TRACKING HIV-1 GENETIC VARIATION:

RECOMBINATION AND N-LINKED GLYCOSYLATION SITES

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultäten

der Georg-August-Universität zu Göttingen

vorgelegt von

Ming Zhang

aus Qingdao, P.R.China

Göttingen 2007

D 7

Referent: Prof. Dr. Burkhard Morgenstern

Korreferent: Dr. Thomas Leitner

Tag der mündlichen Prüfung:

Ph.D. thesis advisors:

From University of Goettingen, Germany
Prof. Dr. Burkhard Morgenstern

From Los Alamos National Laboratory, USA
Dr. Bette Korber
Dr. Thomas Leitner

Committee in charge:

From University of Goettingen, Germany
Prof. Dr. Burkhard Morgenstern
Prof. Dr. Edgar Wingender
Prof. Dr. Thomas Friedl
Prof. Dr. Stephan Waack

From Los Alamos National Laboratory, USA
Dr. Bette Korber
Dr. Thomas Leitner

Dedicated to my parents

ABSTRACT

The last 26 years saw a rampant global epidemic of HIV-1. HIV-1's extraordinary diversity is seeded by a high mutation rate, rapid replication, frequent recombination, and strategic placement and loss and gain of N-linked glycosylation sites. In this thesis, the genetic variation of HIV-1 was investigated with special focus on recombination and N-linked glycosylation sites.

Using phylogenetic analyses, distance methods, and HIV-1 subtying tools including one called jumping profile hidden Markov model, HIV-1 recombinants dominating HIV epidemic in three different geographical regions were examined. We found that CRF13_cpx includes sections of the rare subtype J, and that breakpoint inference can be greatly improved using all available sequences within a CRF family. We confirmed that CRF02_AG, a recombinant between subtype A and G that is prevalent in West and West Central Africa, is an old recombinant. The main recombination events that generated CRF02 took place before the 1970's, before HIV-1 had started to spread worldwide and the currently recognized subtypes had formed. Recombinants consisting of subtypes B and C are frequently found in the HIV-1 epidemic of Asia, especially in southwest China where they are associated with different drug trafficking routes. Our study suggested that CRF07 was derived from a recombination between CRF08 and subtype B. However, it is possible that the currently defined CRF07 is not the direct product of this recombination event. Lastly, we found that recent recombination between subtypes B and F in Argentina and Brazil, two epicenters in South America, has created many different, but related, recombinant forms. Taken together, it appears as if the HIV-1 epidemic is becoming more complex as it moves ahead into the future. Recombination among co-circulating forms creates new forms of HIV-1 that are now starting to dominate the epidemic in certain parts of the world.

We developed methods to track N-linked glycosylation sites (sequons) in HIV-1 as they shift positions and vary in local densities. Comparing primate lentiviruses, hepatitis C virus, and influenza A viruses showed that generating and tolerating shifting sequons is a unique evolutionary avenue for HIV-1 immune evasion. In addition, we found the primate lentiviral lineages have host species - dependent levels of sequon shifting, with HIV-1 in humans the most extreme. Further, unlike influenza A hemagglutinin H3 HA1 that accumulates sequons over time, HIV does not have a net increase in the number of sites over time at the population level, indicating that variation in number and placement, not accumulation of N-linked glycosylation sites, is more critical for HIV-1 immune evasion.

The studies detailed in this thesis, together with our great effort in re-subtyping > 150,000 sequences in the Los Alamos HIV sequence database, enables us to draw a more comprehensive and dynamic picture of the global HIV-1 epidemic.

LIST OF ORIGINAL PAPERS

The thesis is based on the following original papers. They are referred to in the text by the Roman numerals.

All published papers were reprinted with the permission with the publishers.

Main Content:

I. **Zhang M**, Wilbe K, Wolfe ND, Gaschen B, Carr JK, Leitner T (2005) HIV type 1 CRF13_cpx revisited: identification of a new sequence from Cameroon and signal for subsubtype J2. *AIDS research and human retroviruses* **21:** 955-960

II. **Zhang, M**., et al. (2007) Evidence for old and new recombinants in different epidemiological settings. *Manuscript*.

III. **Zhang M**, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C, Korber B (2004) Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology* **14:** 1229-1246

Appendix:

IV. **Zhang M**, Schultz AK, Calef C, Kuiken C, Leitner T, Korber B, Morgenstern B, Stanke M (2006) jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. *Nucleic acids research* **34:** W463-465

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

VI. Li M, Salazar-Gonzalez JF, Derdeyn CA, Morris L, Williamson C, Robinson JE, Decker JM, Li Y, Salazar MG, Polonis VR, Mlisana K, Karim SA, Hong K, Greene KM, Bilska M, Zhou J, Allen S, Chomba E, Mulenga J, Vwalika C, Gao F, **Zhang M**, Korber BT, Hunter E, Hahn BH, Montefiori DC (2006) Genetic and neutralization properties of subtype C human immunodeficiency virus type 1 molecular env clones from acute and early heterosexually acquired infections in Southern Africa. *J Virol* **80:** 11776-11790

CONTENTS

INTRODUCTION

AIDS (Acquired Immune Deficiency Syndrome) epidemic was discovered in 1981 (11). In the following couple of years, human immunodeficiency virus (HIV) that belongs to the lentiviral genus of the Retroviridae family was identified to be the etiologic agent of this deadly disease (6, 44, 60). Despite the late discovery of HIV, both retrospective studies (33, 36, 88) and epidemiological modeling (39) suggest that HIV has been present in humans for a long time, at least since the 1930s.

Important events of HIV epidemic during the years 1981-2006 are shown in Figure. 1.

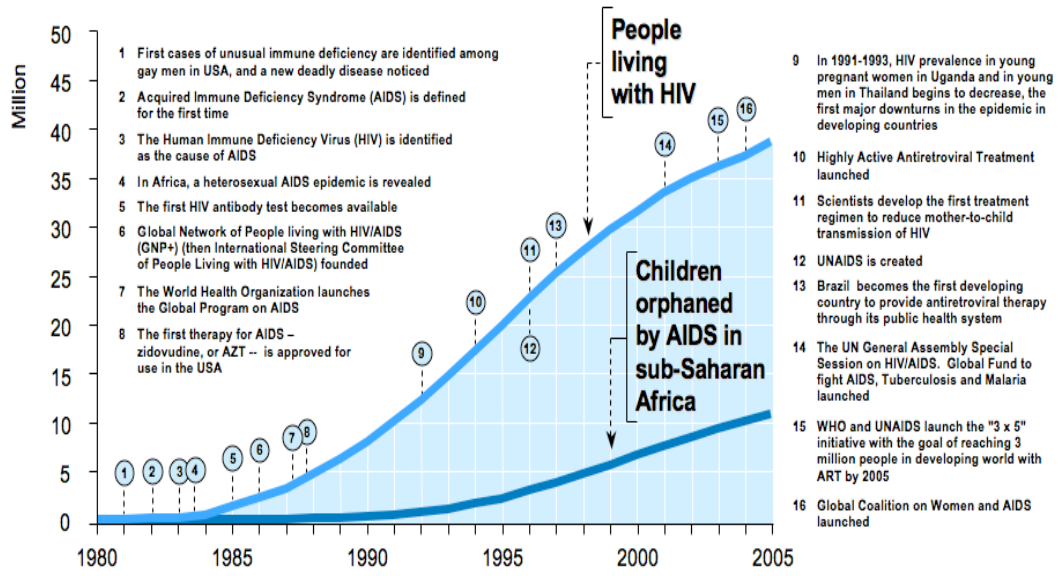


Figure 1. Important events during 25 years of AIDS. From (2).

1. Based on genetic similarities, HIV strains are classified into types, groups, subtypes, and sub-subtypes. This classification also reflects important zoonotic transmission events.

HIV type 1 (HIV-1) and HIV type 2 (HIV-2) are two types of HIV. HIV-2 appears to have resulted from at least seven zoonotic transmissions from sooty mangabey monkeys. It is primarily restricted to West Africa, and less virulent than HIV-1 (reviewed in ref (62)).

HIV-1 is classified into three groups, M (main group), O (outlier group), and N (not-M/not-O group). Group M and N are believed to have derived from at least two independent cross-species transmissions from chimpanzees (*Pan troglodytes troglodytes)* (21, 35), and group O is thought to have originated in a transmission from wild gorilla to humans (75). Group O and N are very rare and essentially limited to West Central Africa (30, 71). The M group dominates the global HIV epidemic. Currently ~ 40 million people are infected with HIV-1 and ~ 25 million have died of AIDS (2).

HIV-1 M group is divided into nine genetically distinct subtypes, A-D, F-H, J and K (reviewed in ref (63)). These subtypes are believed to have diverged in humans following one single zoonotic transmission: their last common ancestor was dated to be 1931, with 95% confidence interval 1915-1941 (39). Within subtype A and F, there are distinctive sub-lineages that are defined as sub-subtypes. Subtype A has two well-established sub-subtypes A1 and A2, though other A sub-subtypes have been suggested (52, 56). Subtype F is divided into F1 and F2. For historical reasons the B and D clades are called subtypes, but in fact the genetic distance between these two clades corresponds to a sub-subtype distance (17, 63).

Global HIV-1 subtypes prevalence and distributions are shown in Figure 2.

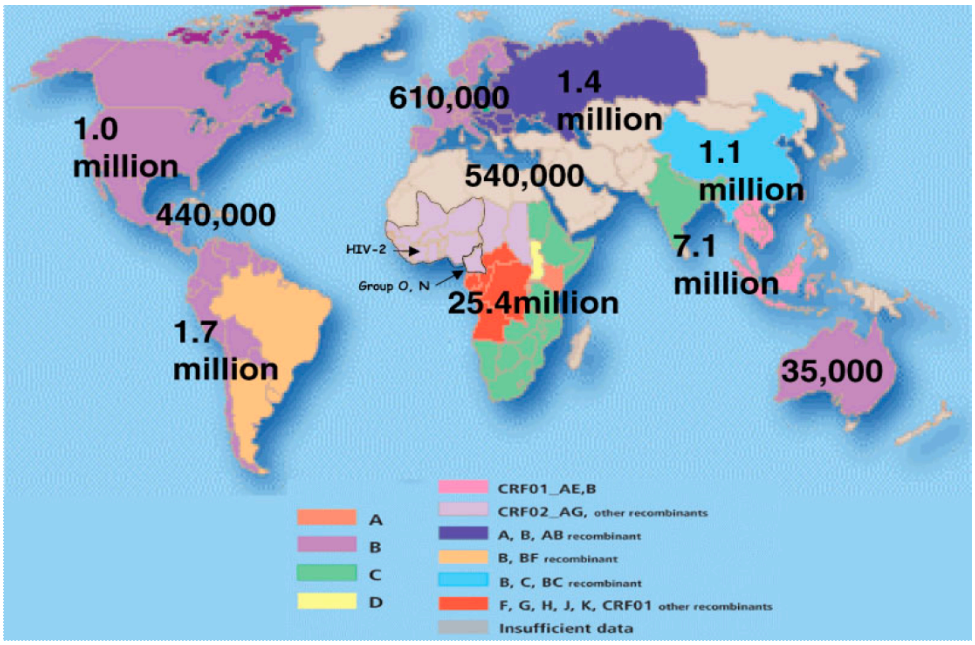


Figure 2. Global HIV-1 prevalence and distribution. From ref (51).

- The estimated numbers of HIV-infected individuals in North America, the Caribbean, South America, Western Europe, Central Asia, East Asia, Southeast Asia, North Africa and the Middle East, sub-Saharan Africa, and Australia are indicated.
- The colors depict regional patterns of HIV variation as follows: subtype A in East Africa; subtype B in the America, Europe, and Australia; subtype C in Southern and Eastern Africa, and in India; subtype D in East Africa; CRF01_AE and subtype B in Southeast Asia; CRF02_AG and other recombinants in West Africa; A, B, and AB recombinants in Central Asia; subtype B and BF recombinants in South America; subtype B and C, and BC recombinants in East Asia; rare subtypes, CR01_AE, and other recombinants in Central Africa and areas where there is insufficient data. The principal concentrations of HIV-1 groups O and N in Cameroon, and of HIV-2 in West Africa, are indicated by arrows (51).

2. HIV-1 is one of the most variable human pathogens. Its great variability is driven by at least three main mechanisms.

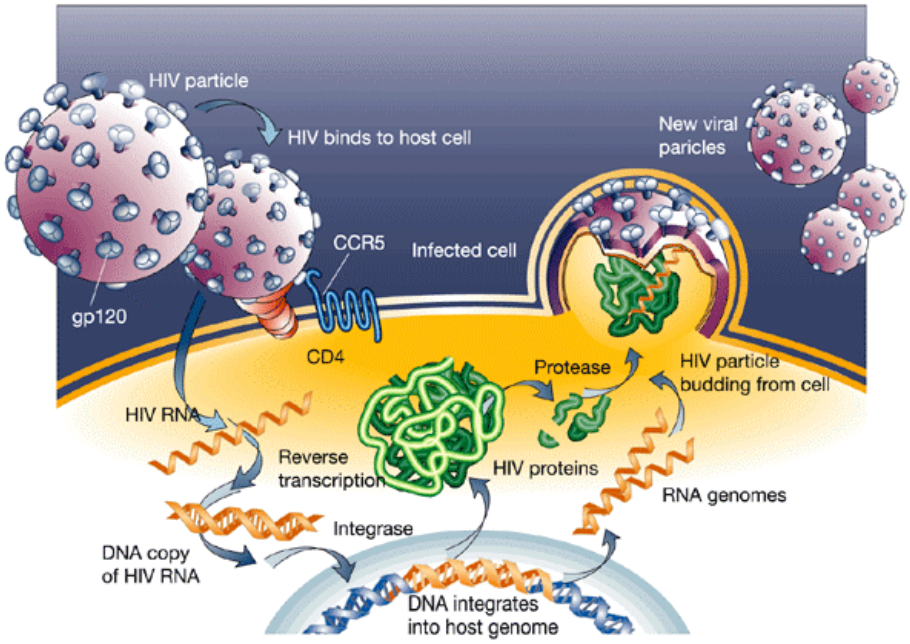
Like all RNA viruses, HIV-1 is characterized by very high mutation rates. The poor fidelity of reverse transcriptase and lack of proof-reading lead to a high error frequency, estimated to be $3.4x10^{-5}$ during reverse transcription (48), thus introducing almost one substitution per genome per replication cycle (22). The mutations are predominated by base substitutions, among those

G-to-A mutations are most frequently observed (48).

Compared to point mutations, recombination provides a more efficient means to explore sequence space and it is an important evolutionary mechanism in the global epidemic (Fig. 2). As with other retroviruses, HIV-1 recombines during reverse transcription (see Fig.3 “HIV-1 life cycle”). When a host cell is infected with two different HIV-1 strains, one RNA strand from each virus can be co-packed into a heterozygous virion. If this virion subsequently infects a new cell, the reverse transcriptase switches between the two different RNAs resulting in a mosaic viral genome. It has been estimated that HIV-1 undergoes recombination at a rate of ≥ 2.8 crossovers per genome per cycle (91). Inter-subtype recombinations are easier to detect, although recombination does take place between strains of the same subtype, and between different HIV-1 groups (57, 74).

Rapid replication (see Fig.3 “HIV-1 life cycle”) coupled with a relatively high intrinsic error rate provides the underlying means for rapid evolution and high levels of viral diversity. Estimates of the viral replication rate *in vivo* suggest an estimated 10^{10} virions are generated per day within an infected individual (31).

(A)



(B)

Inhibitors	Brand Name	Generic Name(s)	Approval Date
Fusion Inhibitors	Fuzeon	enfuvirtide, T-20	13-Mar-03
Nucleoside Reverse Transcriptase Inhibitors (NRTIs)	Combivir	lamivudine and zidovudine	27-Sep-97
	Emtriva	emtricitabine, FTC	02-Jul-03
	Epivir	lamivudine, 3TC	17-Nov-95
	Epzicom	abacavir and lamivudine	02-Aug-04
	Hivid	zalcitabine, dideoxycytidine, ddC	19-Jun-92
	Retrovir	zidovudine, azidothymidine, AZT, ZDV	19-Mar-87
	Trizivir	abacavir, zidovudine, and lamivudine	14-Nov-00
	Truvada	tenofovir disoproxil fumarate and emtricitabine	02-Aug-04
	Videx EC	enteric coated didanosine, ddI EC	31-Oct-00
	Videx	didanosine, dideoxyinosine, ddI	9-Oct-91
	Viread	tenofovir disoproxil fumarate, TDF	26-Oct-01
	Zerit	stavudine, d4T	24-Jun-94
	Ziagen	abacavir sulfate, ABC	17-Dec-98
Nonnucleoside Reverse Transcriptase Inhibitors (NNRTIs)	Rescriptor	delavirdine, DLV	4-Apr-97
	Sustiva	efavirenz, EFV	17-Sep-98
	Viramune	nevirapine, NVP	21-Jun-96

Protease Inhibitors (PIs)	Agenerase	amprenavir, APV	15-Apr-99
	Aptivus	tipranavir, TPV	22-Jun-05
	Crixivan	indinavir, IDV,	13-Mar-96
	Fortovase	saquinavir (no longer marketed)	7-Nov-97
	Invirase	saquinavir mesylate, SQV	6-Dec-95
	Kaletra	lopinavir and ritonavir, LPV/RTV	15-Sep-00
	Lexiva	Fosamprenavir Calcium, FOS-APV	20-Oct-03
	Norvir	ritonavir, RTV	1-Mar-96
	Prezista	darunavir	23-Jun-06
	Reyataz	atazanavir sulfate, ATV	20-Jun-03
	Viracept	nelfinavir mesylate, NFV	14-Mar-97
Multi-class Combination Products	Atripla	efavirenz, emtricitabine and tenofovir disoproxil fumarate	12-July-06

Figure 3. HIV-1 life cycle and the life cycle inhibitors.

(A) HIV-1 life cycle. Image from ref (79). The detailed HIV life cycle is reviewed in (29).

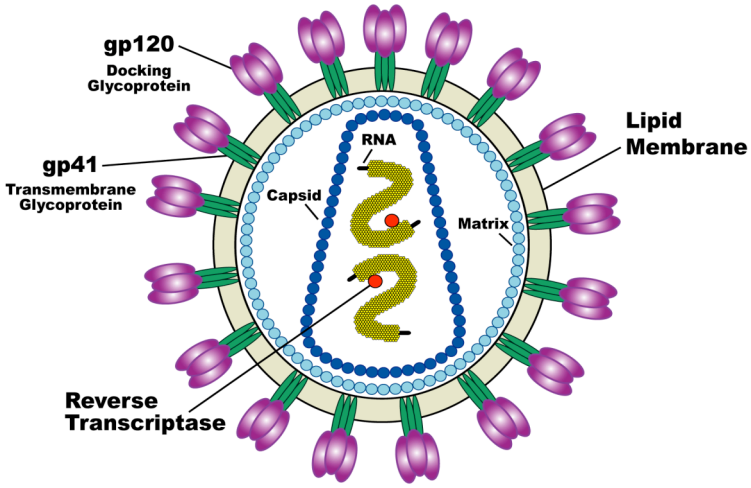
- Binding – HIV-1 attaches to CD4 antigen and a specific chemokine co-receptor. The virus and cell membrane fuse, and the viral core enters the cell. In most cases, HIV-1 uses either CCR5 or CXCR4 co-receptor. CCR5 is a necessary co-receptor for monocytotropic (M-tropic) HIV-1 isolates, and CXCR4 for T-cell-tropic (T-tropic) HIV isolates. The virus uses CCR5 is also called R5 virus, mostly appear in early disease. The virus uses CXCR4 is called X4 virus. It is more prevalent in late disease. X4 virus is associated with CD4 decline, HIV RNA increase, and clinical progress.
  (This binding step can be inhibited by fusion and entry inhibitors)
- Reverse transcription - HIV is uncoated inside the cell. Reverse transcriptase copies genomic RNA into DNA.
  (This step can be inhibited by transcriptase inhibitors)
- Integration - The double-stranded viral DNA is carried into the cell's nucleus by viral integrase, and it is integrated into the cellular DNA. HIV DNA is now called a provirus. This step makes the infection irreversible.
  (Integrase inhibitors can block this step)
- Transcription - Viral RNA is synthesized by the cellular enzyme RNA polymerase II using integrated viral DNA as a template.
- Translation - The mRNA is transported to the cytoplasm where the virus uses cellular machinery to synthesize viral proteins and enzymes.
- Assembly - RNA and viral enzymes gather at the edge of the cell. Viral protease cuts the polypeptides into viral proteins before viral matures.
  (Protease inhibitors can block this viral maturation step)
- Budding – Mature virus buds off from the cell.

(B) Inhibitors interfere with HIV-1 life cycle, and are used in the treatment of HIV infection. (Data source: www.fda.gov. As of May, 2007)

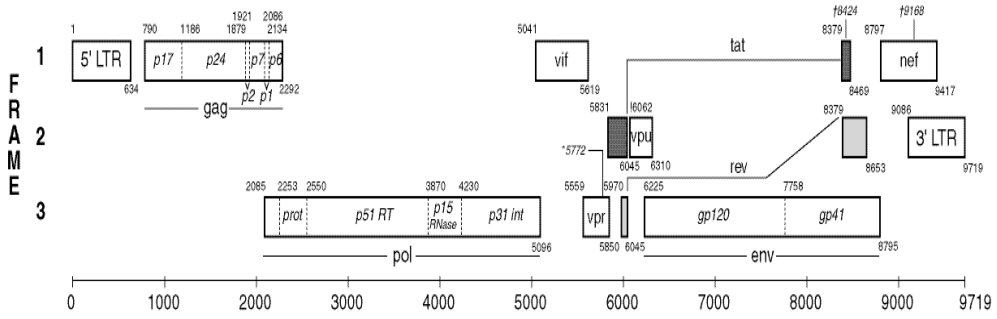
Finally, parts of the HIV genome are under more or less diversifying selection pressures. Most notably, *env* is under strong pressure from the antibodies generated by human immune system, encouraging mutation as escape from this pressure warrants survival. Other proteins evolve to escape T-cell immunity. Similarly, antiviral drug treatment causes positive selection on mutants that confer resistance. While positive/diversifying selection is important to amplify the genetic variation, other sites are under negative/purifying pressures as well as neutral evolution.

The aforementioned mechanisms are major forces in HIV-1 evolution. As a consequence, the rate of evolution was estimated to be 0.0024 substitutions per base pair per year for gp160 envelope and 0.0019 for gag (39), and even higher for variable parts of these genes (42). The genetic divergence of envelope protein (see Fig.4 “HIV-1 genome and proteins”) is as high as 35% among inter-subtype viruses, and 20% among inter-subtype viruses (24, 38, 63). Within a HIV-1 infected individual, highly related but non-identical viruses co-exist (“quasispecies”), and the viral diversity may reach levels of 10% or more during chronic infection (69).

