DOCTORAL THESIS 2013

Evolution of the Human Immunodeficiency Virus type I Protease and Integrase: Effects on Viral Replication Capacity and Robustness.

Elena Capel Malo

IrsiCaixa
Institut de Recerca de la Sida

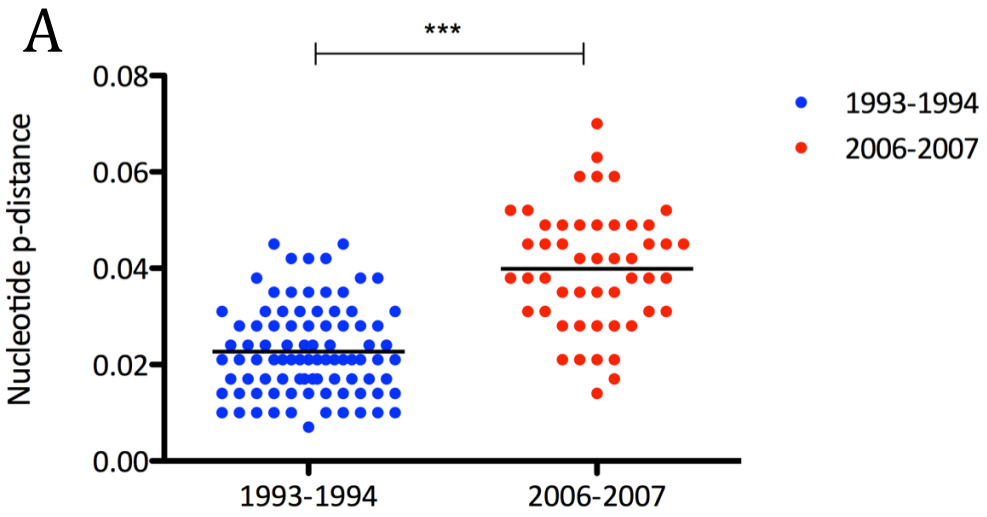
UAB
Universitat Autònoma de Barcelona

Results

Comparison of the sequence conservation of the Protease, Integrase and Gag genes

Decrease in HIV-1 protease sequence conservation over time

HIV-1 diversification over time was explored by randomly selecting 139 protease sequences from two sets of samples isolated 15 years apart. The "early" proteases (n = 89) were obtained from samples isolated in 1993 and 1994 from naïve patients (mean viral load ± SE = 123183 ± 23813 copies/ml, mean CD4+ cell count ± SE = 352 ± 26 cells/µl). The "late" proteases (n = 50) were obtained from samples isolated in 2006 and 2007 from recently infected (within the previous year) naïve patients (mean viral load ± SE = 335741 ± 98540 copies/ml, mean CD4+ cell count ± SE = 380 ± 30 cells/µl). Compared to the subtype B protease ancestral sequence (http://www.hiv.lanl.gov), the mean diversification values for the circulating nucleotide sequences increased significantly from 0.025 ± 0.001 in 1993–1994 (p-distance ± SE), to 0.041 ± 0.002 in 2006–2007 ($p < 0.0001$, unpaired t test) (Fig 20, A). Amino acid sequence diversification also significantly increased from 0.037 ± 0.002 in 1993–1994 to 0.053 ± 0.003 in 2006–2007 ($p < 0.0001$) (Fig 20, B). Similar results were obtained when sequences were compared to the subtype consensus sequence (data not shown). Genetic distances not only increased significantly compared to an ancestral or consensus sequence but also increased as determined by intra-group p-distance diversity. There was an increase from 0.032 ±0.004 to 0.062 ± 0.006, respectively, for nucleotide distances ($p < 0.0001$), and from 0.049 ±0.012 to 0.077 ± 0.016, respectively, for amino acid distances ($p < 0.0001$).



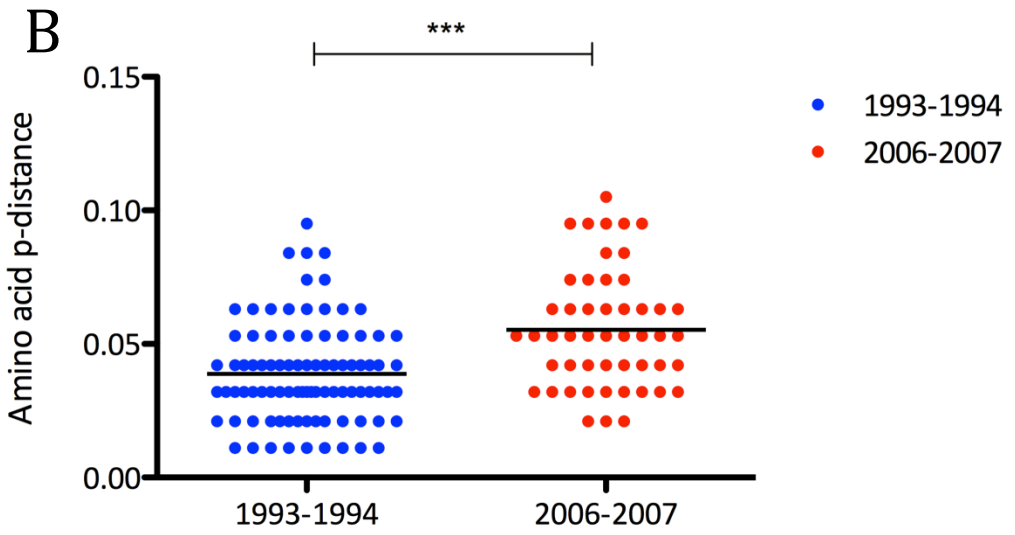


Figure 20: (A) Nucleotide and (B) amino acid p-distances to the HIV-1 protease subtype B ancestral sequence (http://www.hiv.lanl.gov/). The solid horizontal line indicates the mean p-distance. The mean diversification values for circulating sequences increased significantly from 1993–1994 to 2006–2007 ($p < 0.0001$, unpaired t test).

The amino acid entropy was significantly higher ($p \leq 0.05$) in the late proteases at residues 36, 41, 60, 61, 63, 65, 67, and 72. A higher synonymous-to-non-synonymous (dS/dN) ratio was found in late sequences (5.0 ± 0.4 and 4.4 ± 0.4, respectively); however, this did not reach statistical significance ($p = 0.2886$). The study high dS/dN ratio implies that the protease gene was mainly under purifying selection. Positive selection pressures were also evaluated using FEL, which was used to directly estimate synonymous and non-synonymous substitution rates at each site (Pond & Frost, 2005). We identified a total of 7 positively selected codons that are frequent polymorphisms in HIV-1 subtype B proteases: three positively selected codons (T12, E35, and L63) were identified in the early samples, and four codons (L33, N37, V77, and I93) were identified in the late samples.

The evolutionary history of the protease sequences from early and late isolates was inferred by using the Maximum Likelihood method based on the Data specific model (Nei & Kumar, 2000). The bootstrap consensus tree inferred from 500 replicates (Felsenstein, 1985) is taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985) (Fig 21). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. Initial tree(s) for the heuristic search were obtained automatically as follows. When the number of common sites was < 100 or less than one fourth of the total number of sites, the

maximum parsimony method was used; otherwise BIONJ method with MCL distance matrix was used. A discrete Gamma distribution was used to model evolutionary rate differences among sites [5 categories (+G, parameter = 0.2903)]. The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 47.6262% sites). The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 140 nucleotide sequences: 89 sequences from the early isolates, 50 form the late isolates and the subtype B ancestral sequence. All positions containing gaps and missing data were deleted. There were a total of 296 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 (Tamura *et al.*, 2011).

