Highly Conserved Structural Properties of the C-terminal Tail of HIV-1 gp41 Protein Despite Substantial Sequence Variation among Diverse Clades

IMPLICATIONS FOR FUNCTIONS IN VIRAL REPLICATION

Received for publication, May 6, 2011, and in revised form, June 3, 2011 Published, JBC Papers in Press, June 9, 2011, DOI 10.1074/jbc.M111.258855

Jonathan D. Steckbeck, Jodi K. Craigo, Christopher O. Barnes, and Ronald C. Montelaro

From the Center for Vaccine Research and the Department of Microbiology and Molecular Genetics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15261

Although the HIV-1 Env gp120 and gp41 ectodomain have been extensively characterized in terms of structure and function, similar characterizations of the C-terminal tail (CTT) of HIV gp41 remain relatively limited and contradictory. The current study was designed to examine in detail CTT sequence conservation relative to gp120 and the gp41 ectodomain and to examine the conservation of predicted physicochemical and structural properties across a number of divergent HIV clades and groups. Results demonstrate that CTT sequences display intermediate levels of sequence evolution and diversity in comparison to the more diverse gp120 and the more conserved gp41 ectodomain. Despite the relatively high level of CTT sequence variation, the physicochemical properties of the lentivirus lytic peptide domains (LLPs) within the CTT are evidently highly conserved across clades/groups. Additionally, predictions using PEP-FOLD indicate a high level of structural similarity in the LLP regions that was confirmed by circular dichroism measurements of secondary structure of LLP peptides from clades B, C, and group O. Results demonstrate that LLP peptides adopt helical structure in the presence of SDS or trifluoroethanol but are predominantly unstructured in aqueous buffer. Thus, these data for the first time demonstrate strong conservations of characteristic CTT physicochemical and structural properties despite substantial sequence diversity, apparently indicating a delicate balance between evolutionary pressures and the conservation of CTT structure and associated functional roles in virus replication.

The envelope (Env) proteins of retroviruses in general, and HIV-1 in particular, have been shown to serve as a major determinant of viral replication, antigenic, and pathogenic properties, such that natural and experimental variations in envelope sequence can markedly alter these viral phenotypes (1–3). Fundamentally, HIV-1 Env is the mediator of viral infection. Env is translated as a full-length gp160 polypeptide that is post-translationally cleaved into non-covalently associated gp120 and gp41 subunits that yield the functional Env monomer that then forms the virion-associated Env trimer complex. The gp120 and gp41 subunits of Env each mediate specific functions during the infection process. The function of gp120 is to bind the primary receptor, CD4, and subsequently the co-receptor, primarily CCR5 or CXCR4. Receptor binding leads to conformational changes in gp41 such that the previously sequestered fusion peptide at the N terminus of gp41 inserts into the cellular membrane. This is followed by refolding events that function to bring the cellular and viral membranes into proximity, finally leading to fusion and infection.

The gp41 protein is organized into three distinct domains: (i) the ectodomain, including the fusion peptide, the N- and C-terminal heptad repeat regions, and the membrane proximal external region (MPER), a target for broadly neutralizing monoclonal antibodies; (ii) the membrane-spanning domain (MSD), the sequence that anchors Env into the membrane; (iii) the C-terminal tail (CTT), demonstrated to function in cellular Env trafficking (4–6) and virion incorporation (7–9). The gp41 ectodomain has historically been the focus of many studies examining its role in the fusion process and as a target for neutralizing antibodies, whereas the MSD has been the subject of detailed studies to determine its precise length and structure (10–14). In contrast to these gp41 domains, the CTT has been relatively less studied, having only recently been demonstrated as playing a major role in Env trafficking and viral Env incorporation (4–9).

The CTT first came to prominence in the 1980s when it was reported that serum targeting a peptide from the CTT could bind gp160 (15) and neutralize virus infectivity (16, 17). This apparent exposure of the CTT to neutralizing antibodies implied a surface exposure not predicted from the intracytoplasmic location predicted for the CTT sequences. In the following years, however, the CTT was thought to be dispensable for virus replication based on studies indicating the replication competence of proviral constructs in which the CTT was deleted from the Env gene (18–20). More recently the CTT has been demonstrated to contain a number of trafficking signals that control the level of Env expression at the cellular mem-

© 2011 by The American Society for Biochemistry and Molecular Biology, Inc. Printed in the U.S.A.
brane. The CTT contains two functional endocytic motifs that lead to the rapid removal of Env from the surface of Env-expressing cells into late endosomes (4, 6). The CTT has also been shown to contain sequences involved in virion incorporation (5, 7–9), particularly through a proposed interaction with the matrix domain of HIV Gag (21, 22). Additionally, various studies have demonstrated that the CTT is a critical determinant of overall Env conformational properties, as variations in CTT sequences can alter Env fusogenicity and antigenicity mediated by the gp120 protein and the gp41 ectodomain (7). Finally, the CTT has been demonstrated to contain sequences required for association with detergent-resistant membranes, including a crucial role for the α-helical nature of the LLP regions in maintaining lipid raft association (23).

