Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

A molecular study on the evolution of a subtype B variant frequently found in Brazil

M.E. Pinto[1], C.G. Schrago[1], A.B. Miranda[2] and C.A.M. Russo[1]

[1]Departamento de Genética, Laboratório de Biodiversidade Molecular, Instituto de Biologia, UFRJ, Rio de Janeiro, RJ, Brasil
[2]Departamento de Bioquímica e Biologia Molecular, Fiocruz, Rio de Janeiro, RJ, Brasil

Corresponding author: C.A.M. Russo
E-mail: claudia@biologia.ufrj.br

Genet. Mol. Res. 7 (4): 1031-1044 (2008)
Received February 7, 2008
Accepted June 5, 2008
Published October 14, 2008

ABSTRACT. In spite of the remarkable diversity of HIV-1 *env* genes, several amino acids are extremely conserved, probably due to functional constraints. One example is the proline found at the second position of the GPGR motif. Several viruses, however, bear substitutions at this site, for instance, GWGR subtype B variant. GWGR viruses are described in Brazil since the beginning of the epidemics, but the extent of their dispersion or the geographical origin of the variant remains unknown. In the present study, phylogenetic trees were constructed in order to study the origin and spread of this variant. All GWGR sequences as well as a subset of subtype B sequences available were included in the analyses. Analyses of differential selection were also performed on GWGR and non-GWGR sequences in order to unveil evolutionary novelties due to the action of positive selection. Although the GWGR variant was found at least in 23 countries, its expansion probably has a single origin, and Brazil is the epicenter.

Key words: HIV-1; Subtype B; GWGR variant; Molecular evolution; Origin; Selective pressure

INTRODUCTION

A remarkable feature of the HIV-1 is their great genetic diversity. This unexpectedly high variability is due to the instability of their RNA genomes. From all the protein-encoding genes, the most variable is the *env* gene. The *env* gene encodes for the envelope proteins associated with the host cell-HIV interaction (Shioda et al., 1991). Nevertheless, due to their functional relevance, several amino acid residues are extremely conserved among HIV-1 variants. Indeed, the vast majority of HIV-1 sequences, and their putative common SIVcpz ancestor, have the GPGX signature pattern (motif) at the central portion of the V3 loop of the *env* gene (Kuiken et al., 2000).

The well-conserved motif in a variable region suggests a strong purifying selection pressure. Some lineages, however, have been found worldwide with alternate signature patterns, such as the GWGR subtype B variant in which the proline residue is substituted by a tryptophan (Potts et al., 1993; Desgranges et al., 1998; Tanuri et al., 1999). This variant was first recorded in Japan in the early 1990's (Shimizu et al., 1992), but a serologic study suggests its presence in Brazil since 1983 (Hendry et al., 1996).

Although rarely described worldwide, GWGR viruses seem to account for a relatively early and important proportion of subtype B infections in Brazil. Changes in this highly conserved residue provide an interesting case of study to test whether selective pressure was altered with the substitution. Therefore, the aim of the present study was to determine the origin and spread of the GWGR variant in the world and possible differences in selective pressures on lineages and along the codons.

MATERIAL AND METHODS

The presence of other subtype B variants in the course of the pandemic has been previously shown. Subtype B', found in Thailand (Ou, 1993) and other Asian countries (Brown et al., 1996; Cassol et al., 1996; Kusagawa et al., 1998; Chen et al., 1999), for instance, has the subtype B signature pattern GPGR preserved or, alternatively, the GPGQ motif. Another variant, described in Trinidad and Tobago, is characterized by a threonine deletion, terminal to the crown of the V3 loop, and the replacement of the fourth-position arginine by other amino acid residues (Cleghorn et al., 2000). Finally, a third variant, found in Korea, has the unusual GPGS motif at the crown of the V3 loop (Kim et al., 1999).

In order to ascertain the origin and the spread of the GWGR subtype B variant in the world, two sets of available subtype B sequences were carefully assembled to be included in the phylogenetic analysis: the GWGR motif set and a sample that included GPGR and other sequence variants of subtype B.

GWGR sequences

On account of the large number of sequences available in GenBank, the search of GWGR sequences was carried out using Perl scripts. The script searched for the CX(n) GWGRX(n)C pattern along the HIV sequences, in which letters correspond to amino acids. The two bordering C's represent the two cysteines that flank the V3 loop and X(n) is any particular amino acid residues repeated n times. To improve sensitivity, variations in the basic pattern were created to identify sequences with expected amino acid substitutions at the 1st, 3rd, or 4th sites, substituting X by the corresponding letter, namely G, G or R. Clones and non-B sequences were excluded.