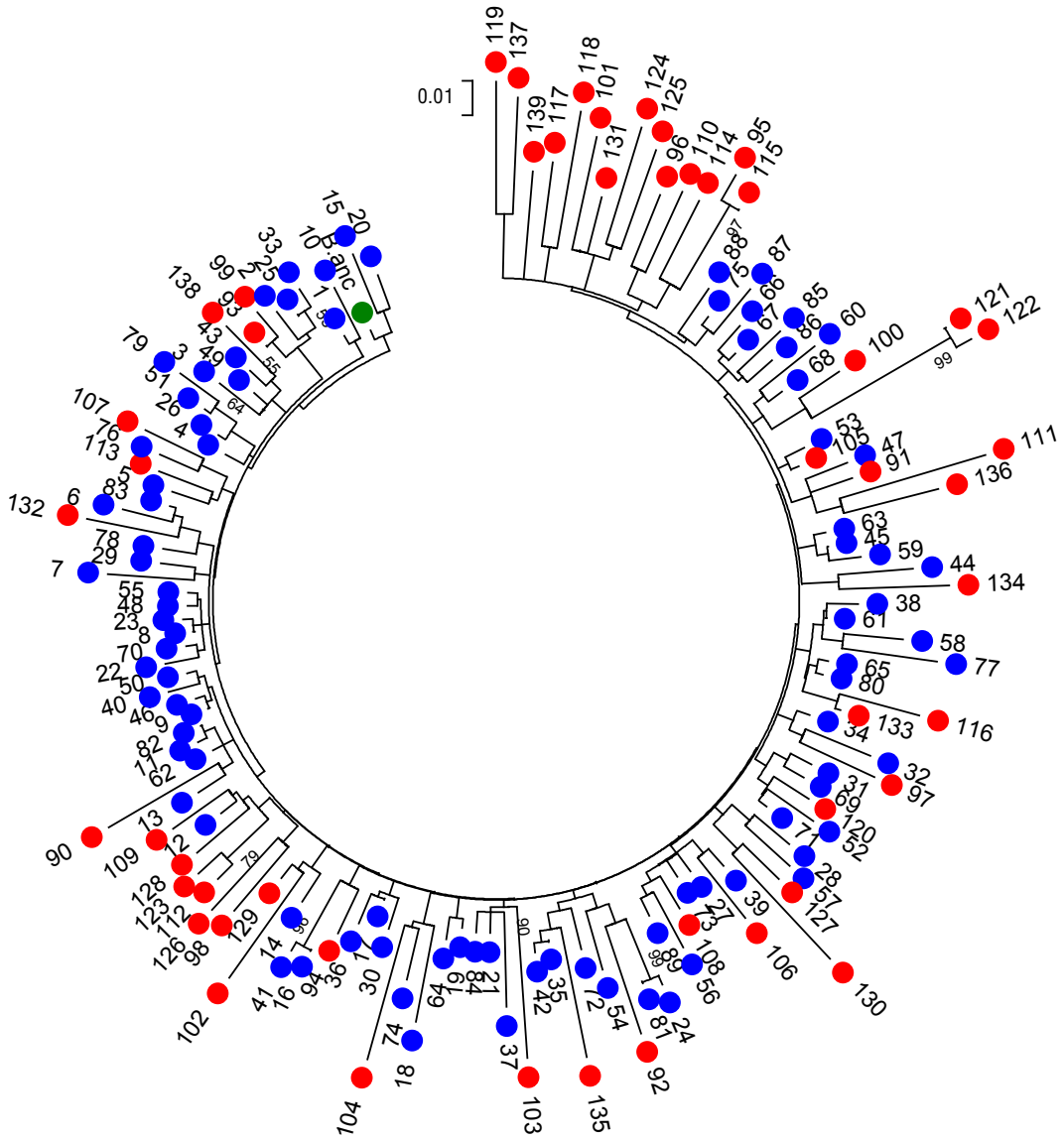


Figure 21: Phylogenetic tree of the protease sequences from patients infected in 1993-1994 and 2006-2007. The tree was drawn by the General Time Reversible method using the protease nucleotide sequences. Blue and red symbols indicate viruses isolated in 1993–1994

and in 2006–2007, respectively; the ancestral subtype B sequence is shown in green. The scale of the genetic distance is indicated next to a branch on the top of the figure.

Decrease in HIV-1 gag sequence conservation over time

Full-length gag sequences from early (n = 40) samples also had significantly lower nucleotide diversification values than those from late samples (n = 41) (0.033 ± 0.001 in 1993–1994 versus 0.050 ± 0.001 in 2006–2007, $p < 0.0001$, unpaired t test)(Fig 22, A). Similarly, amino acid sequence diversification significantly increased from 0.045 ± 0.001 in 1993–1994 to 0.061 ± 0.002 in 2006–2007 ($p < 0.0001$)(Fig 22, B).

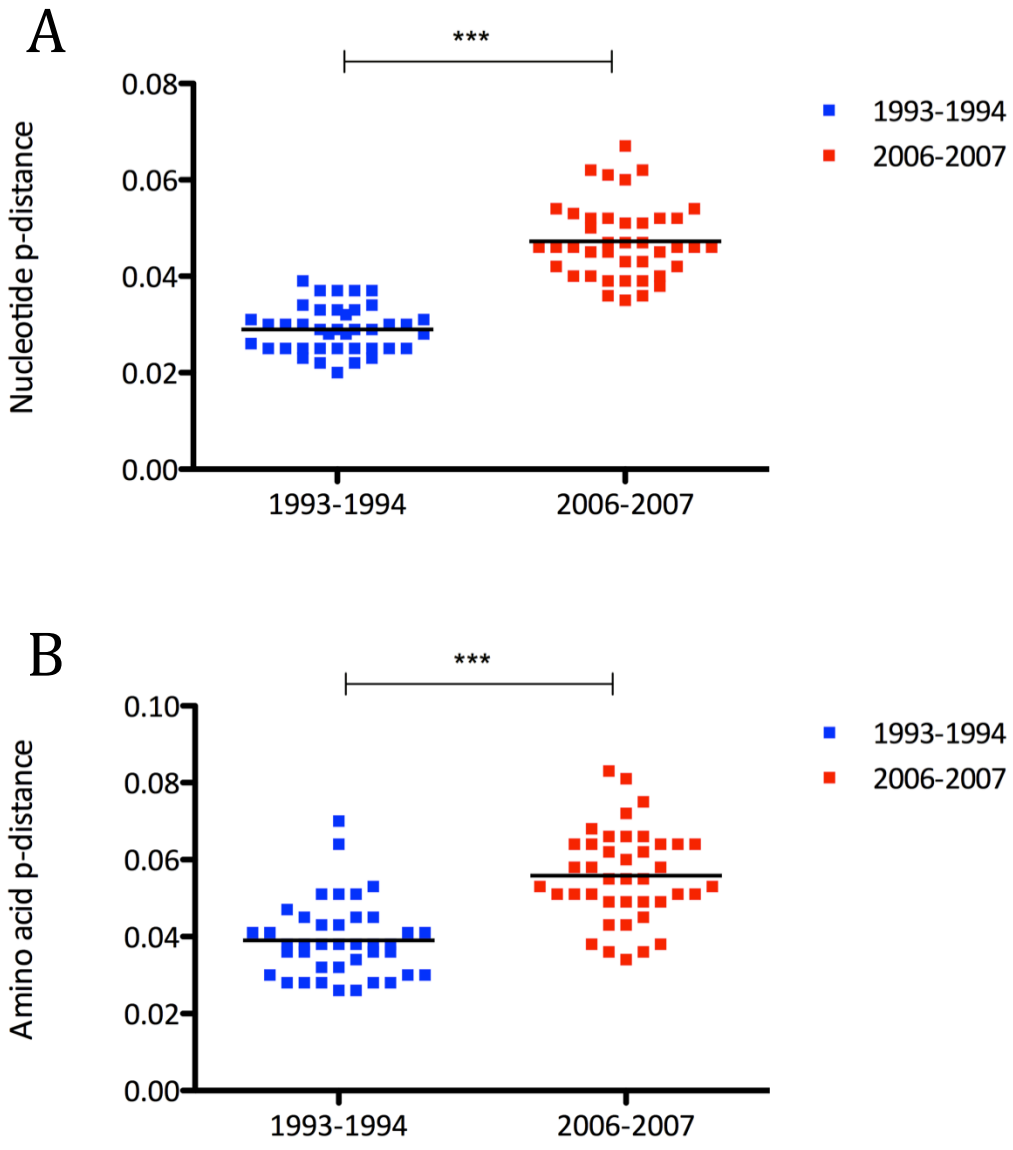


Figure 22: (A) Nucleotide and (B) amino acid p-distances to the HIV-1 gag subtype B ancestral sequence (http://www.hiv.lanl.gov/). The solid horizontal line indicates the mean p-distance. The mean diversification values for circulating sequences increased significantly from 1993–1994 to 2006–2007 ($p < 0.0001$, unpaired t test).

The amino acid entropy was significantly higher ($p \leq 0.05$) in the recent integrases at residues 55, 90, 94, 112, 126, 141, 143, 169, 170, 191, 230, 251, 271, 278, 291, 309, 325, 398, 405, 408, 410, 412, 414, 422, 451, 462, 507, 513, 520 and 533. A higher synonymous-to-non-synonymous (dS/dN) ratio was found in late sequences (5.2 ± 0.03 and 3.9 ± 0.03, respectively); this result reached statistical significance ($p < 0.001$). The high dS/dN ratios imply that the *gag* gene was mainly under purifying selection.

The evolutionary history of the gag sequences of the study patients was inferred by using the Maximum Likelihood method based on the Data specific model (Nei & Kumar, 2000). The bootstrap consensus tree inferred from 500 replicates (Felsenstein, 1985) is taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985) (Fig 23). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. Initial tree(s) for the heuristic search were obtained automatically as follows. When the number of common sites was < 100 or less than one fourth of the total number of sites, the maximum parsimony method was used; otherwise BIONJ method with MCL distance matrix was used. A discrete Gamma distribution was used to model evolutionary rate differences among sites [5 categories (+G, parameter = 0.2988)]. The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 41.1768% sites). The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 82 nucleotide sequences, 40 from the patients isolated earlier, 41 from the recent isolates and the B ancestral sequence. All positions containing gaps and missing data were eliminated. There were a total of 1413 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 (Tamura *et al.*, 2011).

