# Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic

Elodie Cassan, Anne-Muriel Arigon Chifolleau, Jean-Michel Mesnard, Antoine Gross, Olivier Gascuel

# Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic

Elodie Cassan[a,b,1], Anne-Muriel Arigon-Chifolleau[a], Jean-Michel Mesnard[b], Antoine Gross[b,1], and Olivier Gascuel[a,c,1]

[a]Institut de Biologie Computationnelle, Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM)–UMR 5506 CNRS and Université de Montpellier, 34095 Montpellier, France; [b]Centre d'Études d'Agents Pathogènes et Biotechologies pour la Santé (CPBS)–FRE 3689 CNRS and Université de Montpellier, 34293 Montpellier, France; and [c]Unité Bioinformatique Evolutive, Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI)–USR 3756 CNRS and Institut Pasteur, 75015 Paris, France

Edited by Robert C. Gallo, Institute of Human Virology, University of Maryland, School of Medicine, Baltimore, MD, and approved August 5, 2016 (received for review April 8, 2016)

**Recent experiments provide sound arguments in favor of the in vivo expression of the AntiSense Protein (ASP) of HIV-1. This putative protein is encoded on the antisense strand of the provirus genome and entirely overlapped by the *env* gene with reading frame −2. The existence of ASP was suggested in 1988, but is still controversial, and its function has yet to be determined. We used a large dataset of ∼23,000 HIV-1 and SIV sequences to study the origin, evolution, and conservation of the *asp* gene. We found that the ASP ORF is specific to group M of HIV-1, which is responsible for the human pandemic. Moreover, the correlation between the presence of *asp* and the prevalence of HIV-1 groups and M subtypes appeared to be statistically significant. We then looked for evidence of selection pressure acting on *asp*. Using computer simulations, we showed that the conservation of the ASP ORF in the group M could not be due to chance. Standard methods were ineffective in disentangling the two selection pressures imposed by both the Env and ASP proteins—an expected outcome with overlaps in frame −2. We thus developed a method based on careful evolutionary analysis of the presence/absence of *stop* codons, revealing that ASP does impose significant selection pressure. All of these results support the idea that *asp* is the 10th gene of HIV-1 group M and indicate a correlation with the spread of the pandemic.**

HIV-1 | *asp* and *env* genes | overlapping genes | phylogenetic analyses | selection pressure

It is well established that retroviruses are able to perform antisense transcription from the 3′ long terminal repeat (LTR) of their proviral genome (1, 2). In 1988, the existence of an ORF on the antisense strand of the HIV type 1 (HIV-1) genome was suggested (3). This ORF encodes the putative AntiSense Protein (ASP). The existence of this ORF and of the encoded protein was controversial for many years, but now several pieces of evidence argue in favor of its expression (see ref. 4 for an extensive review): (*i*) several polyadenylated antisense transcripts capable of encoding ASP have been characterized within HIV-1–infected cells (1, 5, 6); (*ii*) it was demonstrated that the full-length ASP protein can be expressed ex vivo from the HIV-1 3′ LTR (7); (*iii*) ASP has been detected in freshly infected cells (2, 8, 9); and (*iv*) two recent independent clinical studies have shown the in vivo expression of ASP by detecting a cell-mediated immune response against several ASP epitopes within 30% of individuals infected with subtype B viruses (10, 11) [a percentage similar to those observed with other HIV-1 proteins, e.g., Tat and Pol (12)]. Moreover, experimental results suggested that ASP could form stable aggregates, be located partially at the plasma membrane, and be associated with autophagy (4, 7, 8). Despite this accumulation of evidence, the existence of ASP is still questioned because, for example, defective ribosome products with immune response have been reported for several viruses including HIV-1 (13, 14). Elucidating the function of ASP is thus a major goal, but studying the evolutionary forces acting on ASP is also crucial.

A striking fact with ASP ORF (and a challenge in terms of bioinformatics analyses) is its location on the provirus genome, as it overlaps the *env* (envelope) gene. Overlapping genes are a common feature of viruses to "compress" their genome (15). However, as the same portion of DNA encodes for several proteins, their adaptability is strongly lowered (16). Proteins encoded by overlapping genes are generally accessory proteins that play a role in viral pathogenicity or spreading (17). ASP ORF overlaps *env* on the frame −2: the codon positions 1, 2, and 3 in *env* face positions 2, 1, and 3 in *asp*, respectively (Fig. 1*B*). Because the two most important positions of *env* and *asp* codons are opposite each other, there is particularly little flexibility to encode amino acids (18).

The aims of this study were to assess the presence and conservation of the ASP ORF in the HIV-1 and SIVcpz/gor (chimpanzee and gorilla) groups and subtypes and to demonstrate the selection pressure induced by ASP to confirm its importance in some of the mechanisms of the virus.

## Results

HIV-1 strains are classified into four phylogenetic groups: M, N, O, and P. These four groups resulted from four separate cross-species transmission events of Simian Immunodeficiency Virus (SIV) to humans (19). Group M is the pandemic group. It is divided into nine distinct subtypes, and more than 70 circulating recombinant forms (CRFs).

The ASP ORF is entirely overlapped by the *env* gene (Fig. 1*A*), which has several overlapping ORFs on different reading frames. The *env* gene contains five variable regions, separated by constant regions (21), and the Rev Response Element (RRE) (22), which is a

### Significance

**HIV-1 is commonly assumed to have nine genes. However, in 1988 a 10th gene was suggested, overlapped by the *env* gene, but read on the antisense strand. The corresponding protein was named AntiSense Protein (ASP). Several pieces of evidence argue in favor of ASP expression in vivo, but its function is still unknown. We performed the first evolutionary study of ASP, using a very large number of HIV-1 and SIV (simian) sequences. Our results show that ASP is specific to group M of HIV-1, which is responsible for the pandemic. Moreover, we demonstrated that evolutionary forces act to maintain the *asp* gene within the M sequences and showed a striking correlation of *asp* with the spread of the pandemic.**

Author contributions: E.C., A.-M.A.-C., J.-M.M., A.G., and O.G. designed research; E.C. and O.G. performed research; E.C. and O.G. contributed new analytic tools; E.C. and A.G. analyzed data; and E.C. and O.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: All of our data, alignments, methods, and detailed results are available from FigShare at https://figshare.com/s/9668ef62e84488d4787a.

