viruses

MDPI

Article

Covariation of Amino Acid Substitutions in the HIV-1 Envelope Glycoprotein gp120 and the Antisense Protein ASP Associated with Coreceptor Usage

Angelo Pavesi [1] and Fabio Romerio [2,*]

[1] Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parco Area delle Scienze 23A, I-43124 Parma, Italy; angelo.pavesi@unipr.it

[2] Department of Molecular and Comparative Pathobiology, Johns Hopkins University School of Medicine, 733 North Broadway, Baltimore, MD 21205, USA

* Correspondence: fromeri2@jhmi.edu

Abstract: The tropism of the Human Immunodeficiency Virus type 1 (HIV-1) is determined by the use of either or both chemokine coreceptors CCR5 (R5) and CXCR4 (X4) for entry into the target cell. The ability of HIV-1 to bind R5 or X4 is determined primarily by the third variable loop (V3) of the viral envelope glycoprotein gp120. HIV-1 strains of pandemic group M contain an antisense gene termed *asp*, which overlaps *env* outside the region encoding the V3 loop. We previously showed that the ASP protein localizes on the envelope of infectious HIV-1 virions, suggesting that it may play a role in viral entry. In this study, we first developed a statistical method to predict coreceptor tropism based on Fisher's linear discriminant analysis. We obtained three linear discriminant functions able to predict coreceptor tropism with high accuracy (94.4%) when applied to a training dataset of V3 sequences of known tropism. Using these functions, we predicted the tropism in a dataset of HIV-1 strains containing a full-length *asp* gene. In the amino acid sequence of ASP proteins expressed from these *asp* genes, we identified five positions with substitutions significantly associated with viral tropism. Interestingly, we found that these substitutions correlate significantly with substitutions at six amino acid positions of the V3 loop domain associated with tropism. Altogether, our computational analyses identify ASP amino acid signatures coevolving with V3 and potentially affecting HIV-1 tropism, which can be validated through in vitro and in vivo experiments.

Keywords: HIV-1; antisense gene asp; cell tropism; coevolution; coreceptor usage; envelope protein; Fisher's linear discriminant analysis; prediction algorithm

check for updates

Academic Editor: Carsten Münk

Received: 14 January 2025
Revised: 20 February 2025
Accepted: 24 February 2025
Published: 26 February 2025

Citation: Pavesi, A.; Romerio, F. Covariation of Amino Acid Substitutions in the HIV-1 Envelope Glycoprotein gp120 and the Antisense Protein ASP Associated with Coreceptor Usage. *Viruses* **2025**, *17*, 323. https://doi.org/10.3390/v17030323

1. Introduction

The entry of the human immunodeficiency virus 1 (HIV-1) into human host cells is initiated by interaction of the viral envelope glycoprotein gp120 with the cellular CD4 receptor [1]. This primary interaction induces a conformational change in the gp120 ectodomain of the protein [2] that enables viral binding to one of the cell surface co-receptors, CCR5 or CXCR4 [3,4]. This sequence of events results in membrane fusion and penetration of the virus into the host cell [5]. The third hypervariable loop domain (V3) of gp120, a sequence of ~35 amino acids, is the major determinant of coreceptor tropism [6]. Viral strains are classified as R5-tropic when using the CCR5 coreceptor, X4-tropic when using CXCR4, and R5X4-tropic or dual tropic when using both coreceptors. While R5-tropic viruses are present at all stages of infection, progression towards AIDS disease is often associated with the emergence of X4-tropic strains [7].

Two types of methods have been developed for testing viral tropism: (i) cell-based in vitro phenotypic tests such as ES-Trofile [8]; and (ii) in silico genotypic tests, based on sequencing the V3 loop domain of gp120 and using bioinformatics methods to predict coreceptor usage. Although phenotypic tests have a high sensitivity in distinguishing R5-, R5X4-, and X4-tropic viral strains, they are expensive and time consuming. As an alternative, genotypic tests are easily accessible, fast, and inexpensive.

The first genotypic method to predict coreceptor usage is the 11/25 charge rule, which classifies the virus as X4-tropic if a positively charged amino acid is observed in positions 11 or 25 of the V3 sequence [9]. As reported by Lengauer et al. [10], this simple rule shows good specificity (93% of the non-X4-tropic strains predicted as non-X4-tropic) but low sensitivity (only 60% of the X4-tropic strains predicted as X4-tropic), making it unsuitable for routine clinical use. Other methods use more complex statistical models, such as neural networks [11], support vector machines [12,13], position-specific scoring matrices [14], incorporation of physicochemical and structural properties of the V3 sequence into a numerical descriptor [15], coreceptor-specific weight matrices [16], and machine learning/hidden Markov model [17]. The method Geno2pheno [10] combines two machine learning approaches, support vector machine and decision tree, and uses clinical information such as viral loads and CD4 cell counts if available. Currently, Geno2pheno is the only genotypic method recommended for use in clinical routines by the European Consensus Group [18].

Since establishing HIV-1 tropism is crucial to address antiviral treatment, significant efforts have been made to improve the performance of genotypic tests in terms of sensitivity and specificity. Overall, the performance is high when predicting R5-tropic sequences but relatively low for X4-tropic sequences, likely because most of the V3 sequences from viral samples are in the R5-tropic group.

With the exception of the multiple linear regression [19], multivariate statistical methods have been overlooked as a method for the prediction of viral tropism. In the present study, we developed a genotypic test based on Fisher's linear discriminant analysis [20,21] to predict coreceptor tropism in a dataset of HIV-1 strains that contain the antisense gene *asp*, which overlaps the gp120 gene on the frame -2 and encodes the antisense protein ASP [22–25]. Using this test, we identified amino acid changes in ASP that are preferentially associated with R5 or X4 usage. Interestingly, we found that these substitutions in ASP correlate significantly with substitutions within the principal determinants of coreceptor tropism, the variable loops V3, and, to a lesser extent, V1/V2 of gp120. Altogether, our study identifies ASP residues coevolving with V3 and potentially affecting HIV-1 tropism, which can be validated through in vitro and in vivo experiments.

2. Materials and Methods

2.1. Assembly of a Training Dataset of V3 Sequences with Known Tropism

We downloaded the dataset of V3 sequences developed by Shen et al. [16]. It contains 2679 V3 sequences from HIV-1 strains with a tropism determined by a variety of phenotypic tests, as reported in the Los Alamos HIV sequence database (www.hiv.lanl.gov/content/index, accessed on 10 December 2014). By exploring Shen's dataset to select V3 sequences from strains of genotype B or C, we assembled a training dataset of 1701 R5-tropic sequences (1229 of genotype B and 472 of genotype C) and 137 purely X4-tropic sequences (96 of genotype B and 41 of genotype C). Purely X4-tropic means that we did not include in the training dataset V3 sequences from viral strains that use both coreceptors.

2.2. Linear Discriminant Analysis (LDA) of the Training Dataset

To convert each V3 sequence in the training dataset into a numerical format, we extracted from the AAindex database [26] a subset of 46 indices that quantitatively evaluate the physicochemical properties of amino acids. Within the subset of 149 hydrophobicity indices, we selected the most widely used hydropathy index by Kyte and Doolittle [27]. The numerical descriptor of each V3 sequence was a vector of 7 components; the first is the arithmetic mean of the hydropathy index, and the others are the arithmetic means of 6 indices of physicochemical properties randomly selected from the 46 available. Thus, the input data for LDA were a matrix of V3 R5-tropic sequences (1701 rows and 7 columns) and a matrix of V3 sequences X4-tropic (137 rows and 7 columns). LDA yielded a linear function with 7 coefficients and the corresponding degree of accuracy. Accuracy was the number of R5-tropic sequences predicted correctly plus the number of X4-tropic sequences predicted correctly, divided by the total number of sequences, and multiplied by 100. To find the combination(s) of 7 amino acid indices with linear function(s) having the highest accuracy, we carried out 100,000 LDAs and selected the cases with an accuracy of ∼94.0%.

To evaluate the robustness of LDAs showing the best accuracy, we carried out two cross-validation tests. The first depends on the fact that the training dataset contains ten times as many R5-tropic sequences as X4-tropic sequences (1701 vs. 137). For each combination of amino acid indices, the validation test consisted of (i) 1000 LDAs, each of them between the 137 X4-tropic sequences and a subset of R5-tropic sequences of equal size randomly taken from the 1701 ones; (ii) calculation of the mean accuracy in the prediction of R5 or X4 usage.