Although high resolution structures have been determined for sequences in the gp41 ectodomain, little is known about the structure of the CTT. There are no high resolution structures for any sequences of the CTT, although various studies have indicated that discrete domains in the CTT, known as the lentivirus lytic peptides (LLP), are helical in membrane mimetic environments (24–26). Topologically, gp41 is generally considered to be a type I membrane protein with an extracellular N terminus and a cytoplasmic C-terminal tail following the predicted membrane spanning domain. The model of the CTT as an exclusively intracytoplasmic domain is consistent with the location of functional endocytic motifs that affect Env trafficking. However, this generalization is challenged by conflicting data indicating that the Kennedy epitope is the target for neutralizing antibodies, suggesting an external localization of CTT sequences, perhaps most evident after receptor binding (16, 17, 27). Indeed, recent studies have also demonstrated differential exposure of CTT sequences on viral and cellular membranes (28), and biochemical studies have revealed transient exposure of CTT sequences during the Env fusion to cellular membranes (29).

Sequence analyses of Env have focused almost exclusively on the gp120 subunit and the gp41 ectodomain (30, 31) and the effects of sequence variation on viral phenotypes, e.g. tropism (32) and neutralization (33). To date, there have been no comprehensive analyses of the sequence variability of the CTT domain. In light of recent observations that even minor variations in CTT sequence can markedly alter Env structural and functional properties, the current study was designed to determine the extent of amino acid sequence variation in the CTT of diverse HIV-1 strains and to analyze the effects of this sequence variation on the physicochemical and structural properties of variant CTT species. The results of these studies for the first time provide evidence for evolutionary pressures on CTT sequences that are balanced by a remarkable conservation of uniquely characteristic physicochemical and structural properties, apparently to maintain critical functional roles in virus replication. These data indicate the need for further characterization of these CTT-mediated functions.

**EXPERIMENTAL PROCEDURES**

**Sequence Acquisition and Processing**—Sequences were obtained from the Los Alamos National Laboratory HIV sequence data base. Full-length (gp160) HIV-1 Env amino acid alignments were collected from years including 1997 through 2008 and were separated into clades A, B, C, D, F, and G and Groups N and O for eight total sequence groupings. The sequences in each clade/group were further processed as full-length Env sequences to remove duplicate sequences using the ElimDupes tool from Los Alamos National Laboratory (www.hiv.lanl.gov). The sequences were then separated into the three main domains of Env (gp120, the gp41 ectodomain, and the gp41 CTT) and aligned as described in the following section. The collection of sequences and subsequent processing yielded a total of ~2,000 unique sequences. Total sequences in each clade/group were as follows: 150 in clade A; 888 in clade B; 626 in clade C; 118 in clade D; 34 in clade F; 48 in clade G; 6 in group N; 126 in group O.

**Sequence Alignments and Evolutionary Analyses**—Sequences from each clade/group were individually aligned within the respective domain (gp120, gp41 ectodomain, and gp41 CTT) for a total of 24 alignments (8 clades/groups and 3 protein domains). Global alignments were performed using Geneious Pro 5.0.4 (Biomatters Ltd., NZ) using the Pam250 cost matrix. The gap opening/extension penalties were set to 16/8, with 30 refinement iterations performed for each alignment. Sequences were hand-edited where necessary, particularly around the variable regions in gp120. Evolutionary analyses included amino acid diversity computations and phylogenetic tree construction. The diversity of each domain (gp120, gp41 ectodomain, gp41 CTT) was measured by comparing the calculated distances of all sequences within a group in the alignment to one another. Genetic distances of aligned sequences were determined utilizing Kimura population distance calculations for amino acid sequences with no gap weighting, as executed in the program distmat from the EMBOSS package of software. Statistical significance values of the mean diversity of the domains were calculated in GraphPad Prism 5.0 (GraphPad Software, San Diego, CA). The statistical significance of diversity was calculated using two-tailed unpaired *t* tests, as implemented in GraphPad InStat Version 3.00. Phylogenetic analyses were performed on the individual domain alignments of the hand-edited, aligned sequences. The phylogenetic trees were constructed by the neighbor-joining method of Kimura-corrected measurements with the optimality criterion set to distance as measured in PAUP (35) and implemented in Geneious Pro Version 5.0.4. Statistical significance of branchings and clustering were assessed by bootstrap re-sampling of 1000 pseudoreplicates on the complete data set, and the confidence level was set to 75. The trees were edited for publication using FigTree Version 1.1.2 (36).

**Characterization of Physicochemical Properties of Consensus LLP Peptide Sequences**—MPER, MSD, and LLP sequences from all clades/groups were extracted from consensus sequences determined from alignments generated with Geneious Pro 5.0.4. The physicochemical characteristics (hydrophobic...
Structural Conservation of HIV-1 gp41 C-terminal Tail

moment, net charge, and hydrophobicity) and helical wheel diagrams of the extracted regions were determined using HeliQuest. One-way analysis of variance using Tukey’s multiple comparison test and Bartlett’s test for equal variances was performed using GraphPad Prism 5.0 (GraphPad Software, San Diego, CA).