[1]To whom correspondence may be addressed. Email: olivier.gascuel@pasteur.fr or antoine.gross@cpbs.cnrs.fr or elodie.cassan@lirmm.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1605739113/-/DCSupplemental.

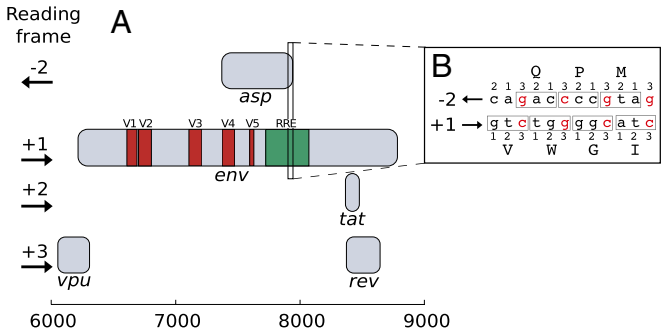



**Fig. 1.** Structure of the HIV-1 genome in the *env* gene region. (*A*) This region contains five overlapping ORFs: the *env* gene, the exon 2 of *tat*, the exon 2 of *rev*, the C-terminal extremity of *vpu*, and the ASP ORF. The *env* gene contains five variable regions (V1 to V5, reddish) and the RRE (greenish). (*B*) Overlapping sequences on the frames −2 and +1 [start of the *asp* region, HXB2 (20; GenBank accession no. K03455)].

highly structured RNA element that plays a role in the export of HIV-1 mRNAs.

**Data.** We downloaded all available and complete HIV-1/SIVcpz/gor *env* sequences and data annotations from the Los Alamos HIV Sequence Database (www.hiv.lanl.gov/content/index). We also used GenBank to retrieve the original version of some of the sequences. After deleting problematic sequences, we obtained 22,992 *env* sequences belonging to 3,931 individuals. Codon-based, multiple alignments were performed on the *env* gene and on the frame −2 of this gene. To avoid counting several times sequences that are very close to each other and belong to the same individual, we used two strategies: (*i*) we used the complete multiple alignment, but weighted the sequences so that each individual had a total weight of 1 (we then obtained the "weighted" alignment); and (*ii*) we randomly selected one sequence per individual when it was required for computational reasons. Most of our results and statistics are based on weighted sequences, except where otherwise specified. Details are provided in *SI Text*.

**Detection of the ASP ORF.** We based our analyses on the presence/absence of *start* and *stop* codons in frame −2 of the *env* region. The ASP ORF of the reference sequence HXB2 (20; GenBank accession no. K03455) has a length of 188 codons and is located between reference *env* positions 1,717 and 1,151. We thus searched all of our sequences for long DNA segments (>150 codons) with a *start* codon and no *stop* codon, read in frame −2 and located between these two reference positions. The analysis was carried out in the group M and an "out-of-M" dataset comprising all nonpandemic (N, O, P) HIV-1 and SIVcpz/gor sequences. For the sequences in group M and using the above criteria, we detected the ASP ORF for 77% of the (weighted) sequences. We clearly observed (Fig. 2) a region that is nearly free of *stop* codons and located between the reference positions of *asp*. At the beginning of the *asp* region, we note the presence of a *stop* codon for 14.5% of the sequences, located 12 codons after the *start* codon. However, most of these sequences (90%) belong to subtype A and its recombinants. One of the A recombinants in the *asp* region, namely CRF02_AG (112 individuals), has the early *stop* for ∼100% of the sequences, but only 7% of its sequences have the ASP ORF using our criteria. In contrast, in subtype A (240 sequences), ∼100% of sequences have the early *stop*, but a large percentage of them (∼90%) have an alternative *start* codon located 17 codons after the early *stop*; These sequences thus have a shorter version of ASP ORF, but still with a length of more than 150 codons.

In the out-of-M sequences, a number of *stop* codons are observed inside the *asp* region (Fig. 2). The *start* codon is not conserved (38% of sequences have a *start*/methionine codon), and less than 1.5% of out-of-M sequences have an ORF in *asp* region with length >150 codons.

**Recent Emergence of the ASP ORF.** The contrasting results between the pandemic group M and the other groups (out-of-M) led us to study the emergence and evolution of the ASP ORF using a phylogenetic approach. For this purpose, we inferred a maximum-likelihood tree using PhyML (23) (GTR+Γ4+I model, 1,000 bootstrap replicates) on a selection of sequences extracted from our complete alignment. We used 33 reference sequences (24) of group M subtypes and CRFs (A, B, C, D, F, G, H, J, CRF01_AE, CRF02_AG), 10 randomly selected sequences from group O, and the 40 sequences of the other out-of-M groups. To complete this phylogeny, we computed statistics from the complete, weighted alignment for all groups, subtypes, and CRFs. By using the same detection criteria as above, we measured the length of the longest ORF in the *asp* region and the fraction of sequences that had the ASP ORF.

