Biol Res 43: 149-163, 2010

Fixations of the HIV-1 env gene refute neutralism: New evidence for pan-selective evolution

Carlos Y Valenzuela, Sergio V Flores and Javier Cisternas

Programa de Genética Humana, ICBM, Facultad de Medicina, Universidad de Chile, Independencia 1027, Santiago, Chile.

ABSTRACT

We examined 103 nucleotide sequences of the HIV-1 *env* gene, sampled from 35 countries and tested: I) the random (neutral) distribution of the number of nucleotide changes; II) the proportion of bases at molecular equilibrium; III) the neutral expected homogeneity of the distribution of new fixated bases; IV) the hypothesis of the neighbor influence on the mutation rates in a site. The expected random number of fixations per site was estimated by Bose-Einstein statistics, and the expected frequencies of bases by matrices of mutation-fixation rates. The homogeneity of new fixations was analyzed using χ^2 and trinomial tests for homogeneity. Fixations of the central base in trinucleotides were used to test the neighbor influence on base substitutions. Neither the number of fixations nor the frequencies of bases fitted the expected neutral distribution. There was a highly significant heterogeneity in the distribution of new fixations, and several sites showed more transversions than transitions, showing that each nucleotide site has its own pattern of change. These three independent results make the neutral theory, the nearly neutral and the neighbor influence hypotheses untenable and indicate that evolution of *env* is rather highly selective.

Key terms: Bose-Einstein distribution, fixations, HIV-1 *env* gene, neutral evolution, pan-selective evolution.

INTRODUCTION

The Neutral Theory of evolution proposed that the evolutionary state (fixation, loss or polymorphism) of most alleles in loci is acquired and maintained by mutation and random genetic drift (Kimura, 1968; King and Jukes, 1969; Kimura, 1979, 1991, 1993). Later on this same proposal about alleles was extended to nucleotide bases in nucleotide sites. Under the neutralist viewpoint, most alleles or bases are selectively equivalent (relative selection coefficients of alleles or bases are zero, or complementarily, the relative fitness is 1; Crow and Kimura, 1970) and are rarely acquired or maintained by positive selection; purifying (lethal and sub lethal) selection occurs infrequently. Thus neutral fixation is a fundamental concept for the Neutral Theory. The value of «rarely» has never been given and we will assume it means lower than 5%. Pure neutralism could not be supported based on results from studies of synonymous or non-synonymous codon replacement (Kreitman, 1996a; Ohta, 1996; Nei, 2005) and comparative genomics (Clark, 2006). Near-neutralism replaced neutralism, as well presenting unviable expectations. Near-neutralism accepts selective processes with selection coefficients of the order of the mutation rate or the reciprocal of the population size (Crow and Kimura, 1970; Kreitman, 1996a, 1996b; Ohta, 1996; Nei, 2005). Neutralism and near-neutralism entail an isotropic expected distribution of mutants in nucleotide sites along the DNA (Valenzuela, 2009). This expectation has been conclusively refuted in known genomes by isochores and signatures that have been maintained for billions of generations (Bernardi, 1993; Karlin and Mrazek, 1997; Valenzuela, 1997, 2009; Mrazek and Karlin, 2007). As well, the sequence of nucleotides is not what would be expected by random mutation (Gatlin, 1976; Valenzuela, 2009). Neutralists counter-argued that there was not sufficient time for base mutation rates (of the order of 10^{-8} mutation/site/ generation = m/s/g) to reach equilibrium, and the anisotropic distribution of bases could be the result of the influence of a base's neighborhood on the mutation rate at a specific nucleotide site (Jukes, 1976; Kimura and Ohta, 1977). However, neutralism, near-neutralism and the neighbor influence hypothesis still imply an isotropic random distribution of bases in DNA segments. By refuting DNA (or RNA) isotropy, we conclusively refuted neutralism, near-neutralism and the neighbor hypothesis (Valenzuela, 2009). This auxiliary hypothesis (neighborhood) can also be tested by maintaining a specific fixed neighborhood [for example, mutation analysis of the central adenine c(A) of the triplet AAA, c(T) of TTT, c(G) of GGG and c(C) of CCC], as we shall test in this article. Precision is needed at this point. The different bases (amino acids) found in a nucleotide site (amino acid locus) in different

Corresponding author: Carlos Y Valenzuela, Programa de Genética Humana, Instituto de Ciencias Biomédicas (ICBM), Facultad de Medicina, Universidad de Chile, Independencia 1027, Casilla 70061, Independencia, Chile, Fax: (56-2) 7373158, Phone: (56-2) 9786302, E-mail: cvalenzu@med.uchile.cl

Received: May 6, 2009. In revised form: May 11, 2010. Accepted: May 13, 2010.

lineages (strains or species) have been erroneously named substitutions or replacements. They are fixations or substitutions that reached a frequency of 100% in that specific lineage (strain, population or clone for parasites within one host), some time ago, and so remained until they were sampled. It is clear that fixation (permanence) is antithetical to substitution (continuous replacement; Valenzuela and Santos, 2006; Valenzuela, 1997, 2000, 2002, 2007, 2009). The substitution rate is a turnover rate, which for neutral mutations is (non-dimensionally) equal to the value of the mutation rate (which is also a turnover rate). Dimensionality is important because it indicates different processes. The dimension of mutations is mutation/site/generation = m/s/g. Since the probability for a mutation to reach the substitution level (frequency 1.0) is dimensionally substitution/mutation (sub/m), and equating both was obtained by their product (King and Jukes 1969) we have m/s/g times sub/m = sub/s/g; which is the rate of neutral mutation equate numerically the rate of neutral substitution, but they are different processes (substitutions include copies, population diffusion, etc.). Fixations cannot be expressed dimensionally "per generation", because they remain an undefined number of generations in that taxon [only fixations/(site or locus)]. Contrary to any observation, recurrent neutral substitutions in a site are expected to lead always to polymorphisms, not to fixations. Fixations can only be generated and maintained by selection, even in populations with only one bacterium (Valenzuela and Santos, 1996; Valenzuela, 1997, 2000, 2002, 2007, 2009). Since this conceptual mistake is generalized (King and Jukes, 1969; Nei, 2005; Clark, 2006), we use fixation and substitution properly and never as synonymous.

Viruses evolve fast and can be used to test neutral and nearly neutral expectations. Their mutation rates are near 10^{-4} to 10^{-6} m/s/g (Drake, 1993, 1999; Drake et al., 1998) for RNA lytic viruses, and one order of magnitude lower in retroviruses (10^{-5} to 10^{-7}) (Drake et al., 1998, 1999). Consequently, for lytic viruses with neutral mutation (nm) rates near 10^{-4}-10^{-5} nm/s/g, the neutral substitution (ns) rate is 10^{-4}-10^{-5} ns/s/g. The human immunodeficiency virus (HIV), the agent of AIDS has about 10,000 nucleotide sites and yields 10^{10} to 10^{11} viruses a day (Fauci and Lane, 2001), leading to a fast turnover of bases (Valenzuela and Santos, 1996; Valenzuela, 2000, 2002). Thus, analyses of molecular evolution of HIV-1 genomes are appealing opportunities for testing neutralism and near-neutralism. Both selective (Mani et al., 2002; Yang, 2001; Kitrinos et al., 2003; Travers et al., 2005; MacNeil et al., 2007) and neutral (Leigh-Brown 1997; Zhang 2004) evolution of the *env* gene of HIV has been proposed in several studies, either for HIV-1 or HIV-2. The ability of HIV genes to mimic those of the host proteins has been reported (Reiher et al., 1986; Serres, 2001).

