A single early introduction of HIV-1 subtype B into Central America accounts for most current cases

Wendy Murillo1,2‡, Nazle Veras3‡, Mattia Prosperi3, Ivette Lorenzana de Rivera1, Gabriela Paz-Bailey4, Sonia Morales-Miranda4, Sandra I. Juarez5, Chunfu Yang6, Joshua DeVos6, José Pablo Marín7, Mattias Mild2,8, Jan Albert2,9, Marco Salemi3*

1Microbiology Department, National Autonomous University of Honduras, Tegucigalpa, Honduras. 2Department of Microbiology, Cell and Tumor Biology, Karolinska Institute, Stockholm, Sweden. 3Department of Pathology, Immunology and Laboratory Medicine, Emerging Pathogens Institute, University of Florida, Gainesville, Florida, US. 4Del Valle University of Guatemala, Guatemala City, Guatemala. 5Division for Global HIV/AIDS, Central America Region, Centers for Disease Control and Prevention, Guatemala City, Guatemala. 6Division of Global HIV/AIDS, Center for Global Health, Centers for Disease Control and Prevention, Atlanta, Georgia, US. 7Laboratory of Molecular Biology, Costa Rican Social Security, San Juan de Dios Hospital, San José, Costa Rica. 8Swedish Institute for Communicable Disease Control, Stockholm, Sweden. 9Department of Clinical Microbiology, Karolinska University Hospital, Stockholm, Sweden.

‡ These authors equally contributed to the work.

* Corresponding author:
Marco Salemi
Emerging Pathogens Institute
University of Florida
2055 Mowry Road
Gainesville, FL 32610
Tel./Fax +1 352-273-9567/8284
salemi@pathology.ufl.edu

Running title: HIV-1 Subtype B in Central America
Abstract word counts: 211
Text word counts: 4,236
Human immunodeficiency virus type 1 (HIV-1) variants show considerable geographical separation across the world, but there is limited information from Central America. We provide the first detailed investigation of the genetic diversity and molecular epidemiology of HIV-1 in six Central American countries. Phylogenetic analysis was performed on 625 HIV-1 pol gene sequences collected between 2002-2010 in Honduras, El Salvador, Nicaragua, Costa Rica, Panama, and Belize. Published sequences from neighboring countries (n=57) and the rest of the world (n=740) were included as controls. Maximum-likelihood methods were used to explore phylogenetic relationship. Bayesian coalescent-based methods were used to time HIV-1 introductions. Nearly all (98.9%) Central American sequences were of subtype B. Phylogenetic analysis revealed that 437 (70%) sequences clustered within five significantly supported monophyletic clades, formed essentially by Central American sequences. One clade contained 386 (62%) sequences from all six countries; the other four clades were smaller and more country-specific, suggesting discrete sub-epidemics. The existence of one large well-supported Central American clade provides evidence that a single introduction of HIV-1 subtype B in Central America accounts for most current cases. An introduction during the early phase of the HIV-1 pandemic may explain its epidemiological success. Moreover, the smaller clades suggest a subsequent regional spread related to specific transmission networks within each country.
Central America has a significant human immunodeficiency virus type 1 (HIV-1) epidemic with an estimated adult prevalence that ranges from 0.2% to 2.3% (43). Belize has the highest prevalence (2.3%) and Nicaragua and Costa Rica the lowest (0.2% and 0.3%, respectively) (44). WHO estimates that approximately 200,000 people in Central America live with HIV, with Guatemala, Honduras and El Salvador contributing the largest number of cases (21-23) (see Table S1 in the supplemental material). The first cases of HIV-1 in Central America were reported among men who have sex with men (MSM) in the middle 1980’s, but since then the virus has primarily spread heterosexually in the general population (3). Today, heterosexual transmission accounts for 70% of the cases even though HIV-1 prevalence is still higher among MSM, female sex workers (FSWs), prisoners, and some ethnic groups such as the Kuna population in Panama and the Garifuna population in Honduras (20, 21, 28, 29, 34, 37, 42).

The distribution of HIV-1 group M subtypes and circulating recombinant forms (CRFs) differs considerably across the world. HIV-1 group M subtype B dominates the epidemics in North America, Western and Central Europe, whereas subtype B and BF recombinants predominate in South America (13). The information on the molecular epidemiology of HIV-1 in Central America is still limited, but available data suggest that HIV-1 group M subtype B predominates (1, 17, 25, 41).