The phylogeny (Fig. 3) clearly shows the four introductions of HIV-1 in the human population, corresponding to the four groups O, P, N, and M. The *start* codon corresponding to *asp* is present in most of the studied sequences. However, group O sequences and some exceptions (e.g., subtype H, prevalence ∼0.1%) do not have this *start* codon. The median length of the ORF in the *asp* region of the out-of-M groups increases when approaching group M: there are 66 codons in SIVcpz_Pts (from *Pan troglodytes schweinfurthii*), which increases to 125 codons in SIVcpz_Ptt (from *Pan troglodytes troglodytes*) that is closest to the group M. For group M sequences, the ASP ORF is present in 77% of the sequences with a median length of 182 codons. All of this indicates that the ASP ORF was created recently and that its emergence in HIV-1 is concomitant with the emergence of the group M. This recent de novo creation is further supported by the fact that ASP does not have any known homologs (*SI Text*). Interestingly, among SIVcpz_Ptt sequences, there is one sequence that possesses the ASP ORF in its entirety. This simian ASP has the same structural features (4) as the human ones. However, it is phylogenetically remote, and we do not have enough SIVcpz_Ptt sequences to figure out whether ASP appeared in the HIV and SIV genomes independently or, rather, if ASP first appeared in the SIVcpz_Ptt genome and was maintained when the HIV-1 group M emerged from it (*SI Text* and Fig. S1).

However, the fraction of M sequences that have ASP ORF varies among subtypes and CRFs. The less prevalent subtypes (D, F, J, H, K, total prevalence ∼3%) have the ASP ORF for less than 45% of their sequences. As already mentioned, only a few sequences (7%) in CRF02_AG (prevalence 7.7%) have the ASP

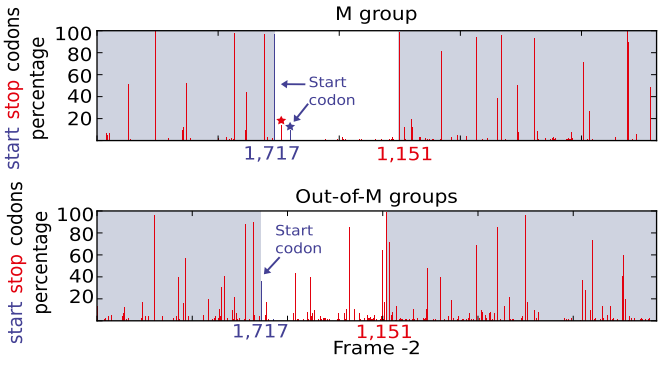
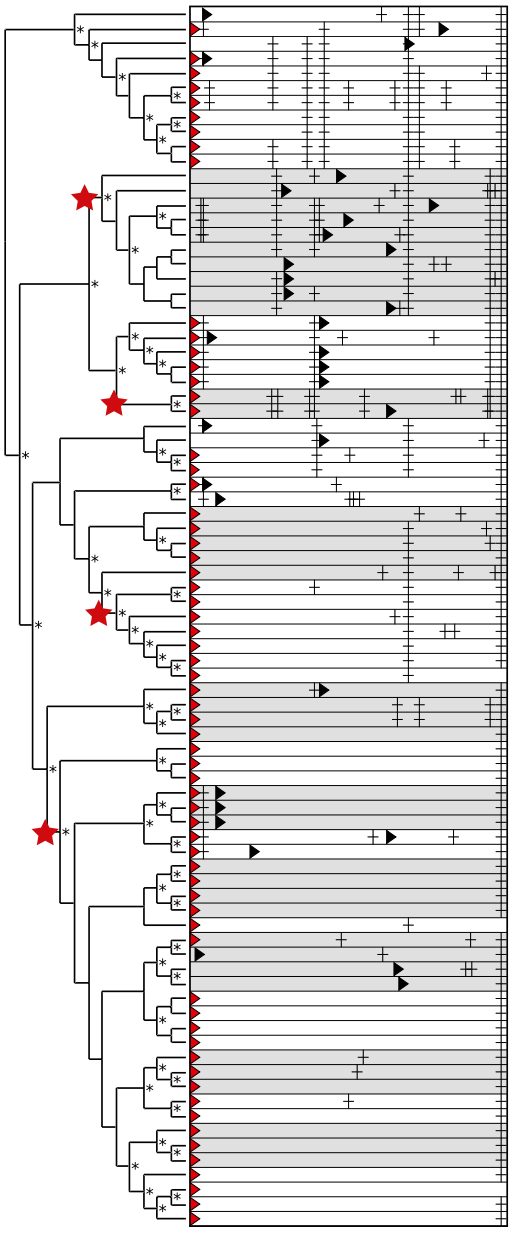



**Fig. 2.** Detection of the ASP ORF. Weighted percentages of *start* (blue) and *stop* (red) codons in frame −2, in the groups M and out-of-M. The *asp* region (white area) is located between the *env* positions 1,717 and 1,151 (HXB2 reference). The red star indicates an early *stop* codon that is specific to subtype A and A recombinants. This early *stop* codon is followed by an alternative *start* codon (blue star) in most of the A sequences and certain A recombinants.



| Group | Subtype | Median length | # sequences with ASP ORF | | Prevalence |
|---|---|---|---|---|---|
| SIVcpz Pts | | 66 | 0/8 | (0%) | \ |
| O | | 35 | 0/44 | (0%) | < 0.1 % |
| SIVgor | | 83 | 0/3 | (0%) | \ |
| P | | 65 | 0/1 | (0%) | ~ 0.0 % |
| SIVcpz Ptt | | 89 | 0/5 | (0%) | \ |
| SIVcpz Ptt | | 139 | 0/5 | (0%) | \ |
| N | | 134 | 0/7 | (0%) | ~0.0 % |
| SIVcpz Ptt | | 125 | 1/3 | (33%) | \ |
| M | CRF01_AE | 180 | 395/443 | (89%) | 5.1 % |
| | A | 160 | 177/240 | (74%) | 12 % |
| | CRF02_AG | 63 | 8/112 | (7%) | 7.7 % |
| | G | 174 | 56/63 | (88%) | 4.6 % |
| | J | 189 | 1/2 | (50%) | 0.1 % |
| | H | 78 | 0/4 | (0%) | 0.1 % |
| | C | 178 | 728/864 | (84%) | 48.2 % |
| | F | 114 | 11/33 | (32%) | 0.4 % |
| | K | 143 | 1/2 | (50%) | 0.1 % |
| | D | 168 | 53/107 | (50%) | 2.5 % |
| | B | 186 | 1,111/1,307 | (85%) | 11.3 % |

