Evolutionary rate shifts suggest species-specific adaptation events in HIV-1 and SIV

Maoz Gelbart and Adi Stern

School of Molecular Cell Biology and Biotechnology, George S. Wise faculty of Life Sciences, Tel Aviv University, Israel

* Correspondence to AS: sternadi@post.tau.ac.il

Abstract

HIV-1 is an emerging virus that entered the human population several times independently. Two host jumps from chimpanzees resulted in HIV-1 groups M and N, whereas two host jumps, most likely from gorillas, resulted in HIV-1 groups O and P. The virus infecting gorilla (SIVgor), is in itself likely to be a derived form of a host hump from the virus of chimpanzees (SIVcpz). How HIV-1 was able to cross the species barrier is an outstanding question. At the macro-molecular level, insights have been made about the relationships between specific antiviral restriction factors and the ability of SIV and HIV to restrict them. Nevertheless, the molecular view of the adaptation process of SIV and HIV to their hosts across a series of host jumps is still mostly unclear.

Here, we utilized the availability of full and diverse HIV-1, SIVcpz and SIVgor sequences to identify sites in the proteins of HIV-1 that underwent a change in the rate of substitution. Rate deceleration events point at sites that became conserved in a specific lineage, and indicate a possible gain-of-function. We found a total of 137 rate deceleration events across the HIV-1/SIV phylogeny. The majority of rate deceleration events occurred in parallel to speciation and were mostly different across strains, indicating different adaptation strategies independently used by the different viral lineages. The rate shift distribution pattern found supports abundant adaption of the virus to the first stage of entry, followed by adaptation of the virus to host restriction factors. Surprisingly, we observed that HIV-1 group O exhibited the most rate shift events, and suggest that since the virus underwent two consecutive host jumps, from chimpanzee to gorilla to human, this entailed more adaptation on the part of the virus.

Introduction

The Human Immunodeficiency Viruses HIV-1 and HIV-2 are the causative agents of AIDS in humans, infecting millions of people worldwide. Those viruses emerged from a clade of lentiviruses known as the Simian Immunodeficiency Virus (SIV), which naturally infects a variety of non-human primate species. HIV in humans arose from several independent transmission events of primates SIV that resulted in HIV-1 groups M and N (from SIV naturally infecting chimpanzees, SIVcpz), HIV-1 groups O and P (from SIV naturally infecting gorillas, SIVgor), and HIV-2 groups A through H (from SIV naturally infecting sooty mangabeys, SIVsmm) [1-3]. The gorilla infecting lentivirus, SIVgor, is itself a result of a transmission of SIVcpz to gorillas. Phylogenetic analyses date the oldest common ancestors of HIV groups M and O to the beginning of the 20th century, making this pathogen a relatively new one in humans [4, 5]. Similar analyses on SIVgor date the inception of this virus in the western lowland gorilla population to the 19th century [6]. SIVcpz was also found to be a transmission from other primates, dated at roughly 1500 and leading to the two lineages of SIVcpz: SIVcpz*ptt* infecting the chimpanzee subspecies *Pan troglodytes troglodytes* of central Africa, and SIVcpz*pts* infecting the *Pan troglodytes schweinfurthii* chimpanzee subspecies of eastern Africa [7-12].

How viruses are able to cross species barrier is a subject of much interest, since many pandemic human viruses arose from zoonosis events such as the influenza strain of "Spanish flu" H1N1, measles virus (MeV), and SARS coronavirus [13]. Due to genetic differences between the hosts, virus adaptation occurs at multiple levels: at the level of entry; at the level of interaction with the host adaptive immune system; at the level of interaction with host antiviral restriction factors; and at the level of viral recruitment of host cellular machinery for its own needs [14, 15]. Protein adaptations are reflected in the history of genomes and may be manifested in changes in the amino acid substitution rates of the adaptive sites, which are expected to be more conserved in the clade where adaptation happened, and less conserved in the other clades [16]. A reciprocal phenomenon where a position in one clade is less conserved than in others is also possible, and may reflect a loss of function or relaxation of a prior constraint. Identification of such rate shifting positions can thus reveal virus adaptation events, and is expected to promote a better understanding of host jumps in general and the HIV pandemic in particular.

A method for identification of evolutionary rate changes has been previously established [17] and was used to study intra-subtype adaptation events. To date, the lack of SIVcpz and group O genomic sequences limited the ability to study adaptation across cross species transfer events. Here, we utilized the growing availability of diverse and full HIV-1 and SIV genomes to identify all sites in various SIVcpz/SIVgor/HIV-1 clades whose evolutionary rate changed across clades, possibly due to adaptation events. We demonstrate how cross-species transmissions are correlated with abundant evolutionary rate shifts and how known adaptation events are manifested in different amino acid substitution rates between lineages. Based on the rate shifts patterns, we suggest possible as yet unknown adaptation events, and highlight the relatively exceptional amount of evolutionary rate shifts observed in HIV-1 group O, possibly due to a more extreme host species barrier.

