficiency virus infection. Virology **354:** 15–27.
9. **Javanbakht, H., F. az-Griffero, M. Stremlau, Z. Si, and J. Sodroski.** 2005. The contribution of RING and B-box 2 domains to retroviral restriction mediated by monkey TRIM5α. J. Biol. Chem. **280:**26933–26940.
10. **Kaiser, S. M., H. Malik, and M. Emerman.** 2007. Restriction of an extinct retrovirus by the human TRIM5α antiviral protein. Science **316:**1756–1758.
11. **Kaumanns, P., I. Hagmann, and M. T. Dittmar.** 2006. Human TRIM5α mediated restriction of different HIV-1 subtypes and Lv2 sensitive and insensitive HIV-2 variants. Retrovirology **3:**79.
12. **Keckesova, Z., L. M. Ylinen, and G. J. Towers.** 2004. The human and African green monkey TRIM5α genes encode Ref1 and Lv1 retroviral restriction factor activities. Proc. Natl. Acad. Sci. USA **101:**10780–10785.
13. **Lee, Y. N., and P. D. Bieniasz.** 2007. Reconstitution of an infectious human endogenous retrovirus. PLoS. Pathog. **3:**e10.
14. **Li, Y., X. Li, M. Stremlau, M. Lee, and J. Sodroski.** 2006. Removal of arginine 332 allows human TRIM5α to bind human immunodeficiency virus capsids and to restrict infection. J. Virol. **80:**6738–6744.
15. **Mangeat, B., P. Turelli, S. Liao, and D. Trono.** 2004. A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action. J. Biol. Chem. **279:**14481–14483.
16. **Naldini, L., U. Blomer, P. Gallay, D. Ory, R. Mulligan, F. H. Gage, I. M. Verma, and D. Trono.** 1996. In vivo gene delivery and stable transduction of nondividing cells by a lentiviral vector. Science **272:**263–267.
17. **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 TRIM5α. Proc. Natl. Acad. Sci. USA.
18. **Ohkura, S., M. W. Yap, T. Sheldon, and J. P. Stoye.** 2006. All three variable

regions of the TRIM5α B30.2 domain can contribute to the specificity of the retrovirus restriction. J. Virol. **80:**8554–8565.
19. **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.
20. **Perron, M. J., M. Stremlau, M. Lee, H. Javanbakht, B. Song, and J. Sodroski.** 2007. The human TRIM5α restriction factor mediates accelerated uncoating of the N-tropic murine leukemia virus capsid. J. Virol. **81:**2138–2148.
21. **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 J. **20:**2140–2151.
22. **Rice, P., I. Longden, and A. Bleasby.** 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. **16:**276–277.
23. **Sakuma, R., J. A. Noser, S. Ohmine, and Y. Ikeda.** 2007. Rhesus monkey TRIM5α restricts HIV-1 production through rapid degradation of viral Gag polyproteins. Nat. Med. **13:**631–635.
24. **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 TRIM5α in humans. Curr. Biol. **16:**95–100.
25. **Sawyer, S. L., L. I. Wu, M. Emerman, and H. S. Malik.** 2005. Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. Proc. Natl. Acad. Sci. USA **102:**2832–2837.
26. **Schrofelbauer, B., D. Chen, and N. R. Landau.** 2004. A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif). Proc. Natl. Acad. Sci. USA **101:**3927–3932.
27. **Sheehy, A. M., N. C. Gaddis, J. D. Choi, and M. H. Malim.** 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature **418:**646–650.
28. **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 TRIM5α exhibits lineage-specific length and sequence variation in primates. J. Virol. **79:**6111–6121.
29. **Song, B., H. Javanbakht, M. Perron, D. H. Park, M. Stremlau, and J. Sodroski.** 2005. Retrovirus restriction by TRIM5α variants from Old World and New World primates. J. Virol. **79:**3930–3937.
30. **Speelmon, E. C., D. Livingston-Rosanoff, S. S. Li, Q. Vu, J. Bui, D. E. Geraghty, L. P. Zhao, and M. J. McElrath.** 2006. Genetic association of the antiviral restriction factor TRIM5α with human immunodeficiency virus type 1 infection. J. Virol. **80:**2463–2471.
31. **Stremlau, M., C. M. Owens, M. J. Perron, M. Kiessling, P. Autissier, and J. Sodroski.** 2004. The cytoplasmic body component TRIM5α restricts HIV-1 infection in Old World monkeys. Nature **427:**848–853.
32. **Stremlau, M., M. Perron, M. Lee, Y. Li, B. Song, H. Javanbakht, F. az-Griffero, D. J. Anderson, W. I. Sundquist, and J. Sodroski.** 2006. Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5α restriction factor. Proc. Natl. Acad. Sci. USA **103:**5514–5519.
33. **Stremlau, M., M. Perron, S. Welikala, and J. Sodroski.** 2005. Species-specific variation in the B30.2(SPRY) domain of TRIM5α determines the potency of human immunodeficiency virus restriction. J. Virol. **79:**3139–3145.
34. **Thornton, J. W.** 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. Nat. Rev. Genet. **5:**366–375.
35. **Williams, P. D., D. D. Pollock, B. P. Blackburne, and R. A. Goldstein.** 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS. Comput. Biol. **2:**e69.
36. **Xu, H., E. S. Svarovskaia, R. Barr, Y. Zhang, M. A. Khan, K. Strebel, and V. K. Pathak.** 2004. A single amino acid substitution in human APOBEC3G antiretroviral enzyme confers resistance to HIV-1 virion infectivity factor-induced depletion. Proc. Natl. Acad. Sci. USA **101:**5652–5657.
37. **Yang, Z.** 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:**555–556.
38. **Yap, M. W., S. Nisole, C. Lynch, and J. P. Stoye.** 2004. Trim5α protein restricts both HIV-1 and murine leukemia virus. Proc. Natl. Acad. Sci. USA **101:**10786–10791.
39. **Yap, M. W., S. Nisole, and J. P. Stoye.** 2005. A single amino acid change in the SPRY domain of human Trim5α leads to HIV-1 restriction. Curr. Biol. **15:**73–78.
40. **Ylinen, L. M., Z. Keckesova, S. J. Wilson, S. Ranasinghe, and G. J. Towers.** 2005. Differential restriction of human immunodeficiency virus type 2 and simian immunodeficiency virus SIVmac by TRIM5α alleles. J. Virol. **79:**11580–11587.
41. **Yohn, C. T., Z. Jiang, S. D. McGrath, K. E. Hayden, P. Khaitovich, M. E. Johnson, M. Y. Eichler, J. D. McPherson, S. Zhao, S. Paabo, and E. E. Eichler.** 2005. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. PLoS. Biol. **3:**e110.
42. **Zhang, F., T. Hatziioannou, D. Perez-Caballero, D. Derse, and P. D. Bieniasz.** 2006. Antiretroviral potential of human tripartite motif-5 and related proteins. Virology **353:**396–409.
43. **Zheng, Y. H., and B. M. Peterlin.** 2005. Intracellular immunity to HIV-1: newly defined retroviral battles inside infected cells. Retrovirology **2:**25.

4.4.2 Original article

Role of common human TRIM5α variants in HIV-1 disease progression

Valérie Goldschmidt [†1], Gabriela Bleiber[†1], Margaret May[2] Raquel Martinez[1], **Millán Ortiz[1]**, Amalio Telenti[*1] and The Swiss HIV Cohort Study

Address: [1]Institute of Microbiology and University Hospital, University of Lausanne, Switzerland and [2]Department of Social Medecine, University of Bristol, UK

Retrovirology 2006 Aug 22;3:54 [66].

Retrovirology

BioMed Central

Research

Open Access

Role of common human TRIM5α variants in HIV-1 disease progression

Valérie Goldschmidt[†1], Gabriela Bleiber[†1], Margaret May[2], Raquel Martinez[1], Millàn Ortiz[1], Amalio Telenti[*1] and The Swiss HIV Cohort Study

Address: [1]Institute of Microbiology and University Hospital, University of Lausanne, Switzerland and [2]Department of Social Medicine, University of Bristol, UK

Email: Valérie Goldschmidt - Valerie.Goldschmidt@chuv.ch; Gabriela Bleiber - Gabriela.x.Bleiber@gsk.com; Margaret May - M.T.May@bristol.ac.uk; Raquel Martinez - Raquel.Martinez@chuv.ch; Millàn Ortiz - Millan.Ortiz-Serrano@chuv.ch; Amalio Telenti* - Amalio.Telenti@chuv.ch; The Swiss HIV Cohort Study - martin.rickenbach@chuv.ch

* Corresponding author †Equal contributors

Published: 22 August 2006

Retrovirology 2006, 3:54 doi:10.1186/1742-4690-3-54

Received: 10 April 2006
Accepted: 22 August 2006

This article is available from: http://www.retrovirology.com/content/3/1/54

Abstract

Background: The retroviral restriction factor tripartite motif protein (TRIM)5α, is characterized by marked amino acid diversity among primates, including specific clusters of residues under positive selection. The identification of multiple non-synonymous changes in humans suggests that TRIM5α variants might be relevant to retroviral pathogenesis. Previous studies have shown that such variants are unlikely to modify susceptibility to HIV-1 infection, or the course of early infection. However, the longterm effect of carrying Trim5α variants on disease progression in individuals infected with HIV-1 has not previously been investigated.

Methods: In a cohort of 979 untreated individuals infected with HIV-1 with median follow up 3.2 years and 9,828 CD4 T cell measurements, we analysed common amino acid variations: H43Y, V112F, R136Q, G249D, and H419Y. The rate of CD4 T cell decline before treatment was used as the phenotype. In addition, we extended previous work on the *in vitro* susceptibility of purified donor CD4 T cells (n = 125) to HIV-1 infection, and on the susceptibility of HeLa cells that were stably transduced with the different *TRIM5* variants. Haplotypes were analysed according to the most parsimonious evolutionary structure, where two main human TRIM5α groups can be defined according to the residue at amino acid 136. Humans present both Q136 and R136 at similar frequency, and additional TRIM5α amino acid variants are almost exclusively derived from R136-carrying haplotypes.

Results: We observed modest differences in disease progression for evolutionary branches carrying R136-derived haplotypes, and with the non-synonymous polymorphisms G249D and H419Y. *In vitro* analysis of susceptibility of donor CD4 T cells, and of the various transduced HeLa cell lines supported the absence of significant differential restriction of HIV-1 infection by the various huTRIM5α alleles.

Conclusion: Common human variants of TRIM5α have no effect or modest effect on HIV-1 disease progression. These variants occur at sites conserved throughout evolution, and are remote from clusters of positive selection in the primate lineage. The evolutionary value of the substitutions remains unclear.