(A)



(B)



(C)

Category	Function	Individual proteins		
		Name	**Function**	**Review Article**
Structural proteins	Essential components of the retroviral particle	Gag	Encoding capsid proteins. Precursor p55 is processed to p17 (Matrix), p24 (Capsid), p7 (NucleoCapsid), and p6 proteins, by the viral protease.	(26, 29)
		Env	The most variable region of the genome. Precursor gp160 is processed to gp120 and gp41. Both are important for viral entry. The V3 loop of gp120 is the principal target for neutralizing antibodies that block HIV-1 infectivity.	(29, 84)
Enzymatic Proteins	Encode viral enzymes	Pol	Encoding protease (Pro), reverse transcriptase (RT), ribonuclease H (RNase H) and integrase (IN). These enzymes are expressed within the context of a Gag-Pol precursor, which is processed by the viral protease.	(29)
Regulatory proteins	Modulate transcriptional and posttranscrip-tional steps of virus gene expression and are essential for virus propagation	Tat	Trans-activator of HIV gene expression. It activates transcription initiation and elongation from the LTR promoter.	(29, 34, 73)
		Rev	Regulator of virion. Promoting the nuclear export, stabilization and utilization of the viral mRNAs.	(29, 73)
Accessory proteins	Are not absolutely required for viral replication in all *in vitro* systems, but represent critical virulence factors *in vivo*.	Vif	Viral infectivity factor. Promoting the infectivity but not the production of viral particles.	(29, 73, 86)
		Vpr	Viral protein R. Its proposed functions include facilitating the nuclear localization of the preintegration complex, cell growth arrest, transactivation of cellular genes, and induction of cellular differentiation.	(9, 29, 53)
		Vpu	Viral protein U. Down-modulating CD4 and enhancing the virion release.	(29)
		Nef	The most immunogenic of the accessory proteins. Essential for efficient viral spread and disease progression *in vivo*. It increases viral infectivity.	(29, 58)

Figure 4. HIV-1 genome and proteins.

(A) Schematic representation of an HIV-1 virion. The virus is ~120 nm in diameter (60 times smaller than a red blood cell), and is spherical in shape. The figure is from reference (54). Gp120 proteins bind gp41 via non-covalent interactions and are associated as a trimer on the cell surface. Gp120, heavily glycosylated, is an active site for CD4 binding. Reviewed in (29, 84).

(B) A schematic view of the HIV-1 genome. Here the HXB2 strain (GenBank accession number: K03455) is shown. HXB2 strain is broadly used as the reference sequence in HIV-1 research (37). Figure from (43), and the figure legend is adapted from (43).

- Rectangles – open reading frames.
- Shaded rectangles - the tat and rev spliced exons.
- Small number in the upper left corner of each rectangle – the position of the a in the atg start codon. For pol, the start is taken to be the first t in the sequence ttttttag which forms part of the stem loop that potentiates ribosomal slippage on the RNA and a resulting -1 frameshift and the translation of the gag-pol polyprotein.
- Number in the lower right corner of each rectangle - the last position of the stop codon.
- Others: In HXB2, *5772 marks position of frameshift in the vpr gene caused by an "extra" t relative to most other subtype B viruses; !6062 indicates a defective acg start codon in vpu; †8424, and †9168 mark premature stop codons in tat and nef.
- LTR - Long terminal repeat. It contains important regulatory regions, especially those for transcription initiation and polyadenylation (40).

(C) Summary of HIV-1 proteins depicted in (A) and (B).

3. Viral evolution during HIV-1 disease progression is a complex process.

Three main transmission routes contribute to HIV-1 epidemic: sexual contact, blood-to-blood contact (mostly intravenous drug use), and mother-to-child transmission. The transmission efficiency is correlated to the virus inoculum and access to target cells, among many other reported factors, and possibly there is a selection of viruses at transmission. During the first few months of infection, HIV-1 increases in copy number but the viral population is homogenous (82, 87). Studies of mother-to-child transmission and sexual transmission have demonstrated that a restricted subset of viruses exists in receipts soon after infection (76, 81, 83, 89, 90). Following this acute infection period, strong HIV-specific cell-mediated immunity acts as the most dominant factor in specifically reducing viral load and, as a consequence, fitness and diversity (7, 27, 28, 64). It is reported that greater HIV-1 genetic diversity during the acute and early infection has been associated with faster disease progression (65). During chronic infection, viruses that have mutated to alter more than 10% of their DNA bases in the *env* arise (46, 69, 82). The overall viral diversity, as well as the divergence from the founder (transmitted) strain, is significantly increased (Fig.5).

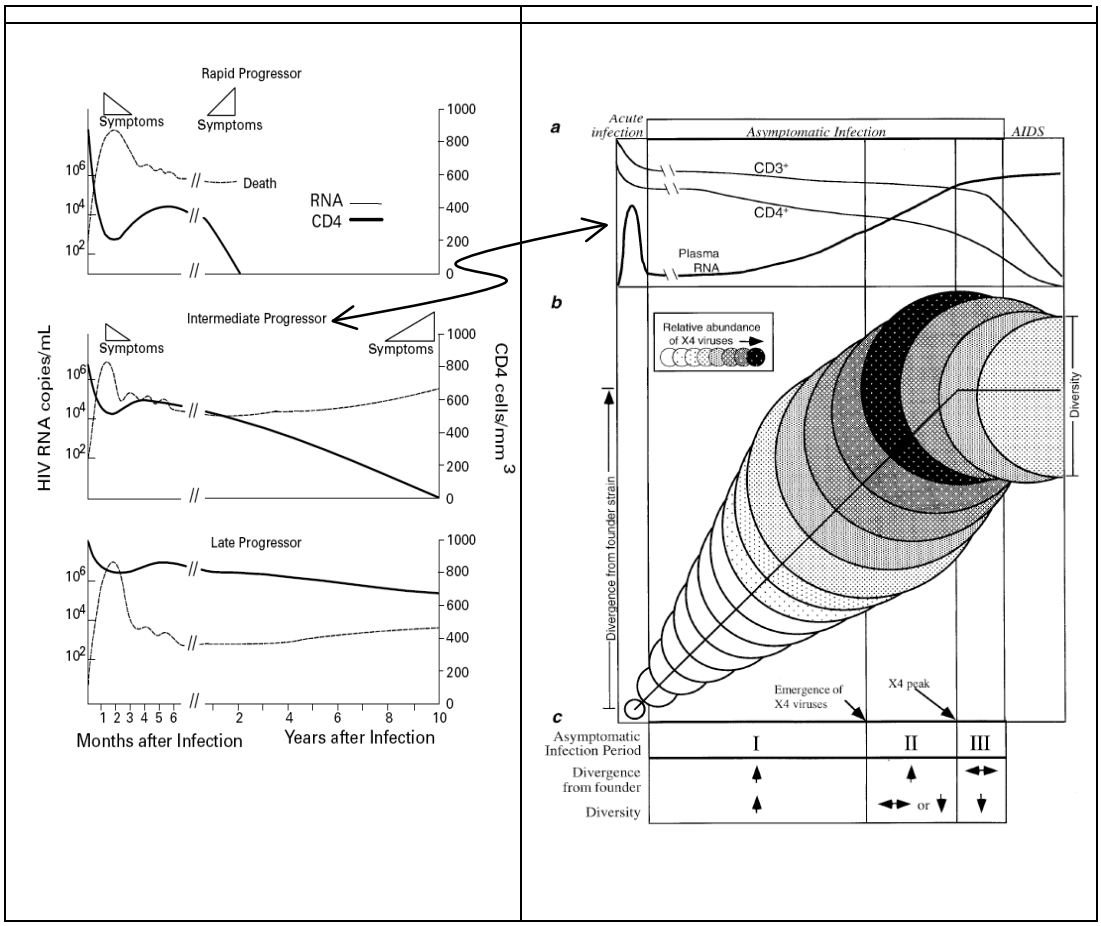


Figure 5. Genetic evolution during HIV-1 disease progression.

Figure on left: Time course of HIV-1 infection and disease. Figure From (1).
A small proportion of HIV-1 patients are rapid progressors, while approximately 5% of HIV-infected individuals exhibit no signs of disease progression even after 12 or more years (10, 55).

Figure on right: Schematic illustration of proposed patterns during HIV-1 disease progression in intermediate progressors. From (69). (a) Clinical phases of HIV infection as well as typical patterns of CD4+ and CD3+ T cells and plasma viral RNA loads. (b) Viral sequence evolution within the asymptomatic period of infection. Circle diameters represent the mean viral population diversities. Vertical displacement of the circles represents the extent of viral population divergence from the founder strain. Shading represents the proportion of the viral population comprised of viruses with an X4 genotype. (c) Characteristic changes in viral evolution in three proposed periods of the asymptomatic phase (↑ increasing; ↓ decreasing; ↔ stable). From (69).

The rate of viral evolution and the extent of genetic variation may vary from patient to patient over the disease course (Fig.6). Immune pressure (3, 59, 61), selective forces elicited by antiretroviral drugs (16, 66, 85), the random genetic drift of the viral population (neutral evolution) (16, 66, 85) and effect of compartmentalization (12, 13, 49, 67, 70) greatly contribute to the overall HIV-1 diversity for the duration of an infection. Neutral evolution and the extent of selective pressure is usually measured and compared by calculating the ratio of synonymous to non-synonymous substitutions. The continual generation of diversity may be important for viral persistence within the patient, as new immune responses are developed and the virus is under continuously changing immune responses (78). Thus at any given time, viral population is dominated by strains that are most fit at that time (47).

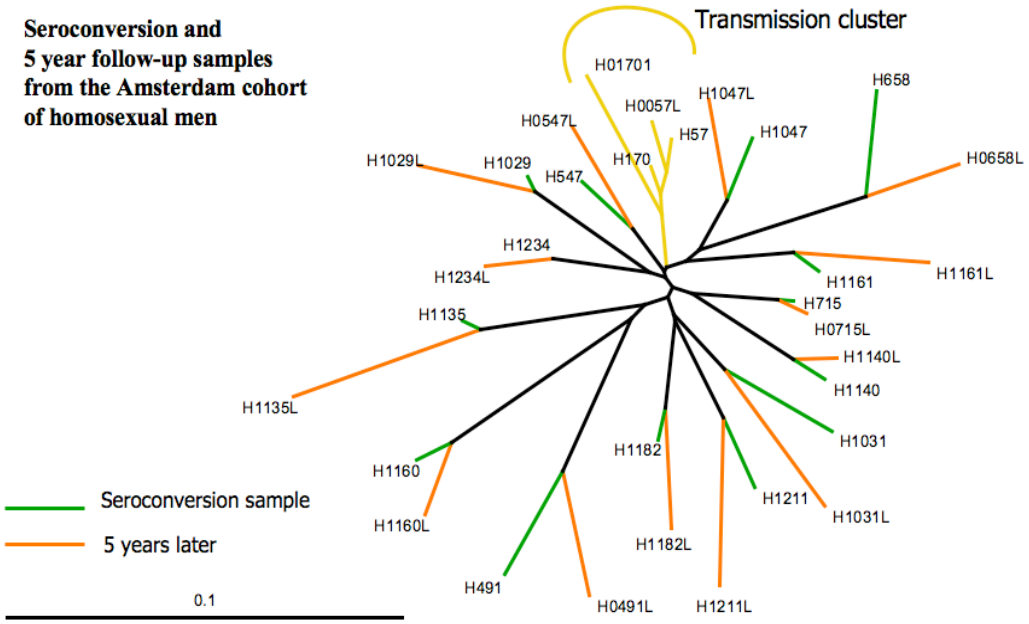


Figure 6. Amsterdam cohort: within-person evolution of HIV after five years. From (41).

4. The variability of HIV-1 is one of the major concerns in vaccine development.

Influenza vaccine has been proposed to be a model system for HIV vaccine design. Influenza A is also a highly variable virus. It diverges by 1–2%/year, making frequent vaccine updates necessary in order to elicit protection from one year to the next (38). However, this influenza vaccinology doesn't apply on HIV-1 due to the extraordinary contemporaneous variability of HIV-1 (Fig. 7). An Immune response against a HIV vaccine variant might not be active against the infecting variant. Furthermore, the latent reservoirs of HIV-1 make it impossible to completely eradicate the virus (8, 15). Recent HIV-1 vaccine designs, using either consensus sequences or ancestral sequences (18, 23, 24, 77) as vaccine antigen sets, can be expected to be more widely recognized than natural strains. This strategy, nevertheless, needs to be improved in order to provide broader coverage of HIV-1 variants and to enhance detection of HIV-1 specific T-cell responses (4, 20). A more recent HIV-1 vaccine study proposed a polyvalent vaccine comprising a small number of mosaic proteins, or genes encoding these proteins, that have been optimized to include common and exclude the most rare potential epitopes. The mosaics resemble natural proteins (or genes) as they were generated from *in silico* recombined natural sequences, thus *in vivo* processing of this kind of vaccine antigens will more closely resemble processing in natural infections (20).

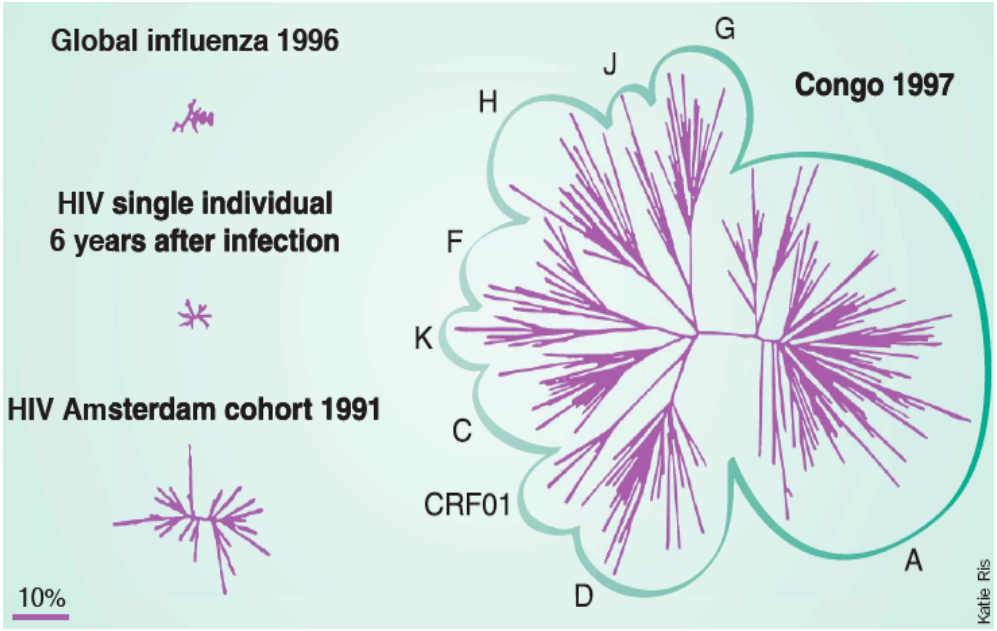


Figure 7. Sequence variation of HIV-1 (gp120, V2-C5) vs. influenza A H3N2 (HA, hemaglutinnin). Figure was originally from (38), and was adapted by (80).

HIV variation within a host 6 years after infection is similar to that of the global influenza A in a single year. A remarkable diverse set was observed in HIV sequences from Democratic Republic of Congo (DRC), where almost all HIV-1 subtypes were found.

AIMS OF THIS THESIS

HIV-1 recombination study

- To characterize a complex circulating recombinant form, CRF13_cpx.
- To track the dynamics of recombination in HIV-1 epidemic in different epidemiological settings.

HIV-1 gp120 study

- To track global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins in different HIV lineages.

HIV-1 sequence variability study

- To develope appropriate HIV-1 analysis methods and tools for the purpose of
  - Subtyping;
  - Evaluating sequence variation.

RESULTS AND DISCUSSION

1. Characterization of CRF13_cpx (paper I)

The molecular epidemiology and evolutionary history of the HIV-1 epidemic is muddled and complicated by HIV-1 recombination. The HIV-1 recombinants, mostly found in regions where multiple subtypes co-circulate (45, 51), are classified into circulating recombinant forms (CRF, the recombinants found in ≥ 3 unrelated patients) or unique recombinant forms (URF, the recombinants usually only found in a single patient) (63). HIV-1 CRFs are numbered sequentially in the order they were first adequately described in peer-reviewed publications.

Here we characterized a complex HIV-1 recombinant sequence (02CM.3226MN) sampled from Cameroon. It was found to cluster together with two previously reported CRF13 sequences, 96CM-4164 and 96CM-1849 (Fig. 8). Similarity plotting, bootscanning, breakpoint analysis, and phylogenetic trees (neighbor-joining and maximum likelihood analyses) also confirmed similar genomic structures with almost identical breakpoint positions among these three isolates. Thus, CRF13 now fulfills the HIV-1 nomenclature requirements; as three isolates have been obtained that share the same recombinant lineage.

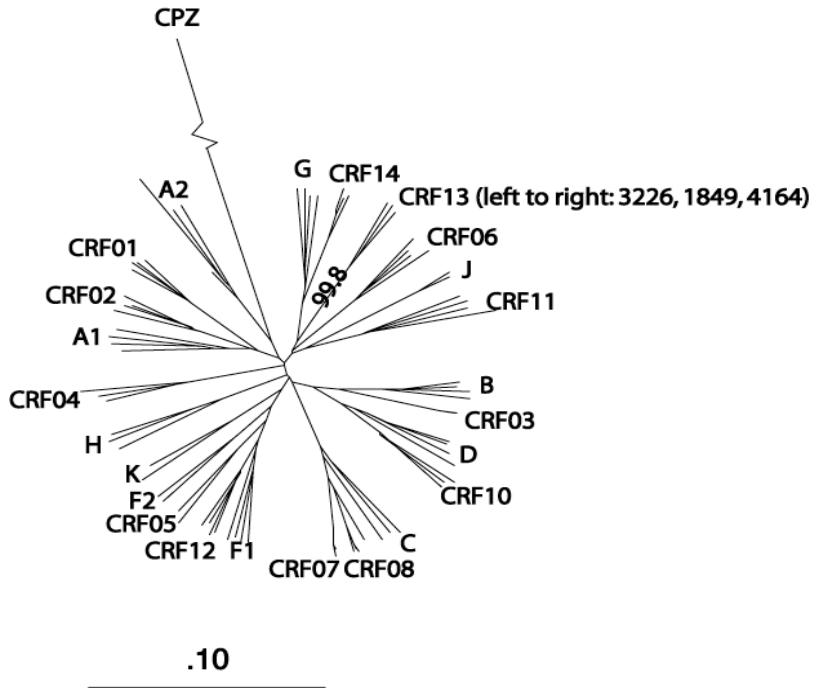


Figure 8. Phylogenetic tree showing the CRF13 cluser, as well as the known HIV-1 subtypes, sub-subtypes, CRF01-14, and SIVcpz.

The tree was made from full-length sequences (neighbor-joining tree with F84 model). The subtype sequences are the reference sequences retrieved from the Los Alamos HIV sequence database web site (http://www.hiv.lanl.gov).

The CRF13 genome consists of fragments of subtypes A1, G, J, CRF01 and one unclassified region. The J segment was found to be closer to J fragments of CRF11 similar to the way that A1/A2 and F1/F2 sub-subtypes associate. This suggests that further sampling may eventually result into re-classification of subtype J into sub-subtypes J1 and J2, as two rather divergent forms of J may be circulating. Such a re-classification, however, would depend on identification of three complete genomes of J1 and J2, under the current nomenclature requirements, and here we have simply identified divergent forms in a recombinant fragment. The unclassified region in CRF13 is an example of a situation where the current subtyping methods, based contemporary sequences, sometimes fail to detect sequences that either are of old origin, or their contemporary sequences haven't been identified.

We also developed a χ^2-based method that optimizes the breakpoints for all three CRF13 sequences simultaneously. Applying this method makes it feasible to locate the breakpoint uncertainty regions in all CRFs. Thus the breakpoint uncertainty regions defined by our method can provide a good reference for the uncertainty regions predicted by other subtyping tools.

2. Old and contemporary HIV-1 recombinants co-exist in the current HIV-1 epidemic (paper II)

Accurate HIV-1 subtyping information is vital for all kinds of HIV research. It gives insights into molecular epidemiology, viral evolution, and facilitates subtype-specific vaccine antigens and testing reagents. For varied reasons (different subtying methods, historical reasons, limited sequencing, etc.), some HIV sequences were misclassified or not classified at all (short sequences, etc.). Also, some sequences' subtyping information hasn't been updated after new subtypes, sub-subtypes, and recombinants were identified. Thus we developed several subtyping tools that can handle high throughput and still are capable of detection of recombinant forms accurately. These subtyping tools have been applied in an automated re-subtyping > 150,000 sequences of the Los Alamos HIV sequence database (www.hiv.lanl.gov) as an effort to enhance the HIV sequence quality control in the rapidly expanding HIV database and to better vaccine design. Two finished subtyping tools, are listed in Table I.

Tools	Algorithms	Features	Note
WBLAST (Window-BLAST based subtyping tool)	- Sliding a window along the query sequences and comparing them with a well-defined HIV DB reference set in each window	- Fast subtyping against the HIV DB or a set of user-defined sequences - Sequence submission in a batch mode - Vector or HIV lab-strain strain contamination detection - Preliminary detection of the breakpoints for recombinants.	Unpublished (Zhang, M., Leitner, T., and Korber, B)

		- Accuracy improved by adding trained sequence sets	
jpHMM (jumping profile hidden Markov model)	- Profile HMM (19)+ jumping alignments (72) - Profile HMMs were built for each HIV subtype and are connected by subtype transitions (jumps) between them	- Accurate in determining the involved subtypes and breakpoints for recombinants	**Paper IV and V in APPENDIX**

Table I. Two HIV-1 subtyping tools, among others, were used in re-subtyping the Los Alamos HIV sequence database.

Both tools were extensively tested on artificial recombinant and real recombinant sequences. jpHMM test result is detailed in **paper V**. The web-interfaced jpHMM is available at http://jphmm.gobics.de, and is described in **paper IV**.

Phylogenetic analyses, distance methods, and jpHMM were applied in all recombinants (CRFs and URFs) involving subtypes A/G, or B/C, or B/F. The CRFs included in this study are CRF02_AG, CRF07_BC, CRF08_BC, CRF12_BF, CRF17_BF, CRF28_BF, and CRF29_BF. AG CRFs (CRF02) cluster together, with long branches to their most recent common ancestor, in every phylogenetic trees we examined. This tree result, together with the comparison with a sequence (Z321. Accession number U76035) sampled in 1976 (14, 25), suggests that AG CRFs (CRF02) are from an old recombinant lineage, with the latest recombination events and founder virus occurring possibly before mid-1970s. In BC CRFs, the phylogenetic trees indicate that CRF07 was derived from a recombination event between CRF08 and subtype B. However, the immediate CRF07 descendent of CRF08 and B is unlikely to have been found, and the currently defined CRF07 set is a variation. The BF epidemic in South America is very unique, as CRFs are easily outnumbered by URFs. In this study, we proposed that BF epidemics in two HIV-1 epicenters in South America, Argentina and Brazil, are not independent/separate as it was thought before. Shared breakpoints among Argentinean and Brazilian BF sequences may fuel the rampant spread of HIV-1 in South America. Regarding the BF origin in Argentina, we suggest that BF in Argentina did not necessarily originated in Brazil, as a composite

of all Argentinean B and F fragments can cover a full “Argentinean” HIV-1 genome of each subtype.

The AG, BC, BF recombinants analyses also indicate that all HIV-1 sequences known today are just some samples from a big complex and dynamic pool of HIV-1 sequences, in which old and new recombinants co-exist and still has some traces left by the extinct sequences (Fig. 9). In this regard, it may be more appropriate to define recombinants using recombinant families rather than CRFs. The biggest difference between these two is that, the CRF definition is more focused on having the exact or very similar breakpoints among all members within a CRF, and thus be the consequence of a single lineage from an initial recombinant form. Exact or very similar breakpoints may be easily blurred by rapid evolution of HIV-1, or by being embedded in a conserved region where precise breakpoints are difficult to resolve. Therefore the sequences defined in a CRF are possibly snapshots during a given time period of a dynamic HIV-1 picture. We suggest to use a “recombinant family” definition to define recombinants consisting of the same subtypes to reflect the dynamic feature of the HIV-1 epidemic.