For each combination of amino acid indices, the other cross-validation test consisted of (i) creating a training subset composed of 65 R5-tropic sequences and 65 X4-tropic sequences, randomly selected from the 1701 and 137 sequences, respectively; (ii) calculating the linear discriminant function; and (iii) evaluating its accuracy on a validation subset composed of 65 R5-tropic sequences and 65 X4-tropic sequences, randomly taken from the 1701 and 137 ones, respectively, and different from those of the training subset. We repeated this procedure 1000 times and calculated the mean accuracy in the prediction of R5 or X4 usage.

2.3. Prediction of the Tropism in a Dataset of Viral Strains of Genotype B or C Containing an Antisense Gene asp

Selection of LDAs with an accuracy ∼94.0%, followed by evaluation of their robustness by cross-validation testing, yielded 3 combinations of 7 amino acid indices in which LDA showed the highest prediction accuracy (Section 3.1). Using the corresponding 3 linear functions, we predicted the tropism in a dataset of 2593 *env* sequences (1625 of genotype B and 968 of genotype C) taken from our previous study on the antisense gene *asp* [22]. These *env* sequences, aligned for a total of 1274 codon positions, contain an antisense gene termed *asp* [24], which overlaps the *env* gene at the surface and transmembrane boundary and outside the *env* region encoding the V3 loop domain [25,26]. We extracted the V3 region (nucleotide positions 7110 to 7217 in the HIV-1$_{\text{HXB2}}$ reference sequence, acc. number K03455) from each aligned *env* sequence and, after translation into amino acids, predicted its tropism. We classified a given V3 sequence as R5-tropic (or X4-tropic) only if predicted as R5-tropic (or X4-tropic) by all 3 linear discriminant functions. By this rule, we assigned the tropism to 2496 out of 2593 V3 sequences (Section 3.2).

2.4. Comparative Analysis of the Amino Acid Composition in the V3 Region Between R5-Tropic and X4-Tropic Viral Strains

Once predicted the tropism, we compared the amino acid content in the V3 region of strains R5-tropic with that of strains X4-tropic, with the aim to detect amino acid positions significantly associated with coreceptor usage. We analyzed the dataset of 2496 V3 regions, aligned for a total of 53 amino acid positions, and selected 35 positions in which the frequency of gaps was <5% for statistical analysis. At each position, we compared the amino acid content in R5-tropic sequences with that in X4-tropic sequences using the χ^2 contingency-table test. We performed the same analysis on the V1/V2 region (nucleotide positions 6579–6812 of HIV-1_{HXB2}), on the *env* region upstream V1/V2 and not overlapping the *vpu* gene (nucleotide positions 6312–6578 of HIV-1_{HXB2}), on the *env* region between V1/V2 and V3 (nucleotide positions 6813–7109 of HIV-1_{HXB2}), and on the *env* region downstream V3 and not overlapping the antisense *asp* gene (nucleotide positions 7218–7370 of HIV-1_{HXB2}).

2.5. Comparative Analysis of the Amino Acid Composition in the env/asp Overlap Between R5-Tropic and X4-Tropic Viral Strains

The region of *env* overlapping *asp* is an antisense −2 coding region, in which a codon of *asp* overlaps with two contiguous codons of *env* on the opposite strand. In detail, the third base of *asp* overlaps with the third base at codon *n* of *env*, while the second and the first base of *asp* overlap, respectively, with the first and the second base at codon (n + 1) of *env*. We extracted the region that overlaps with *asp* (nucleotide positions 7371–7943 of HIV-1_{HXB2}) from each of the 2496 aligned *env* sequences and selected 170 codon positions in which the frequency of gaps was <5% for statistical analysis. At each position, we compared the amino acid content in R5-tropic sequences with that in sequences X4-tropic using the χ^2 contingency-table test. We performed the same analysis on each codon position of *asp* that overlaps with two contiguous codons of *env* on the opposite strand.

2.6. Identification of Significant Patterns of Pairwise Associations Between Amino Acid Substitutions in ASP and Amino Acid Substitutions in V3 Loop Region

We detected, both in ASP and V3, a number of positions in which the amino content in R5-tropic strains differs significantly from that in X4-tropic strains (Sections 3.2 and 3.3). To test if there was a significant association between V3 and ASP substitutions, we calculated for each pair of amino acid changes the φ binomial correlation coefficient. It is similar to Pearson's correlation coefficient but is specifically used for categorical data arranged in a 2 × 2 contingency table. A φ value between 0.3 and 0.7 indicates a weak-medium pairwise association, while a φ value between 0.7 and 1.0 indicates a strong pairwise association.

3. Results

3.1. Linear Discriminant Analysis (LDA) Accurately Predicts HIV-1 Coreceptor Usage

We carried out LDA on a training dataset of 1701 R5-tropic V3 sequences (1229 of genotype B and 472 of genotype C) and 137 X4-tropic V3 sequences (96 of genotype B and 41 of genotype C). Sequence data are reported in Supplementary File S1. The mean length of V3 sequences was 34.8 amino acids, with a very low standard deviation (0.6). We converted each V3 sequence into numerical form using the Kyte–Doolittle hydropathy index [27] and 6 indices of physicochemical properties of amino acids, taken at random from the subset of 46 indices in the AAindex database [26].

We randomly generated 100,000 combinations of seven amino acid indices and found that, in most of them (86,892), LDA predicts the tropism with an accuracy of between 87.3 to 94.7%. We also found, however, that the accuracy is biased toward a more accurate prediction for R5 than X4 usage. Indeed, we found a very small number of combinations (11 cases) in which LDA predicts the tropism with an accuracy ∼94% both for R5 and X4 usage. We evaluated the robustness of these LDAs by two cross-validation tests (Section 2.2) and found the best performance for three combinations of amino acid indices. Indeed, sensitivity, specificity, and accuracy of LDA#1, LDA#2, and LDA#3 are all ∼94% (Table 1).

The value of MCC (Matthew's correlation coefficient) [28] is moderate (0.71), where value '1' corresponds to the perfect prediction and '0' to a completely random prediction. The moderate value of MCC is due to the robustness of this parameter when dealing with different class sizes, just like in this case where the number of R5-tropic sequences is one order of magnitude greater than that of sequences X4-tropic.

Table 1. Performance of LDA on the training datasets of V3 sequences from 1701 R5-tropic and 137 X4-tropic HIV-1 isolates.

LDA	Sensitivity [1]	Specificity [2]	Accuracy [3]	MCC [4]	Mean Sensitivity by First Cross-Validation Test	Mean Specificity by First Cross-Validation Test	Mean Accuracy by First Cross-Validation Test	Mean Sensitivity by Second Cross-Validation Test	Mean Specificity by Second Cross-Validation Test	Mean Accuracy by Second Cross-Validation Test
LDA#1	94.2	94.5	94.5	0.71	93.7	94.5	94.1	92.3	93.0	92.6
LDA#2	94.2	94.2	94.2	0.71	93.8	94.3	94.0	92.8	92.7	92.7
LDA#3	94.2	94.7	94.6	0.72	93.3	94.3	93.8	92.3	93.1	92.7

[1] Sensitivity is (TP/(TP + FN)) × 100, where TP is the number of true positives (number of correctly predicted X4-tropic sequences) and FN is the number of false negatives (X4-tropic sequences predicted as R5-tropic). [2] Specificity is (TN/(FP + TN)) × 100, where TN is the number of true negatives (number of correctly predicted non-X4-tropic sequences) and FP is the number of false positives (R5-tropic sequences predicted as X4-tropic). [3] Accuracy is [(TP + TN)/(TP + FP + TN + FN)] × 100. [4] MCC, the Matthew's correlation coefficient [27], is: $\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$.

In addition to the fixed index of hydropathy (AAindex entry KYTJ8201201), the three combinations of seven amino acid indices have four physicochemical properties in common: residue volume (entry GOLD730102 for LDA#1 and entry BIGC670101 for LDA#2 and LDA#3), relative mutability (entry DAYM780201 for LDA#1 and LDA#3, and entry JOND920102 for LDA#2), entropy of formation (HUTJ700103), and side chain volume (KRIW790103). Unique to LDA#1 were indices OOBM850102 (optimized propensity to form reverse turn) and ANDN9290101 (Alpha-CH chemical shifts), to LDA#2 indices GARJ730101 (partition coefficient) and FAUJ880104 (length of the side chain), and to LDA#3 indices CHAM830105 (number of atoms in the side chain) and OOBM770102 (short- and medium-range non-bonded energy per atom).

The high discriminant power of LDA#1, LDA#2, and LDA#3 is evident when examining the distribution of the corresponding LDA scores in R5-tropic and X4-tropic V3 sequences (Figure 1A–C). A detailed description of the method (LDA#1, LDA#2, and LDA#3) is reported in Supplementary File S2.

