Infection, Genetics and Evolution 10 (2010) 1155–1164

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

An introduction to a novel population genetic approach for HIV characterization

Beatriz Perez-Sweeney [a,*], Rob DeSalle [a], John L. Ho [b,**]

[a] American Museum of Natural History, Sackler Institute for Comparative Genomics, 79th/Central Park West, New York, NY, USA
[b] The Hospitalist Service, Southern Main Medical Center, One Medical Center Drive, Biddeford, Main 04005, USA

ARTICLE INFO

Article history:
Received 15 April 2010
Received in revised form 6 July 2010
Accepted 6 July 2010
Available online 14 July 2010

Keywords:
HIV
Codon
Population genetics
Variation
Selection
Phylogenetic
Haiti
Adaptation
Parallel evolution

ABSTRACT

The rapid evolution of the HIV genome is influenced in part by host selection pressure, which may cause parallel evolution among strains under shared selection pressures. To understand the mechanisms behind HIV-host immune escape across host populations, researchers have compared signatures of positive selection pressure on HIV codons across HIV subtypes and across phylogenetic groups of isolates within major subtypes, all relying on a criterion of phylogenetic separation. The HIV codon sites that retain diversity, evolve convergently among sets of hosts (cohorts) and diverge between cohorts may be phylogenetically undiagnostic (reveal little information about the relationship of the strains) and thus undetectable on a tree. We propose a new approach to characterizing genetic divergence among isolates using existing population genetic methods to better understand HIV response to host selection pressures. The approach combines population genetic statistical methods with codon analysis to identify putative amino acid sites evolving convergently. To illustrate the approach, we compared the C2–V3–C3 region of the envelope protein of HIV-1 clade B isolates between Haiti and USA hosts. This region showed no phylogenetic separation between host populations. Still, we identified codon sites in the C2–V3–C3 HIV-1 region that may have evolved differently between the two host populations. The sites are localized in human leukocyte antigen (HLA) class I binding epitopes, N-glycosylation motifs or both and are limited to the C2 and C3 regions. Our method provides a potential means to reveal candidate sites actively involved in HIV-1 immune escape that would otherwise be missed if a requisite for phylogenetic distinctiveness was made *a priori*. This strategy may prove to be a helpful way to characterize HIV genetic variation among hosts with suspected selection pressure differences, like progressors versus non-progressors.

1. Introduction

HIV-1 evolves in part under adaptive selection pressure presented by the host immune response during the course of infection (Wyatt et al., 1993; Bonhoeffer et al., 1995; Zanotto et al., 1999; Ross and Rodrigo, 2002; Williamson, 2003; Sanjuan et al., 2004). Signatures of this selection pressure are detectable on the HIV genome using codon-based analysis (e.g. Nei and Gojobori, 1986; Yang and Nelson, 2000; Pond and Frost, 2005b). Many genetic studies involving codon analysis have compared HIV subtypes (A through G) to understand the role of selection pressure in HIV evolution (Yang et al., 2003; Choisy et al., 2004; Oliveira et al., 2004; Travers et al., 2005). Other codon-based studies have considered the role of selection pressure on HIV-1 evolution in the same major clade (subtype) but from host populations with phylogenetic distinction (Daniels et al., 2003; Pond et al., 2006). All of these studies relied on cladistic or phylogenetic distinction among isolates as a criterion for addressing their hypothesis. Yet, selection pressure creates nucleotide differences among isolates that may not contribute to phylogenetic separation across cohorts but instead represent homoplasies at one site resulting in shared parallel (convergent) evolutionary sites. These non-phylogenetically diagnostic nucleotide differences among groups of isolates may reflect contrasting evolutionary events specific to host evasion. HIV genetic sequence variation at certain sites have been found to be a product of differential adaptation to host HLA environments resulting in co-evolutionary sites within cohorts (e.g. Moore et al., 2002; Kiepiela et al., 2007; Brumme et al., 2008; Kawashima et al., 2009; Berger et al., 2010). We developed a simple approach to finding putative differentially adapting HIV codons that may be evolving in parallel among certain host individuals, without *a priori* knowledge of HLA population frequencies. In order to characterize amino acid distribution at each site among hosts, the approach employs population genetic statistical methods and treats amino acid

* Corresponding author. Fax: +1 212 313 7819.
** Corresponding author.
E-mail addresses: bperez-sweeney@amnh.org (B. Perez-Sweeney), millennium.john@gmail.com, jho@smmc.org (J.L. Ho).

1567-1348/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2010.07.010

residues as independent units. Once differences in amino acid composition are found, those sites are tested for selection pressure.

We performed a comparative study of an HIV-1 envelope region from Haiti and USA isolates to describe the approach in the context of a contrasting sample set and locus subject to high selection pressure. We illustrate the approach with the aim to promote additional testing and evaluation of the approach, particularly by researchers with relevant proprietary data. Once further substantiated, the approach may prove useful for identifying sites to include in second generation HIV-1 vaccine and may further our understanding of HIV-1 immune escape in different hosts.

The HIV epidemic in Haiti and USA presents several unique features for a comparative analysis of HIV-1 adaptation. First, in terms of race/ethnicity, Haitians are mostly of African decent, which differs from the USA, where early in the AIDS epidemic (before 1994), most cases occurred in Caucasians. Observed differences between African and Caucasian groups in the frequency of the homozygous Δ32CCR5 allele (Samson et al., 1996), in the distribution of known HLA types (Botarelli et al., 1991) and possibly in unidentified HLA types in Africans may provide differential immune selection that manifest as host population specific HIV-1 adaptation. Second, early in the AIDS epidemic (before 1994), HIV was predominantly transmitted by bisexual and/or heterosexual contact in Haiti, while in the USA, male homosexual activities and intravenous drug use were the predominant modes of transmission (Pape et al., 1983). We selected HIV-1 isolates that had been obtained before 1994, a time in which antiretroviral drugs were not extensively used. Thus, the differences in the "life history" of the viruses in the two host populations may be a product of differential host selection pressure from two geographically distinct populations.

We analyzed the C2–V3–C3 region in the gp120 surface envelope (env) glycoprotein for putative differential selection pressure on HIV-1 isolates from Haiti and USA populations. This region mediates viral entry and has been shown to exhibit adaptive changes (de Jong et al., 1992; Fouchier et al., 1992; Shioda et al., 1992; Wyatt et al., 1993; Douek et al., 2003; Jensen et al., 2003; Choisy et al., 2004; Moore et al., 2004). We expected that this envelope region of intense selection pressure and evolution would be of little phylogenetic use, but still offer evolutionary differences between Haiti and USA under our methods. We hypothesized that our method would identify codon sites that may be involved in shared and divergent HIV-1 immune escape strategies in Haitian hosts versus USA hosts that would otherwise be missed if a prerequisite for phylogeographic or cladistic separation were made.

2. Materials and methods

2.1. Data collection

GenBank sequences of the V3 region with part of the conserved flanking blocks, two and three, were downloaded for 38 USA and 30 Haiti isolates collected mostly from blood before 1994 (Supplemental Table 1 for sample list with GenBank accession numbers). Fourteen of the Haitian isolates were from patient samples submitted to the Centers for Disease Control and Prevention by investigators (John L. Ho and Jean W. Pape) from Cornell-GHESKIO (Groupe Haitien d'Etudes du Sarcome de Kaposi et des Infections Opportunistes) Centers, Port-au-Prince. Over 90% of Haitians are of African descent therefore it was assumed that most Haitian samples were from patients of African origin. The sequence used for analysis, excluding gaps, spanned 198 nucleotides and included 60 nucleotides of conserved block 2 and 48 nucleotides of conserved block 3 along with the 90 nucleotides of the V3 region. The data were used to measure "genetic distinction" and "selection pressure" as discussed below.

2.2. Measures of genetic distinction among HIV-1 isolates from Haiti and the USA

Genetic distinction was measured to determine if the C2–V3–C3 region of HIV-1 isolates from Haiti and USA isolates are genetically different and to evaluate whether selection pressure is a contributor to genetic distinction. The two methods used to determine genetic distinction were the phylogenetic method and the Fixation Index estimation for both nucleotide and amino acid data.