Results

We queried the Los Alamos HIV database for all available HIV-1, SIVcpz and SIVgor full sequences, from which we took the sequences of all nine HIV-1 proteins. Due to high representation of group M sequences, we downsampled this group to the size of available group O sequences, retaining the strains with the most internal variation (N=126, see Methods). We used RASER tool to identify sites that displayed different rates of evolution from the two sides of some branch along the HIV-1/SIVcpz/SIVgor phylogeny (see example in Figure 1) [17]. A different downsample of group M sequences to a bigger size was also conducted, to validate results robustness to cohort size variation (N=223, see Suppl. text). All sites reported are in HIV-1 subtype B reference strain HXB2 coordinates.

Most rate shift events were identified in speciation branches. We first sought to characterize the branches where rate shift events were suggested. A total of 271 rate-shifting positions were identified along the phylogeny (Table S1). 137 of them were identified as rate-deceleration events and additional 134 identified as rate-accelerations. As listed in Table 1, 230 out of the 271 rate-shifting positions were attributed to four branches: the branch separating group M from SIVcpz (with 51 identified rate shifts), the branch separating group O from SIVgor (97 rate shifts), the branch separating SIVgor and group O from SIVcpz (24 rate shifts) and the branch separating SIVcpz*pts* from other viruses (58 rate shifts). Log likelihood tests were found to be very significant for a rate shift model explaining all proteins' evolutionary rates (Table 2).

Different rate deceleration patterns identified in different branches. A special focus was given to rate deceleration events, as they likely indicate a gain of function in a protein. In Figure 1b we mapped the percent of rate shifts in each protein observed in each lineage, corresponding to different virus speciation events. While rate deceleration distribution patterns differed between the different lineages (Figure 1 and Table 3), one common theme was that all prominent branches underwent a significant portion of their rate deceleration events in the env polyprotein. The next two common rate decelerating proteins were gag and pol identified in all four highlighted branches, which is expected given that these are the longest genes in the genome. When controlling for gene size (Table 1), a different pattern emerges: vpu seems to have undergone a disproportionate portion of rate deceleration events, in particular in the branch leading to group M. In general, rate deceleration patterns for accessory proteins seemed to be branch-specific: nef and vif consisted 29% and 14%, respectively, of all rate deceleration events in the branch leading to SIVgor and group O. Vpu composed 14% of all rate deceleration events in the branch leading to group M, and rev contributed 20% of all rate decelerations in the branch leading to group O. The branch separating SIVcpz*pts* from other viruses displayed a more diverse pattern with relatively less rate decelerations in the env protein and more rate decelerations in the nef, vif and vpu proteins.

Figure 1. Proposed rate deceleration patterns in prominent branches, (A) shown on the phylogeny, (B) shown along the genome of the branch leading to SIVcpzpts (I), HIV-1 group M (II), SIVgor+P+O (III) and HIV-1 group O (IV). (C) Residues distribution for env_{458} along the HIV-1/SIV phylogeny. The evolutionary rate of this residue was found to be slower in HIV-1 group M than in the rest of the phylogeny. (D) Residues distribution for gag_{30} along the HIV-1/SIV phylogeny.