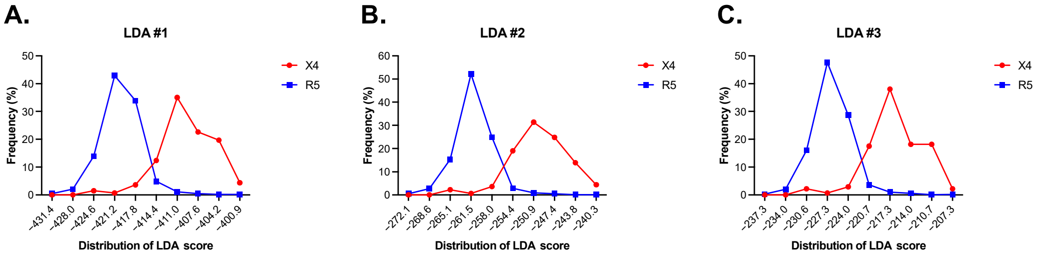

Figure 1. (**A**) Percent frequency distribution of the LDA#1 score in 1701 R5-tropic V3 sequences and in 137 X4-tropic V3 sequences. With a cut-off score of −417.38, 1608 out of 1701 R5-tropic sequences (94.5%) were predicted as R5 (score below the cut-off), and 129 out of 137 X4-tropic sequences (94.2%) as X4 (score above the cut-off). (**B**) Percent frequency distribution of the LDA#2 score in the same dataset. With a cut-off score of −258.33, 1603 out of 1701 R5-tropic sequences (94.2%) were predicted as R5 (score below the cut-off), and 129 out of 137 sequences X4-tropic (94.2%) as X4 (score above the cut-off). (**C**) Percent frequency distribution of the LDA#3 score in the same dataset. With a cut-off score of −224.01, 1610 out of 1701 R5-tropic sequences (94.6%) were predicted as R5 (score below the cut-off), and 129 out of 137 sequences X4-tropic (94.2%) as X4 (score above the cut-off).

3.2. Prediction of Tropism in a Dataset of V3 Sequences and Detection of Amino Acid Positions Associated with Coreceptor Usage

Using the three linear functions yielded, respectively, by LDA#1, LDA#2, and LDA#3, we predicted the tropism in a dataset of 2593 aligned *env* sequences (1625 of genotype B and 968 of genotype C) taken from our previous study [22]. We extracted from each *env* sequence the V3 region and predicted the tropism based on the following rule: a V3 region was classified as R5-tropic (or X4-tropic) only if predicted as such by all 3 linear functions. By this stringent approach, we assigned the tropism to 2496 out of 2593 V3 regions (96.2%): 2261 were predicted as R5-tropic (1377 of genotype B and 884 of genotype C), and 235 as X4-tropic (180 of genotype B and 55 of genotype C). The sequence data are reported in Supplementary File S3.

At each position of V3, we compared the amino acid content in R5-tropic sequences with that in sequences X4-tropic using the χ^2 contingency-table test. Over a total of 35 amino acid positions suitable for statistical analysis (frequency of gaps <5%), we found 20 showing a significant difference between R5-tropic and X4-tropic. To highlight the substitutions most closely related to the tropism, we considered only amino acid changes with prevalence $\geq$ 5% in R5- or X4-tropic sequences, respectively. We found that the 20 amino acid positions account for 42 different substitutions, 34 showing a significant prevalence in strains of the X4-tropic group and 8 in strains of the R5-tropic group (Figure 2).

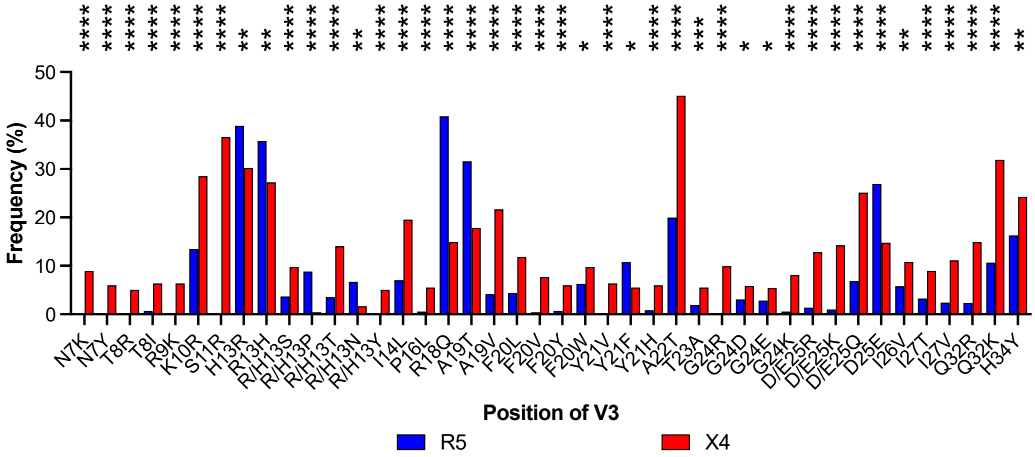

Figure 2. Percent frequency of the 42 amino acid substitutions in the V3 loop significantly associated with coreceptor tropism. In total, 34 of them show a statistically significant prevalence in strains of the X4-tropic group (red column). *, $p < 0.05$; **, $p < 0.005$; ***, $p < 0.0005$; ****, $p < 0.0001$.

A closer inspection of these results revealed several interesting points that pertain to the so-called "11/24/25 charge rule" [29]. Indeed, our results show that at position 11 of the V3 loop, there is Ser in 85.2% of R5-tropic strains. On the contrary, position 11 in X-tropic strains contains Arg at 36.6% of frequency (37% for Ser) versus only 0.09% of R5-tropic strains. Similar results were observed at position 25, where we found Asp + Glu at a frequency of 74.4% in R5-tropic strains. However, we found Arg + Lys at 24.7% in X-tropic strains (Asp + Glu at 27.6%) versus 2.4% in R5-tropic strains. Again, at position 24, we found Gly in 84.6% of R5-tropic strains and Arg + Lys in 17.3% of X4-tropic strains (Gly down to 57.9%) versus 0.8% in R5-tropic strains. Similar results were also observed at positions 18 and 32, where Arg was found at significantly higher frequency in X4-tropic strains. Specifically, at position 18, we found Arg in 42.1% of R5-tropic strains compared to 75.7% of X4-tropic strains. At position 32, we found Arg + Lys in 13.0% of R5-tropic strains compared to 33.8% of X4-tropic strains. Altogether, our results are in line with those of previous studies [30,31], showing an increased frequency of positively charged amino acids at key positions of the V3 loop of gp120 in X-tropic compared to R5-tropic strains. The quantitative assessment of the global net charge in the V3 loop has allowed the development of more accurate algorithms for the prediction of coreceptor usage [31–33] compared to the simple "11/25" rule [9].

We calculated the percent frequency of the 42 amino acid substitutions; it significantly associated with coreceptor tropism in genotypes B and C. To highlight the substitutions most closely related to the genotype, we considered only amino acid changes with prevalence > 5% in sequences of genotype B or C, respectively. Out of a total of 21 amino acid substitutions meeting the rule, we found that the great majority of them (19) showed a significant difference between the frequency of occurrence in genotype B and that in genotype C (Supplementary Table S1, Section A). For example, substitution H13R and substitution R18Q had a frequency of occurrence around 95% in genotype C, whereas they were extremely rare (around 4%) in genotype B. The percentage conservation in R5-tropic, X4-tropic, and all viral strains of each of the 20 amino acid residues undergoing substitution in Figure 2 is reported in Supplementary Table S2 (Section A).

3.3. Detection of Amino Acid Positions in the ASP Protein Associated with Coreceptor Usage

We used the dataset of 2496 *env* sequences with predicted tropism to test the presence in ASP of amino acid sites associated with R5 or X4 usage. We extracted, from each aligned sequence, the region of *env* that overlaps with *asp* (289 codon positions) and selected 170 positions suitable for statistical analysis (frequency of gaps < 5%). The *env*/*asp* is an antisense -2 overlapping region, in which two contiguous codons of *env* overlap with a single codon of *asp*. By considering $170-1$ pairs of contiguous codons in *env*, we obtained the same number of codon positions in the *asp* antisense frame.

At each codon position of *asp*, we compared the amino acid content in R5-tropic sequences with that in sequences X4-tropic using the χ^2 contingency-table test. We repeated the analysis on each codon position of *env*. This analysis was crucial for detecting the amino acid positions of ASP, in which the presence of a significant difference in composition between R5 and X4 usage was observed, while no significant difference was found between R5 and X4 usage in the two contiguous antisense positions of *env*.