As a consequence of rapid evolution, HIV-1 gene genealogies (phylogenetic trees) contain important information concerning the patterns, processes and dynamics of viral spread at the population level (12, 18, 32). In the present study we employed phylogenetic
methods to obtain, for the first time, detailed information about HIV-1 origin and epidemiological history in Central America. We found five monophyletic HIV-1 group M subtype B clades suggesting multiple independent introductions into Central America, possibly dating back to the 1960-1970s and followed by early population subdivision in several country-specific sub-epidemics.
MATERIALS AND METHODS

Study population. The study included plasma, serum or dried-blot-spots samples from 632 individuals who were diagnosed as HIV-1-positive in Belize, Honduras, El Salvador, Nicaragua, Costa Rica and Panama during 2002–2010. For the purpose of this study we generated 302 new sequences and included 330 published sequences, recently obtained from Honduras (24, 25), Belize (27) and Panama (2). The study included samples from general population in Honduras (n=305), Costa Rica (n=38), Nicaragua (n=52), Panama (n=37) and Belize (n=21); MSM and FSW from El Salvador (n=78 and n=38, respectively) and Honduras (n=36 and 27, respectively).

Ethical approvals were obtained from the Regional Medical Ethics Board in each Central American country, as well as in Sweden, and by the Associate Director for Science at Division of Global HIV/AIDS, CDC. All study participants signed an informed consent and completed a study questionnaire.

Pol gene sequencing. Sequence analysis of the protease (PR) and the first part of the reverse transcriptase (RT) of HIV-1 pol gene (nucleotides position 2268-3257 of reference strain HXB2; PR nucleotides 16-297 and RT nucleotides 1-708) was performed on the plasma or serum samples by using a published protocol (16, 24, 27). Dried-blood-spots samples were sequenced using another published genotyping assay developed by the CDC (47, 48).

HIV-1 subtype classification and analysis of drug resistance mutations. Subtype was determined by uploading sequences individually into the REGA HIV-1 Automated Subtyping Tool v2.0 (5). Subtyping was confirmed with phylogenetic analysis using Mega
v5 and the neighbor-joining method with GTR+G estimated distances and 1000 bootstrap replicates (40). Drug resistance mutations were defined on the basis of the standardized list of mutations for treated (14) and untreated patients (4).

Data sets for molecular epidemiology of HIV-1 group M subtype B. Detailed phylogenetic analyses were performed on 625 study subjects infected with HIV-1 group M subtype B. Non-subtype B sequences were too few for meaningful analyses. Subtype B control sequences were obtained in the following way: 1) Related sequences to each study sequence were retrieved by BLAST searches in Los Alamos HIV (LANL) database (www.hiv.lanl.gov), which generated 137 sequences after removal of duplicates; 2) All subtype B sequences spanning our PR+RT sequence fragment with known sampling year and country were downloaded from the LANL and Genbank (www.ncbi.nlm.nih.gov) (n=15,084). Sequences that had already been retained from the BLAST searches were omitted and from the remaining sequences we first selected most of the available strains from other Central American neighboring countries (Bahamas, Cuba, Dominican Republic, Haiti, Jamaica, Mexico, and Trinidad & Tobago) (n=57). Then we made three random selections of 740 sequences from the rest of the world in order to have approximately a 1:1 ratio with our Central America data set. Consequently, we analyzed three independent data sets in which our Central American sequences, sequences from the BLAST search and sequences from neighboring countries were kept constant, but the random sequences from the rest of the world varied. We also analyzed a fourth data set were all the available pol sequences from Haiti (n=21) were included. Finally, in order to avoid the ascertainment bias due to the overrepresentation of North American sequences in the HIV databases and better investigate the relationship between the Central American epidemic and the epidemic
in Haiti, which has been reported as the oldest HIV-1 epidemic in the Americas (11), we created five additional data sets, each one including: all available pol sequences from Haiti (n=21) and a random sub-sample of sequences from North America (42 strains), Central America (21 strains), South America (42 strains) and Europe/Oceania (42 strains). For each data set, two different outgroups were tested: HIV-1 group M subtype D sequences (K03554 and EF633445), or subtype C (U46016) and H (AF005496) sequences.

**Sequence alignment and phylogenetic analysis.** Multiple alignments were generated using MUSCLE (9), either using all positions or excluding codons associated with drug resistance, according to the 2010 International AIDS Society (IAS) list of NRTI, NNRTI, and PI mutations, with the exception of minor PI mutations (14).