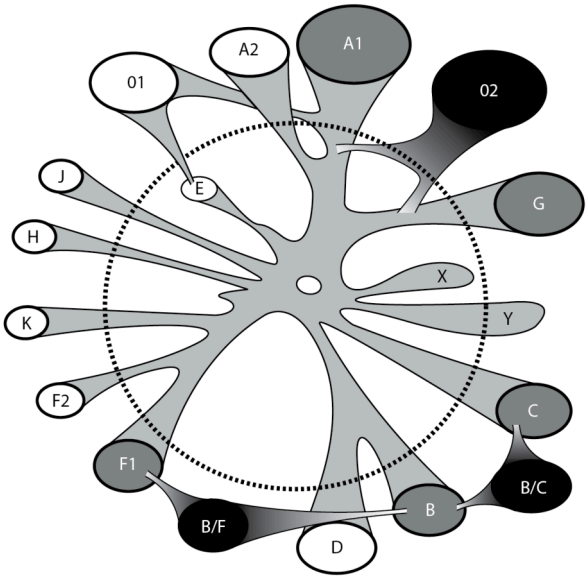


Figure 9. Contemporary sequences co-exist with some old sequences in the current HIV-1 epidemic.

The dashed circle differentiates the old and contemporary sequences. Inside the circle, the old sequences, like subtype E strains, may be no longer exist in the epidemic. We can only deduce its old presence based on CRF01_AE, a recombinant between subtype A and E. "X" represents an extinct strain, "Y" represents an old strain still circulating in the current epidemic, but it hasn't been identified. CRF02 is an old recombinant derived from old A and old G. BF and BC recombinants are rather new. Their parental sequences are contemporary sequences.

3. N-linked glycosylation site (sequon) variation in HIV, HCV, and influenza glycoproteins (paper III).

N-linked glycosylation sites (or sequons, enabled by the amino acid patten NXS or NXT, N: asparagines, X: any amino acid, S: serine, T: threonine) (50) are a critical component of the external proteins of primate lentiviruses, influenza, and hepatitis C viruses, and their modification can be important for evolution of escape from the immune response. The gain or loss of such sites can play a key role in viral infectivity, antigen conformation, and immune escape.

We explored N-linked glycosylation site (sequon) variation at the population level in aforementioned viruses using a web-based glycosite tool (http://www.hiv.lanl.gov/content/hiv-db/GLYCOSITE/glycosite.html) developed to facilitate the sequon tracking and to define patterns in evolution. Two distinctive patterns of sequon variation were identified in HIV-1, HIV-2, and SIV CPZ. The first pattern (fixed) describes readily aligned sites that are either simply present or absent. These sites tend to be occupied by high mannose glycans, and are involved with binding to DC-SIGN, a lectin that facilitates HIV-1 infections of cells (32). The second pattern (shifting) refers to sites embedded in regions of extreme local length variation and is characterized by shifts in terms of the relative position and

local density of sequons; these sites tend to be populated by complex carbohydrates. HIV, with its extreme variation in number and precise location of sequons, does not have a net increase in the number of sites over time at the population level. Primate lentiviral lineages have host species - dependent levels of sequon shifting, with HIV-1 in humans the most extreme (Fig. 10). HCV envelope proteins, despite evolving extremely rapidly through point mutation, show limited sequon variation, although two shifting sites were identified. Human influenza A hemagglutinin H3 HA1 is accumulating sequons over time, but this trend is not evident in any other avian or human influenza A serotypes. Among these studied viruses, HIV-1 seems to have a unique evolutionary avenue for immune evasion partially due to its ability of generating and tolerating shifting sequons in viral proteins (78).

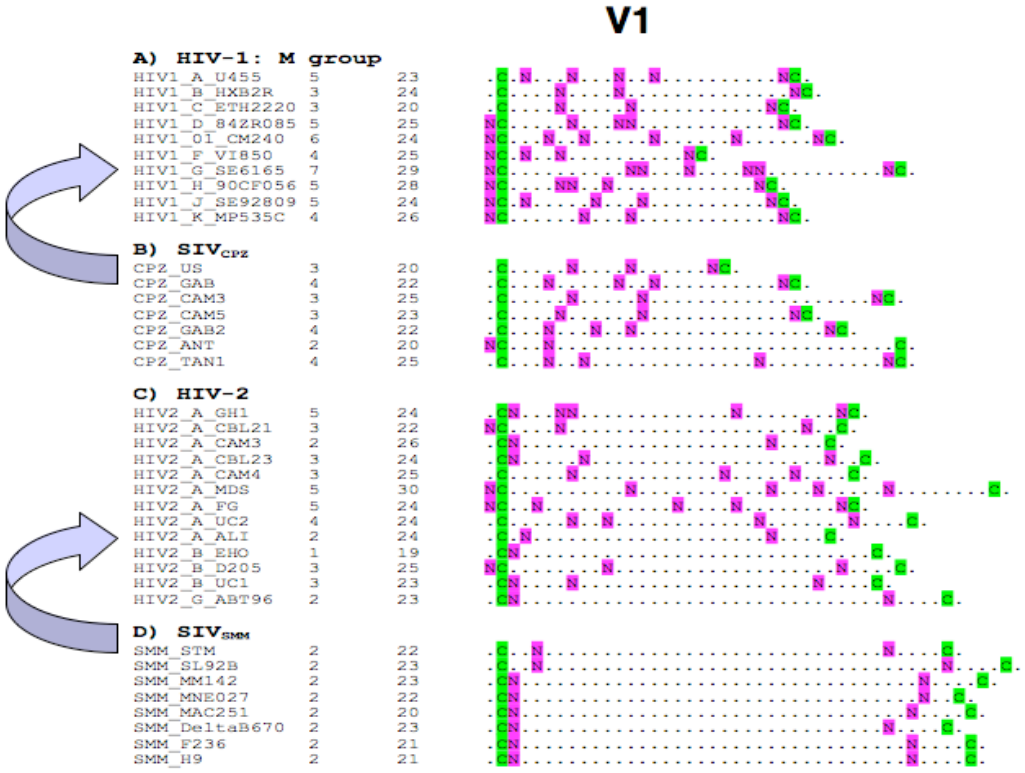


Figure 10. N-linked glycosylation patterns in primate lentiviruses.

V1 loop in gp120 is shown as an example. Only N-linked glycosylation sites (N, in red) and

Cys (C, in green), which close the base of the V1 loop, are marked, all other amino acids are indicated by a period (.), to highlight the relative change of position of the N-linked sites and the length variation of the V1 loop. Part A includes one randomly selected sequence from each clade to represent HIV-1 M group diversity. Part B includes available viral sequences isolated from chimpanzees, thought to be the source of the human HIV-1 epidemic. Part C lists viruses from HIV-2 clades A and B, and part D from sooty mangabey, which are thought to be the source of the HIV-1 epidemic. Arrows, possible zoonotic transmissions.

We also applied the analyses of N-linked glycosylation site, together with sequence length, in a study aimed at creating a well-characterized panel of subtype C gp160 reference clones to facilitate standardized assessments of vaccine-elicited neutralizing antibody responses (**paper VI in Appendix**). Eighteen subtype C sequences from acute/early heterosexually infections (C-ref) were compared with newly transmitted subtype B viruses (B-ref) and the subtype B (B-db) and C sequences (C-db) from the Los Alamos HIV sequence database which sequences are mostly from chronic viruses. The gp120 sequence lengths in C-ref are shorter (most obvious in V1 and V4) and less glycosylated than B-ref (Table II, and Fig.11). C-ref is underglycosylated but no different in length compared to C-db of which most are chronic subtype C viruses. All C-ref retain sequon at HXB2 position 301, an important position known to mask V3 epitopes on subtype B viruses (5, 68), and possess a V2 loop that on average was the same size as B-ref. These structural features may confer an effective masking of V3 epitopes on C-ref viruses. It was also found that there is a significant trend toward a greater number of sequons on the gp120 of B-ref compared with B-db (most are chronic B). These observations, however, need further support from a larger number of gp120 sequences from acutely and chronically infected individuals in order to confirm if newly transmitted viruses have unique genetic features.

Genomic region	Analysis group[a]	Sequence length interquartile	Glycan no. interquartile	Comparison	*P* value	
					Sequence length comparison[b]	Glycan no. comparison[b]
gp120	B-ref	509–519	26–28	B-ref/C-ref	0.01	0.01
	C-ref	492–509	24–26	B-db/C-db	**8.44×10^{-15}**	0.004
	B-db	507–515	24–27	B-ref/B-db	0.44	0.02
	C-db	499–510	23–26	C-ref/C-db	0.45	0.86
V1	B-ref	25–31	4–5	B-ref/C-ref	0.03	0.06
	C-ref	17–27	3–4	B-db/C-db	**3.37×10^{-13}**	**0.0008**
	B-db	26–32	4–5	B-ref/B-db	0.75	0.43
	C-db	22–28	3–5	C-ref/C-db	0.31	0.45
V2	B-ref	40–46	4–5	B-ref/C-ref	0.52	0.85
	C-ref	42–45	4–5	B-db/C-db	**3.33×10^{-5}**	0.50
	B-db	40–44	4–5	B-ref/B-db	0.23	0.52
	C-db	41–45	4–5	C-ref/C-db	0.16	0.88
V3	B-ref	35–35	3–3	B-ref/C-ref	0.36	0.0036
	C-ref	35–35	1–2	B-db/C-db	0.02	**$<2.2 \times 10^{-16}$**
	B-db	35–35	2–3	B-ref/B-db	0.61	0.4259
	C-db	35–35	2–2	C-ref/C-db	0.60	0.7318
V4	B-ref	31–33	4–5	B-ref/C-ref	0.005	0.02
	C-ref	23–30	3–4	B-db/C-db	**1.74×10^{-12}**	0.02
	B-db	30–33	4–5	B-ref/B-db	0.50	0.75
	C-db	27–31	4–5	C-ref/C-db	0.06	0.01
V5	B-ref	9–10	2–2	B-ref/C-ref	0.57	0.09
	C-ref	9–11	1–2	B-db/C-db	0.13	0.004
	B-db	9–11	1–2	B-ref/B-db	0.26	0.12
	C-db	9–11	1–2	C-ref/C-db	0.83	0.71

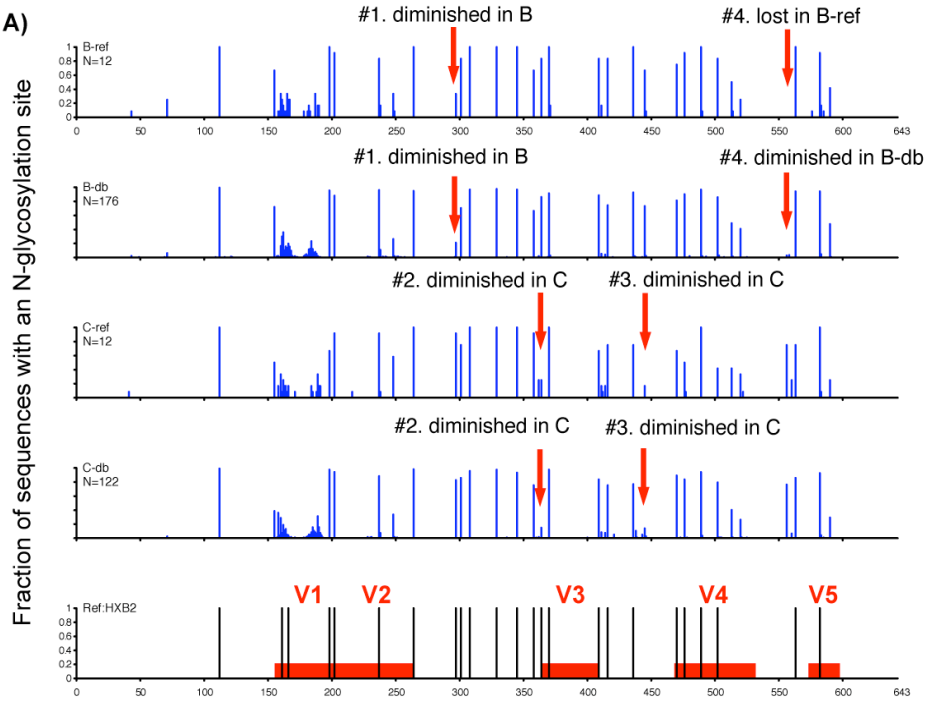
[a] B-ref, 12 selected sequences from the subtype B panel of reference strains; B-db, 176 well-characterized subtype B database sequences; C-ref, 12 selected sequences from the C panel of reference strains; C-db, 122 well-characterized subtype C database sequences.

[b] Bolded *P* values ($P < 0.002$) are significant differences after Bonferroni correction. Underlined *P* values ($0.002 < P < 0.05$) are significantly different trends after Bonferroni correction.

Table II. Comparison of sequence lengths and N-linked glycosylation sites (sequons) between HIV-1 subtype B and C gp120.

N-glycosylation Site Numbers In B and C Sequences (gp120 region)

A)



B)

Highlighted # in the plot	Position in the alignment	Position relative to the protein start in HXB2 gp160	Contingency table summary			
			Groups	Seq # WITH N-glyco site	Seq # WITHOUT N-glyco site	P-values (two-sided fisherís exact test)
1	297	230	B-db	37	139	$P < 2.2x10^{-16}$
			C-db	101	21	
2	364	295	B-db	152	24	
			C-db	18	104	
3	445	362	B-db	129	47	
			C-db	17	105	
4	556	442	B-db	5	171	
			C-db	93	29	

Figure 11. Comparison of N-linked glycosylation sites (sequons) by sequence positions.

Some loss/gain of sequons (highlighted in part A) were statistically significant (part B).

CONCLUSIONS

HIV-1 recombination study

- Subtype J may actually reflect more divergent sub-subtypes J1 and J2, where J2 is represented by fragments in CRF11 and CRF13. But full length genomes of the ancestral subtype would need to be identified to validate the diversity that is suggested by the fragments.
- Recombinants blur subtype phylogeny.
- Recombinant families may be a useful epidemiological concept.
- Old and new recombinants co-exist in the current HIV-1 epidemic.

HIV-1 gp120 study

- The extent of N-linked glycosylations site variation is host and lineage dependent.
- The mechanism that generate shifting sites and tolerance of the shifting N-linked glycosylation sites in viral proteins provides a unique evolutionary avenue for immune evasion in HIV-1. It needs to be considered in vaccine design, and may have subtype-specific constrains.
- N-linked glycosylation sites do not accumulate over time in HIV, indicating the evolutionary importance and selective balance of both the loss, gain, and relative position of N-linked glycosylation sites in HIV.

Developed web-based tools in the study

- Subtyping
  - Window-BLAST: fast strategy for detecting HIV contaminations and recombinants. (Internally used in HIV sequence database group. Will be released to the public soon.)

- Web-interfaced jpHMM: good graphic representation of predicted recombinants. http://jphmm.gobics.de.

- Nglyco: tracking N-linked glycosylation sites. http://www.hiv.lanl.gov/content/hiv-db/GLYCOSITE/glycosite.html
- Shannon Entropy: statistically evaluate sequence variation at each sequence position within or between sequence alignments. http://www.hiv.lanl.gov/content/hiv-db/ENTROPY/entropy.html
    - Tree-based entropy: implementing phylogenetic relationships in Shannon Entropy (Internally used in HIV sequence database group. Will be released to the public soon).

ACKNOWLEDGEMENT

The work of this thesis was carried out at the Department of Bioinformatics, Institute of Microbiology and Genetics, University of Goettingen, Germany and the HIV Sequence Database group at the Los Alamos National Lab, USA. During the last four years I was working on this thesis, I've been accompanied and supported by many great people that I'd like to say THANK YOU here.

My advisors - Drs. Bette Korber, Thomas Leitner, and Burkhard Morgenstern. I have been lucky enough to study under the excellent guidance of three brilliant scientists! Thank you for your enthusiastic supervision, for sharing your outstanding scientific expertise and ideas, for teaching me, for inspiring me, for your great efforts of making me grow!!

And thanks to Bette and Thomas for making the database group a stimulating and fun environment! My database group colleagues, Charles Calef, Carla Kuiken, Brian Foley, Brian Gaschen, Werner Abfalterer, Jennifer Macke, James Szinger, Will Fischer, Peter Hraber, …, thank you all for the warm-hearted help and support, for the encouragement, and for your valuable hints.

I am also indebted to my Germany colleagues. Anne-Kathrin Schultz, Britta Leinemann, Rasmus Steinkamp, Maike Tech, Peter Meinicke, Thomas Lingner, Katharina Hoff, Fabian Schreiber, Isabelle Heinemeyer,…., thank you for providing a lot of help during my stay in Germany, and it has been a pleasure to work with you all.

My friends and former intern advisors at CDC. Robert Wohlhueter, Joe Esposito, and Scott Sammons. Thank you for introducing me into the fascinating world of viruses!

To Profs. Thomas Friedl, Edgar Wingender, and Stephan Waack. Thank you for being my thesis committee member, and for your encouragement and support!

And the most important, I own everything to my parents. Thank you for your infinite love, and for always being there for me!

My brother, thank you for the encouragement and help whenever I need it!

And to my husband, thank you for your love, patience, and understanding!

REFERENCES

1. 1998. Report of the NIH Panel to Define Principles of Therapy of HIV Infection. MMWR Recomm Rep **47:**1-41.
2. **-** 2006, posting date. UNAIDS/WHO 2006 report on the global AIDS epidemic. (www.unaids.org). [Online.]
3. **Allen, T. M., D. H. O'Connor, P. Jing, J. L. Dzuris, B. R. Mothe, T. U. Vogel, E. Dunphy, M. E. Liebl, C. Emerson, N. Wilson, K. J. Kunstman, X. Wang, D. B. Allison, A. L. Hughes, R. C. Desrosiers, J. D. Altman, S. M. Wolinsky, A. Sette, and D. I. Watkins.** 2000. Tat-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia. Nature **407:**386-90.
4. **Altfeld, M., M. M. Addo, R. Shankarappa, P. K. Lee, T. M. Allen, X. G. Yu, A. Rathod, J. Harlow, K. O'Sullivan, M. N. Johnston, P. J. Goulder, J. I. Mullins, E. S. Rosenberg, C. Brander, B. Korber, and B. D. Walker.** 2003. Enhanced detection of human immunodeficiency virus type 1-specific T-cell responses to highly variable regions by using peptides based on autologous virus sequences. J Virol **77:**7330-40.
5. **Back, N. K., L. Smit, J. J. De Jong, W. Keulen, M. Schutten, J. Goudsmit, and M. Tersmette.** 1994. An N-glycan within the human immunodeficiency virus type 1 gp120 V3 loop affects virus neutralization. Virology **199:**431-8.
6. **Barre-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier.** 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science **220:**868-71.
7. **Betts, M. R., D. R. Ambrozak, D. C. Douek, S. Bonhoeffer, J. M. Brenchley, J. P. Casazza, R. A. Koup, and L. J. Picker.** 2001. Analysis of total human immunodeficiency virus (HIV)-specific CD4(+) and CD8(+) T-cell responses: relationship to viral load in untreated HIV infection. J Virol **75:**11983-91.
8. **Blankson, J. N., D. Persaud, and R. F. Siliciano.** 2002. The challenge of viral reservoirs in HIV-1 infection. Annu Rev Med **53:**557-93.
9. **Bukrinsky, M., and A. Adzhubei.** 1999. Viral protein R of HIV-1. Rev Med Virol **9:**39-49.
10. **Cao, Y., L. Qin, L. Zhang, J. Safrit, and D. D. Ho.** 1995. Virologic and immunologic characterization of long-term survivors of human immunodeficiency virus type 1 infection. N Engl J Med **332:**201-8.
11. **CDC.** 1981. Pneumocystis pneumonia--Los Angeles. MMWR Morb Mortal Wkly Rep **30:**250-2.
12. **Chang, J., R. Jozwiak, B. Wang, T. Ng, Y. C. Ge, W. Bolton, D. E. Dwyer, C. Randle, R. Osborn, A. L. Cunningham, and N. K. Saksena.** 1998. Unique HIV type 1 V3 region sequences derived from six different

regions of brain: region-specific evolution within host-determined quasispecies. AIDS Res Hum Retroviruses **14:**25-30.

13. **Cheynier, R., S. Henrichwark, F. Hadida, E. Pelletier, E. Oksenhendler, B. Autran, and S. Wain-Hobson.** 1994. HIV and T cell expansion in splenic white pulps is accompanied by infiltration of HIV-specific cytotoxic T lymphocytes. Cell **78:**373-87.
14. **Choi, D. J., S. Dube, T. P. Spicer, H. B. Slade, F. C. Jensen, and B. J. Poiesz.** 1997. HIV type 1 isolate Z321, the strain used to make a therapeutic HIV type 1 immunogen, is intersubtype recombinant. AIDS Res Hum Retroviruses **13:**357-61.
15. **Chun, T. W., and A. S. Fauci.** 1999. Latent reservoirs of HIV: obstacles to the eradication of virus. Proc Natl Acad Sci U S A **96:**10958-61.
16. **Condra, J. H.** 1998. Resistance to HIV protease inhibitors. Haemophilia **4:**610-5.
17. **Cornelissen, M., R. van den Burg, F. Zorgdrager, V. Lukashov, and J. Goudsmit.** 1997. pol gene diversity of five human immunodeficiency virus type 1 subtypes: evidence for naturally occurring mutations that contribute to drug resistance, limited recombination patterns, and common ancestry for subtypes B and D. J Virol **71:**6348-58.
18. **Doria-Rose, N. A., G. H. Learn, A. G. Rodrigo, D. C. Nickle, F. Li, M. Mahalanabis, M. T. Hensel, S. McLaughlin, P. F. Edmonson, D. Montefiori, S. W. Barnett, N. L. Haigwood, and J. I. Mullins.** 2005. Human immunodeficiency virus type 1 subtype B ancestral envelope protein is functional and elicits neutralizing antibodies in rabbits similar to those elicited by a circulating subtype B envelope. J Virol **79:**11214-24.
19. **Durbin, R., S. Eddy, A. Krogh, and G. Mitchison.** 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press.
20. **Fischer, W., S. Perkins, J. Theiler, T. Bhattacharya, K. Yusim, R. Funkhouser, C. Kuiken, B. Haynes, N. L. Letvin, B. D. Walker, B. H. Hahn, and B. T. Korber.** 2007. Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nat Med **13:**100-6.
21. **Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn.** 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature **397:**436-41.
22. **Gao, F., Y. Chen, D. N. Levy, J. A. Conway, T. B. Kepler, and H. Hui.** 2004. Unselected mutations in the human immunodeficiency virus type 1 genome are mostly nonsynonymous and often deleterious. J Virol **78:**2426-33.
23. **Gao, F., E. A. Weaver, Z. Lu, Y. Li, H. X. Liao, B. Ma, S. M. Alam, R. M. Scearce, L. L. Sutherland, J. S. Yu, J. M. Decker, G. M. Shaw, D. C. Montefiori, B. T. Korber, B. H. Hahn, and B. F. Haynes.** 2005. Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group m consensus envelope glycoprotein. J Virol **79:**1154-63.
24. **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.

