LA-UR- 09-00327

Approved for public release;
distribution is unlimited.

Title:	Emergence of Recombinant forms in geographic regions with co-circulating HIV subtypes in the dynamic HIV-1 Epidemic (Tentative Tile)
Author(s):	M. Zhang, Z#: 188651, T-6/T-Division T. Leitner, Z#: 120084, T-6/T-Division B. Korber, Z#: 108817, T-6/T-Division B. Foley, Z#: 097029, T-6/T-Division
Intended for:	Journal of Virology

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

EMERGENCE OF RECOMBINANT FORMS IN GEOGRAPHIC REGIONS WITH CO-CIRCULATING HIV SUBTYPES IN THE DYNAMIC HIV-1 EPIDEMIC (TENTATIVE TITLE)

Ming Zhang[1,2]*, Brian Foley[1], Anne-Kathrin Schultz[3], Jennifer Macke[1], Mario Stanke[3], Burkhard Morgenstern[3], Bette Korber[1,4], Thomas Leitner[1]*

[1] Theoretical Division, [2] Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [3] Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik Goldschmidtstraße 1, 37077 Göttingen, Germany [4]The Santa Fe Institute, Santa Fe, NM 87501, USA

RUNNING TITLE: ***Old and new HIV recombinants co-circulate***

Abstract: xxx words

Text: xxx words

Running title: xx characters (with space)

*Corresponding authors. Mailing address: T-10, MS K710, Los Alamos National Lab, Los Alamos, NM 87544, USA. Phone: 505-665-6604 (Ming Zhang), and 505-667-3898 (Thomas Leitner). Fax: 505-665-3493. E-mail: Ming Zhang (mingzh@lanl.gov) and Thomas Leitner (tkl@lanl.gov).

ABSTRACT (TENTATIVE)

We have re-examined the subtype designations of ~10,000 subtype A, B, C, G, and AG, BC, BF recombinant sequences. We compared the results of the new analysis with their published designations. Intersubtype recombinants dominate HIV epidemics in three different geographical regions. The circulating recombinant from (CRF) CRF02_AG, common in West Central Africa, appears to result from a recombination event that occurred early in the divergence between subtypes A and G, although additional more recent recombination events may have contributed to the breakpoint pattern in this recombinant lineage as well. The Chinese recombinant epidemic strains CRF07 and CRF08, in contrast, result from recent recombinations between more contemporary strains. Never-the-less, CRF07 and CRF08 contributed to many subsequent recombination events. The BF recombinant epidemics in two HIV-1 epicenters in South America are not independent and BF epidemics in South America have an unusually high fraction of unique recombinant forms (URFs) that have each been found only once and carry distinctive breakpoints. Taken together, these analyses reveal a complex and dynamic picture of the current HIV-1 epidemic, and suggest a means of grouping and tracking relationships between viruses through preservation of shared breakpoints.

INTRODUCTION

Retrovirus recombination results from strand switching during reverse transcription between two RNA copies co-packaged in one virion (reviewed in (46)). Recombination in lentiviruses introduces rapid large genetic alternations (31, 32, 69), and can repair genome damage (15, 64). Occurring at an estimated rate of at least 2.8 crossovers per genome per cycle (79), recombination between HIV-1 subtypes may result in virulence changes at the epidemic level (55, 56), and may contribute to patterns of immune escape or act as an efficient approach for selecting variants resistant to HIV-1 specific drugs and immune pressure within a single host (33, 44, 48, 72).

It has been estimated that at least 20% HIV-1 isolates sequenced worldwide are inter-subtype recombinants (26, 49-51). Those recombinants are classified into two categories, CRFs (circulating recombinant forms) and URFs (unique recombinant forms), referring to recombinants that have established recurrent and transmitted forms in populations, and to those only identified in one individual, respectively (58). To define a CRF, at least 3 near full length sequences that share the same mosaic genomic breakpoint structure must be obtained from epidemiologically unlinked patients (58). The CRFs are numbered sequentially in the order in which they are first adequately described in the peer-reviewed literature. Currently, more than 40 CRFs have been identified worldwide (http://www.hiv.lanl.gov, CRF page) as well as more than 100 URFs (http://www.hiv.lanl.gov, sequence database search interface), and these numbers are still increasing as an outcome of new infections and HIV-1 entering new social networks,

64 large-scale near full-length genome sequencing, and the availability of more advanced
65 recombination detection techniques.
66
67 CRF02_AG is estimated to have caused at least 9 million infections worldwide (43). It is
68 thus a very successful lineage in Africa, and it continues to diverge and contribute to
69 novel recombinants. First identified in Nigeria in 1994 (30), but then later in samples
70 from 1984 from the Democratic Republic of the Congo (34), it is now the most prevalent
71 strain in West and West Central Africa. In Cameroon, where the original HIV-1 M group
72 zoonotic transmissions are believed to have taken place (21, 24), CRF02 was already
73 prevalent in the early 1990s (27), and it is currently the dominant lineage (12, 19). It is
74 possible that CRF02's high prevalence in Africa is explained by its long presence in the
75 epidemic. Comparing the genetic diversity within CRF02 to that of within pure subtypes,
76 Carr *et al* suggested that CRF02 may be as old as the pure subtypes (10). A recent study
77 further suggested that CRF02_AG was the parent of subtype G (2). Both studies suggest
78 that the CRF02 lineage was established early in the epidemic, and here we evaluate these
79 observations further using jpHMM and the database of available A, G, and CRF02
80 genomes.
81
82 The most common BC recombinants are CRF07_BC and CRF08_BC. CRF07 was first
83 identified in the Xinjiang province of China in 1997 (21, 52, 65). It is believed to have
84 migrated to Xinjiang along a northern drug trafficking route (52, 76). CRF08 is a
85 predominant subtype among intravenous drug users (IDUs) in Guangxi and the east part
86 of the Yunnan province in China (52, 77). These CRFs presumably were generated in

87 Yunnan, the epicenter of the AIDS epidemic in China where subtypes B and C were co-
88 circulating in the early 1990s (8, 39, 60, 61), or in Myanmar and then imported from
89 there into China (14, 37, 67, 68). The uneven distribution of CRF07, CRF08 in Yunnan
90 suggests the presence of independent transmission networks and clusters among IDUs in
91 Yunnan (76). It has not been established if other BC recombinants in Myanmar and
92 China regions are epidemiologically linked to CRF07 and CRF08 HIV-1 epidemic in
93 Southern China (13, 38).
94
95 AG recombinants in western Africa are most often seen as CRF02s, which are more
96 frequently identified than URFs, but there are many regional exceptions in Africa to this
97 pattern (29) (74). BC recombinants in China are dominated by CRF07 and CRF08. In
98 contrast, BF recombinants in South America are dominated by a large number of URFs
99 (http://www.hiv.lanl.gov, geography page). The origin of BF recombinants in South
100 America is not clear, but it seems as at least one of the main introductory routes of HIV-1
101 into South America is through Brazil (4). BF recombinants in South America are
102 represented by a disperse distribution radiating from at least two genetic centers. One is
103 represented by CRF12_BF and related genomes that are more frequently found in
104 Argentina; and the other by CRF28_BF and CRF29_BF, and a collection of BF URFs
105 that have been found in Brazil (17). Recently, a Bayesian hierarchical method analysis of
106 CRF12_BF indicated extensive ongoing recombination among CRF_12 viruses (41).
107
108 Accurate subtyping and recombination identification techniques are important to the HIV
109 field for many reasons, including epidemiological tracking, targeting vaccines to regional

110 epidemics, and defining potential phenotypic differences in different subtypes or inter-
111 subtype recombinants (57)). Here we use the jumping profile hidden Markov model
112 (jpHMM) method (59, 78) as one of several approaches we are incorporating into an
113 automated procedure to re-subtype 250,000 sequences in the Los Alamos HIV sequence
114 database. In jpHMM, each HIV-1 subtype is defined by a profile hidden Markov model
115 and all profile models are connected by empirical probabilities, allowing the detection of
116 possible recombinants and related breakpoints by jumping from one profile to another.
117 JpHMM performs best in predicting recombinants that involve subtypes that have had
118 adequate sampling and that thus have well-informed profiles, i.e., not subtypes H, J, and
119 K, because too few genome sequences are available from those subtypes (N=3, 3, and 2,
120 respectively) (59, 78). JpHMM is also effective at identifying breakpoint positions and is
121 computationally fast, allowing thousands of comparisons (78). In the present study,
122 jpHMM was used to detect the recombination patterns in recombinants that are
123 exclusively composed of subtypes A and G, or B and C, or B and F; and each subtype
124 considered here has enough data to form a good model of sequence variation.
125

MATERIALS AND METHODS

Sequences

The following sequence sets were retrieved from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov sequence database search interface). Set 1: All near full-length sequences (>7000 nucleotides [nt]) of subtype A, B, C, F, G, CRF02, CRF07, CRF08, CRF12, CRF17, CRF28, CRF29, and all URFs composed exclusively of subtypes A and G, or BC, or BF. Set 2: Fragments of HIV sequences that are between 300 nt and 7000 nt including all BC recombinants from Asia and BF from South America. Set 3: For additional BC analyses, worldwide sampling of two additional fragments were also retrieved from the database. These two fragments were the longest subtype B section shared by all near full-length BC recombinants (HXB2 positions 3497-4473), and the longest subtype C section shared by all near full-length BC recombinants (HXB2 positions 6582-7349). To avoid redundancy and reduce issues related to non-independence of data points, only 1 sequence per patient was included in the analyses of the fragment sequences. All sequences were aligned with the HIV-1 subtype reference sequences (http://www.hiv.lanl.gov sequence alignment page) using the Gene Cutter tool (http://www.hiv.lanl.gov Gene Cutter page). Alignment quality was checked manually using BioEdit (1) to ensure the alignments did not contain obvious problems and that they were correctly codon aligned. Risk factor information for the near full-length BF recombinant sequences from South America was also retrieved from the database.

Recombination detection and phylogenetic analyses

147 The jpHMM program (59, 78) was used to analyze the subtype assignment of the
148 aforementioned sequences. For the recombinants detected, jpHMM also provides detailed
149 information of subtype composition and breakpoint locations. The jpHMM source code is
150 available at jpHMM Web interface http://jphmm.gobics.de. The near full-length
151 sequences were grouped together if the sequences had similar subtype composition and
152 breakpoint patterns. Sub-genomic regions delimited by shared breakpoints in the majority
153 of AG recombinants (including jpHMM-confirmed CRF02 and AG URFs) were further
154 subjected to phylogenetic analysis. PhyML (25) was used to build maximum likelihood
155 (ML) trees using a GTR model with gamma distributed rates across sites. Similarly,
156 among the BC recombinants, sub-regions delimited by shared jpHMM-confirmed
157 breakpoints of CRF07 and CRF08 and further sub-regions shared by CRF07, CRF08, and
158 most BC URFs, were analyzed by ML tree reconstructions. In addition, subtype B and C
159 sequences, collected worldwide, of the largest identified B and C genomic sub-regions in
160 all near full-length BC recombinants were included in large tree analyses using neighbor
161 joining (under a F84 model). In the BF recombinants set, sub-regions delimited by shared
162 breakpoints between jpHMM-confirmed CRF12, CRF28, and CRF29 were subjected to
163 ML analysis. The statistical robustness and the reliability of the clustering patterns were
164 evaluated by non-parametric bootstrap analyses in PAUP (66) (neighbor-joining, F84
165 model, 1000 replicates) A bootstrap value of ≥70% was considered significant for
166 subtype clustering (28).
167
168 **CRF02 origin detection**

169 In addition to the ML tree analyses, we used Recombination Identification Program
170 version 3 (RIP3; http://www.hiv.lanl.gov RIP page) to examine the relationship between
171 CRF02 and contemporary sequences and inferred ancestor sequences of subtypes A and
172 G. Also, a CRF02 consensus sequence was analyzed against an alignment that included
173 maximum likelihood inferred ancestral sequences (22) (M group, A1, and G) and
174 consensus sequences (M group, A1, and G). Near full-length CRF02 sequences
175 confirmed by jpHMM results and phylogenetic analysis were used to build the CRF02
176 consensus using Consensus Maker (http://www.hiv.lanl.gov Consensus Maker page).
177 Other consensus and ancestor sequences were retrieved from the HIV sequence database
178 alignment page. All consensus and ancestor sequences were aligned using Gene Cutter,
179 followed by manual editing.
180
181 **Breakpoint frequency calculations**
182 Breakpoint frequency calculations were performed in the following sequence sets: 1)
183 Near full-length BC and BF recombinant sequences; 2) Fragmental BC recombinant
184 sequences from Asia; and 3) Fragmental BF recombinant sequences from South America.
185 The subtyping and recombination patterns in these sequences were based on jpHMM.
186 The breakpoint frequencies of all sequences in each alignment were calculated and
187 plotted. In BC CRFs, > 95% breakpoints were 16 nt off breakpoint median. In BF CRFs,
188 > 95% breakpoints were 98 nt off breakpoint median. These two numbers (16 nt and 98
189 nt) were used as breakpoint certainty regions for BC and BF CRFs, respectively.
190