**Fig. 3.** Recent emergence of the ASP ORF using a phylogenetic approach. This phylogeny (* = bootstrap >80%) of the *env* gene contains reference sequences from HIV-1 and SIVcpz/gor groups, subtypes, and CRFs. The four distinct simian/human transmissions are indicated by red stars. For each of the sequences, we show the distribution of *start* codons (red triangle; black triangle for alternative *start*) and *stop* codons (black cross) in the *asp* region. For each group, subtype, and CRF, the table provides the median length of the ASP ORF, the fraction of sequences with the ASP ORF (length > 150 codons), and the prevalence in the human population (44).

ORF. In contrast, the other prevalent subtypes and CRFs (A, B, C, G, and CRF01_AE, total prevalence = 81%) have ASP for 84% of their sequences. We thus see a clear correlation ($P$ value = 0.003) (*Materials and Methods*): prevalent M subtypes and CRFs (except CRF02_AG) have the ASP ORF for a large majority of sequences, whereas low-prevalence subtypes and nonpandemic groups (N, O, P) have the ASP ORF in a minority of sequences (none in some groups/subtypes). This correlation is confirmed when accounting for phylogenetic correlation (Bayes factor = 3.8) (*Materials and Methods*).

The ASP ORF is present in 84% of sequences for prevalent subtypes (A, B, C, G) and CRF01_AE. This fraction is quite high and is not likely to be due to chance, as we shall see. However, 16% of sequences in these subtypes and CRF do not have the ASP ORF. This level of absence is similar to the one observed with *nef* [13.5% (25)], an accessory gene, and higher than the ~5% that we found for *env* and *pol* (two obligatory genes) by scanning for the presence of *stop* codons [all available Los Alamos database sequences (December 2015)]. This nonnegligible fraction of *stop*s in *env* and *pol* is explained by both sequencing errors (26) and the fact that some of the sequences are defective (27). The higher level of absence with ASP ORF is explained by the fact that *asp* is an accessory gene. As expected, ASP ORF was lost not only in some of the M subtypes and CRFs, but also

in some of the individuals of prevalent subtypes and CRF01_AE, where 12% of individuals in our dataset do not have any sequence with the ASP ORF, whereas 81% of individuals have the ASP ORF for all of their sequences. These 12%, added to the 5% of sequencing errors and defective sequences observed with *env* and *pol*, roughly explain the 16% of *asp* absence in prevalent subtypes and CRF01_AE.

**Conservation of the ASP ORF.** Previous analyses indicate that the ASP ORF is present in a large fraction of the group M sequences. We used computer simulations to demonstrate that there is a very low probability that this is due to chance.

We first estimated the probability of observing an ORF with ASP length overlapping the *env* gene in frame −2. For this purpose, we randomly generated sequences with the same length (856 codons) as the *env* gene of HXB2 and the same codon usage as HIV-1 (www.kazusa.or.jp/codon/). In this case, the probability of an ORF of length 180 in frame −2 is ∼3%. This is a low probability, but ASP is the longest overlapping ORF present in HXB2 in any reading frame (3), and one could argue that having such an ORF in the whole HIV-1 genome is quite likely. Using the same method as above, we thus generated sequences having the same length (3,239 codons) as HXB2 and searched for an ORF of length 180 in the five possible reading frames. The probability in this case is ∼19%. This is still a relatively low probability, but clearly we cannot reject that the presence of the ASP ORF at the root of HIV-1 M is merely due to chance. However, we observed the ASP ORF in 77% of our M sequences. The question then is: if we assume the ASP ORF presence at the root of HIV-1 M, would we have a significant chance of observing its presence in so many sequences at the phylogeny tips?

To answer this question, we simulated the evolution of the *env* gene along phylogenies inferred using 350 randomly selected strains. We used PhyML and codonPhyML (28) to infer 10 such phylogenies. For each one, we performed 100 codon-based simulations using Alf (29), starting from the *env* gene of HXB2 at the tree root (*Materials and Methods*). The maximum percentage of the tip sequences where the ASP ORF was still present (across 1,000 datasets) was equal to 67%, and on average, the ASP ORF was conserved in only 42% of sequences.

These results show that there is an extremely low probability that our observation of 77% on the conservation of the ASP ORF in the M group is due to chance, thus revealing a selection pressure that tends to conserve ASP. In the following section, we show that this selection pressure is also detected at the sequence level. We first used standard methods (evolutionary rate, nonsynonymous versus synonymous substitutions, codon usage), but none of these approaches provided a significant signal due to the specificity of frame −2 (30, 31) (*SI Text*). Thus, we developed a method dedicated to the frame −2.

