JOURNAL OF VIROLOGY, Feb. 2005, p. 1836–1841 Vol. 79, No. 3
0022-538X/05/$08.00+0 doi:10.1128/JVI.79.3.1836–1841.2005

Evidence for Heterogeneous Selective Pressures in the Evolution of the *env* Gene in Different Human Immunodeficiency Virus Type 1 Subtypes

Simon A. A. Travers, Mary J. O'Connell, Grace P. McCormack, and James O. McInerney*

Biology Department, National University of Ireland, Maynooth, County Kildare, Ireland

Received 30 June 2004/Accepted 10 September 2004

Recent studies have demonstrated the emergence of human immunodeficiency virus type 1 (HIV-1) subtypes with various levels of fitness. Using heterogeneous maximum-likelihood models of adaptive evolution implemented in the PAML software package, with *env* sequences representing each HIV-1 group M subtype, we examined the various intersubtype selective pressures operating across the *env* gene. We found heterogeneity of evolutionary mechanisms between the different subtypes with a category of amino acid sites observed that had undergone positive selection for subtypes C, F1, and G, while these sites had undergone purifying selection in all other subtypes. Also, amino acid sites within subtypes A and K that had undergone purifying selection were observed, while these sites had undergone positive selection in all other subtypes. The presence of such sites indicates heterogeneity of selective pressures within HIV-1 group M subtype evolution that may account for the various levels of fitness of the subtypes.

It has been hypothesized that human immunodeficiency virus type 1 (HIV-1) may have entered humans in three independent transmissions of simian immunodeficiency virus from infected chimpanzees from which the three HIV-1 M, N, and O lineages arose (9). Within group M, nine phylogenetically distinct subtypes have been proposed (subtypes A to D, F to H, J, and K), with subsubtypes being proposed for subtypes A and F (18). Subtype distribution varies worldwide, with subtype B predominating in North America and Europe (15) and subtype C accounting for more than 55% of worldwide infections (7) due mainly to its prevalence in Southern and Eastern Africa (1, 2, 14, 17, 26) and India (21) and its increasing prevalence in Brazil (22) and China (19). Biological differences including low CXCR4 coreceptor usage in subtype C (15), decreased protease susceptibility in subtype G (5), and varying subtype reactivity to monoclonal antibodies (13, 25) have been observed among the subtypes. The production of broadly neutralizing or subtype-specific vaccines requires an in-depth understanding of the inter- and intrasubtype evolution.

The study of the selective pressures governing the evolution of protein-coding DNA sequences has traditionally been carried out by comparing dN (nonsynonymous substitutions per nonsynonymous site) to dS (synonymous substitutions per synonymous site), resulting in a dN-to-dS ratio (ω) (see reference 31 for a review). An ω of >1 is indicative of positive selection, an ω of 1 indicates neutral evolution, and an ω of <1 indicates purifying (negative) selection. However, if there is strong purifying selection operating on the majority of amino acid positions, averaging ω over an entire sequence could misleadingly indicate purifying selection for the entire molecule even in the presence of a small number of sites undergoing positive selection. Recently, more biologically realistic methods have been developed to allow for identification of heterogeneous selection pressure across amino acid sites and also heterogeneity across both sites and lineages within the phylogeny (32, 33).

Previous studies (4, 28) focused on searching for positive selection within the HIV-1 group M subtypes by analyzing each subtype independently and identifying amino acid sites with a high probability of having undergone positive selection. However, Drummond et al. (6), referring to work by Seo et al. (20) as an example, suggested that positive selection seems to be a minor contributor to the overall molecular evolution of HIV-1 and that negative (purifying) selection imposed by functional constraints in HIV-1 is more important than positive selection. Here, we present an analysis of the likely selective pressures that have affected HIV-1 group M *env* sequences in their diversification from the original group M founder virus. We have carried out this analysis by comparing each individual subtype to all other group M subtypes in an attempt to identify amino acid sites whose evolutionary history appears to be unique in terms of selective constraints for that subtype. The identification of such sites yields information as to unique subtype-specific molecular traits that may also manifest as unique biological traits.

MATERIALS AND METHODS

Alignments. All available full-length envelope gene sequences were downloaded from the Los Alamos National Laboratory HIV sequence database (http://hiv-web.lanl.gov) and aligned by using MacClade (12), and neighbor-joining trees were produced for each subtype by using PAUP* (23). A subset of the full-length envelope gene sequences for each subtype was selected by choosing as diverse a range of sequences as possible within each subtype based on their spread through the subtype-specific trees (Table 1). Sequences with a large degree of similarity contain much the same information, whereas divergent sequences will contain more information about the intrasubtype diversity. An alignment of the representative sequences of each subtype was produced by using MacClade (12). Ambiguous regions of the alignment were removed to avoid possible false detection of positive selection due to alignment of nonhomologous sites. The resulting *env* data set contained 40 sequences and was 764 codons in length.

* Corresponding author. Mailing address: Bioinformatic and Pharmacogenomics Laboratory, Biology Department, NUI Maynooth, Maynooth, County Kildare, Ireland. Phone: 353-1-708 3860. Fax: 353-1-708 3845. E-mail: james.o.mcinerney@may.ie.

TABLE 1. Representative sequences of selected subtypes present in the *env* data set

Subtype	Representative sequences (GenBank accession no.)
A1	AF004885, AF457080, AF069673, AF19327, AB098333, AF457075, AF484478, U51190
A2	AF286237, AF286238
B	U63632, AY037270, U69589, AF042102
C	AF110967, AF110971, AF443091, AF286227
D	U88822, AY237166, AF484505, AF484519
F1	AF077336, AF005494, AJ249238, AY173957
F2	AJ249236, AF377956, AJ249237
G	AF061642, AF061640, AF084936, AF423760
H	AF190128, AF005496, AF190127
J	AF082394, AF082395
K	AJ249235, AJ249239

A phylogenetic analysis of the data was done by using the maximum-likelihood criterion as implemented with PAUP* (23) using the GTR+I+G substitution model as selected by Modeltest (16). In order to assess confidence in each of the internal nodes of the constructed phylogeny, a bootstrap resampling (1,000 replicates) of the data using the neighbor-joining method based on maximum-likelihood distances was performed with PAUP* (23). Tests for saturation of synonymous sites throughout the phylogenetic tree were performed by using SWAPSC (8).