2.3. Phylogenetic analysis

Phylogenetic methods were employed to determine the branching patterns of the USA and Haiti isolates and to perform codon selection pressure analysis. Each unique viral isolate was used as a terminal or individual taxon. We added three SIV sequences from *Pan troglodytes* (chimpanzee) as outgroups, and challenged the HIV-1 isolates from Haiti and USA with representatives of seven major HIV-1 sequences (A through G). For nucleotide sequences, three phylogenetic methods, the Neighbor Joining method using the general time reversible distance matrix (GTR), Maximum Parsimony using equal weighting of all characters and the Maximum likelihood method using the best fit rate-matrix model estimated by Modeltest 3.6 (Posada and Crandall, 1998), were used in the program package PAUP (Swofford, 2001). For amino acid sequences, a phylogenetic tree was constructed under the Neighbor Joining Method using raw amino acid differences and under the Maximum Parsimony Method using a successive weighting procedure, which utilized five rounds of re-weighting under the rescaled consistency index as implemented in PAUP (Swofford, 2001). The overall lack of a large number of characters for this analysis (only 66 amino acids and close to 200 nucleotides) limits the robustness of inferences that can be made at the base of the tree. Consequently, only general inferences can be made from these phylogenetic analyses. We used this method to confirm the subtype to assign these sequences and to evaluate phylogenetic clustering of Haiti and USA isolates.

To perform an initial evaluation of the influence of selection pressure on the envelope region, we compared the degree of clustering of HIV-1 sequences to USA and Haiti geographic origins in the nucleotide data-based and the amino acid-based trees and determined if the two types of data yield differences in clustering. Clustering of isolates in the amino acid-based trees but not nucleotide based trees offers one signature of selection pressure (Agosti et al., 1996). In this test, the country of origin was scored as a different character state (Haiti = 1 and USA = 0). We assumed that the amino acid character was gained once and then calculated the probability that the observed clustering pattern was significantly different from random using the concentrated changes test in MacClade (Maddison and Maddison, 2003).

Additional analysis on amino acid data were run using Mr. Bayes, ProtTest and RAxML (Ronquist and Huelsenbeck, 2003; Abascal et al., 2005; Stamatakis et al., 2008, respectively). The best fit data models for tree analysis were determined using algorithms from ProtTest, Hyphy, RAxML and Mr. Bayes. HIVw+I+G and HIVw+G (Nickle et al., 2007) were chosen as the best fit models by Prottest and Hyphy and applied to tree searches under ProtTest. GAMMA-P-Invar with substitution rate matrix JTT was chosen by RAxML and applied in analysis. The WAG model was chosen using the MCMC approach in Mr. Bayes. Five million generations were run, with the first 20% disregarded as burn-in.

2.4. Fixation Index calculation and population divergence

The Fixation Index (Fst), a measure of divergence among predefined groups, was calculated using a nested analysis of molecular variance (covariance) called AMOVA (Excoffier et al., 1992; Excoffier, 2000), in the software package Arlequin v2.0 (Schneider et al., 2000). The significance of Fst is tested using a non-parametric re-sampling permutation of the data to get a null distribution that replaces an ANOVA normal distribution (Excoffier et al., 1992).

To determine if selection pressure is a major component of population divergence, we calculated two additional Fst values using (1) a distance matrix of nucleotides responsible for non-synonymous mutations and (2) a distance matrix of nucleotides responsible for synonymous nucleotide mutations. Pairwise non-synonymous mutation (causes a change in amino acid) and synonymous (cause no change in amino acid) rates were estimated under the Nei and Gojobori method (Nei and Gojobori, 1986) in the PAML package (Yang and Nelson, 2000). The two rate matrices produced were then input into Arlequin for two Fst estimates. We compared Fst calculated for non-synonymous versus synonymous changes.

2.5. Adaptive selection pressure and adaptively divergent amino acids at specific sites

We determined which codon sites on the C2–V3–C3 Env protein among the combined HIV-1 isolates from Haiti and USA was divergent and experiencing adaptive selection pressure. These sites were identified as adaptively diverging sites. We analyzed the adaptively divergent amino acids for selection pressure within the Haiti or USA HIV-1 populations.

Divergence at codon site was measured by testing the codon sites for a statistical difference in amino acid composition. GenAlex (Peakall and Smouse, 2006) was used to summarize amino acid residue composition and frequencies at each amino acid site for each population. A statistical difference in amino acid composition was measured by two tests using the amino acid frequency data set: (1) a test analogous to Analysis of Variance (ANOVA) called Analysis of Molecular Variance (AMOVA), described above (Excoffier et al., 1992) and (2) a test analogous to Fisher's Exact test on an $r \times k$ contingency table called Fisher's Exact Test for Population Differentiation (Raymond and Rousset, 1995). Both AMOVA and the Fisher's Exact Test for Population Differentiation are performed under the null hypothesis of panmixia (a random distribution of amino acids) among the populations, (Excoffier et al., 1992; Raymond and Rousset, 1995).

To identify sites with signatures of positive selection pressure, the Maximum Likelihood approach implemented in "SLAC" (Pond and Frost, 2005b) in the DataMonkey package (Pond and Frost, 2005a) was performed. The HKY85 nucleotide substitution model was selected to account for transition/transversion bias (Hasegawa et al., 1984; Hasegawa et al., 1985) and was used along with the MG94 codon model (Muse and Gaut, 1994). The application of MG94 model with HKY85 model is analogous to F3 × 4 MG model in Yang and Nelson (Yang and Nelson, 2000). SLAC uses a modification of the Suzuki–Gojobori method (Suzuki and Gojobori, 1999) to calculate observed and expected synonymous and non-synonymous substitutions. The rates of non-synonymous and synonymous substitutions (dN and dS, respectively) were calculated from the ratio of observed to expected. To determine if dN was significantly greater than dS, a two-tailed binomial distribution was used and statistical significance was set at $p < 0.05$. Amino acid sites indicative of positive selection are sites showing statistical significance. A trend towards positive selection was acknowledged when $p < 0.1$.

Codon-site analysis was run separately on three Neighbor-Joining gene trees from three HIV-1 datasets consisting of: USA and Haiti combined, USA alone, and Haiti alone. To determine whether codons that are adaptively diverging among populations are under different selection pressures within populations, amino acid sites in the C2–V3–C3 region were compared for signatures of adaptive selection pressure among the HIV isolate from the USA and Haiti. There are three types of selection pressure to consider – adaptive (or positive), purifying and relaxed. Purifying selection occurs when most amino acid variants are detrimental and may be evidenced by fixed or nearly fixed amino acid residues in the population at a given site. Relaxed selection permits variation at a site without functional consequence. There are three possible outcomes per codon expected (Fig. 2): (1) both populations at a codon exhibit adaptive selection pressure (+,+), (2) one of the two populations at a codon exhibits adaptive selection pressure, ((+,−) or (−,+)) and (3) neither population at a codon exhibits adaptive selection pressure (−,−). The outcomes were interpreted as follows. For a given adaptively divergent site, if both populations exhibit adaptive selection pressure (+,+), then HIV was diverging under positive selection pressures in the USA and Haiti at the codon site. For a given adaptively divergent site, if one population does not show adaptive selection pressure as in cases 2 (+,−; −,+) and 3 (−,−) that population may be either under purifying or relaxed selection. Purifying selection is considered if the population has little to no amino acid variation at the codon. Relaxed selection pressure is considered if the population has amino acid variation at the codon in the absence of adaptive selection pressure. Alternatively, it may be that samples sizes were insufficient to detect adaptation. Insufficient sample size is considered in cases where the population's dN/dS ratio is nearly 1.0 or if the dS = 0 and the dN value is not negligible.