RESULTS

Subtype assignment and CRF grouping using jpHMM

In total, 9435 near full-length and fragmental sequences from the Los Alamos HIV sequence database were reevaluated (Table 1). Overall in ≈95% of the assignments, the jpHMM subtyping results were consistent with the database assigned subtype (generally taken from the primary literature). The classification in the current database of sequence fragments was more often in disagreement with our jpHMM results than the near full-length sequences were. The largest disagreement occurred for BC and BF fragments, where up to 60% of the CRF08 fragment were assigned differently using jpHMM as compared to the original author assignments (Table 1). This difference is, however, not as dramatic as it may seem; all of these differences were explained by the fact that the fragments were in fact pure subtypes in their sequenced parts. Thus, it becomes a philosophical question which assignment that is best, i.e., "CRF08" or "C" or "B". Given that we do not know for sure what the subtype is in un-sequenced regions, we will only assign the sequences based on the information we have, e.g., a C fragment will be assigned C even if it is suggested that this C is closer to the C in CRF08 than to a pure C (note also that this distinction not always can be done). Next, we grouped near full-length AG, BC, and BF recombinants into common groups if the sequences had similar genomic structure and breakpoints (Fig. 1). Our results suggested that revisions of some CRF designations may be needed. For instance, some database-assigned BF CRF sequences in this analysis appear to be unique BF URFs (Fig. 1C). We noted that more sequences are needed to confirm the "circulating" designation of CRF17, which does not fulfill the currently accepted minimal criterion of 3 independent cases (58); only two near full-

213 length sequences have been found in this "CRF" and they are epidemiologically linked
214 (9). The CRF and URF sequences described below refer to the sequences that have their
215 recombinant status confirmed by jpHMM.
216
217 **CRF02 is a recombinant of old and contemporary subtypes A and G**
218 To examine the evolutionary relationships among recombinants that are exclusively
219 composed of subtypes A and G, as well as their relationships with all subtype A and G
220 sequences, first we performed phylogenetic analyses in 8 common sub-regions (Fig. 2C,
221 regions I-VIII) delimited by the breakpoints of the majority of the 53 AG sequences
222 depicted in figure 1A. Second, the phylogenetic analyses were performed in four smaller
223 sub-regions (regions I', II', V', and VI'). Again, each of these smaller sub-regions was
224 delimited by as many AG recombinants as possible that didn't have additional
225 breakpoints inside the examined (smaller) sub-region. The results of the smaller sub-
226 regions (regions I', II', V', and VI') were used to verify the results of the bigger sub-
227 regions (regions I, II, V, and VI).
228
229 The maximum likelihood trees of the different sub-regions are shown in figure 2A. All
230 CRF02, as well as some additional AG recombinant sequences, always clustered together
231 regardless of subtype and genomic region, but not with the same subtype in all regions,
232 thus indicating a common and recombinant origin (Fig. 1A group 1-4 and some
233 sequences in the URF group). Interestingly, some regions suggest that CRF02 is an old
234 recombinant clade derived from ancient representatives of subtypes A and G. There, the
235 CRF02 clade is a sibling lineage to contemporary subtype A and G sequences (sibling of

236 A in regions I, I', III; and sibling of G in regions II, II', V'. Fig. 2 A and B). The topology
237 of the trees also suggest that the current CRF02 has undergone multiple recombination
238 events, and some genomic regions of the first generation of CRF02 sequences were
239 replaced by more contemporary sequences (Regions V, VI, VI' are descendent lineages
240 of A; Region VII are descendent lineages of G. Fig. 2 A and B). Of particular interest, it
241 has been previously suggested that region IV of CRF02 is a parent of contemporary G
242 (2). In our analysis, however, both the ML tree (Fig. 2 A and B) and the RIP results (Fig.
243 2E) of region IV demonstrate that this fragment of CRF02 was derived from an ancestral
244 G sequence. A RIP analysis (Fig. 2E) comparing maximum likelihood estimates of
245 ancestral sequences of the A and G clades with consensus sequences derived from
246 contemporary A and G isolates, further supported that some sections of the CRF02
247 genome may have involved old recombination events from a time when the clades were
248 beginning to diverge, and other regions that involved more recent subtype A and G
249 sequences.
250
251 Importantly, the inference of whether CRF02 was a descendant or sibling linage to A or
252 G (but never parent) was supported by high bootstrap values (≥70%). Also, the alignment
253 quality across the genome was fairly even, i.e., the gap count did not adversely affect the
254 phylogenetic signal more in some regions than in others (Suppl Fig 1).
255
256 **The Chinese BC epidemic involves subtypes circulating in China and neighboring**
257 **countries**

To characterize the relationships among BC recombinants from China, Asia, and worldwide, we first investigated the relationship between CRF07 and CRF08. Sequences classified as CRF07 or CRF08 are summarized in figure 1B, and ML trees were constructed for sub-regions delimited by all CRF07 and CRF08 sequences (Fig. 3). While most of the examined sub-regions show a sibling relationship between CRF07 and CRF08, two sub-regions (HXB2 positions 794-2064 and 2547-2846) suggest that, at least these parts of, CRF08 could be the parent of CRF07 because CRF07 sequences are clustered inside the CRF08 clade (bootstrap support >70). Further, CRF07 and CRF08 were derived from multiple recombination events, as indicated by unequal breakpoint frequencies in CRF07 and CRF08 (Fig. 4 BC recombinant panel). However, the breakpoint at HXB2 position 8866 was consistent among CRF07 and CRF08 and subsequent recombinants, and thus was likely introduced into the common ancestor of CRF07 and CRF08.

To investigate Chinese BC recombinants and BC recombinants from China's neighboring countries, phylogenetic analyses were performed on consensus sub-regions delimited by most near-full-length BC recombinants in figure 1B. The results, not shown here, demonstrated a close relationship between Yunnan B and Myanmar B. Limited sampling from these two geographic regions (Yunnan BC: 6 sequences; Myanmar BC: 2 sequences), however, prevented us from deducing the direction in which B had moved between Yunnan and Myanmar.

280 Finally, the influence of worldwide B and C epidemics on the Chinese BC recombinants
281 was analyzed. For this, subtype B sequences from worldwide were retrieved from the
282 HIV database in the genomic region that was the biggest subtype B sub-region shared by
283 all CRF07, CRF08 and most near full-length BC recombinants (HXB2 positions 3497-
284 4473). Similarly, the biggest subtype C sub-region shared by all CRF07, CRF08, and
285 most near full-length BC recombinants was used to retrieve all subtype C sequences
286 worldwide (HXB2 6582-7349). One sequence per patient was included in both sets. In
287 the sub-region subtype B tree China B appears to be a local epidemic only involving
288 neighboring countries Thailand and Myanmar (Suppl Fig 2), possibly through drug
289 trafficking routes (52, 76). Other Asian countries, for instance, Korea, Japan, and
290 Thailand, appear to have had more frequent contacts with each other, suggesting multiple
291 HIV introductions in these countries. Finally, South America seems to have had multiple
292 HIV contacts with Europe and North America. The sub-region subtype C tree also
293 suggests that China C is a local epidemic, with C moving in from India, while India C has
294 multiple contacts with Africa (Suppl Fig 2). Finally, South America C appears to come
295 from a single introduction from Africa (Suppl. Fig 2, and 7, 20).
296
297 **Contemporary Argentinean and Brazilian HIV epidemics are not independent**
298 Our results of the phylogenetic analyses of CRF12, CRF28, and CRF29 are consistent
299 with the results reported elsewhere (9, 17) thus the data is not shown here. The
300 breakpoint frequencies of all near full-length BF sequences are summarized in figure 4,
301 BF recombination panel. Although the HIV-1 epidemic in Argentina is represented by
302 CRF12, and in Brazil by CRF28 and CRF29, all BF breakpoints that were identified were

303 found in more than one country, including near full-length (Fig. 4 BF panel) and
304 fragmental BF sequences (Suppl. Fig 3). These frequently shared breakpoints indicate a
305 BF epidemic that has moved back and forth between Argentina and Brazil. Furthermore,
306 the BF non-full length genome fragments carry the information that fills the gap between
307 the two extremes of BF CRFs represented in Argentina and Brazil, enabling us to track
308 the movement of the lineages between these countries. Finally, sequence V62 (accession
309 number AY536236) had the same genomic structure and breakpoints as CRF28:
310 Accordingly, V62 was included as a CRF28 sequence in our analysis. Previously, V62
311 was assigned as a URF, as it was submitted to the database before CRF28 and CRF29
312 were identified. In contrast to the other two CRF28 sequences that were sampled from
313 Brazil, sequence V62 came from a patient in Argentina (62), and is a further single
314 example illustrating movement of HIV between the Argentina and Brazil. Thus, the HIV
315 epidemics in Argentina and Brazil are not independent.
316
317 Next, we did not find evidence for that Argentinean B and F were derived from Brazil. ,
318 as has previously been suggested ((62, 70)). The trees, which agreed with previous
319 publications ((17, 18, 23)), showed that B and F fragments from CRF12, CRF28, and
320 CRF29 were mingled together, and thus could not support a single direction of HIV-1
321 flow. Also, we found that Argentinean B and F fragmental sequences in the HIV database
322 cover the full HIV-1 genome of each subtype, meaning that there was potential to form
323 any B/F recombinant in Argentina and that there was no need to import already
324 recombined genomes from Brazil *per se*. In addition, a recently identified near full-length
325 Argentinean pure F sequence, ARE933 (accession number DQ189088), was found to be

326 closer to Argentinean BF than any other F strains (3, 4). Thus, we cannot rule out the
327 possibility that Argentinean BF recombinants were formed in Argentina rather than
328 imported from Brazil, and further that the direction might have been from Argentina to
329 Brazil, or generated in Uruguay and spread both north and south from there ((73). It is
330 also possible that the shared breakpoints among Argentinean and Brazilian BF
331 recombinants may be indicative of breakpoint hot spots. Overall, the most likely scenario
332 is that there were HIV-1 transmissions in both directions, with recombination of
333 circulating strains in all three countries.
334
335

DISCUSSION

Here we present a large-scale sequence re-subtyping effort of 9435 HIV-1 sequences that involve subtypes A, B, C, G, and F. We found strong evidence that the contemporary HIV-1 epidemic has recombinants mixed with strains of old and new origin, and that shared breakpoints can be used for tracking patterns in the epidemic.

We found that CRF02 is a complex recombinant. Its old origin, as well as the subsequent recombination events that occurred prior to the establishment of the contemporary CRF02 lineage, can easily confound the analysis of CRF02. It explains the low bootstrap values in some trees (Fig. 2A), and further, it explains why jpHMM and some other HIV-1 subtyping tools, mostly based on contemporary sequences, failed the CRF02 classification in some genomic regions. Of note, our phylogenetic analyses also show that evolutionary signals in smaller regions may have been easily lost in bigger regions.

We also showed that the BC epidemic in China is unique compared to most other Asian countries; CRF07 and CRF08 were recently introduced to the epidemic, but both have undergone multiple recombination events. Subtypes B and F in South America seem to appear earlier than B and C in China, with an estimated introduction time in the late 1960s and 1970s, respectively (5, 6, 9). Shared breakpoints in BF recombinants indicate that the HIV-1 epidemics in Argentina and Brazil are not independent, but it does not necessarily mean that B and F in Argentina originated in Brazil.