Background

The tripartite motif (TRIM) family is a well conserved family of proteins characterized by a structure comprising a RING domain, one or two B-boxes and a predicted coiled-coil region [1]. In addition, most TRIM proteins have additional C-terminal domains. Members of the TRIM protein family are involved in various cellular processes, including cell proliferation, differentiation, development, oncogenesis and apoptosis (for recent review [2,3]). Some TRIM proteins display antiviral properties, targeting retroviruses in particular [4].

TRIM5α is a retroviral restriction factor targeting the early steps of cellular infection [4]. TRIM5α restricts retroviral infection by specifically recognizing the capsid and promotes its premature disassembly [5]. Human TRIM5α (huTRIM5α) has limited efficacy against HIV-1, while some primate TRIM5α orthologues can potently restrict this particular lentivirus (for review see [2,3]). Considerable inter-species sequence diversity characterizes TRIM5α and might underlie differences in the pattern and breadth of restriction of multiple lentiviruses. Evolutionary analysis reveals that up to 2% of codons of TRIM5α are predicted to be under positive selection with high confidence [6,7]. Residues under positive selection cluster in the C-terminal B30.2 domain. A first cluster resides between amino acids 322 to 340 in the variable region v1 [7,8], a region previously described as a "patch" of positive selection [6]. Replacement of the v1 region, or of specific amino acids within v1, modifies the restriction pattern of TRIM5α [9,10]. The second cluster, localized between amino acids 381 to 389 [7], corresponds to the previously described variable region v2 of the B30.2 domain [8]. Substitution of the human v2 region by a v2 from Rhesus monkeys exhibits no inhibitory activity against HIV-1 [9,10]. However, v2 variants are thought to result in species-specific lentiviral restriction patterns [11]. An additional region of considerable variation among Sooty mangabeys and Rhesus monkeys has been mapped to the coiled-coil motif [12].

Recently, two studies have addressed the potential role of huTRIM5α variants in modulating susceptibility to HIV-1 [13,14]. Sawyer et al. identified several non-synonymous SNPs in huTRIM5α, but only one of these (H43Y) was found to have a functional consequence [13]. H43Y lies in the RING domain of TRIM5α and may negatively affect its putative E3 ubiquitin ligase activity. Although huTRIM5α weakly restricts HIV-1, H43Y might further reduce viral restriction to a level similar to that of cells expressing no exogenous huTRIM5α [13]. To assess whether the impaired retroviral restriction seen with exogenously expressed huTRIM5α H43Y resulted in altered susceptibility in human cells, Sawyer et al. tested B-lymphocytes from four individuals: one homozygous for H43, two homozygous for 43Y, and one heterozygous at this residue. Challenge with HIV-1 failed to demonstrate a significant effect of the H43Y change, although 43Y homozygous cells could be infected with N-MLV about 100-fold more efficiently than cells with the common H43 allele [13]. In a second study, Speelmon et al. assessed the association of various non-synonymous variants with susceptibility to HIV-1 among 110 HIV-1 infected subjects and 96 exposed seronegative persons [14]. This study identified possible associations between specific haplotypes and alleles and susceptibility to infection and viral setpoint after acute infection.

In our study, we analysed data from a large cohort of subjects infected with HIV-1 to explore whether different huTRIM5α variants are associated with long-term disease evolution. The study is complemented by analysis of CD4 cell susceptibility to HIV-1, and the *in vitro* functionality of selected huTRIM5α variants. We conclude from our study that there is no impact for some and negligible to modest impact for other common human variants of TRIM5α on disease progression.

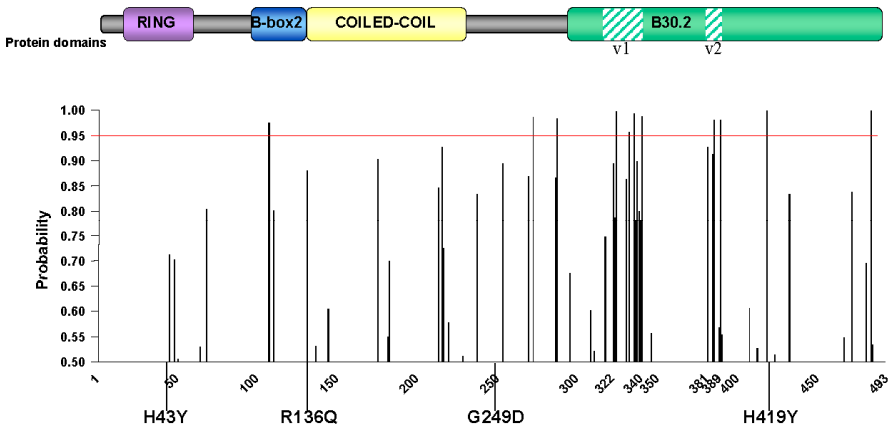
Results and discussion

TRIM5α polymorphism

TRIM5α is characterized by important sequence diversity in humans, as shown in this and in previous studies [13,14]. Analysis of huTRIM5α polymorphism in blood donors identified 21 genetic variants (Additional file 1), including four SNPs leading to non-synonymous changes: 127C>T (rs3740996, H43Y, allelic frequency f = 0.11), 407G>A (rs10838525, R136Q, f = 0.35), 12468G>A (rs11038628, G249D, f = 0.08), and 15142C>T (rs28381981, H419Y, f = 0.05). One additional common variant (rs11601507, V112F, f = 0.08), and several rare non-synonymous variants have been described in the other studies [13,14]. Changes involve evolutionary conserved positions(Figure 1). None of these changes are within patches of positive selection in primates, in particular none are in the proximity of variable regions v1 and v2 of the B30.2 domain (Figure 1), and thus their evolutionary significance is uncertain. We speculate that R136Q might result from balancing selection because humans carry Q136, the ancestral amino acid, and R136, an amino acid shared only with chimpanzees, with similar frequencies.

Association of genetic variants with in vitro *susceptibility to infection*

Data from Sawyer et al. indicated that H43Y results in reduced capacity to restrict N-MLV, but has minimal or no impact on HIV-1 susceptibility to infection *in vitro* in feline fibroblasts (CRFK) cells [13]. We extend and confirm these results by showing that HeLa cells stably transduced with the different human variants of *TRIM5* do not



Human	H	R	G	H
Bonobo	H	R	G	H
Chimpanzee	H	R	G	H
Gorilla	H	Q	G	H
Orangutan	H	Q	G	H
Lar Gibbon	H	Q	G	H
Nomascus	H	Q	G	H
Siamang	H	Q	G	H
Rhesus monkey	H	Q	G	H
African Green Monkey	H	Q	G	H
Tamarin	H	E	G	S

Figure 1
Position of common human TRIM5α amino acid variants in the context of primate sequence conservation and of the v1 and v2 patches of positive pressure. Y-axis: posterior probabilities of positively selected codons. X-axis: human TRIM5α amino acid numbering. The evolutionary analysis is adapted from reference [7].

differ in susceptibility to HIV-1 infection (Figure 2). We also confirmed that the H43Y variant failed to restrict N-MLV in the HeLa background (Additional file 2). Thus, results in Hela cells, that express TRIM5α endogenously, are in full concordance with data from CRFK cells. HeLa cells have the most common TRIM5 alleles (-2CC, H43, V112, heterozygous R136Q, G249, H419).

Speelmon et al. [14] reported on the permissiveness of purified CD4 T cells from 77 seronegative donors. Analysis included assessment of H43Y, R136Q, H419Y, and a series of less common variants. We performed similar experiments by infecting purified CD4 T cells from 125 Caucasian healthy blood donors with replicating HIV-1. Alleles tested included H43Y, R136Q, H419Y, and the common variant G249D (not tested in the above study). There was no significant association of specific variants or haplotypes with *in vitro* p24 production after 7 days (Additional file 3). None of the additional SNPs investigated *in vitro* were associated with differences in cell permissiveness (Additional file 1).

Association of genetic variants with disease progression in vivo

The reports by Sawyer et al and Speelmon et al. [13,14] suggested that some of the alleles could indeed have an impact on HIV-1 susceptibility *in vivo*. We extended their analyses by investigating the association of the various huTRIM5α variants with long-term disease progression in a large cohort of individuals infected with HIV-1. The clinical phenotype was defined as the patient-specific rate of CD4 T cell decline, a recognized marker of disease progression [15]. Analysis excluded any CD4 T cell values after initiation of treatment. The median follow up time was 3.2 years, during which 979 cohort participants, not receiving antiretroviral treatment, contributed 9,828 CD4

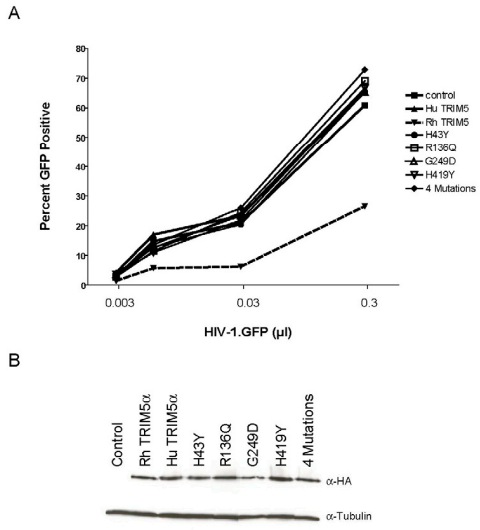


Figure 2

Restriction of HIV-1 by common human TRIM5α variants. HeLa cells were stably transduced by oncoretroviral vectors expressing the Rhesus (Rh) TRIM5α, the common huTRIM5α and its variants, separately or in a hypothetical four-mutation protein. Panel A: Single-cycle infectivity assays used VSV-pseudotyped recombinant viruses (HIV.1-GFP) at various m.o.i. After 48 h, cells were analysed by fluorescence-activated cell sorter, and scored for number of GFP-positive cells. Panel B: Expression of HA-tagged TRIM5α proteins was determined by western blotting using an anti-HA antibody. Tubulin was detected with the anti-α tubulin antibody. Control: HeLa cells transduced with an empty oncoretroviral vector.

T cell determinations to the analysis (median 7 CD4 T cell determinations per participant). We first tested for associations of individual non-synonymous polymorphisms with differences in the natural history of disease progression. H43Y and R136Q had no effect on disease progression. Participants who had one or two copies of G249D or H419Y had slower progression although confidence intervals were wide due to small numbers. Compared to non-carriers who had a square root transformed CD4 gradient of -2.02, participants who were carriers of G249D and H419Y had gradients of -1.74 (95% CI -1.39 to -2.09, p = 0.11) and -1.64 (95% CI -1.33 to -1.95, p = 0.02) respectively.