2.6. Functional relevance of adaptively divergent sites

To determine whether specific amino acid sites are associated with known biological function, we mapped the amino acid sites to known epitopes that bind to HLA (human leukocyte antigens) variants and putative glycosylation motifs. HLA epitopes were located on population consensus sequences using motif scan on the web (http://www.hiv.lanl.gov/content/immunology/motifscan/motifscan). Only epitopes with experimental evidence of HLA binding were used. The program N-Glycosylate at http://www.hiv.lanl.gov/content/hiv-db/GLYCOSITE/glycosite.html was used to locate likely N-glycosylation regions.

3. Results

3.1. Overall genetic divergence among Haiti and USA

Three different tree building approaches (Maximum Likelihood, Neighbor-Joining and Parsimony) were used to evaluate phylogenetic signal in the C2–V3–C3 *env* region of HIV-1 from Haiti and USA. HIV-1 sequences from both countries were clade B associated as expected (Fig. 1 and Supplemental Fig. 1A and B). Neither HIV-1 isolates from Haiti nor from the USA were monophyletic within clade B for this sequence fragment (Fig. 1). The nucleotide and amino acid data both provided phylogenetic signal and resolution under most analysis, but did not provide phylogeographic separation between Haiti and USA strains.

Nucleotide data by three phylogenetic methods did not reveal clustering using the concentrated changes test, $p = 0.0$ for all three phylogenetic methods (Fig. 1(A) and Supplemental Fig. 1A). Amino acid data analysis via Mr. Bayes yielded unresolved trees. Analysis of amino acid data under ProtTest or RAxML yielded resolved trees but no clustering of Haiti versus USA env sequence

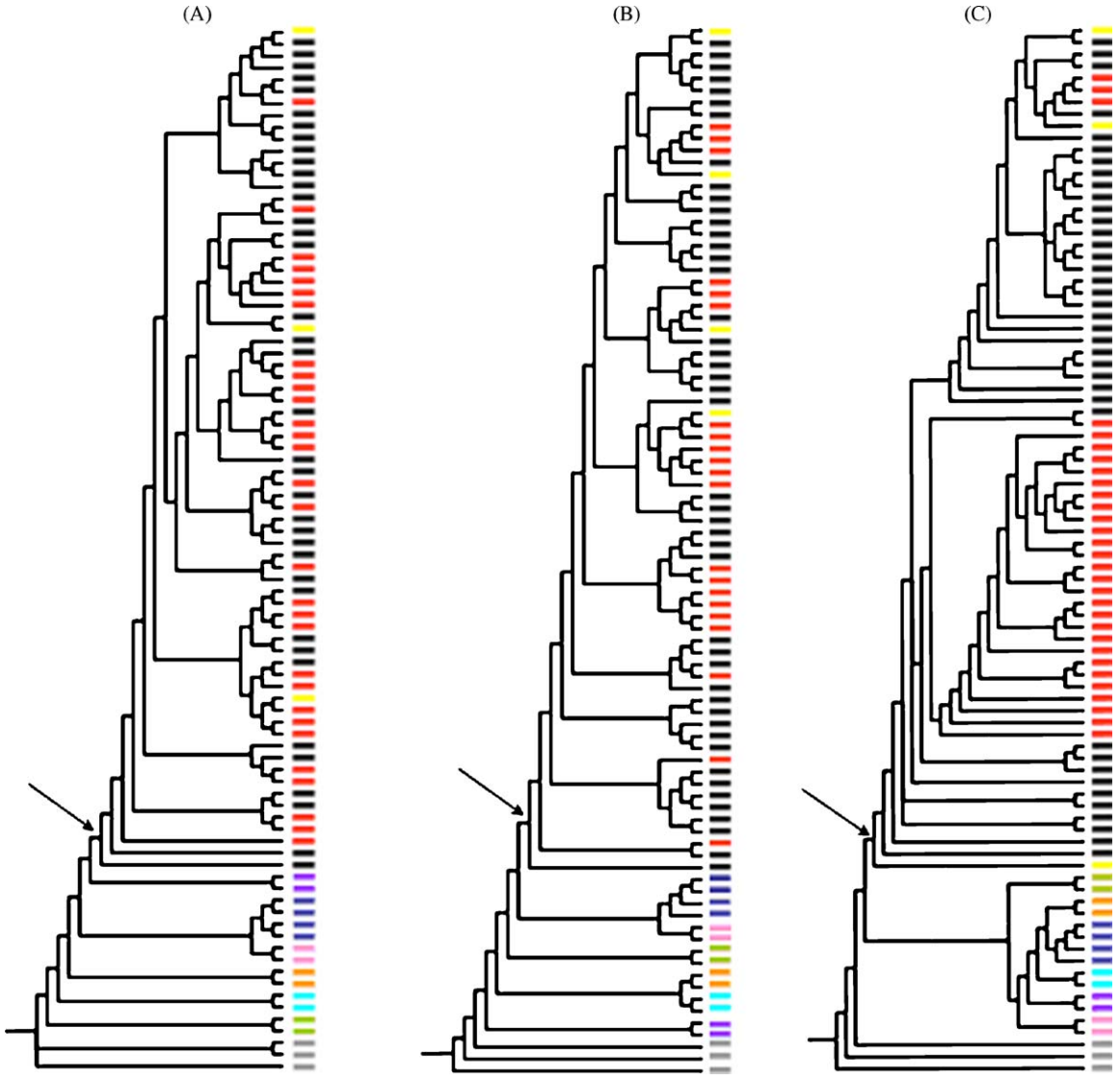

Fig. 1. Cladograms of the C2–V3–C3 envelope region of the HIV-1 from Haiti and the USA. Phylogenetic trees of the HIV-1 C2–V3–C3 region produced from Neighbor Joining and Maximum Parsimony Methods. In all three trees (A through C) HIV-1 isolates from Haiti (in red) clustered into the same large B clade as HIV-1 isolates from the USA (in black) and reference B clade (in yellow) isolates. The arrow indicates the node of the B clade. Reference sequences for clades A and C through G isolates are color coded as: A, green; C, orange; D, purple; E, turquoise; F, blue; G, pink; Chimpanzee SIV isolates, gray. The corresponding name of each isolate, accession number and citation/reference are provided in Supplemental Materials: Supplemental Fig. 1 and Supplemental Table 1. (A) Analysis of nucleotides sequences using the Neighbor Joining Method (for analysis with Maximum Likelihood and Maximum Parsimony Methods, see Supplemental Materials). The Haiti HIV-1 isolates are interspersed with the USA HIV-1 isolates. (B) Analysis of amino acid sequences using the Neighbor Joining method. Significant clustering of Haiti and USA isolates occurred within clade B ($p = 0.99$, concentrated changes test, MacClade). (C) Analysis of amino acid sequences used the Maximum Parsimony Method. Significant clustering of Haiti and USA isolates occurred within clade B ($p = 0.96$, concentrated changes test, MacClade). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

sets. These incongruent amino acid trees showed unusual subtype relationships unlike nucleotide and amino acid trees under the reported phylogenetic methods (trees available upon request). Additional amino acid characters are likely needed under these maximum likelihood and Bayesian approaches for reliable analysis of our approach. These trees were thus not included in the overall discussion of our approach. In contrast to nucleotide data, statistically significant clustering under Neighbor Joining and Maximum Parsimony of Haiti isolates was observed in the trees constructed from amino acid data ($p = 0.99$ for Neighbor Joining Method and $p = 0.96$ for Maximum Parsimony Method, Fig. 1(B) and (C)). Such a contrast in phylogenetic signal between nucleotide and amino acid clustering has previously been suggested as a possible product of different selection pressures between these two data levels (Agosti et al., 1996).