358 The current HIV-1 epidemic involves lineages that are composed by both old and recent
359 recombination events (Fig 5). Recombination involving early lineages in the epidemic,
360 involving clades that may still circulate in the current HIV-1 epidemic, imposes
361 difficulties in recombinant detection. Co-evolution of sites due to fitness constrains and
362 HLA imposed immune pressure giving rise to distinct but potentially coordinant patterns
363 of immune escape can also confound recombination analsysis. Sometimes the history of
364 old lineages can be recovered by extrapolating backward from surviving viruses (like
365 subtype E (11, 45)), while some lineages presumably can never be found (like lineage X
366 in Fig 5). Tracking the history of strains with a new origin is much easier and more
367 accurate, because most of the existing HIV-1 subtyping tools are based on contemporary
368 sequences.
369
370 The HIV-1 epidemic may display different features in different epidemiological settings
371 (16, 24, 40). In Africa where the HIV epidemic is of a predominantly heterosexual
372 character, the ancient history of CRF02, with its higher replicative capacity than some
373 contemporary subtypes (36, 47) and its high prevalence (42), makes CRF02 an active
374 participant in generating more complex recombinants, for instance, the newly identified
375 CRF36_cpx (53). BC recombinants in China will also continue to evolve. Super-infection
376 of IDUs by CRF07 and CRF08 viruses (76), as well as continual influx of B and C into
377 Yunnan from China's surrounding countries (54, 75), is currently contributing greatly to
378 the emergence of new BC recombinants, especially BC URFs. Another important factor
379 is the rapid transitions in the HIV-1 epidemic in some regions of China. In Yunnan,
380 subtype B was found to be the dominant subtype in the late 1980s, but it was soon

381 replaced by Thai B; in 1992, subtype C was found in this region, thus Thai B and C co-
382 circulated; in 1994, CRF01 was identified in Yunnan; in 2000 and 2001, subtype C was
383 not detected among IDU samples in the same region (37, 39, 75, 76). While some of
384 these apparent transitions in regional prevalence might have been a consequence of
385 sampling bias, still they trace an intriguing pattern of transitions. BF recombinants in
386 South America are possibly moving toward the direction of having more URFs,
387 considering a long circulation record of subtypes B and F in South America (5, 6, 9),
388 and/or a tight HIV-1 transmission network with high incidence rates found in some South
389 American regions that would favor an elevated number of dual or super infections (70). A
390 possible outcome of this dynamic pattern in evolution is that the pure subtype F may
391 disappear after being gradually diluted from the South American epidemic; it is currently
392 relatively rare. The geographic distribution of subtypes and recombinant lineages in any
393 epidemic is dynamic and difficult to predict. Complicated host related behavior and
394 social network structures and possibly viral factors dictate the molecular epidemiology
395 (35, 63, 71), where tracking the genetic lineages and patterns in recombination
396 breakpoints can shed light on these issues.
397
398 Based on the dynamic picture of the HIV-1 epidemic, it is likely that the current pure
399 subtypes are recombinants that were formed a long time ago, but because the "pure"
400 parental lineages have been lost, we cannot trace their origin any more. Thus the current
401 subtype nomenclature does not mean that "pure" subtypes as currently defined are not
402 consequences of earlier recombination events, rather that they serve as good background
403 references for use in studying the current HIV-1 epidemic and their relative genetic

404 relatedness provides a basis for understanding the immunological consequences of
405 diversity. Therefore most current tools are not well designed to infer old recombination
406 events or those that involve unknown parents. Current CRF nomenclature requires all
407 sequences of one CRF to bear identical or very similar breakpoints, and thus originate
408 from a single lineage of an initial recombinant form. Such breakpoints may be easily
409 blurred by the rapid substitution rate of HIV-1 as well as further recombination events.
410 Hence, the sequences defined in a CRF are merely snapshots of the dynamic HIV-1
411 evolution. One solution to this problem is to define "families" that track recombination
412 break points among sets that are composed of the same subtypes, but the recombinants'
413 genomic structures and breakpoints may not be identical due to successive recombination
414 events or our capacity to accurately describe them. Sequences would belong to one
415 family as long as they are closer to a defined central strain of that family than to any other
416 family, including "pure" subtypes. Using such a "family" concept makes it feasible to
417 dynamically track the HIV diversity and epidemiologically important families through
418 (evolutionary) time, regardless of their precise phylogenetic history.
419

ACKNOWLEDGMENTS

We greatly thank to Dr. William Fisher for his helpful discussions. This work was supported by an NIH-DOE interagency agreement (Y1-AI-8309).

REFERENCES:

1. **A.Hall, T.** 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser **41:**95-98.
2. **Abecasis, A. B., P. Lemey, N. Vidal, T. de Oliveira, M. Peeters, R. Camacho, B. Shapiro, A. Rambaut, and A. M. Vandamme.** 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J Virol **81:**8543-51.
3. **Aulicino, P. C., G. Bello, C. Rocco, H. Romero, A. Mangano, M. G. Morgado, and L. Sen.** 2007. Description of the first full-length HIV type 1 subtype F1 strain in Argentina: implications for the origin and dispersion of this subtype in South America. AIDS Res Hum Retroviruses **23:**1176-82.
4. **Aulicino, P. C., J. Kopka, A. M. Mangano, C. Rocco, M. Iacono, R. Bologna, and L. Sen.** 2005. Circulation of novel HIV type 1 A, B/C, and F subtypes in Argentina. AIDS Res Hum Retroviruses **21:**158-64.
5. **Bello, G., W. A. Eyer-Silva, J. C. Couto-Fernandez, M. L. Guimaraes, S. L. Chequer-Fernandez, S. L. Teixeira, and M. G. Morgado.** 2007. Demographic history of HIV-1 subtypes B and F in Brazil. Infect Genet Evol **7:**263-70.
6. **Bello, G., M. L. Guimaraes, and M. G. Morgado.** 2006. Evolutionary history of HIV-1 subtype B and F infections in Brazil. Aids **20:**763-8.
7. **Bello, G., C. P. Passaes, M. L. Guimaraes, R. S. Lorete, S. E. Matos Almeida, R. M. Medeiros, P. R. Alencastro, and M. G. Morgado.** 2008. Origin and evolutionary history of HIV-1 subtype C in Brazil. AIDS **22:**1993-2000.
8. **Beyrer, C., M. H. Razak, K. Lisam, J. Chen, W. Lui, and X. F. Yu.** 2000. Overland heroin trafficking routes and HIV-1 spread in south and south-east Asia. Aids **14:**75-83.
9. **Carr, J. K., M. Avila, M. Gomez Carrillo, H. Salomon, J. Hierholzer, V. Watanaveeradej, M. A. Pando, M. Negrete, K. L. Russell, J. Sanchez, D. L. Birx, R. Andrade, J. Vinoles, and F. E. McCutchan.** 2001. Diverse BF recombinants have spread widely since the introduction of HIV-1 into South America. Aids **15:**F41-7.
10. **Carr, J. K., M. O. Salminen, J. Albert, E. Sanders-Buell, D. Gotte, D. L. Birx, and F. E. McCutchan.** 1998. Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants. Virology **247:**22-31.
11. **Carr, J. K., M. O. Salminen, C. Koch, D. Gotte, A. W. Artenstein, P. A. Hegerich, D. St Louis, D. S. Burke, and F. E. McCutchan.** 1996. Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. J Virol **70:**5935-43.
12. **Carr, J. K., J. N. Torimiro, N. D. Wolfe, M. N. Eitel, B. Kim, E. Sanders-Buell, L. L. Jagodzinski, D. Gotte, D. S. Burke, D. L. Birx, and F. E. McCutchan.** 2001. The AG recombinant IbNG and novel strains of group M HIV-1 are common in Cameroon. Virology **286:**168-81.
13. **Chang, S. Y., W. H. Sheng, C. N. Lee, H. Y. Sun, C. L. Kao, S. F. Chang, W. C. Liu, J. Y. Yang, W. W. Wong, C. C. Hung, and S. C. Chang.** 2006. Molecular epidemiology of HIV type 1 subtypes in Taiwan: outbreak of HIV type

1 CRF07_BC infection in intravenous drug users. AIDS Res Hum Retroviruses **22:**1055-66.

14. **Chu, T. X., and J. A. Levy.** 2005. Injection drug use and HIV/AIDS transmission in China. Cell Res **15:**865-9.

15. **Clavel, F., M. D. Hoggan, R. L. Willey, K. Strebel, M. A. Martin, and R. Repaske.** 1989. Genetic recombination of human immunodeficiency virus. J Virol **63:**1455-9.

16. **Clewley, J. P.** 2004. Phylodynamics: a conjunction of epidemiology and evolution? Commun Dis Public Health **7:**83-5.

17. **De Sa Filho, D. J., M. C. Sucupira, M. M. Caseiro, E. C. Sabino, R. S. Diaz, and L. M. Janini.** 2006. Identification of two HIV type 1 circulating recombinant forms in Brazil. AIDS Res Hum Retroviruses **22:**1-13.

18. **de Souza, A. C., C. M. de Oliveira, C. L. Rodrigues, S. A. Silva, and J. E. Levi.** 2008. Short communication: Molecular characterization of HIV type 1 BF pol recombinants from Sao Paulo, Brazil. AIDS Res Hum Retroviruses **24:**1521-5.

19. **Fonjungo, P. N., E. N. Mpoudi, J. N. Torimiro, G. A. Alemnji, L. T. Eno, J. N. Nkengasong, F. Gao, M. Rayfield, T. M. Folks, D. Pieniazek, and R. B. Lal.** 2000. Presence of diverse human immunodeficiency virus type 1 viral variants in Cameroon. AIDS Res Hum Retroviruses **16:**1319-24.

20. **Fontella, R., M. A. Soares, and C. G. Schrago.** 2008. On the origin of HIV-1 subtype C in South America. AIDS **22:**2001-11.

21. **Gao, F., D. L. Robertson, C. D. Carruthers, S. G. Morrison, B. Jian, Y. Chen, F. Barre-Sinoussi, M. Girard, A. Srinivasan, A. G. Abimiku, G. M. Shaw, P. M. Sharp, and B. H. Hahn.** 1998. A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J Virol **72:**5680-98.

22. **Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber.** 2002. Diversity considerations in HIV-1 vaccine selection. Science **296:**2354-60.

23. **Gomez-Carrillo, M., S. Pampuro, A. Duran, M. Losso, D. R. Harris, J. S. Read, G. Duarte, R. De Souza, L. Soto-Ramirez, and H. Salomon.** 2006. Analysis of HIV type 1 diversity in pregnant women from four Latin American and Caribbean countries. AIDS Res Hum Retroviruses **22:**1186-91.

24. **Grenfell, B. T., O. G. Pybus, J. R. Gog, J. L. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes.** 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science **303:**327-32.

25. **Guindon, S., and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:**696-704.

26. **Hemelaar, J., E. Gouws, P. D. Ghys, and S. Osmanov.** 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS **20:**W13-23.

27. **Heyndrickx, L., W. Janssens, L. Zekeng, R. Musonda, S. Anagonou, G. Van der Auwera, S. Coppens, K. Vereecken, K. De Witte, R. Van Rampelbergh, M. Kahindo, L. Morison, F. E. McCutchan, J. K. Carr, J. Albert, M. Essex, J. Goudsmit, B. Asjo, M. Salminen, A. Buve, and G. van Der Groen.** 2000.

Simplified strategy for detection of recombinant human immunodeficiency virus type 1 group M isolates by gag/env heteroduplex mobility assay. Study Group on Heterogeneity of HIV Epidemics in African Cities. J Virol **74:**363-70.

28. **Hillis, D. M., and Bull, J. J.** 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic trees. Syst. Biol. **42:**182-192.
29. **Hoelscher, M., B. Kim, L. Maboko, F. Mhalu, F. von Sonnenburg, D. L. Birx, and F. E. McCutchan.** 2001. High proportion of unrelated HIV-1 intersubtype recombinants in the Mbeya region of southwest Tanzania. AIDS **15:**1461-70.
30. **Howard, T. M., D. O. Olaylele, and S. Rasheed.** 1994. Sequence analysis of the glycoprotein 120 coding region of a new HIV type 1 subtype A strain (HIV-1IbNg) from Nigeria. AIDS Res Hum Retroviruses **10:**1755-7.
31. **Hu, W. S., and H. M. Temin.** 1990. Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. Proc Natl Acad Sci U S A **87:**1556-60.
32. **Hu, W. S., and H. M. Temin.** 1990. Retroviral recombination and reverse transcription. Science **250:**1227-33.
33. **Jung, A., R. Maier, J. P. Vartanian, G. Bocharov, V. Jung, U. Fischer, E. Meese, S. Wain-Hobson, and A. Meyerhans.** 2002. Multiply infected spleen cells in HIV patients. Nature **418:**144.
34. **Kalish, M. L., K. E. Robbins, D. Pieniazek, A. Schaefer, N. Nzilambi, T. C. Quinn, M. E. St Louis, A. S. Youngpairoj, J. Phillips, H. W. Jaffe, and T. M. Folks.** 2004. Recombinant viruses and early global HIV-1 epidemic. Emerg Infect Dis **10:**1227-34.
35. **Kiwanuka, N., O. Laeyendecker, M. Robb, G. Kigozi, M. Arroyo, F. McCutchan, L. A. Eller, M. Eller, F. Makumbi, D. Birx, F. Wabwire-Mangen, D. Serwadda, N. K. Sewankambo, T. C. Quinn, M. Wawer, and R. Gray.** 2008. Effect of human immunodeficiency virus Type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection. J Infect Dis **197:**707-13.
36. **Konings, F. A., S. T. Burda, M. M. Urbanski, P. Zhong, A. Nadas, and P. N. Nyambi.** 2006. Human immunodeficiency virus type 1 (HIV-1) circulating recombinant form 02_AG (CRF02_AG) has a higher in vitro replicative capacity than its parental subtypes A and G. J Med Virol **78:**523-34.
37. **Laeyendecker, O., G. W. Zhang, T. C. Quinn, R. Garten, S. C. Ray, S. Lai, W. Liu, J. Chen, and X. F. Yu.** 2005. Molecular epidemiology of HIV-1 subtypes in southern China. J Acquir Immune Defic Syndr **38:**356-62.
38. **Lim, W. L., H. Xing, K. H. Wong, M. C. Wong, Y. M. Shao, M. H. Ng, and S. S. Lee.** 2004. The lack of epidemiological link between the HIV type 1 infections in Hong Kong and Mainland China. AIDS Res Hum Retroviruses **20:**259-62.
39. **Luo, C. C., C. Tian, D. J. Hu, M. Kai, T. Dondero, and X. Zheng.** 1995. HIV-1 subtype C in China. Lancet **345:**1051-2.
40. **Maljkovic Berry, I., R. Ribeiro, M. Kothari, G. Athreya, M. Daniels, H. Y. Lee, W. Bruno, and T. Leitner.** 2007. Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. J Virol **81:**10625-35.