25. **Getchell, J. P., D. R. Hicks, A. Svinivasan, J. L. Heath, D. A. York, M. Malonga, D. N. Forthal, J. M. Mann, and J. B. McCormick.** 1987. Human immunodeficiency virus isolated from a serum sample collected in 1976 in Central Africa. J Infect Dis **156:**833-7.
26. **Göttlinger, H. G.** 2001. HIV-1 Gag: a Molecular Machine Driving Viral Particle Assembly and Release, p. **2-28**. *In* C. Kuiken, B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber (ed.), HIV Sequence Compendium 2001. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 02-2877.
27. **Goulder, P. J., M. A. Altfeld, E. S. Rosenberg, T. Nguyen, Y. Tang, R. L. Eldridge, M. M. Addo, S. He, J. S. Mukherjee, M. N. Phillips, M. Bunce, S. A. Kalams, R. P. Sekaly, B. D. Walker, and C. Brander.** 2001. Substantial differences in specificity of HIV-specific cytotoxic T cells in acute and chronic HIV infection. J Exp Med **193:**181-94.
28. **Goulder, P. J., and D. I. Watkins.** 2004. HIV and SIV CTL escape: implications for vaccine design. Nat Rev Immunol **4:**630-40.
29. **Greene, W. C., Peterlin, B. M.** 2005. Molecular insights into HIV biology. *In* L. Peiperl, S. Coffey, O. Bacon, and P. Volberding (ed.), HIV Insite Knowledge Base. Univ. of California San Francisco and San Francisco General Hospital, San Francisco.
30. **Gurtler, L. G., L. Zekeng, J. M. Tsague, A. van Brunn, E. Afane Ze, J. Eberle, and L. Kaptue.** 1996. HIV-1 subtype O: epidemiology, pathogenesis, diagnosis, and perspectives of the evolution of HIV. Arch Virol Suppl **11:**195-202.
31. **Ho, D. D.** 1997. Perspectives series: host/pathogen interactions. Dynamics of HIV-1 replication in vivo. J Clin Invest **99:**2565-7.
32. **Hong, P. W., S. Nguyen, S. Young, S. V. Su, and B. Lee.** 2007. Optimal DC-SIGN binding to HIV-1 gp120 involves specific N-glycans within the 2G12 epitope. J Virol.
33. **Hooper, E.** 1997. Sailors and star-bursts, and the arrival of HIV. Bmj **315:**1689-91.
34. **Karn, J.** 2000. Tat, a novel regulator of HIV transcription and latency, p. 2-18. *In* C. Kuiken, F. McCutchan, B. Foley, J. Mellors, B. Hahn, J. Mullins, P. Marx, and S. Wolinsky (ed.), HIV Sequence Compendium 2000. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.
35. **Keele, B. F., F. Van Heuverswyn, Y. Li, E. Bailes, J. Takehisa, M. L. Santiago, F. Bibollet-Ruche, Y. Chen, L. V. Wain, F. Liegeois, S. Loul, E. M. Ngole, Y. Bienvenue, E. Delaporte, J. F. Brookfield, P. M. Sharp, G. M. Shaw, M. Peeters, and B. H. Hahn.** 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science **313:**523-6.
36. **Kolata, G.** 1987. Boy's 1969 death suggests AIDS invaded U.S. several times, The New York Times, New York.
37. **Korber, B., B. Foley, C. Kuiken, S. Pillai, and J. Sodroski.** 1998. Numbering Positions in HIV Relative to HXB2CG, p. III-102-111. *In* B. Korber, C. Kuiken, B. Foley, B. Hahn, F. McCutchan, J. Mellors, and J. Sodroski (ed.), Human Retroviruses and AIDS 1998. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

38. **Korber, B., B. Gaschen, K. Yusim, R. Thakallapally, C. Kesmir, and V. Detours.** 2001. Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull **58:**19-42.
39. **Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya.** 2000. Timing the ancestor of the HIV-1 pandemic strains. Science **288:**1789-96.
40. **Krebs, F., T. Hogan, S. Quiterio, S. Gartner, and W. B.** 2001. Lentiviral LTR-directed Expression, Sequence Variation, and Disease Pathogenesis, p. 29-70. *In* C. Kuiken, B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber (ed.), HIV Sequence Compendium 2001. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, LA-UR 02-2877.
41. **Kuiken, C. L., V. V. Lukashov, E. Baan, J. Dekker, J. A. Leunissen, and J. Goudsmit.** 1996. Evidence for limited within-person evolution of the V3 domain of the HIV-1 envelope in the amsterdam population. Aids **10:**31-7.
42. **Leitner, T., and J. Albert.** 1999. The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc Natl Acad Sci U S A **96:**10752-7.
43. **Leitner, T., B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber.** 2005. HIV Sequence Compendium 2005, HIV Sequence Compendium 2005. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, LA-UR number 06-0680.
44. **Levy, J. A., A. D. Hoffman, S. M. Kramer, J. A. Landis, J. M. Shimabukuro, and L. S. Oshiro.** 1984. Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. Science **225:**840-2.
45. **Los Alamos HIV Sequence Database Group**, posting date. HIV-1 geography site. http://www.hiv.lanl.gov/components/hiv-db/new_geography/geography.comp?region=world&form=all. [Online.]
46. **Lukashov, V. V., C. L. Kuiken, and J. Goudsmit.** 1995. Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period. J Virol **69:**6911-6.
47. **Malim, M. H., and M. Emerman.** 2001. HIV-1 sequence variation: drift, shift, and attenuation. Cell **104:**469-72.
48. **Mansky, L. M., and H. M. Temin.** 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol **69:**5087-94.
49. **Marras, D., L. A. Bruggeman, F. Gao, N. Tanji, M. M. Mansukhani, A. Cara, M. D. Ross, G. L. Gusella, G. Benson, V. D. D'Agati, B. H. Hahn, M. E. Klotman, and P. E. Klotman.** 2002. Replication and compartmentalization of HIV-1 in kidney epithelium of patients with HIV-associated nephropathy. Nat Med **8:**522-6.
50. **Marshall, R. D.** 1974. The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. Biochem Soc Symp**:**17-26.
51. **McCutchan, F. E.** 2006. Global epidemiology of HIV. J Med Virol **78 Suppl 1:**S7-S12.
52. **Meloni, S. T., B. Kim, J. L. Sankale, D. J. Hamel, S. Tovanabutra, S. Mboup, F. E. McCutchan, and P. J. Kanki.** 2004. Distinct human

immunodeficiency virus type 1 subtype A virus circulating in West Africa: sub-subtype A3. J Virol **78:**12438-45.

53. **Muthumani, K., A. Y. Choo, W. X. Zong, M. Madesh, D. S. Hwang, A. Premkumar, K. P. Thieu, J. Emmanuel, S. Kumar, C. B. Thompson, and D. B. Weiner.** 2006. The HIV-1 Vpr and glucocorticoid receptor complex is a gain-of-function interaction that prevents the nuclear localization of PARP-1. Nat Cell Biol **8:**170-9.
54. **NIH - National Institute of Allergy and Infectious Disease** 2006, posting date. HIV fact sheets. [Online.]
55. **Pantaleo, G., S. Menzo, M. Vaccarezza, C. Graziosi, O. J. Cohen, J. F. Demarest, D. Montefiori, J. M. Orenstein, C. Fox, L. K. Schrager, and et al.** 1995. Studies in subjects with long-term nonprogressive human immunodeficiency virus infection. N Engl J Med **332:**209-16.
56. **Peeters, M.** 2005. Unpublished data. Communication with HIV sequence database group.
57. **Peeters, M., F. Liegeois, N. Torimiro, A. Bourgeois, E. Mpoudi, L. Vergne, E. Saman, E. Delaporte, and S. Saragosti.** 1999. Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient. J Virol **73:**7368-75.
58. **Piguet, V., and D. Trono.** 1999. A Structure-function analysis of the Nef Protein of Primate Lentiviruses, p. 448-459. *In* C. Kuiken, B. Foley, B. Hahn, B. Korber, F. McCutchan, P. Marx, J. Mellors, J. Mullins, J. Sodroski, and S. Wolinksy (ed.), Human Retroviruses and AIDS 1999. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.
59. **Poignard, P., R. Sabbe, G. R. Picchio, M. Wang, R. J. Gulizia, H. Katinger, P. W. Parren, D. E. Mosier, and D. R. Burton.** 1999. Neutralizing antibodies have limited effects on the control of established HIV-1 infection in vivo. Immunity **10:**431-8.
60. **Popovic, M., M. G. Sarngadharan, E. Read, and R. C. Gallo.** 1984. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science **224:**497-500.
61. **Price, D. A., P. J. Goulder, P. Klenerman, A. K. Sewell, P. J. Easterbrook, M. Troop, C. R. Bangham, and R. E. Phillips.** 1997. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. Proc Natl Acad Sci U S A **94:**1890-5.
62. **Reeves, J. D., and R. W. Doms.** 2002. Human immunodeficiency virus type 2. J Gen Virol **83:**1253-65.
63. **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.
64. **Rosenberg, E. S., M. Altfeld, S. H. Poon, M. N. Phillips, B. M. Wilkes, R. L. Eldridge, G. K. Robbins, R. T. D'Aquila, P. J. Goulder, and B. D. Walker.** 2000. Immune control of HIV-1 after early treatment of acute infection. Nature **407:**523-6.

65. **Sagar, M., L. Lavreys, J. M. Baeten, B. A. Richardson, K. Mandaliya, B. H. Chohan, J. K. Kreiss, and J. Overbaugh.** 2003. Infection with multiple human immunodeficiency virus type 1 variants is associated with faster disease progression. J Virol **77:**12921-6.
66. **Sala, M., and S. Wain-Hobson.** 2000. Are RNA viruses adapting or merely changing? J Mol Evol **51:**12-20.
67. **Sala, M., G. Zambruno, J. P. Vartanian, A. Marconi, U. Bertazzoni, and S. Wain-Hobson.** 1994. Spatial discontinuities in human immunodeficiency virus type 1 quasispecies derived from epidermal Langerhans cells of a patient with AIDS and evidence for double infection. J Virol **68:**5280-3.
68. **Schonning, K., B. Jansson, S. Olofsson, J. O. Nielsen, and J. S. Hansen.** 1996. Resistance to V3-directed neutralization caused by an N-linked oligosaccharide depends on the quaternary structure of the HIV-1 envelope oligomer. Virology **218:**134-40.
69. **Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins.** 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol **73:**10489-502.
70. **Simmonds, P., L. Q. Zhang, F. McOmish, P. Balfe, C. A. Ludlam, and A. J. Brown.** 1991. Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 env sequences in plasma viral and lymphocyte-associated proviral populations in vivo: implications for models of HIV pathogenesis. J Virol **65:**6266-76.
71. **Simon, F., P. Mauclere, P. Roques, I. Loussert-Ajaka, M. C. Muller-Trutwin, S. Saragosti, M. C. Georges-Courbot, F. Barre-Sinoussi, and F. Brun-Vezinet.** 1998. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat Med **4:**1032-7.
72. **Spang, R., M. Rehmsmeier, and J. Stoye.** 2002. A novel approach to remote homology detection: jumping alignments. J Comput Biol **9:**747-60.
73. **Strebel, K.** 2003. Virus-host interactions: role of HIV proteins Vif, Tat, and Rev. Aids **17 Suppl 4:**S25-34.
74. **Takehisa, J., L. Zekeng, E. Ido, Y. Yamaguchi-Kabata, I. Mboudjeka, Y. Harada, T. Miura, L. Kaptu, and M. Hayami.** 1999. Human immunodeficiency virus type 1 intergroup (M/O) recombination in cameroon. J Virol **73:**6810-20.
75. **Van Heuverswyn, F., Y. Li, C. Neel, E. Bailes, B. F. Keele, W. Liu, S. Loul, C. Butel, F. Liegeois, Y. Bienvenue, E. M. Ngolle, P. M. Sharp, G. M. Shaw, E. Delaporte, B. H. Hahn, and M. Peeters.** 2006. Human immunodeficiency viruses: SIV infection in wild gorillas. Nature **444:**164.
76. **van't Wout, A. B., N. A. Kootstra, G. A. Mulder-Kampinga, N. Albrecht-van Lent, H. J. Scherpbier, J. Veenstra, K. Boer, R. A. Coutinho, F. Miedema, and H. Schuitemaker.** 1994. Macrophage-tropic variants initiate human immunodeficiency virus type 1 infection after sexual, parenteral, and vertical transmission. J Clin Invest **94:**2060-7.
77. **Weaver, E. A., Z. Lu, Z. T. Camacho, F. Moukdar, H. X. Liao, B. J. Ma, M. Muldoon, J. Theiler, G. J. Nabel, N. L. Letvin, B. T. Korber, B. H. Hahn, B. F. Haynes, and F. Gao.** 2006. Cross-subtype T-cell immune

responses induced by a human immunodeficiency virus type 1 group m consensus env immunogen. J Virol **80:**6745-56.

78. **Wei, X., J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw.** 2003. Antibody neutralization and escape by HIV-1. Nature **422:**307-12.
79. **Weiss, R. A.** 2001. Gulliver's travels in HIVland. Nature **410:**963-7.
80. **Weiss, R. A.** 2003. HIV and AIDS: looking ahead. Nat Med **9:**887-91.
81. **Wolfs, T. F., G. Zwart, M. Bakker, and J. Goudsmit.** 1992. HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. Virology **189:**103-10.
82. **Wolinsky, S. M., B. T. Korber, A. U. Neumann, M. Daniels, K. J. Kunstman, A. J. Whetsell, M. R. Furtado, Y. Cao, D. D. Ho, and J. T. Safrit.** 1996. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science **272:**537-42.
83. **Wolinsky, S. M., C. M. Wike, B. T. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. L. Munoz.** 1992. Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. Science **255:**1134-7.
84. **Wyatt, R., P. Kwong, W. Hendrickson, and J. Sodroski.** 1998. Structure of the core of the HIV-1 gp120 Exterior Envelope Glycoprotein., p. III-3-9. *In* B. Korber, C. Kuiken, B. Foley, B. Hahn, F. McCutchan, J. Mellors, and J. Sodroski (ed.), Human Retroviruses and AIDS 1998. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.
85. **Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen.** 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:**431-49.
86. **Yu, Q., N. Landau, and R. König.** 2003. Vif and the Role of Antiviral Cytidine Deaminases in HIV-1 Replication, p. 2-13. *In* T. Leitner, B. Foley, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber (ed.), HIV Sequence Compendium 2003. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, LA-UR number 04-7420.
87. **Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Brown, and P. Simmonds.** 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J Virol **67:**3345-56.
88. **Zhu, T., B. T. Korber, A. J. Nahmias, E. Hooper, P. M. Sharp, and D. D. Ho.** 1998. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature **391:**594-7.
89. **Zhu, T., H. Mo, N. Wang, D. S. Nam, Y. Cao, R. A. Koup, and D. D. Ho.** 1993. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. Science **261:**1179-81.
90. **Zhu, T., N. Wang, A. Carr, D. S. Nam, R. Moor-Jankowski, D. A. Cooper, and D. D. Ho.** 1996. Genetic characterization of human immunodeficiency virus type 1 in blood and genital secretions: evidence for viral compartmentalization and selection during sexual transmission. J Virol **70:**3098-107.

91. **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.

GLOSSARY

antigen - the protein or part of a protein that is used to stimulate an immune response.

B cells - white blood cells of the immune system that produce infection-fighting proteins called antibodies.

CD4+ T cells - white blood cells that orchestrate the immune response, signaling other cells in the immune system to perform their special functions. Also known as **T helper cells**, these cells are killed or disabled during HIV infection.

CD8+ T cells - white blood cells that kill cells infected with HIV or other viruses, or transformed by cancer. These cells also secrete soluble molecules that may suppress HIV without killing infected cells directly. Also known as **cytotoxic T cells**.

cytokines - proteins used for communication by cells of the immune system. Central to the normal regulation of the immune response.

DC-SIGN - a C-type lectin receptor present on both macrophages and dendritic cells. Also known as CD209.

dendritic cells - immune system cells with long, tentacle-like branches. Some of these are specialized cells at the mucosa that may bind to HIV following sexual exposure and carry the virus from the site of infection to the lymph nodes. See also follicular dendritic cells.

epitope - the small part of a protein that an antibody or T cell recognizes

gp41 - glycoprotein 41, a protein embedded in the outer envelope of HIV. Plays a key role in HIV's infection of CD4+ T cells by facilitating the fusion of the viral and cell membranes.

gp120 - glycoprotein 120, a protein that protrudes from the surface of HIV and binds to CD4+ T cells.

gp160 - glycoprotein 160, an HIV precursor protein that is cleaved by the HIV protease enzyme into gp41 and gp120.

immune deficiency - the inability of the immune system to work properly, resulting in susceptibility to disease.

integrase - an HIV enzyme used by the virus to integrate its genetic material into the host cell's DNA.

lectin - carbohydrate-binding proteins or glycosprotein which are highly specific for their sugar moieties.

lentivirus - "slow" virus characterized by a long interval between infection and the onset of symptoms. HIV is a lentivirus as is the simian immunodeficiency virus (SIV), which infects nonhuman primates.

LTR - long terminal repeat, the RNA sequences repeated at both ends of HIV's genetic material. These regulatory switches may help control viral transcription.

macrophage - a large immune system cell that devours invading pathogens and other intruders. Stimulates other immune system cells by presenting them with small pieces of the invaders.

opportunistic infection - an illness caused by an organism that usually does not cause disease in a person with a normal immune system. People with advanced HIV infection suffer opportunistic infections of the lungs, brain, eyes, and other organs.

pathogenesis - the production or development of a disease. May be influenced by many factors, including the infecting microbe and the host's immune response.

pathogens - disease-causing organisms.

protease - an HIV enzyme used to cut large HIV proteins into smaller ones needed for the assembly of an infectious virus particle.

provirus - DNA of a virus, such as HIV, that has been integrated into the genes of a host cell.

rational vaccine design - make modified proteins designed to improve their ability to stimulate useful immune responses. These proteins are not found in nature.

retrovirus - HIV and other viruses that carry their genetic material in the form of RNA and that have the enzyme reverse transcriptase.

reverse transcriptase - the enzyme produced by HIV and other retroviruses that allows them to synthesize DNA from their RNA.

(Source: Most are from U.S. National Institute of Allergy and Infectious Diseases web. URL: http://www.niaid.nih.gov/)

MAJOR PAPERS

AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 21, Number 11, 2005, pp. 955–960

Sequence Note

HIV Type 1 CRF13_cpx Revisited: Identification of a New Sequence from Cameroon and Signal for Subsubtype J2

MING ZHANG,[1,2] KARIN WILBE,[3] NATHAN D. WOLFE,[4] BRIAN GASCHEN,[1] JEAN K. CARR,[5] and THOMAS LEITNER[1]

ABSTRACT

A nearly full-length genome sequence of an HIV-1 isolate originating from Cameroon, 02CM.3226MN, was found to cluster together with previously reported CRF13 sequences 96CM-4164 and 96CM-1849. Similarity plotting, bootscanning, breakpoint analysis, and phylogenetic trees confirmed similar genomic structures with almost identical breakpoint positions among these three isolates. Thus, CRF13 now fulfills the HIV-1 nomenclature requirements. A χ^2 analysis across all three genomes simultaneously was applied to more accurately determine breakpoints and address the uncertainty in such estimates. Some fragments were found to be difficult to classify, as indicated by a low branching index (BI), due to limited knowledge about parental and reference subtype sequences. One fragment with low BI association to reference subtype J sequences (BI = 0.27, cut-off for subtype classification >0.55) was found to be closer to J fragments of CRF11 similar to the way that A1–A2 and F1–F2 subsubtypes associate. This suggests that subtype J may need to be reclassified into subsubtypes J1 and J2. The CRF13 genome consists of fragments from subtypes A1, G, and both J1 and J2 as well as CRF01 and one region that was left unclassified.

INTERSUBTYPE RECOMBINATION IS AN IMPORTANT MECHANISM driving the rapid evolution of HIV-1 and it plays a major role in global and regional HIV epidemics. According to the Los Alamos HIV database, more than 20 circulating recombinant forms (CRFs) are currently recognized or have been proposed (http://hiv-web.lanl.gov/CRFs/CRFs.html), although not all sequence data are presently published and available.[1,2] The majority of the CRFs have been identified in geographic regions where multiple HIV-1 subtypes cocirculate. In Cameroon, where various HIV-1 subtypes of group M, O, and N coexist, CRF01_AE, CRF02_AG, CRF11_cpx, and CRF13_cpx have been found.[1,3–5] We have identified a third CRF13_cpx isolate (02CM.3226MN), and thus the general nomenclature requirements for defining CRFs have been met.

02CM.3226MN was sampled from a healthy, married female in Manyamen, Cameroon, who, in terms of transmission, had no high-risk behaviors. 02CM.3226MN displayed the same genetic organization as two previously reported CRF13 isolates (96CM-4164, GenBank accession number AF460974, and 96CM-1849, GenBank accession number AF460972).[6] The two previous CRF13_cpx sequences were reanalyzed together with 02CM.3226MN in all analyses described. The GenBank accession number for 02CM.3226MN is AY371154.

A neighbor-joining tree using F84 model distances was used to investigate overall sequence similarities between 02CM.3226MN and previously reported full-length nonrecombinant and recombinant genomes in the year 2001 HIV-1 full-length reference strains of M-group subtypes and CRFs (http://www.hiv.lanl.gov/content/hiv-db/SUBTYPE_REF/align.html). PHYLIP programs DNADIST and NEIGHBOR were used for tree reconstruction and the reliability of the constructed phylogenetic trees was assessed by nonparametric

[1]Theoretical Biology and Biophysics Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM, 87545.
[2]Department of Bioinformatics, Institute of Microbiology and Genetics, University of Göttingen, 37077 Göttingen, Germany.
[3]Department of Virology, Swedish Institute for Infectious Disease Control, SE-171 82 Solna, Sweden.
[4]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21218.
[5]U.S. Military HIV Research Program, Rockville, MD 20850.

bootstrap from 500 alignment replicates with the PHYLIP programs SEQBOOT, DNADIST, NEIGHBOR, and CONSENSE.[7] The obtained trees were visualized using TREEVIEW.[8] The previously described CRF13 sequences 96CM-1849, 96CM-4164, together with 02CM.3226MN formed a separate cluster with a 99.8% bootstrap value and did not associate with any specific subtype or other recombinant virus. This suggested that this new sequence, 02CM.3226MN, had unique sequence similarities shared with CRF13_cpx.