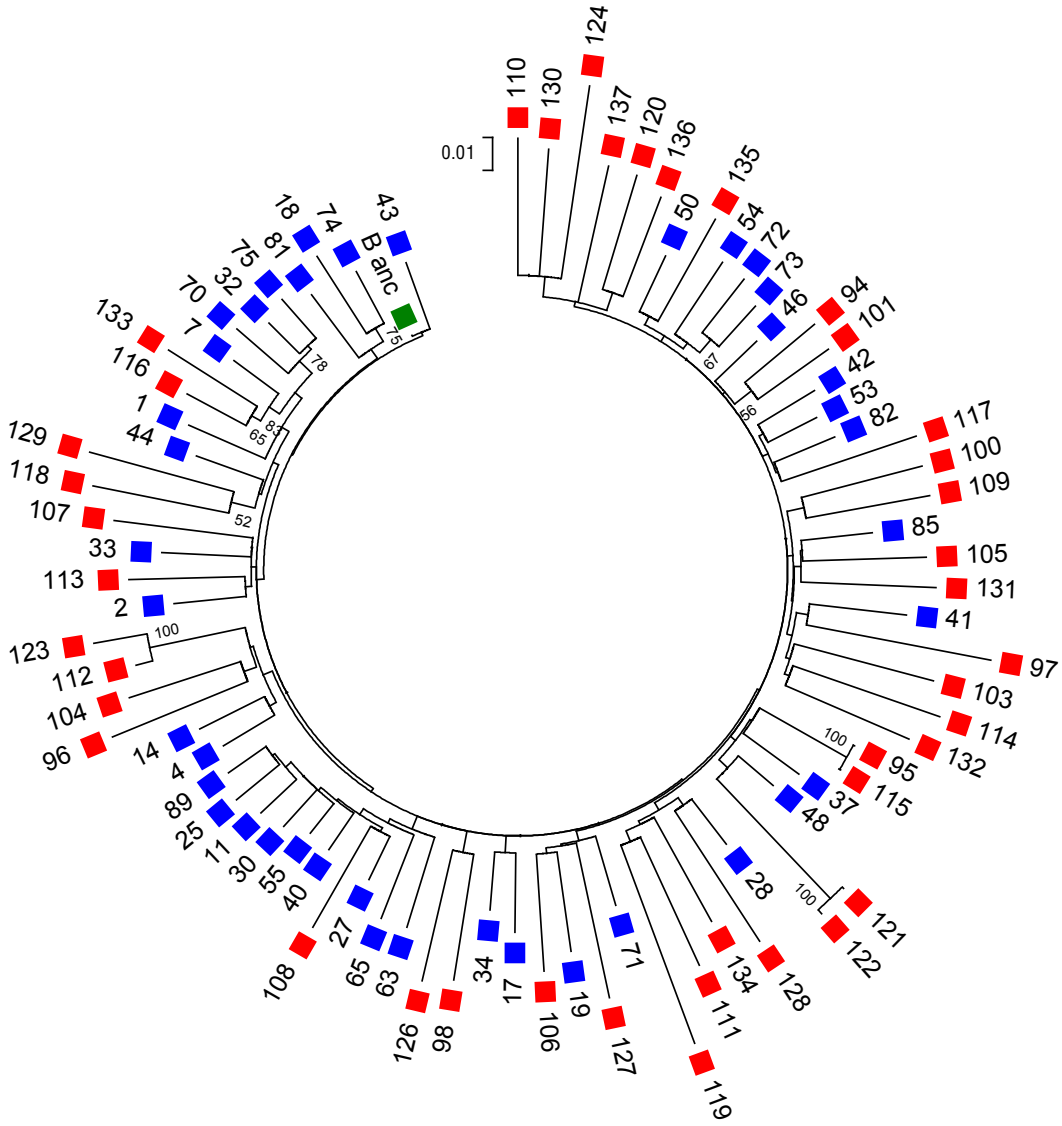


Figure 23: Phylogenetic tree of the gag sequences from patients infected in 1993-1994 and 2006-2007. The tree was drawn by the General Time Reversible method using the gag nucleotide sequences. Blue and red symbols indicate viruses isolated in 1993–1994 and in 2006–2007, respectively; the ancestral B sequence is shown in green. The scale of the genetic distance is indicated next to a branch on the top of the figure.

As expected, protease nucleotide diversification was higher than gag nucleotide diversification ($r = 0.4303$, $p < 0.0001$, linear regression) (Fig 24, A), as well as amino acid diversification ($r = 0.1760$, $p < 0.0001$, linear regression) (Fig 24, B). The relationship between these two genes diversification was confirmed by the Spearman's rank correlation coefficient in both nucleotide ($r = 0.6517$, $p < 0.0001$) and amino acid p-distances data ($r = 0.3845$, $p < 0.0001$).

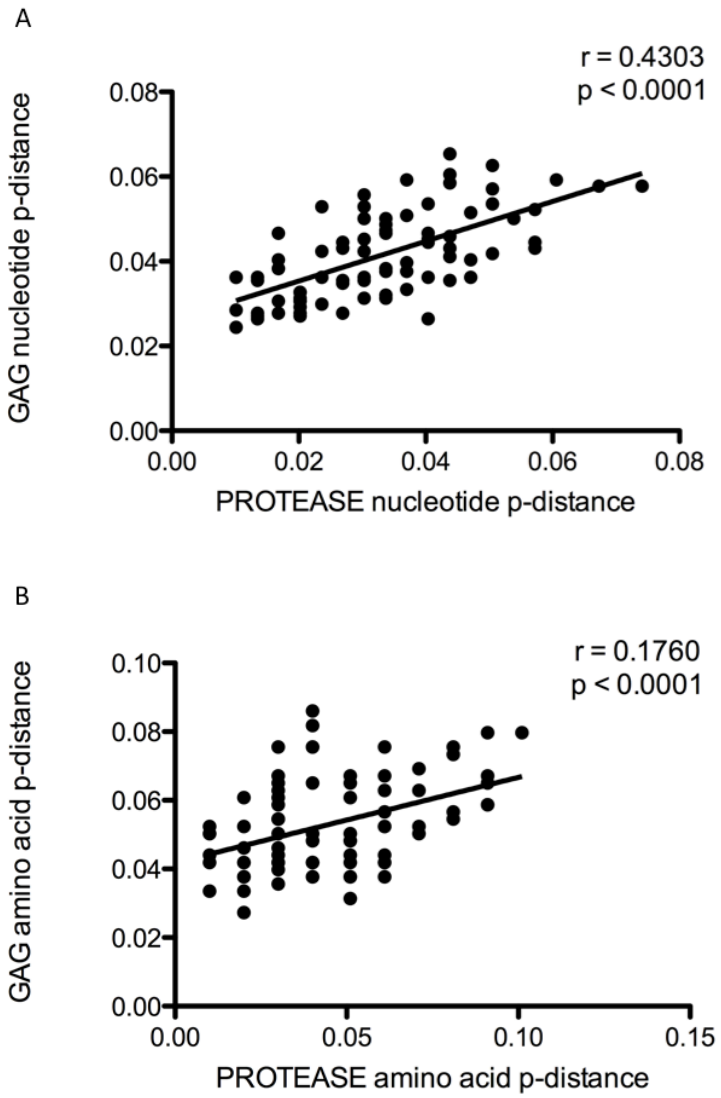


Figure 24: Comparison of the (A) Nucleotide and (B) amino acid p-distances from the HIV-1 gag and protease sequences. The solid line indicates the linear regression (r). Correlation between both circulating sequences from 1993–1994 to 2006–2007 was significant ($p < 0.0001$, unpaired t test).

Decrease in HIV-1 integrase sequence conservation over time

To study the HIV-1 integrase diversification, 47 integrase sequences were randomly selected from the "early" samples and 47 from the "late" samples. Compared to the subtype B integrase ancestral sequence (http://www.hiv.lanl.gov), the mean diversification values for the circulating nucleotide sequences increased significantly from 0.025 ± 0.001 in 1993–1994 (p-distance ± SE), to 0.034 ± 0.001 in 2006–2007 ($p < 0.0001$, unpaired t test) (Fig 25, A). Amino acid sequence diversification also increased significantly from 0.028 ± 0.001 in 1993–1994 to 0.034 ± 0.002 in 2006–2007 ($p = 0.0051$) (Fig 25, B). Similar results were obtained when sequences were compared to the subtype consensus sequence (data not shown). Genetic distances

not only increased significantly compared to an ancestral or consensus sequence but also increased as determined by intra-group p-distance diversity. There was an increase from 0.032 ± 0.002 to 0.050 ± 0.003, respectively, for nucleotide distances ($p < 0.0001$), and from 0.038 ±0.006 to 0.049 ± 0.007, respectively, for amino acid distances ($p = 0.2359$).