Intersubtype evolutionary analysis. The software program Codeml from the PAML package (30) was used for evolutionary analysis of the data set. A number of site-specific models of codon substitution that allow for rate heterogeneity among sites were employed, namely model 0, model 1, model 2, model 3, model 7, and model 8 (M0, M1, M2, M3, M7, and M8, respectively). The null models M0, M1, and M7, with dN-to-dS ratios (ω) limited between 0 and 1, do not allow for the existence of positively selected sites. The alternate models M2, M3, and M8 allow for the detection of positive selection by enabling the estimated ω to be greater than 1. For each of the site-specific models, all sites in the data set under examination are allocated to one of the constrained or estimated ω values using maximum likelihood with the proportion of sites allocated to that category being described by using *P* values with p_0 pertaining to the proportion of sites allocated to ω_0, p_1 pertaining to the proportion of sites allocated to ω_1, and so on.

Also, branch site-specific models (model A and model B), which allow for rate heterogeneity across sites and across the tree, were employed. Model A computes three ω values and is an extension of M1 in that it limits the first two ω values (ω_0 and ω_1) to 0 and 1 and allows the final ω (ω_2), which is estimated, a value greater than 1. Model B is an extension of M3 in that all three ω values are estimated. For both of the branch site models, four proportions of sites are allocated to the data set. p_0 is the proportion of sites throughout the alignment allocated to ω_0 with p_1 being the proportion of sites allocated to ω_1. p_2 corresponds to the proportion of sites with a ω_0 value in the background and a ω_2 value in the foreground, while p_3 corresponds to the proportion of sites with a ω_1 value in the background and a ω_2 value in the foreground. The models that allow for all parameters to be estimated are more biologically realistic than the ones that constrain certain parameters since the evolutionary mechanisms operating within a data set are never simple. Constraining certain parameters within the analysis provides a poor representation of the data, while allowing all parameters to be estimated from the data will be a much better and more realistic representation of the data.

The significance of the alternate models (whether the alternate model is a significantly better representation of the data than the null model) were tested by using a likelihood ratio test (LRT) which involves taking twice the difference of the log likelihood between the nested models and testing for significance using the χ^2 distribution with the degrees of freedom being the difference in the number of free parameters between the two models. Models compared in this study using LRT were M0 and M3, M1 and M2, M7 and M8, M1 and model A, and M3 and model B (for more information on the models used, see references 32 and 33).

Since the branch site models operate by allowing the user to examine the evolutionary mechanisms occurring in a particular lineage in the tree (the foreground) against the other lineages (the background), they provide a unique method of analysis by allowing the selective constraints operating on certain sequences to be compared to the selective constraints operating on all the other sequences present in the data set. The site-specific models do not allow for this kind of analysis, and therefore, labeling the internal node leading to each HIV-1 group M subtype allowed comparison of the evolution of that particular subtype contrasted with all other subtypes using the branch site-specific models. The branches leading to subtypes A, B, C, D, F, G, H, J, and K were labeled in a separate analysis, as were the branches leading to the A1, A2, F1, and F2 lineages. To ensure stable results, each model was run four times using different starting ω values, and results from the run with the best likelihood score were taken.

Detection of significant sites. Codeml uses a Bayesian approach to infer the posterior probability that a particular codon in an alignment is in a particular category (i.e., undergoing a specific selective pressure), and generally, codon sites with *P* values of >0.95 are accepted as being significantly allocated to that class. At times, especially with the branch site models, the likelihood ratio test may be significant, yet no sites allocated to a particular category will consider a *P* value of >0.95 as being in that category. The significant result from the LRT indicates the presence of a class of sites causing significance of the model; however, the Bayesian approach for the identification of these sites has been suggested to be inadequate using the branch-specific models (33). In order to identify these sites causing LRT significance, we used a site-stripping method to remove the sites with the highest Bayesian posterior probability, and the resulting stripped alignment was then reanalyzed (using the same models and parameters). This process was repeated iteratively until the LRT failed. Sites removed before the LRT failed were taken to be the sites contributing to the significance of the alternate model.

RESULTS

ω estimates. All subtype clades in the maximum-likelihood tree produced from the data were strongly supported by bootstrapping (Fig. 1), and no saturation of synonymous sites was observed within the data. For the site-specific models, all LRTs were significant with a *P* of 0.0005. The biologically more realistic models detected positive selection occurring at sites in the data with M3 and M8 allocating 14% of sites with an ω of 2.4814 and 12% of sites with an ω of 2.58819, respectively. Purifying selection was observed to have occurred in the ma-

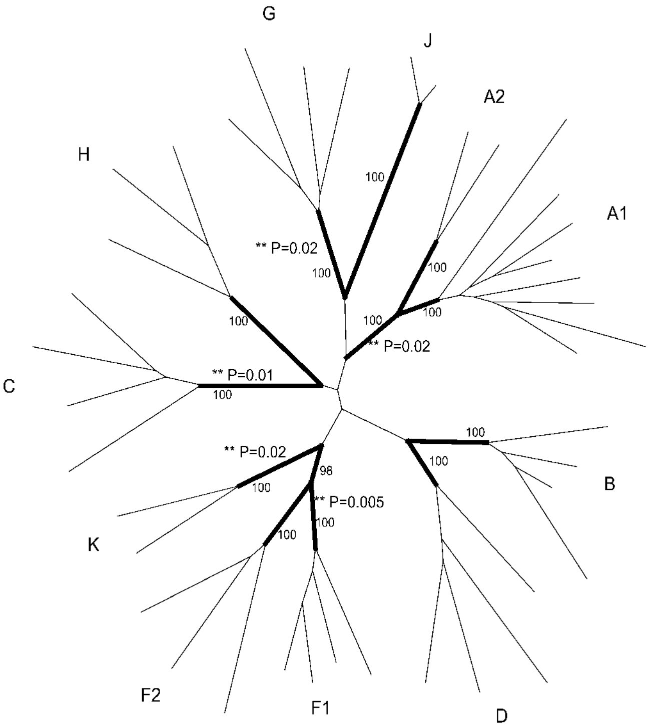


FIG. 1. Constructed phylogeny of the *env* data. Branches labeled in boldface are the branches leading to each subtype lineage analyzed in this study. *P* values for significant branch site models are marked (** P), as are the bootstrap supports for each subtype lineage.