Haplotypes were assessed according to the most parsimonious evolutionary analysis, where two main huTRIM5α groups can be defined according to the residue at amino acid 136 (Figure 3). With the notable exception of chimpanzees and humans that carry an arginine at position 136, all other primates code for a glutamine at codon 136 (glutamic in tamarins), which therefore represents the ancestral sequence for old world monkeys, gibbons and apes. However, humans have similar frequencies of Q136 and R136, and additional TRIM5α amino acid variants are almost exclusively derived from R136-carrying haplotypes (data from this study and from [14]). We did not observe differences in HIV-1 disease progression for evolutionary branches carrying Q136- or R136-containing haplotypes (Figure 3). Weak associations of some haplotypes with disease progression were found, but the p values did not reach the experiment wide corrected significance level of p = 0.0028 (Simes modified Bonferroni p value).

Whilst our study was being completed, Sawyer et al. [13] and Speelmon et al [14] reported on an additional common variant V112F (allelic frequency of 7%). We re-genotyped the cohort and identified the presence of 112F in R136-carrying haplotypes. We confirm the absence of significant effect of this particular amino acid on disease progression [CD4 T cell gradient: -2.21 (95% CI -1.89 to -2.534) compared to mean reference slope -2.01]. In addition, Speelmon et al. included in their analysis the 5'UTR -2C>G SNP [14]. The -2C represents the ancestral allele and is present in high linkage disequilibrium with Q136. Speelmon et al. identified a rare haplotype, where individuals carried -2G in the context of Q136, possibly associated with susceptibility to HIV infection (OR 5.49, p = 0.02) [14]. We also identified this rare haplotype (frequency of 1.4%), but did not observe any association with disease progression [CD4 T cell gradient: -2.04 (95% CI -1.20 to -2.88)].

Conclusion

We have extended previously reported findings on huTRIM5α by showing that the amino acid variant H43Y which results in loss of restriction of N-MLV [13] is not associated with significant differences in HIV-1 disease progression in a large human cohort. The present study also underscores the modest or negligible effect of human variants G249D and H419Y and of some haplotypes on HIV-1 susceptibility and disease progression. Despite the conserved nature of these residues in primates, the evolutionary relevance of the variants in humans is uncertain. However, it is possible that the polymorphisms found in TRIM5α might have been selected in past epidemics by viruses unrelated to HIV-1.

Materials and methods

Cells

CD4 T cells from 125 healthy Caucasian blood donors were isolated by anti-CD4 magnetic beads (Miltenyi Biotech) and cultured *ex vivo* in RPMI-1640 (Gibco-Invitro-

A

Haplotype	127C>T	407G>A	12468G>A	15142C>T	Haplotype Frequency
H1	C	G	G	C	**0.45**
H2	C	G	G	T	**0.01**
H3	C	G	A	C	**0.08**
H4	C	A	G	C	**0.34**
H5	C	A	G	T	**<0.01**
H6	C	A	A	C	**<0.01**
H7	T	G	G	C	**0.07**
H8	T	G	G	T	**0.04**
H9	T	G	A	C	**<0.01**
H10	T	A	G	C	**<0.01**
Allele Frequency	**0.11**	**0.35**	**0.08**	**0.05**	
Aminoacid change	H43Y	R136Q	G249D	H419Y	

B

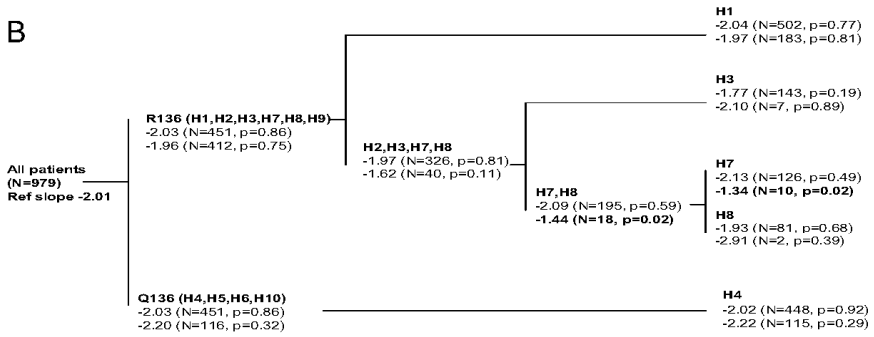


Figure 3

Association of human TRIM5α haplotypes with HIV-1 disease progression *in vivo*. Panel A, Inferred haplotypes carrying non-synonymous variants. Panel B, Analysis of haplotypes according to the most parsimonious evolutionary analysis, where two main huTRIM5α groups can be defined according to the residue at codon 136. Shown are square root CD4 gradient (reference slope for all patients is -2.01). Top set of slope estimates corresponds to one copy of haplotype(s) group, bottom set is for two copies of haplotype(s) group. Only the H7 haplotype presented a slope significantly different from that of all patients (uncorrected p value). However, p values did not reach the experiment-wide significance of 0.003 (Simes modified Bonferroni p value).

gen), supplemented with 20% fetal calf serum (FCS), 20 U/ml human IL-2 (Roche) and 50 μg/ml gentamicin, following stimulation with 2 μg/ml phytohaemagglutinin (PHA) during two days. CD4 T cells (10^6 cells) were infected with R5 clone HIV-1 NL4-3Ba*Lenv* (1000 pg p24 antigen) for 2 hours at 37°C, 5% CO_2, in 1 ml final volume. Cells were washed and cultured for 7 days. Virus-containing supernatant was harvested and p24 antigen production was monitored by ELISA (Abbott). Permissiveness was defined as the ability of cells to be infected and sustain replication of HIV-1 [16]. The *ex vivo* viral replication for each genotype was represented by the median p24 antigen production at day 7.

Identification of SNPs, and allelic discrimination

Single nucleotide polymorphism (SNP) discovery used single strand conformation polymorphism and sequencing of 94 chromosomes (47 Caucasian blood donors). For this, a total of 21 PCR reactions were designed to cover exons, putative promoter regions, and intron-exon

boundaries (6771 bp/subject). SNPs resulting in non-synonymous substitutions were then genotyped by using TaqMan allelic discrimination (Additional file 4).

Biological analysis of huTRIM5α variants

The pLPCX oncoretroviral vector containing the human and Rhesus *TRIM5α* gene with an HA epitope tag was obtained from the NIH AIDS Reagent Program (donated by J. Sodroski). Variants of huTRIM5α were made by using the QuikChange protocol (Stratagene). Retroviral vectors were packaged by co-transfecting the various pLPCX constructs with the pNB-tropic MLV Gag-Pol and pVSV-G packaging plasmids [17]. Supernatants were concentrated and used to transduce HeLa cells. Seventy two hours after transduction, cells were selected in 0.5 mg/ml puromycin. Expression of HA-tagged TRIM5α proteins were determined by Western blotting using an anti-HA antibody (Roche). Tubulin was detected with the anti-α tubulin antibody (Sigma). Single-cycle infectivity assays in HeLa cells used the VSV-pseudotyped recombinant viruses HIV.1-GFP and N-MLV-GFP at various m.o.i. Cells were analysed by fluorescence-activated cell sorter (FACS) 48 h after transduction.

In vivo *analysis: CD4 cell count decline*

Study participants (n = 979) were recruited within the genetics project of the Swiss HIV Cohort Study [18]. The ethics committees of all participant centers approved the study. Patients gave written informed consent for genetic testing. DNA from PBMCs was used for genotyping. Their characteristics are shown in Additional file 5. The rate of decline in CD4 T cell count during the natural history of disease progression was defined as study phenotype as previously reported [19]. The CD4 T cell trajectories were modeled using a repeated measures hierarchical approach using Mlwin software [20]. Square root transformed CD4 T cell counts were modeled as a linear function of time since estimated date of seroconversion with random effects for both the intercept and the gradient with additional terms for sex, age, and risk group [19]. For each genotype, the average square root CD4 decline per year was estimated in dominant and recessive models. Haplotypes were attributed using PHASE [21].

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

VG and GB carried out re-sequencing, genotyping studies, and construction of expression vectors. VG performed the transduction assays. MM did statistical analyses, data modeling and revised the manuscript. RM did SNP discovery and genotyping studies. MO performed the bioinformatics analyses. AT conceived the study, supervised the molecular genetic analysis, secured funding, and drafted the manuscript.

Additional material

Additional file 1

Genetic variants and their association with HIV-1 cell permissiveness in vitro *in purified CD4 T cells from 125 healthy blood donors.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4690-3-54-S1.pdf]

Additional file 2

Restriction of N-MLV by common human TRIM5α variants.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4690-3-54-S2.pdf]

Additional file 3

Analysis of association of specific human TRIM5α variants or haplotypes and in vitro *p24 production 7 days post infection of purified CD4 T cells from healthy blood donors with an R5-tropic viral strain.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4690-3-54-S3.pdf]

Additional file 4

TaqMan allelic discrimination primers and probes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4690-3-54-S4.pdf]

Additional file 5

Characteristics of 979 subjects infected with HIV-1.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1742-4690-3-54-S5.pdf]

Acknowledgements

This study has been financed in the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation. (Grant no 3345-062041), and by Swiss National Science Foundation grant no. 3100-0-110012/1 to A.T.

The members of the Swiss HIV Cohort Study are S. Bachmann, M. Battegay, E. Bernasconi, H. Bucher, Ph. Bürgisser, M. Egger, P. Erb, W. Fierz, M. Fischer, M. Flepp (Chairman of the Clinical and Laboratory Committee), P. Francioli (President of the SHCS, Centre Hospitalier Universitaire Vaudois, CH-1011- Lausanne), H.J. Furrer, M. Gorgievski, H. Günthard, P. Grob, B. Hirschel, L. Kaiser, C. Kind, Th. Klimkait, B. Ledergerber, U. Lauper, M. Opravil, F. Paccaud, G. Pantaleo, L. Perrin, J.-C. Piffaretti, M. Rickenbach (Head of Data Center), C. Rudin (Chairman of the Mother & Child Substudy), J. Schupbach, R. Speck, A. Telenti, A. Trkola, P. Vernazza (Chairman of the Scientific Board), R. Weber, S. Yerly.