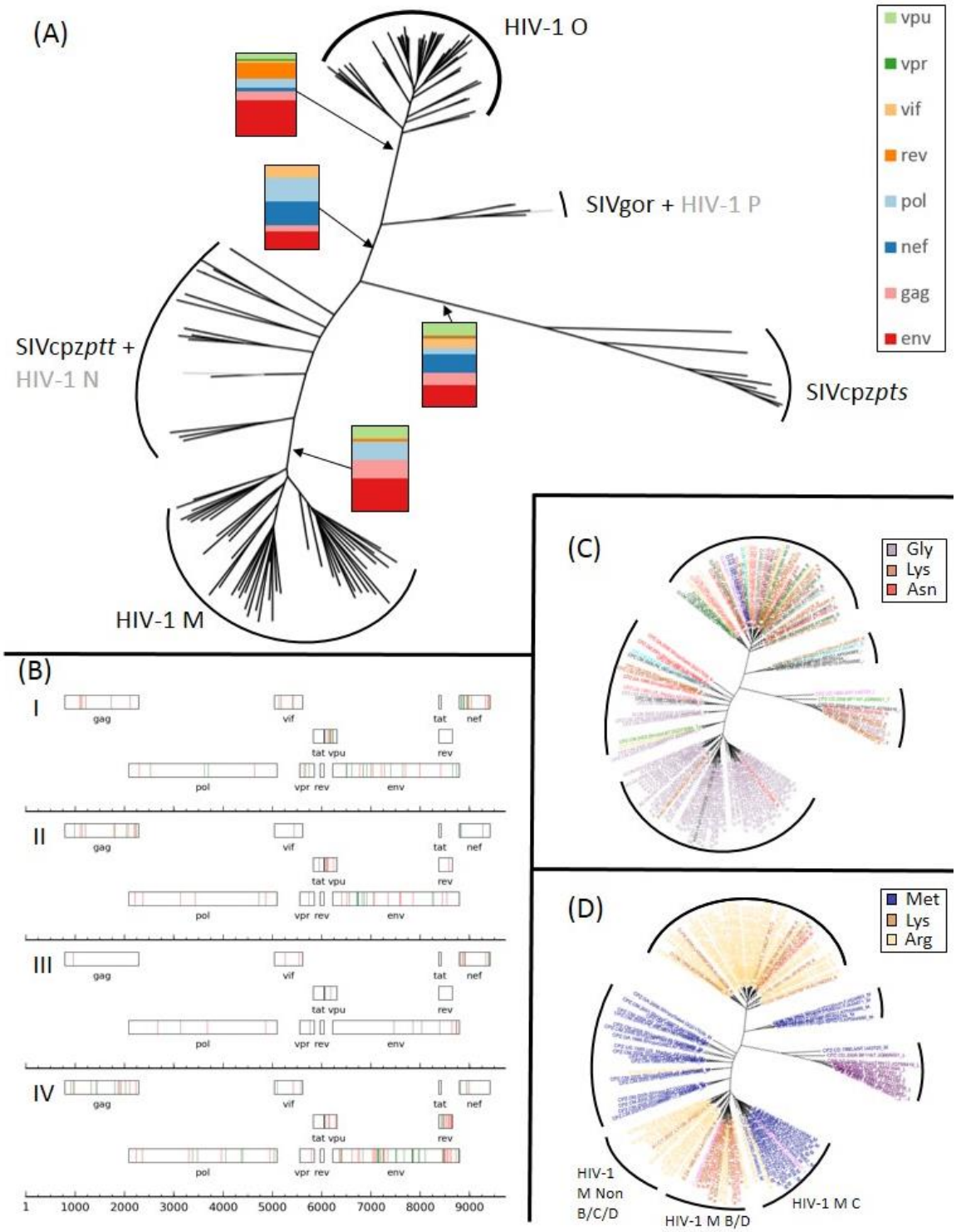


Table 1. Rate shifts as *percentage from total protein size* for prominent branches, for rate decelerating sites (upper) and rate accelerating sites (lower), colored by intensity.

DEC Branch/protein	env	gag	nef	pol	rev	tat	vif	vpr	vpu	Total mutations for branch
SIVcpz*pts*	1%	1%	3%	0%	0%	0%	2%	1%	5%	27
SIVgor+P+O	0%	0%	2%	0%	0%	0%	1%	0%	0%	14
Group O	2%	1%	1%	0%	8%	0%	1%	1%	4%	46
Group M	1%	1%	0%	1%	1%	0%	0%	0%	5%	28
ACC Branch/protein	**env**	**gag**	**nef**	**pol**	**rev**	**tat**	**vif**	**vpr**	**vpu**	**Total mutations for branch**
SIVcpz*pts*	1%	1%	3%	0%	0%	0%	1%	2%	5%	31
SIVgor+P+O	0%	0%	2%	0%	0%	0%	0%	1%	1%	10
Group O	3%	2%	1%	1%	3%	0%	2%	0%	1%	51
Group M	1%	1%	1%	0%	0%	1%	1%	1%	1%	23

Table 2: Maximum log-likelihood (LL) values for the analysis of the nine HIV-1/SIVcpz/SIVgor proteins under the rate shift and null models.

ORF	Rate Shift Model LL	Null Model LL	2ΔLL	P-value (χ^2_3)
gag	-29,478.3	-29,671.4	386	$<10^{-50}$
pol	-40,813.1	-41,097.2	568	$<10^{-100}$
Vif	-12,677.6	-12,756.3	157	$<10^{-30}$
Vpr	-5,328.1	-5,355.76	55	$<10^{-10}$
Tat	-8,918.23	-8,947.61	59	$<10^{-10}$
Rev	-12,662	-12,733.6	143	$<10^{-30}$
Vpu	-10,548.7	-10,664.1	231	$<10^{-50}$
Env	-92,347.6	-93,028.8	1362	$<10^{-100}$
Nef	-15,611.4	-15,804.1	385	$<10^{-50}$

Parallel rate shift events in HIV-1 groups M and O. Rate shift events common to several HIV-1 groups are of special interest, since they may better reflect the species barrier and not lineage-specific adaptation event. As is evident from Table 4, most parallel rate shift events of HIV-1 groups O and M (or at their closest precursor branch) were inferred as rate deceleration events, the majority of them in the highly conserved pol gene. Notably, the consensus amino acid identity in those parallel-occurring rate deceleration events varied between the M and O clade for all identified positions. Only a single parallel rate acceleration event was identified, in nef_8 of

HIV-1 groups M and O. This possibly indicates a relaxation of prior immune constraint as the ancestral SIV allele was maintained in many HIV-1 sequences.

Table 3. Rate shifts as *percentage from total rate shifts for prominent branches*, for rate decelerating sites (upper) and rate accelerating sites (lower), colored by intensity.

DEC	**env**	**gag**	**nef**	**pol**	**rev**	**tat**	**vif**	**vpr**	**vpu**
SIVcpz*pts*	26%	15%	22%	7%	0%	0%	11%	4%	15%
SIVgor+P+O	21%	7%	29%	29%	0%	0%	14%	0%	0%
Group O	43%	11%	4%	11%	20%	0%	2%	2%	7%
Group M	39%	21%	0%	21%	4%	0%	0%	0%	14%
ACC	**env**	**gag**	**nef**	**pol**	**rev**	**tat**	**vif**	**vpr**	**vpu**
SIVcpz*pts*	35%	10%	23%	10%	0%	0%	3%	6%	13%
SIVgor+P+O	40%	0%	40%	0%	0%	0%	0%	10%	10%
Group O	55%	18%	4%	12%	6%	0%	4%	0%	2%
Group M	43%	22%	13%	4%	0%	4%	4%	4%	4%

Table 4. Rate shifting sites identified independently in parallel in HIV-1 groups M and O or their immediate ancestor.