Phylogenetic signal was tested by means of Xia’s test (46) and likelihood mapping analysis (38). Model selection was executed by using PhyML implemented in TOPALi (19). TVM+I+G was the model with the lowest Bayesian Information Criterion (BIC) (73322) followed by the GTR+I+G model (73344), whilst all other models had BICs > 20 points compared to GTR+I+G. Maximum-likelihood (ML) phylogenetic analysis was performed on the alignment. The parallel version of the FastTree software (30) with the GTR+I+G model (a posteriori 20-parameter gamma optimization) was used (since the TVM+I+G model is not implemented in the FastTree software). The tree topology search was carried out using a mix of nearest-neighbor interchanges and sub-tree/prune/re-graft moves. The reliability of each tree split was calculated by a Shimodaira-Hasegawa (S-H) test (36). For the smaller data sets we also inferred Bayesian trees with MrBayes, using the TVM+I+G model. For each data set, eight Markov Chain Monte Carlo (MCMC) runs were carried out in parallel for $10^7$ generation with sampling every 10,000 generations.
Convergence among independent runs was assessed by checking that the standard deviation of split frequencies was below 0.01 at the end of the runs (which was the case for each run). Maximum clade credibility trees were obtained from posterior distribution of trees with TreeAnnotator v1.6.1 included in the BEAST software package (8).

Epidemiological/transmission clusters were inferred using the phyloPart program (31), with a patristic distance threshold that would maximize the number of clusters and would not be higher that the 25th percentile of the overall patristic distance distribution (see Table S2 in the supplemental material). Only sub-trees of the phylogenetic tree with a p-value ≤0.1 as obtained by the Shimodaira-Hasegawa test were considered eligible for phyloPart clustering. Furthermore, a cluster was defined as a monophyletic clade only if it included at least 10 sequences.

**Molecular clock analysis.** To obtain a Bayesian estimate of HIV-1 group M subtype B origin in Central America, a data set was assembled including all sequences from the five Central American monophyletic clades, as well as several HIV-1 non-B, group M strains used as controls. The evolutionary rate (nucleotide substitutions per site per year) and the time of the most recent common ancestor (TMRCA) were inferred by the Markov Chain Monte Carlo (MCMC) approach implemented in the BEAST software package version 1.6.1 (8). The analysis was performed assuming either a strict or a relaxed molecular clock, with different coalescent priors (constant population size, and Bayesian skyline plot), and a uniform prior for the root height with mean and standard error chosen according to previously published estimates for group M origin (1921±13) (45). An MCMC chain was run for 100,000,000 generations with sampling every 10,000th generation. Proper mixing of the MCMC was assessed by calculating the effective sampling size (ESS) for each
parameter (8). All ESS values were >200, indicating sufficient mixing of the Markov chain.

The results were visualized in Tracer v.1.5.

Different clock models were compared by calculating the Bayes Factor (BF), which is the ratio of the marginal likelihoods of the two models (39). We calculated approximate marginal likelihoods for each model via importance sampling (1000 bootstraps) using the harmonic mean of the sampled likelihoods (with the posterior as the importance distribution). The difference (in loge space) of marginal likelihood between any two models is the loge of the Bayes Factor, $\log_e(BF)$. Evidence against the null model (i.e. the one with lower marginal likelihood) is indicated by: $6 > [2 \cdot \log_e(BF)] > 2$ (positive); $10 > [2 \cdot \log_e(BF)] > 6$ (strong) and $[2 \cdot \log_e(BF)] > 10$ (very strong). BF calculations were performed with Tracer v.1.5.
HIV-1 subtype classification. Almost all sequences (625 of 632; 98.9%) were classified as subtype B (Table 1). Five (0.8%) sequences were classified as subtype C, and two (0.3%) sequences had previously been identified as unique recombinant forms (URFs), URF_AD and URF_AK (24). Table 1 lists the baseline characteristics of the 632 patients. The median age of the participants was 29 years (range: 0 month – 71 years); the proportion of female and male was almost equal, 49% and 51%, respectively. The predominant transmission route was heterosexual (n=443; 70.6%), 346 of these patients were classified as belonging to the general population, whereas 65 patients were FSW. The remaining patients reported the following routes of transmission: homosexual (18.1%); mother-to-child (10.6%) and by blood products (0.6%). Two-thirds of the study participants (66%) reported that they never had been exposed to antiretroviral therapy (ART). Interestingly, 32 patients were Garifunas, a population descending from Caribe, Arawak and West African people living in relatively isolated communities, about which little HIV epidemiological data are available to date.

Antiretroviral drug resistance-associated mutations. Mutations associated with antiretroviral drug resistance were detected in 161 (26%) of the sequences: 130 (62%) among 211 individuals with previous ART and 31 (7.5%) among treatment naïve individuals. In the ART group the prevalence of resistance mutations to any NRTIs was 53%, to any NNRTIs was 49%, to any PIs was 23%, to the combination to at least one NRTI and at least one NNRTI was 42%, and to at least one drug from each of the three drug classes (NRTIs + NNRTIs + PIs) was 14%. The most commonly observed mutations
were: M184V (46%), T215Y (26%), M41L (23%), K103N (23%), D67N (18%) and K70R (16%).