Whenever sequence subtype was unavailable, the subtyping was performed by phylogenetic analysis using the reference sequences from the Los Alamos National Laboratory database (http://hiv-web.lanl.gov). A total of 116 subtype B sequences with W at the crown of the V3 loop were found circulating in 23 countries (Table 1) from Latin America (Argentina, Bolivia, Brazil, Chile, and Paraguay), Central and North America (Canada, Cuba, Mexico, and the USA), Western and Eastern Europe (Czech Republic, Denmark, France, Germany, Italy, Lithuania, The Netherlands, Spain, Switzerland, and United Kingdom), Asia (China, Japan and Philippines) and Africa (Reunion Islands). As mentioned previously, however, the majority of sequences (67%) were found in Brazil (Table 1).

Table 1. Distribution of the subtype B sequences used to construct the phylogenetic tree on Figure 1 according to their country of origin and amino acid displayed at the second position of the GXGX motif.

Country	Amino acid		Country	Amino acid	
	Any	Trp (W)		Any	Trp (W)
Angola	1	No	Italy	3	1
Argentina	3	3	Ivory Coast	1	No
Australia	9	No	Japan	4	1
Barbados	2	No	Lebanon	1	No
Belgium	2	No	Lithuania	1	1
Bolivia	2	2	Malaysia	1	No
Brazil	16	78	Martinique	1	No
Cameroon	4	No	Mexico	1	1
Canada	4	1	Myanmar	4	No
Chile	No	2	The Netherlands	10	2
China	6	1	New Zealand	1	No
Colombia	1	No	Paraguay	No	1
Congo	1	No	Peru	1	No
Cuba	2	2	Philippines	No	3
Cyprus	1	No	Portugal	2	No
Czech Republic	1	2	Puerto Rico	1	No
Denmark	1	1	Reunion Islands	1	1
Ecuador	1	No	Russian Federation	2	No
Egypt	1	No	Senegal	2	No
Estonia	1	No	Singapore	2	No
Ethiopia	1	No	South Africa	10	No
Finland	1	No	Spain	7	4
France	12	4	Sudan	1	No
French Guiana	1	No	Sweden	4	No
Gabon	1	No	Switzerland	1	2
Gambia	1	No	Taiwan	1	No
Germany	2	1	Thailand	11	No
Greece	2	No	Trinidad Tobago	2	No
Guinea-Bissau	1	No	UK	9	1
Haiti	3	No	Ukraine	1	No
Honduras	1	No	Uruguay	2	No
Hungary	1	No	USA	38	1
India	3	No	Venezuela	3	No
Indonesia	2	No	Vietnam	1	No
Israel	1	No	Total	222	116

Subtype B sample

Our sample of subtype B sequences was obtained as follows. First, all HIV-1 sequences classified as subtype B were downloaded from the Los Alamos National Laboratory database. Pseudogenes (as noted in GenBank), clones and sequences with less than 300 bp were excluded from the following analyses. Sequences were ordered according to their country of origin and, then, to their GenBank accession number. Also, due to the large number of sequences, for each country subset, only the first sequence submitted to the databank was selected.

The subtype B sample subset comprised 222 sequences, of which 211 had a proline at the V3 loop motif. The remaining 11 sequences had any other amino acid. No GWGR sequence was included in this subset. The geographic distribution of the sequences is also shown in Table 1.

Evolutionary analyses

The final set included the 338 subtype B sequences described and two non-B sequences, accession numbers AF268919 (subtype F) and AF069670 (subtype A), which were used as outgroups. The 340 sequences were aligned using the ClustalW algorithm (Higgins et al., 1996). After visual inspection, utmost care was taken in the manual adjustment of insertions and deletions with the BioEdit program. This was a particularly difficult task due to the extreme variability of the region. The final alignment was 486 bp long and it is presented as supplemental information. After the alignment, the MEGA2 program was used to run all phylogenetic analyses. The mean *p*-distance value and the transition/transversion ratio (R) were 0.133 and 1.576, respectively (pairwise deletion). The average nucleotide compositions were 23.9% (T), 14.7% (C), 42.2% (A), and 19.1% (G). A neighbor-joining (NJ) tree (Saitou and Nei, 1987) was constructed using the Jukes-Cantor substitution model. The statistical support for each interior branch of the phylogenetic tree was performed with the interior branch test (Sitnikova et al., 1995).

We have also performed an evolutionary pathway analysis on the origin of the GWGR variant from a GPGR ancestor. For this analysis, sequences with putative intermediary substitutions were downloaded from databanks and our strategy was as follows. A secondary databank was created with the phpMyAdmin program in which *env* sequences were downloaded in FASTA format from the GenBank. For the download, the following key words were used “HIV-1 AND *env* NOT HIV-2 NOT SIV”. Whenever available, the subtype classification and the geographic origin of the sequences were also included for our records. Third, sequences were aligned together with the V3 fragment of the RF subtype B reference sequence using the ClustalX program in order to exclude those that did not span the V3 region. Finally, signature patterns of the V3 sequences were identified using a set of Perl scripts in which the central X of the CX(n)GXGRX(n)C expression was substituted by any other amino acid, except proline and tryptophan (sequences with these amino acids were previously identified as described above).