Here we use *env* gene sequences from 103 HIV strains to test the random distribution of bases because we assume that neutral evolution implies a random distribution of mutations, substitutions or fixations. Under neutralism, viruses with high mutation rates are expected to rapidly reach the equilibrium of base frequencies. For neutral evolution, this equilibrium is attained when the four bases have frequencies near 0.25 (Jukes and Cantor, 1969; Valenzuela and Santos, 1996; Li, 1997). For nearly neutral evolution, the equilibrium frequency of the positive selected site is near 0.43 and those of the other (negatively selected) three bases are near 0.19. Although these neutral or nearly neutral expected frequencies for large populations have never been found, and most nucleotide sites remain monomorphic, the neutral and nearly-neutral theory are not considered to have been definitively refuted, probably because several misconceptions on neutral evolution do not allow for seeing these contradictions (Valenzuela 2000, 2002, 2007, 2009). Our aim is to test neutralism, near-neutralism and the neighborhood hypotheses by using independent tests for the random distribution of bases in the HIV-1 *env* gene, and to estimate how far the distribution of fixations is from randomness.

MATERIAL AND METHODS

Virus Sequences and Subtypes (Strains)

We analyzed 103 sequences belonging to 35 subtypes of HIV-1; a single stranded RNA lentivirus (subfamily) of the Retroviridæ family (Fauci and Lane, 2001) retrieved from Genbank (http://www.ncbi.nlm.nih.gov/Genbank/HIV (see accession numbers in APPENDIX 1). They were chosen, as far apart as possible, from 35 countries of the five continents, in order to avoid over-representation of sequences with the same recent origin, and to increase the origin variability, between 1984 and 2001. Consequently, according to neutralism and taking into account high mutation rates, a large number of sites should be highly polymorphic for A, T, G and C. The DNA segment that codes for this envelope protein (*env* GP120) was aligned by using ClustalX (1.8) (http://www-igbmc.u-strasbg.fr/BioInfo/). The consensus fragment had 3,239 nucleotide sites, including deletions, insertions and ambiguous bases of the 103 strains. We worked with sites whose bases were determined unambiguously among the 103 strains (2105 sites), but numeration from the 1st to the 3,239th site was conserved. The longest DNA segment was present in the sequence

AF119820 with 2,627 sites, 901 A (34.3%), 636 T (24.2%), 621 G (23.6%) and 469 C (17.9%). This sequence was taken as a reference to estimate the expected number of bases and triplets and the correlation between neighbor sites (also used in Valenzuela 2009). We demonstrated previously that the correlation of the bases between pairs of sites discriminated between consecutive sites (0-site separation) and pairs separated by 1, 2, 3 or more sites (Valenzuela, 2009).

Rationale to Study the Distribution of Fixations

Our aim is to test whether evolution is neutral (or nearly neutral) or selective. If evolution is neutral (or nearly neutral), then the distribution of mutations, substitutions and fixations (or complementarily losses) on nucleotide sites can be expected to be random (or nearly random). If evolution is selective, the distribution of mutations, substitutions and fixations on loci or sites can be expected to deviate from randomness or to be random. Thus, finding a non-random distribution of fixations refutes neutral and nearly neutral evolution, but finding a random distribution of fixations affirms neutral or nearly neutral evolution, but does not refute selective evolution. Thus, we use the powerful *modus tollens* logical method to refute neutral evolution; that is the proposition p (neutral evolution) implies q (random distribution of fixations). If q results false, then p is necessarily false. This logical method is conclusive, but this is not the case of searching for adaptive conditions where the fallacy of *modus ponens* is always present; p (selective processes) implies q (molecular constraints or correlations with structure or functions); we find those constraints or correlations, but this does not mean that p is true.

The study of the distribution of fixations at any DNA or RNA site of a genome or genome segment is a pre-transcriptional test for the evolutionary mode that also tests most transcriptional or post-transcriptional deviations from randomness. This occurs because an amino acid constraint of a protein is mostly a biased composition of amino acids of this protein, that is, a non-random distribution of amino acids among all the possible distributions of amino acids this protein can have. A non-random distribution of amino acids implies a non-random distribution of codons, which implies a non-random distribution of messenger RNA nucleotides, and this in turn implies a non-random distribution of DNA nucleotides. Our search for determining the deviation from randomness of the DNA nucleotides includes most amino-acid constraints correlated with non-random distribution of nucleotides. However, our study also includes all the pre-transcriptional selective causes correlated with non-random nucleotide distribution. We are not testing or searching for the particular causes of non-randomness, found in the distribution of fixations, coming from molecular mechanics (Hamacher, 2008), structure or functions of proteins, drug or immune resistance, differences in codon positions or in synonymous or non-synonymous substitutions, or any other transcriptional or post-transcriptional process that imply clearly selective or adaptive conditions (Chen et al., 2004). These features or molecular traits correspond to one or to a few amino acids or nucleotide sequences among a huge number of all the possible ones that are included in our basic analyses of all the possible fixations. Most of these invariable structures or functions have been considered as un-debatable "constraints" (Chen et al., 2004; Valenzuela, 2007, 2009). However, the acquisition and maintenance of these constraints should be determined prior to any analysis, but at present, this is not done at all [as for example: Was the genetic code (or any constraint) acquired by drift or selection? Was it (Were they) maintained by drift?). Thus, if we find a basic non-random distribution of fixations, we refute neutral and nearly neutral evolution, and our present purpose is satisfied; the transcriptional or post-transcriptional causes of the acquisition and maintenance of these non-random distributions are unnecessary. Other research should address these aims. If we find random distributions of fixations, then, since selective evolution is still possible (type II epistemic and statistical error), these studies will be necessary to complete the demonstration, but they could only increase the power of our demonstration by a little if we find a major deviation from random distribution. However, our previous longitudinal analyses of the total HIV-1 genome and the longest *env* GP 120 gene sequences (Valenzuela, 2009) demonstrated a degree of deviation from random distributions, so to start with this hypothesis it is consistent with that result.

As well, our analyses are addressed to the dynamic evolution of a nucleotide site (not a base or an allele). If we know the history of processes (mutations, substitutions or fixations) that occurred in a site, the conditions of coding or non-coding (for amino-acid), synonymous or non-synonymous substitutions, or other transcriptional or post-transcriptional properties are irrelevant, because neutral evolution (the null hypothesis) assumes that these processes are equally distributed for evolutionary purpose on the sites that lead to these properties. Unfortunately, this history cannot be known directly and the method of sequence alignment should be used as a sufficient and acceptable determination of sites where these processes occur. However, insertions, deletions

and micro-arrangements (micro-inversions) can confound sites and lead alignment methods to blur the true site history. This is why choosing different methods of alignment (parsimony, maximum likelihood, Bayesian, etc.) are also irrelevant for our purpose. Moreover, our method tolerates confounding of sites because the history of sites, under neutral evolution, is expected, as an average, to be equal. Our method was prepared for the analysis of complete genomes or genome segments regardless of the distinction between coding and non-coding regions and for longitudinal analyses of sequences (Valenzuela, 2009).