To further assess genetic divergence, the Fixation Index (Fst) was calculated. The HIV strains were statistically different between Haiti and USA (Fixation index = 0.03, $p < 0.01$ with AMOVA). A test was run to determine if the small difference was a product of selection pressure. Fst was measured between non-synonymous (causing amino acid changes) and synonymous changes (no amino acid changes). The Fixation Index measurements of the C2–V3–C3 sequence of HIV-1 from Haiti and USA showed statistically significant divergence for non-synonymous mutations (Fst = 0.036, $p < 0.01$); in contrast to no statistically significant divergence for synonymous mutations (Fst = 0.016, $p < 0.06$). The difference in divergence between synonymous and non-synonymous nucleotides further supports selection pressure acting at the amino acid level. The population divergence observed, based on the Fixation Index, largely reflects the differences in amino acid frequencies between viral isolates from Haiti and the USA (Table 1).

Table 1
List of codon sites showing adaptive divergence and the percentage of amino acids found in HIV-1 isolates from USA and Haiti.

Site	Amino acid	United States of America	Haiti	Site	Amino acid	United States of America	Haiti
4	A	–	3.3	53	E	2.6	3.3
	D	68.4	36.7		G	2.6	13.3
	G	–	3.3		I	5.3	10.0
	K	–	3.3		K	18.4	3.3
	N	31.6	36.7		R	71.1	56.7
	S	–	13.3		S	–	10.0
	Y	–	3.3		T	–	3.3
6	A	89.5	83.3	61	A	–	6.7
	I	2.6	–		D	–	3.3
	Q	–	3.3		E	10.5	6.7
	S	–	3.3		G	–	20.0
	T	–	10.0		H	–	3.3
	V	7.9	–		I	–	3.3
15	A	5.3	6.7		K	63.2	26.7
	D	2.6	3.3		L	–	6.7
	E	81.6	56.7		N	2.6	6.7
	K	5.3	10.0		Q	10.5	6.7
	N	–	6.7		R	13.2	10.0
	Q	–	3.3				
	R	–	3.3				
	T	2.6	10.0				
	V	2.6	–				

3.2. Adaptive selection pressure and divergence at specific codons

A partial listing of the frequencies of various amino acids at specific sites in the C2–V3–C3 is shown in Table 1. Differences in amino acid composition between USA and Haiti HIV-1 were evaluated at each residue site for amino acid divergence using AMOVA (Excoffier et al., 1992) and Fisher's Exact Test for Population Differentiation (Raymond and Rousset, 1995). Of importance, AMOVA and Fisher's Exact Test for Population Differentiation were performed under the null hypothesis of panmixia or a random distribution of amino acids among the population. Codon sites were tested for evidence of adaptive selection pressure using SLAC (Pond and Frost, 2005b). Sites with both amino acid divergence and evidence of adaptive selection pressure were considered adaptively divergent sites. Five sites showed both adaptive selection pressure ($p < 0.05$, Fig. 3) and amino acid divergence as shown by AMOVA ($p < 0.05$) and Fisher's Exact Test for Population Differentiation (Table 1). Therefore, these five amino acid sites were considered adaptively divergent sites. All five adaptively divergent sites were located in the conserved sequence blocks flanking the variable 3 region of the Env protein (Fig. 3). The five adaptively divergent sites were further evaluated for selection pressure within each of these two HIV-1 populations (Fig. 3). One of the five codons (site 61) exhibited evidence of adaptive selection pressure in both Haiti and USA HIV-1 populations (Case 1, (+,+); Fig. 2). The remaining four codons (sites 4, 6, 15, and 53) in C2–V3–C3 *env* showed adaptive selection in only one of the two HIV-1 populations (Case 2, (−, +) or (+, −); Fig. 2).

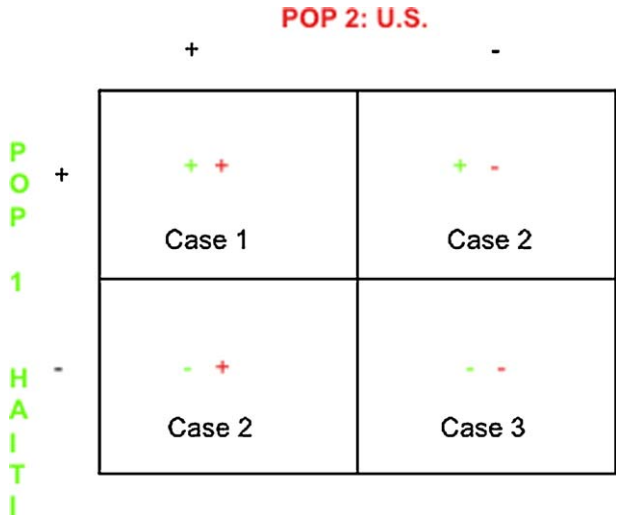

Fig. 2. Outcomes in the comparison of selection pressure. Illustrated are possible outcomes from the analysis of adaptive selection pressure between HIV-1 isolates from Haiti and the USA. The symbols '+' and '−' indicate evidence or absence of adaptive selection pressure, respectively. Evidence for adaptive selection pressure is observed, in both populations (Case 1), in one of the two populations (Case 2) or in neither population (Case 3).

3.3. Functional relevance of adaptively divergent codon sites

The adaptively divergent codon sites identified in the C2–V3–C3 *env* region were mapped onto immune functional domains; all were involved with either HLA class I binding epitope and N-glycosylation motifs or both. Four of the five adaptively divergent codons (site 4, 6, 53 and 61) were located in two HIV-1 Env epitopes known to bind to HLA class I, A24 and B35 (Fig. 3). The fifth adaptively divergent site flanked epitopes that bind the B35 and B18/B40 HLA class 1 variants. A total of eight N-glycosylation motifs were identified among the USA and Haiti strains; all of these sites lie within HLA binding epitopes. Four glycosylation motifs (motif regions 1, 2, 5 and 6) contained an identified adaptively divergent site (4, 15, and 53, respectively, Fig. 3).

4. Discussion

4.1. Detection of divergent sites under putative positive Darwinian selection

Our analysis identified genetic differences in the HIV-1 C2–V3–C3 envelope region among Haiti and USA isolates obtained before 1994 despite a lack of phylogeographic separation. Adaptive selection pressure appeared to largely account for the detected difference between HIV-1 C2–V3–C3 *env* sequence from Haiti and USA as indicated by phylogenetic, population genetic and codon analyses. The data also suggest that specific codon sites experience adaptive divergence that may also have functional importance.