We found that 31 of 417 (7.5%; 95% CI 5.1-10.4%) of the sequences from the treatment naïve patients had at least one mutation associated with transmitted drug resistance (TDR) (4). The prevalence of resistance mutations to any NRTI was 5.1%, to any NNRTI was 3.6%, to any PI was 0.2%. The most commonly observed mutations were: M41L (3.4%), K103N (2.4%), M184V (1.4%), P255H (1%), and T215Y (0.7%).

Five HIV-1 subtype B monophyletic clades in Central America. Detailed phylogenetic analyses were carried out on the 625 subtype B sequences. There was no evidence of substitution saturation in the full or the reduced (i.e. after exclusion of positions associated to drug resistance) alignment, as demonstrated by the Xia’s saturation test (p<0.0001). Likelihood mapping analysis showed that the exclusion of drug resistance mutations improved the phylogenetic signal (52.2% vs. 64.7% of the quartets localized in the center of the triangle). Therefore, all subsequent phylogenetic analyses were done on the reduced alignment.

To place the sequences from the six Central American countries in the broader context of the HIV-1 group M subtype B pandemic we inferred eight different ML phylogenetic trees. Each tree was rooted either with HIV-1 group M subtype D sequences or with subtype C and H sequences and included the 625 subtype B sequences from the six study countries and different subsets of subtype B control sequences obtained by a combination of BLAST searches, selection of sequences from neighboring countries, as well as random sampling of worldwide sequences (see Materials and Methods). As
expected, each tree showed in general a star-like topology with little resolution, but all trees resulted in identical clustering for the Central American sequences (not shown). The ML tree including all 21 available \textit{pol} sequences from Haiti and rooted with HIV-1 group M subtype D sequences is shown in Fig. 1. By using each of our 625 Central American sequences in a BLAST search, 100 of the 137 (73\%) sequences found were from the US, while very few sequences were found from other American countries (the origin of all control sequences is given in Table S3 in the supplemental material). However, it is difficult to draw firm (or any) conclusions from this finding because of the ascertainment bias due to the overrepresentation of US sequences in the databases.

The Phylopart analyses of the ML tree identified five statistically supported monophyletic clades formed almost exclusively by Central American sequences (Fig. 1 and supplemental material Fig. S1). These five clades were robust to the selection of worldwide control sequences or the sequences used as outgroup, as they were present also in the other ML trees (not shown). Clade I was very large and contained a majority of our Central American subtype B sequences (n=389, 62\%) (Fig. 1) and was represented in all six Central American study countries. Despite our efforts to find closely related database sequences using BLAST searches, Clade I was almost exclusively Central American and included only six non-Central American sequences (four from US and two from Canada).

There were four additional smaller sequence clusters that also fulfilled our criteria for statistically supported clades (Clades II-V). Clade II included 10 strains (five from Belize and five from Honduras); Clade III included 17 Central American sequences (15 from Costa Rica, one from Nicaragua and one from El Salvador); Clade IV included 14 sequences (12 from El Salvador and two from Honduras); and Clade V included 11
Nicaraguan sequences (Fig. 1). In the ML tree rooted with subtype D sequences (Fig. 1), clade III shared a most recent common ancestor with a sequence from Japan (see Fig. S1 in the supplemental material) and, in turn, with clade II and clade I. However, the monophyletic origin of these clades, indicating that clade II and III are just sub-clades of the major Central American clade I, should be taken with caution since support for this clustering pattern was extremely low (S-H test p<0.5) and the clustering was not present in the tree rooted with subtype C and H sequences (data not shown).

Multivariate testing showed that there were significant higher odds (p<0.05) for strains from MSM and mother to child transmission cases to be part of a cluster when compared to strains from heterosexual subjects. However, the distribution of risk behavior within the five Central American clades showed no significant differences compared to the complete data set. The remaining Central American sequences fell outside the five monophyletic clades and where either intermixed in the trees or belonged to additional clades that did not meet our criteria for significant support (see Methods).