Analysis of differential selection

Also, we aimed to investigate whether GWGR and GPGR sequences were submitted to differential selective pressures. In this case, heterogeneity of selective pressures within groups and along the sequences was analyzed by inferring the non-synonymous to synonymous rate ratio (d_N/d_S, also dubbed ω). This statistic offers a measure of how much non-synonymous substitutions (amino acid-changing) accumulated compared to synonymous (silent) substitutions. Since only amino acid-changing substitutions are visible to natural selection, an ω ratio >1 indicates that substitutions that effectively changed the protein sequence were maintained in the virus population at a rate above the random background expectation (synonymous substitutions). Such phenomenon is known as positive (or diversifying) selection because the new molecule has escaped the selective sieve and was kept in the population. On the other hand, if the ω ratio is <1, substitutions that modified the protein sequence were eliminated, clearly indicating that there is a selective force to hold protein structure unchanged. When new molecules are swept from the population, this is called negative (or purifying) selection. Finally, we may theoretically realize what biological phenomenon is represented when ω = 1. This occurs when both rates are approximately equal, i.e., all substitutions are maintained in the population without any noticeable preference. Such kind of accumulation of substitutions is known as neutral evolution.

The d_N/d_S ratios were studied using maximum likelihood (ML) methods, available in the PAML 3.15 program (Yang, 1997), that estimate lineage-specific ω (Yang, 1998) and codon-specific ω (Yang et al., 2000). To investigate d_N/d_S ratios on lineages, sequences were separated into two groups - GWGR and non-GWGR. Since ML algorithms are time-consuming, we reduced the original sample to obtain 21 and 22 sequences of GWGR and non-GWGR groups, respectively. The choice of sequences was made to maximize the number of codons analyzed and to obtain a global genetic diversity in which ML methods have both power and accuracy (Anisimova et al., 2002).

To infer lineage-specific ω's, we constructed the NJ tree of all 43 sequences using the Jukes Cantor model. An average ω was then estimated for the whole topology using the CODEML program (available in the PAML package). This null hypothesis was tested against the alternative hypothesis in which the GWGR clade was allowed to have its own separate ω value. If the log-likelihood (lnL) of the alternative hypothesis is significantly higher than H_0, the scenario of two distinct, lineage-specific ω's, is accepted. The test was performed via the likelihood ratio test (LRT), in which twice the difference of lnL is compared with the significance level from a χ^2 distribution with one degree of freedom.

In codon-specific analyses, each group was analyzed independently to test if ω heterogeneity along codons differed in each set. NJ trees were estimated for GWGR and non-GWGR sequences separately, and each topology was entered on CODEML together with sequence alignments. The Yang et al. (2000) method tests between two nested models of ω evolution - the nearly neutral and positive selection. In the nearly neutral model, codons are allowed to belong to an ω < 1 category or to a neutral (ω = 1) category. In the positive selection model, a third category is created for positively selected codons (ω > 1). Again, LRT is used to decide if the lnL increase in the parameter-richer model (positive selection) is significant. If so, codons that were assigned to the ω > 1 category with posterior probability >95% (a = 5%) are considered to be under positive selection. Moreover, in the positive selection model, posterior probabilities for codon assignment are calculated for all three categories, ω < 1, ω = 1 and ω > 1. Thus, it is possible to estimate an average ω for a given codon by summing up the products of posterior probabilities and inferred ω:

$$\overline{\omega}_i = \sum_{K=1}^{3} P(\hat{\omega}_K) \cdot \hat{\omega}_K \qquad \text{(Equation 1)}$$

where $P(\hat{\omega}_K)$ is the posterior probability that codon *i* belongs to class *K*. There are three classes (negative selection, neutral evolution and positive selection) of codon assignment. This mean ω for a codon allows the overall heterogeneity along codons between groups to be tested via the Wilcoxon rank sum test (Choisy et al., 2004), in which the null hypothesis is that the overall shift of ω values between pairs is equal to zero. Also, we can estimate how well mean ω's correlate between groups with the Pearson product-moment correlation.

RESULTS

The origin of GWGR

The phylogenetic analysis of the subtype B sequences is shown in Figure 1. Interestingly, all but three GWGR sequences were clustered in a single interior branch with an 82% confidence probability (CP), regardless of their geographic origin. This branch will be referred henceforth as the "GWGR branch". The three remaining GWGR sequences clustered with other GPGR sequence branches and will be considered in the following.

A few hypotheses would explain the position of these three sequences. The first is the phylogenetic instability of the interior branches between the lineages they were inserted and the GWGR branch. These sequences, however, did not cluster close to the GWGR branch and the branches that do separate them from the main GWGR cluster are occasionally very well supported, indicating their consistent position outside the branch.