PEP-FOLD Structure Prediction for LLP Peptides—LLP and Kennedy epitope (KE) sequences for all clades/groups were extracted from the consensus sequences generated by the global multiple sequence alignments. Sequences were submitted to the PEP-FOLD server (37–39) for peptide folding including PSIPRED predictions (40). The number of clusters and the percentage of total structures occupying each cluster were graphed to demonstrate conformational diversity for each peptide sequence. Conservation of secondary structure for the peptides was determined by finding the root mean square deviation (r.m.s.d.) for the peptides as follows. For KE, LLP1, and LLP3, the models were aligned as described (41, 42) using the r.m.s.d. align function of VMD (43). Due to the predicted random coil nature of the C-terminal half of the group O LLP2 model, the LLP2 models were aligned using the STAMP structural alignment method (44) instead of aligning to the minimum r.m.s.d. To take into consideration the differences in the predicted differences in LLP2 lengths, the LLP2 analysis was done using the first 14 residues of clades/groups A, C, G, N, and O and residues 3–16 of clades B, D, and F. Average r.m.s.d. was then determined for all residues on the STAMP-aligned models.

Circular Dichroism (CD) Spectroscopy to Determine Peptide Secondary Structure—LLP sequences for clades B and C and group O were extracted from the consensus sequences generated by the global multiple sequence alignments. In addition to the domains derived from the consensus sequences, we also determined the sequence with the lowest average divergence (LAD) from all sequences within each clade/group to select a representative natural isolate for secondary structure characterization. LLP sequences were also extracted from these LAD sequences. All 17 peptides ( LLP2 from clade B consensus and LAD were identical) were synthesized using standard Fmoc (N-(9-fluorenyl)methoxycarbonyl) chemistry with a carboxylated N terminus and amidated C terminus to mimic their nature chemistry within a longer polypeptide sequence. Peptides were resuspended in 10 mM HEPES, 10 mM SDS in 10 mM HEPES, or 60% TFE in 10 mM HEPES to a concentration of 50 /mM. CD spectra were determined at 37 °C on a Jasco 810 CD spectrometer from 195 to 250 nm. Spectra are presented as the average of 10 acquisitions. Peptide mean residue ellipticity was determined using CDPro (45, 46).

RESULTS

CTT Sequence Variation Is Intermediate between gp120 and the gp41 Ectodomain—Initial analyses of the variation present in the individual domains of the Env protein involved a comparative alignment of the amino acid sequences of the ~2000 isolates available in the Los Alamos data base (see “Experimental Procedures” for inclusion/exclusion determinations). Alignments of the respective gp120, gp41 ectodomain, and the CTT domains were performed within and between all sequences of all clades and groups (data not shown). For simplicity and because data from all clades/groups followed similar patterns, only clades B and C (the two most widely studied) and group O (the most variant) are presented here, with the remaining clades/group (clades A, D, F, and G, and group N) presented in supplemental Figs. S1–S4. Subsequently, phylogenetic analyses by construction of neighbor-joining trees were performed to demonstrate the fundamental ancestral associations between the clades/groups at these three domains (Fig. 1 and supplemental Fig. S1). All phylogenetic inferences were made by comparing the resultant ancestral relationships of the three domains at a similar scale. The results demonstrated the least genetic relatedness for the gp120 sequences within a clade/group, as evidenced by the greatest average distance between sequences. In contrast, the gp41 ectodomain sequences were most closely related to each other within a clade/group, as exhibited by the least average distance between sequences. Finally, the gp41 CTT sequences were more related to each other than gp120 sequences but less related to each other than the gp41 ectodomain sequences.