TABLE 2. Selected sites for each subtype[a]

Subtype	Selected sites					
A	4K	75V	305K	307I	641L[b,c]	832V[b,f]
C	665K	853R				
F1	94N[e]	210F				
G	25L	732P	824E			
K	46K	49T	62D[b]	65V	87V[b,c,g]	92N[f]
	99D	161I[c]	164S[c]	169V[b,c]	171K	229N
	231K[b,e,f]	240T[b,c,d,g]	252R	268E	271V	289N[b,f]
	290T	293E[b,e]	295N[b,f]	297T	305K	306R[b,c,d]
	309I	314G	322K[d]	334S	335R[d]	339N[c]
	440S[b,c,d,g]	471G	474D	515I	535M[b]	606T
	648E[b]	668S	671N	726G	732G	818T
	821A	829V[b]	845R			

[a] Category I sites were observed for subtypes C, F1, and G, while category II sites were observed for subtypes A and K. Sites are described using the HXB2 reference sequence.

[b] Site undergoing positive selection in at least one HIV-1 group M subtype as determined by Choisy et al (4).

[c] Site undergoing positive selection as determined by Yang (29).

[d] Site undergoing positive selection as determined by Yamaguchi-Kabata and Gojobori (27).

[e] Site undergoing purifying selection as determined by Yamaguchi-Kabata and Gojobori (27).

[f] Site undergoing positive selection as determined by Yang et al (28) for their combined data set.

[g] Site undergoing positive selection as determined by Yang et al (28) in a separate analysis of subtypes A, B, and C.

jority of sites (86% for M3 and 88% for M8) through the *env* gene for all subtypes.

The branch site models were implemented to detect any sites that have evolved uniquely to a particular subtype when compared to the other subtypes. The branch site results were significant for the branches leading to subtypes A (P = 0.02), C (P = 0.01), F1 (P = 0.005), G (P = 0.02), and K (P = 0.02), suggesting that these subtypes contain a category of sites that have evolved differently from the other subtypes. From the branch-specific model, the subtypes' results fell into two categories. In subtypes C, F1, and G, a proportion of their sites were observed to have undergone positive selection, whereas all other subtypes had undergone purifying selection at that site (described herein as category I sites). In subtypes A and K, a proportion of sites was observed to have undergone purifying selection with positive selection having occurred in the other subtypes at those sites (described herein as category II sites). For subtypes C and F1, two sites each were allocated to category I, while three sites were identified in this category for subtype G. Six codons were allocated to category II for subtype A, and 45 codons were allocated to category II for subtype K (Table 2).

The significant sites for each subtype were labeled on amino acid alignments using the known protein secondary structures for gp120 (11) and gp41 (3) (Fig. 2 and 3).

gp120 structural amino acids. For the gp120 structure (Fig. 2), the majority of sites observed in both category I and category II were structural sites not directly involved in known gp120 functions. Within subtype K, however, a number of the identified category II gp120 sites are functionally significant. Amino acid sites 295N, 297T, and 334S correspond to a cluster of nonlinear sites located on the outer domain of gp120 associated with the binding of the 2G12 antibody (25). Twenty-six residues spanning six segments of the gp120 molecule are involved in direct contact with the host cell CD4 receptor (11), one of which (474D) was identified in category II for subtype K. Sites 305K, 306R, and 322K, also identified as category II sites in subtype K, are sites directly involved in or adjacent to sites directly involved in the switch from the CXCR4 to the CCR5 coreceptor.