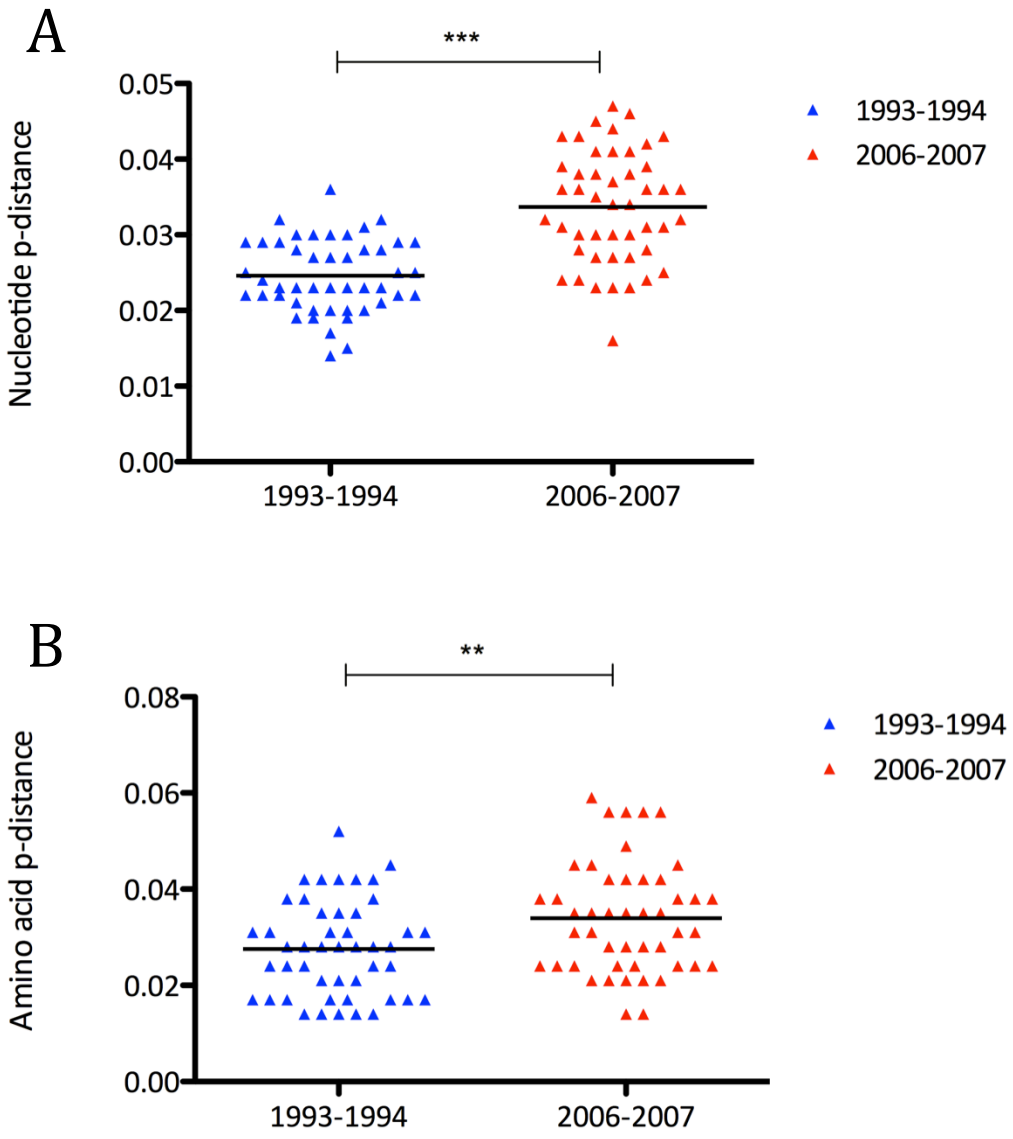


Figure 25: (A) Nucleotide and (B) amino acid p-distances to the HIV-1 integrase subtype B ancestral sequence (http://www.hiv.lanl.gov/). The solid horizontal line indicates the mean p-distance. The mean diversification values for circulating sequences increased significantly from 1993–1994 to 2006–2007 ($p < 0.0001$ for nucleotide p-distance and $p = 0.0051$ for amino acid p-distance, unpaired t test).

The amino acid entropy was significantly higher ($p \leq 0.05$) in the late integrases at residues 27, 45, 50, 203, and 265. A higher synonymous-to-non-synonymous (dS/dN)

ratio was found in late sequences (8.1 ± 0.1 and 6.2 ± 0.1, respectively) ($p < 0.001$). The high dS/dN ratios imply that the integrase gene was mainly under purifying selection. Positive selection pressures were also evaluated using FEL, which was used to directly estimate synonymous and non-synonymous substitution rates at each site (Pond & Frost, 2005). We identified a total of 7 positively selected codons that are frequent polymorphisms in HIV-1 subtype B integrases: six positively selected codons (L28, I72, T122, T124, T206 and A265) were identified in the early samples, and six codons (T112, S119, V201, T206, S230 and L234) were identified in the late samples.

The evolutionary history of the integrase sequences from the study patients was inferred by using the Maximum Likelihood method based on the Data specific model (Nei & Kumar, 2000). The bootstrap consensus tree, inferred from 500 replicates (Felsenstein, 1985), was taken to represent the evolutionary history of the taxa analyzed (Felsenstein, 1985) (Fig 26). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches (Felsenstein, 1985). Initial tree(s) for the heuristic search were obtained automatically as follows. When the number of common sites was < 100 or less than one fourth of the total number of sites, the maximum parsimony method was used; otherwise BIONJ method with MCL distance matrix was used. A discrete Gamma distribution was used to model evolutionary rate differences among sites [5 categories (+G, parameter = 0.2336)]. The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 49.9764% sites). The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 95 nucleotide sequences: 47 sequences from the early isolates, 47 from late isolates, and the subtype B ancestral sequence. All positions containing gaps and missing data were eliminated. There were a total of 864 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 (Tamura *et al.*, 2011).

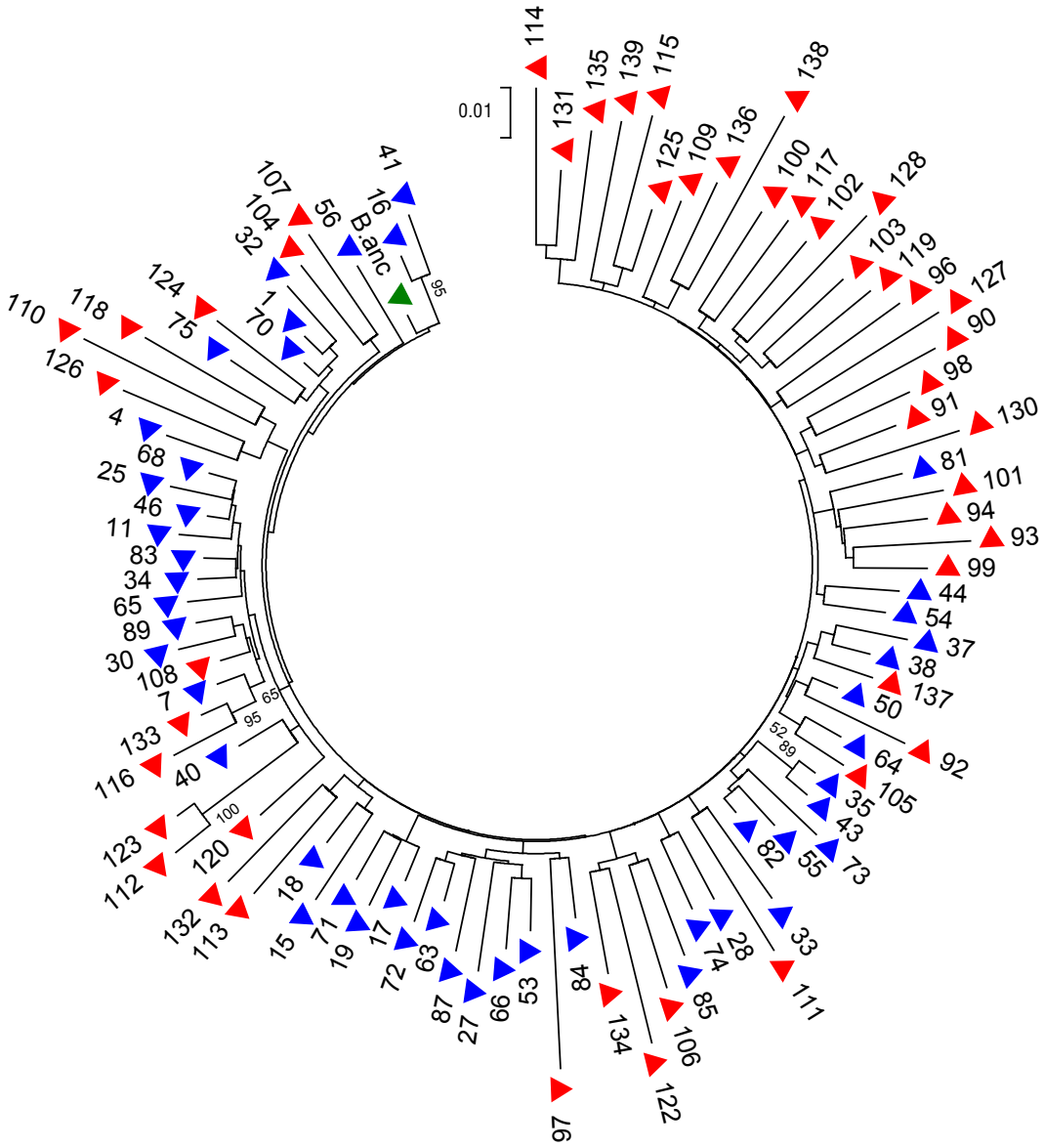


Figure 26: Phylogenetic tree of the integrase sequences from patients infected in 1993-1994 and 2006-2007. The tree was drawn by the General Time Reversible method using the integrase nucleotide sequences. Blue and red symbols indicate viruses isolated in 1993–1994 and in 2006–2007, respectively; the ancestral B sequence is shown in green. The scale of the genetic distance is indicated next to a branch on the top of the figure.

As expected, protease diversification was highly correlated with integrase nucleotide diversification ($r = 0.3705$, $p < 0.0001$, linear regression) (Fig 27, A). However this correlation was lower for amino acid diversification ($r = 0.0730$, $p = 0.0084$, linear regression) (Fig 27, B). The relationship between these two genes diversification was confirmed by the Spearman's rank correlation coefficient in both nucleotide ($r = 0.5971$, $p < 0.0001$) and amino acid p-distances data ($r = 0.2630$, $p = 0.0104$).

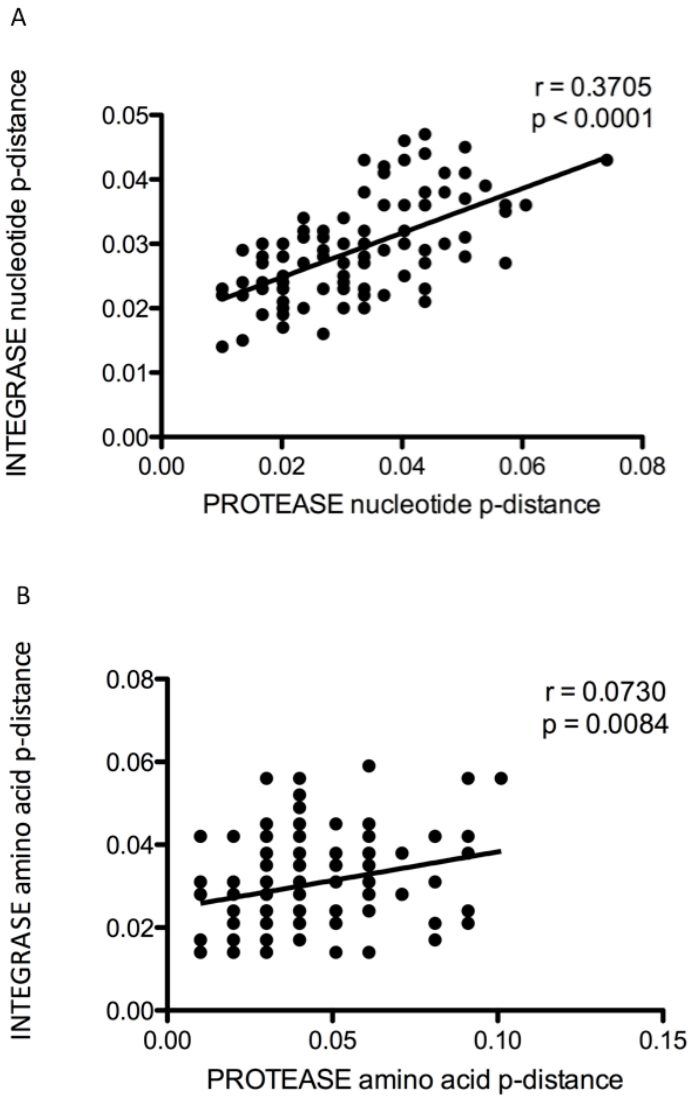


Figure 27: Comparison of the (A) Nucleotide and (B) amino acid p-distances from the HIV-1 protease and integrase sequences. The solid line indicates the linear regression (r). Correlation between both circulating sequences from 1993–1994 to 2006–2007 was significant ($p < 0.0001$ and $p = 0.0084$, unpaired t test).