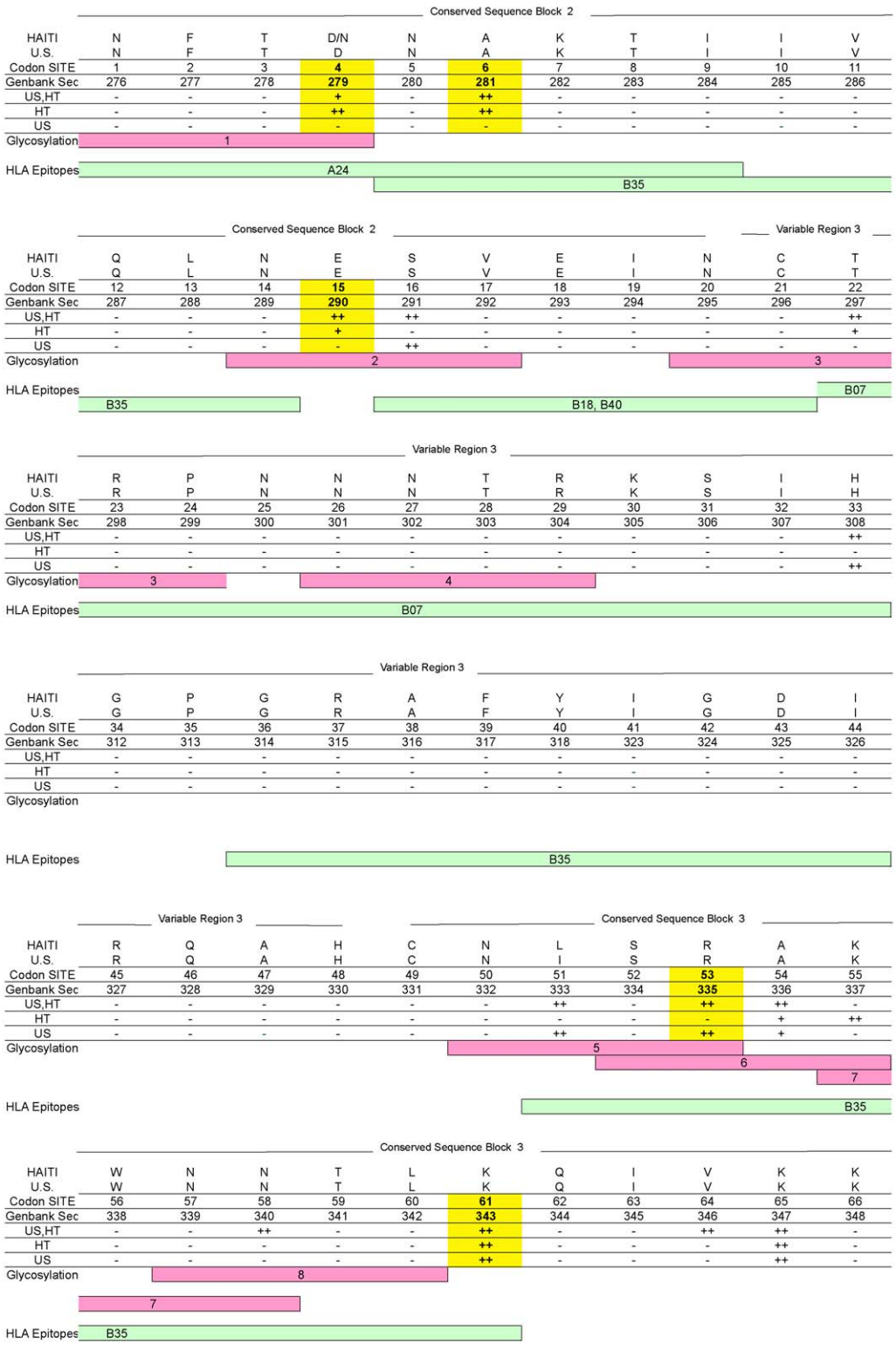

Fig. 3. Selection pressure in the C2–V3–C3 region of HIV-1 within and among isolates from Haiti and the USA. HIV-1 C2–V3–C3 consensus amino acid sequence and codon or amino acid sites with and without adaptive selection pressure are shown for HIV-1 isolates from Haiti and the USA. The codon site number is the amino acid position along the amino acid sequence. "GenBank seq" number corresponds to the GenBank amino acid site of the *env* reference sequence (accession number K03455). Codons in which dN was significantly greater than dS at $p < 0.05$ are marked with '++', at $p < 0.1$ are marked with '+', and '–' were sites that lacked significance. Codon sites showing statistically significant dN > dS values were designated as sites with adaptive selection pressure. Blocked in yellow are five codon sites showing adaptive selection pressure and a statistically significant difference in amino acid composition ($p < 0.05$ AMOVA and Fisher's Exact Test for Population Differentiation). These five amino acid sites were designated as adaptively divergent codons sites. Four of the five adaptively divergent codon sites mapped to the HLA binding epitopes (aquamarine) and the fifth flanks two HLA binding epitopes. Three of the five adaptively divergent codon sites mapped to N-glycosylation motif (pink). Glycosylation motif 1 was present in nearly all isolates while the other motifs were present in two or more isolates. Putative HIV epitopes found in the USA and Haiti consensus sequences that bind to specific HLA alleles were identified using http://www.hiv.lanl.gov/content/immunology/motifscan/motifscan. Only epitopes with experimentally supported binding to HLA were included. Putative N-glycosylation regions were numbered consecutively and were found for each isolate using the program N-Glycosylate, http://www.hiv.lanl.gov/content/hiv-db/GLYCOSITE/glycosite.html. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

A near to complete phylogeographic separation between Haiti and USA for nucleotide data of under ten HIV-1 strains from Haiti was previously reported when a larger portion of the HIV-1 genome was analyzed (Li et al., 1988; Korber et al., 2000; Gilbert et al., 2007). Our phylogenetic analysis of the C2–V3–C3 *env* nucleotide nucleotide sequence placed HIV-1 isolates from Haiti and the USA in clade-B as expected from their GenBank designations, but without phylogeographic clustering, as expected because of the short gene sequence read and high variability. Despite the lack of C2–V3–C3 env Haiti/USA phylogeographic distinction on the nucleotide trees, we found a statistically significant clustering between Haiti and USA isolates in the amino acid phylogentic analysis and population genetic divergence of dN but not dS mutations between the Haiti and the USA isolates. The above results taken together, (1) a paradoxically high level of phylogeographic clustering of amino acids as opposed to nucleotides and (2) the population genetic differentiation of non-sysnomymous mutations as opposed to synonymous mutations between HIV-1 isolates from Haiti and the USA, suggest that divergence among HIV-1 sequences may be largely due to host selection pressure.

The methods employed in this study revealed five putative adaptively divergent sites between the HIV-1 isolates from the two host groups (codon positions 4, 6, 15, 53 and 61). Site 61 appeared to be under adaptive selection pressure for both the Haitian and USA populations. Four of the five adaptively divergent sites (at positions 4, 6, 15 and 53) differed in selection pressure between HIV-1 isolates from Haiti and the USA. The lack of adaptive selection pressure signature at sites 4, 6, and 15 in the HIV-1 isolates from the USA compared to those from Haiti (Fig. 3 and Table 1) may likely be a product of purifying selection in the USA. This is evident by the lower amount of variation in amino acids at these sites in the USA compared to Haiti (Table 1). Conversely, the lack of adaptive selection pressure signature at site 53 in Haitian isolates may be a product of small sample size or relaxed selection pressure in Haiti given the high variance in amino acid composition in Haitian isolates at this site (Fig. 3 and Table 1).

4.2. Functional significance of divergent sites under positive selection

Several studies show that sites with signatures of adaptive selection pressure have evolved as a means to escape immune elimination by neutralizing antibodies, by T-helper cells or by cytotoxic lymphocytes (Phillips et al., 1991; Ogg et al., 1998; Kelleher et al., 2001; Ross and Rodrigo, 2002; Yusim et al., 2002; Richman et al., 2003; Wei et al., 2003). To discern whether the adaptively divergent sites among the HIV-1 isolates from the USA and Haiti have functional significance, we mapped the C2–V3–C3 amino acid sites to known HLA class I binding epitopes and putative N-glycosylation motifs. These putative functional domains have been suggested targets for immune selection (Phillips et al., 1991; Wyatt et al., 1993; Ogg et al., 1998; Zanotto et al., 1999; Kelleher et al., 2001; Ross and Rodrigo, 2002; Yusim et al., 2002; Richman et al., 2003; Wei et al., 2003; Sanjuan et al., 2004; Yamaguchi and Gojobori, 1997). All five adaptively divergent sites found here map to putative HLA epitopes making it plausible that these sites may in fact be responding to immune selection. An interplay between host immune response and variation in HIV-1 epitopes was suggested by the strong association found between human HLA class I types and HIV-1 epitope variants that escape binding to the cognate HLA class I types (Moore et al., 2002; Brumme et al., 2008; Kawashima et al., 2009; Berger et al., 2010).

Three adaptively divergent sites found here also map to putative glycosylation sites. Glycosylation alters the conformation of the HIV-1 envelope protein and has been shown to prevent recognition by T cell receptor or binding by neutralizing antibody (Botarelli et al., 1991; Hwang et al., 1991; Papandreou et al., 1996; Pollakis et al., 2001; Wei et al., 2003). Although the adaptively divergent sites did not appear to alter gain or loss of glycosylation, it is possible that they may obscure the antigenicity of these areas because they also lie within HLA epitopes (Botarelli et al., 1991).