References

1. Reymond A, Meroni G, Fantozzi A, Merla G, Cairo S, Luzi L, Riganelli D, Zanaria E, Messali S, Cainarca S, Guffanti A, Minucci S, Pelicci PG,

Ballabio A: **The tripartite motif family identifies cell compartments.** *EMBO J* 2001, **20**:2140-2151.
2. Nisole S, Stoye JP, Saib A: **TRIM family proteins: retroviral restriction and antiviral defence.** *Nat Rev Microbiol* 2005, **3**:799-808.
3. Towers GJ: **Restriction of retroviruses by TRIM5 alpha.** *Future Virol* 2006, **1**:71-78.
4. Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, Sodroski J: **The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys.** *Nature* 2004, **427**:848-853.
5. Stremlau M, Perron M, Lee M, Li Y, Song B, Javanbakht H, az-Griffero F, Anderson DJ, Sundquist WI, Sodroski J: **Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor.** *Proc Natl Acad Sci U S A* 2006, **103**:5514-5519.
6. Sawyer SL, Wu LI, Emerman M, Malik HS: **Positive selection of primate TRIM5{alpha} identifies a critical species-specific retroviral restriction domain.** *Proc Natl Acad Sci U S A* 2005, **102**:2832-2837.
7. Ortiz M, Bleiber G, Martinez R, Kaessmann H, Telenti A: **Patterns of evolution of host proteins involved in retroviral pathogenesis.** *Retrovirology* 2006, **3**:11.
8. Song B, Gold B, O'Huigin C, Javanbakht H, Li X, Stremlau M, Winkler C, Dean M, Sodroski J: **The B30.2(SPRY) domain of the retroviral restriction factor TRIM5alpha exhibits lineage-specific length and sequence variation in primates.** *J Virol* 2005, **79**:6111-6121.
9. Stremlau M, Perron M, Welikala S, Sodroski J: **Species-Specific Variation in the B30.2(SPRY) Domain of TRIM5{alpha} Determines the Potency of Human Immunodeficiency Virus Restriction.** *J Virol* 2005, **79**:3139-3145.
10. Yap MW, Nisole S, Stoye JP: **A Single Amino Acid Change in the SPRY Domain of Human Trim5alpha Leads to HIV-1 Restriction.** *Curr Biol* 2005, **15**:73-78.
11. Stoye JP: **Restriction of retrovirus replication.** *13th Conference on Retroviruses and Opportunistic Infections, Denver* 2006-Abstract 59.
12. Newman R, Hall L, Connole M, Chen GL, Kaur A, Miller G, Johnson W: **Balancing selection, gene duplication and functional polymorphism in the Rhesus macaque and Sooty mangabey TRIM5alpha locus.** *13th Conference on Retroviruses and Opportunistic Infections, Denver* 2006-Abstract 141LB.
13. Sawyer SL, Wu LI, Akey JM, Emerman M, Malik HS: **High-Frequency Persistence of an Impaired Allele of the Retroviral Defense Gene TRIM5alpha in Humans.** *Curr Biol* 2006, **16**:95-100.
14. Speelmon EC, Livingston-Rosanoff D, Li SS, Vu Q, Bui J, Geraghty DE, Zhao LP, McElrath MJ: **Genetic association of the antiviral restriction factor TRIM5alpha with human immunodeficiency virus type 1 infection.** *J Virol* 2006, **80**:2463-2471.
15. Carpenter CC, Cooper DA, Fischl MA, Gatell JM, Gazzard BG, Hammer SM, Hirsch MS, Jacobsen DM, Katzenstein DA, Montaner JS, Richman DD, Saag MS, Schechter M, Schooley RT, Thompson MA, Vella S, Yeni PG, Volberding PA: **Antiretroviral therapy in adults: updated recommendations of the International AIDS Society-USA Panel.** *JAMA* 2000, **283**:381-390.
16. Williams LM, Cloyd MW: **Polymorphic human gene(s) determines differential susceptibility of CD4 lymphocytes to infection by certain HIV-1 isolates.** *Virology* 1991, **184**:723-728.
17. **Didier Trono laboratory** 2006 [http://tronolab.epfl.ch].
18. **The Swiss HIV Cohort Study** 2006 [http://www.shcs.ch/].
19. Bleiber G, May M, Martinez R, Meylan P, Ott J, Beckmann J, Telenti A: **Use of a combined ex vivo/in vivo population approach for screening of human genes involved in the Human immunodeficiency virus type 1 life cycle for variants influencing disease progression.** *J Virol* 2005, **79**:12674-12680.
20. **MLwiN: a visual interface for multilevel modelling** 2006 [http://www.mlwin.com/].
21. **PHASE: software for haplotype reconstruction, and recombination rate estimation from population data** 2006 [http://www.stat.washington.edu/stephens/software.html].

4.4.3 Original article

Model Structure of Human APOBEC3G

Kun-Lin Zhang[1], Bastien Mangeat[2], **Millan Ortiz**[1], Vincent Zoete[3], Didier Trono[2], Amalio Telenti[1]*, Olivier Michielin[3]*

[1] Institute of Microbiology, University Hospital Center, University of Lausanne, Lausanne, Switzerland, [2] Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, [3] Swiss Institute of Bioinformatics, Lausanne, Switzerland

PLos One 2007 Apr 18;2(4) [67]

OPEN ACCESS Freely available online

PLoS one

Model Structure of Human APOBEC3G

Kun-Lin Zhang[1], Bastien Mangeat[2], Millan Ortiz[1], Vincent Zoete[3], Didier Trono[2], Amalio Telenti[1]*, Olivier Michielin[3]*

1 Institute of Microbiology, University Hospital Center, University of Lausanne, Lausanne, Switzerland, 2 Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 3 Swiss Institute of Bioinformatics, Lausanne, Switzerland

Background. **APOBEC3G (apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like 3G) has antiretroviral activity associated with the hypermutation of viral DNA through cytosine deamination. APOBEC3G has two cytosine deaminase (CDA) domains; the catalytically inactive amino-terminal domain of APOBEC3G (N-CDA) carries the Vif interaction domain. There is no 3-D structure of APOBEC3G solved by X-ray or nuclear magnetic resonance.** ***Methodology/Principal Findings.*** **We predicted the structure of human APOBEC3G based on the crystal structure of APOBEC2. To assess the model structure, we evaluated 48 mutants of APOBEC3G N-CDA that identify novel variants altering ΔVif HIV-1 infectivity and packaging of APOBEC3G. Results indicated that the key residue D128 is exposed at the surface of the model, with a negative local electrostatic potential. Mutation D128K changes the sign of that local potential. In addition, two novel functionally relevant residues that result in defective APOBEC3G encapsidation, R122 and W127, cluster at the surface.** ***Conclusions/Significance.*** **The structure model identifies a cluster of residues important for packaging of APOBEC3G into virions, and may serve to guide functional analysis of APOBEC3G.**

Citation: Zhang K-L, Mangeat B, Ortiz M, Zoete V, Trono D, et al (2007) Model Structure of Human APOBEC3G. PLoS ONE 2(4): e378. doi:10.1371/journal.pone.0000378

INTRODUCTION

Primate APOBEC3G has antiretroviral activity associated with the hypermutation of viral DNA through cytosine deamination (for recent review see [1 3]). Human APOBEC3G (huAPOBEC3G) fails to restrict HIV-1 due to the degradation imposed by the HIV-1 Vif [4]. In contrast, a number of primate APOBEC3G orthologs display activity against HIV-1 [5 8]. APOBEC3G has a duplicated catalytic deaminase domain (CDA); the amino-terminal domain (N-CDA) of APOBEC3G is required for viral encapsidation but not cytosine deamination [9 11].

There is no 3-D structure solved by X-ray or NMR nor an accurate model of APOBEC3G available. APOBEC3G relates to APOBEC family and AID (activation-induced deaminase) at the sequence level. Recent comparative modeling work for APOBEC-1 and AID [12 17] led to the proposition of a secondary structure alignment between APOBEC3G and cytidine deaminase [18]. The recent publishing of the crystal structure of APOBEC2 [19] provides the template to build a reliable model structure of APOBEC3G by theoretical methods.

In the present work, we model human APOBEC3G, with particular emphasis on the N-terminal domain. Using mutant data of the N-CDA, we mapped critical residues for packaging of APOBEC3G into viral particles on the new structure model of the huAPOBEC3G N-CDA. We completed the analysis by mapping N-CDA residues that are under positive evolutionary pressure in primate APOBEC3G.

While this work was concluded, Huthoff and Malim provided a detailed molecular genetic analysis of the N-CDA region spanning amino acids residues 119 to 146 (Ref [20]). This analysis defined residues 124 to 127 as having a role in APOBEC3G packaging into HIV-1 virions, and residues 128 to 130 as crucial for the interaction with HIV-1 Vif. Our current results confirm and extend this work, and place the findings in a detailed structural model that may serve to advance rational drug design.

MATERIALS AND METHODS

Target sequence and template for structural model

The huAPOBEC3G sequence (residues 1-384; NCBI accession: NP_068594, see http://www.ncbi.nih.gov/) was defined as target sequence. The newly crystallized Human APOBEC2 dimer [19] (PDB ID 2NYT obtained from Xiaojiang S. Chen, see http://www.rcsb.org/pdb/Welcome.do) served as template.

Target-template alignment

We used the align2d function of MODELLER program (http://salilab.org/modeller/) [21 23], to align huAPOBEC3G N-CDA sequence (residues 1-194) and huAPOBEC3G C-CDA sequence (residues 195-384) to the huAPOBEC2 dimer. Additional assessment of the target-template alignment compared structurally determined (DSSP, http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html) [24] and predicted (PSIPRED, http://bioinf.cs.ucl.ac.uk/psipred/) [25 27] secondary structures. The alignment was analyzed and viewed by Jalview (http://www.jalview.org/) [28].

Model building, evaluation and mapping of key residues

The target-template alignment was used to build the model by satisfaction of spatial restraints. The ANOLEA program (http://swissmodel.expasy.org/anolea/) [29 31], that estimates the fold-

Academic Editor: Alejandro Aballay, Duke University Medical Center, United States of America

Received February 9, 2007; **Accepted** March 26, 2007; **Published** April 18, 2007

Funding: Supported by Swiss National Science Foundation grant no. 310000-110012/1 (to A.T.), SCORE funds 3232B0-103172 and 3200B0-103173 (to O.M.) and a grant for interdisciplinary research from the Faculty of Biology and Medicine of the University of Lausanne (to A.T. and O.M.). KZ received a postdoctoral award from the Foundation for advancement in Microbiology and Infectious Diseases. The funding agencies had no role in the design or interpretation of the study.

Competing Interests: The authors have declared that no competing interests exist.