We chose the *env* gene because it shows more fixations than other HIV-1 proteins According to the neutral theory, a gene with more polymorphic sites (more substitutions, mutations or fixations) is a better candidate for being neutral than a highly monomorphic gene (Kimura, 1979) due to functional constraints, in spite of the contradictory proposition that more varied fixations implies a more selective genome region. The number of sites having a specific ancestral base was estimated by assuming that the ancestral base was the most frequent base in the site among the 103 strains used here. Ancestral states were also reconstructed using the parsimony criterion as implemented in PAUP (Swofford, 2002). Results from both approaches were convergent, and this approach gives advantage to neutral and nearly neutral evolution. For example, we found 741 sites with A as the ancestral base (ancestral fixations) that presented 8,524 different new fixations among the 103 strains; so our expected random distribution of 8,524 fixations on 741 sites has an expected mean equal to 11.5 fixations per site. To calculate the expected number of sites with 0, 1, 2, fixations, we need the following analysis.

I) Bose-Einstein distribution analyses for nucleotide sites having 0, 1, 2, 3, n fixations different from the ancestral base:

We assume that neutral base mutations that originate neutral base fixations, by drift, are evolutionarily undistinguishable events that occur in evolutionarily distinguishable nucleotide sites. The random distribution of undistinguishable balls (events) in distinguishable boxes follows Bose-Einstein statistics (Feller, 1968; Valenzuela 2009). Thus, the neutral distribution of fixations found in the 103 strains can be tested by Bose-Einstein statistics (Valenzuela, 2009). Because this study assumes neutral evolution, as the null hypothesis, and compares the inter-site frequency vector of fixations, it is not affected by phylogeny relationships, the method of alignment and assigning the ancestral base, mutation rates, the chosen segment of DNA, the codon position, and other evolutionary conditions. This because any of these conditions is equal for all the sites, regardless of the gene to which they belong or the coding position, as we showed for longitudinal analyses of base sequences (Valenzuela, 2009). It is important to realize that the present study is a pre-transcriptional analysis. In the case of the distribution of fixations among sites with A as the ancestral base, we have: 59.2 expected sites with 0 A; 54.5 expected sites with 1 A; 50.1 with 2 A; … and so on.

Independently, neutral evolution can be tested by the base composition at equilibrium, with the frequency of bases in the DNA or RNA segment, or at disequilibrium by examining the tendency of the frequencies of new fixations in relation to the frequencies at equilibrium, as follows:

II) Analysis of the expected frequencies of bases at equilibrium and disequilibrium

They were obtained by applying the matrix of base mutation (Nei, 1987, see APPENDIX 2). This is the matrix that gives the mutation rate of a base, into any base; for example, the first row shows the rates of mutation of A into A (it remains unchanged), T, G or C. However, we do not have mutation rates, or substitution rates, but fixation rates. Thus, we homologated the fixation matrix with the mutation matrix to profit of its mathematical properties. One of its properties is the expected base frequency at equilibrium, when a large number of base changes have occurred these frequencies are equal for neutral mutation and fixation rates, provided that an exact proportionality between mutation and fixation rates is conserved (this is the neutralist expectation). Since HIV-1 mutates quickly, we may assume that its bases at any site have reached equilibrium. However, to cover all the cases (equilibrium and disequilibrium), we tested the equilibrium base frequencies with the ancestral base frequencies, the observed base frequencies, the longest segment, and the base distribution among new fixations (the tendency at disequilibrium). Appendix 2 shows that fixation rates ranged from 0.01 to 0.09 [f/s/(set of data)]. Note that fixations do not have generation as a dimensional variable like mutations and substitutions. Moreover, viral mutation rates that range from 0.0001 to 0.00001 (m/s/g) demonstrate that mutation rates are very different from fixation rates. Basic population genetics demonstrated that in 10/m (m=mutation rate) generations the equilibrium of gene frequencies is attained with an error less than $5x10^{-5}$ (Valenzuela and Santos 1996). If we assume that HIV-1 has more than 2/m generations per year, in five years the virus sequences should reach the equilibrium of frequencies.

By examining the homo-heterogeneity distribution of fixations, we can perform a third independent test for neutralism and neutral neighbor influence.

III) Testing the homogeneity of new fixations from one ancestral base to the other three in a context of three consecutive bases (neighborhood influence)

For this test, we assumed that the neutralist neighbor influence is true and studied new base fixations in a site within its consecutive context of one upstream and one downstream site (a base triplet). Second, we assumed that most (95% or more) or all mutations are neutral, so their recurrent "mutation rates" are equal to their (recurrent) "substitution rates", as neutralists demonstrated (King and Jukes, 1969; Hey, 1999), and, we add equal to neutral "fixation rates". Third, we assumed that the best estimate of the neutral fixation rate, in a set of sites, is given by the total new fixations from the ancestral base among all the virus strains in the set of nucleotide sites. These three assumptions (besides that of parsimony, for assigning the ancestral base) concede the maximal advantage to the neutralist model, although we have demonstrated they are theoretically and empirically unsupported (Valenzuela and Santos, 1996; Valenzuela, 1997, 2000, 2002, 2007, 2009). Thus, the test consists of comparison between the observed fixations in each site (or subset of sites) to the expected neutralist values estimated from the whole set of fixations. This test is also independent of the alignment method, the chosen DNA segment (or protein), the codon position and phylogenetic relationships. This occurs because we are again comparing the vector of fixations in a site to the vector of the other sites; that is, this is an inter-site comparison, and all these sources of possible differences are equal for all the sites. With the observed new fixated base proportion in the set of sites, we calculated the exact trinomial probability of the distribution (or a more extreme one) in each site of this set. Our procedure is like the one-tailed Fisher's exact test (Maxwell, 1961). We used the log-likelihood ratio χ^2_k test (k=degrees of freedom) to study the heterogeneity of substitution rates among sets of sites with small (<5) expected numbers (Howell, 2002).

We know that it is difficult for a reader, habituated to the standard studies, to understand that our analyses are independent of the method of alignment, assignment of the ancestral base, phylogeny, codon position and any condition (especially transcriptional and post-transcriptional ones) that is equal (neutral) for every site. This is because we are testing neutral or nearly neutral evolution by comparing fixations in a site to the expected vector of fixations estimated from all the sites and, under the neutral or nearly neutral assumptions, all the sites are expected to behave equally for this analysis and for alignment methods. Thus, the expected distribution of neutral or nearly neutral fixations (that is equal to neutral or nearly neutral mutations) is the same for every site. We prepared an example of this procedure based on Appendix 3. It shows 9 consecutive sites with very different vectors of fixations. The five central sites have adenine as an ancestral base. Thus, the three central sites correspond to AAA triplets (equal neighborhood) of consecutive equal ancestral bases. These three central sites show very different fixation vectors. Only site1804 agrees with the expected transition>transversion fixations. The contiguous sites 1805 and 1806 showed more transversions than transitions, but 1805 had more Thymine and 1806 more Cytosine. The χ^2_4 test for the homogeneous distribution of fixations among these 3 consecutive sites was 87.65 ($p<<10^{-6}$). Since these represent 103 different virus strains from 5 continents and 35 countries, the data show evolutionary convergence towards the distribution of fixations in the entire world. The χ^2_8 log-likelihood test (some cells have expected values under 5) for the homogeneous distribution of fixations among the 5 A central consecutive sites was 101.23 ($p<<10^{-6}$). However, the flanking A sites 1803 and 1807 are closer to monomorphism than to polymorphism. If we accept that the mutation and fixation rates are those of the 3 central ancestral A sites, it is not possible to account for the low mutation and fixation rates of the flanking ancestral A sites. If we explain this difference by the neighbor influence of non-A in the sites 1803 and 1807, we cannot account for the heterogeneous distribution of fixations in the 3 central A with the AAA neighborhood. It is evident that the codon position, the precision of alignment, as well as other transcriptional or phylogenetic conditions of these 9 sites, are irrelevant for the refutation of the neutral and nearly neutral theory or the neighbor influence hypothesis.