Integrase diversification was also highly correlated with gag nucleotide diversification ($r = 0.3605$, $p < 0.0001$, linear regression) (Fig 28, A). However this correlation remained lower for amino acid diversification ($r = 0.0657$, $p = 0.0275$, linear regression) (Fig 28, B). Again, the relationship between these two genes diversification was confirmed by the Spearman's rank correlation coefficient in both nucleotide ($r = 0.5589$, $p < 0.0001$) and amino acid p-distances data ($r = 0.2360$, $p = 0.0429$).

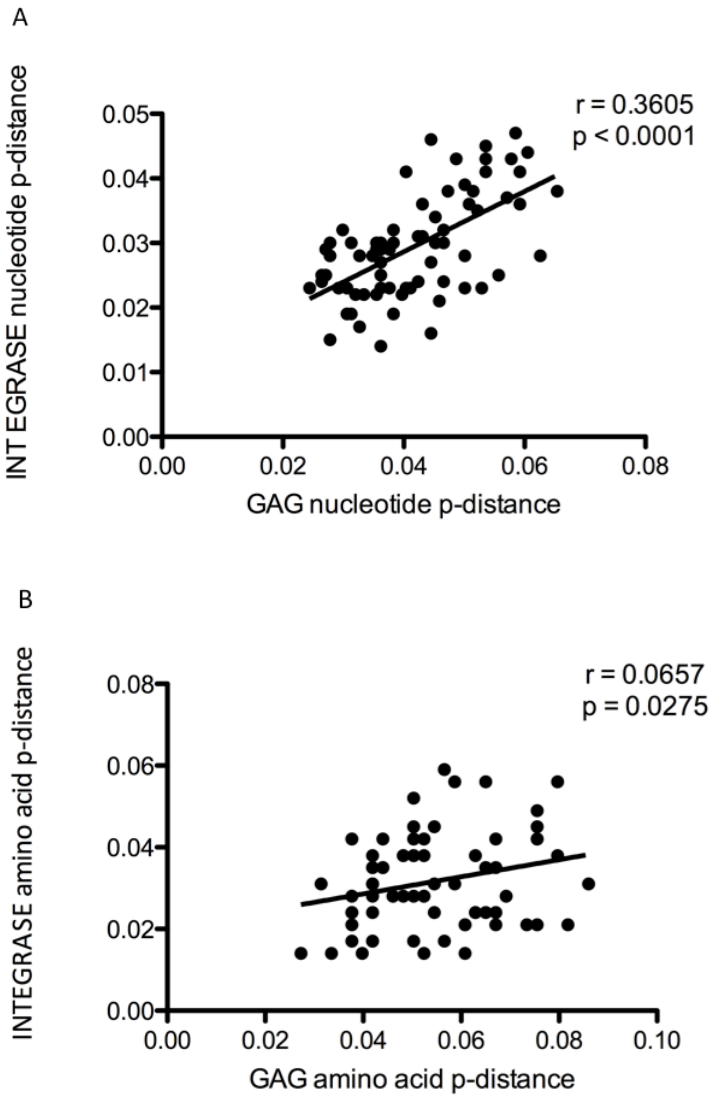


Figure 28: Comparison of the (A) Nucleotide and (B) amino acid p-distances from the HIV-1 integrase and gag sequences. The solid line indicates the linear regression (r). Correlation between both circulating sequences from 1993–1994 to 2006–2007 was significant ($p < 0.0001$ and $p = 0.0275$, unpaired t test).

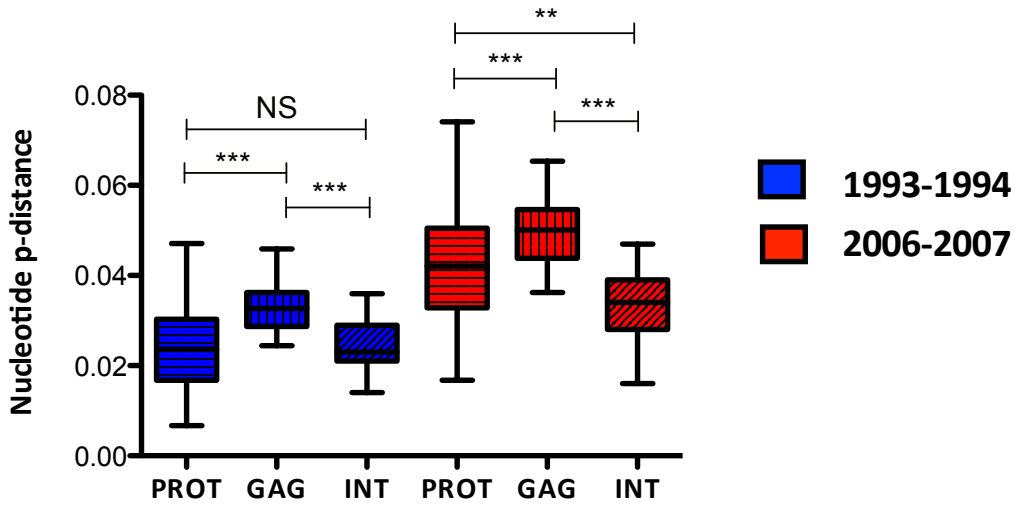
Whereas protease and gag genes had a high sequence conservation reduction, this reduction was lower in the integrase gene (Fig 29; Table 6). The mean ± SD increase of nucleotide sequence diversification of integrases, 0.009 ± 0.002, was significantly lower to the mean increase of nucleotide sequence diversification of proteases and gags (0.017 ± 0.003 and 0.016 ± 0.003 respectively; $p < 0.0001$; $p < 0.0001$), indicating that integrases were more conserved. Moreover, the mean ± SE increase of amino acid sequence diversification of integrases, 0.006 ± 0.002, was significantly lower to the mean increase of amino acid sequence diversification of proteases and gags (0.016 ± 0.003 and 0.016 ± 0.002 respectively; $p < 0.0001$; $p < 0.0001$), confirming that integrases were more conserved. Although the mean increase of

nucleotide sequence diversification of proteases was also significantly higher than the one of gag sequences (p = 0.0164), the mean increase of amino acid sequence diversification of proteases was not significantly higher than the one of gag sequences (p = 1.000).

Table 6. Comparison of the HIV-1 protease, gag and integrase sequence diversification between 1993-1994 and 2006-2007. Mean ± SE p-distance values to the relative subtype B sequence and the difference between the two groups (Δ) are indicated.

	Nucleotide			Amino acid		
	1993-1994	2006-2007	Δ	1993-1994	2006-2007	Δ
PROT	0.025 ± 0.001	0.041 ± 0.002	0.017 ± 0.003	0.037 ± 0.002	0.053 ± 0.003	0.016 ± 0.003
GAG	0.033 ± 0.001	0.050 ± 0.001	0.016 ± 0.003	0.045 ± 0.001	0.061 ± 0.002	0.016 ± 0.002
INT	0.025 ± 0.001	0.034 ± 0.001	0.009 ± 0.002	0.028 ± 0.001	0.034 ± 0.002	0.006 ± 0.002

A



B

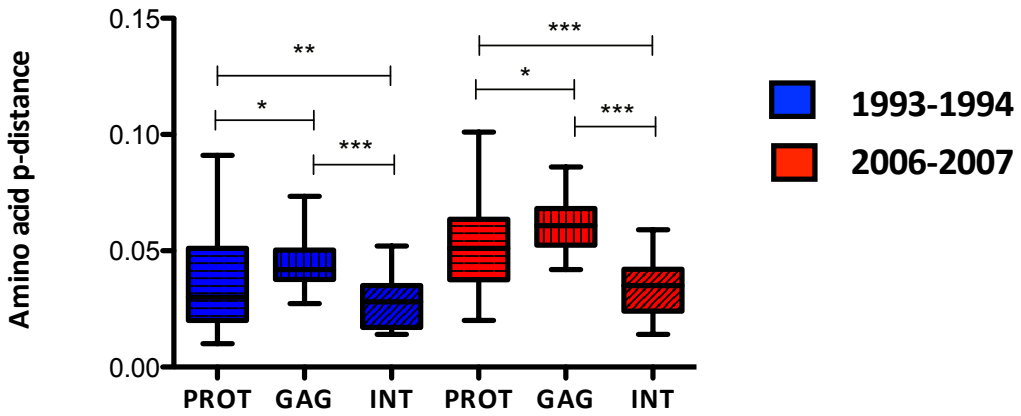


Figure 29: Comparison of the (A) Nucleotide and (B) amino acid p-distances to the HIV-1 subtype B ancestral sequences (http://www.hiv.lanl.gov/) of the HIV-1 protease, gag and integrase genes. The solid horizontal line indicates the mean p-distance.