41. **Martins Lde, O., E. Leal, and H. Kishino.** 2008. Phylogenetic detection of recombination with a Bayesian prior on the distance between trees. PLoS ONE **3**:e2651.
42. **McCutchan, F. E.** 2006. Global epidemiology of HIV. J Med Virol **78 Suppl 1**:S7-S12.
43. **McCutchan, F. E.** 2000. Understanding the genetic diversity of HIV-1. Aids **14 Suppl 3**:S31-44.
44. **Moutouh, L., J. Corbeil, and D. D. Richman.** 1996. Recombination leads to the rapid emergence of HIV-1 dually resistant mutants under selective drug pressure. Proc Natl Acad Sci U S A **93**:6106-11.
45. **Murphy, E., B. Korber, M. C. Georges-Courbot, B. You, A. Pinter, D. Cook, M. P. Kieny, A. Georges, C. Mathiot, F. Barre-Sinoussi, and et al.** 1993. Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the central African Republic. AIDS Res Hum Retroviruses **9**:997-1006.
46. **Negroni, M., and H. Buc.** 2001. Mechanisms of retroviral recombination. Annu Rev Genet **35**:275-302.
47. **Njai, H. F., Y. Gali, G. Vanham, C. Clybergh, W. Jennes, N. Vidal, C. Butel, E. Mpoudi-Ngolle, M. Peeters, and K. K. Arien.** 2006. The predominance of Human Immunodeficiency Virus type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness. Retrovirology **3**:40.
48. **Nora, T., C. Charpentier, O. Tenaillon, C. Hoede, F. Clavel, and A. J. Hance.** 2007. Contribution of recombination to the evolution of human immunodeficiency viruses expressing resistance to antiretroviral treatment. J Virol **81**:7620-8.
49. **Osmanov, S., C. Pattou, N. Walker, B. Schwardlander, and J. Esparza.** 2002. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J Acquir Immune Defic Syndr **29**:184-90.
50. **Peeters, M., and P. M. Sharp.** 2000. Genetic diversity of HIV-1: the moving target. Aids **14 Suppl 3**:S129-40.
51. **Peeters, M., C. Toure-Kane, and J. N. Nkengasong.** 2003. Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials. Aids **17**:2547-60.
52. **Piyasirisilp, S., F. E. McCutchan, J. K. Carr, E. Sanders-Buell, W. Liu, J. Chen, R. Wagner, H. Wolf, Y. Shao, S. Lai, C. Beyrer, and X. F. Yu.** 2000. A recent outbreak of human immunodeficiency virus type 1 infection in southern China was initiated by two highly homogeneous, geographically separated strains, circulating recombinant form AE and a novel BC recombinant. J Virol **74**:11286-95.
53. **Powell, R. L., J. Zhao, F. A. Konings, S. Tang, A. Nanfack, S. Burda, M. M. Urbanski, D. R. Saa, I. Hewlett, and P. N. Nyambi.** 2007. Identification of a novel circulating recombinant form (CRF) 36_cpx in Cameroon that combines two CRFs (01_AE and 02_AG) with ancestral lineages of subtypes A and G. AIDS Res Hum Retroviruses **23**:1008-19.
54. **Qiu, Z., H. Xing, M. Wei, Y. Duan, Q. Zhao, J. Xu, and Y. Shao.** 2005. Characterization of five nearly full-length genomes of early HIV type 1 strains in

Ruili city: implications for the genesis of CRF07_BC and CRF08_BC circulating in China. AIDS Res Hum Retroviruses **21:**1051-6.

55. **Quinones-Mateu ME, A. E.** 2001. HIV-1 Fitness: Implications for Drug Resistance, Disease Progression, and Global Epidemic Evolution, HIV Sequence Compendium Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos.

56. **Quinones-Mateu ME, A. E.** 1999. Recombination in HIV-1: Update and implications. AIDS Reviews **1:**89-100.

57. **Rainwater, S., S. DeVange, M. Sagar, J. Ndinya-Achola, K. Mandaliya, J. K. Kreiss, and J. Overbaugh.** 2005. No evidence for rapid subtype C spread within an epidemic in which multiple subtypes and intersubtype recombinants circulate. AIDS Res Hum Retroviruses **21:**1060-5.

58. **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:**55-6.

59. **Schultz, A. K., M. Zhang, T. Leitner, C. Kuiken, B. Korber, B. Morgenstern, and M. Stanke.** 2006. A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. BMC Bioinformatics **7:**265.

60. **Shao Y, Z. Q., Wang B, et al.** 1994. Sequence analysis of HIV env ene among HIV infected IDUs in Yunnan epidemic area of China. Chin J Virol **10:**291–299.

61. **Shao Y, Z. Y., Chen Z, et al.** 1991. Isolation of viruses from HIV infected individuals in Yunnan. Chin J Epidemiol **12:**129.

62. **Sierra, M., M. M. Thomson, M. Rios, G. Casado, R. O. Castro, E. Delgado, G. Echevarria, M. Munoz, J. Colomina, R. Carmona, Y. Vega, E. V. Parga, L. Medrano, L. Perez-Alvarez, G. Contreras, and R. Najera.** 2005. The analysis of near full-length genome sequences of human immunodeficiency virus type 1 BF intersubtype recombinant viruses from Chile, Venezuela and Spain reveals their relationship to diverse lineages of recombinant viruses related to CRF12_BF. Infect Genet Evol **5:**209-17.

63. **Skar, H., S. Sylvan, H. B. Hansson, O. Gustavsson, H. Boman, J. Albert, and T. Leitner.** 2008. Multiple HIV-1 introductions into the Swedish intravenous drug user population. Infect Genet Evol **8:**545-52.

64. **Stahl, F. W.** 1987. Genetic recombination. Sci Am **256:**90-101.

65. **Su, L., M. Graf, Y. Zhang, H. von Briesen, H. Xing, J. Kostler, H. Melzl, H. Wolf, Y. Shao, and R. Wagner.** 2000. Characterization of a virtually full-length human immunodeficiency virus type 1 genome of a prevalent intersubtype (C/B') recombinant strain in China. J Virol **74:**11367-76.

66. **Swofford, D.** 1991. PAUP: phylogenetic analysis using parsimony, version 3.1.

67. **Takebe, Y., K. Motomura, M. Tatsumi, H. H. Lwin, M. Zaw, and S. Kusagawa.** 2003. High prevalence of diverse forms of HIV-1 intersubtype recombinants in Central Myanmar: geographical hot spot of extensive recombination. AIDS **17:**2077-87.

68. **Tee, K. K., O. G. Pybus, X. J. Li, X. Han, H. Shang, A. Kamarulzaman, and Y. Takebe.** 2008. Temporal and spatial dynamics of human immunodeficiency

virus type 1 circulating recombinant forms 08_BC and 07_BC in Asia. J Virol **82:**9206-15.

69. **Temin, H. M.** 1993. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. Proc Natl Acad Sci U S A **90:**6900-3.

70. **Thomson, M. M., E. Delgado, I. Herrero, M. L. Villahermosa, E. Vazquez-de Parga, M. T. Cuevas, R. Carmona, L. Medrano, L. Perez-Alvarez, L. Cuevas, and R. Najera.** 2002. Diversity of mosaic structures and common ancestry of human immunodeficiency virus type 1 BF intersubtype recombinant viruses from Argentina revealed by analysis of near full-length genome sequences. J Gen Virol **83:**107-19.

71. **Tovanabutra, S., C. Beyrer, S. Sakkhachornphop, M. H. Razak, G. L. Ramos, T. Vongchak, K. Rungruengthanakit, P. Saokhieo, K. Tejafong, B. Kim, M. De Souza, M. L. Robb, D. L. Birx, J. Jittiwutikarn, V. Suriyanon, D. D. Celentano, and F. E. McCutchan.** 2004. The changing molecular epidemiology of HIV type 1 among northern Thai drug users, 1999 to 2002. AIDS Res Hum Retroviruses **20:**465-75.

72. **Vijay, N. N., Vasantika, R. Ajmani, A. S. Perelson, and N. M. Dixit.** 2008. Recombination increases human immunodeficiency virus fitness, but not necessarily diversity. J Gen Virol **89:**1467-77.

73. **Vinoles, J., M. Serra, J. C. Russi, D. Ruchansky, S. Sosa-Estani, S. M. Montano, G. Carrion, L. M. Eyzaguirre, J. K. Carr, J. G. Olson, C. T. Bautista, J. L. Sanchez, and M. Weissenbacher.** 2005. Seroincidence and phylogeny of human immunodeficiency virus infections in a cohort of commercial sex workers in Montevideo, Uruguay. Am J Trop Med Hyg **72:**495-500.

74. **Yang, C., M. Li, Y. P. Shi, J. Winter, A. M. van Eijk, J. Ayisi, D. J. Hu, R. Steketee, B. L. Nahlen, and R. B. Lal.** 2004. Genetic diversity and high proportion of intersubtype recombinants among HIV type 1-infected pregnant women in Kisumu, western Kenya. AIDS Res Hum Retroviruses **20:**565-74.

75. **Yang, R., S. Kusagawa, C. Zhang, X. Xia, K. Ben, and Y. Takebe.** 2003. Identification and characterization of a new class of human immunodeficiency virus type 1 recombinants comprised of two circulating recombinant forms, CRF07_BC and CRF08_BC, in China. J Virol **77:**685-95.

76. **Yang, R., X. Xia, S. Kusagawa, C. Zhang, K. Ben, and Y. Takebe.** 2002. On-going generation of multiple forms of HIV-1 intersubtype recombinants in the Yunnan Province of China. AIDS **16:**1401-7.

77. **Yu, X. F., W. Liu, J. Chen, W. Kong, B. Liu, J. Yang, F. Liang, F. McCutchan, S. Piyasirisilp, and S. Lai.** 2001. Rapid dissemination of a novel B/C recombinant HIV-1 among injection drug users in southern China. AIDS **15:**523-5.

78. **Zhang, M., A. K. Schultz, C. Calef, C. Kuiken, T. Leitner, B. Korber, B. Morgenstern, and M. Stanke.** 2006. jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. Nucleic Acids Res **34:**W463-5.

79. **Zhuang, J., A. E. Jetzt, G. Sun, H. Yu, G. Klarmann, Y. Ron, B. D. Preston, and J. P. Dougherty.** 2002. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J Virol **76:**11273-82.

TABLE 1. Re-subtyping results of subtype A, B, C, F, G, and AG, BC and BF recombinants.

	AG set				BC set										BF set												
	Full length (world) N=140				Full length (world) N=509					Fragments (Asia) N=4413					Full length (world) N=220							Fragments (S. America) N=4153					
Num of sequences																											
Database subtype	**A**	**G**	**02**	**AG**	**B**	**C**	**07**	**08**	**BC**	**B**	**C**	**07**	**08**	**BC**	**B**	**F**	**12**	**17**	**28**	**29**	**BF**	**B**	**F**	**12**	**17**	**28**	**29**	**BF**
Num of sequences	72	12	48	8	152	334	7	4	12	3133	1048	17	171	44	152	12	11	2	3	4	36	3070	242	261	0	0	0	580
Num of problematic sequences [1]	1	0	2	0	15	12	0	0	3	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0
Num of disagreed sequences [2]	0	0	1	0	2	0	0	0	2	24	6	6	102	27	2	2	6	2	1	1	1	74	19	31	0	0	0	107

Footnotes

1. Problematic sequence that could not be unequivocally assigned. They meet one of the following criteria: 1) Contain unusually high contents of IUPAC code N (meaning any nucleotide), either have > 100 continuous Ns, or > 7% N for sequences of length < 1000 nt, or > 5% N for sequences of length 1000-2999, or > 3% N for sequences of length 3000 or above; 2) Contain an artifactual deletion of > 100 nt.
2. Classification of the sequences was compared between the database assignments (of which the majority were extracted from the literature) and the jpHMM predictions. The jpHMM classification in > 95% of these sequences were confirmed by the NCBI HIV genotype tool (http://www.ncbi.nlm.nih.gov/projects/genotyping).

FIGURE LEGENDS

FIG 1. Genome maps of all near full-length sequences composed exclusively of subtype A and G, or B and C, or B and F.