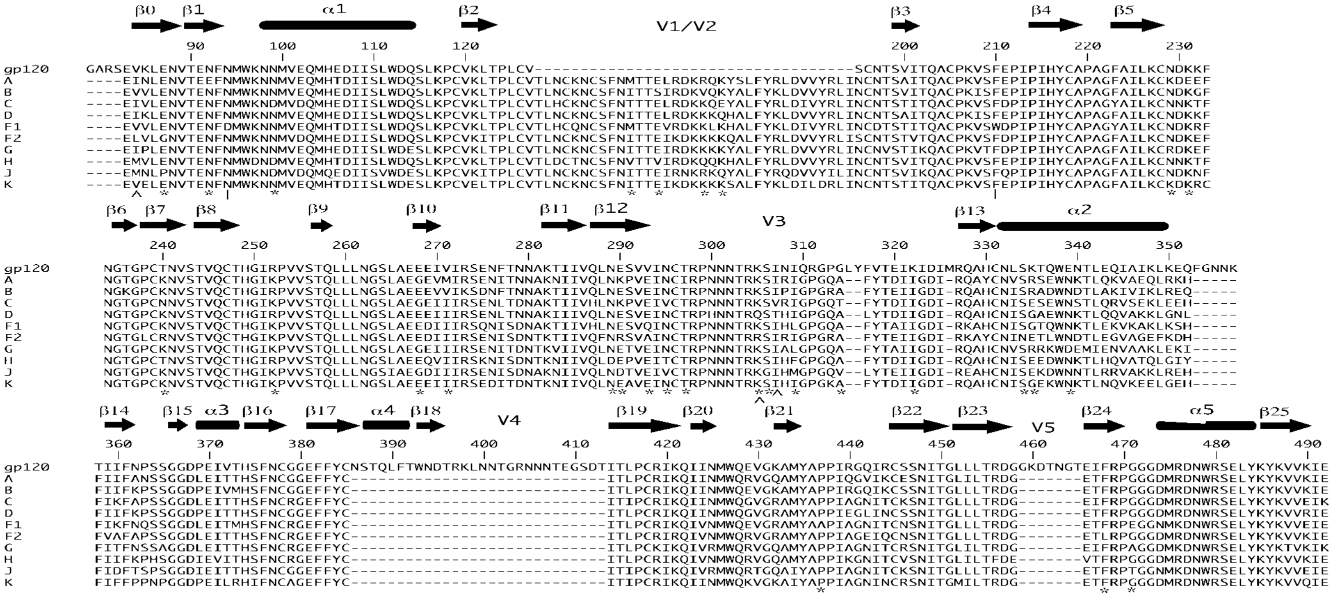
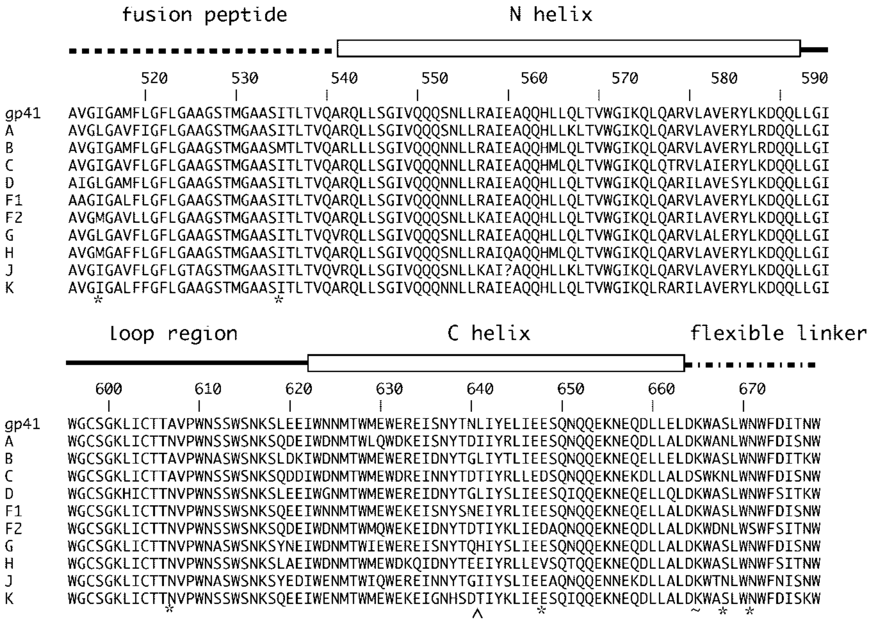