Comparison of the protease, gag and integrase genes nucleotide composition

HIV-1 base composition of the 3 different genes over time was analysed by randomly selecting 139 protease sequences, 81 gag sequences and 94 integrase sequences from the two sets of samples isolated 15 years apart (Table 7). All the 3 genes were particularly rich in Adenosine (A) nucleotide content (37%) while poor in Cytosine (C) nucleotide content (17%). Thus fulfilling the general property of the lentivirus family to which HIV belongs, that describes their A-nucleotide richness and C-nucleotide poorness (Berkhout & van Hemert, 1994; Bronson & Anderson, 1994; van der Kuyl & Berkhout, 2012; van Hemert & Berkhout, 1995). As expected, the gag gene maintained its nucleotide composition constant over time (Table 7; Fig 30, B). Nevertheless, the protease sequences from late isolates had a significantly higher proportion of G-nucleotide (22.7% in 1993-1994 versus 22.9% in 2006-2007, $p = 0.0474$) and a significantly lower proportion of U-nucleotide content (25.1% in 1993-1994 versus 24.9% in 2006-2007, $p = 0.0250$) compared to early isolates (Table 7; Fig 30, A). The integrase sequences from recent isolates displayed a significantly higher amount of C-nucleotides (14.9% in 1993-1994 versus 15.0% in 2006-2007, $p = 0.0234$) and, like the recent protease sequences, a significantly lower amount of U-nucleotides (21.5% in 1993-1994 versus 21.4% in 2006-2007, $p = 0.0320$) compared to early isolates (Table 7; Fig 30, C).

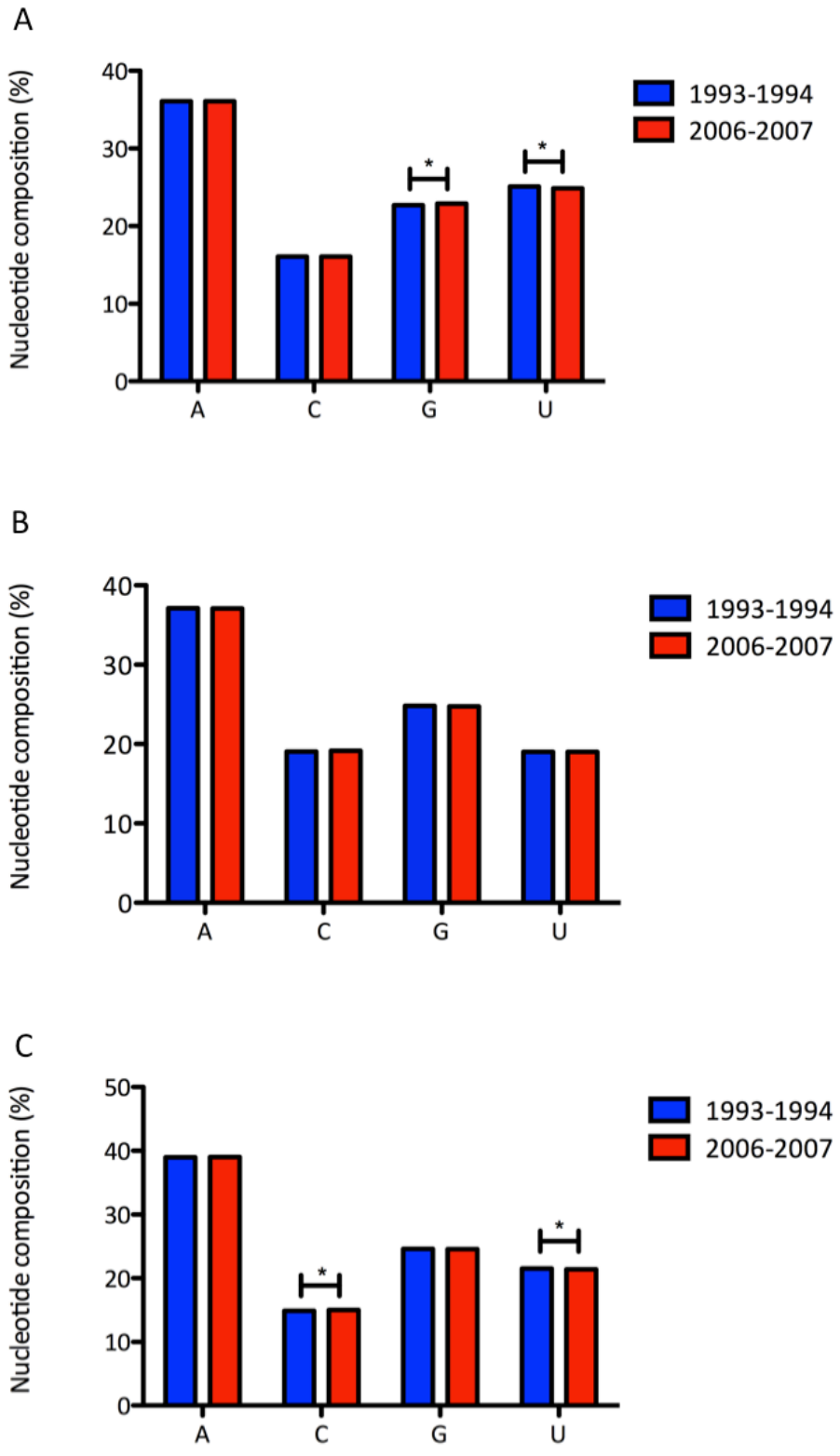


Figure 30: Comparison of the HIV-1 (A) protease, (B) gag and (C) integrase nucleotide composition. (*: p <0.05, unpaired t-test).

Table 7. Nucleotide composition of HIV-1 RNA genome segments over time (1993-2007)

HIV-1 genome segment	Length (nt)	Number of genomes analysed	Average nucleotide composition (%) 1993-1994	Range	Number of genomes analysed	Average nucleotide composition (%) 2006-2007	Range	p-value
Gag	1502	40	A 37.1	36.6-38.3	41	A 37.1	36.3-37.7	0.7101
			C 19.0	18.5-19.5		C 19.2	18.3-19.9	0.1147
			G 24.8	23.9-25.5		G 24.7	24.1-25.4	0.4206
			U 19.0	18.5-19.4		T 24.9	18.5-19.6	0.7598
Protease	297	89	A 36.1	34.7-37.4	50	A 36.1	34.0-38.0	0.9956
			C 16.1	14.8-17.5		C 16.1	14.5-17.8	0.9939
			G 22.7	21.5-23.9		G 22.9	21.9-25.3	**0.0474**
			U 25.1	24.2-26.3		T 24.9	23.2-26.6	**0.0250**
Integrase	864	47	A 39.0	37.7-39.8	47	A 39.0	38.0-39.0	0.6055
			C 14.9	14.5-15.5		C 15.0	14.2-15.9	**0.0234**
			G 24.6	24.1-25.3		G 24.6	23.3-25.3	0.4461
			U 21.5	20.8-22.0		T 21.4	20.7-22.2	**0.0320**

Bold letters indicate p-value < 0.05.

Natural variability of the HIV-1 protease, gag and integrase genes

Analysis of all the amino acid sequences from the 139 naïve patients revealed that the protease of HIV-1 subtype B is a conserved protein with 72% (71/99) of amino acid residues achieving ≥97% conservation and 65% (64/99) completely conserved (Table 8). As expected, when comparing the early versus late sequences, we found that there was a higher amount of early sequences carrying residues fully conserved, 71% (70/99) compared to 45% (45/99) in late protease sequences. The catalytic triad (Asp 25, Thr 26 and Gly 27) was 100% conserved and the regions described to be essential for the enzyme functionality [amino terminal (1-9), catalytic site (21-32), flap [44-56], substrate union site (78-88), and carboxyl terminal (94-99)] were ≥99% conserved. No deletions or insertions were seen in the protease sequence.

Analysis of all the amino acid sequences from the 81 naïve patients revealed that the gag of HIV-1 subtype B is a conserved protein with 84% (458/545) of amino acid residues achieving ≥97% conservation and 53% (287/545) completely conserved (Table 8). Although gag precursor is composed of 500 amino acids, our samples carried approximately 45 insertions. Therefore conservation of the different residues has been calculated using 545 residues. Comparison of the fully conserved amino acids in both group of patients revealed again that the gag sequences from early patients were more conserved, 62% (337/545), than the ones from late patients, 55% (302/545).

Analysis of all the amino acid sequences from 97 naïve patients revealed that the integrase of HIV-1 subtype B is a conserved protein with 91% (263/288) of amino acid residues achieving ≥97% conservation and 66% (189/288) completely conserved (Table 8). The HHCC zinc-binding site, the DDE catalytic triad and the LEDGF/p75 IN-binding domain between residues 128–132 and 161–173 were all ≥97% conserved. No deletions or insertions were seen in the integrase sequence. Taken altogether, integrase appeared to be the more conserved protein of the three HIV-1 subtype B analysed proteins.

Table 8. Conservation of the HIV-1 genes at the amino acid level.

			100% conserved aa		≥97% conserved aa	
	# total aa	Patients	#	%	#	%
Protease	99	All	64	65	71	72
		1993/1994	70	71		
		2006/2007	45	45		
Gag	545	All	287	53	458	84
		1993/1994	337	62		
		2006/2007	302	55		
Integrase	288	All	189	66	263	91
		1993/1994	217	75		
		2006/2007	203	70		

#: number; aa: amino acid.