(A) AG recombinants were classified into 4 groups (with > 1 sequence) and 22 URFs. (B) BC recombinants were classified into 2 groups and 7 URFs. (C) BF recombinants were classified into 9 groups and 29 URFs. The genomic compositions and breakpoint positions were defined by the jpHMM program as described in the Method and Material session.

"Country" refers to the sampling country. Two-letter country code is used here. [AR: Argentina. BE: Belgium, BO: Bolivia, BR: Brazil, CD: Dem Rep of the Congo, CL: Chile, CM: Cameroon, CN: China, EC: Ecuador, ES: Spain, FR: France, GH: Ghana, KE: Kenya, MM: Myanmar. NG: Nigeria, SE: Sweden, SN: Senegal, US: United States, UY: Uruguay, UZ: Uzbekistan, VE: Venezuela.]

"Sequence source" refers to the HIV database/literature-assigned subtypes. The digits in brackets are the sequence numbers.

"Risk factor" in (C): MtoM, men to men; MtoB, mother to baby; heterosexual, heterosexual contact; biosexual, biosexual contact; IDU, injecting drug use. The transmission information was retrieved from the Los Alamos HIV sequence database.

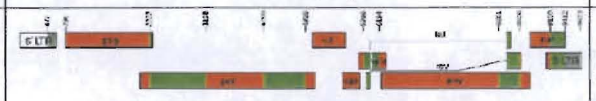
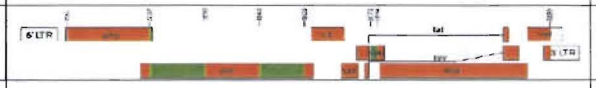
FIG 2. CRF02 sequence analysis results. (A) The ML trees of consensus sub-regions delimited by the breakpoints in the majority of CRF02 and AG recombinant sequences. Bootstrap supports for clustering are also shown. (B) The relationship between CRF02 and subtype A, G inferred from the ML results shown in (A). As: A's sibling. Gs: G's sibling. Gp: G's parent. Ad: A's descendent. Gd: G's descendent. A/G: mixture between A and G,

733 but not able to cluster CRF02 with either A or G. (C) The consensus sub-regions were
734 mapped onto the genome of HXB2. (D) CRF02-IBNG genome composition – jpHMM
735 result. Genomic regions in red: subtype A. Genomic regions in green: subtype G. (E) The
736 relationship between CRF02 with contemporary subtype A, G, and subtype A, G ancestors.
737 A Jukes-Cantor distance plot from the RIP result is shown.
738
739 **FIG 3. ML trees of consensus sub-regions delimited by the breakpoints in CRF07 and**
740 **CRF08 CRFs.**
741
742 **FIG. 4. Breakpoint frequency in near full-length BC and BF recombinants.**
743 The breakpoint positions are based on HXB2-numbering. Highlighted grey regions. Left
744 and middle: breakpoints are less present in BC than in BF recombinants. Right: both BC
745 and BF recombinants have few breakpoints in portion of gp120. Red line: breakpoints
746 present in 3 sequences. Above red line: breakpoints shared in > 3 sequences. Below red
747 line: breakpoints shared in < 3 sequences.
748
749 **FIG. 5. Contemporary sequences co-exist with some old sequences in the current**
750 **HIV-1 epidemic.**
751 The dashed circle differentiates the old and contemporary sequences. Inside the circle, the
752 old sequences, like subtype E strains, may be no longer exist in the epidemic. We only can
753 deduce subtype E's old presence based on CRF01_AE, a recombinant between subtype A
754 and E. "X" represents an extinct strain, "Y" represents an old strain that is still circulating
755 in the current epidemic, but it hasn't been identified. CRF02 is an old recombinant derived

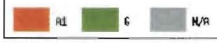
from old and contemporary subtype A and G. BF and BC recombinants are rather new. Their parental sequences are contemporary sequences.

Figure 1

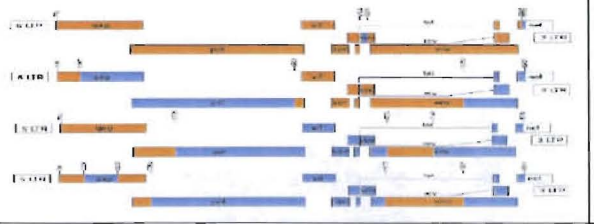
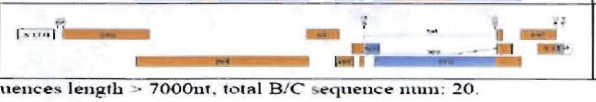
(A)

Group	Genome map	Country	Seq Source
1 (IBNG group)		CM(7) NG(2) EC(2) FR(2) US(1) UZ(1)	CRF02(14) AG(1)
2		CM(7) US(1) SN(1)	CRF02(9)
3		CM(2) UZ(2)	CRF02(4)
4		CM(2) SN(1)	CRF02(3)
22 URFs		CM(10) GH(3) SN(2) NG(2) US(1) SE(1) CD(1) BE(1) KE(1)	CRF02(15) AG(7)

- Sequences length > 7000nt, total sequence num: 53

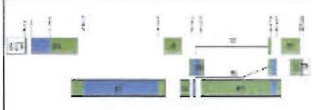
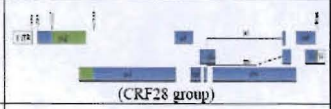
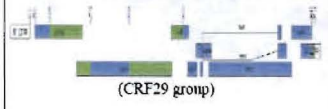
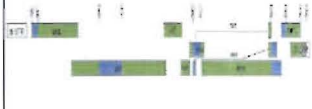
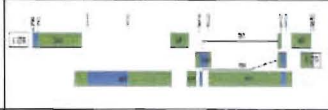
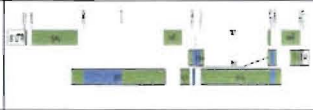
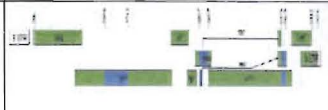


(B)

Group	Genome map	Country (city)	Sequence source
1 (CRF07 group)		CN (7)= Xinjiang (6) + Guangxi (1)	CRF07(7)
2 (CRF08 group)		CN(6)= Guangxi (3) + Yunnan (2) + Gansu (1)	CRF08(4) BC(2)
7 URFs		CN(4)= Yunnan (4) All from Ruili	BC(7)
		MM (2)	
		AR (1)	

- Sequences length > 7000nt, total B/C sequence num: 20.

763 (C)

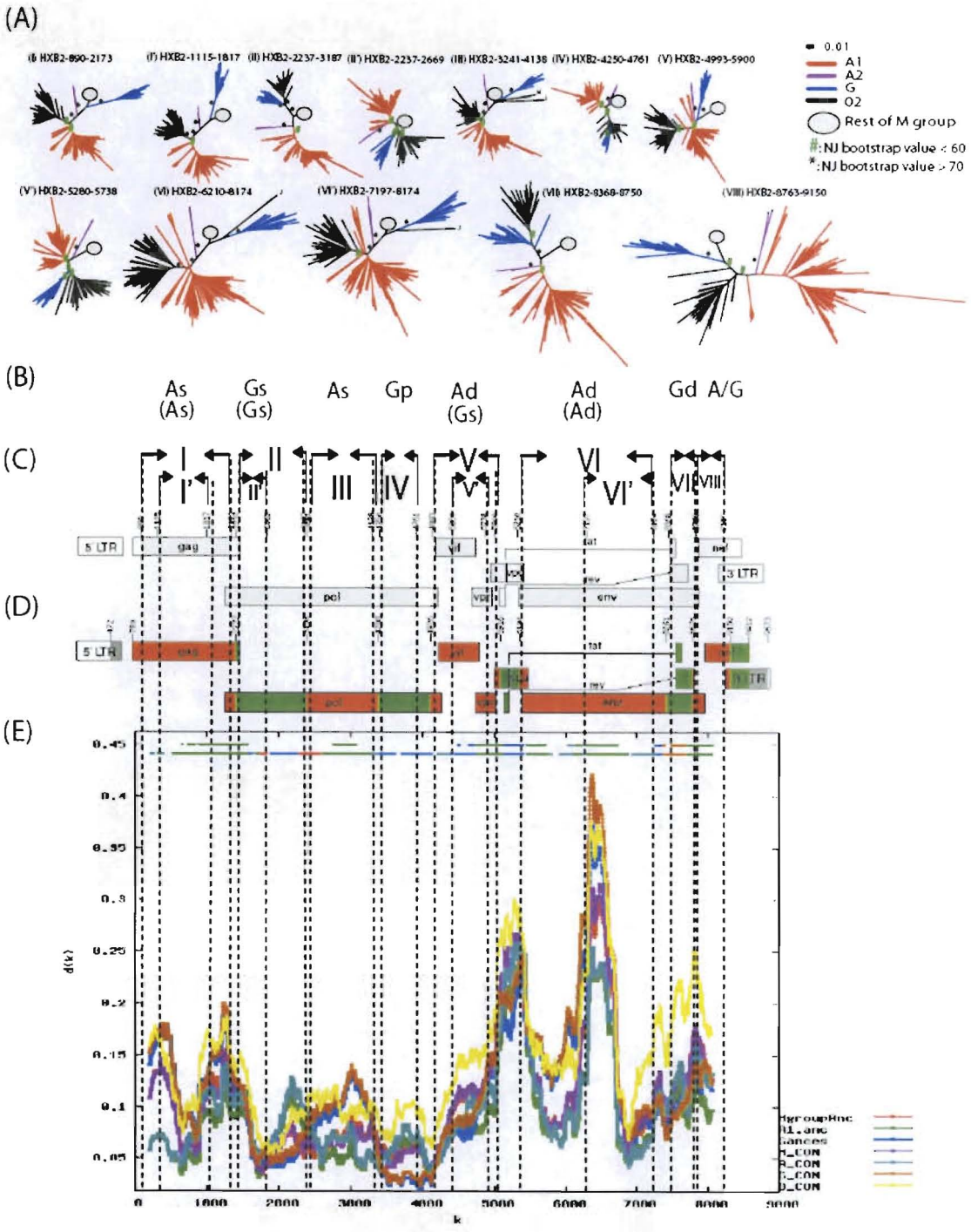
Group	Genome map	Country	Seq source	Risk factor	Group	Genome map	Country	Seq source	Risk factor
1	 (CRF12 group)	AR(4) UY(1)	CRF12 (5)	Hetero sexual (2) IDU(1) MtoM (1) N/A(1)	6		CL(2) AR(1)	BF(3)	MtoB (2) Heterosex ual (1)
2	 (CRF28 group)	BR(2) VE(1)	CRF28 (2) BF(1)	Hetero sexual (3)	7		AR(2)	BF(1) BF1(1)	Heterosex ual (1) N/A (1)
3	 (CRF29 group)	BR(3)	CRF29 (3)	Hetero sexual (1) MtoB (1) N/A (1)	8		AR(1) UY(1)	CRF12 (1) BF(1)	Heterosex ual (1) MtoM (1)
4		AR(1) UY(1) BO(1) ES(1)	CRF12 (2) BF(2)	Hetero sexual (2) MtoM (2)	9		AR(1) UY(1)	CRF12 (1) BF(1)	IDU (1) MtoM (1)
5		AR(3)	BF(2) BF1(1)	Hetero sexual (1) IDU (1) N/A (1)	29 URF		BR (18) AR(9) CL(1) ES(1)	BF(19) BF1(4) CRF12 (2) CRF17 (2) CRF28 (1) CRF29 (1)	Heterosexual (15) MtoB (3) IDU(2) Biosexual (1) N/A(7) Heterosexual and transfusion (1)

- Sequences length > 7000nt, total sequence num: 56.