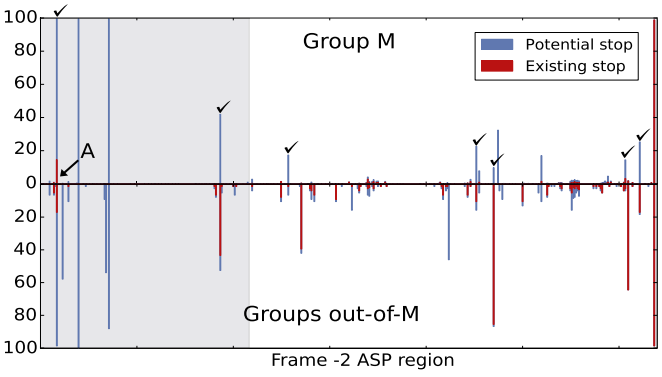



**Fig. 4.** Percentage of potential (blue) and existing (red) *stop* codons in the *asp* region. Shaded area = RRE region. In the top panel, the *stop* codon that is conserved in 14.5% of sequences is characteristic of subtype A and A recombinants. The labeled sites contain potential *stop* codons for more than 10% of M sequences and existing *stop* codons in some of the out-of-M sequences.

**Measuring Selection Pressure.** This method is based on a careful analysis of the presence/absence of *start* and *stop* codons. Let us first discuss the *start* codon (included in the RRE) (Fig. 1). Having a *start* codon (*atg*) on frame −2 implies the presence of the two codons on (*env*) frame +1: *xxc* followed by *atx* (*x* represents any nucleotide) (Fig. 1*B*). Any mutation on the third base of the first codon (*xxc*) leads to the disappearance of the *start* codon on frame −2. However, for all amino acids that are encoded by a codon ending with *c*, a mutation of *c* into *t* leaves the amino acid unchanged. Thus, on frame −2, the *start* codon is never imposed by the sense gene and may appear/disappear synonymously. In fact, we observe (Fig. 2) that the *start* codon is highly frequent (97%) among the sequences of HIV-1 M. This is a first indication of selection pressure acting on *asp*. This selection effect is not found in out-of-M sequences, where the *start* codon is present in only 38% of sequences. In other words, this selection effect is most likely attributable to the maintenance of the ASP ORF in HIV-1 M, and not to some other cause (e.g., RRE structural constraints), which would impact both M and out-of-M sequences. However, the *start* codon is not included in further statistical significance calculations, as the RRE secondary structure could differ between M and out-of-M sequences.

Moreover, ∼90% of A sequences (240 individuals) with an early *stop* codon have an alternative *start*/methionine codon (Fig. 2), which is observed in only 4.8% of the sequences that do not have this early *stop*. This is another indication of the pressure to maintain the ASP ORF, acting specifically in subtype A and some A recombinants.

We used the same type of reasoning with the *stop* codons. There are actually two types of *stop* codons on frame −2 (see *Materials and Methods* for details). Some are imposed by the coding of Env, as is the case for the terminal *stop*. This explains why nearly 99% of our sequences do have the terminal *stop* codon, even when they do not have the ASP ORF. The other type of *stop*s may appear/disappear without modifying Env, just as with the *start* codon. When they are absent, we call them "potential" *stop*s, meaning that they can mutate into a *stop* codon synonymously for Env. Fig. 4 displays the frequency of the potential/existing *stop* codons in the *asp* region of both M and out-of-M sequences. For 11 sites, we observe potential *stop*s for more than 10% of M sequences, but, on average, these sites contain *stop*s for only 0.5% sequences (disregarding A and A recombinant sequences with early *stop* and alternative *start*). Moreover, for seven of these sites (labeled in Fig. 4), *stop* codons are actually observed in out-of-M sequences. When we remove the RRE region, which clearly imposes strong structural constraints, five of these seven sites remain, with (on average) <0.7% and >23% *stop*s in M and out-of-M sequences, respectively. We thus have a second, strong indication of the selection pressure to maintain ASP: *stop* codons could be observed and are actually observed in out-of-M sequences, but not in M sequences. Note, however, that we cannot exclude the existence of some yet-unknown constraint (e.g., large-scale RNA structure), differing in M and out-of-M groups and inducing such an effect.

To measure the statistical significance of these findings, we used a method inspired by Firth (32), but dedicated to frame −2 and the analysis of potential *stop* codons. The original principle is to count the number of synonymous mutations on the reference gene and check whether this number is significantly less than expected. However, in frame −2 most synonymous mutations in the reference frame are also synonymous in the overlapping frame due to the fact that the third codon positions face each other (Fig. 1*B*). This makes the original method unable to detect any global pressure induced by ASP, although it performs well with other reading frames (Fig. S2). Our adaptation involves focusing solely on the *stop* codons and counting the synonymous mutations corresponding to the change of potential *stop*s into existing *stop*s. We compute a statistic that is equal to the expected number of such mutations minus the observed number of

mutations; then we perform a Z-score–like test to derive a $P$ value (*Materials and Methods*).

For sequences of group M (3,855 sequences), we observe a positive result ($Z = 6.47$, $P$ value $= 10^{-10}$) in the *asp* region. In other words, there are fewer *stop*s than expected if the null model was true, and we clearly see the presence of selection pressure. When excluding the RRE region, which induces strong structural constraints (Fig. 4), the difference is still highly significant, with fewer *stop*s than expected ($Z = 2.9$, $P$ value $= 0.006$).

As a negative control, we applied the same method to the rest of the *env* gene (i.e., the *env-asp* region). In this case, the number of observed *stop*s in group M was greater than expected ($Z = -1.9$, $P$ value $= 0.06$), which further supports the significance of our observations in the *asp* region. Moreover, for out-of-M groups (76 sequences) we did not detect any significant difference between the observed and expected number of *stop*s, both in the *asp* region ($Z = 1.53$, $P$ value $= 0.13$) and in the *env-asp* region ($Z = 1.04$, $P$ value $= 0.3$). When excluding the RRE, we obtained a negative score ($Z = -0.17$, $P$ value $= 0.71$), meaning that the positive score ($Z = 1.53$) observed in the whole *asp* region was induced by the RRE.