Tree analysis including recombinant sequences is, however, problematic. While it may display overall similarities, it does not reveal any details about the recombinatory nature of the sequences and it may also confuse the tree-building method. To overcome these problems, we reanalyzed all three sequences in the new CRF13 cluster using methods that specifically investigate HIV recombination, including similarity plotting and bootscanning, an expanded version of the breakpoint analysis by a χ^2 test of informative sites, the

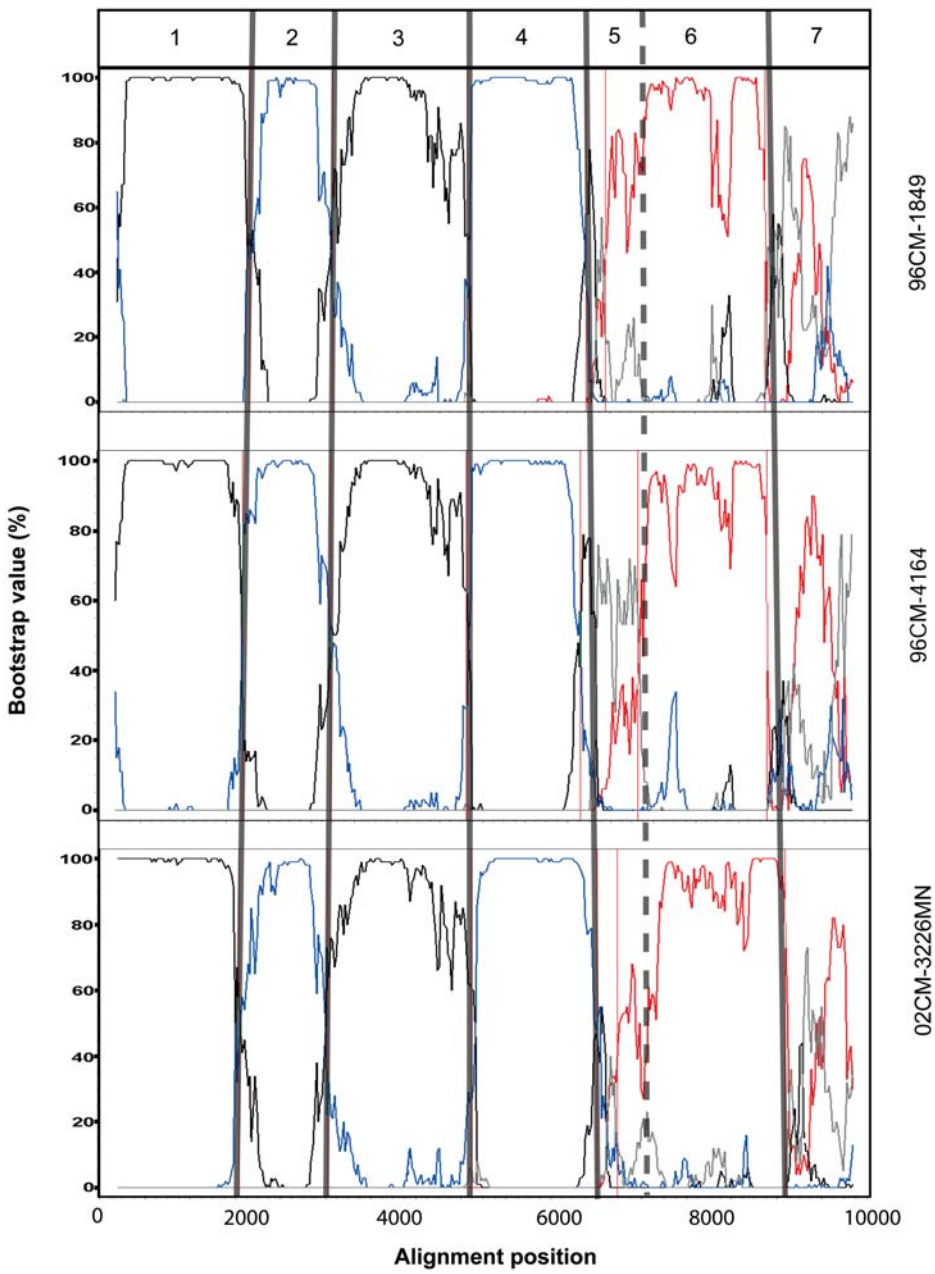


FIG. 1. Bootscanning results of 96CM-1849MN, 96CM-4164MN, and 02CM.3226MN genomes. Nucleotide positions of the alignment are shown on the *x*-axis and bootstrap values (*y*-axis) are plotted in the midpoint of each sliding window. The bootstrap values of 96CM-1849MN, 96CM-4164MN, and 02CM.3226MN clustering to subtype G (black), subtype J (blue), subtype A (red), and CRF01-AE (gray) are shown. The vertical lines indicate the breakpoints that divide the genome into the seven fragments (shown in numbers).

branching index, and separate bootstrapped trees for each subtype fragment.

Recombination was first investigated using SimPlot v3.5[9] (window-size: 400 bp, step-size: 20 bp, 500 replicates) with three different subtype reference sequence sets to avoid effects of bias due to the selection of reference sequences. The first two sets, with one sequence representing each subtype, were selected from the full-length alignment of the HIV sequence database[10] and analyzed independently. The first reference set included (GenBank accession numbers in parenthesis) A1.UG.92.92UG037 (ACC #: U51190), B-US.WR27 (U26546), C.ET.86.ETH2220 (U46016), D.CD.84.84ZR085 (U88822), F1.FR.96.MP411 (AJ249238), F2.CM.95.MP257 (AJ249237), G.FI.93.HH8793_12_1 (AF061641), H.BE.93. VI991 (AF190127), J.SE.94.SE7022 (AF082395), K.CD.97. EQTB11C (AJ249235), and CRF01_H93TH253 (U51189). The second set included AUGSE6594 (AF069672), B-FR.HXB2R (K03455), C.BW.96.96BW0502 (AF110967), D.CD.83.ELI (K03454), F1.BR.93.93BR020_1 (AF005494), F2.CM.95. MP255 (AJ249236), G.SE.93.SE6165 (AF061642), H.BE.93. VI997 (AF190128), J.SE.93.SE7887 (AF082394), K.CM.96. MP535 (AJ249239), and 01_AE.TH.90.CM240 (U54771). The third set included consensus sequences of subtype A–D, F1, F2, G, H, J, K, and CRF01_AE obtained from the HIV sequence database consensus archive year 2002 (http://www.hiv.lanl.gov/content/hiv-db/CONSENSUS/M_GROUP/Consensus.html).

The similarity plotting and bootscanning analyses revealed that 02CM.3226MN had a mosaic sequence pattern nearly identical to 96CM-1849 and 96CM-4164. On the basis of these preliminary analyses, subtypes A, G, J, and CRF01-AE were included in the subsequent bootscanning analyses in order to more precisely identify the recombination breakpoints of 02CM.3226MN. Figure 1 shows the resulting mosaic structure for all three CRF13 sequences from bootscanning using the consensus sequence as the reference set. Bootscanning analysis results from the two other reference sequence sets corroborated these results (results not shown).

Recombination breakpoints were determined based the bootscanning plots by identification of informative sites and maximization of the χ^2. As can be seen in Table 1, the breakpoints estimated from each of the three CRF13 sequences individually[11] showed some variation. This is not surprising as we would expect the evolution to continue along the CRF13 lineage since its formation, leading to divergence between sequences over time. Thus, a random error will be added and affect our ability to accurately reconstruct the original recombination breakpoints. To address this problem and to better estimate the true breakpoints, we applied a modified version of

Table 1. Individual Results from Sequence Fragments of CRF13 Isolates (96CM-1849MN, 96CM-4164MN, and 02CM.3226MN)

HIV-1 genome	*Genome fragment*	*Fragment (HXB2 position)*[a]	*Potential subtype*[b]	p *value for 3′ breakpoint*[c]	*Bootstrap value*[d]	*Branching index*
96CM-1849MN	1	623-2312	G	0.0018	100	1.00
	2	2313-3282	J	0.0359	97	0.27
	3	3283-4870	G	0.0096	100	0.86
	4	4871-6258	J	0.0002	92	0.12
	5	6259-6491	?[e]	0.0359	N/A[f]	N/A
	6	6492-8278	A	0.0032	100	1.00
	7	8279-9501	?[e]		N/A	N/A
96CM-4164MN	1	623-2247	G	0.0339	100	1.00
	2	2248-3282	J	0.0544	97	0.27
	3	3283-4870	G	0.0190	100	0.86
	4	4871-6183	J	0.0007	92	0.12
	5	6184-6833	?[e]	0.0226	N/A	N/A
	6	6834-8278	A	0.0013	100	1.00
	7	8279-9430	?[e]		N/A	N/A
02CM.3226MN	1	796-2303	G	0.0339	100	1.00
	2	2304-3282	J	0.1294	97	0.27
	3	3283-4870	G	0.0213	100	0.86
	4	4871-6258	J	0.0013	86	0.12
	5	6259-6483	?[e]	0.0652	N/A	N/A
	6	6484-8278	A	0.0006	100	1.00
	7	8279-9190	?[e]		N/A	N/A

[a]The breakpoints estimated from bootscanning analysis and maximization of chi-square values are shown as corresponding HXB2 positions.

[b]The closest subtype cluster to the analyzed fragment as suggested by maximum likelihood tree analysis.

[c]The probability of the 3′ breakpoint of the fragment determined by a chi-square test for heterogeneity of informative sites between adjacent fragments.

[d]Shown as percentages of 500 neighbor joining trees. For unclassified regions, the bootstrap values are not given.

[e]In these regions the investigated sequences are unclassified, clustering with multiple subtypes. Therefore, BIs are not given.

[f]N/A, not applicable.

the standard test that optimizes the breakpoints for all three sequences simultaneously (Fig. 2). Briefly, each phylogenetically informative site supports one of three possible phylogenetic associations among four taxa (the query sequence, two candidate subtypes, and one outgroup) and the χ^2 test is used to locate the maximal difference in the number of informative sites on either side of a putative breakpoint across all sequences simultaneously. Because most sites are uninformative and the informative sites are not evenly distributed, a step-like curve will describe the χ^2 function across sites (Fig. 2). The breakpoint point estimate was chosen as the midpoint of the region with the highest χ^2 value. A region of uncertainty was defined with boundaries given by the location of an informative site that lowered the χ^2 value from the maximum by at least 11 χ^2 units (which in standard statistics correspond to $p < 0.001$).

Each fragment, as suggested by the above breakpoint analysis, was further analyzed individually by maximum likelihood tree analysis for subtype classification (using DNAML in PHYLIP[7]) and by neighbor joining (using PHYLIP as above) for calculating bootstrap values and the branching index as described previously.[12] Year 2001 subtype reference sequences obtained from the HIV sequence database (http://www.hiv.lanl.gov/content/hiv-db/SUBTYPE_REF/align.html) were used in these analyses. A previously defined cutoff value of 0.55, above which subtype designation is supported, was used in this study.[12]

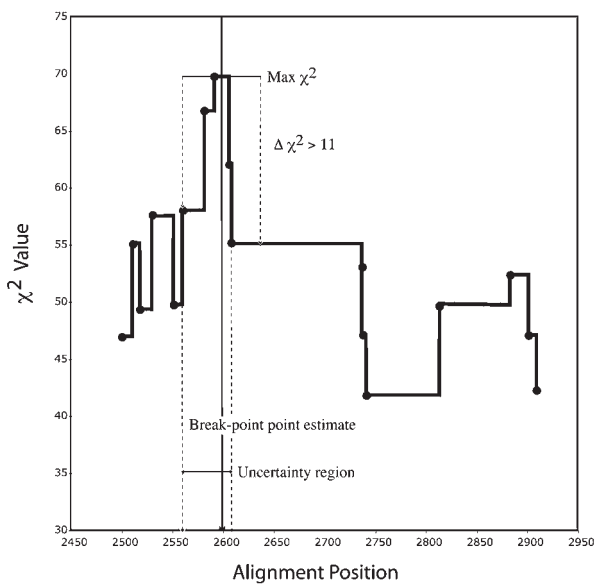


FIG. 2. χ^2 optimization curve and definitions of our breakpoint analysis. Dots on the line mark the positions of informative sites. When a putative breakpoint crosses one such site then the χ^2 value changes, resulting in a discrete step. The breakpoint point estimate is defined as the midpoint of the maximum region in which the χ^2 value does not change (Max χ^2). The region of uncertainty is defined as the range between the two informative sites that lower the Max χ^2 value by at least 11 units. The figure shows the 3′-breakpoint analysis of fragments 2 as an example. Note that the alignment positions are not in HXB2 coordinates at this stage, they are according to our specific alignment.

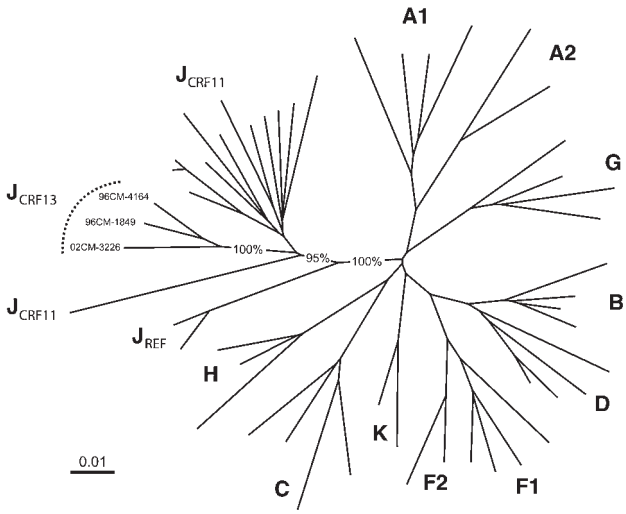


FIG. 3. Maximum likelihood tree of CRF13_cpx (J_{CRF13}), CRF11-cpx (J_{CRF11}), and subtype J reference (J_{REF}) sequences in fragment 2 of CRF13_cpx. The subtype reference sequences[10] and all available sequences from CRF11 and CRF13 were included in the analysis. For simplicity, the individual sequence names were removed and only subtype designations are shown. Bootstrap values are shown for relevant subtype J clades. Based on a BI analysis of subsubtypes A1/A2 and F1/F2 in this genomic region, we hypothesize that subtype J may also be subdivided into J1 (= J_{REF}) and J2 (= $J_{CRF11+CRF13}$). The scale bar indicates 0.01 substitutions/site according to an F84 substitution model.

The genomes of the three CRF13 sequences were divided into seven fragments according to the bootscanning and breakpoint analyses (Fig. 1 and Table 1). Fragment 1 covered most of *gag* and the beginning of *pol* and was classified as subtype G; fragment 2, which included most of *pol* protease, and about half of RT, was classified as subtype J; the other half of RT and most of *pol* integrase (p31) constituted fragment 3 and was identified as subtype G; fragment 4, comprising the last part of *pol* p31 and *vif*, *vpr*, *tat*, *rev*, and most of *vpu*, was classified as subtype J; fragment 5, which included the end of *vpu* and the beginning of *env* gp120, was not clearly associated with any defined subtype, but the majority of *env* that was included in fragment 6 consisted of subtype A1; and finally, fragment 7, which covered the last part of *env*, the second exons of *tat* and *rev*, and most of *nef*, showed a complex pattern with weaker associations to subtypes A, G, and CRF01-AE. In addition, neighbor-joining bootstrap and the branching index (BI) were applied to further test the classification of each fragment. In most fragments, as shown in Table 1, high bootstrap values were consistent with good BI values. For fragments 2 and 4, however, low BI values were observed despite the high bootstrap values. So, fragments 2 and 4, as well as the unclear associations of fragments 5 and 7, were subjected to further analyses.

Fragments 2 and 4 were classified as subtype J based on similarity plotting, bootscanning, and phylogenetic tree analyses. The low BI and the fact that only two J full-length sequences have been sequenced so far, however, raised the question of

TABLE 2. FINAL BREAKPOINTS AND SUBTYPE ASSIGNMENTS OF CRF13_cpx

Fragment	*Subtype*	*3′-Breakpoint*[a,b]	*χ^2 Range*[a,c]
1	G	2410	2311–2453
2	J2	3146	3115–3163
3	G	4801	4717–5096
4	J1	6169	6032–6283
5	U	6370[d]	N/A[e]
6	A1	8279	8215–8396
7	CRF01	N/A	N/A

[a]Positions are according to HXB2 coordinates.
[b]Position is midpoint estimate of maximum χ^2 region (Fig. 2).
[c]Range as defined by $\Delta\chi^2 > 11$ (Fig. 2).
[d]Breakpoint estimated by bootstrap support <70%; see text for details.
[e]N/A, not applicable.

whether these two J reference sequences were good representatives of subtype J. Thus, 96CM-1849MN, 96CM-4164MN, and 02CM.3226MN were compared with all possible J-containing CRFs (queried from Los Alamos HIV sequence database) in regions corresponding to fragments 2 and 4, respectively, again using similarity plotting, bootscanning, and phylogenetic tree reconstruction.

Among other described CRFs besides CRF13_cpx, only CRF11-cpx has a J part in the region of fragment 2 of CRF13_cpx. The maximum likelihood tree result demonstrated that fragments 2 from 02CM.3226MN, 96CM-1849MN, and 96CM-4164MN clustered with CRF11–cpx rather than the subtype J reference sequences (Fig. 3). This showed that the J fragments of CRF13 and CRF11–cpx are more closely related to each other than to the subtype J reference sequences in the region of fragment 2 of CRF13, which is also consistent with the previous result.[6] Thus, this suggests that fragment 2 of CRF13 and CRF11 may have originated from a common ancestor. An analysis of BIs in fragment 2 showed that J(ref) and J(CRF11 + CRF13) related to each other in a similar way as does A1 to A2 and F1 to F2 ($BI_{A1/A2} = 0.35$; $BI_{A2/A1} = 0.55$; $BI_{F1/F2} = 0.46$; $BI_{F2/F21} = 0.67$; $BI_{J(ref)/J(CRF)} = 0.35$; $BI_{J(CRF)/J(ref)} = 0.63$). Thus, although it is somewhat unclear which of the subsubtypes should be regarded as the "main" cluster, subtype J sequences in CRF11 and CRF13 seem to have derived genomic material from J clade representatives as diverse as subsubtypes are. This may indicate that a subsubtype J2 may exist or have existed when CRF11 and CRF13 were created.

A similar analysis was performed for fragment 4 of CRF13_cpx and other J-containing CRFs (CRF06 and CRF11) in the corresponding region. The J fragment in this region covers HXB2 positions 4096–6048 in CRF06 and 5058–5454 and 5848–6258 in CRF11. Hence, the J segments covered by CRF06, CRF11, and CRF13 within fragment 4 were 5058–5454 and 5848–6048. In contrast to fragment 2, in fragment 4 no closer association was detected to other J-containing CRFs than to the standard reference J sequences (data not shown).

Fragments 5 and 7 showed weak similarities with any of the reference sequences (Fig. 1). The phylogenetic analyses also did not reveal clear classification of these two fragments with any other subtypes[6] (data not shown). Our new sequence, 02CM.3226MN, was intensely investigated in fragment 5 because the span of this fragment was somewhat ambiguous among the three sequences (Fig. 1 and Table 1). Thus, fragment 5 was broken into two smaller parts based on an additional breakpoint according to the results from all three CRF13 sequences (Fig. 1). The first part (HXB2 positions 6183–6376) showed that all three CRF13 were more similar to CRF01 instead of to subtype G as shown in the bootscanning plot (Fig. 1). Awkwardly, however, CRF01 shared a most recent common ancestor with subtype G split by the CRF13 sequences. In the phylogenetic analysis of the second part (6377–6811), CRF13 isolates did not show close relation with CRF01, which was again inconsistent with the bootscanning results. Also unexpected in this tree was that subtype J clustered inside subtype G. All these results were supported by good bootstrap results. Considering the unsatisfactory tree results, most likely explained by the short lengths of these two parts within fragment 5, we left fragment 5 unclassified. To determine the breakpoints, i.e., the start and end of fragment 5, we could not use the above described χ^2 method, however. Instead, and to be conservative, the 3′-boundary of this region was determined by the right-most position among the three CRF13 sequences where the bootstrap fell below 70% (HXB2 position 6370). To justify this bootstrap method, the ranges of where the bootstrap fell below 70% in each of the three sequences were compared to the results of the overall χ^2 method at all other breakpoints, and reassuringly it was found to agree very well (data not

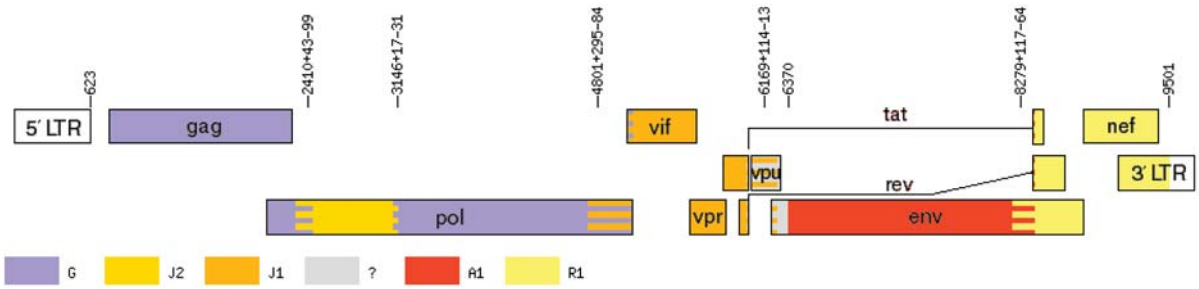


FIG. 4. Mosaic subtype organization of CRF13_cpx. Breakpoints are according to the HXB2 numbering system, and uncertainty regions are shown as interlaced fields. The colored fields correspond to subtype designations; R1 in the figure stands for CRF01. This map was generated using Recombinant HIV-1 Mapper, a new tool at the Los Alamos HIV database.

shown). Thus, this seemed to be a reasonable alternative to define the range of unclassified sequence regions. Finally, fragment 7 was reanalyzed and found most closely associated with CRF01 (data not shown). The low bootscanning curves were explained by the fact that fragment 7 contained subtype A parts that caused competition between subtype A and CRF01 references, as previously reported.[6]

The resulting breakpoints and subtype classifications after our reanalysis of CRF13 are shown in Table 2 and Fig. 4. The breakpoints are estimates of the parental CRF13 with regions of uncertainty. It is expected that when more full genomes become available, the breakpoints will be better estimated and the uncertainty region will become closer to a true confidence interval.

It has been reported that both 96CM-4164MN and 96CM-1849MN have an insertion of 10 amino acids (SRPEPTAPPA) in *gag* p6 (corresponding to amino acids 460/461 in HXB2), and this insertion is unique for these isolates.[6] Here we found that 02CM.3226MN had almost the same insertion of SRPEPTAPPV in this region. However, a number of subtype B and C sequences have similar inserts in this region, which seems to be due to a repetitive element. Thus, although all three CRF13 sequences have this insert, it is not as unique as was originally suggested.

In conclusion, the new isolate 02CM.3226MN was classified as the third member of the CRF13 family. The subtype fragment ranges were somewhat revised to the previously reported CRF13_cpx's sequences (Fig. 4). A modified χ^2 breakpoint analysis was applied to optimize breakpoints across all available sequences, improving the accuracy as well as adding a measure of certainty to the analysis of breakpoint locations. In fragment 2 high bootstrap values coexisted with low branching index values. Our explanation for this discrepancy is that the two standard reference J sequences may not be good representatives of this J-containing CRF. Instead, CRF13 and CRF11 were closer to each other in this fragment. Because no such relation was observed in fragment 4, however, the picture is more complicated. It is possible that CRF13 got material form both what is represented by the subtype J reference sequences (J1) and as of now undiscovered representatives of subsubtype J2. Indeed, the lack of more J reference sequences confounds the identification and analysis of mosaic genomes, in particular in areas where multiple sequence subtypes and CRFs cocirculate (like Cameroon from which all CRF13 were sampled).

ACKNOWLEDGMENTS

We thank Charles Calef for valuable discussions and contributions to this study. We thank the government of Cameroon for permission to undertake this study. This work was supported in part by an NIH-DOE interagency agreement (Y1-AI-1500-04), an award from the U.S. Military HIV Research Program (to Donald Burke), and an International Research Scientist Development Award grant from the National Institutes of Health Fogarty International Center (Grant 5 K01 TW000003-05) to N.W.