The fact that a single well-supported clade included the 62% of the Central American sequences suggests that a single introduction was responsible for a majority of the current HIV-1 cases in the region. Moreover, the existence of other four smaller clades where strains belonged to either a single (clade V), two (clade II and IV) or three countries (clade III) may represent later and/or independent introductions of subtype B into Central America. There was no well-supported pattern clustering the Central American clades with control sequences from a specific region or country. In order to examine the positioning of our Central American sequences relative to Haitian sequences and the root of the tree, we needed to take into account the strong ascertainment bias in the pol alignment that included...
650 Central American sequences but only 21 (available) Haitian sequences, even though both epidemics are roughly the same size (200,000 individuals). To provide a proper comparison, phylogenies were inferred from five additional data sets, each one including the Haitian sequences and a random sub-sample of sequences from North America (42 strains), Central America (21 strains), South America (42 strains) and Europe/Oceania (42 strains). In each tree rooted with HIV-1 subtype D sequences, the Haitian strains were basal with high posterior probability in the Bayesian analysis (Fig. 2 and supplemental material Fig. S2). This is in agreement with Gilbert et al. (2007) scenario of a Haitian origin of HIV/AIDS in the Americas (11). Interestingly, the result could not be reproduced when more divergent sequences (subtype C and H) were used as outgroups (data not shown). The large Central American clade I usually clustered in proximity to US and Haitian strains (see Fig. S2 in the supplemental material), although the support was low (p<0.5) in each tree. Taken together our data indicate that the five HIV-1B Central American clades belong to the US/Pandemic subtype B clade.

**TMRCA and evolutionary rate of HIV-1 subtype B in Central America.** Different molecular clock models (strict and relaxed) were tested to infer the timing of the five HIV-1 subtype B clades introductions in Central America. The Bayes Factor (BF) using either a constant population size or Bayesian Skyline Plot (BSP) coalescent prior favored the relaxed over the strict clock model (see Table S4 in the supplemental material). Moreover, under the relaxed molecular clock, the coefficient of variation 95% highest posterior density (95%HPD) intervals did not include the value zero using either a constant (95%HPD 0.2818-0.4628) or BSP (95%HPD 0.2133-0.3339) coalescent prior, in agreement with a better fit of the relaxed over the strict clock model. The BSP coalescent
prior also performed significantly better than the constant population size prior both in the case of a strict (BF=13.3) or relaxed (BF=18.1) molecular clock. Therefore, the relaxed clock model with BSP prior was used to investigate the evolutionary rate and the time of the TMRCA of the five Central American HIV1-B clades. Several studies, on the basis of different methods, estimated HIV-1 group M TMRCA in 1921-1931 using env (15, 45) and in 1902-1939 (median 1920) using pol sequences (35), while the origin of subtype B has been placed around the early 1960s (7). These estimates were used as internal controls to verify the robustness of the Bayesian inference. The median evolutionary rate estimated for the five clades ranged from 8.0x10^{-4} to 1.1x10^{-3} nucleotide substitutions/site/year (Table 2). The TMRCA median estimates for Central American clade I was 1966, while clades II-IV estimates ranged from 1971 to 1976 (Table 2 and Fig. 1). However, 95% highest posterior density (95%HPD) intervals were rather large, possibly due to the strong star-like signal in the pol alignments (see Table S5 in the supplemental material), and placed the origin of the five Central American clades between the mid-1950s and the mid-1980s (Table 2).
The present study is the first in-depth description of HIV-1 molecular epidemiology in Central America based on a large number of samples collected in six of the seven Central American countries. The findings show that subtype B is predominant in the region, which is congruent with reports from neighboring countries such as Mexico, Colombia and Venezuela (6, 10, 33), as well as smaller reports from Central America (1, 17, 25, 41).

Interestingly, however, while HIV-1 group M subtype B strains worldwide usually intermix in a star-like phylogeny, indicative of the panmictic structure of the epidemic, strains from Central America appeared to be highly compartmentalized. The phylogenetic analysis showed that the majority of the current subtype B cases in Central America appeared to originate from a single early introduction (95%HPD 1955-1977). Furthermore, four additional statistically supported monophyletic clades consisting of at least 10 Central American sequences appeared to have evolved independently, suggesting the existence of discrete sub-epidemics within different countries, which originated from separate introductions. The fact that the largest clade included 62% of the strains may suggest that, of the possible multiple subtype B introductions in the region, one was very successful and gave rise to a regional epidemic, particularly affecting Honduras and El Salvador. The origin of the major Central American clade occurred relatively soon (median estimate: 1966) after the emergence of the most recent common ancestor of HIV-1 group M subtype B (median estimate 1957), which may explain in part the early compartmentalization of the epidemic. It is important to note, however, that our median estimate of HIV-1 group M subtype B TMRCA (1957), based on pol sequences, is one decade earlier than current estimates (1966) based on env (11). This may be due, at least in part, to the higher star-like
signal (phylogenetic noise) in the pol gene, resulting in an unrealistically deep divergence date, although the 95%HPD intervals (1944-1968) of our estimates are still overlapping with the ones reported in the literature. The existence of independent monophyletic clades also implies the presence of separate transmission networks for subtype B spread within the Central American region. Therefore, our data may provide valuable information for targeted intervention and prevention.