ORF	Rate Shifting site	Rate shift type	Branches	Consensus Amino Acid
pol	349	Deceleration	M, Pre O	M: E, Pre O: P
pol	406	Deceleration	Pre M, O	Pre M: S, O: V
pol	926	Deceleration	M, Pre O	M: K, Pre O: T
env	61	Deceleration	M, O	M: Y, O: T
env	772	Deceleration	M, O	M: R, O: S
rev	71	Deceleration	Pre M, O	Pre M: V, O: N
nef	8	Acceleration	M, O	M: S, O: S

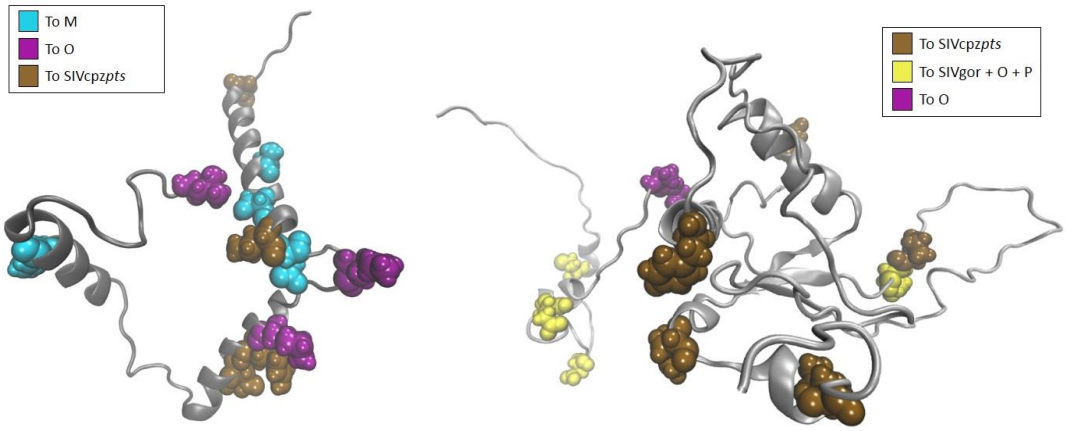
Rate shift events are often parallel to known macromolecular adaptation events. The intensive research around the HIV-1 pandemic and its sources identified several adaptation events of HIV-1 to its human host. These events include adapting to cellular variations such as differences in host receptor CD4; overcoming host innate immune restriction factors such as Tetherin (BST-2) and APOBEC3G; and of course, evasion from the adaptive immune system. We sought to evaluate whether the proposed rate-shifts agree with the known adaptation events.

Anti Tetherin/BST-2 adaptations. Tetherin is a trans-membrane antiviral restriction factor that disrupts the budding phase of the retroviral lifecycle, thus preventing the infection from spreading [18, 19]. SIVcpz and many more SIV's antagonize tetherin through interactions of their nef protein with the cytoplasmic tail of tetherin [20, 21]. Tetherin is a fast evolving protein,

and differs between chimpanzee, gorilla and human [22, 23]. In humans, a significant deletion in the cytoplasmic tail of tetherin rendered it invulnerable to the SIV's nef-based counteraction [21]. Therefore, in order to regain infectivity HIV-1 had to develop alternative way to encounter tetherin. In groups M and N, vpu protein adapted to antagonize tetherin through its transmembrane domain [24]. In group O the virus adapted its nef protein to counteract the human tetherin, but in a less efficient manner than in group M [25, 26]. No anti-tetherin adaptation event is known to have occurred in the rare group P [27]. In gorilla, the nef protein adapted to encounter its host's tetherin but also maintained its ability to encounter the chimpanzee tetherin [28, 29].

Our results agree to some extent with the described anti-tetherin adaptation events: 14% of all identified group M rate decelerations were found in the vpu protein, most of them in the tetherin-binding domain [Table 3, Figure 1b and Figure 2]. In addition, in HIV-1 group M the nef protein was identified with three rate accelerations, possibly indicating the loss of function of its anti-tetherin activity. This is also supported by a reintroduction experiment of HIV-1 group M strain into chimpanzee in which M-nef readapted to chimpanzee host, as the mutations required for M-nef to regain anti-tetherin activity in chimpanzees (nef_{163} and nef_{169}) were located proximally to nef_{157}, one of the rate accelerating sites of M-nef identified in this study [30]. Numerous O-nef rate acceleration events were suggested as well, perhaps as a proxy for the loss of chimpanzee/gorilla anti-tetherin function. A preponderance of acceleration and deceleration events were also found in the lineages leading to the ancestor of SIVgor and O and to SIVcpz*pts*; it is currently unclear what these events specify.

Figure 2. Projection of the identified rate decelerations onto the tertiary structure of vpu protein (left, PDB ID: 2N28) and nef protein (right, obtained from [31]) in the major lineages where it was identified.