The second hypothesis would consider that the three outcast sequences are intra-subtype recombinants. In this case, the V3 regions were inherited from GWGR parental viruses, whereas the remaining fragments were derived from GPGR ancestors. In order to test this hypothesis, a recombination test was applied to the largest GWGR fragment that clustered outside the GWGR branch.

The selected query sequence is 882 bp long and it derives from Mexico (accession #AF200859). It was aligned using the BioEdit program along with 16 GPGR sequences from the branch the query was clustered, four GWGR sequences that clustered in the GWGR branch, four subtype B reference sequences, and a GPGR Mexican sequence. An analysis performed with the RDP and Bootscan methods implemented in the RDP program (http://darwin.uvigo.es/rdp/rdp.html; Martin and Rybicki, 2000) found no evidence for recombination in the query sequence (alignment and analysis parameters used are available upon request). This hypothesis, however, could not be completely ruled out due to the small extension of the sequence alignment analyzed.

Finally, a third hypothesis that explains the three GWGR sequences that were positioned outside the GWR branch predicts that these sequences are, in fact, representatives of distinct evolutionary origins for the GWGR viruses. In this case, GWGR viruses have evolved at least four times in the course of the HIV-1 epidemics. If it were the case, however, there would be evidence supporting that their world expansion resulted from a single common ancestor that circulated in Brazil before 1983, as discussed below (Hendry et al., 1996).

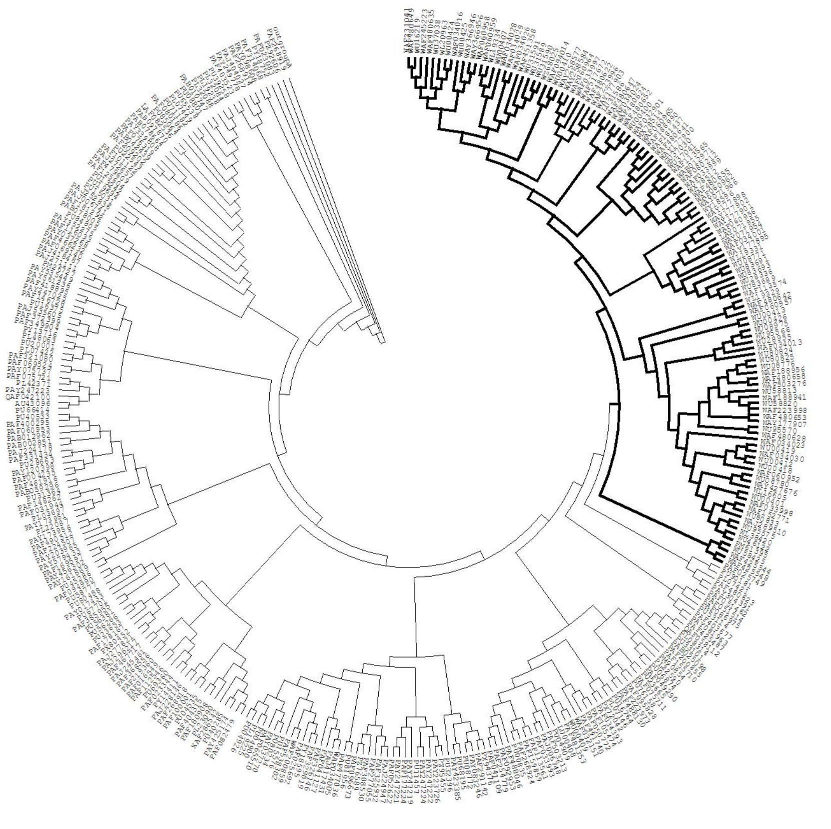

Figure 1. Neighbor-joining tree constructed with the Jukes-Cantor model. The tree includes all available GWGR sequences and a subset of subtype B sequences. Subtypes F and A were used as outgroups (accession numbers AF268919 and AF069670).

GPGR to GWGR

In spite of the outcast sequences, the GWGR branch clearly demonstrates that the vast majority of worldwide GWGR sequences had a single origin (Figure 1). Interestingly, this branch also includes sequences in which tryptophan at the GWGR motif was substituted by glycine (G), methionine (M), phenylalanine (F), and leucine (L). This high motif diversity clustered within the GWGR branch suggests a longstanding presence of GWGR in the course of the epidemics. Not surprisingly, 70% of sequences with amino acids other than tryptophan in the GWGR branch came from Brazil (Table 2), where the virus seems to have circulated at least since 1983 (Hendry et al., 1996). Conversely, sequences with amino acids for which the codons are close to the proline (codon CCA), that is, glutamine (GQGX motif, codon CAA for Q), alanine (GAGX, codon GCA for A) and leucine (GLGX, codon CTA for L) that most probably had GPGX ancestors, clustered outside the GWGR branch.

Table 2. Geographic distribution of the subtype B sequences with amino acids other than proline and tryptophan at the second position of the GXGX motif.