β0 β1 α1 β2 V1/V2 β3 β4 β5
90 100 110 120 200 210 220 230
gp120 GARSEVKLENVTENFNMWKNNMVEQMHEDIISLWDQSLKPCVKLTPLCV-----------------------------------SCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCNDKKF
A ----EINLENVTEEFNMWKNNMVEQMHTDIISLWDQSLKPCVKLTPLCVTLNCKNCSFNMTTELRDKRQKYSLFYRLDVVYRLINCNTSAITQACPKVSFEPIPIHYCAPAGFAILKCKDEEF
B ----EVVLENVTENFNMWKNNMVEQMHEDIISLWDQSLKPCVKLTPLCVTLNCKNCSFNITTSIRDKVQKYALFYKLDVVYRLINCNTSVITQACPKISFEPIPIHYCAPAGFAILKCNDKGF
C ----EIVLENVTENFNMWKNDMVDQMHEDIISLWDQSLKPCVKLTPLCVTLHCKNCSFNTTTELRDKKQEYALFYRLDIVYRLINCNTSTITQACPKVSFDPIPIHYCAPAGYAILKCNNKTF
D ----EIKLENVTENFNMWKNNMVEQMHEDIISLWDQSLKPCVKLTPLCVTLNCKNCSFNITTELRDKKKQHALFYKLDVVYRLINCNTSAITQACPKISFEPIPIHYCAPAGFAILKCNDKKF
F1 ----EVVLENVTENFDMWKNNMVEQMHTDIISLWDQSLKPCVKLTPLCVTLHCQNCSFNMTTEVRDKKLKHALFYKLDIVYRLINCDTSTITQACPKVSWDPIPIHYCAPAGYAILKCNDKRF
F2 ----ELVLGNVTENFNMWKNNMVDQMHEDIISLWDQSLKPCVKITPLCVTLNCKNCSFNITTEIKDKKKKQALFYRLDVVYRLISCNTSTVTQACPKVSFDPIPIHYCAPAGFAILKCNDKEF
G ----EIPLENVTENFNMWKNNMVEQMHEDIISLWDESLKPCVKLTPLCVTLNCKNCSFNITTEIRDKKKKYALFYRLDVVYRLINCNVSTIKQACPKVTFDPIPIHYCAPAGFAILKCRDKEF
H ----EMVLENVTENFNMWDNDMVEQMHTDIISLWDQSLKPCVKLTPLCVTLDCTNCSFNVTTVIRDKQQKHALFYRLDVVYRLINCNTSVITQACPKVSFEPIPIHYCAPAGFAILKCNNKTF
J ----EMNLPNVTENFNMWKNDMVDQMQEDIISVWDESLKPCVKITPLCVTLNCKNCSFNITTEIRNKRKQYALFYRQDVVYILINCNTSVIKQACPKVSFQPIPIHYCAPAGFAILKCNDKNF
K ----EVELENVTENFNMWKNNMVEQMHTDIISLWDESLKPCVELTPLCVTLNCKNCSFNITTEIKDKKKKSALFYKLDILDRLINCNTSTITQACPKVSFEPIPIHYCAPAGFAILKCKDKRC
^ * * | * * * * * | * *
β6 β7 β8 β9 β10 β11 β12 V3 β13 α2
240 250 260 270 280 290 300 310 320 330 340 350
gp120 NGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEEIVIRSENFTNNAKTIIVQLNESVVINCTRPNNNTRKSINIQRGPGLYFVTEIKIDIMRQAHCNLSKTQWENTLEQIAIKLKEQFGNNK
A NGTGPCKNVSTVQCTHGIKPVVSTQLLLNGSLAEGEVMIRSENITNNAKNIIVQLNKPVEINCIRPNNNTRKSIRIGPGQA--FYTDIIGDI-RQAYCNVSRSEWNKTLQKVAEQLRKH-----
B NGKGPCKNVSTVQCTHGIRPVVSTQLLLNGSLAEEEVVIKSDNFTNNAKTIIVQLNESVEINCTRPNNNTRKSIPIGPGRA--FYTEIIGDI-RQAHCNISRADWNDTLAKIVIKLREQ-----
C NGTGPCNNVSTVQCTHGIKPVVSTQLLLNGSLAEGEIIIRSENLTDNAKTIIVHLNKPVEIVCTRPNNNTRKSVRIGPGQT--FYTDIIGDI-RQAHCNISESEWNSTLQRVSEKLEEH-----
D NGTGPCKNVSTVQCTHGIRPVVSTQLLLNGSLAEEEIIIRSENLTNNAKIIIVQLNESVEINCTRPHNNTRQSTHIGPGQA--LYTDIIGDI-RQAHCNISGAEWNKTLQQVAKKLGNL-----
F1 NGTGPCKNVSTVQCTHGIKPVVSTQLLLNGSLAEEDIIIRSQNISDNAKTIIVHLNESVQINCTRPNNNTRKSIHLGPGQA--FYTAIIGDI-RKAHCNISGTQWNKTLEKVKAKLKSH-----
F2 NGTGLCRNVSTVQCTHGIKPVVSTQLLLNGSLAEEDIIIRSENISDNTKTIIVQFNRSVAINCTRPNNNTRRSIRIGPGRA--FYTEIIGDI-RKAYCNINETLWNDTLEGVAGEFKDH-----
G NGTGPCKNVSTVQCTHGIKPVVSTQLLLNGSLAEGEIIIRSENITDNTKVIIVQLNETVEINCTRPNNNTRKSIALGPGQA--FYTAIIGDI-RQAHCNVSRRKWDEMIENVAAKLEKI-----
H NGTGPCTNVSTVQCTHGIRPVVSTQLLLNGSLAEEQVIIRSKNISDNTKNIIVQLDEPVEITCTRPNNNTRKSIHFGPGQA--FYTDIIGDI-RQAHCNISEEDWNKTLHQVATQLGIY-----
J NGTGPCKNVSTVQCTHGIKPVVSTQLLLNGSIAEGDIIIRSENISDNAKNIIVQLNDTVEIVCTRPNNNTRKGIHMGPGQV--LYTEIIGDI-REAHCNISEKDWNNTLRRVAKKLREH-----
K NGTGPCKNVSTVQCTHGIKPVVSTQLLLNGSLAEEEIIIRSEDITDNTKNIIVQLNEAVEINCTRPNNNTRKSIHIGPGKA--FYTDIIGDI-RQAHCNISGEKWNKTLNQVKEELGEH-----
* * * * ** * * * **^ * * * ** *
^
β14 β15 α3 β16 β17 α4 β18 V4 β19 β20 β21 β22 β23 V5 β24 α5 β25
360 370 380 390 400 410 420 430 440 450 460 470 480 490
gp120 TIIFNPSSGGDPEIVTHSFNCGGEFFYCNSTQLFTWNDTRKLNNTGRNNNTEGSDTITLPCRIKQIINMWQEVGKAMYAPPIRGQIRCSSNITGLLLTRDGGKDTNGTEIFRPGGGDMRDNWRSELYKYKVVKIE
A FIIFANSSGGDLEITTHSFNCGGEFFYC----------------------------ITLPCRIKQIINMWQRVGQAMYAPPIQGVIKCESNITGLILTRDG-------ETFRPGGGDMRDNWRSELYKYKVVKIE
B FIIFKPSSGGDPEIVMHSFNCGGEFFYC----------------------------ITLPCRIKQIINMWQEVGKAMYAPPIRGQIKCSSNITGLLLTRDG-------ETFRPGGGDMRDNWRSELYKYKVVKIE
C FIKFAPSSGGDLEITTHSFNCRGEFFYC----------------------------ITIPCRIKQIINMWQGVGRAMYAPPIAGNITCKSNITGLLLTRDG-------EIFRPGGGDMRDNWRSELYKYKVVEIK
D FIIFKPSSGGDPEITTHSFNCGGEFFYC----------------------------ITLPCRIKQIINMWQGVGRAMYAPPIEGLINCSSNITGLLLTRDG-------ETFRPGGGDMRDNWRSELYKYKVVRIE
F1 FIKFNQSSGGDLEITMHSFNCRGEFFYC----------------------------ITLPCRIKQIVNMWQEVGRAMYAAPIAGNITCNSNITGLLLTRDG-------ETFRPEGGNMKDNWRSELYKYKVVEIE
F2 FVAFAPSSGGDLEITTHSFNCRGEFFYC----------------------------ITLPCRIRQIVNMWQRVGRAMYAPPIAGEIQCNSNITGLLLTRDG-------ETLRPGGGDMRDNWRSELYKYKVVKIE
G FITFNSSAGGDLEITTHSFNCRGEFFYC----------------------------ITLPCKIKQIVRMWQRVGQAMYAPPIAGNITCRSNITGLILTRDG-------EIFRPAGGDMKDNWRSELYKYKTVKIK
H FIIFKPHSGGDIEVITHSFNCRGEFFYC----------------------------ITLPCRIKQIVNMWQRVGQAMYAPPIKGNITCVSNITGLILTFDE-------VTFRPGGGDMRDNWRSELYKYKVVKIE
J FIDFTSPSGGDIEITTHSFNCGGEFFYC----------------------------ITIPCKIKQIVRMWQRTGQAIYAPPIAGNITCKSNITGLLLTRDG-------ETFRPTGGNMKDNWRSELYKYKVVEIE
K FIFFPPNPGGDPEILRHIFNCAGEFFYC----------------------------ITIPCRIKQIINMWQKVGKAIYAPPIAGNINCRSNITGMILTRDG-------ETFRPGGGDMRDNWRSELYKYKVVQIE
* * *