Anti APOBEC3G adaptations. APOBEC3G (A3G) is a broad-range antiviral protein that is packaged into HIV-1 virions, and upon reverse transcription adds C→U mutations to the synthesized DNA strand, thus potentially creating nonviable genomes [32]. Lentiviruses use their vif protein to overcome this factor by degrading it as well as other antiviral proteins from the APOBEC3 family (such as A3F, A3D) [33]. A3G variants of human, chimpanzee and gorilla have

been studied and the vif recognition domain has been identified at A3G residues 126-132 [34]. Of those species, gorilla has a different residue at A3G position 129 as compared to that of human and chimpanzee A3Gs and indeed the SIVgor adaptation event to the different host A3G has been demonstrated by [35, 36]. It has been shown that SIVgor can replicate in the presence of human-, gorilla- and chimpanzee-A3G, but HIV-1 group M or SIVcpz cannot replicate in the presence of gorilla-A3G, indicative of a gain-of-function in the adaptation process of SIVgor to gorilla [36].

In line with these findings, our analysis shows that the branch leading to SIVgor and HIV-1 groups O and P experienced 14% of its rate decelerations in the vif protein, particularly in positions vif_{73} and vif_{167}. In HIV-1, vif_{73} is in close proximity to the A3G-binding motif vif_{69-72} that is also responsible in-part for interaction with APOBEC3F (which in itself contains differences in the vif interacting loop between chimpanzee and human/gorilla) [37-39]. O-vif experienced a single rate deceleration in position vif_{127}, which is located in the Cullin-5 interacting domain [38]. Lack of rate deceleration events in M-vif suggest no adaptation of this strain to the human host; this is supported by the observation that SIVcpz-vif can encounter human A3G [36].

Env adaptations. Another major barrier for host jumps lies at the cell entry level. CD4 is the target receptor of SIV and HIV, while CCR5 or CXCR4 act as co-receptors, recognized by the viral env protein which is exposed to the virus external surface. Previous studies mapped the genetic diversity between different chimpanzees' CD4 and revealed that there are several differences between chimpanzee subspecies as well as differences between chimpanzee versus gorilla and human, especially in the regions that are in close contact with env subunit gp120 [40, 41]. Some of those changes affect the glycosylation patterns of CD4, making these differences even more distinguishable [40, 41]. It is therefore likely that HIV-1 had to adapt to recognize these differences in CD4.

Our analysis overall supports a high level of deceleration and acceleration events found in Env, possibly reflecting the changes that occurred in the host CD4. For instance, many of the changes in M-env occurred in CD4 interacting residues of env_{374}, env_{455}, env_{456} and env_{458} [42]. However, we cannot rule out that the very high rate of evolution in Env, driven by the adaptive immune system [43], have led to changes that “mimic” a rate-shift. Indeed when accounting for the size of the Env protein, it seems that this protein is less pronounced in its rate shift distribution (Table 1)

Gag_{30} human adaptation marker. Position 30 of gag polyprotein has been suggested to be an adaptation event of HIV-1 to its human host, as it diversified from Methionine in SIVcpz*ptt* and SIVgor strains into Lysine or Arginine in many HIV-1 strains [44]. Reintroduction of HIV-1 to chimpanzees resulted in reversion of this position back to the conserved Met of SIVcpz*ptt* and therefore it was suggested as an adaptation event to the human host [44]. However, the ancestor of subtype C of group M probably contained Met in gag_{30}, suggesting complex interactions governing this human adaptation marker.

Our results indeed suggest an evolutionary rateshift event in gag_{30} yet are not conclusive about the branch in which the rateshift happened: we identified a rate acceleration event in the branch leading to HIV-1 group O with a marginal probability, while next two suspected branches were the one leading to HIV-1 group M and to HIV-1 group M subtypes B

and D [Figure 1d]. This is likely since this site displays a “content-shift” rather than a strong “rate-shift” (elaborated in the discussion).

Discussion

Understanding the micro-molecular changes in a pathogen when adapting to infect a new host species is of high importance. In this study we aimed to identify evolutionary rate shifts throughout the SIVcpz/SIVgor/HIV-1 phylogeny that may indicate a gain of function or loss of function in the various lentiviral proteins. A total of 271 rate shift events have been identified, most of them in the branches where the viruses crossed species or subspecies. We have shown that known adaptation events to the different hosts are supported by the unique rate shift patterns identified in prominent branches.

Rate decelerations identified in distinguishable branches may indicate adaptations to new hosts. This is most prominent in HIV-1 group O, where the rev protein was identified as having a relatively large amount of rate decelerations. To some extent, this may be interpreted as reflections of rate shifts in the overlapping reading frame of env [Figure 2b and Table S1], but the possibility of rev rate decelerating in group O due to possible adaptation event cannot be ruled out. In addition, the relative shortage of rate shift events in O-vif raises the possibility that O-vif may also antagonize gorilla and chimpanzee A3G, as the closely-related vif of SIVgor.

Since group M is the pandemic strain of HIV-1, one could hypothesize it will experience the most rate shift events among all HIV-1 groups, indicating a better adaptation to its human host. Surprisingly, our analysis revealed that group O had almost twice the amount of rate shift events, despite being a non-pandemic strain that remained localized mainly to infections in west-central Africa [45, 46]. This could indicate that the amount of rate shift events is not the determinant of a pandemic strain; it may rather reflect the relative height of the species barrier a strain had to overcome, as chimpanzee is genetically closer to human than gorilla [47, 48]. Furthermore, the lack of parallel rate decelerating sites in groups M and O of HIV-1 with the same amino acid for both clades may depict the independence of evolutionary processes on relatively different genetic background as those zoonotic transfers emerged from relatively divergent strains of SIV.