Diversity analyses were performed to provide a more quantitative distance measure between the sequences of each clade/group (Fig. 2 and supplemental Fig. S2). Similar to the results from the phylogenetic analyses, gp120 was observed to have the highest overall diversity within each clade/group, with the mean diversity ranging from 20.32 to 37.75%. The gp41 ectodomain sequences displayed the lowest mean diversity, ranging from 6.63 to 18.96%. Finally, CTT sequences exhibited intermediate diversity between gp120 and gp41 ectodomain, with a mean diversity range of 12.36–26.78%. Together these results demonstrate an intermediate level of amino acid sequence relatedness and diversity for the CTT in comparison to gp120 and gp41 ectodomain domains, suggesting a previously unrecognized evolutionary pressure on the CTT domain.
Physicochemical Characteristics of CTT Domains Are Conserved—In light of the relatively high level of diversity calculated for the CTT domain, we next sought to analyze the effects of CTT sequence variation on the physicochemical properties of defined segments of the CTT domain. To achieve a practical means of comparing the large number of CTT sequences in the data base, we chose to additionally perform an alignment of the consensus sequences of the CTT domain for all of the clades (Fig. 3). Consensus sequences of the entire CTT region of each clade were derived from hand-edited alignments and aligned to one another as described under “Experimental Procedures.” To more thoroughly understand the potential impact of amino acid variability, we investigated the conservation of the physicochemical characteristics of discrete, putative membrane-associated domains of the CTT compared with two other defined membrane-associated domains of gp41, the MPER and the MSD. We calculated the hydrophobic moment (µH), net charge, and hydrophobicity (H) of the peptide domains derived from the consensus sequences for the MPER, MSD, and the LLP regions of the CTT for all clades/groups (Fig. 4). The results of these analyses reveal that the LLP domains have significantly higher hydrophobic moments (mean µH = 0.6003, 0.3640, and 0.3945 for LLPs 1, 2, and 3, respectively) than the MPER and MSD (mean µH = 0.1520 and 0.1153, respectively) and that the calculated LLP1 hydrophobic moment is significantly higher than LLP2 and LLP3 (Fig. 4A). Both LLP1 and LLP2 (+5 and +3 mean charges, respectively) are significantly more cationic than the MPER, MSD, or LLP3 (−0.25, +1, and +0.25 mean charges, respectively) (Fig. 4B). The LLP regions were also found to be significantly less hydrophobic (0.3643, 0.4974, and 0.6650 mean H for LLPs 1, 2, and 3, respectively) than either the MPER (0.8268 mean H) or the MSD (1.133 mean H) (Fig. 4C). Comparison of the variances of the calculated hydrophobic moments and net charge between the clades/groups indicate in general a similar level of variation in the sequences across clades/groups. However, the variances between clades/groups did differ for hydrophobicity, with the MPER and MSD exhibiting lower variance than the LLP regions across all clades/groups, indicating that the hydrophobicity of the MPER and MSD peptide sequences is more conserved than for the LLP peptide sequences.

As the consensus LLP sequences were found to have high hydrophobic moments and, thus, high amphiphatic potential, we modeled the LLP peptide sequences using helical wheel diagrams (Fig. 5, supplemental Fig. S3). Inspection of the helical wheel diagrams confirms the amphipathic nature of the peptide sequences, with the charged and polar residues concentrated on one helical face and the hydrophobic and nonpolar residues concentrated on the opposite helical face. This amphipathic nature is highly conserved across all groups and clades, indicated by the similar incidence and position of the charged residues on the polar helical face.

Although overall charge for the LLP sequences was conserved (Fig. 4B), the consensus sequences suggest a preference for arginine relative to lysine in the cationic LLP1 and LLP2 sequences. To test this apparent preference for arginine in LLP sequences, we calculated the frequency of arginine and lysine in the CTT of HIV-1 Env using 1675 sequences from clades A, B, C, D, F, and G and groups N and O from the Los Alamos National Laboratory 2009 HIV-1 Env sequence alignment (www.hiv.lanl.gov) and compared the calculated frequencies to the arginine and lysine frequencies for all proteins represented in the UniProtKB/Swiss-Prot protein knowledgebase release 2010_09. Overall, arginine was found to occur at a frequency of 12.7% in the CTT (5.9% for all of Env) compared with a frequency of 5.5% observed for all proteins in the UniProtKB/Swiss-Prot knowledgebase. This increased frequency of arginine was accompanied by a marked decrease in the frequency of lysine (1.9% CTT versus 5.9% UniProtKB/Swiss-Prot; 5.0% for all of HIV Env), implying an important functional role for the arginine residue relative to lysine in the CTT. To ascertain position-specific arginine/lysine substitutions, we prepared logo diagrams for the LLP regions to determine the most
Conserved positively charged positions (Fig. 6). Positions where the probability of arginine/lysine was 0.9 across all clades/groups were considered to be highly conserved. For the most highly conserved arginine residues in LLP1 and LLP2, lysine was substituted for arginine 0.40% of the time (40 of 10,044 possible instances), whereas in LLP3 the conserved lysines were replaced by arginine in 4.84% of instances (162 of 3348). Collectively these results indicate that the physicochemical properties of the LLP regions are conserved despite high sequence variation and suggest a preference for arginine over lysine, particularly in LLP1 and LLP2.

Secondary Structure of the LLP Regions of the CTT Is Predicted to Be Highly Conserved—After determining that the physicochemical characteristics of the LLP regions were conserved, we examined the conservation of LLP secondary structure using PEP-FOLD computational peptide folding (37–39). Because peptide analogs of the LLP regions are proposed to be helical in conformation (24–26), we included as a control another reference CTT sequence, the KE, which is predicted to exist in a non-helical conformation (47). Each PEP-FOLD run included 50 simulations followed by a comparison of the returned models by cluster analysis to assess conformational diversity, with models grouped together in a cluster (39). Fig. 7 and supplemental Fig. S4 show the number of clusters returned by PEP-FOLD for each CTT domain and the relative percentage of the total population comprised by each cluster. The conformational diversity of the KE was high for all the clades/groups, with between 13 and 33 clusters identified, consistent with a random coil conformation. The LLP regions, on the other hand, were more conformationally homogeneous, with 21 of the 24 LLP regions tested returning only one cluster. Models from the three LLP domains that returned more than one cluster are shown in supplemental Fig. S5. Clade A LLP2 (supplemental Fig. S5A) and clade C LLP1 (supplemental Fig. S5B) both returned two clusters of high helical content, with a kinked second, higher energy structure. Group O LLP2 returned seven clusters. However, the 16 N-terminal amino acids (LLSNLASHGIQKVISHL) were consistently helical, whereas the remaining nine C-terminal amino acids (GLGLWILGQ) adopted a random coil conformation (supplemental Fig. S5C). Finally, to assess overall conservation of the predicted secondary structure of the LLP domain, we compared the lowest energy PEP-FOLD models from the LLP regions and the Kennedy epitope are boxed.