RESULTS

I) Bose-Einstein analyses

The number of ancestral sites (AS_K), the total number of new fixations (TF_K) and the average number of new fixations per site (AF_K), according to its K base (K subscript) were: for A AS_A = 741, TF_A = 8,524, AF_A = 11.5; for T AS_T = 526, TF_T = 4,561, AF_T = 8.7; for G AS_G = 461, TF_G = 5,246, AF_G = 11.4; for C AS_C = 373, TF_C = 4,608, AF_C = 12.4. Table 1 shows the expected and observed number of sites with the ancestral fixation (0 new fixations), with 1 new fixation, 2 new fixations and so on, until 68 fixations excluding empty (0 observed numbers for the 4 bases) and non-significant rows. For the 4 bases, the general distribution departed greatly from the expected random neutral distribution given by the

Bose-Einstein model. There were a large number of conserved sites for the four ancestral bases (N=0); the significance of this row is sufficient to make the whole table of 69 rows significant (some of them not presented). The number of significant values ($P<0.05$) were: A = 13; T = 16; G = 19; C = 11. For 69 independent comparisons per column, 4 columns of data, at 0.05 level of significance, there were 0.05x69 x 4 = 13.8 expected significant values occurring by random fluctuations; there were 59 significant ones ($\chi^2_1 = 148.0$, $P<10^{-30}$), most of them with probabilities lower than 0.01.

II) The Fixation-Matrix analysis

For neutral evolution, the mutation matrix, as well as the fixation matrix (APPENDIX 2), yielded equal expected frequencies of bases at equilibrium (a mathematical property of these matrices). These equilibrium frequencies were tested against the observed distribution of the ancestral bases, the distribution obtained with the total set of 103 strains, the distribution found in the longest segment and the base distribution found among the new fixated bases that indicates the tendency towards some equilibrium (Table 2). The first comparison is tested by a χ^2_3 with observed and expected numbers given by the frequencies multiplied by the number of ancestral bases (741A + 526T + 461G + 373C = 2101 bases), and the others with their respective numbers. All the comparisons resulted in major deviations from the frequencies expected under neutral equilibrium, because in the first three comparisons, a greater excess of observed A and deficiencies of G and C; T resulted similar to the expected values. New fixated bases showed a highly significant deviation due to a moderate deficiency of A, vast deficiency of T and excesses of G and C. These latter deviations cannot compensate for the former one in relation to the frequencies at equilibrium; they indicate another direction of variation of base frequencies that is also distant from the expected equilibrium frequencies. These three analyses indicate a non-nomological trend; thus they indicate that the base frequency follows a contingent history.

III) Testing the homogeneity of fixation rates among the sites in a fixed triplet context

We analyzed whether the distribution of fixations in each site is randomly (or homogeneously) distributed in comparison to the expected distribution among all the sites (which we assumed is the best neutral estimate, see again Appendix 3). For each of the 4 ancestral bases, there are 16 neighbors given by the four upstream and downstream bases. We studied the 64 triplets, but only AAA, TTT, GGG and CCC are presented, because the others (not shown) had similar patterns of deviations from the neutral expectations.

None of all the 2,105 sites (among the 103 possible bases for each one) had proportions of bases near the expected equilibrium frequencies (Appendix 2) A (31%), T (25%), G (25%) and C (19%). This result is impossible under neutral or nearly neutral evolution, considering that the average number of new fixations per site (AF_K) was over 8.6. In agreement with this result, it was possible to unambiguously assign an ancestral base for 2,103 of the 2,105 sites. The expected numbers of AAA, TTT, GGG, and CCC triplets in 2,103 sites, according to the model strain, were 84.7, 29.7, 27.7 and 11.9, respectively. We had 75 (10 without new fixations), 34 (10 without new fixations), 31 (6 without new fixations) and 21(7 without new fixations), respectively (the significance of the distribution of the number of fixations was performed for the whole segment in analysis I). The only significant excess of CCC was assumed to be due to neighbor influence (to give additional advantage to the neutral theory). Table 3 shows the distribution of new fixations for the original central A, T, G, and C in the AAA, TTT, GGG and CCC context, respectively. Sites were clustered in sets from that with smallest to that with the largest number of new fixations. In total, transitions (tsi) were more frequent than transversions (tve), a well-known expectation. A (a purine) was replaced by new fixations more frequently by G (56%, purine) than by T and C (44%, pyrimidines). T was replaced more by C (56%) than by A and G (44%), G more by A (80%) than by T and C (20%) and C more by T (79%) than by A and G (21%). However, these total fixation rates were extremely heterogeneous among subsets of sites with different numbers of substitutions. The four χ^2 tests were significant and three gave p-values of less than 10^{-5}. Heterogeneities for T and C are dramatic. Three subsets of T sites showed more tve than tsi; the 102-98 subset had 17: 6 (tve: tsi); the 97-93 subset presented 25: 20; the 75-47 subset had 90: 69; the 102-93 subset of C had 19:10. From these subsets we analyzed the most substituted sites of A, T, G and C. Table 4 shows the analysis for Adenine.

Among the 18 sites of the 92-80 ancestral A subset, there were 11 significant sites (trinomial $P<0.022$). All the 16 sites of the 77-43 subset deviated significantly from the expected neutral distribution. There were non-significant sites in the first two subsets (102-93, not shown in Table 4); this does not mean they are not deviated from neutral expectation, because they are deviated in relation to several other sites of the total set (χ^2 analysis). Lack of power, i.e. few substitutions, did not allow a higher significance. The simple inspection of Table

TABLE 1

Expected (Exp) and observed (Obs) number (N) of fixations in a site, according to Bose-Einstein statistics