The five Central American clades appeared to belong to the US/Pandemic clade of HIV-1 group M subtype B (11), but there was no clear pattern clustering subtype B strains from Central America with specific sequences from the rest of the world. Interestingly, our analyses did show that some strains from Haiti were closely related to strains belonging to the major Central American clade, although statistical support was not significant, suggesting a possible common ancestor between these two epidemics. In addition, the Bayesian trees showed Haitian sequences basal to the HIV-1 group M subtype B phylogeny in agreement with the results of Gilbert et al. (2007) (11). It is also important to note that, in spite of the observed phylogenetic noise in the pol data sets analyzed in the present work, most of the Central American sequences still clustered within a highly supported monophyletic clade, which strengthens the main finding that a single introduction is responsible for the majority of the current cases in the region.

The observation that Central American clades and strains may have originated in the US/Haitian strains is expected and in agreement with known early epidemiological data, as well as other phylogenetic analyses (11), geographical proximity, known tourism and legal and illegal immigration patterns. However, given the low resolution in the phylogenetic trees and the unbalanced sampling from other American countries, no firm conclusions can
be drawn and the route of the earliest HIV-1 transmissions in Central America remains to be investigated.

We found evidence of TDR in 7.5% of the 417 treatment naïve patients from the six Central American countries, which is a moderate level according to WHO criteria (26). This is the first comprehensive report of TDR in Central America. The results have to be interpreted in light of the history of antiretroviral therapy in the countries, which initially often involved uncontrolled mono- and dual therapy with antiretroviral drugs that were sold illegally or sent from patients' relatives in the US, whereas in more recent years almost all therapy is provided as combination therapy through national treatment programs.

Some limitations of our study should be mentioned. First, we cannot exclude the possibility that the sampled population is not fully representative of all HIV-1 cases in the six Central American countries included in the study because the samples represent only a fraction of all HIV-infections in the region. Furthermore, sampling was unbalanced between countries. It is possible that a more balanced sampling would show additional clades and/or Clades II-V more widely spread across Central America than suggested by our analyses. Consequently, while our study is the most comprehensive investigation of HIV-1 spread in Central America to date, the full picture will require additional sampling from several other countries, including Guatemala that was not sampled at all. It is also possible that the coalescence times for Clades II-V would be more similar to Clade I (i.e. earlier TMRCAs) had they been more thoroughly sampled. Another limitation was the use of partial, rather than full, genome sequences, which could underestimate the proportion of recombinant strains and possibly may limit the possibilities to resolve completely the
deepest portions of the trees. However, our preliminary results suggest that the subtypes and clustering of Honduran strains is similar in analyses based on env V3 sequences.

In conclusion, this is the first comprehensive study of how and when HIV-1 has entered and spread in Central America, which is a region with a substantial HIV/AIDS burden. HIV-1 group M subtype B predominates, although sporadic non-B strains were identified. Phylogenetic and molecular clock analysis showed one major well-supported monophyletic cluster compatible with a single early subtype B introduction accounting for most current cases, as well as a subsequent expansion into regional sub-epidemics, which deserves further investigation in order to understand the ecological factors driving subtype B successful emergence and dissemination in Central America.
We thank in particular all study participants. A special thanks to Michael Worobey for providing the data sets analyzed in the Gilbert et al. (2007) paper and the two anonymous reviewers for their helpful comments. This study was funded by the Swedish International Development Cooperation Agency (Sida) on behalf of a bilateral collaboration with the National Autonomous University of Honduras, the Network for Research and Training in Tropical Diseases in Central America (NeTropica) under the project No 06-R-2010 and the Swedish Medical Research Council.
REFERENCES


22


USAID. February, 2011. HIV/AIDS health profiles for the Central America Region.

USAID. April, 2011. HIV/AIDS health profiles for the Latin America and the Caribbean Region.


FIGURES LEGEND

FIG. 1. Maximum likelihood phylogenetic analysis of HIV-1 group M subtype B pol sequences. The data set included 625 HIV-1B strains from six Central American countries: Honduras (366), El Salvador (116), Nicaragua (52), Costa Rica (38), Panama (37), and Belize (17); 807 non-Central America control sequences downloaded from GenBank and the Los Alamos HIV database: 67 sequences from Central America neighboring countries and 740 sequences from the rest of the world (including 21 Haitian sequences). The tree was rooted by using subtype D sequences as outgroup. The tree showing the tip labels with the full name of each strain is given in Fig. S1 in the supplemental material. (A) The maximum likelihood tree shows five significantly supported monophyletic clades (numbers inside the monophyletic clades correspond to approximate likelihood-ratio test SH-like values), formed essentially by Central American sequences. Branches are scaled in nucleotide substitutions per site according to the bar at the bottom of the figure and colored to indicate the regional origin of each American strain according to the legend in the figure, while black branches represent strains from the rest of the world (Europe, Asia, Africa). (B) Distribution of the Central American sequences in each of the clades.