REFERENCES

1. Peeters M: Recombinant HIV sequences: Their role in the global epidemic. In: *Human Retroviruses and AIDS 2000: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences* (Kuiken C, Foley B, Hahn B, *et al*., eds.). Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, 2000, pp. 39–54.
2. Carr JK, Avila M, Gomez Carrillo M, *et al*.: Diverse BF recombinants have spread widely since the introduction of HIV-1 into South America. AIDS 2001;15:41–47.
3. Mboudjeka I, Zekeng L, Takehisa J, *et al*.: HIV type 1 genetic variability in the northern part of Cameroon. AIDS Res Hum Retroviruses 1999;15:951–956.
4. Montavon C, Vergne L, Bourgeois A, *et al*.: Identification of a new circulating recombinant form of HIV type 1, CRF11-cpx, involving subtypes A, G, J, and CRF01-AE, in Central Africa. AIDS Res Hum Retroviruses 2002;18:231–236.
5. Paraskevis D, Magiorkinis M, Paparizos V, Pavlakis GN, and Hatzakis A: Molecular characterization of a recombinant HIV type 1 isolate (A/G/E/?): Unidentified regions may be derived from parental subtype E sequences. AIDS Res Hum Retroviruses 2000;16:845–855.
6. Wilbe K, Casper C, Albert J, and Leitner T: Identification of two CRF11-cpx genomes and two preliminary representatives of a new circulating recombinant form (CRF13_cpx) of HIV type 1 in Cameroon. AIDS Res Hum Retroviruses 2002;18:849–856.
7. Felsenstein J: PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 1989;5:164–166.
8. Page RDM: TREEVIEW: An application to display phylogenetic trees on personal computers. Comput Appl Biosci 1996;12: 357–358.
9. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, and Ray SC: Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 1999;73:152–160.
10. Leitner T, Foley B, Hahn B, *et al*.: *HIV Sequence Compendium 2003*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM, 2004.
11. Robertson DL, Hahn BH, and Sharp PM: Recombination in AIDS viruses. J Mol Evol 1995;40:249–259.
12. Wilbe K, Salminen M, Laukkanen T, McCutchan F, Ray S, Albert J, and Leitner T: Characterization of novel recombinant HIV-1 genomes using the branching index. Virology 2003;316:116–125.

Address reprint requests to:
Thomas Leitner
MS K710, T-10
Los Alamos National Laboratory
Los Alamos, NM 87545

E-mail: tkl@lanl.gov

Evidence for old and new recombinants in different epidemiological settings

Ming Zhang et al.

Introduction

As one of the most diversified viruses, human immunodeficiency virus type 1 (HIV-1) possesses multiple ways to generate variants. Recombination, resulting from strand switching between two genetically different RNA copies co-packaged in one virion during reverse transcription (reviewed in (30)), is one important strategy employed by the virus to introduce large genetic alternations (17, 18, 49), and/or to repair genome damages (12, 47). Occurring at an estimated rate of at least 2.8 crossovers per genome per cycle (57), recombination between HIV-1 subtypes may result in antigenic shift and possible virulence changes at epidemic level (38, 39) or acts as an efficient approach for selecting variants resistant to HIV-1 specific drugs and immune pressure within a single host (20).

It has been estimated that at least 20% HIV-1 isolates sequenced worldwide are recombinants (32, 34, 35). Those recombinants can be classified into two categories, CRFs (circulating recombinant forms) and URFs (unique recombinant forms), referring to the recombinants that have established epidemics in a population, and to those only identified in one individual, respectively (40). To define a CRF, at least 3 full length sequences or 2 full length sequences plus one fragmental sequence that share the same mosaic genomic structure must be obtained from epidemiologically unlinked patients (40). The CRFs are numbered sequentially in the order in which they were first adequately described in the peer-reviewed literature. Currently at least 34 CRFs have been identified worldwide (http://www.hiv.lanl.gov/content/hiv-db/CRFs/CRFs.html) in the milieu of a complex picture of more than 100 URFs (www.hiv.lanl.gov web search page), and these numbers are still increasing as an

outcome of rapid HIV-1 evolution, large-scale full-length genome sequencing, and the availability of more advanced recombination detection techniques.

Previously we introduced a jumping profile hidden Markov model (jpHMM) method (43, 55) as one of the subtyping approaches we are currently using to automatically re-subtype more than 150,000 sequences in the Los Alamos HIV sequence database. In jpHMM, each HIV-1 subtype is defined by a profile hidden Markov model. All such subtype models are connected by empirical probabilities, allowing the detection of possible recombinants and related breakpoints. Informed from more than 300 well characterized near full-length sequences to build these subtype models, jpHMM has shown greatly improved accuracy in predicting recombinants, especially in locating recombination breakpoints, than other existing HIV-1 subtyping tools that are based on comparing a few representative subtype sequences. JpHMM performs particularly well in predicting recombinants involving subtypes A-D, F, and G because large sets of sequences are available from these subtypes, while H, J, K subtypes are less useful for these models as they have too few sequences to inform good jpHMM models. Therefore, jpHMM was used as one of the methods in the recombination study described in this paper that involves analysis of AG, BC, and BF recombinants mainly circulating in three continents.

CRF02_AG has caused at least 9 million infections worldwide (27). It is the most prevalent strain in West/West Central Africa, where multiple HIV-1 subtypes and CRFs co-circulate (http://www.hiv.lanl.gov geography site). Some studies have indicated that in this region, especially in Cameroon that has been suggested to be the place the zoonotic transmission took place (21, 24), CRF02 spreads more rapidly than either of the parental strains, subtypes A and G (14, 42). This may be due to, suggested by some studies (5, 23, 31), a founder effect reflecting a longer presence of CRF02 than A or G in West/West Central Africa, or alternatively because the recombination event provided an advantage of CRF02 over its parental subtypes. CRF02 has also been identified in other continents where most of those epidemics are self-contained transmission clusters (7, 9, 51).

BC recombinants are best characterized in CRF07_BC and CRF08_BC. CRF08_BC is a predominant subtype among intravenous drug users (IDUs) in Guangxi and east

Yunnan Province in Southeast China, while CRF07_BC has a low prevalence in Yunnan and is believed to have migrated to Xinjiang along the northern drug trafficking route (36, 54). Some studies have suggested that these two CRFs were generated in Yunnan, the epicenter of the AIDS epidemic in China where subtypes B and C were co-circulating in the early 1990s (4, 26, 44, 45). Currently, new BC recombinants, especially BC URFs that derived from secondary recombination of these two CRFs, are frequently arising in Yunnan, indicating that there is continuous inflow or co-circulation of B and C into Yunnan from its surrounding regions, for instance, from Myanmar where B is one of the most prevalent strains (37, 53). The emergence of BC URFs may also be due to superinfection of IDUs by CRF07 and CRF08 viruses (54). The uneven distribution of B, C, CRF07, CRF08 and BC URFs in Yunnan suggests the presence of more or less independent transmission networks and clusters among IDUs in Yunnan (54). BC recombinants in other Chinese regions may be different than those found in southwest China. CRF07 sampled from Taiwan may have derived from a recent single introduction of a new CRF07 strain or may merely be an expansion of a pre-existing, but not reported, virus strain from the mainland (10). While BC recombinants identified in Hong Kong were found to be epidemiologically unlinked to the mainland B, C or BC recombinants (25).

In contrast to AG and BC recombinants, where CRFs are more prevalent than URFs in the HIV-1 epidemics, BF recombinants are dominated by a large number of URFs. BF recombinants have been mostly sampled in South America, which is represented by a disperse distribution radiating from at least two genetic centers: one from CRF12_BF and similar genomes that are more frequently found in Argentina; and the other from CRF28_BF, CRF29_BF and some other BF recombinants that mostly circulate in Brazil (13). It appears as if none of these BF CRFs are vastly predominant but rather are outnumbered by BF URFs. This could be partially explained by either a relatively long period in which recombinant BF viruses appear to have been in circulation in South America (6), and/or a tight HIV-1 transmission network with high incidence rates found in some South American regions that would favor an elevated number of dual or super infections (50). Although the origin of the Argentinean BF recombinants is still unclear, it seems that at least one of the main introductory routes of HIV-1 into South America passes through Brazil (2).

Materials and Methods

Near full-length sequences (≥ 7000nt) of subtypes A, B, C, F, G, CRF02, CRF07, CRF08, CRF12, CRF17, CRF28, CRF29, and recombinants composed exclusively of AG, or BC, or BF were retrieved from the Los Alamos HIV sequence database (http://www.hiv.lanl.gov web search page). Fragmental sequences (300nt < length< 7000nt) of BC recombinants from Asia and BF recombinants from South America were also obtained from the same source. The details are listed in Table I "Database retrieved sequences" section. All near full-length sequences were aligned to the HIV-1 subtype reference sequences (http://www.hiv.lanl.gov/content/hiv-db/SUBTYPE_REF/align.html) using the GeneCutter tool (http://www.hiv.lanl.gov/content/hiv-db/GENE_CUTTER/cutter.html), followed by a manual quality check. We then used jpHMM analysis to investigate subtype recombination patterns in all near full-length and fragmental sequences. Based on the jpHMM results, the near full-length sequences were classified into different groups based on having similar genomic structures and breakpoints. Within the same recombinant set, for instance, within recombinants composed exclusively of subtypes B and C, the consensus subregions delimited by the breakpoints from all non-URF near full-length sequences were subjected to the phylogenetic analysis (neighbor-joining tree with F84 model distances) in PAUP* (48). Here, URFs identified by jpHMM were excluded in order to obtain long enough genome segments for tree analysis. The statistical robustness of the trees and the reliability of the clustering patterns were confirmed by bootstrap analysis (1000 replicates) in which a value of ≥ 70% (16) was considered significant for subtyping. For the AG recombinants, maximum likelihood trees were also built in PAUP* to confirm the sequence clusters. F84 distances between all AG near full-length sequences were also calculated in PAUP*.

Results

1. jpHMM results

(1) Defining sequence groups among near full-length sequences

All near full-length and fragmental sequences were evaluated using jpHMM. Possibly mis-subtyped sequences (Table I, "Possibly mis-subtyped sequences" section), either because their database assignment disagreed with the jpHMM prediction, or just because the sequence quality was too bad to do any further analysis (for instance, too many Ns in the nucleotide sequences), were excluded from all analyses mentioned below. The total number of jpHMM verified recombinants is shown in Table I "jpHMM verified recombinants" section.

The predicted breakpoint positions in all jpHMM verified recombinants were translated into HXB2 numbering (http://www.hiv.lanl.gov/content/hiv-db/REVIEWS/HXB2.html). This enabled us to compare sequences on the same scale. All sequences were then manually checked to see if they had similar genomic structures. If so, the most common breakpoint in a specific genomic region among these sequences was defined as the "breakpoint seed" in this region. Finally, those sequences bearing similar genome structures and breakpoints < 200nt off the breakpoint seeds were classified as one group. A value of 200nt was chosen as a cut-off in grouping breakpoints, because we've found that 200nt is the best window flanking a breakpoint seed to include more than 95% published breakpoints of AG, BC, BF CRFs that were identified by varied methods in the published literature, enabling us to retain as many of the currently defined CRF designations as possible. Also, with this value, we tended to be inclusive and group as many related sequences as possible, and to avoid assigning too many URFs (sequences with distinct breakpoints).

(2) Grouping AG recombinants

Near full-length recombinants between subtypes A and G were classified into 4 groups (each has >1 sequences) and 22 URFs (Fig.1). Group 1, represented by CRF02 prototype sequence IBNG, was confirmed to be a predominant AG circulating form. In addition, three Uzbekistan sequences were classified into two groups (group 1 and group 3, respectively). One of these three sequence bears an extra subtype G fragment (~1000nt). This appears to be inconsistent with a previous study (7) showing that those three Uzbekistan sequences were possibly associated with a point source introduction and the sequence distances among them were very low. This discrepancy became explainable after we examined the phylogenetic trees (see Discussion section).

(3) Grouping BC recombinants

Twenty near full-length BC recombinants, 17 out of which were from China, were classified into 2 groups and 7 URFs (Fig.2). We identified two more CRF08 sequences that were originally classified as BC URFs. Among the URFs, only one sequence was sampled outside of Asia. This Argentinean BC sequence has shown to be more likely generated locally rather than related to BC from Asia (41), explaining its distinct genomic structure. Four Chinese BC URFs were sampled from Ruili, the HIV epidemic center in Yunnan and the first city in China where an HIV-1 outbreak in IDUs was reported (48). The distinctive patterns found in the Ruili URFs, most display subtype C segments inserted in the B backbone that is different from CRF07 and CRF08 genomes, is partially due to its geographic proximity to Myanmar where subtype B is spreading.

(4) Grouping BF recombinants

All 56 BF recombinants were divided into 9 groups and 29 URFs (Fig.3). Compared with the original classifications described in various sources (summarized in http://www.hiv.lanl.gov/content/hiv-db/CRFs/CRFs.html), sequences included in CRF12, CRF18, and CRF29 groups have been revised in our study (groups 1-3 in Fig.3) but still demonstrate that Argentina and Brazil are two HIV-1 epidemic centers in South America. Group 4 and group 5, each of which has at least 3 epidemiologically unlinked patients, may be considered as new CRFs (after

confirmation in phylogentic analyses described below). CRF17, a CRF identified in other studies (6), needs more sequences to be confirmed. Currently only two near full-length sequences were found in this "CRF" and they are epidemiologically linked (6). Compared with groups 1-9, BF URF groups seem to involve fewer male sex with male transmissions. The revised CRF12 and CRF29 groups are not limited to any risk factors while all CRF28 sequences are associated with heterosexual transmission (Fig.3). However, we don't have enough sequences or appropriate sampling to analyze any risk factor correlations.

2. Phylogenetic analysis

(1) CRF02 group

There are nine consensus subregions shared by all sequences of groups 1-4 (Fig.1): fragment 1 (HXB2 positions 796-2173), fragment 2 (2237-2669), fragment 3 (2801-3187), fragment 4 (3241-4138), fragment 5 (4250-4771), fragment 6 (4993-5900), fragment 7 (6339-8174), fragment 8 (8383-8560), and fragment 9 (8763-9150). The neighbor-joining trees of these fragments are shown in figure 4. Maximum likelihood trees of these fragments yield the same tree topology shown in the NJ trees (thus data not shown).

In all trees, with the exception of the fragment 8 tree, all AG recombinants in groups 1-4 (Fig.1) cluster together, irrespective of their jpHMM-predicted subtypes in each tree region. Tree 8 has three AG recombinants outside of the AG sequence cluster, which we believe is due to the small genomic region studied in this fragment. Since almost all sequences in groups 1-4 are CRF02 sequences, their tree cluster was defined as a CRF02 cluster in our analysis. In all trees, the CRF02 clusters have long branches to their most recent common ancestor, and the genetic diversity within any single CRF02 clusters is at least not lower than the diversity within other pure subtypes. This is in accordance with Carr et al (8) who proposed that CRF02 was as old as pure subtypes. Trees 6 and 9 implicate that CRF02 emerged at roughly the same time as A1 and G. Trees 3 and 5 demonstrate a more recent introduction of subtype G fragments into these two regions. Taken together, the CRF02 cluster didn't

fall in the same topological position in all trees, suggesting that CRF02 has undergone multiple recombination events.

To examine if CRF02's "old origin" observation stands for other AG recombinants, we performed the phylogenetic analysis (NJ tree, F84 model) on all AG recombinants listed in figure 1, and all pure subtypes A and G sequences from the Los Alamos HIV sequence database. Figure 5 depicts the trees of the consensus subregions (delimited by jpHMM predicted breakpoints) in all 53 AG recombinant sequences. Trees 1-4 show that all database assigned CRF02 and several AG URFs always cluster together. However, we didn't observe a significant clustering of these CRF02 sequences in a concatenated consensus tree of fragments 1-4 (data not shown). Thus the CRF02 sequences may not have derived from the same ancestral virus. Trees 1-4 also demonstrate superclusters formed by the CRF02 cluster and its closely related subtype, A1 or G. In trees 1, 2 and 4, the supercluster includes the CRF02 cluster and subtype A1; in tree 3, the CRF02 cluster, subtypes A and G form the supercluster. Those long supercluster lengths indicate that all CRF02 sequences included in our analysis are old recombinants. This is in line with the aforementioned result based on a smaller sequence set (groups 1-4 sequences in figure 1). Trees 1-4 in figure 5 also suggest that the old origin of CRF02 may be related to Cameroonian subtypes A and G, as more than 50% of the CRF02 sequences were sampled from Cameroon.

(2) CRF07 and CRF08 groups

Nine partial genomic segments delimited by consensus breakpoints in CRF07 and CRF08 groups (Fig.6) were analyzed by neighbor-joining trees. These segments are fragment 1 (HXB2 positions 794-2064), fragment 2 (2064-2547), fragment 3 (2547-2846), fragment 4 (2979-3167), fragment 5 (3183-5836), fragment 6 (5851-6429), fragment 7 (6429-8840), fragment 8 (8866-9004) and fragment 9 (9059-9412). In all trees, except those for the short fragments 4 and 8, subtypes B and C form significant sub-clusters (> 94% in 1000 replicates). In all fragments, all CRF07/08's C segments are close to Indian C, while most Bs are closer to Thai B. Fragment 6 is the only region that shows CRF07's B segments are closest to US B, rather than to Thai B. This may reflect a possible entry of US subtype B to some Asian countries in the mid-

1980s (28, 33, 52), leading to its participation in the recombination event generating CRF07.

Of interest, when CRF07 and CRF08's fragments cluster within the same clade, CRF07 is located inside the CRF08 sub-cluster (Fig. 6, trees 1, 3, 5, and 7). Thus it is possible that CRF07 was derived from CRF08. The trees that don't show this close relationship of CRF07 and CRF08 are all short fragment trees: 4, 8 and 9. This may be due to short fragments giving unreliable trees (fragment 4, 8 and 9 are < 300nt), or it implicates that the currently identified CRF07 sequences are not the immediate descendents of CRF08. In the latter case, the currently identified CRF07 have accumulated enough nucleotide changes along the distance of > 1000km when CRF07 spreads from Yunnan (possible origin place) to Xinjiang, leading to a separate sub-cluster from CRF08 in trees 4, 8 and 9.

(3) CRF12, CRF28, and CRF29 groups

The following genomic segments are the consensus subregions delimited by breakpoints in nine BF groups (groups 1-9 in Fig.3): fragment 1 (HXB2 positions 960-1227), fragment 2 (1398-2463), fragment 3 (2565-2973), fragment 4 (2988-3678), fragment 5 (3927-4864), fragment 6 (6458-8450), and fragment 7 (9120-9409). Fragments were analyzed in NJ trees shown in figure 7. In the long fragments (fragments 2, and 4-6), the jphmm-predicted B and F segments cluster into their respective B or F clade with high bootstrap values (> 90% in 1000 replicates). In trees 1 and 7, jphmm-predicted B or F regions are not supported by strong bootstrap values due to short fragments. All trees but fragment 3 are consistent with jpHMM predicted subtypes. In fragment 3, some jpHMM-predicted B segments in groups 4-9 sequences either cluster within F clade or locate between B and F clades, but all CRF sequences (CRF12, CRF28, and CRF29 groups in Fig.3) were correctly clustered with their respective clades. The difference between jpHMM and tree result in this fragment may reflect the limitations in tree analysis in dealing with short and similar sequences. On the other hand, it indicates that fragment 3 covers uncertainty region of the jpHMM-predicted breakpoints in groups 4-9 sequences. Thus fragment 3 was assigned as "undetermined region" in the sequences of groups 4-9.

Since all CRF groups (CRF12, CRF28, and CRF29 groups) were the only groups having confirmed subtypes (results very consistent between jpHMM and tree analysis) in every genomic region we studied, they were subjected to further breakpoint frequency analysis described in "Defining breakpoint frequency" section.

3. Defining the CRF02 family

CRF02's old origin poses a difficult question of how to classify a queried AG recombinant sequence as a CRF02 sequence, with the absence of the CRF02's original parental A and G sequences. Two solutions are possible: either to compare the query sequence with some CRF02 reference sequences, or to compare it with a CRF02 consensus. The problem of the first solution is, how to select an appropriate reference? And the problem of the second solution is, a consensus doesn't represent every bit of information at every sequence position. In case of CRF02 that may have undergone many point mutations and multiple recombination events, a consensus comparison may lead to poorly resolved subtype assignments, in particular confused with A or G, considering the close relationship of A, G, and CRF02. Here we introduced an alternative method besides the tree analysis to classify the CRF02 family sequences.

As shown in figure 8 A, the problem of deciding if a queried sequence belongs to CRF02 or not becomes a question of how to define the radius of the CRF02 family circle. The center of this family is a sequence that has the least distance (F84 distance in our study) to all other well-defined CRF02 sequences, and the IBNG strain (GenBank accession number: L39106) was found to be this central sequence. Three distances are calculated in defining the radius of the CRF02 family circle: the query sequence's averaged distance to all subtype A reference sequences (Aref), to all subtype G reference sequences (Gref), and to the CRF02 central sequence IBNG. All CRF02 family sequences should have shorter distance to IBNG than to A, or to G reference sequences. Thus a sequence defining the radius of the CRF02 family circle is the last sequence that is closer to IBNG than to A or to G reference sequences. Fig 8 B depicts a distance plot in defining the CRF02 family. The CRF02 sequences defined by this family plotting method are the same sequences included in the CRF02

tree clusters. Further, the radius of the CRF02 family circle can be used as a quick reference to check if any queried sequences belong to CRF02 or not.

Based on the phylogenetic analysis and the family plotting method, we confirmed that all previously database assigned CRF02 near full-length sequences belong to the CRF02 family. In addition, three new CRF02 sequences were identified. These sequences are (accession numbers listed here): AB052867, AF184155, and AY371147.

4. Tracking BC and BF recombination breakpoint frequency

In contrast to CRF02, the BC and BF recombinants are of recent origin and have diversified less (Fig. 4, 6, and 7. All trees are drawn on the same scale). Breakpoint predictions in the BC and BF recombinants, therefore, are more accurate than those in the CRF02 sequences. This allows us to look into the following questions. First, which CRF came first in the HIV-1 epidemic, CRF07 or CRF08? Second, to what degree the Brazilian BF and Argentinean BF are related or unrelated? To address these questions, we calculated the breakpoint frequency in two sets of sequences: (1) all BC and BF near full-length sequences; (2) fragmental sequences (300nt< length<7000nt) of BC recombinants from Asia and BF recombinants from South America.

(1) Defining the certainty region of breakpoints

As described in the "Defining sequence groups among near full-length sequences" section, a breakpoint seed ± 200 nt was used as a breakpoint window to group recombinant sequences. Here a breakpoint seed is the median breakpoint of its breakpoint window. Figure 9 depicts the distribution of individual jpHMM-predicted breakpoints off their breakpoint seeds in each recombinant group. A distance of 16nt was found to include > 95% breakpoints in the BC CRF groups, and 98nt in BF CRF groups. We define these two values, 16nt and 98nt, as the breakpoint certainty window (with 95% confidence interval) in BC and BF CRFs, respectively. Biologically, the size of the breakpoint certainty window is impacted by how diverse the region is the breakpoint is embedded in and how long the sequences have had to

evolve. This window, together with its flanking subtypes, determines if a query sequence shares breakpoints with the CRF group.

(2) Breakpoint frequency in BC CRFs

All jpHMM-identified breakpoints in CRF07 and CRF08 are detailed in figure 10. The breakpoint number described below refers to those listed in the figure.

Along the complete genome, the most shared breakpoints in CRF07, breakpoint #4 and #7, are also very often seen in CRF08 (#2 and #3 in CRF08). Two subtype B fragments, HXB2 2979-3167 and 8866-9004, are shared in CRF07 and CRF08. The former fragment is inside *p51* and contains two important catalytically important residues Asp^{185} and Asp^{186} (19, 22), and the latter locates at the beginning of *nef*.

All breakpoints seen in CRF07 and CRF08 groups, with the exception of CRF07 breakpoint #5 (highlighted in red in Fig.10 CRF07 group), were only observed in Chinese BC recombinants (in near full-length and fragmental sequences). CRF07 breakpoint #5 were found in two Myanmar fragmental sequences, indicating a different introduction of this breakpoint compared with other breakpoints in CRF07 and CRF08. This is consistent with our tree analysis: the fragment delimited by CRF07 breakpoint #5 and #6 is closer to US B while all other B fragments of CRF07 and CR08 are closer to Thailand B (tree 6 in Fig.6).