AB	ADENINE			THYMINE			GUANINE			CYTOSINE		
N	Exp	Obs	P	Exp	Obs	P	Exp	Obs	P	Exp	Obs	P
0	59.2	128	$<10^{-20}$	54.3	164	$<10^{-30}$	37.2	90	$<10^{-20}$	27.9	92	$<10^{-30}$
1	54.5	60	NS	48.7	66	.013	34.2	73	$<10^{-20}$	25.8	43	.007
2	50.1	56	NS	43.7	38	NS	31.4	55	$<10^{-4}$	23.9	20	NS
4	42.4	39	NS	35.1	26	NS	26.6	19	NS	20.4	9	.012
5	39.1	38	NS	31.5	14	.002	24.4	8	.001	18.9	11	NS
6	35.9	21	.013	28.3	14	.007	22.5	13	.045	17.5	9	.042
7	33.1	34	NS	25.3	14	.025	20.6	9	.011	16.2	6	.011
8	30.4	15	.005	22.7	9	.004	19.0	7	.006	15.0	11	NS
9	28.0	21	NS	20.4	7	.003	17.5	7	.012	13.9	6	.034
10	25.8	18	NS	18.3	13	NS	16.1	2	.0004	12.8	10	NS
11	23.7	12	.016	16.4	11	NS	14.8	11	NS	11.9	8	NS
12	21.8	17	NS	14.7	7	.045	13.6	4	.009	11.0	6	NS
14	18.5	13	NS	11.8	6	NS	11.5	2	.005	9.4	4	NS
15	17.0	13	NS	10.6	5	NS	10.5	3	.021	8.7	5	NS
16	15.6	9	NS	9.5	4	NS	9.7	2	.013	8.1	6	NS
17	14.4	11	NS	8.5	5	NS	8.9	2	.025	7.5	3	NS
19	12.2	4	.019	6.9	6	NS	7.5	2	.045	6.4	4	NS
23	8.7	4	NS	4.4	0	.036	5.4	2	NS	4.7	2	NS
32	4.1	9	.016	1.7	1	NS	2.5	2	NS	2.3	1	NS
39	2.3	1	NS	.77	5	.001	1.4	3	NS	1.3	2	NS
41	1.9	4	NS	.62	6	$<10^{-4}$	1.2	2	NS	1.1	4	.026
42	1.8	5	.036	.55	3	.018	1.1	3	NS	1.1	1	NS
45	1.4	5	.014	.40	1	NS	.84	2	NS	.84	2	NS
46	1.3	6	.002	.36	4	.001	.77	4	.008	.78	4	.008
47	1.2	2	NS	.32	1	NS	.71	2	NS	.72	4	.006
49	1.0	7	$<10^{-4}$	.26	1	NS	.60	1	NS	.61	1	NS
50	.92	3	NS	.23	0	NS	.55	4	.002	.57	3	.020
51	.84	3	NS	.21	3	.001	.51	6	$<10^{-4}$	.53	1	NS
52	.77	3	.043	.18	0	NS	.46	4	.001	.49	1	NS
53	.71	1	NS	.17	0	NS	.43	1	NS	.45	4	.001
55	.60	3	.023	.13	2	.008	.36	1	NS	.38	0	NS
56	.55	3	.018	.12	1	NS	.33	1	NS	.36	0	NS
57	.51	0	NS	.11	2	.006	.30	0	NS	.33	0	NS

AB = ancestral base in the site; P = Probability. NS = non-significant. Only significant and non-empty rows are shown.

TABLE 2

Testing the expected equilibrium frequencies calculated with the fixation-matrix method, against the frequencies of the ancestral bases, the total ancestral bases and fixations, the longest segment and new fixated mutations.

	BASES				
	Adenine	Thymine	Guanine	Cytosine	Total
CONDITION			FREQUENCIES		
Equilibrium	0.3094	0.2482	0.2478	0.1946	1.0
			NUMBER OF ANCESTRAL BASES		
Expected	650.1	521.5	520.6	408.8	2,101
Observed	741	526	461	373	2,101
			$\chi^2_3 = 22.71$, P = 0.000046		
			NUMBERS IN THE TOTAL SET OF BASES		
Expected	67,089.3	53,813.3	53,726.1	42,186.3	216,815
Observed	74,711	54,134	48,822	39,148	216,815
			$\chi^2_3 = 1{,}131.36$, P << 10^{-50}		
			NUMBERS IN THE LONGEST SEGMENT		
Expected	812.9	652.0	651.2	511.1	2,627
Observed	901	636	621	469	2,627
			$\chi^2_3 = 14.81$, P = 0.000199		
			NUMBERS AMONG ALL FIXATED MUTATIONS		
Expected	7,097.3	5,693.5	5,684.3	4,463.9	22,939
Observed	6,812	4,374	6,522	5,231	22,939
			$\chi^2_3 = 572.07$, P = <10^{-40}		

TABLE 3

Substitutions of original central bases in AAA, TTT, GGG and CCC triplets of the env gene in 103 HIV-1 sequences.

N_A	NS	A substituted by				N_T	NS	T substituted by			
		T	G	C	Tot			A	G	C	Tot
102-100	13	1	16	11	28	102-98	8	12	5	6	28
99-93	18	7	86	36	129	97-93	6	12	13	20	45
92-80	18	15	186	65	266	87-77	6	5	21	107	133
77-43	16	71	320	272	663	75-47	4	87	3	69	159
Tot N	65	94	608	384	1086	Tot N	24	116	42	202	360
Tot %		8.6	56.0	35.4				32.2	11.7	56.1	
		$\chi^2_6 = 45.61$, P < 10^{-10}						$\chi^2_6 = 113.50$, P < 10^{-20}			
N_G	NS	G substituted by				N_C	NS	C substituted by			
		A	T	C	Tot			A	T	G	Tot
102-95	17	34	7	4	45	102-93	8	7	10	12	29
88-57	8	197	45	3	245	92-56	6	11	102	0	113
Tot N	25	231	52	7	290	Tot N	14	18	112	12	142
Tot %		79.7	17.9	2.4				12.7	78.9	8.4	
		$\chi^2_2 = 6.64$, P = 0.0361						$\chi^2_2 = 52.31$, P < 10^{-9}			

N = number of ancestral bases; NS = number of nucleotide sites

4 shows that each site has its own deviation from the expected neutralist proportion (at the bottom, Tot %). A great deal of sites presented 1 or 2 new fixated bases; although they have enough fixations to have the 3 possible ones. Table 5 presents the case of Thymine, and Table 6 that of Guanine and Cytosine. T, G and C showed the similar pattern found for Adenine. The case of T in the 75-47 subset is remarkable; as is G in the 88-57 subset. It is clear that every site has its own pattern of new fixations.

DISCUSSION

It is evident that mutation rates (m/s/g) are not substitution rates (su/s/g) and substitution rates are not fixation rates (f/s); they are dimensionally different. The neutralist error was to assume that once an allele or base reached a frequency of 100% by drift; it remains fixed (by drift) forever. Thus, the number of different bases found when comparing two or more DNA homologous segments, regardless of the time since they are evolutionary separated, was assumed to be the number of neutral substitutions. They are actually the number of selective fixations, because they are maintained by selection; neutral fixation is impossible (Wright, 1931; Feller, 1951; Valenzuela and Santos, 1996; Valenzuela, 2000, 2002, 2007, 2009). The expected frequency of a base that reached the frequency of 100% in the next generation is (1.0-m), being m the mutation rate. No allele or base can remain fixed, regardless of the population size (even with one individual; Valenzuela, 2000, 2002, 2007, 2008). Moreover, neutralists demonstrated that neutral evolution is independent of the population size (N) (Kimura, 1968; King and Jukes, 1969); as drift depends on N, we should conclude that neutral evolution is independent of drift (by logical laws of transitivity). Neutral evolution (as Brownian motion) occurs inexorably, regardless of the number of individuals (particles). For every nucleotide site of all genomes the expected neutral base distribution is approximately: $^1/_4$ A, $^1/_4$ T, $^1/_4$ G and $^1/_4$ C. Actual mutation rates different from equality do not substantively change this expectation; a six-parameter system of mutation rates conserves the equality of frequencies of A and T, and G and C and a similar figures between both pairs of bases (Sueoka, 1995; Valenzuela, 1997). The demonstration that not only neutral evolution, but also selective evolution, is independent of N and drift is outside of the scope of this article. The expected evolutionary movement by absolute (random) drift is always zero (Valenzuela, 2007).