* **To whom correspondence should be addressed.** E-mail: amalio.telenti@chuv.ch (AT); olivier.michielin@unil.ch (OM)

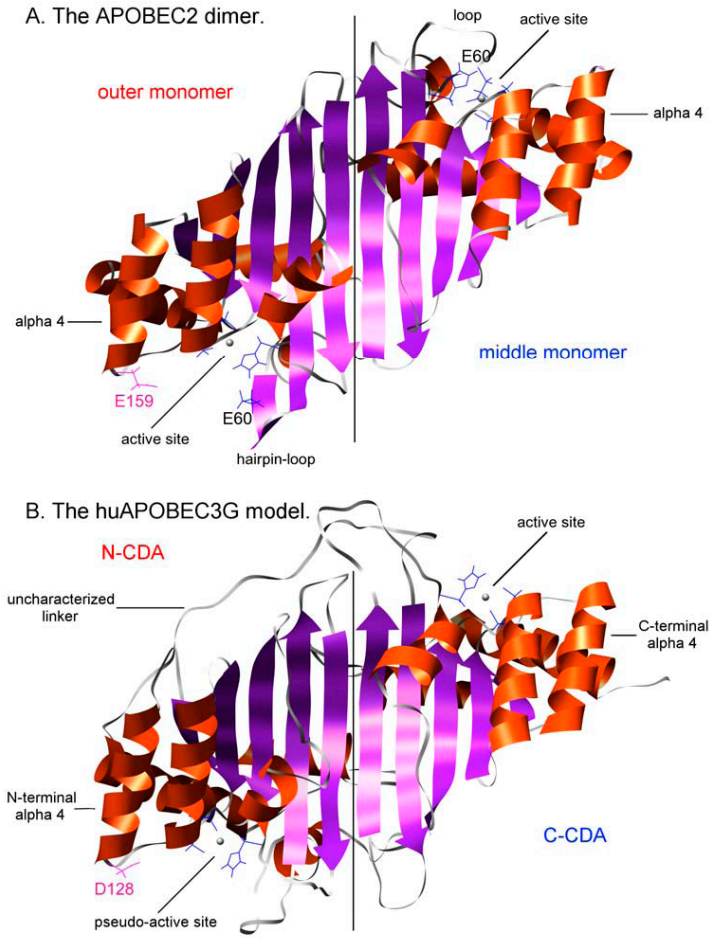


Figure 1. Ribbon view of the huAPOBEC2 dimer and the model of huAPOBEC3G. Panel A. The APOBEC2 "homo"-dimer. Panel B. The huAPOBEC3G model underscoring the six active and pseudo-active site residues in blue, the two zinc ions, and the position of the residue D128, key in the interaction with HIV-1 Vif.
doi:10.1371/journal.pone.0000378.g001

ing free energy of each residue of a protein chain to assess the quality of the predicted structure, was used to score all the models, using the default 5 residue window averaging.

The geometry of the active site/pseudo-active site was based on the corresponding homologous active site regions of the huAPOBEC2 template. The main interaction of anti-parallel $\beta 2$-$\beta 2'$, between huAPOBEC3G N-CDA and C-CDA was defined by restraining segments 48-56 and 235-243 as anti-parallel β-sheet, based on the huAPOBEC2 dimer. For the chain connection between N-CDA and C-CDA, the first 20 amino acid residues of C-CDA were used as a linker and modeled ab initio using the loop routine of the MODELER program.

Generation of 100 models allowed selection of the best model candidate based on the global ANOLEA score. Final refinement for alignment of gap regions was performed by generating 100 additional models, followed by selection of the best model on the basis of the ANOLEA score. The final model was energy-minimized using the CHARMM program (http://www.charmm.org/) [32] and the CHARMM22 all atom force field [33]. The minimization consisted of 200 steps of steepest descent using a dielectric constant of 1 and the Generalized Born GB-MV2 implicit solvent model without cutoff for the solvation free energy. Model evaluation was based on ANOLEA with window 5 averaging, as described above.

Table 1. Mutation, antiviral activity, and surface exposure of residues of the huAPOBEC3G N-CDA.

	Anti-HIV activity*	Protein expression**	Solvent Accessible Surface Area ***
Y13A	+++	+++	14.1+0
F17L	+++	+++	2.4+0.3
F21L	+++	+++	0.4+0
S28A	++	+++	73.9+17.3
R29A	+++	+++	187.2+0
T32A	++	++	6.5+0
Y37A	+++	+++	0+0
K40A	+	+	76.0+0.2
S45A	++	++	33.9+9.5
L49A	+++	+++	24.8+0
L62A	+++	+++	50.1+4.3
H65R	+	+	7.5+0
E67Q	+	+	0+0
F70L	0	++	0.2+0
F74L	++	+++	151.5+15.0
E85Q	+++	+++	104.9+0
Y86A	+++	+++	16.3+0.8
W90L	+++	ND	0+0
Y91A	0	++	16.7+0
I92V	+++	ND	0.1+0
S93A	+++	+++	0+0
W94L	+	+++	35.0+0
P96L	++	++	0+0
C97S	+	+	15.0+1.5
C100S	+	+	0+0.2
M104A	+++	+++	0+0
F107L	+++	+++	23.6+0
L108A	++	++	2.5+0.6
L116A	++	++	0.5+0
T117A	+++	+++	19.1+0
I118A	++	+	0+0
R122A	0	++	76.6+0
L123A	++	++	0+1.6
Y124A	+	+++	36.1+0.7
Y125A	+++	+++	138.6+6.4
F126L	+++	+++	22.6+0
W127L	0	++	187.3+17.8
D130K/N	+++	+++	88.7+2.1
Y131V	++	++	2.2+0
E133Q	+++	+++	112.3+4.9
L135A	+++	+++	0+0
L138A	++	++	0+0
M152A	++	++	0+0
Y154A	++	+++	120.7+0
F157L	++	+++	4.4+0
C160S	+++	+++	0.2+0
F164L	++	++	0+0

Novel functional residues that abolish huAPOBEC3G antiviral activity against wild-type HIV-1 infection are highlighted bold.

*0 = no activity, +residual activity (<10% of wild type),++modest activity (10–50%), +++full activity.

**Semiquantitative western blot

***Estimated by side chain+backbone. Values greater than 20 suggest that the residue is exposed at the surface.

doi:10.1371/journal.pone.0000378.t001

The CHARMM program was used to calculate the solvent accessible surface area of the final model. Additionally, using the FoldX program (http://foldx.embl.de/) [34], in silico alanine-scanning was performed by mutating each residue and calculated the change of folding energy. Molecular surface visualization was done using Chimera (http://www.cgl.ucsf.edu/chimera/) [35,36]. The surface was color-coded according to the Poisson-Boltzmann potential calculated by the UHBD program [37].

Mutation analysis, constructs, viral production and titration

The plasmid expressing a hemagglutinin (HA)-tagged form of APOBEC3G was a kind gift from M. Malim. A series of APOBEC3G alanine and specific mutants was constructed with the QuickChange Mutagenesis kit (Stratagene). The collection was already constituted, and it was not defined on the basis of the new structural data. HIV-1 particles were produced by transient transfection of 293T cells with Fugene (Roche) of a wild-type or of a Vif-defective HIV-1 proviral clone. Viral titers were determined in single-round infectivity assays by applying filtered supernatant from producer cells on HeLa-CD4-LTRLacZ indicator cells. Virion infectivity was derived by dividing the infectious titer by the amount of physical particles.

Packaging assay and Western Blots

HIV-1 particles were produced by transient transfection of 293T cells with Fugene (Roche). 1 ml of virus was then spun in Eppendorf tubes at 13'000rpm in a microfuge at 4° for 90 minutes, without sucrose cushion. Pellets were resuspended in PBS 1% Triton, and the virion amount was measured by a standard RT assay. Normalized amount of virions were then loaded on standard Laemmli protein gels to perform Western Blots. Cell extracts were obtained through a standard RIPA extraction procedure. The HA tag was detected with the mouse hrp-coupled anti-HA 3F10 antibody (Roche). PCNA (proliferating cell nuclear antigen) was detected with the mouse Ab-1 antibody (Oncogene Science). The HIV-1 capsid was detected with the murine anti-p24 antibody produced from the AIDS Research and Reference Reagent Program #183-H12-5C.

RESULTS

The huAPOBEC2 tetramer is composed of two outer monomers and two middle monomers. The outer and middle monomers share the same sequence but a slightly different structure in the region of residues 57-68. For an outer monomer, this region is a hairpin-loop; but for a middle monomer this region is a loop with residue E60 coordinating with Zn^{2+} [19], **Fig. 1A**. We modeled the huAPOBEC3G N-CDA based on the outer monomer of APOBEC2 to resolve the structure between residues 22 and 33 of huAPOBEC3G. The outer monomer of APOBEC2 is more suitable as template because the middle monomer coordinates the Zn^{2+} with E60, a residue not present at the corresponding position in huAPOBEC3G N-CDA. The corresponding region of huAPOBEC3G C-CDA was modeled ab initio. The target-template alignment generated by MODELLER agreed with the secondary structure alignment (Supplementary Fig. S1A). The resulting huAPOBEC3G model is shown in **Fig. 1B**. The structural pattern of this model is very similar to that of the huAPOBEC2 dimer. The final model has a good ANOLEA score profile (indicating reliability of the structure prediction) with a pattern comparable to that of huAPOBEC2 dimer despite the low sequence identity (27%), **Supplementary Fig. S1.**

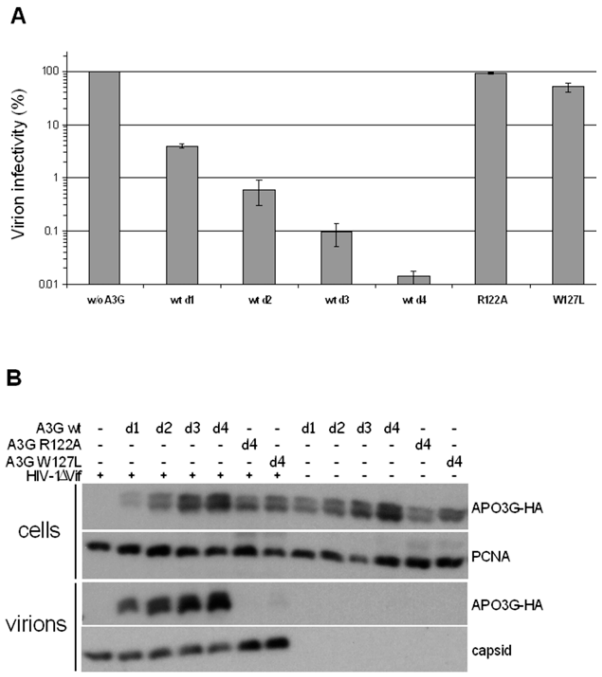


Figure 2. Infectivity, and packaging assay of novel defective huAPOBEC3G mutants. Panel A. Δ Vif HIV-1 particles were produced in presence of different doses of wt APOBEC3G-HA (A3G), or with the highest dose of the R122A and W127L mutants of APOBEC3G-HA. The infectivity of these particles was determined by titration on P4.2 indicator cells. The doses d1 to d4 of APOBEC3G correspond respectively to a molar ratio of APOBEC3G-HA plasmid to virus plasmid of 0.6:1, 1.3:1, 2.7:1 and 5.5:1. The graph is made from one duplicate experiment, and is representative of a total of at least 4 independent duplicate experiments. Panel B. The presence of wt APOBEC3G-HA or the R122A and W127L mutants in ΔVif HIV-1 virions was assessed by western blot. The PCNA and capsid western blots serve to check for consistent loading of cellular and virion extracts, respectively. Representative of at least two independent experiments.
doi:10.1371/journal.pone.0000378.g002

Mutation analysis and mapping of key functional residues

Mutation analysis interrogated 48 of 194 (25%) residues of the huAPOBEC3G N-CDA (**Table 1**). Mutation of the conserved residues constituting the N-CDA Zn^{2+} coordination domain H65, C97 and C100 prevented or reduced encapsidation as previously reported [9]. Mutation of residues F70 and Y91, shown to mediate RNA binding [38], was associated with reduced APOBEC3G encapsidation, as previously reported [9]. Mutation of the pseudo-catalytic site residue E67 (E67Q), resulted in poor protein expression, which limited functional evaluation; E67A was previously reported to reduce the rates of APOBEC3G encapsidation into HIV1 virions [9]. E67 and F70 are buried inside of the model; E67A and F70A may affect folding stability (in silico alanine-scanning). Y91 is a partially exposed at the surface.