Evolution of the human immunodeficiency virus type 1 protease: Effects on viral replication capacity and protease robustness

Relationship between HIV-1 protease sequence conservation and viral RC

To investigate whether HIV-1 protease sequence diversification affected viral RC, 22 proteases were non-randomly selected from the set of 139 proteases. These 22 proteases were selected on the basis of their high or low divergence from the ancestral subtype B protease amino acid sequence (Fig 31, A and B). Eleven proteases were chosen from the 1993–1994 samples (mean viral load ± SE = 158431 ± 78847 copies/ml, mean CD4+ cell count ± SE = 335 ± 56 cells/µl), and the other 11 were selected from the 2006–2007 samples (mean viral load ± SE = 398345 ± 253620 copies/ml, mean CD4+ cell count ± SE = 394 ± 35 cells/µl). Bulk patient plasma-derived PCR protease amplicons were recombined with a protease-deleted HXB2 infectious clone. In addition, 11 proteases carrying PI resistance mutations (mean viral load ± SE = 40364 ± 12201 copies/ml, mean CD4+ cell count ± SE = 149 ± 41 cells/µl) that were isolated in 2009–2011 were also PCR amplified, recombined with the HXB2 infectious clone, and included in the analysis (Fig 31, C). The protease region of the chimeric viruses was then sequenced from the resultant chimeric viral stocks and compared to the original source plasma (Fig 32). A mean ± SD of 0.30 ± 0.34 % and 0.34 ± 0.79 % difference at the nucleotide and amino acid level, respectively, was observed between plasma HIV-1 RNA and chimeric virus. These data indicated that although there may have been some selection of chimeric viruses during recombination and/or culture, protease chimeric viruses represented the

dominant form of plasma viruses and, therefore, selection during culture was unlikely to affect the overall analysis.

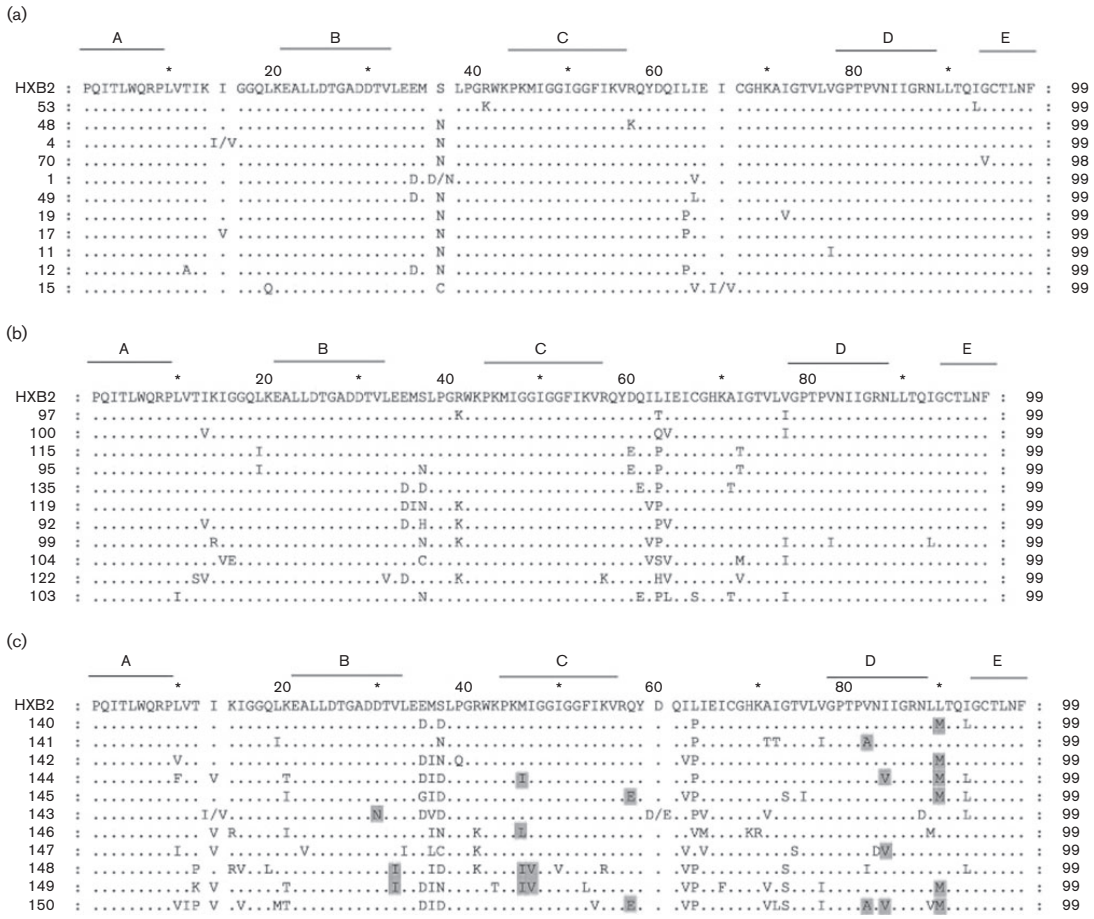


Figure 31: Alignment of the amino acid sequences of the 33 HIV-1 proteases that were recombined with an HXB2 protease-deleted infectious clone. The sequences shown were from viruses recovered 7 days after transfection. (A) Naïve samples isolated in 1993–1994. (B) Naïve samples isolated in 2006–2007. (C) Protease inhibitor (PI) resistant samples isolated in 2009–2011. The sequences shown were from infectious viruses recovered after transfection. The indicated amino acid changes were relative to a subtype B ancestral sequence. Dots indicate amino acid sequence identity. Residues in which a mixture of nucleotides was observed are also indicated. A grey background is used to indicate major PI resistance mutations. The black lines above each sequence alignment indicate the conserved protease sequence regions: A, amino terminal; B, catalytic site; C, flap; D, substrate union site; E, carboxyl terminal (Capel *et al.*, 2012).

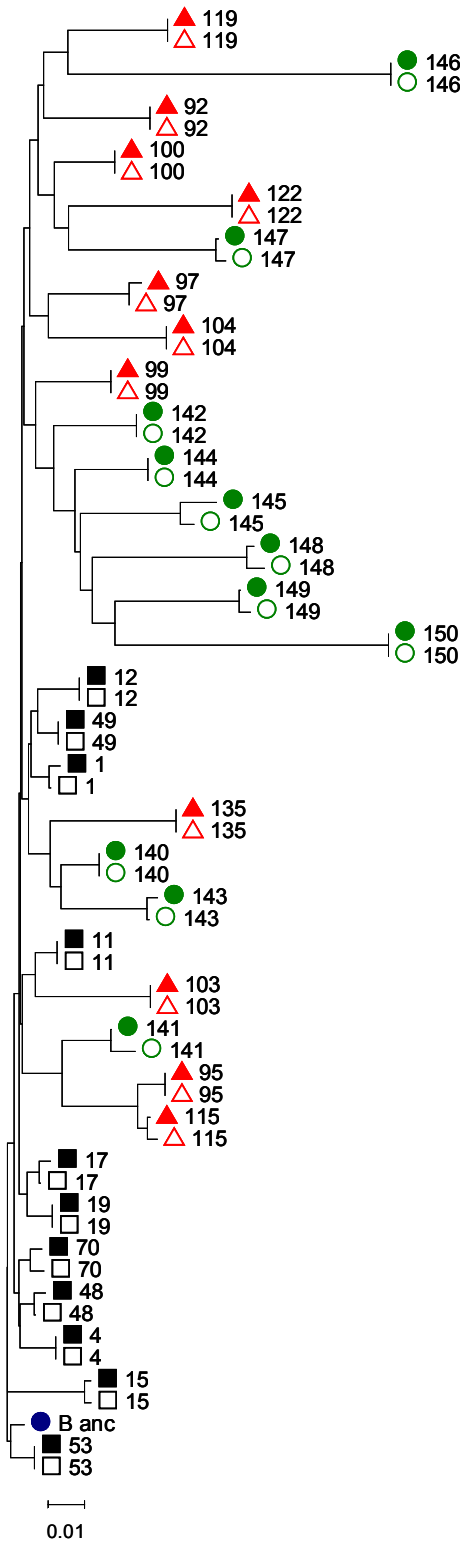
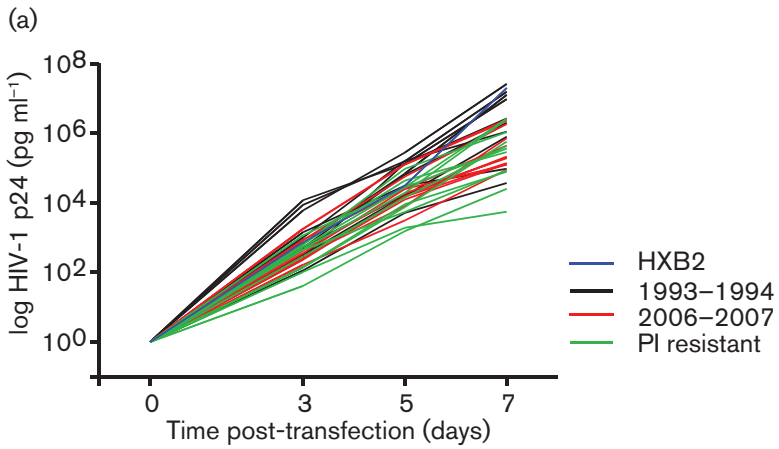


Figure 32: Validation of obtained chimeric HXB2 sequences by phylogenetic analysis. The tree was drawn by the neighbour-joining method using the protease nucleotide sequences. Empty taxa symbols indicate sequences from chimeric viral stocks. Fill taxa symbols indicate the original plasma viral sequence. Black, red, and green symbols indicate viruses isolated in 1993–1994, in 2006–2007, and PI resistant proteases, respectively; the ancestral B sequence is shown in blue. Note that all viral stocks and plasma viral sequences from the same subjects shared the same branches. Modified from (Capel *et al.*, 2012).