Previous studies about HIV-1 adaptation. Wain et al. initiated the first attempt to contrast SIVcpz and HIV-1 genetic sequences and identify positions that differ between the clades, as an indication for human adaptation [44]. This attempt resulted in the identification of gag_{30} as a Met in SIVcpz and SIVgor but basic amino acid (Lys or Arg) in the ancestors of groups M and O, and therefore was suggested as a human adaptation marker of HIV-1. The authors focused on identifying shared substitution events between the ancestors of groups M, O and N, hopefully representing vital adaptation events of SIVcpz and SIVgor to humans. However, adaptation to a host is in many times independent as demonstrated in the different anti-tetherin strategies used by groups M, O and even HIV-2 [49]. In addition, Wain et al. ignored the phylogenetic structure of the sequences. To the best of our knowledge, our study entails the first large-scale approach to detect adaptation events in the transition from non-human primates to humans, and relies on a robust phylogenetic modeling approach. We have discovered many clade-specific

adaptation events in many proteins of HIV-1, SIVgor and SIVcpz*pts*, and have reported sites in where groups M and O suggested to undergone parallel rateshifts.

Soares et al. attempted to measure molecular adaptation following host jumps. That study focused exclusively on accessory proteins (nef, vif, vpr, vpu and HIV-2-vpx) and attempted to identify sites under diversifying selection, however using unequal amount of sequences for each gene, analyzing the different proteins using different phylogeny [50]. In contrast, our approach used equal number of genomes in each protein, and more importantly each host contributed to all 9 ORFs.

Limitations. This study has several limitations. First, the availability of full genetic sequences of SIVcpz and SIVgor is still low, therefore reducing the statistical power of the analysis. Indeed, we noted that increasing the sample size of group M sequences that are much more available, led to detection of more rate shifting sites (see supplementary Text). We expect that the availability of more SIVgor sequences will increase the amount of rate shift events that are unique in that clade, possibly revealing more important sites for this lineage. Second, the method that we utilized to identify rate shifts is calibrated to identify dramatic changes in the evolutionary rate along a lineage. Accordingly, it cannot detect sites where a "content shift" occurred, i.e. the amino-acid changed and remained conserved in two complementary lineages, since this entails only a minor change in evolutionary rate. This is partially demonstrated in gag_{30}, where the rate of evolution changed mildly, while the content of the amino acid changed between chimpanzee and human viruses.

Implications. The compiled list of the positions suggested as rate decelerating can now serve as a guide for future functional studies that aim to understand the differences among HIV-1 and SIV proteins. Furthermore, the ability to track adaptation events by utilizing sequence data highlights the power of the method when studying emerging pandemics, strengthening the need to sequence full genomes of pathogens broadly.

Conclusions. Genetic sequences of viruses and specifically HIV-1 and SIV viruses can be harnessed to identify adaptation events of viral clades to their host. Our results suggest that innate and adaptive immunity serve as strong barriers for cross-species transmission events, and that these barriers imposed a strong selective pressure for viruses to adapt as they succeeded in crossing these barriers with increasing efficiency.

Methods

In order to collect sequences for this study, the Los Alamos HIV sequence database [available online at http://www.hiv.lanl.gov] was queried for HIV-1 sequences from the same strain that spanned all nine HIV-1 open reading frames (ORFs: *gag, pol, vif, vpr, tat, rev, vpu, env* and *nef*) and for SIVcpz and SIVgor strains that spanned the corresponding ORFs (with the exception of vpx). This led to 2004 sequences of HIV-1 group M, 45 sequences of HIV-1 group O, 9 sequences of HIV-1 group N, 2 sequences of HIV-1 group P, 4 sequences of SIVgor and 29 sequences of SIVcpz. Two different datasets were constructed: (i) with a very large number of group M sequences compared to the amount of group O sequences, and (ii) with an equal number of group M and group O sequences. Due to computational reasons, dataset (i) included 200 HIV-1 sequences, most of them group M. In both datasets, the sequences were sampled so that the n most distant strains (in terms of genetic distance) were sampled. Due to extremely high similarity, HIV-1 groups N and P sequences were reduced to a single representative strain from each. The IIIB_LAI strain was added manually as a reference sequence. The results in this study are reported mainly with datasets (ii), chosen since it allows comparing the result from group M and group O. Additional sites found with datasets (i) are reported in the supplementary material.

Initial multiple sequence alignments of the nine ORFs were performed using PRANK and iteratively improved until convergence [51], and then used in a concatenated form as input for reconstruction of the phylogenetic tree with PhyML [52]. We next used the reconstructed phylogeny as a guide tree to realign each ORF with PRANK. JpHMM was used to validate that the strains used in the analysis are not inter-group recombinants [53].

In order to identify evolutionary rate shifts, we used RASER [17] to analyze each of the nine ORFs separately, with the whole proteome phylogeny as input. Briefly, this method determines if the rate of evolution for some site in an input MSA is significantly different in the two sides of some branch of an input phylogeny, and reports the site and candidate branches where the rate shift occurred. In order to test for statistical significance of the identified rate shifting sites, a chi squared log likelihood test against a null model of no rate shifts was performed, similar to [17].