Consider, for example, the amino acid position 106 of ASP, in which we found a significant difference in composition (p = 0.00007) between R5 and X4 usage (39% content in Asn for R5 and 22% for X4). In contrast, the two contiguous amino acid positions of ENV, which overlapped position 106 of ASP, did not show any significance difference in composition between R5 and X4 usage (p = 0.40 and p = 0.86, respectively). This result indicates that the amino acid diversity at position 106 is exclusive to ASP and not imposed by amino acid substitutions in ENV.

Based on this rule and considering only substitutions with prevalence $\geq$ 5% in R5- or X4-tropic sequences, respectively, we found five amino acid positions in ASP with substitutions associated preferentially with coreceptor usage (Figure 3); one is located in the N-terminal intracellular domain (position 20), two in the ectodomain (positions 106 and 119), and two in the C-terminal transmembrane domain (positions 157 and 161). At positions 20 and 157, the amino acid substitutions showed a significant prevalence in strains X4-tropic, while at positions 106, 119, and 161, they significantly prevailed in strains R5-tropic (Figure 3).

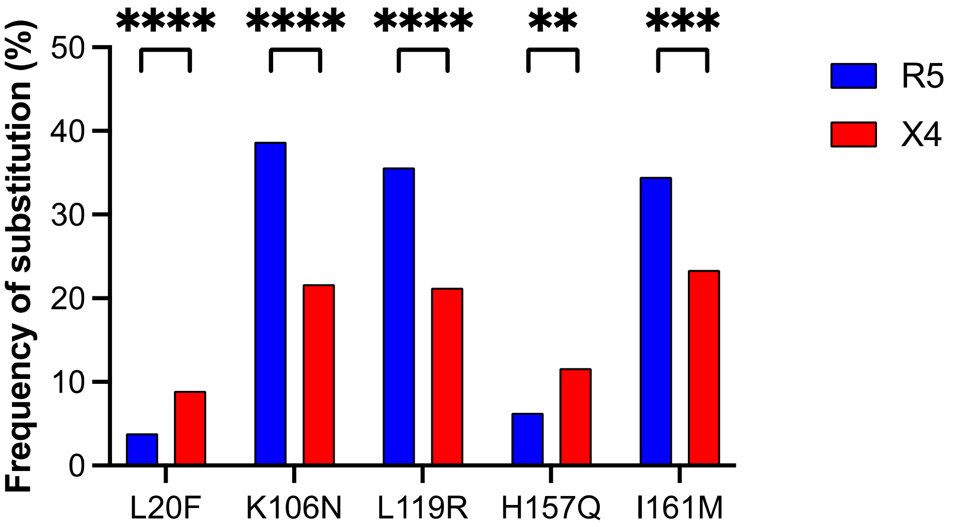

Figure 3. Percent frequency of the five amino acid substitutions in the ASP protein significantly associated with coreceptor tropism. **, $p < 0.005$; ***, $p < 0.0005$; ****, $p < 0.0001$.

We calculated the percent frequency of the five amino acid substitutions significantly associated with coreceptor tropism in the genotypes B and C. All substitutions showed a prevalence > 5% in sequences of genotype B or C, respectively. In all cases, we found a significant difference between the frequency of occurrence in genotype B and that in genotype C (Supplementary Table S1, Section B). For example, substitution K106N and substitution L119R have a frequency of occurrence above 90% in genotype C, whereas they are extremely rare (around 1.5%) in genotype B. The percentage conservation in R5-tropic, X4-tropic, and all viral strains of each of the five amino acid residues undergoing substitution in Figure 3 is reported in Supplementary Table S2 (Section B).

3.4. Detection of Significant Patterns of Association Between V3 and ASP Amino Acid Substitutions

Next, we tested if there is a significant association between the 42 amino acid substitutions in V3 (Figure 2) and the five amino acid substitutions in ASP associated with coreceptor usage (Figure 3). For each pair of substitutions, we calculated the φ correlation coefficient and its statistical significance. A positive and significant correlation (φ > 0.3) indicates that the co-occurrence of two amino acid substitutions is not due to chance and may confer advantage in terms of coreceptor usage.

We found that the highest values of the φ correlation coefficient involve substitutions located in the ectodomain of ASP (Table 2). We found that substitutions K106N and L119R of ASP are both strongly correlated (φ from 0.83 to 0.87) with substitutions H13R and R18Q of V3. This pattern of covariation occurs in ∼90% of strains of genotype C (Table 2), whereas it is extremely rare in strains of genotype B. Substitution I161M of ASP is moderately correlated (φ = 0.56) with substitutions H13R and R18Q of V3. Again, covariation is specific to genotype C, yet in this case it occurs in ∼67% of strains of genotype C (Table 2). All three substitutions in ASP (K106N, L119R, and I161M) are significantly correlated (φ from 0.37 to 0.56) with substitution A19T of V3. In this case, covariation between I161M and A19T (φ = 0.37) occurs in an even smaller fraction (47%) of strains of genotype C (Table 2).

Table 2. Co-variation between amino acid substitutions in ASP and in the V3 loop of gpl20.

ASP Substitution	Location in ASP	V3 Substitution	Prevalent Tropism	φ Binomial Correlation Coefficient [1]	Subtype Prevalence (%)
K106N	Ectodomain	H13R	R5	0.85 (***)	C (93%)
		R18Q	R5	0.87 (***)	C (90%)
		A19T	R5	0.56 (**)	C (63%)
L119R	Ectodomain	H13R	R5	0.83 (***)	C (86%)
		R18Q	R5	0.85 (***)	C (87%)
		A19T	R5	0.56 (**)	C (60%)
I161M	C-terminal transmembrane domain	H13R	R5	0.56 (**)	C (67%)
		R18Q	R5	0.55 (**)	C (68%)
		A19T	R5	0.37 (*)	C (47%)
L20F	N-terminal intracellular domain	I14L	X4	0.45 (*)	B (4.4%)
		F20W	X4	0.67 (**)	B (5.7%)
		D/E25Q	X4	0.37 (*)	B (3.9%)
H157Q	C-terminal transmembrane domain	I14L	X4	0.34 (*)	B (5.6%)
		F20W	X4	0.51 (**)	B (5.8%)

[1] A φ value from 0.3 to 0.7 indicates a weak or medium association (*; **), a φ value from 0.7 to 1.0 indicates a strong association (***).

As shown at the bottom of Table 2, we found a weak/moderate correlation (φ from 0.34 to 0.67) between substitutions L20F and H157Q of ASP and substitutions I14L and F20W of V3. This pattern of covariation is virtually absent in genotype C, and it occurs only in a small fraction (∼5%) of genotype B strains (Table 2). Covariation between substitution L20F of ASP and substitution D/E25Q of V3 (φ = 0.37) occurs in an even smaller fraction (3.9%) of strains of genotype B (Table 2).

3.5. Detection of Amino Acid Positions in the V1/V2 Region Associated with Coreceptor Usage and Detection of Significant Patterns of Association Between V1/V2 and ASP Substitutions

It has long been known that the V1/V2 region of ENV, although to a lesser extent than V3, can determine the selectivity of HIV-1 strains for CCR5 or CXCR4 [34–36]. Therefore, we carried out a comparative analysis of strains R5- and X4-tropic for the amino acid composition in the V1/V2 region, with the aim to detect positions associated with coreceptor usage.

We extracted from each of the 2496 *env* sequences with predicted tropism the V1/V2 region (from codon position 189 to 350). After translation into amino acids, we obtained 2496 V1/V2 regions, aligned for a total of 162 positions. For statistical analysis, we selected 65 positions, in which the frequency of gaps was <5%, and compared their amino acid content in R5-tropic sequences with that in X4-tropic sequences using the χ^2 contingency-table test. We found 10 positions showing a significant difference between R5 and X4 usage. They accounted for 13 different amino acid substitutions, of which 10 significantly prevailed in strains X4-tropic and 3 in strains of the R5-tropic group (Figure 4). We then tested if there was a significant covariation between the 13 substitutions in V1/V2 and the 5 substitutions in ASP associated with coreceptor usage (Figure 3).

As shown in Table 3, we found that substitutions K106N and L119R of ASP are strongly correlated (φ = 0.78) with substitution I/L72M and moderately correlated (φ = 0.60 and φ = 0.57) with substitution S83E of V1/V2. A weaker correlation was observed between substitutions I161M of ASP and I/L72M (φ = 0.51) and S83E (φ = 0.36) of V1/V2. This pattern of covariation is unique to genotype C, where it occurs with a frequency of between 39 and 82% (Table 3). In addition, we found a strong/moderate correlation between substitutions L20F and H157Q of ASP and substitution D55T of V1/V2 (φ = 0.88 and φ = 0.68). This pattern of covariation occurs only in a small fraction (∼5%) of strains of genotype B (Table 3).