To summarize, these results indicate that the ASP ORF in group M is maintained selectively by conserving the *start* codon and avoiding *stop* codons. This finding is highly significant and explains, at the sequence level, the computer simulation results presented in the previous section. Moreover, the same results were not observed in either the *env-asp* region of M sequences or in the *asp* and *env-asp* regions of out-of-M sequences.

## Discussion

By looking at the presence/absence of *start* and *stop* codons, we showed the existence of the ASP ORF in most of the prevalent subtypes and recombinant forms of HIV-1 M (with the notable exception of the CRF02_AG recombinant). In contrast, the ASP ORF appears to be absent in the other nonpandemic subtypes and human groups. These results indicate that the ASP ORF was created recently, concomitantly with the emergence of HIV-1 M, and is specific to this group, which is responsible for the human pandemic.

However, a relatively large fraction (16%) of M sequences do not have the ASP ORF, even in the prevalent subtypes and recombinant forms. This level of absence, and the fact that *asp* is a recent de novo creation, indicate that *asp* is an accessory gene that can be lost without compromising the viability of the virus. This could explain why finding the function of ASP has proven to be so difficult since its discovery in the 1990s.

It has been observed that viral de novo genes often play a role in pathogenicity or spreading, rather than being central to viral replication or structure (17, 33, 34). This scheme most likely applies to *asp*. The striking correlation between the presence of *asp* in nonpandemic groups and M subtypes and CRFs, and their prevalence, strongly supports the idea that ASP could play a role in spreading. This contradicts a common argument that the difference among HIV-1 groups in terms of prevalence and impact in human populations has no molecular basis, but is mostly due to social changes in ∼1960 in central Africa, where the group M was already well established (35).

Our simulations and careful analyses of *start* and *stop* codons all indicate the presence of selection pressure to maintain the ASP ORF. However, we have not been able to reveal selection pressure acting at the amino acid level, which could be related to the structure and function of ASP protein. Similar difficulties were encountered in other studies on overlapping genes (e.g., ref. 36). Although such pressure possibly exists, one hypothesis could be that the function of ASP is essentially related to the expression of its ORF, for example, by interfering with the regulation of *env*. Alternatively, ASP could be involved in some mechanisms that alter the cellular functioning of specific cells, such as dendritic cells (2), for example, by forming aggregates (7). An interaction is possible with the recently discovered HIV antisense long noncoding RNA, which was suggested to modulate viral transcription (5, 37). However, the transcripts seem to be different (38), and the selection pressure acting on ASP ORF to maintain the *start* codon and avoid *stop* codons is clearly in favor of a coding part. Such mechanisms could be sufficient to improve the fitness of viral strains that have ASP and produce the observations reported in this study. Deciphering the function of ASP is now a major goal for further research, as all of our results support the idea that *asp* is the 10th gene of HIV-1 group M and indicate a correlation with the spread of the pandemic.

## Materials and Methods

**Correlation Between Prevalence and Presence of ASP.** We used an exact Spearman rank correlation test to demonstrate the statistical significance of the correlation between the presence of ASP in N, O, and P groups and M subtypes and CRFs and their prevalence (data in Fig. 3). Only the two prevalent CRFs (i.e., CRF01_AE and CRF02_AG, with prevalence 5.1% and 7.7%, respectively) were considered because the other CRFs appeared recently in most cases (hence their low prevalence) and conform with the original subtype from which they derive [e.g., all but one of the CRFs deriving from subtype B in the *asp* region (19) have ASP, just like the subtype B]. Using the *rankor* function ("Spearman," "exact," and "greater" options) of the *pvrank* R package, we obtained a rho value of 0.705 and a one-sided $P$ value of 0.003.

To account for phylogenetic correlation, we used BayesTraits V2 (39). As there is no consensus on the phylogeny of the HIV-1 groups and subtypes, we computed 1,000 PhyML trees (GTR+Γ6) with one randomly selected sequence per subtype, prevalent CRF, and nonpandemic group. We then launched BayesTraits with this set of trees, and the same presence and prevalence values as in previous test. The mean log of the Bayes Factor (correlation model versus independence assumption) was equal to 3.8, that is, again, a strong evidence for correlation.

**Statistical Significance of ASP ORF Conservation Using Simulations.** Our statistics on ASP ORF conservation are based on computer simulations. The basic principle is to assume that the ASP ORF was present at the phylogeny root and then simulate the evolution of sequences (read in *env* frame +1) along the tree and count the number of tip sequences that still have the ASP ORF. For this purpose, we selected 350 *env* sequences of group M at random with at most one sequence per individual. To root the tree, we added one *env* sequence of the closest SIVcpz_ptt group (GenBank entry: DQ373064). Ten samples of 350+1 sequences were obtained in this manner. For each sample, we estimated a phylogenetic tree using PhyML (GTR+FreeRate, six rate categories). Branch lengths were re-estimated for the *asp* region, using CodonPhyML with empirical codon model (40) and *env* codon usage. Having rooted the tree with the SIVcpz_ptt sequence, we ran simulation using Alf (same model options as CodonPhyML) with sequences evolving along this tree, starting with the *asp* region of HXB2 at the tree root. Moreover, specific codon rates were used to account for the variability of rates in the *asp* region. We used three codon categories, corresponding to the RRE, the variable regions V4 and V5, and the remaining codons. The rate of each category was estimated using the tree length estimated for that category by CodonPhyML (same options), divided by the tree length for the whole *asp* region. Finally, we calculated the percentage of tip sequences that have the ASP ORF in frame −2 (ORF length > 150 codons).

**Obligatory and Potential *stop* Codons.** In all HIV-1 groups, the final *stop* codon is highly conserved. This *stop* codon is obligatory in most sequences. It is induced by Env, which contains a phenylalanine followed by a tyrosine at that position for 99% of sequences. At the nucleotide level, we thus have one of the four possibilities: *tt**{c,t}** ta{c,t}* (the overlap is in boldface), and on the opposite strand in frame −2, we necessarily have one of the two *stop* codons *ta{g,a}*.