**FIGURE 3. Comparison of the amino acid sequence of the CTT domain.** Consensus sequences for the CTT region of the Env of each clade or group were derived from hand-edited alignments of all the isolates downloaded from the Los Alamos data base (see “Experimental Procedures”). Derived CTT region sequences were aligned and compared with the clade B consensus sequence for visualization purposes. Only the amino acid residues different from clade B are reported on the alignment. Dots indicate residues identical to the clade B consensus sequence; dashes indicate alignment gaps. Previously described LLP regions and the Kennedy epitope are boxed.

**FIGURE 4. LLP sequences have different physicochemical properties than MPER or MSD sequences.** Physicochemical properties of lipophilic domains from gp41 were determined from consensus sequences from each clade/group. Data presented as the mean and S.D. of clades/groups, and differences between columns are significant at p < 0.001 unless otherwise noted. A, hydrophobic moments were calculated to determine amphipathic potential, and LLPs were found to be significantly more amphipathic than MPER or MSD sequences. B, average net charges demonstrate that LLPs 1 and 2 are highly cationic. C, MPER and MSD are significantly more hydrophobic than LLP regions. n.s., not significant; *, p < 0.05.
FOLD models for each consensus domain sequence from all of the clades/groups (Fig. 8) and determined the average cr.m.s.d. (Table 1). In general, for the LLP peptides the average cr.m.s.d. was <1, indicating a high level of structural conservation. In contrast to the other consensus LLP sequences, a cr.m.s.d. value of \(-6\) was calculated for the group O LLP2 consensus sequence due to the predicted random coil nature of the C-terminal nine amino acids. As expected from the predicted random coil conformation of the KE consensus sequence, the average cr.m.s.d. calculated for the KE ranged from \(-2.9\) to \(6.5\). These predictive experiments and analyses suggest that the \(-\) helical structure of the LLP regions is conserved across all the clades/groups for the consensus sequences generated from multiple sequence alignments.

**LLP Peptide Analogs Adopt Helical Structure in Membrane-mimetic and Hydrophobic Environments**—As the PEP-FOLD experiments predicted that consensus sequences of the LLP regions adopt stable helical conformations, we endeavored to explicitly measure the secondary structure of LLP sequence peptide analogs. Toward this objective, peptides were synthesized corresponding to the LLP regions from the consensus sequences from clades B and C and group O and from the LAD sequences determined from the CTT alignments for the same clades/groups (Table 2). These clades/group were chosen as they represent the most widely studied HIV-1 sequences (clades B and C) and the most structurally divergent as predicted by the PEP-FOLD analyses (group O, particularly for LLP2). CD spectroscopy measurements of the LLP peptides were performed in three solutions: 10 mM HEPES, 10 mM SDS in 10 mM HEPES, and 60% TFE in 10 mM HEPES. For peptides in aqueous solution, spectra were consistent with predominantly random coil conformation with the exception of LLP1 from the clade C LAD sequence and LLP2 from clade B (Figs. 9A and 10). The addition of 10 mM SDS to 10 mM HEPES
resulted in an increase (or a maintenance, in the case of clade B LLP2 and clade C LAD LLP1) in helical content relative to 10 mM HEPES (Figs. 9B and 10). Likewise, the addition of 60% TFE to 10 mM HEPES maintained or moderately increased helical content compared with that observed with 10 mM SDS (Figs. 9C and 10). Results from CD spectroscopy measurements of LLP peptide analogs confirm predictions obtained by PEP-FOLD analysis that the α-helical nature of the LLP peptides in membrane mimetic environments (i.e. SDS micelles and/or TFE) is broadly conserved among clades/groups despite extensive amino acid variation.

**DISCUSSION**

Contemporary research has placed the Env structure and function as a high priority research focus for HIV-1. Although the majority of published studies focus on the gp120 protein or gp41 ectodomain, recent studies have implicated the CTT in the modulation of overall Env structure and function (7, 48, 49). These recent observations increase the need for a more detailed characterization of CTT sequences among diverse isolates of HIV-1 and the effects of sequence variation on CTT structure. Thus, the current study was designed to provide a comprehensive comparative analysis of CTT species among different groups and clades of HIV-1.