Interestingly, all five adaptively divergent sites were limited to the conserved regions surrounding the V3. Our findings are in agreement with prior studies that noted increased levels of variation at the V3 conserved flanking regions compared to the V3 region (Dighe et al., 1997) and a high number of positively selected sites in those conserved regions (Yamaguchi-Kabata and Gojobori, 2000). A study of 25 clade B and 25 clade C HIV-1 isolates by Gaschen et al. (2002) reported that the C3 flanking region showed a higher concentration of strong-positive selected sites for clade C isolates compared to clade B isolates. These findings led Gaschen et al. (2002) to suggest that HIV-1 clades B and C may be evolving differently and should perhaps be treated differently in vaccine design. Our analysis of adaptive selection pressure evaluated clade B HIV-1 isolates from two divergent host populations using sample size of HIV-1 isolates similar to that of Gaschen and colleagues' study. We found specific codons that may be differentially adapting to two host populations and our ability to do so at the sub-phylogenetic level may be due to our choice of HIV-1 isolates from two genetically different host populations.

4.3. Data set limitations

We compared HIV from Haiti to HIV from USA to illustrate our methodological approach, but we recognize that the interpretation of the results must be taken with caution because of limitations the sample set presents. Firstly, the sample size for population comparisons 30 HIV-1 isolates from Haiti and 38 isolates from USA and the sequence lengths are relatively limited 198 nucleotides, 66 amino acids. However, this locus has critical functions for HIV-1 and host interactions enhancing the opportunity for discovering differences even in a limited data set. These samples may have been derived over several years as opposed to a small window of time (i.e. 1 year); even though the samples from GHESKIO were derived over several months time period. These factors may lead to false detection of positively selected sites. Knowing these potential limitations, we utilized the recommended methods for smaller sample sizes (Pond and Frost, 2005b). Conversely, the limited size of the locus evaluated obviated the potential to identify the full dimension of positively selected sites and amino acid divergence, leaving other adaptively divergent sites undetected. Secondly, we evaluated some patients without knowledge of their date of infection and/or CD4 counts (a marker for HIV disease/AIDS stage). Thus, the sample set may be a mixture of isolates from patients infected more recently and patients infected in the distant past and therefore, vary in their stage of disease. Since the HIV strain of a recently infected individual is likely to differ from that of a 'distantly' infected patient (Rambaut et al., 2004; Salazar-Gonzalez et al., 2009; Goonetilleke et al., 2009), the mixing of viral sample types from different stage of HIV-1 disease may complicate the study by adding more variance, creating noise that would prevent identification of adaptively selected sites. Our finding of adaptively divergent sites likely represents HIV-1 isolates obtained mostly from patients with significant HIV disease or AIDS because in the pre-1994 era most patients were typically diagnosed at AIDS onset and few were discovered as part of screening of healthy individuals or those with acute HIV-1 infection. This was certainly the case of isolates obtained from GHESKIO. Lastly, while larger HIV-1 genetic regions indicate some level of cladistic distinction between USA isolates using a small sample set of less than 10 Haitian isolates (Li et al., 1988; Korber et al., 2000; Gilbert et al., 2007), we did not, as

expected, find phylogenetic distinction among the 30 Haiti and 38 USA HIV-1 isolates using a limited sequence of 198 env nucleotides. As a result we did not determined levels of viral exchange between Haiti and the USA.

Despite these shortcomings, our study of these HIV-1 samples show the potential utility of a novel approach for studying adaptive selection pressure among groups of isolates using biologically, functionally and evolutionarily important HIV-1 regions that are phylogenetically not diagnostic or where phylogenetic diagnosis of populations may not be present. We found that even though the C2–V3–C3 *env* region of HIV-1 isolates from Haiti and USA occupy the same clade B lineage, the application of population genetics and statistical analysis allowed the identification of putative adaptively divergent codon sites at putative immune domains that differ between the two human hosts.

4.4. Implications for vaccine design

The incorporation of regions with adaptive selection pressure (Oliveira et al., 2004) and sequence variation in a cocktail design has been proposed as an alternative vaccine strategy to the failed consensus sequence-based vaccines (Gaschen et al., 2002; Slobod et al., 2005). One of the most recent vaccine efforts designed mosaic immunogens by performing in silico recombination of natural HIV sequence strains found in populations. These mosaic vaccines showed improved cellular or CD8 T cell repertoire compared to consensus sequence vaccine when tested in Rhesus macaques and support that inclusion of antigenic variants can elicit better immune response than consensus sequences (Barouch et al., 2010; Corey and McElrath, 2010; Santra et al., 2010). Along these lines, our approach may potentially identify optimal antigenic candidates present in HIV-1-infected populations for vaccine design. In our preliminary analysis, the sites evolving divergently under adaptive selection pressure bind to major histocompatibility antigens [MHC] needed to generate effective T cell response and were differentially undergoing selection in two host populations, possibly provoking differential host immune response.

HIV variation is maintained by the variation present in host immune response, which provides the selection backdrop for HIV evolution. Creating vaccines that can capture or reflect HIV variants subject to the selection pressure of host genetic variants has been advocated by several groups (e.g. Kiepiela et al., 2007). The use of diversity capture methods that include consideration of within-clade diversity has also been encouraged (Yang, 2008). Our method of identifying HIV adaptively divergent sites within and among sequence sets may provide much needed information for choosing portions of sequences to include in designing a preventative or therapeutic HIV-1 vaccine. Inclusion of such sites in a HIV-1 vaccine may increase the chances of an immune induced bottleneck in HIV-1 replication to thereby, abort infection dissemination or delay HIV-1 disease progression.

4.5. Summary and future research

We presented, in this study, an approach to characterize HIV-1 variants by focusing on adaptive sites specific to certain hosts. Our approach identified potential adaptively divergent sites among two host populations defined by geographical distance and host genetic difference. Studies with larger sample sizes evaluating HIV-1 isolates from Haiti and the USA from the current era in which more gene regions are analyzed are needed to confirm these preliminary findings. Moreover, the adaptively selected sites should also be evaluated for putative immune escape function.

Although some focus has been placed on revealing adaptations that are transmitted from host to host (population level adaptations, e.g. Pond et al., 2006), we think characterizing unique adaptive mutations shared among HIV isolates that are not necessarily phylogenetically transmitted is equally important. Such shared unique mutations are a result of parallel (“convergent”) evolution among hosts in the population. Our approach does not differentiate between transmitted versus parallel adaptive mutations and we do not consider this a limitation but instead a complimentary way to characterizing HIV variation.

Our next steps are to test the power and precision of this approach using simulated data sets and isolates from hosts with known immune (e.g. HLA) environments under different conditions of geneflow and recombination. In these tests, we will also include the less commonly used non-phylogenetic methods for dN/dS codon analysis (Nei and Gojobori, 1986 and Ota and Nei, 1994) since our results (data not shown) under such analysis located additional sites with selection pressure. If substantiated, the application of population genetic statistical methods to study HIV-1 loci and strains from different host populations may help us further reveal patterns of HIV-1 evolution and uncover HIV-1 sequence sites applicable for vaccine design.

Acknowledgements

This research was supported by NIH Fogarty International Center AIDS International Training in Research Program D43 TW00018 (WDJ) and D43 TW00018-S3 (WDJ and JLH), U2R TW006901 (WDJ), NIH 5RO1 HL61960 (JLH), NIH R21 AI62332 (JLH) and the Minority Supplement Award R01 HL61960S (BPS, JLH). We thank Drs. Dan Fitzgerald and Warren D. Johnson for their support and and early discussions, the patients involved in the study from USA and Haiti and especially patients from GHESKIO, Dr. Jean W. Pape for making the Haiti HIV samples available and Centers for Disease Control and Prevention (Chin-Yi Ou, Marsha Kalish) and other investigators for their submission of the HIV sequences to GenBank used for these analyses. We also thank Dr. Andrea Gibson for composition of S Table 1. RD thanks the Lewis B and Dorothy Cullman Program in Molecular Systematics, the Sackler Institute for Comparative Genomics and the Korein Family Foundation at the American Museum of Natural History. We thank an anonymous reviewer for valuable comments that improved the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2010.07.010.