In order to test for sequence sampling effects on the identified rate shift patterns, we repeated the analysis with increased amount of group M sequences (n=183), reduced amount of group O sequences (n=13) and no SIVgor sequences, denoted as dataset (i). Analysis revealed more rate decelerations in the branch leading to group M than in dataset (ii) analysis (48 compared to 28, Table S2). 54% of the positions identified as rate-decelerating in group M in dataset (ii) (n=45) were also identified in dataset (i). Chi-squared test for differences in group M rate shift distributions showed no significant difference (p=0.11 and 0.87 for rate decelerations and accelerations, respectively), indicating that the patterns of rate shifts between the large group M sample and the smaller group M sample are similar.

Acknowledgements

The authors would like to thank to Dr. Osnat Penn, Prof. Eran Bacharach (Tel Aviv University) and Prof. Tal Pupko (Tel Aviv University) for their contribution to the inception of this study, and to Prof. Matthias Geyer (University of Bonn) for providing the nef protein model structure. This

work was supported in part by the German-Israeli Foundation (grant number I-1096-411.8-2015), the Israeli Science Foundation (grant number 1333/16), and by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University.

Bibliography

1. Hahn, B.H., et al., *AIDS as a zoonosis: scientific and public health implications.* Science, 2000. **287**(5453): p. 607-614.
2. Van Heuverswyn, F., et al., *Human immunodeficiency viruses: SIV infection in wild gorillas.* Nature, 2006. **444**(7116): p. 164-164.
3. D'arc, M., et al., *Origin of the HIV-1 group O epidemic in western lowland gorillas.* Proceedings of the National Academy of Sciences, 2015. **112**(11): p. E1343-E1352.
4. Korber, B., et al., *Timing the ancestor of the HIV-1 pandemic strains.* science, 2000. **288**(5472): p. 1789-1796.
5. Lemey, P., et al., *The molecular population genetics of HIV-1 group O.* Genetics, 2004. **167**(3): p. 1059-1068.
6. Takehisa, J., et al., *Origin and biology of simian immunodeficiency virus in wild-living western gorillas.* Journal of virology, 2009. **83**(4): p. 1635-1648.
7. Wertheim, J.O. and M. Worobey, *Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2.* PLoS computational biology, 2009. **5**(5): p. e1000377.
8. Bailes, E., et al., *Hybrid origin of SIV in chimpanzees.* Science, 2003. **300**(5626): p. 1713-1713.
9. Leitner, T., et al., *Sequence Diversity among Chimpanzee Simian Immunodeficiency Viruses (SIVcpz) Suggests that SIVcpz Pts Was Derived from SIVcpz Ptt through Additional Recombination Events.* AIDS research and human retroviruses, 2007. **23**(9): p. 1114-1118.
10. Li, Y., et al., *Eastern chimpanzees, but not bonobos, represent a simian immunodeficiency virus reservoir.* Journal of virology, 2012. **86**(19): p. 10776-10791.
11. Etienne, L., et al., *Gene loss and adaptation to hominids underlie the ancient origin of HIV-1.* Cell host & microbe, 2013. **14**(1): p. 85-92.
12. Bell, S.M. and T. Bedford, *Modern-day SIV viral diversity generated by extensive recombination and cross-species transmission.* PLOS Pathogens, 2017. **13**(7).
13. Longdon, B., et al., *The evolution and genetics of virus host shifts.* PLoS pathogens, 2014. **10**(11): p. e1004395.
14. Kuiken, T., et al., *Host species barriers to influenza virus infections.* Science, 2006. **312**(5772): p. 394-397.
15. Taubenberger, J.K. and J.C. Kash, *Influenza virus evolution, host adaptation, and pandemic formation.* Cell host & microbe, 2010. **7**(6): p. 440-451.
16. Knudsen, B. and M.M. Miyamoto, *A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins.* Proceedings of the National Academy of Sciences, 2001. **98**(25): p. 14512-14517.
17. Penn, O., et al., *Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes.* PLoS computational biology, 2008. **4**(11): p. e1000214.
18. Neil, S.J., T. Zang, and P.D. Bieniasz, *Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu.* Nature, 2008. **451**(7177): p. 425.