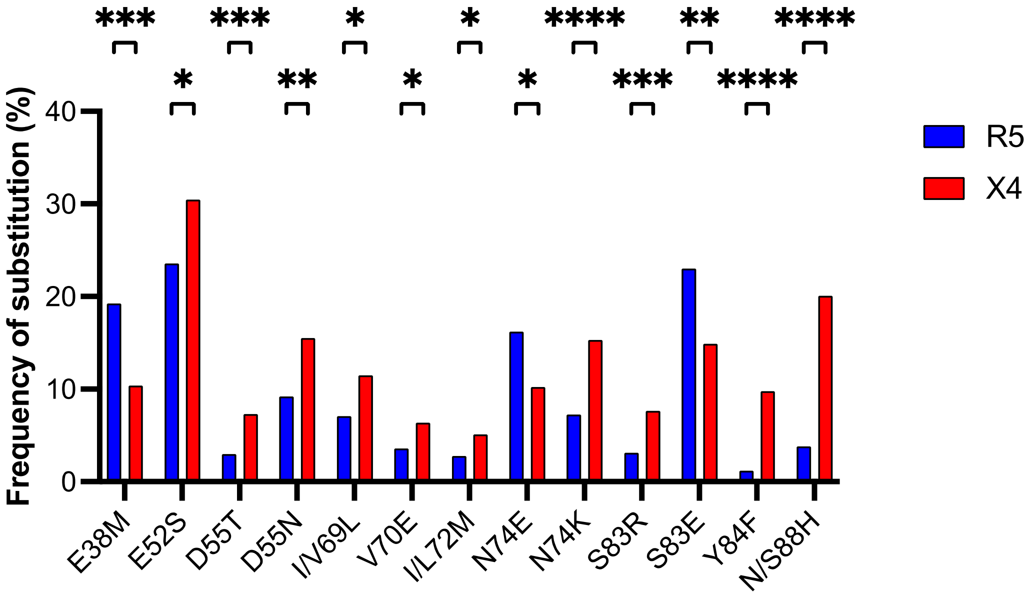

Figure 4. Percent frequency of the 13 amino acid substitutions in the V1/V2 loop significantly associated with coreceptor tropism. In total, 10 of them show a statistically significant prevalence in strains of the X4-tropic group (red column). *, $p < 0.05$; **, $p < 0.005$; ***, $p < 0.0005$; ****, $p < 0.0001$.

Table 3. Co-variation in amino acid substitutions in ASP and the V1/V2 loop of gp120.

ASP Substitution	Location in ASP	V1/V2 Substitution	Prevalent Tropism	φ Binomial Correlation Coefficient [1]	Subtype Prevalence (%)
K106N	Ectodomain	I/L72M S83E	X4 R5	0.78 (***) 0.60 (**)	C (82%) C (54%)
L119R	Ectodomain	I/L72M S83E	X4 R5	0.78 (***) 0.57 (**)	C (78%) C (50%)
I161M	C-terminal transmembrane domain	I/L72M S83E	X4 R5	0.51 (**) 0.36 (*)	B (65%) B (39%)
L20F	N-terminal intracellular domain	D55T	X4	0.88 (***)	C (5.2%)
H157Q	C-terminal transmembrane domain	D55T	X4	0.68 (***)	C (5.3%)

[1] A φ value from 0.3 to 0.7 indicates a weak or medium association (*; **), a φ value from 0.7 to 1.0 indicates a strong association (***).

Finally, we calculated the percent frequency of the 13 amino acid substitutions significantly associated with coreceptor tropism in the genotypes B and C. To highlight the substitutions most closely related to the genotype, we considered only amino acid changes with prevalence > 5% in sequences of genotype B or C, respectively. Out of a total of 11 amino acid substitutions meeting the rule, we found that all of them show a significant difference between the frequency of occurrence in genotype B and that in genotype C (Supplementary Table S1, Section C). The substitutions for which the frequency in either genotype shows greater difference compared to the other were E52S (38.9% in genotype B and 0.0 in genotype C) and S83E (1.9% in genotype B and 56.0% in genotype C). The percentage conservation in R5-tropic, X4-tropic, and all viral strains of each of the 10 amino acid residues undergoing substitution shown in Figure 4 is reported in Supplementary Table S2 (Section C).

3.6. Extension of Statistical Analysis to the Regions of gp120 Protein Outside V3 and V1/V2

We extended the analysis performed on V3 and V1/V2 to the remaining regions of gp120 protein. We considered these regions as controls, because they do not appear to be involved in viral tropism. From each of the 2496 *env* sequences with predicted tropism, we extracted the following regions: (i) the region upstream V1/V2, not overlapping the *vpu* gene (from codon position 51 to 188 in the full-length aligned *env* sequences); (ii) the region between V1/V2 and V3 (from codon position 351 to 478); and (iii) the region downstream V3, not overlapping the antisense *asp* gene (from codon position 532 to 605).

After translation into amino acids, we obtained 3 datasets of 2496 sequences, aligned, respectively, for a total of 138, 128, and 74 positions. Within the 3 datasets, we selected, respectively, 88, 99, and 49 positions in which the frequency of gaps was <5%, and we compared their amino acid content in R5-tropic sequences with that in X4-tropic sequences using the χ^2 contingency-table test. We found 3 positions in the region upstream V1/V2, 7 positions in the region between V1/V2 and V3, and 6 positions in the region downstream V3 showing a significant difference between R5 and X4 usage. Overall, we found that the fraction of amino acid sites associated with coreceptor usage is remarkably low (16 out of a total of 236; 6.7%), about one order of magnitude smaller than that found in V3 (20 out of 35; 57.1%) and half of that found in V1/V2 (10 out of 65; 15.4%). This result is in line with long-established evidence that V3 and, to a lesser extent, V1/V2 are the determinants of co-receptor tropism.

The three positions we found in the region upstream V1/V2 account for three different amino acid substitutions. They show a significant covariation (φ from 0.36 to 0.86) with the ASP substitutions K106N, L119R and I161M (Table 4). The seven positions we found in the region between V1/V2 and V3 account for eight different amino acid substitutions. Four of them show a significant covariation (φ from 0.38 to 0.82) with the ASP substitutions K106N, L119R and I161M (Table 4). The six positions we found in the region downstream V3 account for 12 different amino acid substitutions. Five of them show a significant covariation (φ from 0.32 to 0.65) with the ASP substitutions K106N, L119R and I161M (Table 4). All these covariations were found only in genotype C strains, where they occur with a frequency of between 20 and 89%. Unlike the V3 and V1/V2 regions, none of the amino acid substitutions found in the control regions show a significant covariation with the ASP substitutions L20F and H157Q.

Table 4. Amino acid substitutions in the ASP protein significantly associated with amino acid substitutions in three regions of ENV encoded in env sequences outside the overlap with *asp*.

ASP Substitution	Region of gp120	Substitution	Prevalent Tropism	φ Binomial Correlation Coefficient [1]	Subtype Prevalence (%)
K106N	Upstream of V1/V2	A1V	R5	0.63 (**)	C (64%)
		T36K	R5	0.86 (***)	C (89%)
		T51R	R5	0.37 (*)	C (26%)
L119R	Upstream of V1/V2	A1V	R5	0.66 (**)	C (63%)
		T36K	R5	0.84 (***)	C (85%)
		T51R	R5	0.36 (*)	C (25%)
I161M	Upstream of V1/V2	A1V	R5	0.42 (*)	C (48%)
		T36K	R5	0.57 (**)	C (67%)
K106N	Between V1/V2 and V3	V/T4A	R5	0.40 (*)	C (30%)
		F28Y	R5	0.82 (***)	C (86%)
		F99L	R5	0.67 (**)	C (79%)
		N128V	R5	0.56 (**)	C (45%)

Table 4. *Cont.*

ASP Substitution	Region of gp120	Substitution	Prevalent Tropism	φ Binomial Correlation Coefficient [1]	Subtype Prevalence (%)
L119R	Between V1/V2 and V3	V/T4A	R5	0.41 (*)	C (41%)
		F28Y	R5	0.81 (***)	C (82%)
		F99L	R5	0.69 (**)	C (76%)
		N128V	R5	0.57 (**)	C (44%)
I161M	Between V1/V2 and V3	V/T4A	R5	0.30 (*)	C (24%)
		F28Y	R5	0.55 (**)	C (65%)
		F99L	R5	0.45 (*)	C (59%)
		N128V	R5	0.38 (*)	C (34%)
K106N	Downstream of V3	A/V17S	R5	0.33 (*)	C (22%)
		A/V17G	R5	0.34 (*)	C (20%)
		Q28H	R5	0.65 (**)	C (67%)
		V/I48K	R5	0.36 (*)	C (27%)
		G70R	R5	0.63 (**)	C (69%)
L119R	Downstream of V3	A/V17S	R5	0.32 (*)	C (21%)
		A/V17G	R5	0.34 (*)	C (20%)
		Q28H	R5	0.65 (**)	C (63%)
		V/I48K	R5	0.34 (*)	C (25%)
		G70R	R5	0.61 (**)	C (66%)
I161M	Downstream of V3	Q28H	R5	0.45 (*)	C (51%)
		G70R	R5	0.39 (*)	C (51%)

[1] A φ value from 0.3 to 0.7 indicates a weak or medium association (*; **), a φ value from 0.7 to 1.0 indicates a strong association (***).