Our first demonstration is that new fixations (NF), occurring over nearly 20 years, do not randomly distribute in relation to their number. A great proportion of sites did not allow NF. Unfortunately, these sites are separated from one another and do not indicate the existence of possible epitopes for preparing vaccines. Sites that had 0, 1, 2, 3, … n, NF behave heterogeneously; a great number of them are invariant, some are almost invariant, others are moderately variable and some highly variable, but the spectrum is largely deviated from randomness and cannot be produced by neutral or nearly neutral evolution, nor by the neighbor influence of bases (Tables 3 to 6 and Appendix 3). The second demonstration with the expected frequencies of the bases at equilibrium confirmed the first one. It may be argued that HIV-1 has not reached equilibrium frequencies, but this is untenable, because the ancestral distribution of bases (it represents HIV-1 at the origin of human infection), the total distribution of bases and the one of the longest segment are very similar (with the exception of the non-significant T frequency). Thus, if there is any direction of the process of fixations, this is not towards the equilibrium frequencies. Moreover, the distribution of the bases of the new fixations is not the equilibrium distribution either and it is not going towards its own equilibrium. In summary, the structure of the fixation vector at any site shows rather a specific or contingent behavior, which is compatible with pan-selective evolution.

These conclusions are affirmed by the analyses of the frequency vector of new fixations from the ancestral base, site by site, revealed an extraordinary high heterogeneity. It can be summarized as: -each site has its own pattern of new fixations; even though we homogenized by an equal context of 3 consecutive bases to control for a possible neighbor influence-. Thus, this highly significant heterogeneity refutes neutralism, near-neutralism and the neighbor influence hypothesis, in these virus strains. How would this precise and fine pattern of selection for each site and strain be produced? A detailed model including what is known about the interaction between the host defense system and HIV-1 is beyond the objectives of this article. However, this issue has been partially addressed by Moore et al. (2002) and Martin and Carrington (2005). Those authors proposed that this pattern of molecular evolution in HIV occurs because of the highly polymorphic host responses and defenses, based mostly on the innate immune system and adaptive humoral or cell mediated immunity. The variety, force and specificity of these three responses depend mainly on the HLA, immunoglobulins, immunoglobulin-like receptor systems, immunity cell mediated processes and interleukins. HLA is one of the most polymorphic systems involved in the viral antigen recognition through mechanisms performed by

TABLE 4

Substitutions of the original central A in AAA triplets of the env gene among the most substituted 103 HIV-1 sequences

Site	N_A = 92 to 80				Site	N_A = 77 to 43			
	T	G	C	Prob		T	G	C	Prob
1152	1	6	4	0.1067	161	0	26	0	$6x10^{-9}$
2333	0	0	11	$2x10^{-7}$	986	0	7	22	$4x10^{-5}$
2590	0	9	2	0.2259	1416	2	7	20	$9x10^{-4}$
289	3	4	5	0.0030	416	2	20	8	0.0148
1098	2	7	3	0.0700	1431	0	3	29	$4x10^{-9}$
1341	1	7	4	0.1449	2227	0	37	0	$2x10^{-12}$
2558	0	4	8	0.0018	1993	0	22	16	0.0051
1994	0	13	0	0.0096	2767	1	22	19	0.0074
1995	1	11	0	0.0210	1806	6	5	34	$2x10^{-8}$
277	0	14	0	0.0067	1366	0	13	33	$2x10^{-6}$
2644	0	15	0	0.0047	169	0	46	1	$6x10^{-4}$
3170	1	13	2	0.0963	987	9	32	8	$3x10^{-5}$
302	2	15	0	0.0031	1342	7	11	31	$4x10^{-5}$
2421	1	16	0	0.0054	1804	3	28	18	0.0430
2495	0	17	1	0.0116	1805	38	13	4	$<10^{-23}$
121	0	12	7	0.0664	1811	3	28	29	0.0259
651	1	9	10	0.0060	Tot N	71	320	272	663
1417	2	14	7	0.1001	Tot %	10.7	48.3	41.0	
Tot N	15	186	65	266					
Tot %	5.6	69.9	24.5						

N_A = number of the original Adenine; Prob = trinomial probability of this or a more extreme result.

TABLE 5

Substitutions of the original central T in TTT triplets of the env gene among 103 HIV-1 sequences

Site	N_T = 102 to 98				Site	N_T = 87 to 77			
	A	G	C	Prob		A	G	C	Prob
306	0	0	1	0.2609	413	3	0	13	0.0206
1428	1	0	0	0.5217	3108	0	0	19	0.0160
72	0	1	1	0.1606	684	2	11	8	$8x10^{-6}$
2356	3	0	0	0.1420	359	0	3	22	0.1515
2679	3	0	0	0.1420	187	0	0	26	0.0035
2692	0	4	0	0.0022	2739	0	7	19	0.0447
3231	0	0	4	0.0046	Tot N	5	21	107	133
3176	5	0	0	0.0387	Tot %	3.8	15.8	80.4	
Tot N	12	5	6	23					
Tot %	52.2	21.7	26.1						
	N_T = 97 to 93					N_T = 75 to 47			
1391	2	3	1	0.0623	2447	23	1	4	$4x10^{-4}$
1519	0	0	6	0.0077	935	33	0	0	$2x10^{-9}$
2690	2	1	3	0.1831	283	21	2	19	0.0716
893	0	8	0	$5x10^{-5}$	284	10	0	46	$2x10^{-9}$
2702	3	0	6	0.0254	Tot N	87	3	69	159
1346	5	1	4	0.0333	Tot %	54.7	1.9	43.4	
Tot N	12	13	20	45					
Tot %	26.7	28.9	44.4						

N_T = number of the original Thymine

TABLE 6

Substitutions of the original central G in GGG and C in CCC triplets of the env gene among 103 HIV-1 sequences.

Site	N_G = 100 to 95				Site	N_C = 102 to 56			
	A	T	C	Prob		A	T	G	Prob
2486	3	0	0	0.3713	908	0	1	0	0.7887
3216	3	0	0	0.3713	925	1	0	0	0.1268
2775	4	0	0	0.2669	398	0	2	0	0.6221
3003	0	4	0	0.0012	2397	1	0	1	0.0286
2135	5	0	0	0.1918	1169	1	2	0	0.2766
3007	3	2	0	0.1693	1723	2	0	1	0.0061
3002	5	0	3	0.0105	2770	2	4	2	0.0265
Tot N	23	6	3	32	1789	1	1	8	$3x10^{-8}$
Tot %	71.9	18.7	9.4		397	2	9	0	0.1752
	NG = 88 to 57				2796	1	10	0	0.2034
147	15	0	0	0.0380	1788	3	11	0	0.0886
1465	22	0	0	0.0083	2787	0	18	0	0.0140
2242	26	0	0	0.0034	924	2	45	0	$5x10^{-4}$
1953	25	2	0	0.0707	Tot N	18	112	12	142
3078	28	0	1	0.0010	Tot %	12.7	78.9	8.4	
3152	34	1	1	0.0020					
3141	2	42	0	$8x10^{-29}$					
3227	45	0	1	$4x10^{-5}$					
Tot N	197	45	3	245					
Tot %	80.4	18.4	1.2						