764

765

765 **Figure 2**

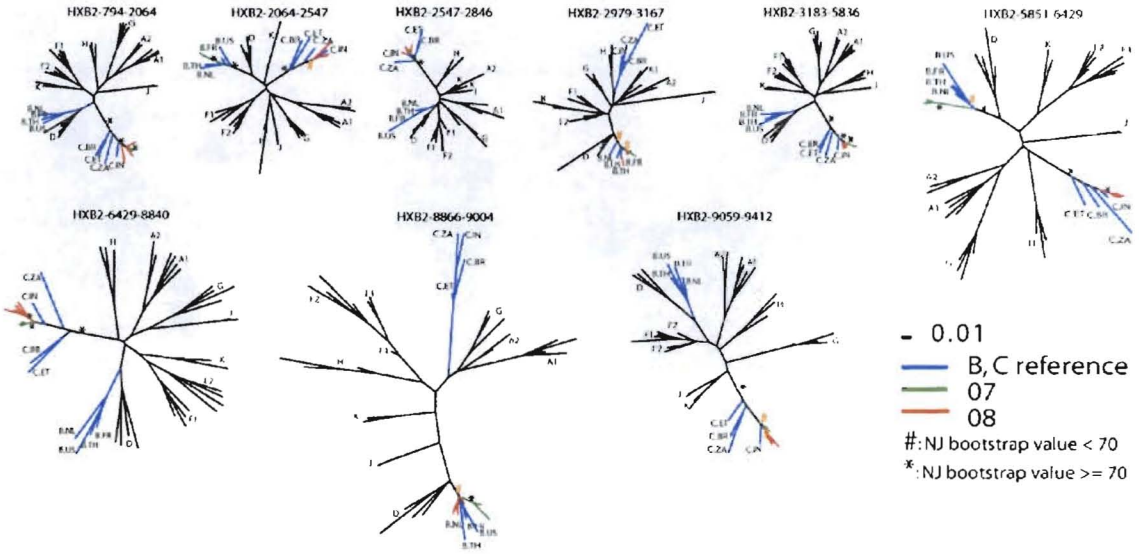


766

767

768

768 **Figure 3**

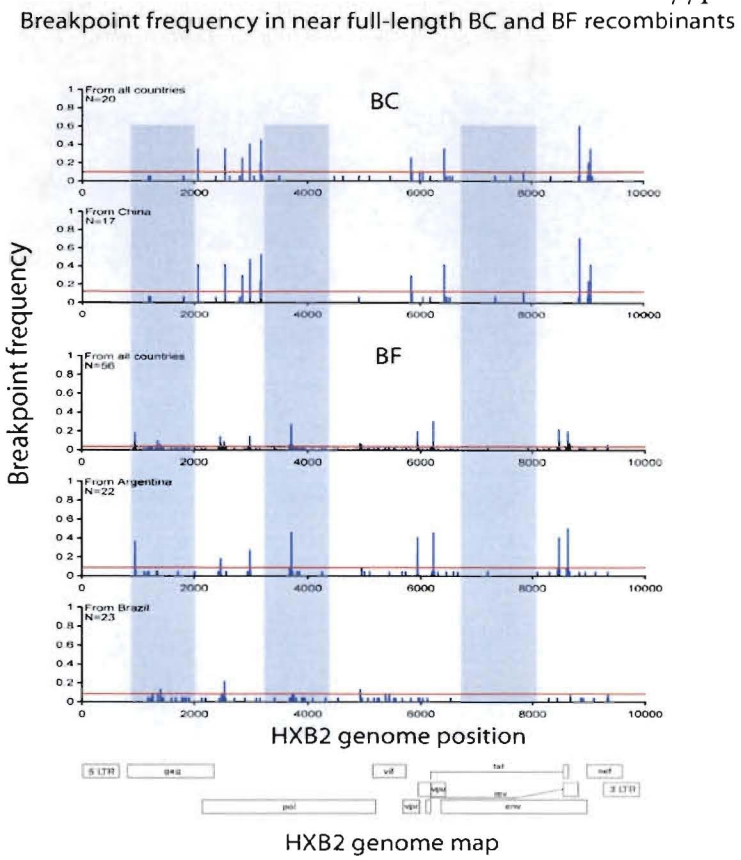


769

770

770 **Figure 4**

771



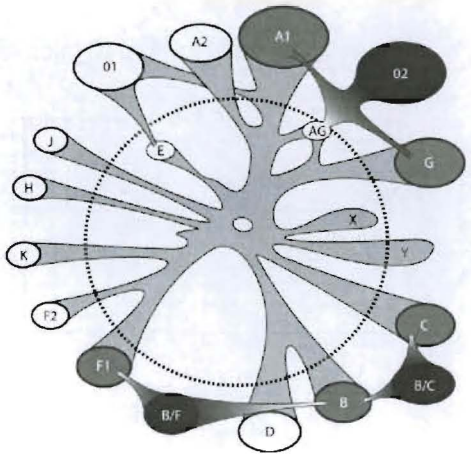
784

785

786

787

787 **Figure 5**



788

789

790

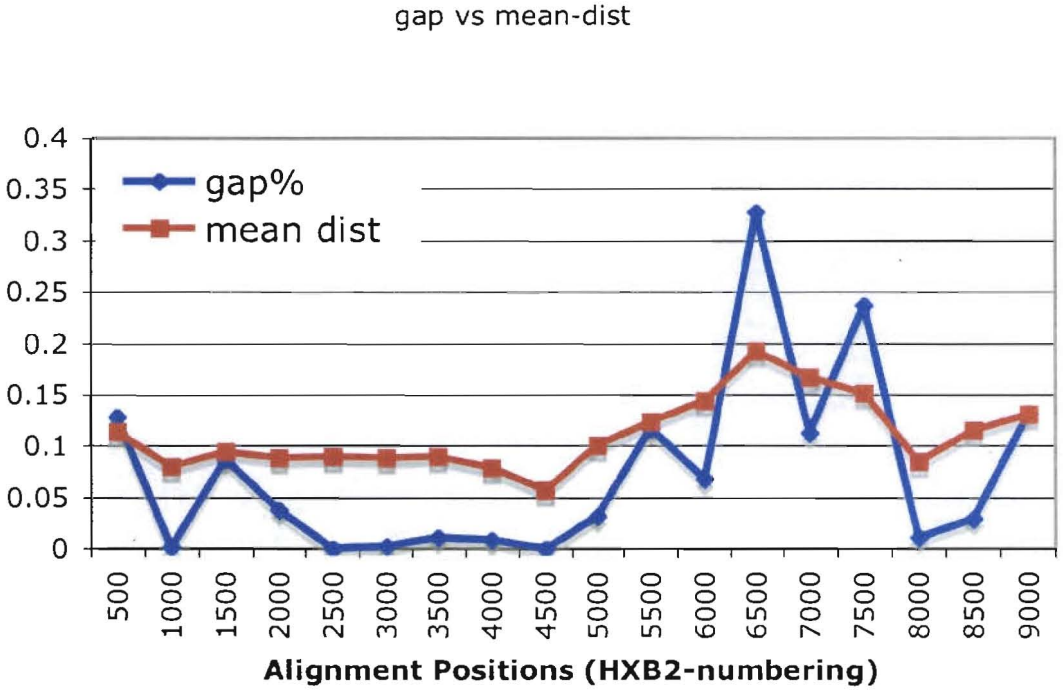
Supplementary figures:

(from Ming: not sure whether we need these figures. So please suggest.)

Suppl figure 1. Plot showing that our CRF02 conclusion, that is, CRF02 was derived from old and new recombination events, is not biased by CRF02 sequence alignment quality.

Gap%: percentage of alignment gaps within every 500 nt window.

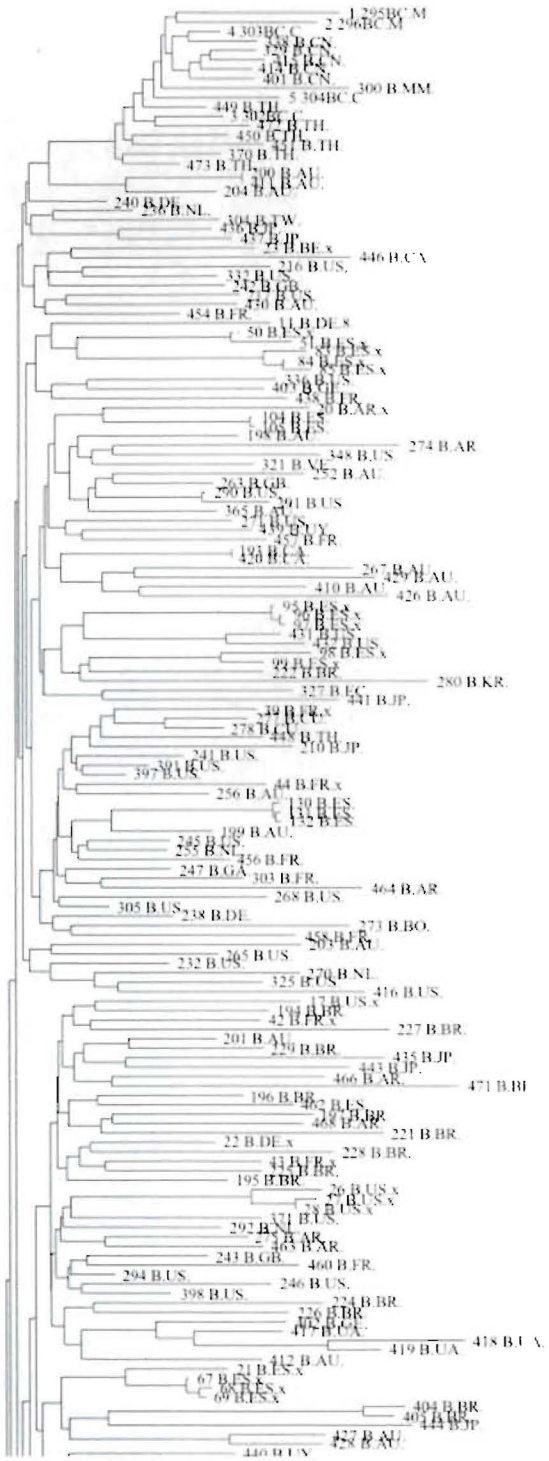
Mean dist: mean evolutionary distance (F84) within every 500 nt window.



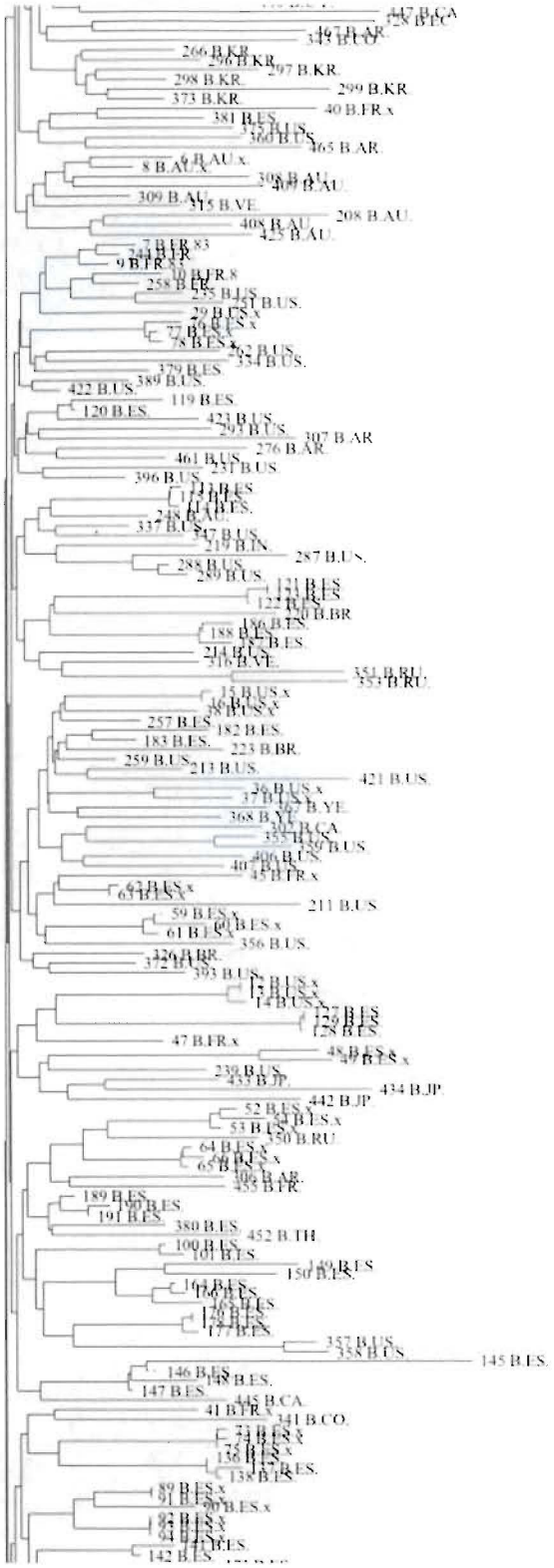
Suppl figure 2. Two neighbor-joining trees shown that China B and C are more restrictly derived from China neighboring countries, while China neighboring countries' B and C have more contacts with B and C from other regions of the world. Sequences used in the subtype B and C fragments analyses, as depicted here,

803 were obtained from worldwide sources. Each sequence is in the following format:
804 "digit" followed by "subtype", followed by "Country code". The digit is a
805 sequential number used in the sequence alignments.
806 (1) Subtype B fragmental sequences (HXB2 positions: 3497-4473)