The results from the analysis of all five regions of gp120 (upstream of V1/V2, V1/V2, between V1/V2 and V3, V3, and downstream of V3) are summarized in Table 5. We found that when the five regions are combined in two groups (group 1 = V1/V2 + V3; group 2 = upstream of V1/V2 + between V1/V2 and V3 + downstream of V3), the mean value of the correlation coefficient φ of group 1 (0.61; sd = 0.18) is significantly greater than that of group 2 (0.52; sd = 0.17; $p = 0.03$, one-tailed *t*-Student test).

Table 5. Sequence analysis for coreceptor tropism and covariation test for five different regions of the gp120 protein.

Region of the ENV Glycoprotein gp120	Positions Examined (gaps < 5%)	Positions Associated to Tropism (%)	Amino Acid Substitutions for Covariation Test	ASP Substitutions Significantly Correlated	Mean Value of φ Correlation Coefficient
V3 loop	35	20 (57.1)	42	5	0.60
V1/V2 loops	65	10 (15.4)	13	5	0.65
Upstream of V1/V2	88	3 (3.4)	3	3	0.59
Between V1/V2 and V3	99	7 (7.1)	8	3	0.55
Downstream of V3	49	6 (12.2)	12	3	0.46
Group 1 [1]	100	30 (30)	55	5	0.61 [3]
Group 2 [2]	236	16 (6.7)	23	3	0.52 [3]

[1] V1/V2 + V3. [2] Upstream of V1/V2 + between V1/V2 and V3 + downstream of V3. [3] Group 1 vs. Group 2; *t*-Student = 1.87; $p = 0.03$.

4. Discussion

The aim of this study was to identify amino acid signatures in the HIV-1 antisense protein ASP that are associated with coreceptor tropism. We have identified five positions of

ASP where specific amino acid substitutions correlated with amino acid substitutions in the V3 and V1/V2 loops of gp120, which, in turn, are significantly associated with R5- or X4-tropism.

Identification of ASP amino acid signatures that are associated with co-receptor tropism required first the development of a computational method capable of predicting HIV-1 tropism from the amino acid sequence of the gp120 V3 loop with high sensitivity and specificity. Several computational methods for predicting viral tropism have been developed [9–19] and significant efforts are still underway to improve the accuracy of genotypic tests based on the sequence of the V3 loop domain. Given the considerable structural flexibility and sequence variability of the V3 loop, individual features of this region distinguishing between the two virus phenotypes R5- and X4-tropic are complex and difficult to define. Multivariate statistical methods [21] have the ability to summarize the information carried by individual variables into a few synthetic variables (e.g., the principal component analysis) or into a single synthetic variable (e.g., the linear discriminant analysis, LDA), leading to a better understanding of the biological process under examination [37–39]. In this study, we used Fisher's linear discriminant analysis [20,40] to find the best combination of amino acid physicochemical features in distinguishing between the two virus phenotypes R5- and X4-tropic. Our study showed that LDA is a promising approach to predict the viral tropism both in terms of sensitivity and specificity (Table 1). The ability of LDA to maximize the variance between groups and minimize the variance within groups can be appreciated by examining the three distributions of the LDA score in R5- and X4-tropic V3 sequences (Figure 1).

Despite the advantages, our method presents a few limitations. The first limitation is that it predicts the tropism only for genotypes B and C. Although they account, respectively, for 11.3% and 50.4% of the global genetic diversity of HIV-1 [41], LDA did not analyze V3 sequences from genotype A (global prevalence of 12.4%), D, F, and G (6.4%), from circulating recombinant forms (15.1%), and from unique recombinant forms (2.0%). Another limitation is the low number of V3 sequences X4-tropic, only 137, in our training dataset. Although most of the V3 sequences from the viral samples are in the R5-tropic group, it is important to increase the number of V3 sequences in the X4-tropic group as much as possible. This should improve, in turn, the ability of the two validation tests to assess the robustness of LDA.

Although LDA was originally developed to classify subjects into one of the two clearly defined groups, it was later expanded to classify subjects into more than two groups. Indeed, as shown in his seminal study [20], Fisher successfully developed a linear discriminant model to distinguish three related species of the genus *Iris* from each other. Therefore, we could extend LDA to three groups: V3 sequences from R5-, X4-, and R5X4-tropic strains. It would be interesting to determine how dual tropic strains stack up against strains R5- and X4-tropic strains and test whether the switch from one tropism to the other is driven by small, but significant, changes in the sequence properties of the V3 loop domain.

After the initial discovery of the *asp* gene by Miller [42], Cassan et al. showed that an intact *asp* ORF is present only in pandemic HIV-1 strains (group M), and it is absent in all other primate lentiviruses, including non-pandemic HIV-1 groups [43]. We showed the existence of selective pressure that maintains an intact *asp* ORF in the HIV-1 genome by conserving the *start* codon and by avoiding early *stop* codons [23]. Recently, we proposed a role for the ASP protein in promoting intra-host viral spread or pathogenesis, based on a significant association between the frequency of viral strains encoding a full-length ASP and disease progression [22]. An extensive sequence analysis by Dimonte first hypothesized an involvement of ASP in coreceptor selection through identification of amino acid substitutions associated with classical signatures of viral tropism in the V3 loop of gp120 [44]. Examples of covariation in amino acid substitutions have also been reported for other HIV-1 proteins [45–49].

The first question in this study was whether there are positions in ASP where amino acid substitutions are preferentially associated with coreceptor usage. From a dataset of 2593 aligned ENV sequences, we extracted the corresponding V3 sequences and predicted the tropism by means of linear discriminant analysis. The percent frequency of V3 sequences with a predicted X4 tropism (235 out of 2496; 9.4%) is comparable to that of V3 sequences with a known X4 tropism (137 out of 1838 in the training dataset; 7.5%). Our analysis identified five amino acid positions in ASP where substitutions were preferentially associated with tropism (Figure 3). The number of positions and substitutions we found in ASP (5 and 5, respectively) is much smaller than the ones reported by Dimonte (36 positions for 58 substitutions) [44]. This large difference is due to the fact that we restricted our analysis exclusively to substitutions in ASP caused by mutations that are non-synonymous in the *asp* ORF but synonymous in *env*. These substitutions are not a byproduct and occur independently of ENV evolution.

The second question of this study was whether there is a significant covariation between the tropism-associated substitutions in ASP and the tropism-associated substitutions in the V3 loop domain. We first identified 20 amino acid positions in V3 that are significantly associated with tropism, including the classical ones 11 and 25 [9], for a total of 42 substitutions (Figure 2). In this case, the result is very similar to the one obtained by Dimonte, who identified 24 amino acid positions in V3 for 44 substitutions [44]. In response to the question, we found that all 5 substitutions in ASP associated with tropism show a significant covariation with 6 of the 42 substitutions in V3 associated with tropism (Table 2). It should be noted that substitutions K106N and L119R map in the ectodomain of ASP and are the ones most significantly associated with substitutions in V3 that determine coreceptor tropism (correlation coefficient φ from 0.56 to 0.87). Altogether, these findings support the hypothesis that ASP may be involved in coreceptor selection, which can be validated through in vitro and in vivo studies.

Crystal and cryo-electron microscopy structure studies on ENV are indicative of a concerted action of V1/V2 and V3 loops during HIV-1 entry [50,51]. They have revealed that V1/V2 loops are located at the trimer apex in connection with V3 loop and that the conformational change in V1/V2 loops upon CD4 binding triggers exposure of V3 loop to the coreceptor. In the V1/V2 region, we identified 10 amino acid positions and 13 substitutions significantly associated with tropism (Figure 4). Despite the greater length of V1/V2, these numbers were remarkably lower than the ones we found in V3 (20 positions and 42 substitutions), in line with the notion that the V3 loop is the major determinant of coreceptor tropism [6]. Interestingly, we found that all the 5 substitutions in ASP associated with tropism show a significant covariation with 3 of the 13 substitutions in V1/V2 associated with tropism (Table 3). Also in this case, the two substitutions in the ASP ectodomain, K106N and L119R, are the ones showing strongest covariation with substitutions in V1/V2, which are associated with coreceptor tropism (correlation coefficient from 0.56 to 0.87). Again, this finding supports the hypothesis of an ASP involvement in viral tropism.