FIG. 2. Phylogenetic analysis of HIV-1 group M subtype B pol sequences using a sub-sample of the reference sequences. Bayesian maximum clade credibility tree inferred from a data set including the 21 available Haitian sequences and a random sub-sample of sequences from North America (42 strains), Central America (21 strains), South America (42 strains) and Europe/Oceania (42 strains). The numbers along the monophyletic branches correspond to approximate likelihood-ratio test SH-like values. Branch lengths in
nucleotide substitutions per site were scaled according to the bar at the bottom of each tree. Trees from four additional data sets with randomly selected sequences are given in Fig. S2 in the supplemental material.
Table 1. Characteristics of the Central America HIV-1 study subjects.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Belize</th>
<th>El Salvador</th>
<th>Honduras</th>
<th>Nicaragua</th>
<th>Costa Rica</th>
<th>Panama</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sequences</td>
<td>21 (3.3)</td>
<td>116 (18.4)</td>
<td>368 (58.2)</td>
<td>52 (8.2)</td>
<td>38 (6)</td>
<td>37 (5.9)</td>
<td>632 (100)</td>
</tr>
<tr>
<td>HIV-1 subtypes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>17 (81)</td>
<td>116 (100)</td>
<td>366 (99.5)</td>
<td>51 (98.1)</td>
<td>38 (100)</td>
<td>37 (100)</td>
<td>625 (98.9)</td>
</tr>
<tr>
<td>Non-B</td>
<td>4 (19)</td>
<td>0</td>
<td>2 (0.5)</td>
<td>1 (1.9)</td>
<td>0</td>
<td>0</td>
<td>7 (1.1)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>12 (57.1)</td>
<td>78 (67.2)</td>
<td>184 (50)</td>
<td>0</td>
<td>31 (18.6)</td>
<td>14 (37.8)</td>
<td>319 (50.5)</td>
</tr>
<tr>
<td>Female</td>
<td>9 (42.9)</td>
<td>38 (32.8)</td>
<td>184 (50)</td>
<td>52 (100)</td>
<td>7 (18.4)</td>
<td>22 (59.5)</td>
<td>312 (49.4)</td>
</tr>
<tr>
<td>Not known</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (2.7)</td>
<td>1 (0.1)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (range)</td>
<td>4* (0*-48*)</td>
<td>25 (18-52)</td>
<td>32 (2-68)</td>
<td>34 (19-57)</td>
<td>32 (20-71)</td>
<td>20 (1*-43)</td>
<td>29 (0*71)</td>
</tr>
<tr>
<td>Transmission route</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterosexual</td>
<td>0</td>
<td>38 (32.8)</td>
<td>290 (78.8)</td>
<td>52 (100)</td>
<td>38 (100)</td>
<td>28 (75.7)</td>
<td>446 (70.6)</td>
</tr>
<tr>
<td>Homosexual</td>
<td>0</td>
<td>78 (67.2)</td>
<td>36 (9.8)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>114 (18.1)</td>
</tr>
<tr>
<td>Mother-to-child</td>
<td>21 (100)</td>
<td>0</td>
<td>38 (10.3)</td>
<td>0</td>
<td>0</td>
<td>8 (21.6)</td>
<td>67 (10.6)</td>
</tr>
<tr>
<td>Blood products</td>
<td>0</td>
<td>0</td>
<td>4 (1.1)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4 (0.6)</td>
</tr>
<tr>
<td>Not known</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (2.7)</td>
<td>1 (0.2)</td>
</tr>
<tr>
<td>ARV exposure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>4 (19)</td>
<td>116 (100)</td>
<td>213 (57.9)</td>
<td>10 (19.2)</td>
<td>38 (100)</td>
<td>36 (97.3)</td>
<td>417 (66)</td>
</tr>
<tr>
<td>Yes</td>
<td>14 (66.7)</td>
<td>0</td>
<td>155 (42.1)</td>
<td>42 (80.8)</td>
<td>0</td>
<td>0</td>
<td>211 (33.4)</td>
</tr>
<tr>
<td>Not known</td>
<td>3 (14.3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (2.7)</td>
<td>4 (0.6)</td>
</tr>
<tr>
<td>Resistance mutations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No resistance</td>
<td>10 (58.8)</td>
<td>106 (91.4)</td>
<td>254 (69.4)</td>
<td>35 (68.6)</td>
<td>25 (65.8)</td>
<td>34 (91.9)</td>
<td>464 (74.2)</td>
</tr>
<tr>
<td>Any resistance</td>
<td>7 (41.2)</td>
<td>10 (8.6)</td>
<td>112 (30.6)</td>
<td>16 (31.4)</td>
<td>13 (34.2)</td>
<td>3 (8.1)</td>
<td>161 (25.8)</td>
</tr>
<tr>
<td>Resistant to any PI</td>
<td>0</td>
<td>1 (0.7)</td>
<td>47 (12.8)</td>
<td>2 (3.9)</td>
<td>0</td>
<td>0</td>
<td>50 (8)</td>
</tr>
<tr>
<td>Resistant to any NRTI</td>
<td>1 (5.9)</td>
<td>4 (3.4)</td>
<td>99 (27)</td>
<td>11 (21.6)</td>
<td>12 (31.6)</td>
<td>3 (8.1)</td>
<td>130 (20.8)</td>
</tr>
<tr>
<td>Resistant to any NNRTI</td>
<td>7 (41.2)</td>
<td>9 (7.8)</td>
<td>97 (26.5)</td>
<td>15 (29.4)</td>
<td>2 (5.3)</td>
<td>1 (2.7)</td>
<td>131 (21)</td>
</tr>
<tr>
<td>Sequences per clade</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clade I</td>
<td>5 (23.8)</td>
<td>54 (46.5)</td>
<td>321 (87.2)</td>
<td>3 (5.8)</td>
<td>2 (5.3)</td>
<td>1 (2.7)</td>
<td>386 (61.1)</td>
</tr>
<tr>
<td>Clade II</td>
<td>5 (23.8)</td>
<td>0</td>
<td>5 (1.4)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10 (1.6)</td>
</tr>
<tr>
<td>Clade III</td>
<td>0</td>
<td>1 (0.9)</td>
<td>14 (3.8)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15 (2.4)</td>
</tr>
<tr>
<td>Clade IV</td>
<td>0</td>
<td>13 (11.2)</td>
<td>1 (0.3)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14 (2.2)</td>
</tr>
<tr>
<td>Clade V</td>
<td>0</td>
<td>0</td>
<td>11 (21.2)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 (1.7)</td>
</tr>
<tr>
<td>Outside clades I-V</td>
<td>11 (52.4)</td>
<td>48 (41.4)</td>
<td>27 (7.3)</td>
<td>38 (73)</td>
<td>36 (94.7)</td>
<td>36 (97.3)</td>
<td>196 (31)</td>
</tr>
</tbody>
</table>