N_G = number of the original base Guanine; N_C = number of the original Cytosine.

antigen processing and presenting cells. The force and specificity of this defense are then, depending on HLA and the T cell receptor (or the natural killer cell receptor of the innate immune systems) that is generated by gene rearrangements similar to that of immunoglobulins (millions of possibilities). Thus, an exquisite immunity cell response against most HIV-1 antigens (epitopes) is prepared. B lymphocyte populations can attack the virus with millions of antibodies also made by gene rearrangements. Viral populations are destroyed, but some mutated viruses with new epitopes escape from this extremely varied and specific immunity arsenal. These viruses produce new populations with new antigens that are again attacked by the host immunity system. Viruses that can evade this host defense, or have envelope proteins similar to host proteins, do so by a highly selective set of new fixations. This fine selective HIV-host molecular struggle leads necessarily to a fine specificity of HIV-1 sequences, which tells us the history of adaptive events of viral populations to evade host defenses, instead of the random drift of viral populations. As well, it has been shown that HIV envelop proteins are similar to host proteins (Reiher et al., 1986; Serres, 2001). These studies, as well as those on protein molecular mechanics (Hamacher, 2008) or drug resistance of HIV (Chen et al., 2004), give strong additional evidence to our conclusion that evolution is rather pan-selective (see also Valenzuela, 2009), but do not change our conclusion that holds without these other evidences.

The researcher in molecular evolution could think our analyses miss references to codons with their first, second and third positions, amino acid protein composition and synonymous or non-synonymous substitutions (Ka and Ks), which are the most frequent analyses found in the literature (Nei, 2005). However, most of those studies of transcriptional and post-transcriptional evolutionary stages are somehow biased since they assume a "neutral molecular background" of bases, genetic code, synonymous substitutions and other "constraints" which are not tested in their selective or neutral origin and maintenance. We scanned all the sites, regardless of whether they are the first, second or third position of the code or originate a synonymous or non-synonymous substitution. This scan of the whole DNA segment revealed that most, if not all, the sites (compared with the universe of sites) had their own non-neutral (selective) pattern of pre-transcriptional evolution. Testing transcriptional or

post-transcriptional stages of molecular evolution are unnecessary, because they are affected by the same non-neutral selective pattern. Post-transcriptional analyses should add different evidence on selective processes to this pre-transcriptional study, only when a random distribution of fixations is found. The reader should realize that if neutral mutations occur at random in the first, second and third position, their destiny (fixation, loss or polymorphism) is the same, regardless of their position, the expected frequency vector of mutation is also the same (see APPENDIX 3, where any other reference to transcriptional or post-transcriptional functions is not relevant to show a major deviation from randomness or neutrality). If the actual distribution of fixations in the codon positions is heterogeneous (as has been found in all the studies), then neutral and nearly neutral evolutions are impossible. The present results are in complete agreement with those of a previous study where the longitudinal total sequence of the HIV-1 and the longest *env* segment were analyzed (Valenzuela, 2009). In longitudinal analyses, it is more evident that transcriptional and post-transcriptional functions are not relevant to show deviations from randomness, because they consider only one sequence. Both completely different studies show only one conclusion: evolution of HIV-1 is close to pan-selective evolution. It is very probable that the heterogeneous pattern of fixations is due to the specific adaptive history of that virus strain to the immune system of the host patients, rather than to historical random mutations and substitutions.

REFERENCES

BERNARDI G (1993) The vertebrate genome: isochores and evolution. Mol Biol Evol 10: 186-204.

CHEN L, PERLINA A, LEE CJ (2004) Positive selection detection in 40,000 human immunodeficiency virus (HIV) type1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. J Virol 78: 3722-3732.

CLARK AG (2006) Genomics of the evolutionary process. Trends Ecol Evol 21: 316-321.

CROW JF, KIMURA M (1971) An Introduction to Population Genetics Theory. New York NY: Harper and Row. pp: 257-258.

DRAKE JW (1993) Rates of spontaneous mutation among RNA viruses. Proc Natl Acad Sci USA 90: 4171-4175.

DRAKE JW (1999) The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. Ann N Y Acad Sci 870: 100-107.

DRAKE JW, CHARLESWORTH B, CHARLESWORTH D, CROW JF (1998) Rates of spontaneous mutation. Genetics 148: 1667-1686.

FAUCI AS, LANE HC (2001) Human immunodeficiency virus (HIV) disease: AIDS and Related disorders. In: BRAUNWALD E, FAUCI AS, KASPER DL, HAUSER SL,

LONGO DL, JAMESON JL (eds) Harrison's Principles of Internal Medicine, 15th ed. New York NY: McGraw-Hill. pp: 1864-1865; 1852-1855.

FELLER W (1951) Diffusion processes in genetics. Proceedings of the 2nd Berkeley Symp Math Stat Prob. Berkeley CA: University of California Press. pp: 227-246.

FELLER W (1968) An introduction to probability theory and its applications. New York NY: John Willey & Sons. pp: 38-42.

GATLIN LL (1976) Counter-examples to a neutralist hipótesis. J Mol Evol 7: 185-195.

HAMACHER K. (2008) Relating sequence evolution of HIV1-protease to its underlying molecular mechanics. Gene 422: 30-36.

HEY J (1999) The neutralist, the fly and the selectionist. Tree 14: 35-38.

HOWELL DC (2002) Statistical Methods for Psychology, 5ºEd. Duxbury, Pacific Grove, CA.

JUKES TH (1976) Comments on Counter-Examples to a Neutralist Hypothesis. J Mol Evol 8: 295-297.

JUKES TH, CANTOR CR (1969) Evolution of protein molecules. In Munro HN (ed) Mammalian protein metabolism. New York NY: Academic Press. (cited in LI 1997, pp 59-62).

KARLIN S, MRAZEK J (1997) Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA 94: 10227-10232.

KIMURA M (1968) Evolutionary rate at the molecular level. Nature 217: 624-626.

KIMURA M (1979) The Neutral Theory of Molecular Evolution. Sci Am 241: 94-104.

KIMURA M (1991) Recent development of the neutral theory viewed from the Wrightian tradition of theoretical population genetics. Proc Natl Acad Sci USA 88: 5969-5973.

KIMURA M (1993) Retrospective of the last quarter century of the neutral theory. Jpn J Genet 68: 521-528.

KIMURA M, OHTA T (1977) Further Comments on «Counter-Examples to a Neutralist Hypothesis. J Mol Evol 9: 367-368.

KING JL, JUKES TH (1969) Non-Darwinian evolution. Science 64: 788-798.

KITRINOS KM, HOFFMAN NG, NELSON JAE, SWANSTROM R (2003) Turnover of *env* variable region 1 and 2 genotypes in subjects with late-stage human immunodeficiency virus type 1 infection. J Virol 77: 6811-6822.

KREITMAN M (1996a) The neutral theory is dead. Long live the neutral theory Bioessays 18: 678-683.

KREITMAN M (1996b) Reply to Ohta. Bioessays 18: 683.

LEIGH BROWN AJ (1997) Analysis of HIV-1 *env* gene sequences reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci USA 94: 1862-1865.