FIG. 2. Positions of category I sites (identified by | for subtype F) and category II sites (identified by ∧ and ∗ for subtypes A and K, respectively) across the gp120-coding sequence. Secondary protein structures are marked above their coding sequences. Site positions are described using the HXB2 reference sequence.



```
gp41    AVGIGAMFLGFLGAAGSTMGAASITLTVQARQLLSGIVQQQSNLLRAIEAQQHLLQLTVWGIKQLQARVLAVERYLKDQQLLGI
A       AVGLGAVFIGFLGAAGSTMGAASITLTVQARQLLSGIVQQQSNLLRAIEAQQHLLKLTVWGIKQLQARVLAVERYLRDQQLLGI
B       AVGIGAMFLGFLGAAGSTMGAASMTLTVQARLLLSGIVQQQNNLLRAIEAQQHMLQLTVWGIKQLQARVLAVERYLRDQQLLGI
C       AVGIGAVFLGFLGAAGSTMGAASITLTVQARQLLSGIVQQQNNLLRAIEAQQHMLQLTVWGIKQLQTRVLAIERYLKDQQLLGI
D       AIGLGAMFLGFLGAAGSTMGAASITLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARILAVESYLKDQQLLGI
F1      AAGIGALFLGFLGAAGSTMGAASITLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARVLAVERYLKDQQLLGI
F2      AVGMGAVLLGFLGAAGSTMGAASITLTVQARQLLSGIVQQQSNLLKAIEAQQHLLQLTVWGIKQLQARILAVERYLKDQQLLGI
G       AVGLGAVFLGFLGAAGSTMGAASITLTVQVRQLLSGIVQQQSNLLRAIEAQQHLLQLTVWGIKQLQARVLALERYLKDQQLLGI
H       AVGMGAFFLGFLGAAGSTMGAASITLTVQARQLLSGIVQQQSNLLRAIQAQQHMLQLTVWGIKQLQARVLAVERYLKDQQLLGI
J       AVGIGAVFLGFLGTAGSTMGAASITLTVQVRQLLSGIVQQQSNLLKAI?AQQHLLKLTVWGIKQLQARVLAVERYLKDQQLLGI
K       AVGIGALFFGFLGAAGSTMGAASITLTVQARQLLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLRARILAVERYLKDQQLLGI
           *                   *

gp41    WGCSGKLICTTAVPWNSSWSNKSLEEIWNNMTWMEWEREISNYTNLIYELIEESQNQQEKNEQDLLELDKWASLWNWFDITNW
A       WGCSGKLICTTNVPWNSSWSNKSQDEIWDNMTWLQWDKEISNYTDIIYRLIEESQNQQEKNEQDLLALDKWANLWNWFDISNW
B       WGCSGKLICTTAVPWNASWSNKSLDKIWDNMTWMEWEREIDNYTGLIYTLIEESQNQQEKNEQELLELDKWASLWNWFDITKW
C       WGCSGKLICTTAVPWNSSWSNKSQDDIWDNMTWMEWDREINNYTDTIYRLLEDSQNQQEKNEKDLLALDSWKNLWNWFDISNW
D       WGCSGKHICTTNVPWNSSWSNKSLEEIWGNMTWMEWEREIDNYTGLIYSLIEESQIQQEKNEQELLQLDKWASLWNWFSITKW
F1      WGCSGKLICTTNVPWNSSWSNKSQEEIWNNMTWMEWEKEISNYSNEIYRLIEESQNQQEKNEQELLALDKWASLWNWFDISNW
F2      WGCSGKLICTTNVPWNSSWSNKSQDEIWDNMTWMQWEKEIDNYTDTIYKLIEDAQNQQEKNEQDLLALDKWDNLWSWFSITNW
G       WGCSGKLICTTNVPWNASWSNKSYNEIWDNMTWIEWEREISNYTQHIYSLIEESQNQQEKNEQDLLALDKWASLWNWFDISNW
H       WGCSGKLICTTNVPWNSSWSNKSLAEIWDNMTWMEWDKQIDNYTEEIYRLLEVSQTQQEKNEQDLLALDKWASLWNWFSITNW
J       WGCSGKLICTTNVPWNASWSNKSYEDIWENMTWIQWEREINNYTGIIYSLIEEAQNQQENNEKDLLALDKWTNLWNWFNISNW
K       WGCSGKLICTTNVPWNSSWSNKSQEEIWENMTWMEWEKEIGNHSDTIYKLIEESQIQQEKNEQDLLALDKWASLWNWFDISKW
                   *                                 ^      *                ~  *  *
```

FIG. 3. Positions of category I sites (identified by ~ for subtype C) and category II sites (identified by ∧ and * for subtypes A and K, respectively) across the gp41-coding sequence. Secondary protein structures are marked above their coding sequences. Site positions are described using the HXB2 reference sequence.

gp41 structural amino acids. Within gp41, many of the identified category I and category II sites are located in the C-terminal transmembrane region for which, as yet, there is no three-dimensional structure. Only one category I site was identified (665K, subtype C) within the known gp41 ectodomain (Fig. 3) and was located in the crucial flexible linker region that connects the ectodomain to the transmembrane region. Within category II, only one subtype A amino acid site (641L) and six subtype K amino acid sites for were identified within the known gp41 structure (Fig. 3).