Potential *stop* codons correspond to particular configurations, easily derived from the genetic code. For example, let us consider *tga*, one of the three *stop* codons. Having *tga* in frame −2 imposes having the two codons *xxt cax* in frame +1. The second codon (*cax*) encodes for histidine or glutamine and cannot be mutated synonymously on the first and second overlapping positions. In contrast, the first codon (*xxt*) can be mutated in a number of ways on the third position without changing the corresponding amino acid. For example, *aac* (asparagine) ↔ *aat* (asparagine), whereas in frame −2 we have *tgg* (tryptophan) ↔ *tga* (*stop* codon). Thus, *aac cax* defines a potential *stop*, whereas *aat cax* corresponds to an existing *stop*. Other potential *stop*s are similar, and their mutations into existing *stop*s always involve the third position of the first codon (hence our restriction to the third codon position; see below).

To measure the statistical significance of findings in Fig. 4, we used some ideas from ref. 32. Our adaptation involves focusing solely on the frame −2 and on synonymous sites corresponding to potential/existing *stop*s. For an aligned

EVOLUTION

sequence pair, we analyzed synonymous sites (i.e., the same Env amino acid pair is encoded in both sequences) for which, in the first sequence (denoted as sequence 1), there is a potential *stop* in frame −2. For each of these sites, we compared the number (0 or 1) of *stop* codons present in the second sequence (denoted as sequence 2) and the expected number of *stop* codons, assuming no selection pressure. For this purpose, we used DNADIST (41) with the F84 substitution model to estimate the evolutionary distance (δ) between both sequences. We restricted ourselves to third codon positions and synonymous sites, using both the *asp* and *env-asp* regions. We thus estimated the expected number of substitutions per site being synonymous regarding Env. The transition/transversion ratio (κ) and nucleotide frequencies (required by DNADIST) were estimated globally from all sequence pairs assuming HKY substitution model (nearly identical to F84, easy formulae). We then estimated, for each synonymous site with a potential *stop*, the expected number of *stop* codons in sequence 2 in frame −2. This estimation was achieved assuming HKY, using the previously estimated evolutionary distance (δ) and parameters (κ, nucleotide frequencies). The difference between this expected number and the observed number of *stop* codons in sequence 2 in frame −2 formed the statistics that we used to assess the significance of our observations (Fig. 4). Assuming no selection pressure, both expected and observed numbers of *stop* codons should be nearly equal. Conversely, assuming that ASP imposes some selection pressure, the expected number of *stop* codons should be larger than the actual number of *stop* codons, which is close to zero in group M (Fig. 4). Assuming a Poisson process (42), the variance of this statistic is equal to its mean. The alignment of the sequence pair being studied was scanned for sites corresponding to this pattern, summing for this pair the expectations, observations, and variances of all occurrences.

To select the list of sequence pairs, we used an unrooted phylogenetic tree [FastTree with GTR+CAT (43); one-per-individual multiple alignment; group M: 3,855 sequences; groups out-of-M: 76 sequences], tracing around the outside of a two-dimensional drawing of the tree. A sequence pair corresponded to two neighboring leaves in the tree. This procedure (derived from ref. 32) was used to account for the evolutionary dependency of the sequences: as every branch in the tree was run twice, we multiplied by 2 (worst case analysis) the variance computed under the independence assumption. We then summed our statistics over all sites and all sequences in the tree and computed a Z-score from which a *P* value was computed under the assumption of a normal distribution. A positive Z-score indicated that we had fewer *stop* codons than expected, assuming no selection pressure. Because the sequence pairs depended on the two-dimensional tree drawing, we randomly rotated subtrees and obtained a new set of sequence pairs from which the same statistics were computed. We obtained 10 replicates in this manner, providing nearly the same results that were averaged (e.g., Z-score of the whole *asp* region = 6.47 ± 0.23 and 1.53 ± 0.21 for groups M and out-of-M, respectively).

**ACKNOWLEDGMENTS.** We thank M. Jung, S. Laverdure, S. Lèbre, and V. Lefort for their help initiating this study and T. Stadler, E. Simon-Loriere, and three anonymous referees for their comments on the manuscript. This project and a PhD grant (to E.C.) were supported by the "Projet Exploratoire Premier Soutien (PEPS) de Site CNRS/Université de Montpellier: Comprendre les maladies émergentes et les épidémies."