To quantitatively compare the ex vivo RC of the resulting 33 chimeric viruses, MT-4 cells were co-transfected and viral growth was measured by quantifying the viral p24 antigen at days 0 to 7 after transfection (Fig 33, A). The RC of the bulk (quasispecies-containing) recombinant viruses was assayed in three independent experiments, and the mean RC values were calculated. No correlation was found between the starting sample quasispecies diversity and ex vivo viral RC (data not shown). Chimeric viruses derived from naïve samples isolated in 1993–1994 showed ex vivo viral RC that was similar to the RC of viruses from naïve samples isolated in 2006–2007 (Fig 33, B). The mean ± SE RC of early (1993–1994) viruses was 98 ± 4%, and that of late (2006–2007) viruses was 96 ± 2% (p = 0.5729). Viral RC was also compared by determining how conserved the proteases sequences were relative to an ancestral subtype B sequence. There was no positive relationship between conservation of sequences and their relative RC ex vivo (r = 0.0322, p = 0.4241, linear regression), nor there was a positive correlation between viral RC and plasma viral load in naïve patients (r = 0.0238, p = 0. 3916, linear regression). Remarkably, some early (1993–1994) viruses were genetically much related to ancestral B sequence (Fig 32), suggesting that early ancestral viruses may have a RC similar to late viruses. As expected, chimeric viruses carrying PI resistance mutations showed significantly lower RC (Fig 33, A and B). The mean RC of the PI resistant proteases was 83 ± 3% (p = 0.0044 and p = 0.0021, for early and late proteases, respectively). As for naïve patients, there was no positive relationship between sequence conservation and relative ex vivo RC (r = 0.0019, p = 0.8978, linear regression). Taken together, these results demonstrate that HIV-1 protease diversification over time has not influenced ex vivo viral RC.



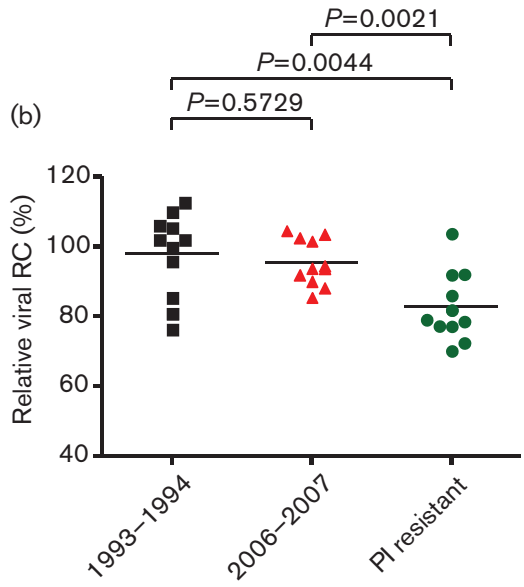


Figure 33: Comparison of ex vivo RC of the 33 chimeric protease viruses from naïve samples isolated in 1993–1994, from naïve samples isolated in 2006–2007, and from protease inhibitor (PI) resistant samples isolated in 2009–2011. (A) The kinetics of viral replication. Black, red, and green lines indicate viruses isolated in 1993–1994, in 2006–2007, and PI resistant proteases, respectively; a blue line indicates HXB2 RC. (B) Comparison of the slopes of the natural logs from days 0 to 7 after transfection. The RC of each variant was compared with that of the virus with the HXB2 recombinant protease (100%); the relative RC values are shown. The values represent the means of at least three independent experiments. P values were determined using the unpaired t test.

Relationship between HIV-1 protease robustness and viral RC

The mutational robustness of the 33 proteases analysed in this study was also investigated. To compare the robustness of these proteases, we used PCR mutagenesis to generate randomly mutated libraries of the different proteases. To quantify the mutant frequency generated by PCR mutagenesis, 132 individual protease clones from 6 out of 33 patient protease libraries were sequenced (Table 9). The mutagenic PCR used in this study had a mean mutation frequency per nucleotide of $1.72 \times 10^{-2} \pm 0.23$ (Table 9).

Table 9. Mean protease mutation frequencies after in vitro PCR mutagenesis and the ex vivo RC of the chimeric protease viruses.

Sample	Mutation frequency*	Ex vivo fitness relative to wild-type protease (%)
12	1.07×10^{-2}	87
15	2.37×10^{-2}	73
92	1.50×10^{-2}	78
104	1.16×10^{-2}	92
140	2.25×10^{-2}	64
142	1.96×10^{-2}	87

*The mean mutation frequency is the number of mutations in a set of 22 clones per sample divided by the length in base pairs of the target clone (i.e. 297 bp).

The 33 PCR mutated libraries were recombined with a protease-deleted HXB2 infectious clone. Similar to the amplified proteases in patient samples, chimeric viruses carrying mutated protease libraries were grown in MT-4 cells and their RC values were compared to those of the corresponding wild type recombinant viruses (Fig 34, A and B). All PCR mutated proteases displayed a significant reduction in virus-specific infectivity. Mutated proteases from each of the three groups of proteases of this study showed a significant reduction in viral growth ($p < 0.0001$, $p < 0.0001$, and $p < 0.0001$, for naïve samples isolated in 1993–1994, naïve samples isolated in 2006–2007, and samples carrying PI resistance mutations, respectively). These results indicate that in the absence of in vivo purifying selection the in vitro accumulation of more diversity caused significant viral RC losses. When randomly mutated proteases from naïve samples isolated in 1993–1994 (mean RC of 78 ± 3%) were compared with those from naïve samples isolated in 2006–2007 (mean RC of 79 ± 3%), there was no significant difference regarding viral infectivity reduction ($p = 0.8035$). This result demonstrates that HIV-1 protease diversification did not affect protease mutational robustness. In other words, the accumulation of amino acid changes over time did not influence the capacity of the protease to buffer additional deleterious mutations. When randomly mutated proteases carrying PI resistance mutations (mean RC of 58 ± 3%) were compared with randomly mutated proteases from naïve samples, a significant mean reduction in viral infectivity was observed ($p < 0.0001$ and $p < 0.0001$, for early and recent samples, respectively). Importantly, the mean ± SE reduction of viral infectivity of randomly mutated naïve proteases was 18 ± 1%, and the mean reduction of viral infectivity of randomly mutated proteases carrying PI resistance mutations was 25 ± 3% ($p = 0.0400$), indicating that proteases carrying PI resistance substitutions were less robust than naïve proteases.

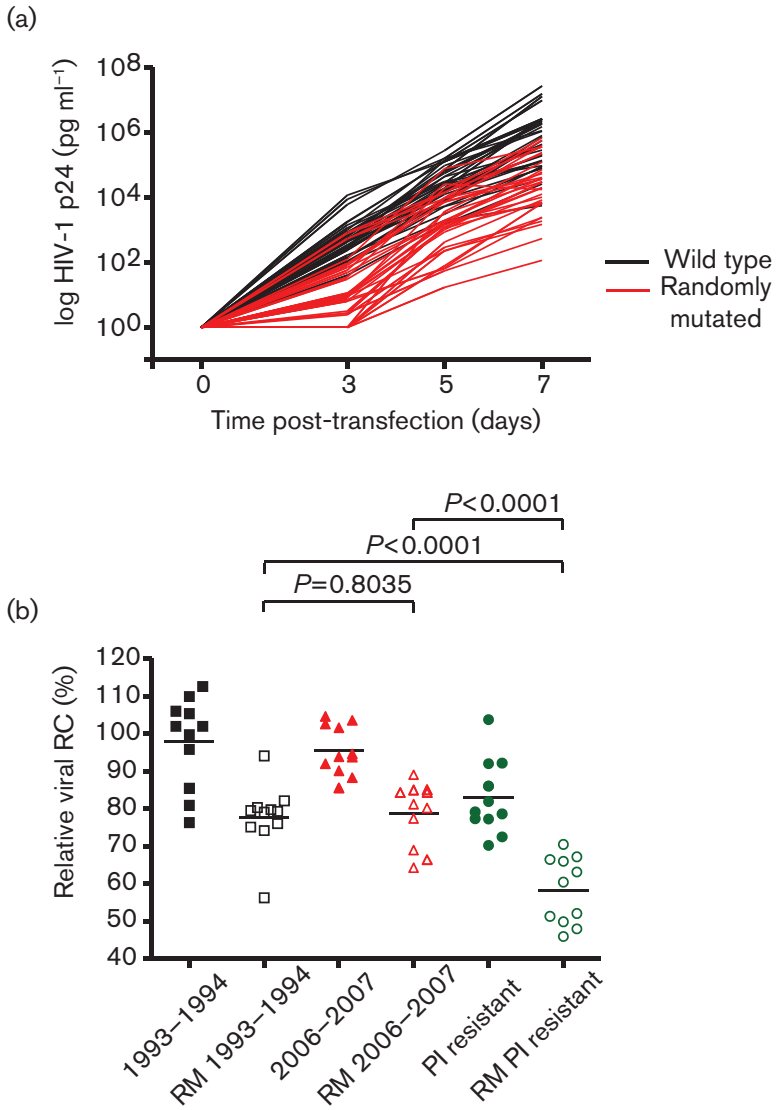


Figure 34: Comparison of the ex vivo RC of the 33 wild type protease viruses and the 33 randomly mutated protease chimeric viruses. (A) The kinetics of viral replication. Black and red lines indicate wild type and randomly mutated viruses, respectively. (B) Comparison of the slope of the natural logs from days 0 to 7 after transfection. The RC of each variant was compared with that of the virus with the HXB2 recombinant protease (100%); the relative RC values are shown. Black symbols represent wild type protease chimeric viruses, and white symbols represent randomly mutated (RM) protease chimeric viruses. The values represent the means of at least three independent experiments. P values were determined using the unpaired t test.