DISCUSSION

We have conducted an analysis of the intersubtype evolution of HIV-1 group M subtypes that identify groups of sites that have been subject to different selective constraints in the lineages leading to each subtype. The use of evolutionary models that incorporate different rates over different lineages allowed for the detection of sites undergoing evolutionary constraints unique to a lineage within the data that would not have been detected in any other method of analysis. Two categories of sites were observed: first, sites that have experienced positive selection in a particular subtype when the same site has experienced purifying selection in all other subtypes (category I); and second, sites that have experienced purifying selection in one subtype while other subtypes have experienced positive selection at that site (category II). Sites with a high probability of being in category I were identified within subtypes C, F1, and G, while sites with a high probability of being in category II were identified within subtypes A and K.

Category I amino acid sites. Upon sequence examination, category I sites are generally composed of amino acids that are conserved in the selected subtype but different in all other subtypes. This indicates positive selection for this particular amino acid at that site in the radiation of the subtype. These amino acid changes were radical, with large physiochemical distances between them when compared using the Grantham indices (10). For example, within subtype C, position 665 in the gp41 flexible linker region contains a serine, while a lysine is present in all other subtypes. Previous studies (33, 34) have shown that functional shifts in a protein are often associated with amino acids that exhibit evidence of positive selection. Therefore, it is possible that these positively selected amino acid changes, observed here in HIV-1 subtypes, may also have induced functional change. However, further analysis of the effects on viral fitness and structure by the observed replacements is needed.

There was little correlation between category I sites and amino acid sites identified in other studies (4, 27–29) as having undergone strong selective pressures. Previous intersubtype studies (4, 28) examined group M subtypes independently, looking at selective pressures within each subtype. In this study, we have examined the selective pressures of the branches leading to each subtype compared to those of all other lineages. Sites we have identified as undergoing positive selection in one subtype compared to all other subtypes may seem to be undergoing purifying selection when only the subtype itself is examined due to the conserved nature of these sites within a subtype.

Category II amino acid sites. Both subtype A and subtype K contain amino acid sites that have been under pressure to retain their current state, while these sites are under pressure to change in all other subtypes. While a small number (six) of such sites was identified for subtype A, a much larger number (45) was identified for subtype K. This finding may indicate a substantial difference between the selective pressures governing the evolution of subtype K and those of all other subtypes.

Available molecular evidence indicates that subtype K has existed for as long as the other subtypes (24), but its observed worldwide prevalence is low. Only two subtype K full-genome sequences were available for use in this analysis, and this may have had some effect on the analysis. The genetic distance between the two subtype K sequences was large (9%), while that of other subtypes used in this study, which also had only two representative sequences (and yet did not yield a significant branch site model LRT), was smaller; for example, there was only a 2% genetic distance between the two A2 representative sequences. When subtype K *env* gene fragments available through the Los Alamos National Laboratory HIV sequence database (http://hiv-web.lanl.gov) were used in a similar analysis, the branch site model was not significant (data not shown). This lack of lineage-specific model significance is most likely due to the much shorter sequence data used (150 codons) and the fact that the fragments covered the V3-V5 region, where only 20% of category II sites were observed in the full gene.

A strong correlation was observed between sites identified as category II amino acids in this study and sites observed as undergoing strong evolutionary constraints in other studies (4, 27–29). For example, 50% of the subtype K category II sites in the gp120 three-dimensional structure were determined to be undergoing strong selective pressures in at least one of the other studies examined (4, 27–29) (Table 2). This is not surprising, as category II sites indicate positive selection in all subtypes other than the one of interest. The other studies also identified positive selection at many of the same sites within the same subtypes as those identified in this study.

Functionality of category I and category II sites. Amino acid sites critically important in gp120 and gp41 function such as receptor binding are for the most part undergoing similar evolutionary mechanisms within the subtypes as determined in this study. The intersubtype evolutionary differences have been observed mostly at amino acid sites involved in structure. A number of exceptions to this observation included sites involved in direct CD4 binding, sites implicated in CXCR4-to-CCR5 coreceptor switch, and a putative glycosylation site (category II sites determined in subtype K). Also, within the flexible linker of the gp41 (important for tethering gp41 to the transmembrane segment, coreceptor binding, and host cell entry), amino acids in category I (665K, subtype C) and category II (668S and 671N, subtype K) were observed. Three category II sites (295N, 297T, and 334S) identified within subtype K correspond to one cluster of the 2G12 epitope, an antibody that has been implicated in neutralization of multiple subtypes (25).

To our knowledge, this is the first study to examine the selective pressures that governed the evolution of the subtypes of HIV-1 group M. We have identified categories of sites that have evolved under unique selective pressures for particular subtypes that may cause subtype-specific genetic characteristics. The presence of such sites indicates heterogeneity of selective pressures within HIV evolution, and this fact should be taken into account in any future HIV vaccine or treatment development.

ACKNOWLEDGMENT

Many thanks to Mario Fares for assistance in the use of SWAPSC for synonymous substitution saturation detection.