	Amino acids															
	Ala (A)		Arg (R)		Phe (F)		Gln (Q)		Gly (G)		Leu (L)		Met (M)		Ser (S)	
	GCA		CGG	CGA/CGG	TTT/TTC		CAG	CAG/CAA	GGG	GGA/GGC	TTG	CTA/CTG	ATG		TCA/TCG	
	W	P	W	P	W	P	W	P	W	P	W	P	W	P	W	P
Argentina	1				1				1				1			
Australia								1				2	1			1
Belgium				1												
Brazil			2		12						7	2	3			
Canada								3				1				
China								1								1
Cuba				1												
Cyprus												1				
Czech Republic								1				2				
France				1					1			2		1		
Germany												1				
Haiti										1						
Hungary				1												
India								1								
Israel								1								
Italy								1				4				2
Japan												1				1
Korea								1								
Myanmar								6				1				
The Netherlands						1		3			1	8				1
South Africa								2								
Spain				1		1		1	1			2				
Switzerland												2				
Taiwan								1				1				
Thailand		1					1	11				6				
Trinidad Tobago		1														
UK								2				1				
Uruguay			1													
USA		3		1		3		4		1		11		2		6
Unknown												1				
Total	1	5	3	6	13	5	1	40	3	2	8	49	5	3		12

All these sequences were included in the phylogenetic tree constructed in order to study the evolutionary pathway proposed for the GWGR variant origin. Columns W and P for each amino acid represented the mean position of the sequences in the tree, that is, inside (W) or outside (P) the GWGR branch.

Regardless of the number of times the GWGR variant appeared in the pandemics' course, any proline-to-tryptophan substitution is the result of, at least, one substitution in each position of the ancestral CCA. This is because tryptophan is only coded by TGG (Kuiken et al., 2000). The most parsimonious evolutionary pathway for the GWGR variant origin, that is, the one with the least number of substitution events, is shown in Figure 2.

In this hypothetical pathway, the first step would be the CCA→CCG synonymous substitution (substitution at the third position of any codon is much more frequent than at the two other positions). In the following there would be non-synonymous substitutions at the first or second positions of CCG, generating TCG or CGG, respectively. The third step is the substitution of cytosine (TCG or CGG) by guanine or thymine to generate TGG that codes for tryptophan. If this hypothesis holds, sequences with serine (TCG) or arginine (CGG) at the second position of the tetrapeptide motif would represent intermediary steps in the GPGR to GWGR evolutionary pathway. Since the set of sequences in the phylogenetic tree (Figure 1) included no GRGR and no GSGR sequences, a second tree was constructed in order to investigate this possibility.

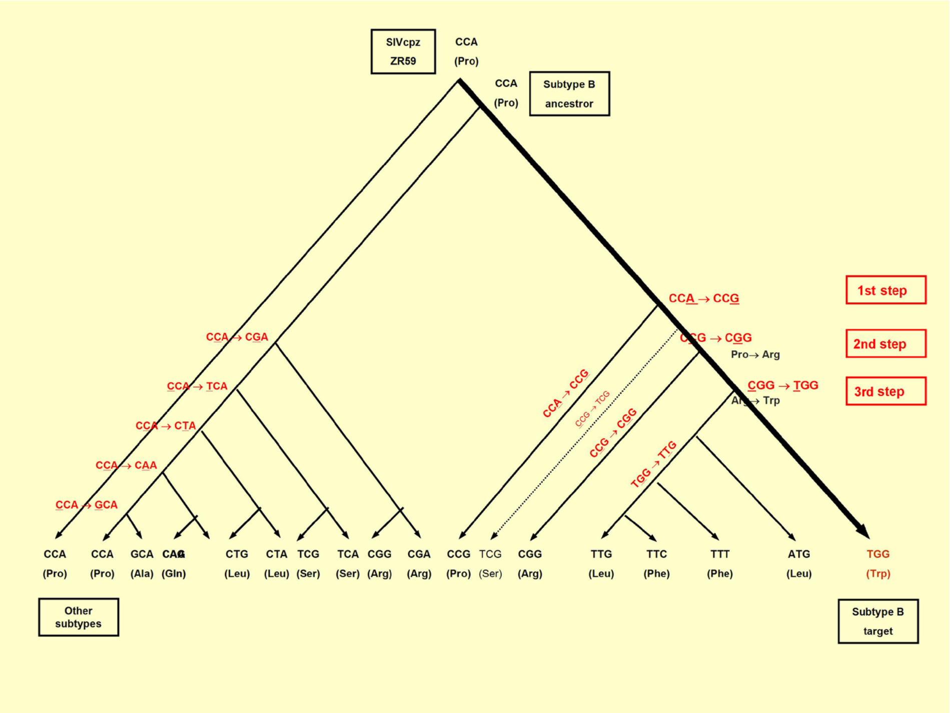

Figure 2. The most parsimonious evolutionary pathway proposed for the GWGR variant origin. The first step is a synonymous substitution at the third base of CCA, generating the codon CCG. Next, non-synonymous substitutions might have occurred at the first or second nucleotide sites generating codons TCG or CGG for serine and arginine, respectively. The third step is the substitution of C (TCG or CGG) by G or T to form codon TGG for proline.