References

Abascal, F., Zardoya, R., Posada, D., 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Agosti, D., Jacobs, D., DeSalle, R., 1996. On combining protein sequences and nucleic acid sequences in phylogenetic analysis: the homeobox protein case. Cladistics 12, 65–82.

Barouch, D.H., O'Brien, K.L., Simmons, N.L., King, S.L., Abbink, P., Maxfield, L.F., Sun, Y.H., La Porte, A., Riggs, A.M., Lynch, D.M., Clark, S.L., Backus, K., Perry, J.R., Seaman, M.S., Carville, A., Mansfield, K.G., Szinger, J.J., Fischer, W., Muldoon, M., Korber, B., 2010. Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. Nat. Med. 16, 319–323.

Berger, C.T., Carlson, J.M., Brumme, C.J., Hartman, K.L., Brumme, Z.L., Henry, L.M., Rosato, P.C., Piechocka-Trocha, A., Brockman, M.A., Harrigan, P.R., Heckerman, D., Kaufmann, D.E., Brander, C., 2010. Viral adaptation to immune selection pressure by HLA class I-restricted CTL responses targeting epitopes in HIV frameshift sequences. J. Exp. Med. 207, 61–75 S61-12.

Bonhoeffer, S., Holmes, E.C., Nowak, M.A., 1995. Causes of HIV diversity. Nature 376, 125.

Botarelli, P., Houlden, B.A., Haigwood, N.L., Servis, C., Montagna, D., Abrignani, S., 1991. N-glycosylation of HIV-gp120 may constrain recognition by T lymphocytes. J. Immunol. 147, 3128–3132.

Brumme, Z.L., Brumme, C.J., Carlson, J., Streeck, H., John, M., Eichbaum, Q., Block, B.L., Baker, B., Kadie, C., Markowitz, M., Jessen, H., Kelleher, A.D., Rosenberg, E., Kaldor, J., Yuki, Y., Carrington, M., Allen, T.M., Mallal, S., Altfeld, M., Heckerman, D., Walker, B.D., 2008. Marked epitope- and allele-specific differences in rates of

mutation in human immunodeficiency type 1 (HIV-1) Gag, Pol, and Nef cytotoxic T-lymphocyte epitopes in acute/early HIV-1 infection. J. Virol. 82, 9216–9227.
Choisy, M., Woelk, C.H., Guegan, J.F., Robertson, D.L., 2004. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J. Virol. 78, 1962–1970.
Corey, L., McElrath, M.J., 2010. HIV vaccines: mosaic approach to virus diversity. Nat. Med. 16, 268–270.
Daniels, R.S., Kang, C., Patel, D., Xiang, Z., Douglas, N.W., Zheng, N.N., Cho, H.W., Lee, J.S., 2003. An HIV type 1 subtype B founder effect in Korea: gp160 signature patterns infer circulation of CTL-escape strains at the population level. AIDS Res. Hum. Retroviruses 19, 631–641.
de Jong, J.J., Goudsmit, J., Keulen, W., Klaver, B., Krone, W., Tersmette, M., de Ronde, A., 1992. Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. J. Virol. 66, 757–765.
Dighe, P.K., Korber, B.T., Foley, B.T., 1997. Global Variation in the HIV-1 V3 Region. Los Alamos National Laboratory, Los Alamos.
Douek, D.C., Picker, L.J., Koup, R.A., 2003. T cell dynamics in HIV-1 infection. Annu. Rev. Immunol. 21, 265–304.
Excoffier, L., 2000. Analysis of population subdivision. In: Balding, D., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley and Sons, Ltd..
Excoffier, L., Smouse, P.E., Quattro, J.M., 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131, 479–491.
Fouchier, R.A., Groenink, M., Kootstra, N.A., Tersmette, M., Huisman, H.G., Miedema, F., Schuitemaker, H., 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. 66, 3183–3187.
Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B.H., Bhattacharya, T., Korber, B., 2002. Diversity considerations in HIV-1 vaccine selection. Science 296, 2354–2360.
Gilbert, M.T., Rambaut, A., Wlasiuk, G., Spira, T.J., Pitchenik, A.E., Worobey, M., 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc. Natl. Acad. Sci. U.S.A. 104, 18566–18570.
Goonetilleke, N., Liu, M.K., Salazar-Gonzalez, J.F., Ferrari, G., Giorgi, E., Ganusov, V.V., Keele, B.F., Learn, G.H., Turnbull, E.L., Salazar, M.G., Weinhold, K.J., Moore, S., Letvin, N., Haynes, B.F., Cohen, M.S., Hraber, P., Bhattacharya, T., Borrow, P., Perelson, A.S., Hahn, B.H., Shaw, G.M., Korber, B.T., McMichael, A.J., 2009. The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. J. Exp. Med. 206, 1253–1272.
Hasegawa, M., Yano, T., Kishino, H., 1984. A new molecular clock of mitochondrial DNA and the evolution of Hominoids. Proc. Jpn. Acad. B 60, 95–98.
Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160–174.
Hwang, S.S., Boyle, T.J., Lyerly, H.K., Cullen, B.R., 1991. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science 253, 71–74.
Jensen, M.A., Li, F.S., van't Wout, A.B., Nickle, D.C., Shriner, D., He, H.X., McLaughlin, S., Shankarappa, R., Margolick, J.B., Mullins, J.I., 2003. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. J. Virol. 77, 13376–13388.
Kawashima, Y., Pfafferott, K., Frater, J., Matthews, P., Payne, R., Addo, M., Gatanaga, H., Fujiwara, M., Hachiya, A., Koizumi, H., Kuse, N., Oka, S., Duda, A., Prendergast, A., Crawford, H., Leslie, A., Brumme, Z., Brumme, C., Allen, T., Brander, C., Kaslow, R., Tang, J., Hunter, E., Allen, S., Mulenga, J., Branch, S., Roach, T., John, M., Mallal, S., Ogwu, A., Shapiro, R., Prado, J.G., Fidler, S., Weber, J., Pybus, O.G., Klenerman, P., Ndung'u, T., Phillips, R., Heckerman, D., Harrigan, P.R., Walker, B.D., Takiguchi, M., Goulder, P., 2009. Adaptation of HIV-1 to human leukocyte antigen class I. Nature 458, 641–645.
Kelleher, A.D., Long, C., Holmes, E.C., Allen, R.L., Wilson, J., Conlon, C., Workman, C., Shaunak, S., Olson, K., Goulder, P., Brander, C., Ogg, G., Sullivan, J.S., Dyer, W., Jones, I., McMichael, A.J., Rowland-Jones, S., Phillips, R.E., 2001. Clustered mutations in HIV-1 gag are consistently required for escape from HLA-B27-restricted cytotoxic T lymphocyte responses. J. Exp. Med. 193, 375–386.
Kiepiela, P., Ngumbela, K., Thobakgale, C., Ramduth, D., Honeyborne, I., Moodley, E., Reddy, S., de Pierres, C., Mncube, Z., Mkhwanazi, N., Bishop, K., van der Stok, M., Nair, K., Khan, N., Crawford, H., Payne, R., Leslie, A., Prado, J., Prendergast, A., Frater, J., McCarthy, N., Brander, C., Learn, G.H., Nickle, D., Rousseau, C., Coovadia, H., Mullins, J.I., Heckerman, D., Walker, B.D., Goulder, P., 2007. CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. Nat. Med. 13, 46–53.
Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., Bhattacharya, T., 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288, 1789–1796.
Li, W.H., Tanimura, M., Sharp, P.M., 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol. Biol. Evol. 5, 313–330.
Maddison, D.R., Maddison, W.P., 2003. MacClade4. Sinauer Associates, Inc., Sunderland.
Moore, C.B., John, M., James, I.R., Christiansen, F.T., Witt, C.S., Mallal, S.A., 2002. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. Science 296, 1439–1443.
Moore, J.P., Kitchen, S.G., Pugach, P., Zack, J.A., 2004. The CCR5 and CXCR4 coreceptors – central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection. AIDS Res. Hum. Retroviruses 20, 111–126.
Muse, S.V., Gaut, B.S., 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11, 715–724.
Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–426.
Nickle, D.C., Rolland, M., Jensen, M.A., Pond, S.L., Deng, W., Seligman, M., Heckerman, D., Mullins, J.I., Jojic, N., 2007. Coping with viral diversity in HIV vaccine design. PLoS Comput. Biol. 3, e75.
Ogg, G.S., Jin, X., Bonhoeffer, S., Dunbar, P.R., Nowak, M.A., Monard, S., Segal, J.P., Cao, Y., Rowland-Jones, S.L., Cerundolo, V., Hurley, A., Markowitz, M., Ho, D.D., Nixon, D.F., McMichael, A.J., 1998. Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma load of viral RNA. Science 279, 2103–2106.
Oliveira, T., Salemi, M., Gordon, M., Vandamme, A.-M., Rensburg, E.J., Engelbrecht, S., Hoosen, M., Cassol, S., 2004. Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design? Genetics 167, 1047–1058.
Ota, T., Nei, M., 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol. Biol. Evol. 11, 613–619.
Papandreou, M.J., Idziorek, T., Miquelis, R., Fenouillet, E., 1996. Glycosylation and stability of mature HIV envelope glycoprotein conformation under various conditions. FEBS Lett. 379, 171–176.
Pape, J.W., Liautaud, B., Thomas, F., Mathurin, J.R., St Amand, M.M., Boncy, M., Pean, V., Pamphile, M., Laroche, A.C., Johnson Jr., W.D., 1983. Characteristics of the acquired immunodeficiency syndrome (AIDS) in Haiti. N. Engl. J. Med. 309, 945–950.
Peakall, R., Smouse, P.E., 2006. Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol. Ecol. Notes 6, 288–295.
Phillips, R.E., Rowland-Jones, S., Nixon, D.F., Gotch, F.M., Edwards, J.P., Ogunlesi, A.O., Elvin, J.G., Rothbard, J.A., Bangham, C.R., Rizza, C.R., et al., 1991. Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature 354, 453–459.
Pollakis, G., Kang, S., Kliphuis, A., Chalaby, M.I., Goudsmit, J., Paxton, W.A., 2001. N-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization. J. Biol. Chem. 276, 13433–13441.
Pond, S.L., Frost, S.D., 2005a. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21, 2531–2533.
Pond, S.L., Frost, S.D.W., 2005b. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22, 1208–1222.
Pond, S.L., Frost, S.D., Grossman, Z., Gravenor, M.B., Richman, D.D., Brown, A.J., 2006. Adaptation to different human populations by HIV-1 revealed by codon-based analyses. PLoS Comput. Biol. 2, e62.
Posada, D., Crandall, K.A., 1998. Modeltest: testing the model of DNA substitution. Bioinformatics 14, 817–818.
Rambaut, A., Posada, D., Crandall, K.A., Holmes, E.C., 2004. The causes and consequences of HIV evolution. Nat. Rev. Genet. 5, 52–61.
Raymond, M., Rousset, F., 1995. An exact test for population differentiation. Evolution 49, 1280–1283.
Richman, D.D., Wrin, T., Little, S.J., Petropoulos, C.J., 2003. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc. Natl. Acad. Sci. U.S.A. 100, 4144–4149.
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Ross, H.A., Rodrigo, A.G., 2002. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. J. Virol. 76, 11715–11720.
Salazar-Gonzalez, J.F., Salazar, M.G., Keele, B.F., Learn, G.H., Giorgi, E.E., Li, H., Decker, J.M., Wang, S., Baalwa, J., Kraus, M.H., Parrish, N.F., Shaw, K.S., Guffey, M.B., Bar, K.J., Davis, K.L., Ochsenbauer-Jambor, C., Kappes, J.C., Saag, M.S., Cohen, M.S., Mulenga, J., Derdeyn, C.A., Allen, S., Hunter, E., Markowitz, M., Hraber, P., Perelson, A.S., Bhattacharya, T., Haynes, B.F., Korber, B.T., Hahn, B.H., Shaw, G.M., 2009. Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. J. Exp. Med. 206, 1273–1289.
Samson, M., Libert, F., Doranz, B.J., Rucker, J., Liesnard, C., Farber, C.M., Saragosti, S., Lapoumeroulie, C., Cognaux, J., Forceille, C., Muyldermans, G., Verhofstede, C., Burtonboy, G., Georges, M., Imai, T., Rana, S., Yi, Y., Smyth, R.J., Collman, R.G., Doms, R.W., Vassart, G., Parmentier, M., 1996. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 382, 722–725.
Sanjuan, R., Codoner, F.M., Moya, A., Elena, S.F., 2004. Natural selection and the organ-specific differentiation of HIV-1 V3 hypervariable region. Evolution 58, 1185–1194.
Santra, S., Liao, H.X., Zhang, R., Muldoon, M., Watson, S., Fischer, W., Theiler, J., Szinger, J., Balachandran, H., Buzby, A., Quinn, D., Parks, R.J., Tsao, C.Y., Carville, A., Mansfield, K.G., Pavlakis, G.N., Felber, B.K., Haynes, B.F., Korber, B.T., Letvin, N.L., 2010. Mosaic vaccines elicit CD8+ T lymphocyte responses that confer enhanced immune coverage of diverse HIV strains in monkeys. Nat. Med. 16, 324–328.
Schneider, S., Roessli, D., Excoffier, L., 2000. Arlequin Ver 2.0: A Software for Population Genetics Data Analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Shioda, T., Levy, J.A., Cheng-Mayer, C., 1992. Small amino acid changes in the V3 hypervariable region of gp120 can affect the T-cell-line and macrophage tropism of human immunodeficiency virus type 1. Proc. Natl. Acad. Sci. U.S.A. 89, 9434–9438.