REFERENCES

1. **Abebe, A., C. L. Kuiken, J. Goudsmit, M. Valk, T. Messele, T. Sahlu, H. Yeneneh, A. Fontanet, F. De Wolf, and T. F. Rinke De Wit.** 1997. HIV type 1 subtype C in Addis Ababa, Ethiopia. AIDS Res. Hum. Retrovir. **13:** 1071–1075.
2. **Abebe, A., G. Pollakis, A. L. Fontanet, B. Fisseha, B. Tegbaru, A. Kliphuis, G. Tesfaye, H. Negassa, M. Cornelissen, J. Goudsmit, and T. F. Rinke de Wit.** 2000. Identification of a genetic subcluster of HIV type 1 subtype C (C′) widespread in Ethiopia. AIDS Res. Hum. Retrovir. **16:**1909–1914.
3. **Caffrey, M., M. Cai, J. Kaufman, S. J. Stahl, P. T. Wingfield, D. G. Covell, A. M. Gronenborn, and G. M. Clore.** 1998. Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. EMBO J. **17:**4572–4584.
4. **Choisy, M., C. H. Woelk, J. F. Guegan, and D. L. Robertson.** 2004. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J. Virol. **78:**1962–1970.
5. **Descamps, D., C. Apetrei, G. Collin, F. Damond, F. Simon, and F. Brun-Vezinet.** 1998. Naturally occurring decreased susceptibility of HIV-1 subtype G to protease inhibitors. AIDS **12:**1109–1111.
6. **Drummond, A., O. G. Pybus, and A. Rambaut.** 2003. Inference of viral evolutionary rates from molecular sequences. Adv. Parasitol. **54:**331–358.
7. **Esparza, J., and N. Bhamarapravati.** 2000. Accelerating the development and future availability of HIV-1 vaccines: why, when, where, and how? Lancet **355:**2061–2066.
8. **Fares, M. A.** 2004. SWAPSC: sliding window analysis procedure to detect selective constraints. Bioinformatics **20:**2867–2868.
9. **Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn.** 1999. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. Nature **397:**436–441.
10. **Grantham, R.** 1974. Amino acid difference formula to help explain protein evolution. Science **185:**862–864.
11. **Kwong, P. D., R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson.** 1998. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature **393:**648–659.
12. **Maddison, W. P., and D. R. Maddison.** 1992. MacClade, version 4.5. Sinauer Associates, Sunderland, Mass.
13. **Moore, J. P., F. E. McCutchan, S. W. Poon, J. Mascola, J. Liu, Y. Cao, and D. D. Ho.** 1994. Exploration of antigenic variation in gp120 from clades A through F of human immunodeficiency virus type 1 by using monoclonal antibodies. J. Virol. **68:**8350–8364.
14. **Novitsky, V. A., M. A. Montano, M. F. McLane, B. Renjifo, F. Vannberg, B. T. Foley, T. P. Ndung'u, M. Rahman, M. J. Makhema, R. Marlink, and M. Essex.** 1999. Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. J. Virol. **73:**4427–4432.
15. **Peeters, M., and P. M. Sharp.** 2000. Genetic diversity of HIV-1: the moving target. AIDS **14**(Suppl. 3)**:**S129–S140.
16. **Posada, D., and K. A. Crandall.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14:**817–818.
17. **Renjifo, B., B. Chaplin, D. Mwakagile, P. Shah, F. Vannberg, G. Msamanga, D. Hunter, W. Fawzi, and M. Essex.** 1998. Epidemic expansion of HIV type 1 subtype C and recombinant genotypes in Tanzania. AIDS Res. Hum. Retrovir. **14:**635–638.
18. **Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber.** 2000. HIV-1 nomenclature proposal. Science **288:**55–56.
19. **Rodenburg, C. M., Y. Li, S. A. Trask, Y. Chen, J. Decker, D. L. Robertson, M. L. Kalish, G. M. Shaw, S. Allen, B. H. Hahn, and F. Gao.** 2001. Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. AIDS Res. Hum. Retrovir. **17:**161–168.
20. **Seo, T. K., J. L. Thorne, M. Hasegawa, and H. Kishino.** 2002. Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. Genetics **160:**1283–1293.
21. **Shankarappa, R., R. Chatterjee, G. H. Learn, D. Neogi, M. Ding, P. Roy, A. Ghosh, L. Kingsley, L. Harrison, J. I. Mullins, and P. Gupta.** 2001. Human immunodeficiency virus type 1 *env* sequences from Calcutta in eastern India: identification of features that distinguish subtype C sequences in India from other subtype C sequences. J. Virol. **75:**10479–10487.
22. **Soares, M. A., T. De Oliveira, R. M. Brindeiro, R. S. Diaz, E. C. Sabino, L. Brigido, I. L. Pires, M. G. Morgado, M. C. Dantas, D. Barreira, P. R. Teixeira, S. Cassol, and A. Tanuri.** 2003. A specific subtype C of human immunodeficiency virus type 1 circulates in Brazil. AIDS **17:**11–21.
23. **Swofford, D. L.** 1998. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.0. Sinauer Associates, Sunderland, Mass.
24. **Triques, K., A. Bourgeois, N. Vidal, E. Mpoudi-Ngole, C. Mulanga-Kabeya, N. Nzilambi, N. Torimiro, E. Saman, E. Delaporte, and M. Peeters.** 2000. Near-full-length genome sequencing of divergent African HIV type 1 sub-

type F viruses leads to the identification of a new HIV type 1 subtype designated K. AIDS Res. Hum. Retrovir. **16:**139–151.

25. **Trkola, A., M. Purtscher, T. Muster, C. Ballaun, A. Buchacher, N. Sullivan, K. Srinivasan, J. Sodroski, J. P. Moore, and H. Katinger.** 1996. Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. J. Virol. **70:** 1100–1108.
26. **Van Harmelen, J. H., E. Van der Ryst, A. S. Loubser, D. York, S. Madurai, S. Lyons, R. Wood, and C. Williamson.** 1999. A predominantly HIV type 1 subtype C-restricted epidemic in South African urban populations. AIDS Res. Hum. Retrovir. **15:**395–398.
27. **Yamaguchi-Kabata, Y., and T. Gojobori.** 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.
28. **Yang, W., J. P. Bielawski, and Z. Yang.** 2003. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J. Mol. Evol. **57:** 212–221.
29. **Yang, Z.** 2001. Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. Pac. Symp. Biocomput. **2001:**226–237.
30. **Yang, Z.** 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:**555–556.
31. **Yang, Z., and J. P. Bielawski.** 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15:**496–503.
32. **Yang, Z., and R. Nielsen.** 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. **19:**908–917.
33. **Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431–449.
34. **Zhang, J., Y. P. Zhang, and H. F. Rosenberg.** 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet. **30:**411–415.