In this case, all available subtype B sequences with GRGR and GSGR motifs were identified in the database created with the phpMyAdmin program as previously described. All sequences with GAGR, GFGR, GGGR, GLGR, GMGR, and GQGR motifs were also included, in order to investigate their evolutionary relationship with the GWGR branch. This new subset included 145 sequences. The initial 340 (Figure 1) plus the 145 new sequences (485 sequences in total) were aligned with the BioEdit program. An NJ tree was constructed with the Jukes Cantor distance model and is available upon request.

Schematically, Table 2 shows the distribution of sequences in the new phylogenetic tree. Sequences that clustered into the GWGR branch were distributed along the W columns, according to the codon for the amino acid found at the second position of the GXGR motif and their country of origin. On the other hand, sequences that clustered with the GPGX sequences were located in the P columns. As may be seen, three of the nine GRGR sequences were found clustered with GWGR variants. For all three sequences, arginine residues (GRGX motif) were coded by CGG, one of the putative intermediates for the GWGR viruses. These sequences were found in Brazil (N = 2) and in Uruguay (N = 1) during the 1990s. Conversely, no GSGR sequence clustered inside the GWGR branch. Table 2 also shows that the majority non-GWGR variants (GFGX, GLGX and GMGX) that probably have GWGR sequences, as ancestors (see Figure 2) came from Brazil.

Differential selection on lineages and codons

The d_N/d_S value estimated for the full topology was 0.55 and the lnL of this null hypothesis was -12555.14. When the GWGR lineage was allowed to have its own ω, it was inferred at 0.54 against a background ω of 0.55 for the remaining branches of the tree. The lnL of the alternative hypothesis was -12555.13, resulting in an LRT statistic of 0.03, which forbids the exclusion of the null hypothesis with a P value close to 1.

Although average ω values for groups were not significantly different, there were differences in codon sites assigned to the positive selection category within lineages. As Tables 3 and 4 show, in both groups analyzed, the positive selection was the model that best explained the data, because lnL values were significantly increased in this model (P < 1%). GWGR sequences showed 8 codon sites under diversifying selection, while non-GWGR sequences were inferred to have 15 sites. Codons 77, 100, 104, and 124 were assigned to the ω > 1 category in both groups.

Table 3. Codon-specific omega (ω) values under different evolutionary models for GWGR sequences and their respective log-likelihoods (lnL).

Model	lnL	ω estimates	PSS
Single ω	-2676.35	$\omega = 0.54$	NA
Nearly neutral	-2591.55	$\omega_0 = 0.13$ $\omega_1 = 1.0$	NA
Positive selection	-2567.82	$\omega_0 = 0.12$ $\omega_1 = 1.0$ $\omega_2 = 3.97$	27, **29**, 77, 96, **99**, **100**, **104**, **124**

PSS = positively selected sites: NA = not applicable. Sites in bold have posterior probability >99% of being assigned to the ω > 1 (ω_2) category, otherwise >95%.

Table 4. Codon-specific omega (ω) values under different evolutionary models for`non-GWGR sequences and their respective log-likelihoods (lnL).

Model	lnL	ω estimates	PSS
Single ω	-2997.65	$\omega = 0.55$	NA
Nearly neutral	-2871.00	$\omega_0 = 0.10$ $\omega_1 = 1.0$	NA
Positive selection	-2840.92	$\omega_0 = 0.12$ $\omega_1 = 1.0$ $\omega_2 = 3.26$	**25**, **60**, 72, 77, 92, **94**, 95, 100, **103**, 104, 106, **107**, 116, **124**, 156

PSS = positively selected sites, NA = not applicable. Sites in bold have posterior probability >99% of being assigned to the ω > 1 (ω_2) category, otherwise >95%.

Parametric estimates for ω's were very similar in both groups. We also observed that the overall distribution of mean d_N/d_S in sites demonstrated the same trends in both sets (Figure 3). Such scenario was confirmed by the Wilcoxon test, which could not find a significant difference in the overall distribution of mean ω along sequences in both groups (P = 0.36), i.e., none of the groups had a tendency to possess higher mean d_N/d_S along the sequence (Figure 4).

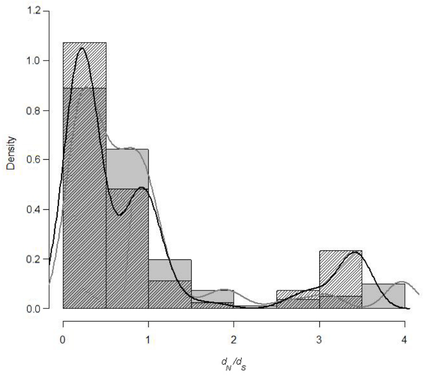

Figure 3. Histogram of d_N/d_S ratios inferred from GWGR (gray bars and lines) and non-GWGR sequences (dark lines and shading).