Extending our analyses to regions of gp120 outside of the first three hypervariable loops (excluding the overlaps with the *vpu* and *asp* genes), we identified a total of 16 positions and 23 substitutions associated with co-receptor tropism. Of these, 12 show covariation with substitutions in ASP, with some significant differences (Table 4). First, covariation was observed with only three of five substitutions in ASP (K106N, L119R, and I161M). Second, the mean coefficient of covariation φ in three regions combined is significantly lower than the one observed for the combined V3 and V1/V2 (Table 5).

It is also noteworthy that the majority of covariations in V3, in V1/V2, and in the regions outside these three hypervariable loops were found almost exclusively in genotype C strains at frequencies as high as >90%. Genotype C is by far the most prevalent HIV-1 clade worldwide [43], suggesting a potential association between specific substitutions in ASP and viral spread within and among hosts. This hypothesis had already been proposed in a previous report [43], which indeed noted a correlation between the frequency of viral strains within each HIV-1 genotype containing an uninterrupted *asp* ORF and the prevalence of the genotype. This hypothesis is also supported by two recent studies. In the first one, we reported that viral isolates with an uninterrupted *asp* ORF are found at significantly higher frequencies in people living with HIV-1 (PLWH) who progress to AIDS in <3 years (rapid progressors) than in those who progress in >12 years (long term non-progressors) [22]. Another recent study described the presence of antibodies to ASP in serum of PLWH [52]. Remarkably, these responses were found in viremic and elite controllers and were found to be directed primarily to epitopes mapping in the ectodomain of ASP [52].

Altogether, our studies suggest that amino acid substitutions in ASP are linked to substitutions in gp120, which are signatures of R5 or X4 coreceptor tropism. Whether the covariations identified in the present study have an actual bearing on R5- versus X4-dependent viral entry remains to be addressed experimentally.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/v17030323/s1: File S1: Training dataset of 1838 amino acid sequences of the V3 loop domain with known tropism; File S2: Linear discriminant analyses on a training dataset of 1838 V3 sequences with known tropism; File S3: Dataset of 2496 amino acid sequences of the V3 loop. Table S1: Frequency within genotypes B and C of amino acid substitutions in the V1/V2 and V3 loops of gp120, and in ASP. Table S2: Percent conservation of the most frequent amino acid in the positions of the V1/V2 and V3 loop of gp120 and in ASP showing a significant difference between R5-tropic and X4-tropic strains

Author Contributions: Conceptualization, A.P. and F.R.; methodology, A.P.; software, A.P.; formal analysis, A.P.; investigation, A.P. and F.R.; data curation, A.P. and F.R.; writing—original draft preparation, A.P. and F.R.; writing—review and editing, A.P. and F.R.; visualization, F.R.; supervision, A.P. and F.R.; project administration, A.P. and F.R.; funding acquisition, A.P. and F.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Institutes of Health grants AI144983 and AI172556 (F.R.).

Data Availability Statement: Data supporting the results reported in this study can be found in the Supplementary Materials Section and at www.hiv.lanl.gov/content/index (accessed on 10 December 2014).

Acknowledgments: This work benefited from the equipment and framework of the COMP-HUB and COMP-R Initiatives, funded by the "Departments of Excellence" program of the Italian Ministry for University and Research (MIUR, 2023–2027).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kwong, P.D.; Wyatt, R.; Robinson, J.; Sweet, R.W.; Sodroski, J.G.; Hendrickson, W.A. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **1998**, *393*, 648–659. [CrossRef] [PubMed]
2. Weissenhorn, W.; Dessen, A.; Harrison, S.C.; Skehel, J.J.; Wiley, D.C. Atomic structure of the ectodomain from HIV-1 gp41. *Nature* **1997**, *387*, 426–430. [CrossRef] [PubMed]
3. Deng, H.; Liu, R.; Ellmeier, W.; Choe, S.; Unutmaz, D.; Burkhart, M.; Di Marzio, P.; Marmon, S.; Sutton, R.E.; Hill, C.M.; et al. Identification of a major co-receptor for primary isolates of HIV-1. *Nature* **1996**, *381*, 661–666. [CrossRef] [PubMed]

4. Lu, Z.; Berson, J.F.; Chen, Y.; Turner, J.D.; Zhang, T.; Sharron, M.; Jenks, M.H.; Wang, Z.; Kim, J.; Rucker, J.; et al. Evolution of HIV-1 coreceptor usage through interactions with distinct CCR5 and CXCR4 domains. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 6426–6431. [CrossRef] [PubMed]
5. Wyatt, R.; Sodroski, J.G. The HIV-1 envelope glycoproteins: Fusogens, antigens, and immunogens. *Science* **1998**, *280*, 1884–1888. [CrossRef] [PubMed]
6. Hwang, S.S.; Boyle, T.J.; Lyerly, H.K.; Cullen, B.R. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science* **1991**, *253*, 71–74. [CrossRef] [PubMed]
7. Regoes, R.R.; Bonhoeffer, S. The HIV coreceptor switch: A population dynamical perspective. *Trends Microbiol.* **2005**, *13*, 269–277. [CrossRef] [PubMed]
8. Whitcomb, J.M.; Huang, W.; Fransen, S.; Limoli, K.; Toma, J.; Wrin, T.; Chappey, C.; Kiss, L.D.; Paxinos, E.E.; Petropoulos, C.J. Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism. *Antimicrob. Agents Chemother.* **2007**, *51*, 566–575. [CrossRef]
9. De Jong, J.J.; De Ronde, A.; Keulen, W.; Tersmette, M.; Goudsmit, J. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: Analysis by single amino acid substitution. *J. Virol.* **1992**, *66*, 6777–6780. [CrossRef]
10. Lengauer, T.; Sander, O.; Sierra, S.; Thielen, A.; Kaiser, R. Bioinformatics prediction of HIV coreceptor usage. *Nat. Biotechnol.* **2007**, *25*, 1407–1410. [CrossRef] [PubMed]
11. Resch, W.; Hoffman, N.; Swanstrom, R. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology* **2001**, *288*, 51–62. [CrossRef]
12. Pillai, S.; Good, B.; Richman, D.; Corbeil, J. A new perspective on V3 phenotype prediction. *AIDS Res. Hum. Retroviruses* **2003**, *19*, 145–149. [CrossRef] [PubMed]
13. Kumar, R.; Raghava, G.P. Hybrid approach for predicting coreceptor used by HIV-1 from its V3 loop amino acid sequence. *PLoS ONE* **2013**, *8*, e61437. [CrossRef]
14. Jensen, M.A.; Li, F.S.; van't Wout, A.B.; Nickle, D.C.; Shriner, D.; He, H.X.; McLaughlin, S.; Shankarappa, R.; Margolick, J.B.; Mullins, J.I. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J. Virol.* **2003**, *77*, 13376–13388. [CrossRef] [PubMed]
15. Bozek, K.; Lengauer, T.; Sierra, S.; Kaiser, R.; Domingues, F.S. Analysis of physicochemical and structural properties determining HIV-1 coreceptor usage. *PLoS Comput. Biol.* **2013**, *9*, e1002977. [CrossRef] [PubMed]
16. Shen, H.S.; Yin, J.; Leng, F.; Teng, R.F.; Xu, C.; Xia, X.Y.; Pan, X.M. HIV coreceptor tropism determination and mutational pattern identification. *Sci. Rep.* **2016**, *6*, 21280. [CrossRef]
17. Chen, X.; Wang, Z.X.; Pan, X.M. HIV-1 tropism prediction by the XGboost and HMM methods. *Sci. Rep.* **2019**, *9*, 9997. [CrossRef] [PubMed]
18. Vandekerckhove, L.P.; Wensing, A.M.; Kaiser, R.; Brun-Vezinet, F.; Clotet, B.; De Luca, A.; Dressler, S.; Garcia, F.; Geretti, A.M.; Klimkait, T.; et al. European guidelines on the clinical management of HIV-1 tropism testing. *Lancet Infect. Dis.* **2011**, *11*, 394–407. [CrossRef] [PubMed]
19. Briggs, D.R.; Tuttle, D.L.; Sleasman, J.W.; Goodenow, M.M. Envelope V3 amino acid sequence predicts HIV-1 phenotype (co-receptor usage and tropism for macrophages). *AIDS* **2000**, *14*, 2937–2939. [CrossRef]
20. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *8*, 376–386. [CrossRef]
21. Morrison, D.F. *Multivariate Statistical Methods*; McGraw-Hill: New York, NY, USA, 1976.
22. Pavesi, A.; Romerio, F. Creation of the HIV-1 antisense gene asp coincided with the emergence of the pandemic group M and is associated with faster disease progression. *Microbiol. Spectr.* **2024**, *12*, e0380223. [CrossRef]
23. Pavesi, A.; Romerio, F. Extending the Coding Potential of Viral Genomes with Overlapping Antisense ORFs: A Case for the De Novo Creation of the Gene Encoding the Antisense Protein ASP of HIV-1. *Viruses* **2022**, *14*, 146. [CrossRef] [PubMed]
24. Affram, Y.; Zapata, J.C.; Gholizadeh, Z.; Tolbert, W.D.; Zhou, W.; Iglesias-Ussel, M.D.; Pazgier, M.; Ray, K.; Latinovic, O.S.; Romerio, F. The HIV-1 Antisense Protein ASP Is a Transmembrane Protein of the Cell Surface and an Integral Protein of the Viral Envelope. *J. Virol.* **2019**, *93*. [CrossRef]
25. Gholizadeh, Z.; Iqbal, M.S.; Li, R.; Romerio, F. The HIV-1 Antisense Gene ASP: The New Kid on the Block. *Vaccines* **2021**, *9*, 513. [CrossRef] [PubMed]
26. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **2008**, *36*, D202–D205. [CrossRef] [PubMed]
27. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132. [CrossRef]
28. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Et Biophys. Acta* **1975**, *405*, 442–451. [CrossRef]