In the present study we examined the conservation of the CTT of HIV-1 gp41 in clades A, B, C, D, F, and G and groups N and O as compared both to gp120 and to the gp41 ectodomain. In addition, we characterized the conservation of physicochemical and structural properties among variant CTT sequences. Our results demonstrate that the CTT is intermediate conserv ed relative to the highly variable gp120 and the more conserved gp41 ectodomain. Despite the relatively high level of amino acid diversity present in the CTT, the physicochemical properties of LLP domains are remarkably conserved across the clades/groups examined. Secondary structure predictions using PEP-FOLD indicate a high level of structure conservation in the LLP regions, results that are confirmed by CD spectroscopy of peptide analogs of the LLP regions from consensus and isolate sequences from clades B and C and group O.

The finding that the CTT exhibits intermediate sequence variation between the highly variable gp120 and the least variable gp41 ectodomain provokes thought as to the source of the evolutionary pressures and significance of the substantial sequence evolution on putative CTT functions. The relatively high level of sequence variation observed in HIV-1 gp120 is due to a combination of random genomic alterations associated with reverse transcription and selective pressure by host immune responses leading to evolution of Env gp120 populations during persistent infection. Although the gp41 gene is also similarly subject to mutations during reverse transcription, the relatively low variation in the gp41 ectodomain is likely due to more strict requirements for structure and function in light of large conformational rearrangements that occur during the fusion process. Intermediate variation observed in the CTT suggests two possibilities; (i) CTT sequences are exposed to the immune system, and therefore, variation is observed as a result
of immune selection or (ii) the CTT is not an important functional domain in Env and as such can vary randomly without detrimental effect. In our opinion, the observed conservation of LLP structure and the selective incorporation of arginine residues indicate a functional importance that is incompatible with random variation. Additionally, when we examined whether the presence of accessory protein coding sequences (Tat and Rev) affected CTT sequence variation, we found no significant differences in sequence variability between regions utilizing one (CTT-only), two (CTT and Rev), or all three (CTT, Rev, and Tat) open reading frames (data not shown). As LLP2 and LLP3 occur in the region including the Rev coding sequence and LLP1 occurs in the CTT-only region, the conservation of the LLP regions is not likely due to extraneous influence (e.g., accessory proteins) and is further indicative of a functional relevance for the sequences. This perspective is further supported by the observations that minor point mutations in LLP sequences can result in major alterations in overall Env structural and functional properties (7).

In stark contrast to the overall observed sequence variability in the CTT is the conservation, and indeed prevalence, of arginine in the CTT, especially relative to the other basic amino acid, lysine (cf. Fig. 5). Although the two highly basic amino acids lysine and arginine are generally considered to be interchangeable and able to readily substitute for one another, the data presented here directly controvert that assumption for the CTT. As indicated by the current analyses of CTT basic residues, the CTT arginine residues are highly conserved and are only very rarely replaced by lysine. In fact, arginine replacement by lysine in the CTT occurs at a much lower incidence as compared with its general prevalence in proteins. This selective preference for arginine in the CTT is likely due to functional requirements for the unique properties of the arginine side chain that terminates in a guanidinium group. Arginine demonstrates properties that make it particularly well suited to occur in membrane-interactive regions of proteins, as exemplified by the highly conserved arginine in the MSD of HIV-1 gp41. Primary is the potential for the arginine side chain to "snorkel" from a membrane-embedded main chain position into the extramembranous space, providing an energetically favorable interaction with aqueous substrates (50). Additional recent controversy surrounds the potential for arginine to reside in or near the membrane center, although the mechanism by which this may occur is unclear (51–54). It is thought that the guanidinium group of arginine may be hydrated by membrane-resident waters (55). More clearly demonstrated and well established is the cation–π interaction of the arginine side chain with proximal phenylalanine, tryptophan, and tyrosine side chains (56, 57). In this interaction the dispersed positive charge of the guanidinium group interacts with the quadrupole moment of planar hydrocarbon rings to form an energetically stable “bond” that can subsequently pass into the membrane hydrocarbon core (56, 58). Any and all of these arginine-specific properties may play a role in the well documented ability of arginine-rich sequences to pass through biological membranes (59–61), and it is this metaproperty that likely is of the greatest significance to the prevalence of arginine in the CTT.

The sequence variation that is characteristic of the CTT appears to indicate an exposure to immune selection perhaps
indicative of surface accessibility of the target sequences. Recent findings have demonstrated transient exposure of LLP2 sequences on the surface of Env-expressing cells (29). In particular, Chen and co-workers (23) determined that LLP2 is exposed to antibody binding during cell-cell fusion under conditions that slow the fusion process (i.e. at 31 °C) but not under physiological conditions (37 °C), suggesting highly dynamic LLP2 exposure. This is consistent with published data demonstrating lipid association of LLP2 both pre- and post-fusion (62). The arginine-rich nature of LLP2 may contribute to the
The C-terminal tail (CTT) of the HIV-1 envelope protein gp41 is a highly conserved region that is involved in the fusogenic activity of the virus. Recent studies have highlighted the ability of the CTT to modulate HIV-1 Env structure and functional properties, but the mechanisms for this control remain to be defined. In addition, the current CTT sequence and structure studies indicate highly conserved physicochemical and structural properties that may mediate critical functions in the life cycle of HIV-1. In light of the genetic economy intrinsic to viruses, there appears to be increasing data for a reevaluation of CTT functional relevance in HIV replication and Env structural and antigenic properties.