1. Landry S, et al. (2007) Detection, characterization and regulation of antisense transcripts in HIV-1. *Retrovirology* 4:71.
2. Laverdure S, et al. (2012) HIV-1 antisense transcription is preferentially activated in primary monocyte-derived cells. *J Virol* 86(24):13785–13789.
3. Miller RH (1988) Human immunodeficiency virus may encode a novel protein on the genomic DNA plus strand. *Science* 239(4846):1420–1422.
4. Torresilla C, Mesnard JM, Barbeau B (2015) Reviving an old HIV-1 gene: The HIV-1 antisense protein. *Curr HIV Res* 13(2):117–124.
5. Kobayashi-Ishihara M, et al. (2012) HIV-1-encoded antisense RNA suppresses viral replication for a prolonged period. *Retrovirology* 9:38.
6. Barbagallo MS, Birch KE, Deacon NJ, Mosse JA (2012) Potential control of human immunodeficiency virus type 1 asp expression by alternative splicing in the upstream untranslated region. *DNA Cell Biol* 31(7):1303–1313.
7. Torresilla C, et al. (2013) Detection of the HIV-1 minus-strand-encoded antisense protein and its association with autophagy. *J Virol* 87(9):5089–5105.
8. Clerc I, et al. (2011) Polarized expression of the membrane ASP protein derived from HIV-1 antisense transcription in T cells. *Retrovirology* 8:74.
9. Briquet S, Vaquero C (2002) Immunolocalization studies of an antisense protein in HIV-1-infected cells and viral particles. *Virology* 292(2):177–184.
10. Bet A, et al. (2015) The HIV-1 antisense protein (ASP) induces CD8 T cell responses during chronic infection. *Retrovirology* 12:15.
11. Berger CT, et al. (2015) Immune screening identifies novel T cell targets encoded by antisense reading frames of HIV-1. *J Virol* 89(7):4015–4019.
12. Frahm N, et al. (2004) Consistent cytotoxic-T-lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities. *J Virol* 78(5):2187–2200.
13. Cardinaud S, et al. (2004) Identification of cryptic MHC I-restricted epitopes encoded by HIV-1 alternative reading frames. *J Exp Med* 199(8):1053–1063.
14. Maness NJ, et al. (2010) CD8+ T cell recognition of cryptic epitopes is a ubiquitous feature of AIDS virus infection. *J Virol* 84(21):11569–11574.
15. Chirico N, Vianelli A, Belshaw R (2010) Why genes overlap in viruses. *Proc Biol Sci* 277(1701):3809–3817.
16. Simon-Loriere E, Holmes EC, Pagán I (2013) The effect of gene overlapping on the rate of RNA virus evolution. *Mol Biol Evol* 30(8):1916–1928.
17. Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D (2009) Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *J Virol* 83(20):10719–10736.
18. Smith TF, Waterman MS (1980) Protein constraints induced by multiframe encoding. *Math Biosci* 49(1):17–26.
19. Hemelaar J (2012) The origin and diversity of the HIV-1 pandemic. *Trends Mol Med* 18(3):182–192.
20. Ratner L, et al. (1985) Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature* 313(6000):277–284.
21. Willey RL, et al. (1986) Identification of conserved and divergent domains within the envelope gene of the acquired immunodeficiency syndrome retrovirus. *Proc Natl Acad Sci USA* 83(14):5038–5042.
22. Cullen BR (2003) Nuclear mRNA export: Insights from virology. *Trends Biochem Sci* 28(8):419–424.
23. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–321.
24. Leitner T, Korber B, Daniels M, Calef C, Foley B (2005) HIV-1 subtype and circulating recombinant form (CRF) reference sequences, 2005. *HIV Seq Compend* 2005:41–48.
25. Pushker R, Jacqué JM, Shields DC (2010) Meta-analysis to test the association of HIV-1 nef amino acid differences and deletions with disease progression. *J Virol* 84(7):3644–3653.
26. Halin M, et al. (2009) Human T-cell leukemia virus type 2 produces a spliced antisense transcript encoding a protein that lacks a classic bZIP domain but still inhibits Tax2-mediated transcription. *Blood* 114(12):2427–2438.
27. Malim MH, Emerman M (2008) HIV-1 accessory proteins: Ensuring viral survival in a hostile environment. *Cell Host Microbe* 3(6):388–398.
28. Gil M, Zanetti MS, Zoller S, Anisimova M (2013) CodonPhyML: Fast maximum likelihood phylogeny estimation under codon substitution models. *Mol Biol Evol* 30(6):1270–1280.
29. Dalquen DA, Anisimova M, Gonnet GH, Dessimoz C (2012) ALF: A simulation framework for genome evolution. *Mol Biol Evol* 29(4):1115–1123.
30. Sabath N, Landan G, Graur D (2008) A method for the simultaneous estimation of selection intensities in overlapping genes. *PLoS One* 3(12):e3996.
31. Mir K, Schober S (2014) Selection pressure in alternative reading frames. *PLoS One* 9(10):e108768.
32. Firth AE (2014) Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Res* 42(20):12425–12439.
33. Li F, Ding SW (2006) Virus counterdefense: Diverse strategies for evading the RNA-silencing immunity. *Annu Rev Microbiol* 60:503–531.
34. Sabath N, Wagner A, Karlin D (2012) Evolution of viral proteins originated de novo by overprinting. *Mol Biol Evol* 29(12):3767–3780.
35. Faria NR, et al. (2014) HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346(6205):56–61.
36. Shi M, et al. (2012) Evolutionary conservation of the PA-X open reading frame in segment 3 of influenza A virus. *J Virol* 86(22):12411–12413.
37. Saayman S, et al. (2014) An HIV-encoded antisense long noncoding RNA epigenetically regulates viral transcription. *Mol Ther* 22(6):1164–1175.
38. Barbeau B, Mesnard JM (2015) Does chronic infection in retroviruses have a sense? *Trends Microbiol* 23(6):367–375.
39. Pagel M (1999) Inferring the historical patterns of biological evolution. *Nature* 401(6756):877–884.
40. Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. *Mol Biol Evol* 24(7):1464–1479.
41. Felsenstein J (1989) PHYLIP: Phylogeny inference package (version 3.2). *Cladistics* 5:163–166.
42. Bulmer M (1991) Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Mol Biol Evol* 8(6):868.
43. Price MN, Dehal PS, Arkin AP (2010) FastTree 2: Approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.
44. Hemelaar J, Gouws E, Ghys PD, Osmanov S (2011) Global trends in molecular epidemiology of HIV-1 during 2000–2007. *AIDS* 25(5):679–689.
45. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* 30(4):772–780.
46. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
47. Finn RD, Clements J, Eddy SR (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res* 39(W1):W29–W37.
48. Hughes AL, Westover K, da Silva J, O'Connor DH, Watkins DI (2001) Simultaneous positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. *J Virol* 75(17):7966–7972.
49. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24(8):1586–1591.
50. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res* 43(W1):W401-7.