Within the CRF07 group, breakpoint #1 and #2 are less shared than breakpoint #3 and #4 in both the fragmental sequences and near full-length sequences. Among other possibilities, this could attribute to a later introduction of the genome region, flanked by breakpoint #1 and #2, than the region delimited by #3 and #4 in the complete genome. In other words, CRF07's B fragments were not introduced into the genome at the same time. Again, this is in line with our tree results suggesting that CRF07 was derived from further recombination(s) between CRF08 and subtype B.

We also identified three BF recombinant sequences that were not sampled from southwest China, where CRF07 and CRF08 were mostly found. These sequences (AY548266, AY548265, and AY548275) were all sampled from Liaoning, a province

in northeast China. All three sequences share CRF07's breakpoint #2 and #3, but they are not long enough (all < 900nt) to see if they belong to CRF07. Lastly, one of these three sequences, AY548275, was mis-classified as a pure B subtype in the original GenBank submission.

(3) Breakpoint frequency in BF CRFs

Though HIV-1 epidemic in Argentina is represented by CRF12, and in Brazil by CRF28 and CRF29, it seems there is no single BF breakpoint limited to one country. Almost all breakpoints from one country's CRF were also identified in the near full-length or fragmental BF sequences from the other country (Fig. 11). This possibly reflects a mingled BF epidemic that had taken place between Argentina and Brazil, considering these two countries' geographic proximity. The fragmental BF sequences very likely bear the information that fills the gap between two extremes of BF CRFs represented in two countries. Furthermore, frequent cases of multiple infections are suggested by the high frequency of URFs identified in this region.

Noticeably, both of CRF28's breakpoints exist in CRF29. However their interspersed locations in all trees didn't reveal a close relationship, as we observed in CRF07 and CRF08. Thus it is unlikely that CRF28 and CRF29 are of relation of parental and progeny sequences. But it can't rule out the possibility that CRF28 and CRF29 share one same ancestor, or breakpoints seen in CRF28 and CRF29 are two breakpoint hotspots in the BF population.

Finally, comparing the breakpoint frequency in BC and BF CRF groups (Fig.12), BC CRF groups seldom have breakpoints in *gag* p17, p24, and pol p51, p15. But both BC and BF groups have few breakpoints in *env*, especially in gp120 region, so that *env* tends to harbor distinct subtypes in its external and internal portions.

Discussion

The study described in this paper presents the first large-scale sequence re-subtyping result. Our purpose of re-subtyping the sequences in the Los Alamos sequence database, on one hand, is to provide the HIV research community with more accurate and systematically consistent sequence information, and on the other hand, to inform us about the driving forces of HIV epidemics in different epidemiological settings.

We confirmed that CRF02 is an old recombinant as proposed by Carr et al (8). CRF02 clusters distinctly from A and G clades in most studied genome regions, and the genetic distance within the CRF02 cluster is approximately the same as those within pure subtypes. We deduced that CRF02 probably was formed immediately after subtype A and G split to establish their individual lineages. After that, CRF02 may have undergone multiple additional recombination events, as the CRF02 cluster didn't fall in the same topological position in all studied trees. The latest recombination events may be inferred from two trees (tree 3 and 5 in Fig.4) in which the CRF02 cluster is associated with contemporary G subtype sequences. Of particular interest, the sequence fragments for these two trees were reported before (15, 29) that they bear significant similarity with Z321, a complex recombinant strain obtained from a 1976 Zairean (now the Democratic Republic of the Congo) serum sample (11, 15, 29). Thus it is possible that CRF02 was firstly formed before 1976, and till today it is still actively circulating in West and West Central Africa due to a fitter status (23) in the long-term competition against other HIV strains in this region.

Our CRF02 analysis also includes subtypes A and G all near full-length sequences from the database. Though the parental A sequences forming CRF02 hasn't come to light, it is not likely that a subtype A sequence containing the information of the parental subtype A is still circulating, as within subtype evolution has had several decades of evolution since it helped form the initial recombinant.

In contrast to CRF02's long existence in the HIV epidemic, CRF07 and CRF08 only joined the HIV epidemic very recently. CRF07 may have derived from CRF08, and its generation involves at least two recombination events between CRF08 and subtype

B. This is because two extra subtype B fragments in CRF07, compared with CRF08, are closer to different subtype B strains (one is closer to Thailand B and the other closer to US B), as we observed this in the BC tree and breakpoint frequency results. After CRF07 was generated, possibly in Yunnan where multiple HIV-1 strains are co-circulating, CRF07 underwent minor nucleotide changes along the way as it migrated to Xinjiang, more than 1000km away from Yunnan. Given HIV's rapid evolutionary rate, the near full-length CRF07 in the sequence database would not be exactly the same as those CRF07 generated in Yunnan. This is clear in the data, given the observation that in short fragment trees 4, 8, and 9 (Fig. 6) when CRF07 and CRF08 cluster within one clade, CRF07 doesn't locate inside CRF08 sub-cluster as we observed in long fragment trees 1, 3, 5, and 7.

Compared with CRF02, CRF07, and CRF08, the most distinctive feature of BF recombinants in South America is the dominance of BF URFs. When all near full-length BF recombinants from this continent were re-subtyped using jpHMM, and later confirmed by phylogenetic analysis, sequences included in CRF12, CRF28 and CRF29 groups were revised, and more BF URFs were identified (Fig.3). Noticeably, V62 (accession number: AY536236), a strain was assigned as a URF before CRF28 and CRF29 were identified, bears the same genomic structure and breakpoints as CRF28 (Fig.3). This V62 sequence was found to have epidemical links with Argentina (46), while the other two CRF28 sequences were both linked to Brazilian HIV epidemic. Thus it is very likely that the epidemics in Argentina and Brazil HIV-1 are not independent. Further evidence came to light from our breakpoint frequency analysis. First, we found that CRF12 and CRF29 do share a breakpoint in the middle of *pol* (Fig.11). Second, the analysis of all near full-length and fragmental BF recombinants in South America shows that no single breakpoint of any BF CRFs is limitedly to one country's BF recombinants (Fig.11). In other words, the Argentinean BF sequences, to some degree, share breakpoints with Brazilian sequences. The sharing of breakpoints among BF recombinants, a relatively long circulation record of BF in South America (6), and tight HIV-1 transmission networks may all contribute to a fluid HIV epidemic in South America.

Another question regarding BF recombinants in South America is whether Argentinean BF was originated from Brazil. In our analysis, we didn't find clear

evidence showing Argentina BF was originated from Brazil. First, CRF12, CRF28, and CRF29 are interspersed among other BF recombinants in all the phylogenetic trees we examined. Second, a near full-length Argentinean pure F strain, ARE933, was reported to be closer to Argentinean BF than other F strains (2). Thus it cannot be ruled out that Argentinean BF recombinants were formed locally in Argentina. If this did happen, then the shared breakpoints among Argentinean and Brazilian BF recombinants may be indicative of breakpoint hotspots.

CRF07, CRF08, CRF12, CRF28 and CRF29 are all B-containing recombinants. Currently we are not clear what's the exact biological consequences of having subtype B in these recombinants. But it was reported that subtype C has less replicative capacity than most other M group subtypes (1, 3), thus the inclusion of subtype B may enable the BC recombinant viruses to replicate more efficiently.

HIV-1 recombination is a dynamic and complex process that has blurred the current HIV subtype phylogeny. The recombinants can be ancient (e.g. CRF02), or be recent (e.g. CRF07, CRF08). Most sequences available today are contemporary (Fig.13), and we use them to attempt to reconstruct the past by extrapolating backward from small sets of surviving clues. A good example is CRF02 detailed in this study. Another problem of the currently identified sequences is that they may not represent a full spectrum of the contemporary virus diversity. For instance, we've found that the parental J sequences of CRF13's J segments are missing (56). Thus the incomplete sequence information used in the HIV-1 subtyping may lead to an underestimate of the HIV-1 genetic space. In this regard, it may be more appropriate to define recombinants using recombinant families rather than CRFs. The biggest difference between these two is that, the CRF definition is more focused on having the exact or very similar breakpoints among all members within a CRF, and thus be the consequence of a single lineage from an initial recombinant form. However, exact or very similar breakpoints may be easily blurred by rapid evolution of HIV-1, or by being embedded in a conserved region where precise breakpoints are difficult to resolve. It is likely that the sequences defined in a CRF are snapshots, taken at a given time, of a dynamic HIV epidemic picture. Therefore, we suggest to use a "recombinant family" definition to reflect the dynamic feature of the HIV-1 epidemic.

Taken together, we've found evidence for old and new recombinants in different epidemiological settings based on more accurate recombinants information. With more reliable sequence information available from our database re-subtyping work, we will recover more lost history of the HIV-1.

References

1. **Arien, K. K., A. Abraha, M. E. Quinones-Mateu, L. Kestens, G. Vanham, and E. J. Arts.** 2005. The replicative fitness of primary human immunodeficiency virus type 1 (HIV-1) group M, HIV-1 group O, and HIV-2 isolates. J Virol **79:**8979-90.
2. **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.
3. **Ball, S. C., A. Abraha, K. R. Collins, A. J. Marozsan, H. Baird, M. E. Quinones-Mateu, A. Penn-Nicholson, M. Murray, N. Richard, M. Lobritz, P. A. Zimmerman, T. Kawamura, A. Blauvelt, and E. J. Arts.** 2003. Comparing the ex vivo fitness of CCR5-tropic human immunodeficiency virus type 1 isolates of subtypes B and C. J Virol **77:**1021-38.
4. **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.
5. **Burda, S. T., F. A. Konings, C. A. Williams, C. Anyangwe, and P. N. Nyambi.** 2004. HIV-1 CRF09_cpx circulates in the North West Province of Cameroon where CRF02_AG infections predominate and recombinant strains are common. AIDS Res Hum Retroviruses **20:**1358-63.
6. **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.
7. **Carr, J. K., Y. Nadai, L. Eyzaguirre, M. D. Saad, M. M. Khakimov, S. K. Yakubov, D. L. Birx, R. R. Graham, N. D. Wolfe, K. C. Earhart, and J. L. Sanchez.** 2005. Outbreak of a West African recombinant of HIV-1 in Tashkent, Uzbekistan. J Acquir Immune Defic Syndr **39:**570-5.
8. **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.
9. **Carrion, G., J. Hierholzer, S. Montano, A. Alava, J. Perez, A. Guevara, V. Laguna-Torres, C. Mosquera, K. Russell, G. Chauca, T. Kochel, D. L. Birx, J. L. Sanchez, and J. K. Carr.** 2003. Circulating recombinant form CRF02_AG in South America. AIDS Res Hum Retroviruses **19:**329-32.
10. **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.
11. **Choi, D. J., S. Dube, T. P. Spicer, H. B. Slade, F. C. Jensen, and B. J. Poiesz.** 1997. HIV type 1 isolate Z321, the strain used to make a therapeutic HIV type 1 immunogen, is intersubtype recombinant. AIDS Res Hum Retroviruses **13:**357-61.

12. **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.
13. **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.
14. **Fischetti, L., O. Opare-Sem, D. Candotti, H. Lee, and J. P. Allain.** 2004. Higher viral load may explain the dominance of CRF02_AG in the molecular epidemiology of HIV in Ghana. Aids **18:**1208-10.
15. **Getchell, J. P., D. R. Hicks, A. Svinivasan, J. L. Heath, D. A. York, M. Malonga, D. N. Forthal, J. M. Mann, and J. B. McCormick.** 1987. Human immunodeficiency virus isolated from a serum sample collected in 1976 in Central Africa. J Infect Dis **156:**833-7.
16. **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.
17. **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.
18. **Hu, W. S., and H. M. Temin.** 1990. Retroviral recombination and reverse transcription. Science **250:**1227-33.
19. **Jacobo-Molina, A., Ding, J., Nanni, R.G., Clark, A.D., Jr., Lu, X., Tantillo, C., Williams, R.L., Kamer, G., Ferris, A.L., Clark, P., Hizi, A., Hughes, S.H., and Arnold, E. .** 1993. Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 Å resolution shows bent DNA. Proc. Natl. Acad. Sci. USA **90:**6320-6324.
20. **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.
21. **Keele, B. F., F. Van Heuverswyn, Y. Li, E. Bailes, J. Takehisa, M. L. Santiago, F. Bibollet-Ruche, Y. Chen, L. V. Wain, F. Liegeois, S. Loul, E. M. Ngole, Y. Bienvenue, E. Delaporte, J. F. Brookfield, P. M. Sharp, G. M. Shaw, M. Peeters, and B. H. Hahn.** 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science **313:**523-6.
22. **Kohlstaedt, L. A., J. Wang, J. M. Friedman, P. A. Rice, and T. A. Steitz.** 1992. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science **256:**1783-90.
23. **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.
24. **Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya.** 2000. Timing the ancestor of the HIV-1 pandemic strains. Science **288:**1789-96.
25. **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.
26. **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.

27. **McCutchan, F. E.** 2000. Understanding the genetic diversity of HIV-1. Aids **14 Suppl 3:**S31-44.
28. **McCutchan, F. E., P. A. Hegerich, T. P. Brennan, P. Phanuphak, P. Singharaj, A. Jugsudee, P. W. Berman, A. M. Gray, A. K. Fowler, and D. S. Burke.** 1992. Genetic variants of HIV-1 in Thailand. AIDS Res Hum Retroviruses **8:**1887-95.
29. **McCutchan, F. E., J. L. Sankale, S. M'Boup, B. Kim, S. Tovanabutra, D. J. Hamel, S. K. Brodine, P. J. Kanki, and D. L. Birx.** 2004. HIV type 1 circulating recombinant form CRF09_cpx from west Africa combines subtypes A, F, G, and may share ancestors with CRF02_AG and Z321. AIDS Res Hum Retroviruses **20:**819-26.
30. **Negroni, M., and H. Buc.** 2001. Mechanisms of retroviral recombination. Annu Rev Genet **35:**275-302.
31. **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.
32. **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.
33. **Ou, C. Y., Y. Takebe, B. G. Weniger, C. C. Luo, M. L. Kalish, W. Auwanit, S. Yamazaki, H. D. Gayle, N. L. Young, and G. Schochetman.** 1993. Independent introduction of two major HIV-1 genotypes into distinct high-risk populations in Thailand. Lancet **341:**1171-4.
34. **Peeters, M., and P. M. Sharp.** 2000. Genetic diversity of HIV-1: the moving target. Aids **14 Suppl 3:**S129-40.
35. **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.
36. **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.
37. **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.
38. **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.
39. **Quinones-Mateu ME, A. E.** 1999. Recombination in HIV-1: Update and implications. AIDS Reviews **1:**89-100.
40. **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.

41. **Sanabani, S., W. K. Neto, D. J. de Sa Filho, R. S. Diaz, P. Munerato, L. M. Janini, and E. C. Sabino.** 2006. Full-length genome analysis of human immunodeficiency virus type 1 subtype C in Brazil. AIDS Res Hum Retroviruses **22:**171-6.
42. **Sarr, A. D., G. Eisen, A. Gueye-Ndiaye, C. Mullins, I. Traore, M. C. Dia, J. L. Sankale, D. Faye, S. Mboup, and P. Kanki.** 2005. Viral dynamics of primary HIV-1 infection in Senegal, West Africa. J Infect Dis **191:**1460-7.
43. **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.
44. **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.
45. **Shao Y, Z. Y., Chen Z, et al.** 1991. Isolation of viruses from HIV infected individuals in Yunnan. Chin J Epidemiol **12:**129.
46. **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.
47. **Stahl, F. W.** 1987. Genetic recombination. Sci Am **256:**90-101.
48. **Swofford, D.** 1991. PAUP: phylogenetic analysis using parsimony, version 3.1.
49. **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.
50. **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.
51. **Ustina, V., K. Zilmer, L. Tammai, M. Raukas, A. Andersson, E. Lilja, and J. Albert.** 2001. Epidemiology of HIV in Estonia. AIDS Res Hum Retroviruses **17:**81-5.
52. **Weniger BG, L. K., Ungchusak K, et al.** 1991. The epidemiology of HIV infection and AIDS in Thailand. AIDS **5:**S71–85.
53. **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.
54. **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.

55. **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.
56. **Zhang, M., K. Wilbe, N. D. Wolfe, B. Gaschen, J. K. Carr, and T. Leitner.** 2005. HIV type 1 CRF13_cpx revisited: identification of a new sequence from Cameroon and signal for subsubtype J2. AIDS Res Hum Retroviruses **21:**955-60.
57. **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.

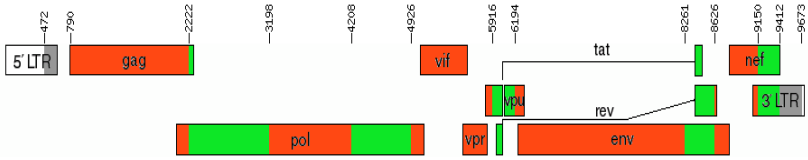
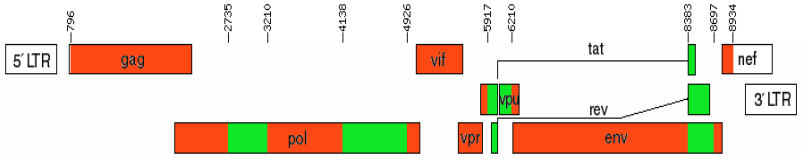
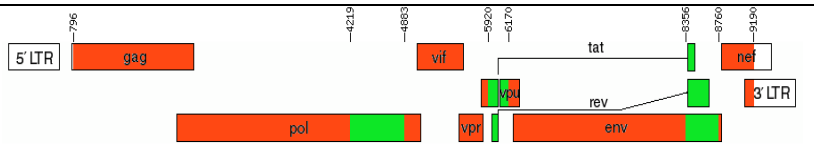
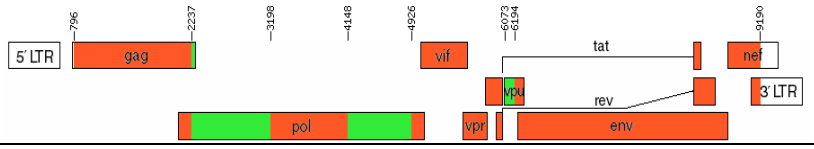
Table I

	AG set				BC set									BF set												
	Full length (world) N=140				Full length (world) N=509					Fragments (Asia) N=4413					Full length (world) N=240							Fragments (South America) N=4153						
Retrieved DB sequences	A	G	02	A G	B	C	07	08	B C	B	C	07	08	B C	B	F	12	17	28	29	B F	B	F	12	17	28	29	BF
	72	12	48	8	152	334	7	4	12	3133	1048	17	171	44	152	12	11	2	3	4	56	3070	242	261	0	0	0	580
Possibly mis-subtyped sequences	2 problematic 02 sequences; plus 1 pure G was mis-subtyped as CRF02				18 problematic sequences; plus 1 BF was mis-subtyped as pure B					~5.3% mis-subtyped					At least 6 CRF sequences appear to be URFs							~6.7% mis-subtyped						
jpHMM verified recombinants	53				20					103					56							751						

Re-subtyping results of the AG, BC and BF recombinants.

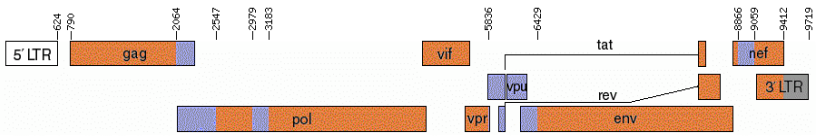
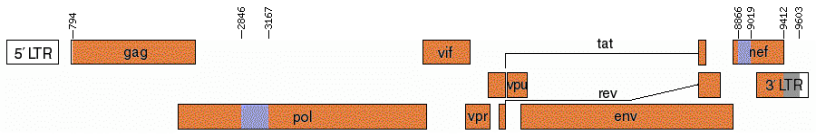
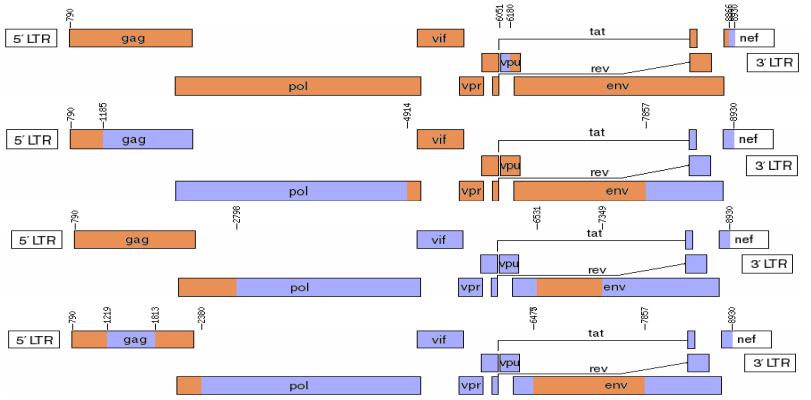
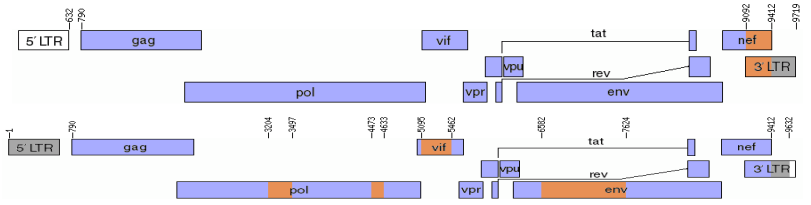
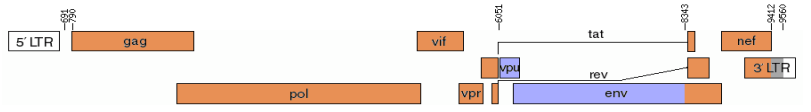
All sequences were retrieved from the Los Alamos Sequence Database (http://www.hiv.lanl.gov). The downloaded sequences were broken down into different categories, based on their database-assigned subtypes. Only those recombinants verified by jpHMM were preceded to other analyses (phylogenetic analysis, breakpoint analysis, etc.).