**REFERENCES**

1. Levy, J. A. (2009) *AIDS* 23, 147–160
2. Evans, D. T., and Desrosiers, R. C. (2001) *Immunol. Rev.* 183, 141–158
3. Roux, K. H., and Taylor, K. A. (2007) *Curr. Opin. Struct. Biol.* 17, 244–252
4. Byland, R., Vance, P. J., Hoxie, J. A., and Marsh, M. (2007) *Mol. Biol. Cell* 18, 414–425
5. Lopez-Vergès, S., Camus, G., Blot, G., Beauvoir, R., Benarous, R., and Berlioz-Torrent, C. (2006) *Proc. Natl. Acad. Sci. U.S.A.* 103, 14947–14952
6. Ohno, H., Aguilar, R. C., Fournier, M. C., Hennecke, S., Cooson, P., and Bonifácio, J. S. (1997) *Virology* 238, 305–315
7. Karla, V., Sarkar, S., Gupta, P., and Montelaro, R. C. (2005) *J. Virol.* 79, 2097–2107
8. Wyma, D. J., Kotov, A., and Aiken, C. (2000) *J. Virol.* 74, 9381–9387
9. Murakami, T., and Freed, E. O. (2000) *Proc. Natl. Acad. Sci. U.S.A.* 97, 343–348
10. Gangopomu, V. K., and Abrams, C. F. (2010) *Biophys J.* 99, 3438–3444
11. West, J. T., Johnston, P. B., Dubay, S. R., and Hunter, E. (2001) *J. Virol.* 75, 9601–9612
12. Yue, L., Shang, L., and Hunter, E. (2009) *J. Virol.* 83, 11588–11598
13. Shang, L., Yue, L., and Hunter, E. (2008) *J. Virol.* 82, 5417–5428
14. Kim, J. H., Hartley, T. L., Curran, A. R., and Engelman, D. M. (2009) *Biochim. Biophys. Acta* 1788, 1804–1812
15. Kennedy, R. C., Henkel, R. D., Pauletti, D., Allan, J. S., Lee, T. H., Essex, M., and Dreesmann, G. R. (1986) *Science* 231, 141–158
16. Kennedy, R. C., Dreesmann, G. R., Kanda, P., Linette, G. P., Sparrow, J. T., Ho, D. D., and Kennedy, R. C. (1986) *EMBO J.* 5, 3065–3071
17. Ho, D. D., Sarnagadharan, M. G., Hirsch, M. S., Schooley, R. T., Rota, T. R., Kennedy, R. C., Chanh, T. C., and Sato, V. L. (1987) *J. Virol.* 61, 2024–2028
18. Hirsch, V. M., Edmondson, P., Murphye-Corbi, M., Arbeille, B., Johnson, P. R., and Mullins, J. I. (1989) *Nature* 341, 573–574
19. Kodama, T., Wookey, D. P., Naidu, Y. M., Kestler, H. W., 3rd, Daniel, M. D., Li, Y., and Desrosiers, R. C. (1989) *J. Virol.* 63, 4709–4714
Structural Conservation of HIV-1 gp41 C-terminal Tail