LI WH (1997) Molecular Evolution. Sunderland: Sinauer Associates.

MACNEIL A, SANKALÉ JL, MELONI ST, SARR AD, MBOUP S, KANKI P (2007) Long-term intrapatient viral evolution during HIV-2 infection J Infect Dis 195: 726-733.

MANI I, GILBERT P, SANKALÉ JL, EISEN G, MBOUP S, KANKI J (2002) Intrapatient Diversity and its correlation with viral setpoint in Human Immunodeficiency Virus Type 1 CRF02_A/G-IbNG infection. J Virol 76: 10745-10755.

MARTIN MP, CARRINGTON M (2005) Immunogenetics of viral infections. Cur Op Immunol 17510-516.

MAXWELL AE (1961) Analysing Qualitative Data. London: Methuen & Co. Ltd. pp: 46-51.

MOORE CB, JOHN M, JAMES IR, CHRISTIANSEN FT, WITT CS, MALLAL SA (2002) Evidence of HIV-1 Adaptation to HLA-Restricted Immune Responses at a Population Level. Science 296: 1439-1433.

MRAZEK J, KARLIN S (2007) Distinctive features of large complex virus genomes and proteomes. Proc Nat Acad Sci USA 104: 5127-5132.

NEI M (1987) Molecular Evolutionary Genetics. New York, NY: Columbia University Press. pp: 67-70.

NEI M (2005) Selectionism and neutralism in molecular evolution. Mol Biol Evol 22: 2318-2342.

OHTA T (1996) Reply to Kreitman. Bioessays 18: 683.

REIHER III WE, BLALOCK JE, BRUNCK TK (1986) Sequence homology between acquired immunodeficiency syndrome virus envelope protein and interleukin 2. Proc Natl Acad Sci USA 83: 9188-9192.

SERRES PF (2001) AIDS: an immune response against the immune system. Role of a precise tridimensional molecular mimicry. J Autoimmunity 16: 287-291.

SUEOKA N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40: 318-325.

SWOFFORD DL (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts

TRAVERS SAA, O'CONNELL MJ, MCCORMAK GP, MCINERNEY JO (2005) Evidence for heterogeneous selective pressures in the evolution of the *env* gene in different human immunodeficiency virus type 1 subtypes J Virol 79: 1836-1841.

VALENZUELA CY (1997) Non random DNA evolution. Biol Res 30: 117-123.

VALENZUELA CY (2000) Misconceptions and false expectations in neutral evolution. Biol Res 33: 187-195.

VALENZUELA CY (2002) A biotic Big Bang. In: PALYI G, ZUCCHI C, CAGLIOTI L (eds) Fundamentals of Life. Paris: Elsevier France. Pp: 197-202.

VALENZUELA CY (2007) Within selection. Rev Chil Hist Nat 80: 109-116.

VALENZUELA CY (2009) Non-random pre-transcriptional evolution in HIV-1. A refutation of the foundational condition for neutral evolution. Genet Mol Biol 32: 159-169.

VALENZUELA CY, SANTOS JL (1996) A model of complete random molecular evolution by recurrent mutation. Biol Res 29: 203-212.

WRIGHT S (1931) Evolution in Mendelian populations. Genetics 16: 97-159.

YANG Z (2001) Maximum likelihood analysis of adaptive evolution in HIV-1 GP120 *ENV* Gene Pacif Symp Biocomput 6: 226-237.

ZHANG J (2004) Frequent False Detection of Positive Selection by the Likelihood Method with Branch-Site Models. Mol Biol Evol 21: 1332-1339.

APPENDIX 1

The 103 virus sequences

Genbank Accession Numbers of the sequences used in this study:

AB070352,	AF070705,	AF070707,	AF070708,	AF362994,	AY082968,
AF362995,	AJ404325,	AF063224,	AB049811,	AF069933,	AF193276,
AF193277,	AF049337,	AF119820,	AF119819,	AJ288982,	AF069934,
AJ288981,	AX149647,	AY008717,	AF286229,	AF289548,	AF289549,
AF289550,	AF492623,	AJ291718,	AF408630,	AY037279,	AF385935,
AF423756,	AF423757,	AF457079,	AF457089,	AF069673,	AF004885,
AF069670,	U51190,	AF286244,	AF286248,	AF286245,	U86768,
AF457060,	AF457072,	AF286239,	AF457088,	AF457064,	U88823,
AF361879,	AF411964,	L22942,	U36881,	U36883,	U85918,
U85917,	L22952,	AF377957,	U88825,	Y13197,	AF069932,
AF069939,	AF069672,	AF042103,	U08445,	AF112562,	AF112560,
AF408628,	AY037271,	AY037272,	AF110978,	L23065,	AF219269,
AF411966,	U39253,	AY074891,	AF361877,	U88822,	AJ401037,
AJ320484,	L22085,	AF005494,	AF075703,	AJ249236,	AJ249237,
AF377956,	AF061641,	AF069937,	U27445,	AF190127,	AF190128,
AF005496,	AJ401041,	AJ401045,	AF082394,	AJ249235,	AJ249239,
AJ271370,	AJ006022,	L20571,	AJ401038,	AJ401039,	AJ401044,
AY046058.					

APPENDIX 2

FIXATION MATRIX

INPUT BASE	OUTPUT BASE			
	A	T	G	C
A	0.888317	0.021595	0.085315	0.041412
T	0.017373	0.915815	0.011204	0.065489
G	0.068210	0.015043	0.889518	0.013040
C	0.026100	0.047547	0.013963	0.880059

The solution for the first eigenvalue with value 1 yields the matrix M_{ij} [according to Nei (1987)]:

0.111683	-0.021595	-0.085315	-0.041412
-0.017373	0.084185	-0.011204	-0.065489
-0.068210	-0.015043	0.110482	-0.013040
-0.026100	-0.047547	-0.013963	0.119941

The frequencies of A (fA), fT, fG or fC at equilibrium ($_e$) are:

$fA_e = [Cof(M_{11})] / [Cof(M_{11})+Cof(M_{22})+Cof(M_{33})+Cof(M_{44})]$
$fT_e = [Cof(M_{22})] / [Cof(M_{11})+Cof(M_{22})+Cof(M_{33})+Cof(M_{44})]$
$fG_e = [Cof(M_{33})] / [Cof(M_{11})+Cof(M_{22})+Cof(M_{33})+Cof(M_{44})]$
$fC_e = [Cof(M_{44})] / [Cof(M_{11})+Cof(M_{22})+Cof(M_{33})+Cof(M_{44})]$

Where Cof = cofactor of the element (i,j) when i=j, Then

$fA_e = 0.309431$
$fT_e = 0.248199$
$fG_e = 0.247797$
$fC_e = 0.194573$

APPENDIX 3

NINE CONSECUTIVE SITES WITH THEIR FIXATIONS

Site	Adenine	Thymine	Guanine	Cytosine	Ancestral Base
1801	1	0	102	0	Guanine
1802	1	0	101	1	Guanine
1803	96	0	6	1	Adenine
1804	54	3	28	18	Adenine
1805	48	38	13	4	Adenine
1806	58	6	5	34	Adenine
1807	98	1	1	3	Adenine
1808	1	102	0	0	Thymine
1809	40	48	0	15	Thymine