Slobod, K.S., Bonsignori, M., Brown, S.A., Zhan, X.Y., Stambas, J., Hurwitz, J.L., 2005. HIV vaccines: brief review and discussion of future directions. Exp. Rev. Vac. 4, 305–313.

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. 57, 758–771.

Suzuki, Y., Gojobori, T., 1999. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16, 1315–1328.

Swofford, D.L., 2001. PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland, MA.

Travers, S.A., O'Connell, M.J., McCormack, G.P., McInerney, J.O., 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.

Wei, X., Decker, J.M., Wang, S., Hui, H., Kappes, J.C., Wu, X., Salazar-Gonzalez, J.F., Salazar, M.G., Kilby, J.M., Saag, M.S., Komarova, N.L., Nowak, M.A., Hahn, B.H., Kwong, P.D., Shaw, G.M., 2003. Antibody neutralization and escape by HIV-1. Nature 422, 307–312.

Williamson, S., 2003. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol. Biol. Evol. 20, 1318–1325.

Wyatt, R., Sullivan, N., Thali, M., Repke, H., Ho, D., Robinson, J., Posner, M., Sodroski, J., 1993. Functional and immunologic characterization of human immunodeficiency virus type 1 envelope glycoproteins containing deletions of the major variable regions. J. Virol. 67, 4557–4565.

Yamaguchi, Y., Gojobori, T., 1997. Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc. Natl. Acad. Sci. U.S.A. 94, 1264–1269.

Yamaguchi-Kabata, Y., Gojobori, T., 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. J. Virol. 74, 4335–4350.

Yang, O.O., 2008. Retracing our STEP towards a successful CTL-based HIV-1 vaccine. Vaccine 26, 3138–3141.

Yang, Z., Nelson, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. 17, 32–43.

Yang, W., Bielawski, J.P., Ziheng, Y., 2003. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J. Mol. Evol. 57, 212–221.

Yusim, K., Kesmir, C., Gaschen, B., Addo, M.M., Altfeld, M., Brunak, S., Chigaev, A., Detours, V., Korber, B.T., 2002. Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. J. Virol. 76, 8757–8768.

Zanotto, P.M., Kallas, E.G., de Souza, R.F., Holmes, E.C., 1999. Genealogical evidence for positive selection in the nef gene of HIV-1. Genetics 153, 1077–1089.