20. Chakrabarti, L., Emerman, M., Tiollais, P., and Sonigo, P. (1989) J. Virol. 63, 4395–4403
21. Murakami, T., and Freed, E. O. (2000) J. Virol. 74, 3548–3554
22. Murakami, T. (2008) Microbiol. Immunol. 52, 287–295
23. Yang, P., Ai, L. S., Huang, S. C., Li, H. F., Chan, W. E., Chang, C. W., Ko, C. Y., and Chen, S. S. (2010) J. Virol. 84, 59–75
24. Kliger, Y., and Shai, Y. (1997) J. Biol. Chem. 272, 7121–7127
25. Srinivas, S. K., Srinivas, R. V., Anantharamaiah, G. M., Segrest, J. P., and Merber, J. (1994) J. Virol. 68, 7115–7123
26. Chernomordik, L., Chanturiya, A. N., Suss-Toby, E., Nora, E., and Zimmerberg, J. (1998) in Virology 315, 283–296
27. Heap, C. J., Reading, S. A., and Dimmock, N. J. (2005) J. Virol. 86, 1499–1507
28. Steckbeck, J. D., Sun, C., Sturgeon, T. J., and Montelaro, R. C. (2010) PLoS ONE 5, e15261
29. Lu, L., Zhu, Y., Huang, J., Chen, X., Yang, H., Jiang, S., and Chen, Y. H. (2008) J. Biol. Chem. 283, 16723–16731
30. Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours, V. (2001) Br. Med. Bull. 58, 19–42
31. Pieniazek, D., Yang, C., and Lal, R. (1998) in Human Retroviruses and AIDS (Korber, B., Foley, B., McCutchan, F., Mellors, J., Hahn, B., Sodroski, J., and Kuiken, C., eds.) pp. 112–117, Los Alamos, NM
32. Scarlatti, G., Tresoldi, E., Björndal, A., Fredriksson, R., Colognesi, C., Deng, H. K., Malnati, M. S., Plebani, A., Siccardi, A. G., Littman, D. R., Fenyo, E. M., and Lusso, P. (1997) Nat. Med. 3, 1259–1265
33. Karlsson Hedestam, G. B., Fouchier, R. A., Phogat, S., Burton, D. R., Sodroski, J., and Wyatt, R. T. (2008) Nat. Rev. Microbiol. 6, 143–155
34. White, S. H., and Wimley, W. C. (1999) Annu. Rev. Biophys. Biomol. Struct. 28, 319–365
35. Swofford, D. L. (2002) PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods, Version 4, Sinauer Associates, Sunderland, MA
36. Morariu, V., Srinivasan, B., Raykar, V., Duraiswami, R., and Davis, L. (2008) Advances in Neural Information Processing Systems, pp. 1113–1120, MIT Press, Boston, MA
37. Maupetit, J., Tuffery, P., and Derreumaux, P. (2007) Proteins 69, 394–408
38. Maupetit, J., Derreumaux, P., and Tuffery, P. (2010) J. Comput. Chem. 31, 726–738
39. Maupetit, J., Derreumaux, P., and Tuffery, P. (2009) Nucleic Acids Res. 37, W498–W503
40. Jones, D. T. (1999) J. Mol. Biol. 292, 195–202
41. Kabsch, W. (1978) Acta Crystallogr. A 34, 827–828
42. Kabsch, W. (1976) Acta Crystallogr. A 32, 922–923
43. Humphrey, W., Dalke, A., and Schulten, K. (1996) J. Mol. Graph. 14, 33–38
44. Russell, R. B., and Barton, G. J. (1992) Proteins 14, 309–323
45. Sreerama, N., and Woody, R. W. (2000) Annu. Rev. Biochem. 69, 251–260
46. Sreerama, N., and Woody, R. W. (2004) Protein Sci. 13, 100–112
47. Hollier, M. J., and Dimmock, N. J. (2005) Virology 337, 284–296
48. Kalia, V., Sarkar, S., Gupta, P., and Montelaro, R. C. (2003) J. Virol. 77, 3634–3646
49. Newman, J. T., Sturgeon, T. J., Gupta, P., and Montelaro, R. C. (2007) Virology 367, 102–116
50. Killian, J. A., and von Heijne, G. (2000) Trends Biochem. Sci. 25, 429–434
51. Dorairaj, S., and Allen, T. W. (2007) Proc. Natl. Acad. Sci. U.S.A. 104, 4943–4948
52. Krishnakumar, S. S., and London, E. (2007) J. Mol. Biol. 374, 1251–1269
53. MacCallum, J. L., Bennett, W. F., and Tieleman, D. P. (2008) Biophys. J. 94, 3393–3404
54. MacCallum, J. L., Bennett, W. F., and Tieleman, D. P. (2007) J. Gen. Physiol. 129, 371–377
55. Kreplik, D., Mihaiescu, M., Freites, J. A., Schow, E. V., Worcester, D. L., Gawrisch, K., Tobias, D. J., White, S. H., and Swartz, K. J. (2009) Nature 462, 473–479
56. Dougherty, D. A. (1996) Science 271, 163–168
57. Ma, J. C., and Dougherty, D. A. (1997) Chem. Rev. 97, 1303–1324
58. Jing, W., Demcoe, A. R., and Vogel, H. J. (2003) J. Bacteriol. 185, 4938–4947
59. Futaki, S. (2005) Adv. Drug Deliv. Rev. 57, 547–558
60. Mitchell, D. J., Kim, D. T., Steinman, L., Fathman, C. G., and Rothbard, J. B. (2000) J. Pept. Res. 56, 318–325
61. Tung, C. H., and Weissleder, R. (2003) Adv. Drug Deliv. Rev. 55, 281–294
62. Viard, M., Ablan, S. D., Zhou, M., Veenstra, T. D., Freed, E. O., Raviv, Y., and Blumenthal, R. (2008) Biochemistry 47, 1977–1983
63. Inomata, K., Ohno, A., Tochio, H., Isogai, S., Tenno, T., Nakase, I., Takeuchi, T., Futaki, S., Ito, Y., Hiroaki, H., and Shirakawa, M. (2009) Nature 458, 106–109
64. Reading, S. A., Heap, C. J., and Dimmock, N. J. (2003) Virology 315, 362–372
65. Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. (1984) J. Mol. Biol. 179, 125–142