Data are no. (%) of patients; *Age in months; PI, protease inhibitor; NRTI, nucleoside reverse transcriptase inhibitor; NNRTI, non-nucleoside reverse transcriptase inhibitor.
Table 2. Bayesian estimates of the mean time to the most recent common ancestor (tMRCA) for the HIV-1 subtype B sequences.

<table>
<thead>
<tr>
<th>Clades</th>
<th>tMRCA *</th>
<th>95% HPD b</th>
<th>Evolutionary rate c (95%HPD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clade I</td>
<td>1966</td>
<td>1955-1977</td>
<td>9.0x10^{-4} (4.0x10^{-4}-1.4x10^{-3})</td>
</tr>
<tr>
<td>Clade II</td>
<td>1974</td>
<td>1964-1983</td>
<td>1.0x10^{-3} (5.0x10^{-4}-1.6x10^{-3})</td>
</tr>
<tr>
<td>Clade III</td>
<td>1971</td>
<td>1961-1980</td>
<td>1.1x10^{-3} (6.0x10^{-4}-1.8x10^{-3})</td>
</tr>
<tr>
<td>Clade IV</td>
<td>1976</td>
<td>1967-1985</td>
<td>8.0x10^{-4} (4.0x10^{-4}-1.4x10^{-3})</td>
</tr>
<tr>
<td>Clade V</td>
<td>1973</td>
<td>1964-1983</td>
<td>1.1x10^{-3} (6.0x10^{-4}-1.7x10^{-3})</td>
</tr>
<tr>
<td>Global B</td>
<td>1957</td>
<td>1944-1968</td>
<td>1.1x10^{-3} (6.0x10^{-4}-1.7x10^{-3})</td>
</tr>
<tr>
<td>Group M</td>
<td>1921</td>
<td>1908-1933</td>
<td>9.0x10^{-4} (4.0x10^{-5}-1.4x10^{-3})</td>
</tr>
</tbody>
</table>

a. Median estimate of tMRCA.
b. HPD: lower and upper boundaries in 95% highest posterior density intervals.
c. Median evolutionary rate (nucleotide substitutions/site/year) and 95% highest posterior density (95%HPD) intervals in parenthesis.