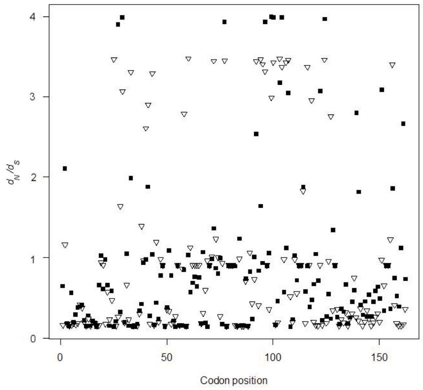

Figure 4. Distribution of d_N/d_S ratios along codons in GWGR (squares) and non-GWGR sequences (triangles).

The majority of codon sites had a mean d_N/d_S smaller than one. There is also a significant proportion of sites with mean ω slightly above neutrality, indicating possible false positives. However, in both lineages, many sites under positive selection were inferred to have ω higher than 3.0, which is clearly observed in Figure 3. Mean d_N/d_S values for codon sites in GWGR and non-GWGR sequences were significantly correlated (P < 1%), with a coefficient of 60.5%.

DISCUSSION

Our results showed that the GWGX motif at the crown of the V3 loop of the *env* gene evolved relatively fast in Brazil since 1983. Moreover, the results suggest that its longstanding presence in the country probably favored expansion of other variants, mainly GLGR and GFGR. Our results strongly indicate that the epidemics of GWGR viruses have a single origin and that they probably evolved in Brazil. The distribution of GPGX sequences into several branches along the phylogenetic tree (Figure 1), on the other hand, characterizes multiple introductions of the ancestral subtype B viruses in each country. In this scenario, it would be expected that GWGR viruses may only account for a small proportion of the Brazilian epidemics, which does not occur.

Several authors have systematically studied the presence of GWGR viruses in Brazil. Curiously, two independent studies suggested that GWGR viruses are associated with a slower progression of the immunodeficiency when compared to the GPGR subtype B infections (Santoro-Lopes et al., 2000; Casseb et al., 2002). A slower progression toward immunodeficiency itself would mean higher fitness, due to a longer non-symptomatic period for the GWGR variant. In this case, longer periods of silent infections would mean greater chances for the viruses to infect new hosts and, thus, survive and reproduce.

If this were the case, we would expect the frequency of GWGR (or GPGR/GWGR ratio) in HIV-1 subtype B viruses in Brazil to be increasing along the course of pandemics. Nevertheless, this is not what we have found in the literature. Serological studies have reported an almost 50% GPGR/GWGR ratio in Brazil since the 1980s. Yet, if GWGR viruses were somewhat restricted to Brazil, we would expect their frequency naturally to diminish along the course of pandemics. This is because we would expect new introductions of HIV-1 subtype B in Brazil to be GPGR, common worldwide.

From the analyses of differential selection, we learn that GWGR and non-GWGR sequences show statistically indistinguishable lineage-specific d_N/d_S ratios. Also, when codon-specific d_N/d_S are compared, there is a general pattern of many codons under purifying selection and a small proportion of sites under diversifying selection with ω ranging from 3.0 to 4.0. The occurrence of false positives could explain the difference found in the sites inferred under positive selection, especially when a large number of codon sites were near the neutrality cutoff. This does not seem to be the case here, as all sites assigned to the ω > 1 category had a mean d_N/d_S ratio higher than 3.0. Actually, codons with d_N/d_S close to 1 did not show posterior probabilities >95% of being positively selected.

Another feasible explanation for the difference in codon compositions under positive selection is that GWGR sequences did not contain enough information to reveal the large number of sites inferred in non-GWGR sequences. If this were true, we would expect that codon sites that were estimated under diversifying selection exclusively in non-GWGR sequences to have posterior probabilities close to the cutoff of 95% used. This may be the case for codons

103 and 107 which displayed posterior probabilities of 75 and 71% to be in the ω > 1 category in GWGR sequences. Conversely, codon sites 29 and 96 had posterior probabilities of 84 and 94%, respectively, to be assigned to the ω > 1 class in non-GWGR sequences. The remaining differences are probably due to biological factors instead of statistical artifacts.

We were prompted to investigate the amino acid variation found in exclusive sites under positive selection in GWGR and non-GWGR sequences. Curiously, codon site 27, which was assigned to the ω > 1 category in GWGR sequences is monomorphic (phenylalanine) in non-GWGR sequences. The same is found in codon 25, which is under positive selection with 99.6% probability in non-GWGR sequences and is fixed for glutamic acid in GWGR. Again, it is unlikely that sampling or statistical artifacts caused such discrepancy.