29. Cardozo, T.; Kimura, T.; Philpott, S.; Weiser, B.; Burger, H.; Zolla-Pazner, S. Structural basis for coreceptor selectivity by the HIV type 1 V3 loop. *AIDS Res. Hum. Retroviruses* **2007**, *23*, 415–426. [CrossRef]
30. Sander, O.; Sing, T.; Sommer, I.; Low, A.J.; Cheung, P.K.; Harrigan, P.R.; Lengauer, T.; Domingues, F.S. Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage. *PLoS Comput. Biol.* **2007**, *3*, e58. [CrossRef]
31. Montagna, C.; De Crignis, E.; Bon, I.; Re, M.C.; Mezzaroma, I.; Turriziani, O.; Graziosi, C.; Antonelli, G. V3 net charge: Additional tool in HIV-1 tropism prediction. *AIDS Res. Hum. Retroviruses* **2014**, *30*, 1203–1212. [CrossRef]
32. Kieslich, C.A.; Shin, D.; Lopez de Victoria, A.; Gonzalez-Rivera, G.; Morikis, D. A predictive model for HIV type 1 coreceptor selectivity. *AIDS Res. Hum. Retroviruses* **2013**, *29*, 1386–1394. [CrossRef] [PubMed]
33. Pramanik, L.; Fried, U.; Clevestig, P.; Ehrnst, A. Charged amino acid patterns of coreceptor use in the major subtypes of human immunodeficiency virus type 1. *J. Gen. Virol.* **2011**, *92*, 1917–1922. [CrossRef] [PubMed]
34. Cho, M.W.; Lee, M.K.; Carney, M.C.; Berson, J.F.; Doms, R.W.; Martin, M.A. Identification of determinants on a dualtropic human immunodeficiency virus type 1 envelope glycoprotein that confer usage of CXCR4. *J. Virol.* **1998**, *72*, 2509–2515. [CrossRef]
35. Lee, M.K.; Heaton, J.; Cho, M.W. Identification of determinants of interaction between CXCR4 and gp120 of a dual-tropic HIV-1DH12 isolate. *Virology* **1999**, *257*, 290–296. [CrossRef] [PubMed]
36. Labrosse, B.; Treboute, C.; Brelot, A.; Alizon, M. Cooperation of the V1/V2 and V3 domains of human immunodeficiency virus type 1 gp120 for interaction with the CXCR4 receptor. *J. Virol.* **2001**, *75*, 5457–5464. [CrossRef]
37. David, C.C.; Jacobs, D.J. Principal component analysis: A method for determining the essential dynamics of proteins. *Methods Mol. Biol.* **2014**, *1084*, 193–226. [CrossRef]
38. Pavesi, A.; Romerio, F. Different Patterns of Codon Usage and Amino Acid Composition across Primate Lentiviruses. *Viruses* **2023**, *15*, 1580. [CrossRef] [PubMed]
39. He, L.; Dong, R.; He, R.L.; Yau, S.S. A novel alignment-free method for HIV-1 subtype classification. *Infect. Genet. Evol.* **2020**, *77*, 104080. [CrossRef] [PubMed]
40. Mardia, K.V. Fisher's pioneering work on discriminant analysis and its impact on Artificial Intelligence. *J. Multivar. Anal.* **2024**, *203*, 105341. [CrossRef]
41. Nair, M.; Gettins, L.; Fuller, M.; Kirtley, S.; Hemelaar, J. Global and regional genetic diversity of HIV-1 in 2010-21: Systematic review and analysis of prevalence. *Lancet Microbe* **2024**, *5*, 100912. [CrossRef] [PubMed]
42. Miller, R.H. Human immunodeficiency virus may encode a novel protein on the genomic DNA plus strand. *Science* **1988**, *239*, 1420–1422. [CrossRef] [PubMed]
43. Cassan, E.; Arigon-Chifolleau, A.M.; Mesnard, J.M.; Gross, A.; Gascuel, O. Concomitant emergence of the antisense protein gene of HIV-1 and of the pandemic. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 11537–11542. [CrossRef] [PubMed]
44. Dimonte, S. Different HIV-1 env frames: gp120 and ASP (antisense protein) biosynthesis, and theirs co-variation tropic amino acid signatures in X4- and R5-viruses. *J. Med. Virol.* **2017**, *89*, 112–122. [CrossRef] [PubMed]
45. Zhao, Y.; Wang, Y.; Gao, Y.; Li, G.; Huang, J. Integrated analysis of residue coevolution and protein structures capture key protein sectors in HIV-1 proteins. *PLoS ONE* **2015**, *10*, e0117506. [CrossRef] [PubMed]
46. Arachchilage, M.H.; Piontkivska, H. Coevolutionary Analysis Identifies Protein-Protein Interaction Sites between HIV-1 Reverse Transcriptase and Integrase. *Virus Evol.* **2016**, *2*, vew002. [CrossRef] [PubMed]
47. Li, G.; Theys, K.; Verheyen, J.; Pineda-Pena, A.C.; Khouri, R.; Piampongsant, S.; Eusebio, M.; Ramon, J.; Vandamme, A.M. A new ensemble coevolution system for detecting HIV-1 protein coevolution. *Biol. Direct* **2015**, *10*, 1. [CrossRef] [PubMed]
48. Rossi, A.H.; Rocco, C.A.; Mangano, A.; Sen, L.; Aulicino, P.C. Sequence variability in p6 gag protein and gag/pol coevolution in human immunodeficiency type 1 subtype F genomes. *AIDS Res. Hum. Retroviruses* **2013**, *29*, 1056–1060. [CrossRef]
49. Travers, S.A.; Tully, D.C.; McCormack, G.P.; Fares, M.A. A study of the coevolutionary patterns operating within the env gene of the HIV-1 group M subtypes. *Mol. Biol. Evol.* **2007**, *24*, 2787–2801. [CrossRef] [PubMed]
50. Julien, J.P.; Cupo, A.; Sok, D.; Stanfield, R.L.; Lyumkis, D.; Deller, M.C.; Klasse, P.J.; Burton, D.R.; Sanders, R.W.; Moore, J.P.; et al. Crystal structure of a soluble cleaved HIV-1 envelope trimer. *Science* **2013**, *342*, 1477–1483. [CrossRef]
51. Lyumkis, D.; Julien, J.P.; de Val, N.; Cupo, A.; Potter, C.S.; Klasse, P.J.; Burton, D.R.; Sanders, R.W.; Moore, J.P.; Carragher, B.; et al. Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. *Science* **2013**, *342*, 1484–1490. [CrossRef] [PubMed]
52. Caetano, D.G.; Napoleao-Pego, P.; Villela, L.M.; Cortes, F.H.; Cardoso, S.W.; Hoagland, B.; Grinsztejn, B.; Veloso, V.G.; De-Simone, S.G.; Guimaraes, M.L. Patterns of Diversity and Humoral Immunogenicity for HIV-1 Antisense Protein (ASP). *Vaccines* **2024**, *12*, 771. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.