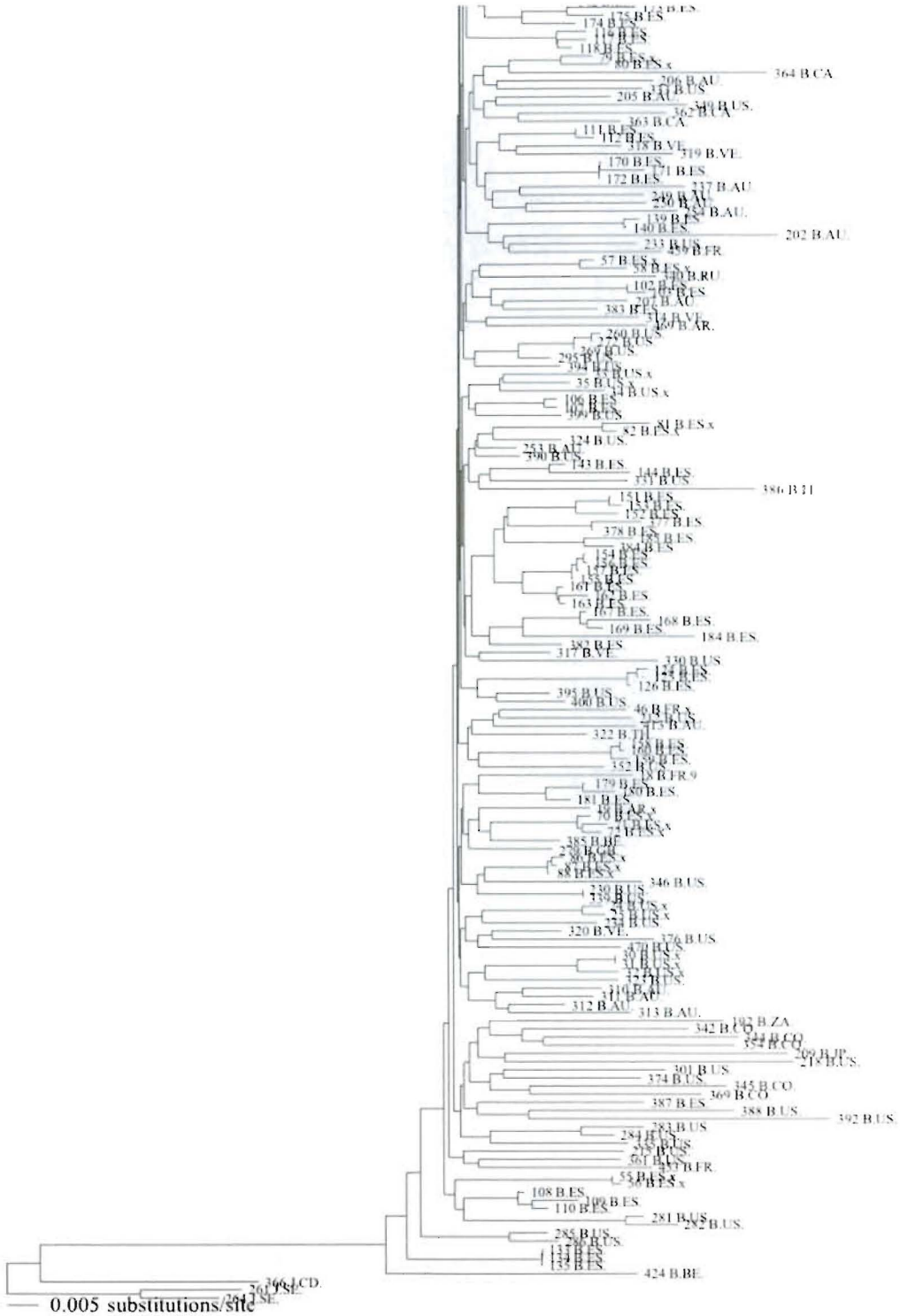
NJ


1 295BC.M
2 296BC.M
4 303BC.C
5 304BC.C
300 B.MM.
446 B.CA
216 B.US.
274 B.AR
348 B.US.
252 B.AU.
291 B.US
280 B.KR.
441 B.JP.
210 B.JP
464 B.AR.
273 B.BO.
416 B.US.
227 B.BR.
435 B.JP
471 B.BI
221 B.BR.
228 B.BR.
460 B.FR.
246 B.US.
418 B.UA
404 B.BR

807


447 B.CA
266 B.KR
297 B.KR
298 B.KR
299 B.KR
373 B.KR
40 B.FR.x
381 B.ES
465 B.AR
6 B.AU.x
8 B.AU.x
309 B.AU
315 B.VE
208 B.AU
408 B.AU
425 B.AU
7 B.FR.83
9 B.FR.83
10 B.FR.8
258 B.ES
255 B.US
451 B.US
29 B.ES.x
78 B.ES.x
302 B.US
334 B.US
379 B.ES
389 B.US
422 B.US
119 B.ES
120 B.ES
423 B.US
293 B.US
307 B.AR
276 B.AR
461 B.US
237 B.US
396 B.US
337 B.US
347 B.US
219 B.IN
287 B.US
285 B.US
269 B.US
121 B.ES
122 B.ES
450 B.BR
186 B.ES
188 B.ES
214 B.ES
316 B.VE
351 B.RU
353 B.RU
15 B.US.x
257 B.ES
182 B.ES
183 B.ES
223 B.BR
259 B.US
213 B.US
421 B.US
36 B.US.x
37 B.US.x
368 B.YE
305 B.CA
355 B.US
359 B.US
400 B.US
45 B.FR.x
62 B.ES.x
211 B.US
59 B.ES.x
61 B.ES.x
356 B.US
393 B.US
14 B.US.x
127 B.ES
129 B.ES
47 B.FR.x
48 B.ES.x
49 B.ES.x
239 B.US
435 B.JP
434 B.JP
442 B.JP
52 B.ES.x
53 B.ES.x
150 B.RU
64 B.ES.x
65 B.ES.x
455 B.FR
189 B.ES
190 B.ES
191 B.ES
380 B.ES
452 B.TH
100 B.ES
101 B.ES
149 B.ES
150 B.ES
357 B.US
358 B.US
145 B.ES
146 B.ES
148 B.ES
147 B.ES
445 B.CA
41 B.FR.x
341 B.CO
138 B.ES
89 B.ES.x
90 B.ES.x
142 B.ES

808


364 B.CA.
202 B.AU.
386 B.JI
346 B.US.
376 B.US.
192 B.ZA.
209 B.JP.
345 B.CO.
392 B.US.
424 B.BE.
0.005 substitutions/site

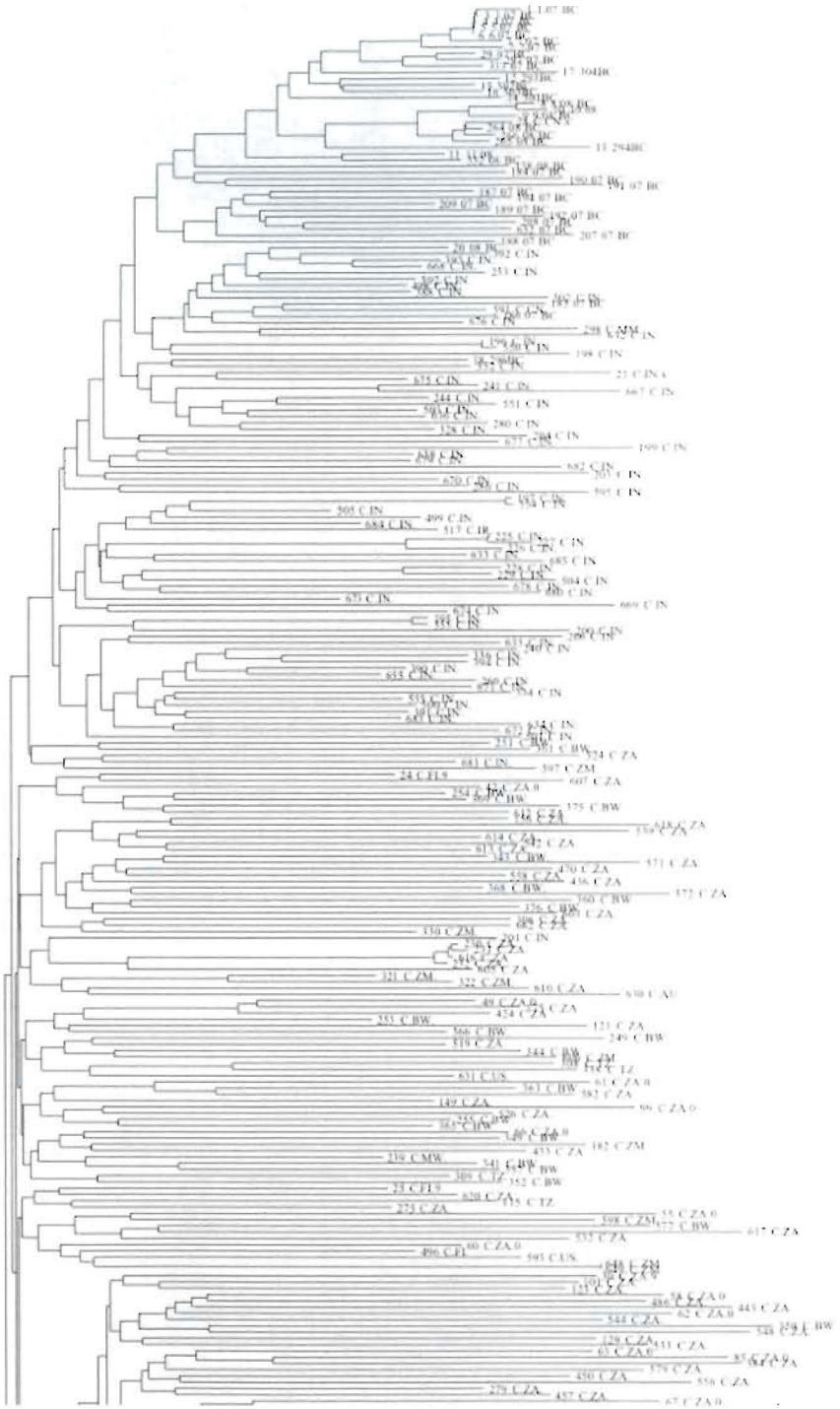
809

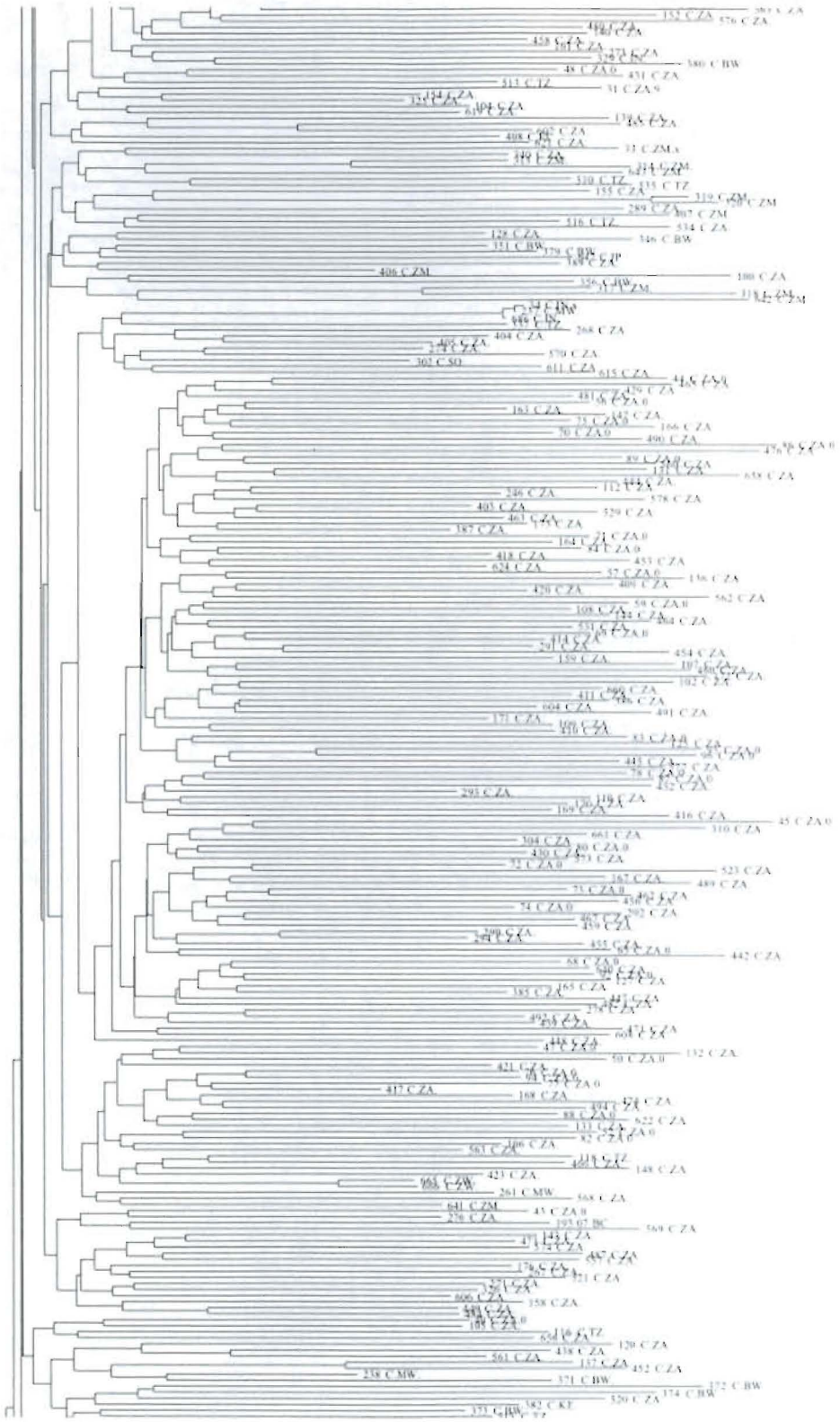
810

(2) Subtype C fragmental sequences: HXB2 positions: 6582-7349.