19. Van Damme, N., et al., *The interferon-induced protein BST-2 restricts HIV-1 release and is downregulated from the cell surface by the viral Vpu protein.* Cell host & microbe, 2008. **3**(4): p. 245-252.
20. Jia, B., et al., *Species-specific activity of SIV Nef and HIV-1 Vpu in overcoming restriction by tetherin/BST2.* PLoS pathogens, 2009. **5**(5): p. e1000429.
21. Zhang, F., et al., *Nef proteins from simian immunodeficiency viruses are tetherin antagonists.* Cell host & microbe, 2009. **6**(1): p. 54-67.
22. McNatt, M.W., et al., *Species-specific activity of HIV-1 Vpu and positive selection of tetherin transmembrane domain variants.* PLoS pathogens, 2009. **5**(2): p. e1000300.
23. Evans, D.T., et al., *BST-2/tetherin: a new component of the innate immune response to enveloped viruses.* Trends in microbiology, 2010. **18**(9): p. 388-396.
24. Sauter, D., et al., *Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains.* Cell host & microbe, 2009. **6**(5): p. 409-421.
25. Kluge, S.F., et al., *Nef proteins of epidemic HIV-1 group O strains antagonize human tetherin.* Cell host & microbe, 2014. **16**(5): p. 639-650.
26. Bego, M.G., et al., *Differential control of BST2 restriction and plasmacytoid dendritic cell antiviral response by antagonists encoded by HIV-1 group M and O strains.* Journal of virology, 2016. **90**(22): p. 10236-10246.
27. Sauter, D., et al., *HIV-1 Group P is unable to antagonize human tetherin by Vpu, Env or Nef.* Retrovirology, 2011. **8**(1): p. 103.
28. Lim, E.S., H.S. Malik, and M. Emerman, *Ancient adaptive evolution of tetherin shaped the functions of Vpu and Nef in human immunodeficiency virus and primate lentiviruses.* Journal of virology, 2010. **84**(14): p. 7124-7134.
29. Yang, S.J., et al., *Lack of adaptation to human tetherin in HIV-1 group O and P.* Retrovirology, 2011. **8**(1): p. 78.
30. Götz, N., et al., *Reacquisition of Nef-mediated tetherin antagonism in a single in vivo passage of HIV-1 through its original chimpanzee host.* Cell host & microbe, 2012. **12**(3): p. 373-380.
31. Geyer, M. and B.M. Peterlin, *Domain assembly, surface accessibility and sequence conservation in full length HIV-1 Nef.* FEBS letters, 2001. **496**(2-3): p. 91-95.
32. Mangeat, B., et al., *Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts.* Nature, 2003. **424**(6944): p. 99.
33. Sheehy, A.M., et al., *Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein.* Nature, 2002. **418**(6898): p. 646.
34. Russell, R.A., et al., *Distinct domains within APOBEC3G and APOBEC3F interact with separate regions of human immunodeficiency virus type 1 Vif.* Journal of virology, 2009. **83**(4): p. 1992-2003.
35. Compton, A.A. and M. Emerman, *Convergence and divergence in the evolution of the APOBEC3G-Vif interaction reveal ancient origins of simian immunodeficiency viruses.* PLoS pathogens, 2013. **9**(1): p. e1003135.
36. Letko, M., et al., *Vif proteins from diverse primate lentiviral lineages use the same binding site in APOBEC3G.* Journal of virology, 2013. **87**(21): p. 11861-11871.
37. Pery, E., et al., *Regulation of APOBEC3 proteins by a novel YXXL motif in human immunodeficiency virus type 1 Vif and simian immunodeficiency virus SIVagm Vif.* Journal of virology, 2009. **83**(5): p. 2374-2381.
38. Feng, Y., et al., *Suppression of APOBEC3-mediated restriction of HIV-1 by Vif.* Frontiers in microbiology, 2014. **5**.

39. Letko, M., et al., *Identification of the HIV-1 Vif and human APOBEC3G protein interface.* Cell reports, 2015. **13**(9): p. 1789-1799.
40. Hvilsom, C., et al., *Genetic subspecies diversity of the chimpanzee CD4 virus-receptor gene.* Genomics, 2008. **92**(5): p. 322-328.
41. Brown, K.E., *Parallel adaptation of CD4 in SIV-endemic Gorilla and Pan lineages*. 2014.
42. Zhou, T., et al., *Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01.* Science, 2010. **329**(5993): p. 811-817.
43. Yamaguchi-Kabata, Y. and T. Gojobori, *Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes.* Journal of Virology, 2000. **74**(9): p. 4335-4350.
44. Wain, L.V., et al., *Adaptation of HIV-1 to its human host.* Molecular biology and evolution, 2007. **24**(8): p. 1853-1860.
45. Peeters, M., et al., *Geographical distribution of HIV-1 group O viruses in Africa.* Aids, 1997. **11**(4): p. 493-498.
46. Leoz, M., et al., *The two-phase emergence of non pandemic HIV-1 group O in Cameroon.* PLoS pathogens, 2015. **11**(8): p. e1005029.
47. Hasegawa, M., H. Kishino, and T.-a. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.* Journal of molecular evolution, 1985. **22**(2): p. 160-174.
48. Ruvolo, M., *Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets.* Molecular biology and evolution, 1997. **14**(3): p. 248-265.
49. Le Tortorec, A. and S.J. Neil, *Antagonism to and intracellular sequestration of human tetherin by the human immunodeficiency virus type 2 envelope glycoprotein.* Journal of virology, 2009. **83**(22): p. 11966-11978.
50. Soares, A.E., M.A. Soares, and C.G. Schrago, *Positive selection on HIV accessory proteins and the analysis of molecular adaptation after interspecies transmission.* Journal of molecular evolution, 2008. **66**(6): p. 598.
51. Löytynoja, A. and N. Goldman, *An algorithm for progressive multiple alignment of sequences with insertions.* Proceedings of the National academy of sciences of the United States of America, 2005. **102**(30): p. 10557-10562.
52. Guindon, S., et al., *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.* Systematic biology, 2010. **59**(3): p. 307-321.
53. Schultz, A.-K., et al., *jpHMM: Improving the reliability of recombination prediction in HIV-1.* Nucleic acids research, 2009. **37**(suppl_2): p. W647-W651.