Amino acids substitutions at positions Y13, F17, F21, R29, T32, Y37, K40, S45, L49, L62, E85, Y86, W90, I92, S93, P96, M104, F107, L108, L116, T117, I118, L123, Y125, F126, D130, Y131, E133, L135, L138, M152, C160 and F164 did not result in changes in anti-viral activity. Mutations S28A, F74L, W94L, Y124A, Y154A resulted in reduced inhibition of ΔVif HIV1, while mutations R122A and W127L completely abolished this activity (**Table 1**). The functionally defective phenotype of these mutants correlated with a failure to become encapsidated into ΔVif HIV-1 particles (**Fig. 2**). R122, W127 and D128 form a cluster at the surface (**Fig. 3**) and contribute to changes in surface charge or structure (**Fig. 4**).

Prior evolutionary analysis of primate APOBEC3G identified a number of residues under diversifying (positive) selection [39]. Most of the residues under positive selection pressure are exposed at the surface of the model. A cluster of residues under positive selection that includes T98, K99, R102, D128, and P129 overlaps with the cluster of functionally important residues around D128 (**Fig. 3**).

DISCUSSION

This work presents a model structure of huAPOBEC3G that captures information from the recently published APOBEC2 structure [19]. Critically, the APOBEC2 template provides a structural reference for the predicted extra helix α4, that carries residue D128–a residue that governs the virus-specific sensitivity of APOBEC3G to Vif-mediated inhibition [5–8]. Previously available CDA structures provided suboptimal templates for APOBEC3G N-CDA. This is not only due to the low sequence identity,

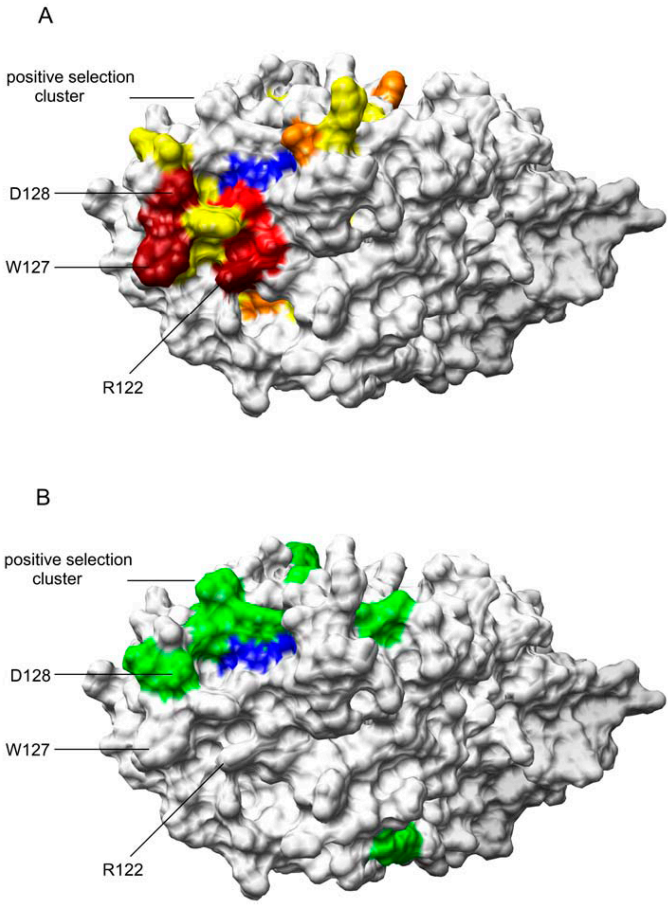


Figure 3. Mapping of functional and evolutionary informative residues at the molecular surface of N-CDA. Panel A, Mapping of functional residues. The color gradient (dark red>red>orange>yellow) reflects the role of various mutations in abolishing anti-HIV activity (see Table 1) ranging from no antiviral activity (dark red) to normal activity (yellow). Panel B. Mapping of evolutionary informative residues at the molecular surface. Green color identifies amino acids under positive selective pressure in primate APOBEC3G. The positive selection cluster includes T98, K99, R102, D128, and P129. Blue color identifies the pseudo–active site.
doi:10.1371/journal.pone.0000378.g003

but also because of the presence in APOBEC3G N-CDA of an extra α helix. In contrast, APOBEC2 has an extra α helix, resulting in a α1-β1-β2-α2-β3-α3-β4-α4-β5-α5-α6 configuration, while CDA structures such as human cytidine deaminase have the pattern α1-β1-β2-α2-β3-α3-β4-β5-α4-α5. In addition, the direction of β4 and β5 are different: parallel in APOBEC2 but antiparallel in cytidine deaminase. We had previously used the yeast Cytosine deaminase (PDB ID 1P6O, chain A) to address the position of the extra helix and the correct direction of β4 and β5.

The model emerging from this analysis allows speculation on various functionally relevant structural details of interest for the understanding of the Vif-APOBEC3G interaction and the process of APOBEC3G encapsidation into HIV-1 virions. In contrast with previous secondary and tertiary structure models, the current model locates the distinctive extra alpha helix on the same planar surface as the pseudo-active site. We evaluated extensive mutation data on this surface, that characterized two functionally relevant residues R122 and W127 that resulted in failure to inhibit infection by HIV-1/*Δvif* due to lack of packaging of APOBEC3G into viral particles. Interaction of APOBEC3G with the NC-domain of HIV-1 Gag and non-specific RNA binding leads to its encapsidation into progeny virions [10,40–43]. The structural model proposes a surface hot-spot domain that includes residues R122, W127 in proximity to the active site, and overlapping with a cluster of residues under positive selective pressure that includes D128 and P129.

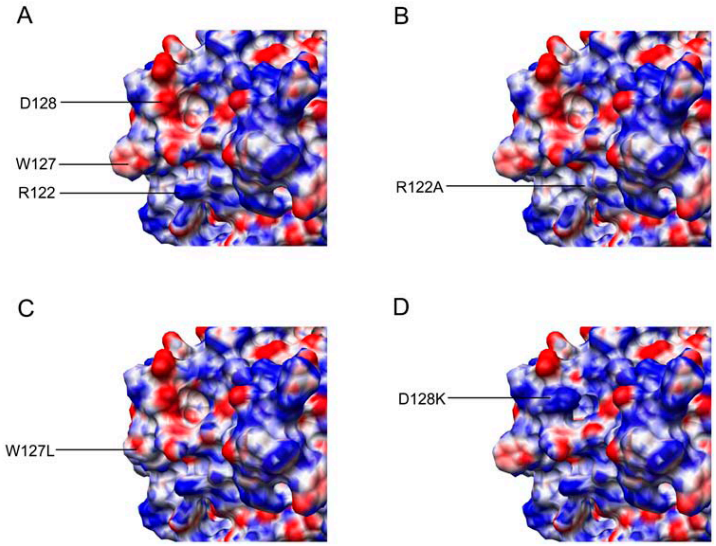


Figure 4. Mapping potential to molecular surface. Zoomed in view on the functional surface region within the N-CDA. Panel A, wild-type huAPOBEC3G N-CDA. Panel B-D, effect of mutations R122A, W127L, and D128K on surface charge and shape. The red color represents negative potential, the blue color expresses positive potential, and the white color expresses zero potential.
doi:10.1371/journal.pone.0000378.g004

Our results are consistent with the work of Huthoff and Malim [20]. While their work concentrated on the molecular genetic analysis of the 28-amino acid region between residues 119 and 146, our analysis interrogated 48 of the 194 residues of the N-CDA domain. This reflects the interest to identify additional motifs capable of interacting with HIV-1 Vif, or participating in packaging [20,44]. Minor differences were observed between the two studies. We observed a defect in the packaging efficiency of R122A mutant protein–that extends the relevant motif to include R122-Y124-Y125-F126-W127. Regarding the motif affecting regulation by Vif, our analysis of D130 mutants (D130K and D130N) did not result in a Vif-resistant phenotype. This difference may reflect a greater dependence of the N-CDA and Vif for the negative charge in position 128 (Ref [20]). While not formally tested in the present study, the biological relevance of P129 is highlighted by its inclusion–together with D128-in a patch of residues under positive selective pressure in primates.

Inspection of surface modifications conferred by various mutations highlights the structural and/or charge differences in this region as the molecular basis for disruption of the APOBEC3G packaging into HIV-1 virions, and modification in the interaction with Vif. The detailed model structure presented here could serve to advance rational drug design. The quality of a homology model is strongly related to the sequence identity with the structural template. The current model, based on a sequence identity of 27%, should be satisfactory in its global fold, as supported by the good ANOLEA energy score profile (**Supplementary Fig. S1**). From such a model, correct positioning of the protein backbone and orientation of the side chains can be expected, allowing reliable conclusion to be made in the design and interpretation of experimental mutation data. Whether or not the accuracy of such a model is sufficient to perform docking simulations is still an open question. However, the fact that critical residues obtained experimentally by mutation analysis do cluster in well-defined patches at the surface of the model argue that side chain packing is correct. In the latter case, docking simulations could be envisioned based on the current model. As recently shown for TRIM5α [45], the model proposed may serve to guide further functional analysis of APOBEC3G.

SUPPORTING INFORMATION

Figure S1 Panel A, Target-template alignment. Alpha helix are shown in red and beta sheets in green. The huAPOBEC3G N-CDA is aligned to the outer (O) monomer of APOBEC2, and the C-CDA to the middle (M) monomer. Panel B, ANOLEA scores. The three ANOLEA profiles and the secondary structure were aligned according to the target-template alignment. The ANOLEA profile excludes the uncharacterized linker (residues 195-214)
Found at: doi:10.1371/journal.pone.0000378.s001 (2.86 MB EPS)

ACKNOWLEDGMENTS

We thank X.S. Chen for APOBEC2 PDB data.

Author Contributions

Conceived and designed the experiments: AT DT OM. Performed the experiments: KZ BM. Analyzed the data: AT DT MO KZ BM OM. Contributed reagents/materials/analysis tools: VZ OM. Wrote the paper: AT KZ OM.