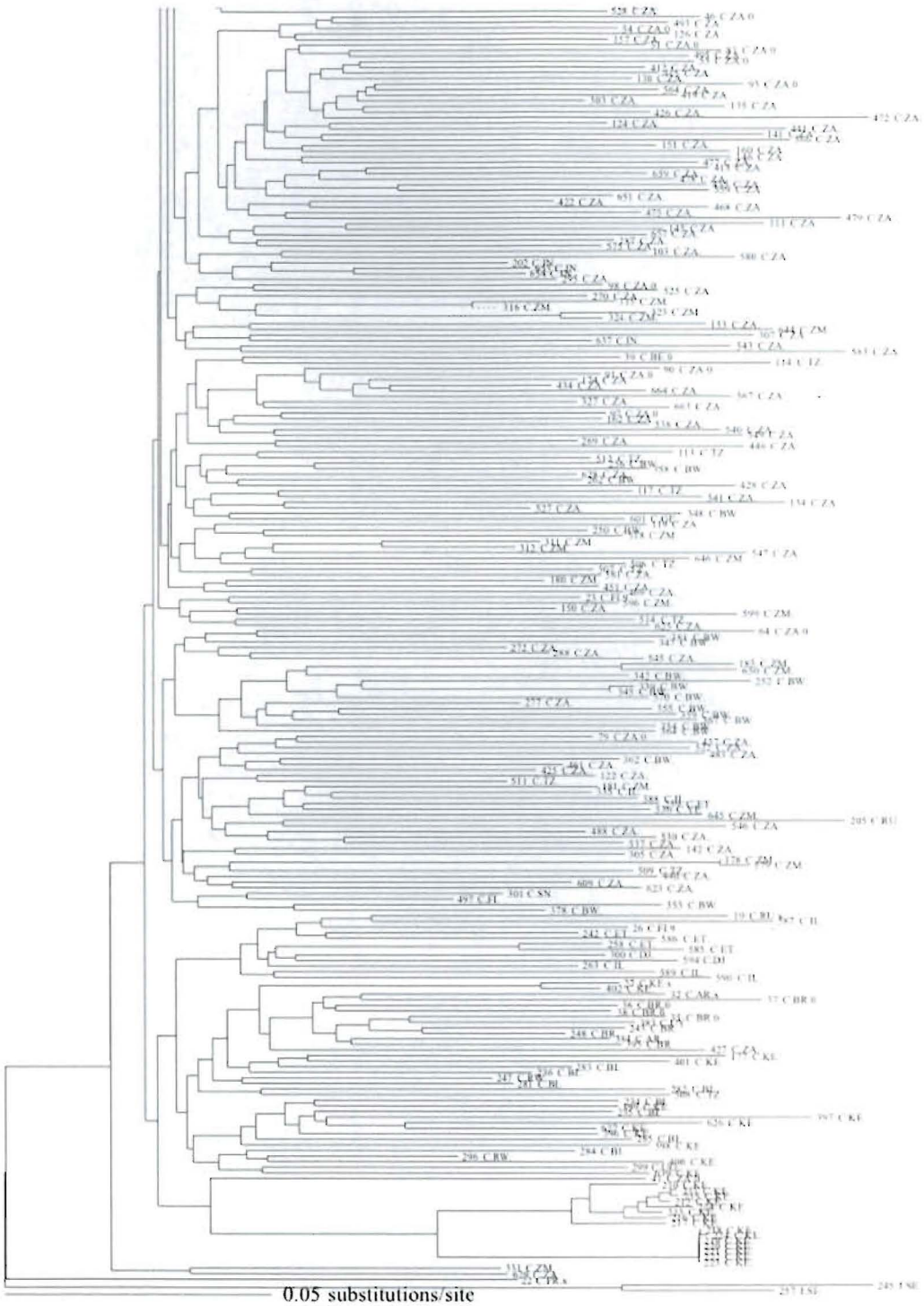
HXB2-6582-7349 (worldwide C)

NJ





813



814

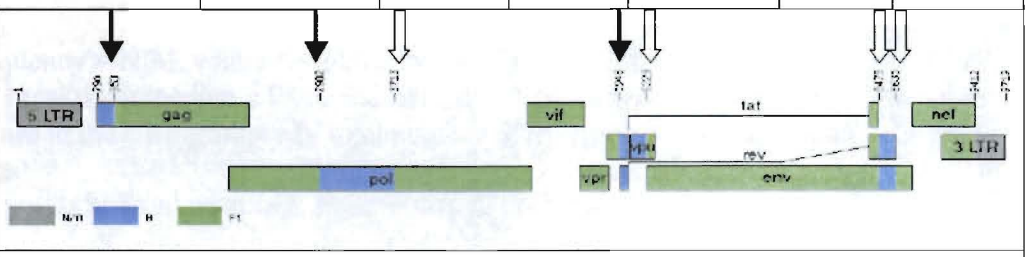
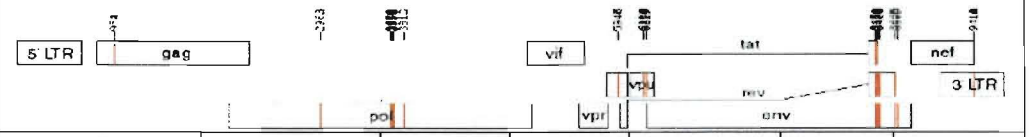
815

816 **Suppl. 3.** The breakpoint range and frequency in CRF12, CRF28, and CRF29 CRF08

817 groups.

818 The breakpoint frequency = N/M, where N=total number of BF recombinants that have
819 breakpoints within breakpoint median ± 98nt, and have the same subtypes flanking the
820 breakpoint as they are in the CRF group; M= total number of BF recombinants that span
821 this genomic region.
822 Black arrow: breakpoints at fixed positions; Hollow arrow: breakpoint region
823

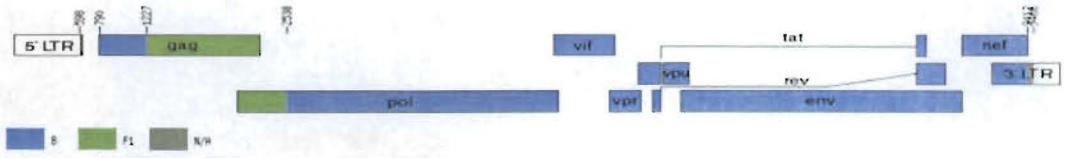
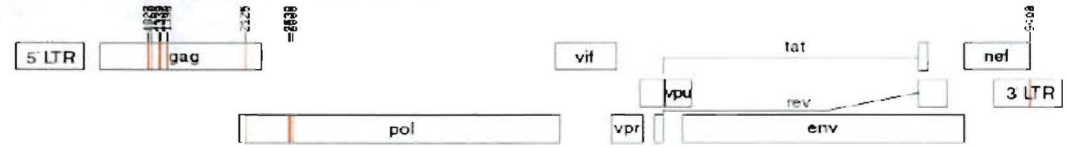
CRF12 group

Breakpoint number	1	2	3	4	5	6	7
Genome graph							
Breakpoint locations Of this CRF							
Breakpoint range	953	2982	3679-3812	5946	6193-6229	8450-8485	8635-8669
Breakpoint median	953	2982	3713	5946	6229	8475	8635
Interquartile range (mean)	const. (953)	const. (2982)	3692-3713 (3722)	const. (5946)	6200-6229 (6216)	8475-8484 (8474)	8635-8669 (8649)
# of complete BF recombinants	Sequence total: 56 (Argentina: 22, Brazil: 23)						
Frequency of full-length BF recombinants that bear bk within median ± 98nt (world)	20/56	13/56	30/56	27/56	24/56	26/56	25/56
Frequency of full-length BF recombinants that bear bk within median ± 98nt (Brazil)	0/23	1/23	8/23	1/23	0/23	0/23	2/23
Frequency of full-length BF recombinants that bear bk within median ± 98nt (Argentina)	15/22	10/22	15/22	18/22	15/22	17/22	15/22
# of South America BF fragments	Sequence total: 751 (Argentina: 639, Brazil: 109)						
Frequency of fragmental BF recombinants that bear bk within median ± 98nt (S. America)	0/2	333/685	11/11	0/0	25/28	1/5	1/5
Frequency of fragmental BF recombinants that bear bk within median ± 98nt (Brazil)	0/2	11/80	0/0	0/0	0/0	4/5	1/5
Frequency of fragmental BF recombinants that bear bk within median ± 98nt (Argentina)	0/0	322/603	11/11	0/0	25/28	0/0	0/0

824

825

CRF28 group:

Breakpoint number	1	2	
Genome graph			
Breakpoint locations Of this CRF			
Breakpoint range	1227-1398	2538-2565	
Breakpoint median	1329	2538	
Interquartile (mean)	1278-1364 (1318)	2538-2552 (2547)	
# of complete BF recombinants	Sequence total: 56 (Argentina: 22, Brazil: 23)		
Frequency of full-length BF recombinants that bear bk within median ± 98nt (world)	15/56	29/56	
Frequency of full-length BF recombinants that bear bk within median ± 98nt (Brazil)	10/23	10/23	
Frequency of full-length BF recombinants that bear bk within median ± 98nt (Argentina)	2/22	10/22	
# of South America BF fragments	Sequence total: 751 (Argentina: 639, Brazil: 109)		
Frequency of fragmental BF recombinants that bear bk within median ± 98nt (S. America)	6/9	313/621	
Frequency of fragmental BF recombinants that bear bk within median ± 98nt (Brazil)	3/4	38/61	
Frequency of fragmental BF recombinants that bear bk within median ± 98nt (Argentina)	3/5	273/558	

826

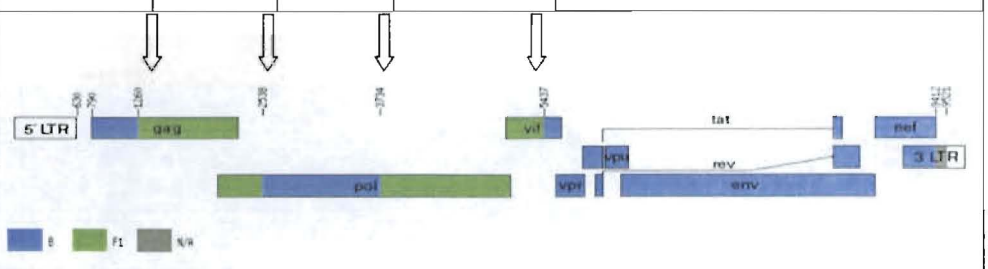
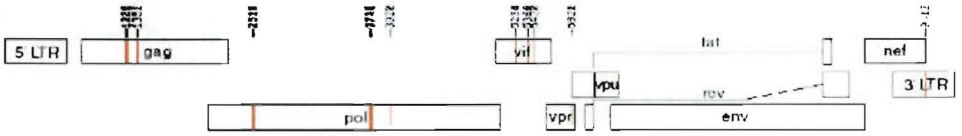
827

828

829

830

831 CRF29 group:

Breakpoint number	1	2	3	4	
Genome graph					
Breakpoint locations Of this CRF					
Breakpoint range	1236-1351	2518-2538	3734-3927	5233-5437	
Breakpoint median	1260	2538	3746	5368	
Interquartile (mean)	1248-1306 (1282)	2528-2538 (2531)	3740-3837 (3802)	5301-5403 (5346)	
# of complete BF recombinants	Sequence total: 56 (Argentina: 22, Brazil: 23)				
Frequency of full-length BF recombinants that bear bk within median ± 98nt (world)	16/56	29/56	30/56	7/56	
Frequency of full-length BF recombinants that bear bk within median ± 98nt (Brazil)	8/23	10/23	8/23	6/23	
Frequency of full-length BF recombinants that bear bk within median ± 98nt (Argentina)	4/22	10/22	15/22	1/22	
# of South America BF fragments	Sequence total: 751 (Argentina: 639, Brazil: 109)				
Frequency of fragmental BF recombinants that bear bk within median ± 98nt (S. America)	4/9	313/621	11/11	0/0	
Frequency of fragmental BF recombinants that bear bk within median ± 98nt (Brazil)	1/2	38/61	0/0	0/0	
Frequency of fragmental BF recombinants that bear bk within median ± 98nt (Argentina)	3/5	273/558	11/11	0/0	

832

833

834