In this sense, if the ratio remains the same, it may possibly indicate a weak form of selection that our tests were not powerful enough to detect. Other studies on syndrome progression and more powerful selection tests may be able to put this hypothesis of GWGR high fitness due to a slower progression toward immunodeficiency to rest.

ACKNOWLEDGMENTS

We are particularly indebted to Prof. Claudio Struchiner. M.E. Pinto was the recipient of a CAPES fellowship and C.A.M. Russo was the recipient of CNPq and FAPERJ fellowships.

REFERENCES

Anisimova M, Bielawski JP and Yang Z (2002). Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* 19: 950-958.

Brown TM, Robbins KE, Sinniah M, Saraswathy TS, et al. (1996). HIV type 1 subtypes in Malaysia include B, C, and E. *AIDS Res. Hum. Retroviruses* 12: 1655-1657.

Casseb J, Komninakis S, Abdalla L, Brigido LF, et al. (2002). HIV disease progression: is the Brazilian variant subtype B' (GWGR motif) less pathogenic than US/European subtype B (GPGR)? *Int. J. Infect. Dis.* 6: 164-169.

Cassol S, Weniger BG, Babu PG, Salminen MO, et al. (1996). Detection of HIV type 1 env subtypes A, B, C, and E in Asia using dried blood spots: a new surveillance tool for molecular epidemiology. *AIDS Res. Hum. Retroviruses* 12: 1435-1441.

Chen J, Young NL, Subbarao S, Warachit P, et al. (1999). HIV type 1 subtypes in Guangxi Province, China, 1996. *AIDS Res. Hum. Retroviruses* 15: 81-84.

Choisy M, Woelk CH, Guegan JF and Robertson DL (2004). Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J. Virol.* 78: 1962-1970.

Cleghorn FR, Jack N, Carr JK, Edwards J, et al. (2000). A distinctive clade B HIV type 1 is heterosexually transmitted in Trinidad and Tobago. *Proc. Natl. Acad. Sci. U. S. A.* 97: 10532-10537.

Desgranges C, Fillon S, Letourneur F, Buzelay L, et al. (1998). HIV-1 subtypes in Santiago, Chile. *AIDS* 12: 1563-1565.

Hendry RM, Hanson CV, Hanson CV, Morgado M, et al. (1996). Immunoreactivity of Brazilian HIV isolates with different V3 motifs. *Mem. Inst. Oswaldo Cruz* 9: 347-348.

Higgins DG, Thompson JD and Gibson TJ (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* 266: 383-402.

Kim EY, Cho YS, Maeng SH, Kang C, et al. (1999). Characterization of V3 loop sequences from HIV type 1 subtype B in South Korea: predominance of the GPGS motif. *AIDS Res. Hum. Retroviruses* 15: 681-686.

Kuiken CL, Foley B, Hahn B, Korber B, et al. (2000). HIV Sequence Compendium 2000. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos.

Kusagawa S, Sato H, Watanabe S, Nohtomi K, et al. (1998). Genetic and serologic characterization of HIV type 1 prevailing in Myanmar (Burma). *AIDS Res. Hum. Retroviruses* 14: 1379-1385.

Martin D and Rybicki E (2000). RDP: detection of recombination amongst aligned sequences. *Bioinformatics* 16: 562-563.

Ou CY, Takebe Y, Weniger BG, Luo CC, et al. (1993). Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand. *Lancet* 341: 1171-1174.

Potts KE, Kalish ML, Lott T, Orloff G, et al. (1993). Genetic heterogeneity of the V3 region of the HIV-1 envelope glycoprotein in Brazil. Brazilian Collaborative AIDS Research Group. *AIDS* 7: 1191-1197.

Saitou N and Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.

Santoro-Lopes G, Harrison LH, Tavares MD, Xexeo A, et al. (2000). HIV disease progression and V3 serotypes in Brazil: is B different from B-Br? *AIDS Res. Hum. Retroviruses* 16: 953-958.

Shimizu N, Takeuchi Y, Naruse T, Inagaki M, et al. (1992). Six strains of human immunodeficiency virus type 1 isolated in Japan and their molecular phylogeny. *J. Mol. Evol.* 35: 329-336.

Shioda T, Levy JA and Cheng-Mayer C (1991). Macrophage and T cell-line tropisms of HIV-1 are determined by specific regions of the envelope gp120 gene. *Nature* 349: 167-169.

Sitnikova T, Rzhetsky A and Nei M (1995). Interior-branch and bootstrap tests of phylogenetic trees. *Mol. Biol. Evol.* 12: 319-333.

Tanuri A, Swanson P, Devare S, Berro OJ, et al. (1999). HIV-1 subtypes among blood donors from Rio de Janeiro, Brazil. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* 20: 60-66.

Yang Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13: 555-556.

Yang Z (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15: 568-573.

Yang Z, Nielsen R, Goldman N and Pedersen AM (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155: 431-449.