REFERENCES

1. Cullen BR (2006) Role and mechanism of action of the APOBEC3 family of antiretroviral resistance factors. J Virol 80: 1067 1076.
2. Yu X-F (2006) Innate cellular defenses of APOBEC3 cytidine deaminases and viral counter-defenses. Curr Opin HIV/AIDS 1: 187 193.
3. Harris RS, Matsuo H (2006) Dancin' deaminase. Nat Struct Mol Biol 13: 380 381.
4. Sheehy AM, Gaddis NC, Choi JD, Malim MH (2002) Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature 418: 646 650.
5. Mangeat B, Turelli P, Liao S, Trono D (2004) A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action. J Biol Chem 279: 14481 14483.
6. Bogerd HP, Doehle BP, Wiegand HL, Cullen BR (2004) A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor. Proc Natl Acad Sci U S A 101: 3770 3774.
7. Schrofelbauer B, Chen D, Landau NR (2004) A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif). Proc Natl Acad Sci U S A 101: 3927 3932.
8. Xu H, Svarovskaia ES, Barr R, Zhang Y, Khan MA, et al. (2004) A single amino acid substitution in human APOBEC3G antiretroviral enzyme confers resistance to HIV-1 virion infectivity factor-induced depletion. Proc Natl Acad Sci U S A 101: 5652 5657.
9. Navarro F, Bollman B, Chen H, Konig R, Yu Q, et al. (2005) Complementary function of the two catalytic domains of APOBEC3G. Virology 333: 374 386.
10. Luo K, Liu B, Xiao Z, Yu Y, Yu X, et al. (2004) Amino-terminal region of the human immunodeficiency virus type 1 nucleocapsid is required for human APOBEC3G packaging. J Virol 78: 11841 11852.
11. Newman EN, Holmes RK, Craig HM, Klein KC, Lingappa JR, et al. (2005) Antiviral function of APOBEC3G can be dissociated from cytidine deaminase activity. Curr Biol 15: 166 170.
12. Scott J, Navaratnam N, Carter C (1998) Molecular modelling and the biosynthesis of apolipoprotein B containing lipoproteins. Atherosclerosis 141 Suppl 1: S17 S24.
13. Navaratnam N, Fujino T, Bayliss J, Jarmuz A, How A, et al. (1998) Escherichia coli cytidine deaminase provides a molecular model for ApoB RNA editing and a mechanism for RNA substrate recognition. J Mol Biol 275: 695 714.
14. Xie K, Sowden MP, Dance GS, Torelli AT, Smith HC, et al. (2004) The structure of a yeast RNA-editing deaminase provides insight into the fold and function of activation-induced deaminase and APOBEC-1. Proc Natl Acad Sci U S A 101: 8114 8119.
15. Zaim J, Kierzek AM (2003) Domain organization of activation-induced cytidine deaminase. Nat Immunol 4: 1153.
16. Ta VT, Nagaoka H, Catalan N, Durandy A, Fischer A, et al. (2003) AID mutant analyses indicate requirement for class-switch-specific cofactors. Nat Immunol 4: 843 848.
17. Wedekind JE, Dance GS, Sowden MP, Smith HC (2003) Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. Trends Genet 19: 207 216.
18. Huthoff H, Malim MH (2005) Cytidine deamination and resistance to retroviral infection: towards a structural understanding of the APOBEC proteins. Virology 334: 147 153.
19. Prochnow C, Bransteitter R, Klein MG, Goodman MF, Chen XS (2006) The APOBEC-2 crystal structure and functional implications for the deaminase AID. Nature.
20. Huthoff H, Malim MH (2007) Identification of amino acid residues in APOBEC3G required for regulation by HIV-1 Vif and virion encapsidation. J Virol PMID. pp 17267497.
21. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234: 779 815.
22. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29: 291 325.
23. Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9: 1753 1773.
24. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577 2637.
25. McGuffin LJ, Bryson K, Jones DT (2000) The PSIPRED protein structure prediction server. Bioinformatics 16: 404 405.
26. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, et al. (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33: W36 W38.
27. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195 202.
28. Clamp M, Cuff J, Searle SM, Barton GJ (2004) The Jalview Java alignment editor. Bioinformatics 20: 426 427.
29. Melo F, Feytmans E (1998) Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 277: 1141 1152.
30. Melo F, Feytmans E (1997) Novel knowledge-based mean force potential at atomic level. J Mol Biol 267: 207 222.
31. Melo F, Devos D, Depiereux E, Feytmans E (1997) ANOLEA: a www server to assess protein structures. Proc Int Conf Intell Syst Mol Biol 5: 187 190.
32. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, et al. (1983) CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comp Chem 4: 187 217.
33. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evaseck JD, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102: 3586 3616.
34. Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320: 369 387.
35. Sanner MF, Olson AJ, Spehner JC (1996) Reduced surface: an efficient way to compute molecular surfaces. Biopolymers 38: 305 320.
36. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera a visualization system for exploratory research and analysis. J Comput Chem 25: 1605 1612.
37. Davis ME, Madura JD, Luty BA, McCammon JA (1991) Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. Comp Phys Comm 62: 187 197.
38. Navaratnam N, Bhattacharya S, Fujino T, Patel D, Jarmuz AL, et al. (1995) Evolutionary origins of apoB mRNA editing: catalysis by a cytidine deaminase that has acquired a novel RNA-binding motif at its active site. Cell 81: 187 195.
39. Ortiz M, Bleiber G, Martinez R, Kaessmann H, Telenti A (2006) Patterns of evolution of host proteins involved in retroviral pathogenesis. Retrovirology 3: 11.
40. Cen S, Guo F, Niu M, Saadatmand J, Deflassieux J, et al. (2004) The interaction between HIV-1 Gag and APOBEC3G. J Biol Chem 279: 33177 33184.
41. Schafer A, Bogerd HP, Cullen BR (2004) Specific packaging of APOBEC3G into HIV-1 virions is mediated by the nucleocapsid domain of the gag polyprotein precursor. Virology 328: 163 168.
42. Svarovskaia ES, Xu H, Mbisa JL, Barr R, Gorelick RJ, et al. (2004) Human apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like 3G (APOBEC3G) is incorporated into HIV-1 virions through interactions with viral and nonviral RNAs. J Biol Chem 279: 35822 35828.
43. Zennou V, Perez-Caballero D, Gottlinger H, Bieniasz PD (2004) APOBEC3G incorporation into human immunodeficiency virus type 1 particles. J Virol 78: 12058 12061.
44. Conticello SG, Harris RS, Neuberger MS (2003) The Vif protein of HIV triggers degradation of the human antiretroviral DNA deaminase APOBEC3G. Curr Biol 13: 2009 2013.
45. Ohkura S, Yap MW, Sheldon T, Stoye JP (2006) All three variable regions of the TRIM5 alpha B30.2 domain can contribute to the specificity of the retrovirus restriction. J Virol 80: 8554 8565.

Chapter 5. Discussion and perspectives

5. Discussion and perspectives:

5.1 Lessons learned

The progress of my thesis work reflects the successive developments in the field, in particular the growing number of genetic information, and the feasibility of amplification and resequencing of gene orthologs.

The first gene targets proved to be excellent models for the power of evolutionary genetics to identify domains important for retroviral restriction. In the case of TRIM5α, evolutionary approaches had the same level of precision in identifying key residues as extensive biological assessment through the construction of chimeras and site-directed mutagenesis gene variants. In the case of APOBEC3G, analysis offered a more general pattern of residues under selective pressure across the gene, and however, enough information to identify residues in the interaction domain with HIV-1 Vif.

The next steps of my research brought me to test how general this approach could be if applied to other proteins that are known to interact with one or more pathogens. For this, the choice was the DC-SIGN family of proteins, because of their role in the recognition of a range of pathogens. The K_A/K_S values were in the range of 0.3 to 0.5, considered as values indicating a process of purifying selection. The absence of strong signs of positive pressure limited the capacity to identify relevant domains. These results made me contemplate the interest of assessing a second family of genes involved in the recognition of molecular patterns from multiple pathogens. Study of the TLR family confirmed the image generated by the analysis of the DC-

SIGN family: a similar range of K_A/K_S values, and lack of precision to identify particular regions or amino acids under positive selection. Thus, the initial results would suggest that only under the unique conditions of strong positive selection would the evolutionary genetic analysis help identify key mutational events of interest of the understanding of the underlying biological process.

The possibility to generalize this approach, with the goal of understanding its potential, was limited by the need to generate high quality of sequence across the various primate lineages. This limitation was in part lifted by the new availability of complete genomes of primates representing the main branches: Hominids, Old and New World monkeys. I first assessed the precision of analyzing 5 primates (human, chimpanzee, bornean orang-utan, rhesus monkey and common marmoset) versus work with a larger set of eleven primate species. The results were encouraging, as they allow the similar estimation of K_A/K_S value averaged over the entire tree. However, it was also apparent that a reduced number of primates would seriously decrease the precision in the identification of regions under positive selection, and of particular residues.

These limitations notwithstanding, I undertook a large scale analysis of 137 genes involved in the HIV-1 life cycle and pathogenesis, and of 100 control genes in the 5-primate set. The large scale analysis confirmed the original impression from the analysis of TRIM5α and APOBEC3G: that evolutionary genetics would identify, without *a priori*, critical residues in the interaction with pathogens. Indeed the analysis identified residues in at least two genes/proteins (CD4 and CCR3) that had already been shown involved in HIV-1 biology. Second, while the average K_A/K_S value was

shown to be 0.2, genes/proteins of the innate immunity consistently exhibited a K_A/K_S average value of 0.4, which then appears as characteristic for these families of genes/proteins. The large scale study also characterized a theoretical “mutational space” for genes involved in cellular activities needed for the viral replication. In particular, genes involved in transcription and late phases of the viral replication cycle were under significantly stronger purifying selection than the average across control genes. We hypothesize that strongly conserved genes in primate evolution will be unlikely to carry (common) polymorphism in modern humans that would result in interindividual differences in disease susceptibility.

5.2 Perspectives

As discussed above, one of the goals of genetic evolutionary studies would be to guide functional analysis of relevant proteins. The outcome of the large scale analysis illustrates this point by the identification of eleven proteins that have amino acids under positive selective pressure. The nature of the protein would guide the most appropriate empirical approach. One of the most innovative approaches is the possibility to use the evolutionary sequence data for ancestral reconstruction. This was done for TRIM5α where we constructed and tested five ancestral TRIM5α variants on the lineage leading to modern humans and spanning 25 million years of primate evolution. The study underscored the feasibility of such approach and resulted in challenging observations: the identification of an ancestral TRIM5α capable of restricting modern HIV-1, and the capacity of this approach to identify new functional residues. This type of approach is now applied to the analysis of APOBEC3G in the laboratory coupled to the reconstruction and analysis of ancestral retroviruses.

A second perspective relates to the increasing availability of complete genomes. This would increasingly improve the precision of the analysis, and will also lead to a more systematic genome-wide cataloguing of genes relevant (or possibly relevant) to a disease. The availability of data from multiple members of a same species will also help precision. The release of members belonging to specific groups, for example the orang-utan and the unexpected Neanderthal genomes will improve the analysis of human-specific selection. In addition, the expected release of the genomes of prosimians will allow the estimation of ancestral proteins up to 40 million years ago.

A third perspective concerns the modelling of co-evolution. We have initiated the analysis of evolution of the HIV-1 viral capsid and of the cognate TRIM5α host partner. These approaches will bring the red-queen principle to the field of functional testing.

A fourth perspective relates to the expected improvements in informatic tools, including the use of better programs for alignment of sequences, better coverage of genomes, and better approaches to the handling of indels, paralogous genes and copy number variation.

Chapter 6. References

Reference List

1. Leigh Van Valen Evolutionary Theory (ed.), pp. 1-301973).

2. Lewis Carroll *Through the looking glass*. Macmillan, London (1872).

3. Hahn, B. H., Shaw, G. M., De Cock, K. M. & Sharp, P. M. AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607-614 (2000).

4. Goodman, M. The genomic record of Humankind's evolutionary roots. *Am. J. Hum. Genet.* **64**, 31-39 (1999).

5. Barre-Sinoussi, F. *et al.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**, 868-871 (1983).

6. Korber, B. *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* **288**, 1789-1796 (2000).

7. Worobey, M. *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**, 661-664 (2008).

8. Preston, B. D., Poiesz, B. J. & Loeb, L. A. Fidelity of HIV-1 reverse transcriptase. *Science* **242**, 1168-1171 (1988).

9. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. & Ho, D. D. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582-1586 (1996).

10. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725-736 (1994).

11. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496-503 (2000).

12. Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107-1118 (2005).

13. Stremlau, M. *et al.* The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature* **427**, 848-853 (2004).

14. Cullen, B. R. Role and mechanism of action of the APOBEC3 family of antiretroviral resistance factors. *J. Virol.* **80**, 1067-1076 (2006).

15. Sharp, P. M., Shaw, G. M. & Hahn, B. H. Simian immunodeficiency virus infection of chimpanzees. *J Virol* **79**, 3891-3902 (2005).

16. Hirsch, V. M. What can natural infection of African monkeys with simian immunodeficiency virus tell us about the pathogenesis of AIDS? *AIDS Rev* **6**, 40-53 (2004).

17. Nisole, S., Stoye, J. P. & Saib, A. TRIM family proteins: retroviral restriction and antiviral defence. *Nat. Rev. Microbiol.* **3**, 799-808 (2005).

18. Song, B. *et al.* The B30.2(SPRY) domain of the retroviral restriction factor TRIM5alpha exhibits lineage-specific length and sequence variation in primates. *J. Virol.* **79**, 6111-6121 (2005).

19. Perez-Caballero, D., Hatziioannou, T., Yang, A., Cowan, S. & Bieniasz, P. D. Human tripartite motif 5alpha domains responsible for retrovirus restriction activity and specificity. *J. Virol.* **79**, 8969-8978 (2005).

20. Yap, M. W., Nisole, S. & Stoye, J. P. A single amino acid change in the SPRY domain of human Trim5alpha leads to HIV-1 restriction. *Curr. Biol.* **15**, 73-78 (2005).

21. Stremlau, M., Perron, M., Welikala, S. & Sodroski, J. Species-specific variation in the B30.2(SPRY) domain of TRIM5alpha determines the potency of human immunodeficiency virus restriction. *J. Virol.* **79**, 3139-3145 (2005).

22. Stremlau, M. *et al.* Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor. *Proc. Natl. Acad. Sci. U. S. A* **103**, 5514-5519 (2006).

23. az-Griffero, F. *et al.* Rapid turnover and polyubiquitylation of the retroviral restriction factor TRIM5. *Virology* **349**, 300-315 (2006).

24. Chelbi-Alix, M. K., Quignon, F., Pelicano, L., Koken, M. H. & de, T. H. Resistance to virus infection conferred by the interferon-induced promyelocytic leukemia protein. *J. Virol.* **72**, 1043-1051 (1998).

25. Turelli, P. *et al.* Cytoplasmic recruitment of INI1 and PML on incoming HIV preintegration complexes: interference with early steps of viral replication. *Mol. Cell* **7**, 1245-1254 (2001).

26. Regad, T. *et al.* PML mediates the interferon-induced antiviral state against a complex retrovirus via its association with the viral transactivator. *EMBO J.* **20**, 3495-3505 (2001).

27. Harris, R. S. *et al.* DNA deamination mediates innate immunity to retroviral infection. *Cell* **113**, 803-809 (2003).

28. Mangeat, B. *et al.* Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* **424**, 99-103 (2003).

29. Lecossier, D., Bouchonnet, F., Clavel, F. & Hance, A. J. Hypermutation of HIV-1 DNA in the absence of the Vif protein. *Science* **300**, 1112 (2003).

30. Marin, M., Rose, K. M., Kozak, S. L. & Kabat, D. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nat. Med.* **9**, 1398-1403 (2003).

31. Yu, X. *et al.* Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science* **302**, 1056-1060 (2003).

32. Mangeat, B., Turelli, P., Liao, S. & Trono, D. A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action. *J. Biol. Chem.* **279**, 14481-14483 (2004).

33. Schrofelbauer, B., Chen, D. & Landau, N. R. A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif). *Proc. Natl. Acad. Sci. U. S. A* **101**, 3927-3932 (2004).

34. Luban, J., Bossolt, K. L., Franke, E. K., Kalpana, G. V. & Goff, S. P. Human immunodeficiency virus type 1 Gag protein binds to cyclophilins A and B. *Cell* **73**, 1067-1078 (1993).

35. Franke, E. K., Yuan, H. E. & Luban, J. Specific incorporation of cyclophilin A into HIV-1 virions. *Nature* **372**, 359-362 (1994).

36. Sokolskaja, E., Sayah, D. M. & Luban, J. Target cell cyclophilin A modulates human immunodeficiency virus type 1 infectivity. *J. Virol.* **78**, 12800-12808 (2004).

37. Berthoux, L., Sebastian, S., Sokolskaja, E. & Luban, J. Cyclophilin A is required for TRIM5{alpha}-mediated resistance to HIV-1 in Old World monkey cells. *Proc. Natl. Acad. Sci. U. S. A* **102**, 14849-14853 (2005).

38. Sayah, D. M., Sokolskaja, E., Berthoux, L. & Luban, J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**, 569-573 (2004).

39. Nisole, S., Lynch, C., Stoye, J. P. & Yap, M. W. A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1. *Proc. Natl. Acad. Sci. U. S. A* **101**, 13324-13328 (2004).

40. az-Griffero, F. *et al.* Requirements for capsid-binding and an effector function in TRIMCyp-mediated restriction of HIV-1. *Virology* **351**, 404-419 (2006).

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

42. Wilson, S. J. *et al.* Independent evolution of an antiviral TRIMCyp in rhesus macaques. *Proc Natl Acad Sci U S A* **105**, 3557-3562 (2008).

43. Luban, J. Cyclophilin A, TRIM5, and resistance to human immunodeficiency virus type 1 infection. *J. Virol.* **81**, 1054-1061 (2007).

44. Koppel, E. A., van Gisbergen, K. P., Geijtenbeek, T. B. & van, K. Y. Distinct functions of DC-SIGN and its homologues L-SIGN (DC-SIGNR) and mSIGNR1 in pathogen recognition and immune regulation. *Cell Microbiol.* **7**, 157-165 (2005).

45. Figdor, C. G., van, K. Y. & Adema, G. J. C-type lectin receptors on dendritic cells and Langerhans cells. *Nat. Rev. Immunol.* **2**, 77-84 (2002).

46. Bashirova, A. A. *et al.* Novel member of the CD209 (DC-SIGN) gene family in primates. *J. Virol.* **77**, 217-227 (2003).

47. Akira, S. TLR signaling. *Curr. Top. Microbiol. Immunol.* **311**, 1-16 (2006).

48. Pomerantz, R. J., Feinberg, M. B., Trono, D. & Baltimore, D. Lipopolysaccharide is a potent monocyte/macrophage-specific stimulator of human immunodeficiency virus type 1 expression. *J Exp Med* **172**, 253-261 (1990).

49. Bafica, A., Scanga, C. A., Schito, M., Chaussabel, D. & Sher, A. Influence of coinfecting pathogens on HIV expression: evidence for a role of Toll-like receptors. *J. Immunol.* **172**, 7229-7234 (2004).

50. Equils, O. *et al.* Toll-like receptor 2 (TLR2) and TLR9 signaling results in HIV-long terminal repeat trans-activation and HIV replication in HIV-1 transgenic mouse spleen cells: implications of simultaneous activation of TLRs on HIV replication. *J Immunol* **170**, 5159-5164 (2003).

51. Schlaepfer, E., Audige, A., Joller, H. & Speck, R. F. TLR7/8 triggering exerts opposing effects in acute versus latent HIV infection. *J. Immunol.* **176**, 2888-2895 (2006).

52. Goff, S. P. Host factors exploited by retroviruses. *Nat Rev Microbiol* **5**, 253-263 (2007).

53. Swanson, C. M. & Malim, M. H. SnapShot: HIV-1 proteins. *Cell* **133**, 742, 742 (2008).

54. Brass, A. L. *et al.* Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319**, 921-926 (2008).

55. Konig, R. *et al.* Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* **135**, 49-60 (2008).

56. Mangeat, B. & Trono, D. Lentiviral vectors and antiretroviral intrinsic immunity. *Hum. Gene Ther.* **16**, 913-920 (2005).

57. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664 (2002).

58. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).

59. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276-277 (2000).

60. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555-556 (1997).

61. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431-449 (2000).

62. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**, 568-573 (1998).

63. Ortiz, M., Bleiber, G., Martinez, R., Kaessmann, H. & Telenti, A. Patterns of evolution of host proteins involved in retroviral pathogenesis. *Retrovirology.* **3**, 11 (2006).

64. Ortiz, M. *et al.* The evolutionary history of the CD209 (DC-SIGN) family in humans and non-human primates. *Genes Immun* **9**, 483-492 (2008).

65. Goldschmidt, V. *et al.* Antiretroviral activity of ancestral TRIM5alpha. *J Virol* **82**, 2089-2096 (2008).

66. Goldschmidt, V. *et al.* Role of common human TRIM5alpha variants in HIV-1 disease progression. *Retrovirology.* **3**, 54 (2006).

67. Zhang, K. L. *et al.* Model structure of human APOBEC3G. *PLoS. ONE.* **2**, e378 (2007).