Figure 1

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** 			

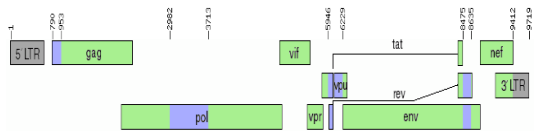
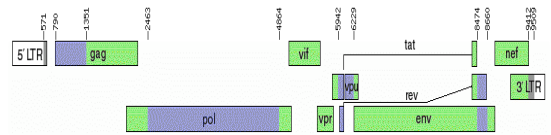
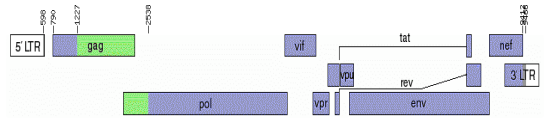
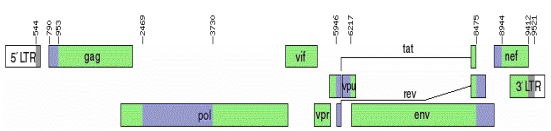
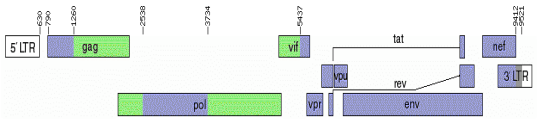
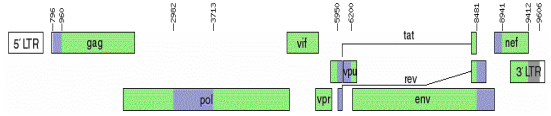
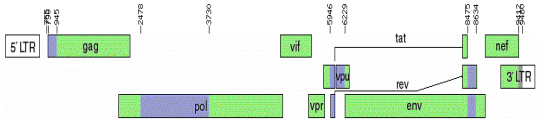
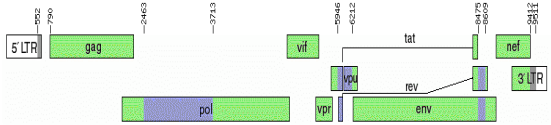
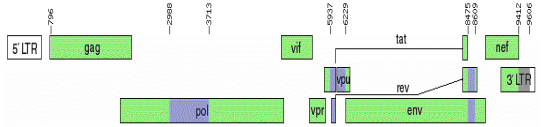
Genome maps of CRF02 and other AG recombinants.
All sequences were classified into 4 groups (with > 1 sequence) and 22 URFs based on genomic compositions and breakpoint locations predicted by jphmm program.
"Country" means the sampling country. Two-letter country code is used here. BE: Belgium, CD: Dem Rep of the Congo, CM: Cameroon, EC: Ecuador, FR: France, GH: Ghana, KE: Kenya, NG: Nigeria, SE: Sweden, SN: Senegal, US: United States, UZ: Uzbekistan.
"Sequence source" refers to the HIV database/literature-assigned CRF02, or AG set.
Digits in brackets are the sequence numbers.

Figure 2

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. 			

Genome maps of CRF07, CRF08 and other BC recombinants.
All sequences were classified into 2 groups (with > 1 sequence) and 7 URFs based on genomic compositions and breakpoint locations defined by jphmm program.
"Country" means the sampling country. Two-letter country code is used here. AR: Argentina. CN: China. MM: Myanmar.
"Sequence source" refers to the HIV database/literature-assigned CRF07, CRF08, and BC sets.
Digits in brackets are the sequence numbers.

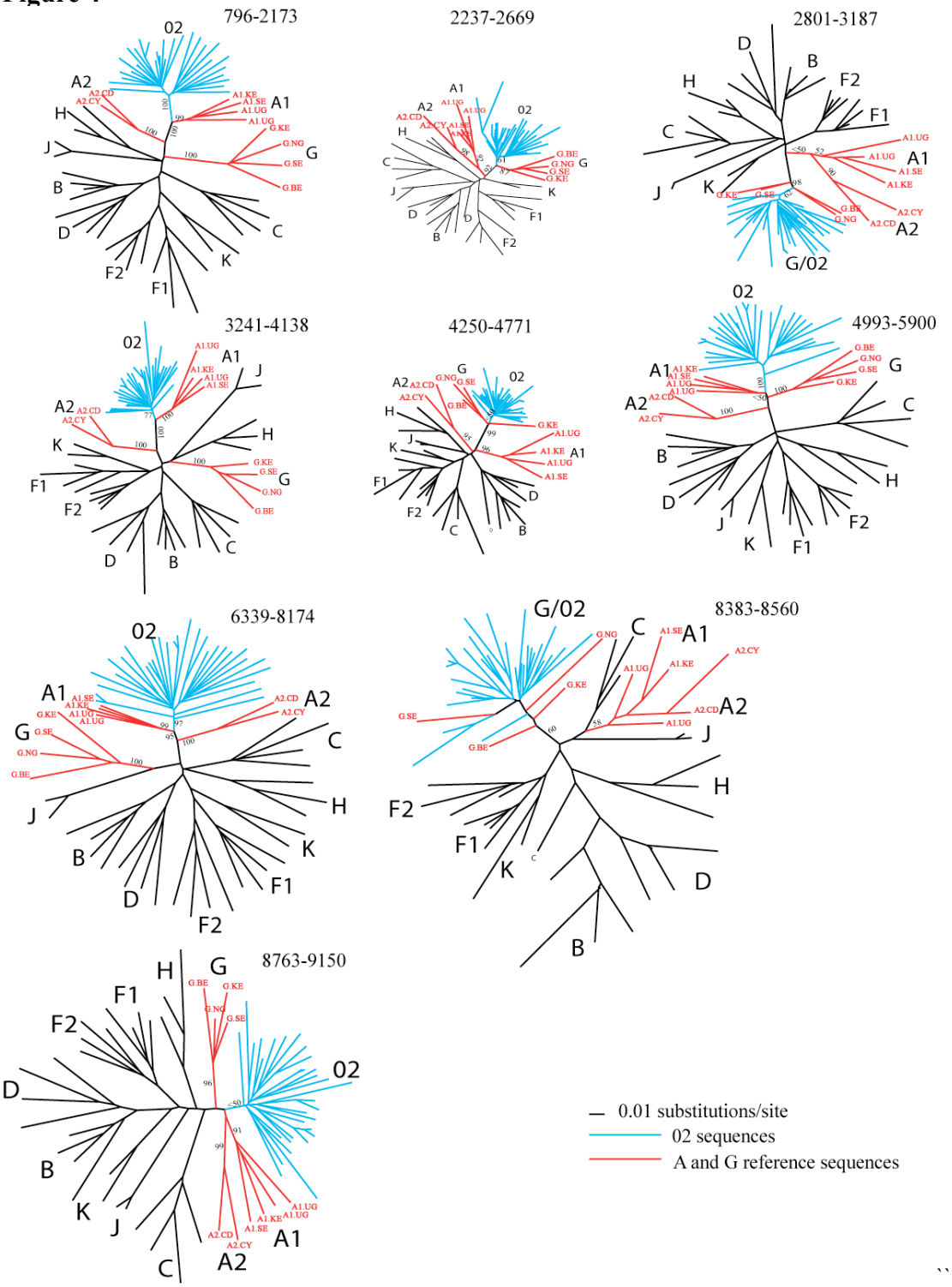
Figure 3

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) MTM (1) N/A(1)	6		CL(2) AR(1)	BF(3)	MTC (2) Hetero-sexual (1)
2	 (CRF28 group)	BR(2) VE(1)	CRF28 (2) BF(1)	Hetero-sexual (3)	7		AR(2)	BF(1) BF1(1)	Hetero-sexual (1) N/A (1)
3	 (CRF29 group)	BR(3)	CRF29 (3)	Hetero-sexual (1) MTC (1) N/A (1)	8		AR(1) UY(1)	CRF12 (1) BF(1)	Hetero-sexual (1) MTM (1)
4		AR(1) UY(1) BO(1) ES(1)	CRF12 (2) BF(2)	Hetero-sexual (2) MTM (2)	9		AR(1) UY(1)	CRF12 (1) BF(1)	IDU (1) MTM (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)	Hetero-sexual (15) MTC (3) IDU(2) Biosexual (1) N/A(7) HAT (1)

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

Genome maps of CRF12, CRF28, CRF29 and other BF recombinants.
All jpHMM –verified BF recombinants were classified into 9 groups (with > 1 sequence) and 29 URFs based on genomic compositions and breakpoint locations defined by jphmm program.
"Country" means the sampling country. Two-letter country code is used here. AR: Argentina, BO: Bolivia, BR: Brazil, CL: Chile, ES: Spain, UY: Uruguay, VE: Venezuela.
"Sequence source" refers to the HIV database/literature-assigned CRF12, CRF27, CRF28, CRF29, BF, and BF1 sets.
Digits in brackets are the sequence numbers.
In risk factor: MTM, men to men; MTC, mother to child; heterosexual, heterosexual contact; biosexual, biosexual contact; IDU, injecting drug use. HAT: heterosexual and transfusion. The transmission information was retrieved from the Los Alamos HIV sequence database.

Figure 4

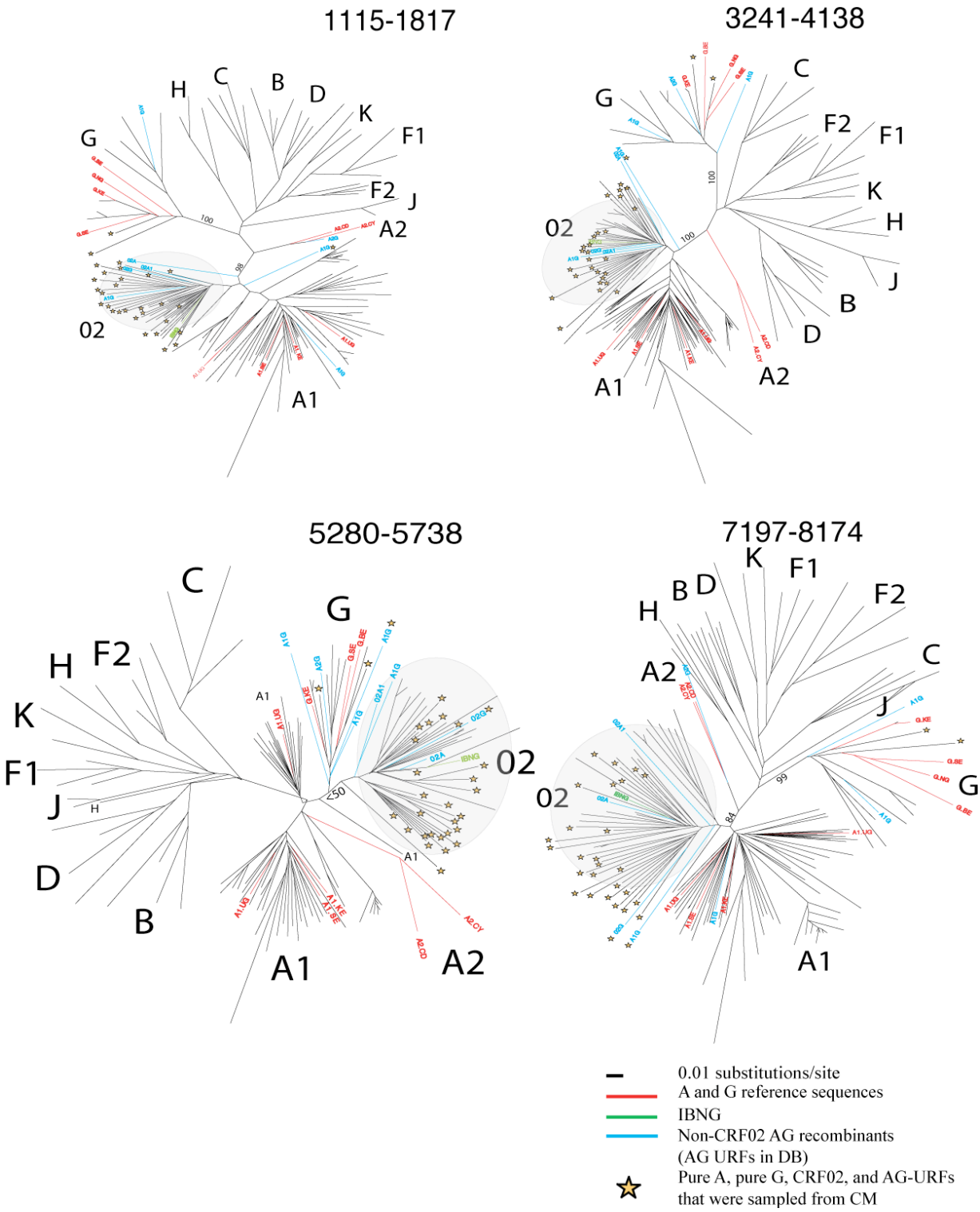


NJ trees of consensus sub-regions delimited by the breakpoints in groups 1-4 AG recombinants.

The genomic regions are based on HXB2 numbering.

The numbers at the node root are bootstrap values (in 1000 replicates).

Figure 5

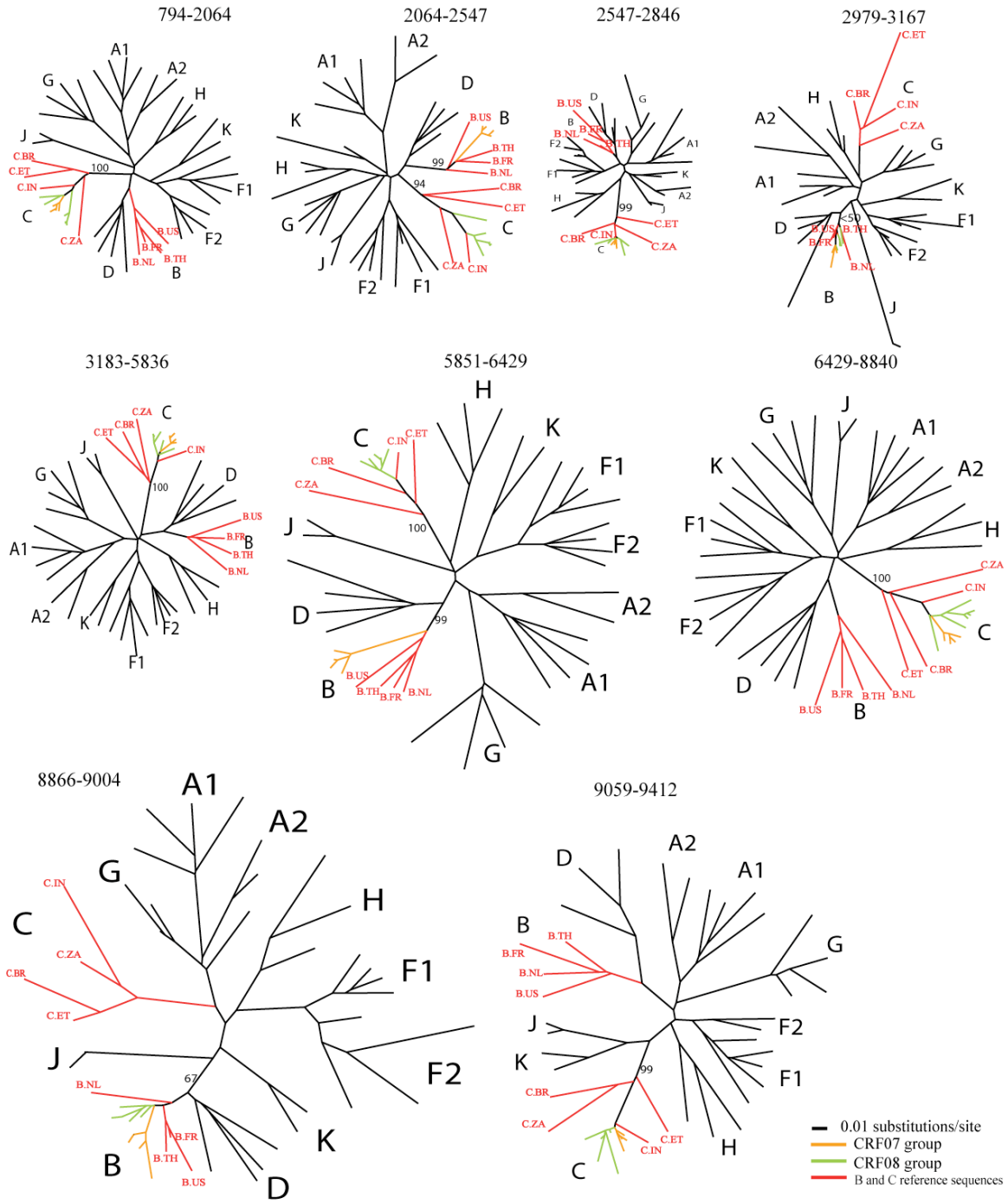


NJ trees of consensus sub-regions delimited by the breakpoints in all jpHMM verified CRF02 and AG sequences. All subtype A and G sequences (near full-length) from the database were also included.

The genomic regions are based on HXB2 numbering.

The numbers at the node root are bootstrap values (in 1000 replicates).

Figure 6

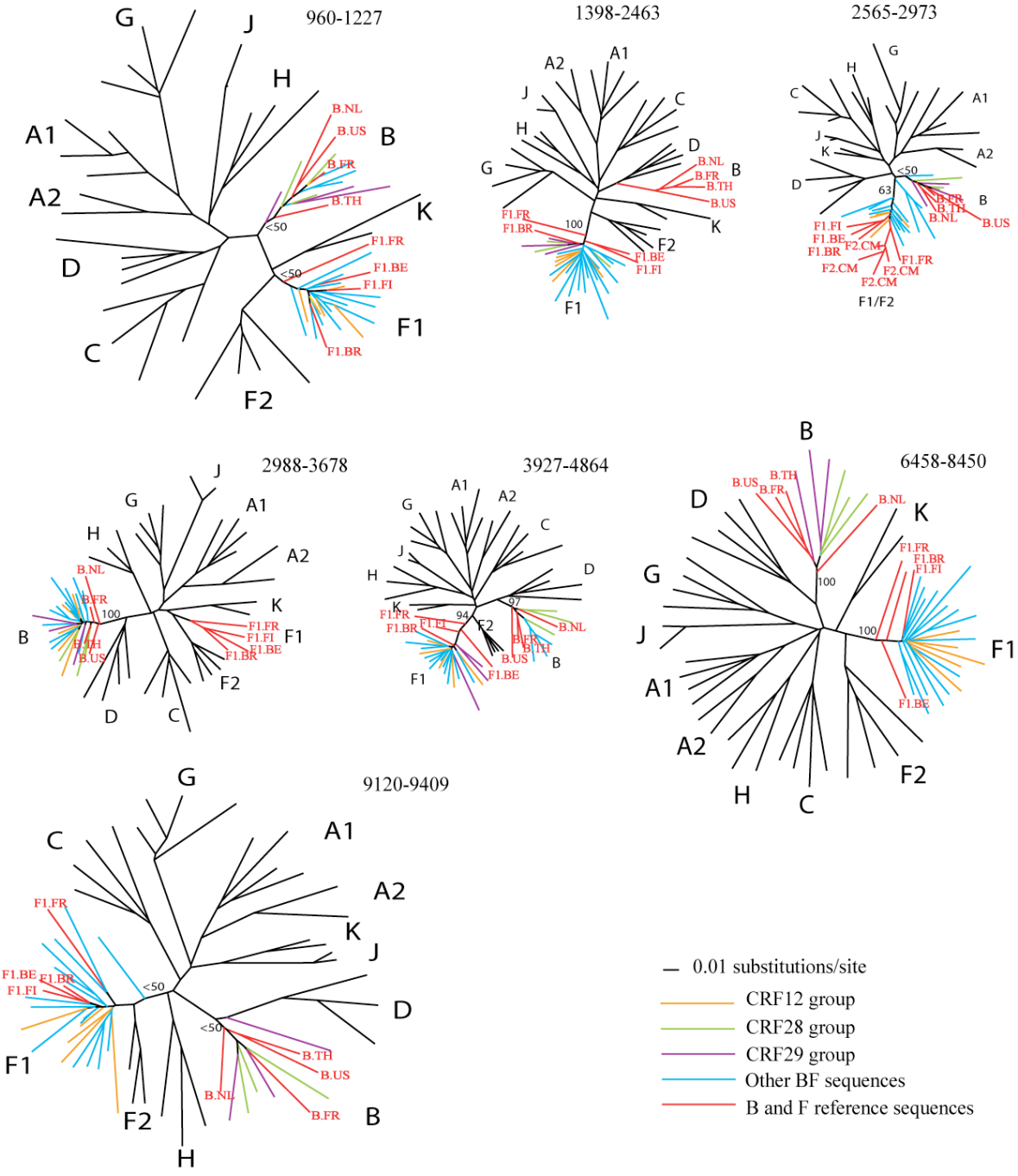


NJ trees of consensus sub-regions delimited by the breakpoints in CRF07 and CRF08 CRFs.

The genomic regions are based on HXB2 numbering.

The numbers at the node roots are bootstrap values (in 1000 replicates).

Figure 7

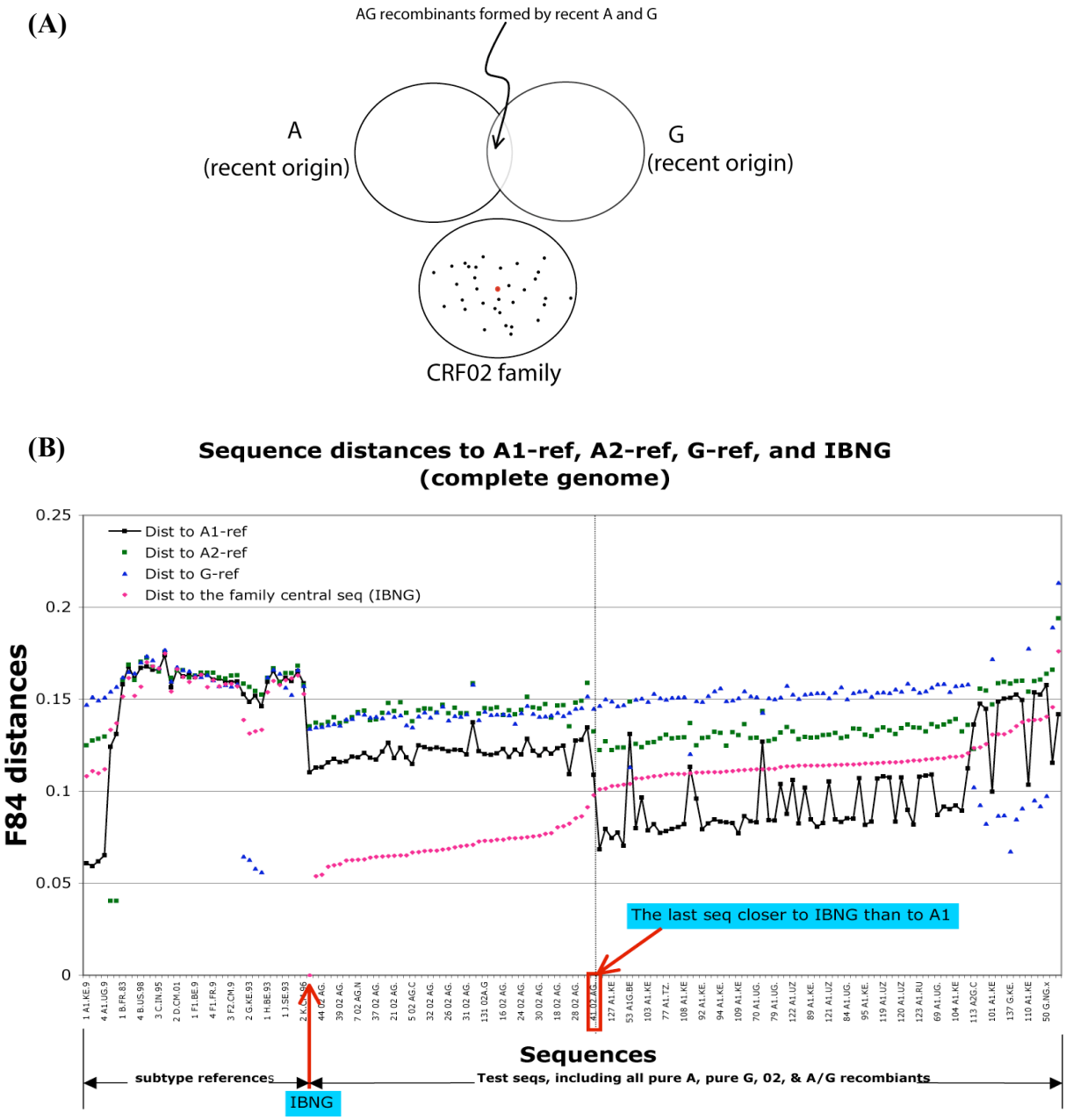


NJ trees of consensus sub-regions delimited by the breakpoints in groups 1-9 BF sequences.

The genomic regions are based on HXB2 numbering.

The numbers at the node root are bootstrap values (in 1000 replicates).

Figure 8

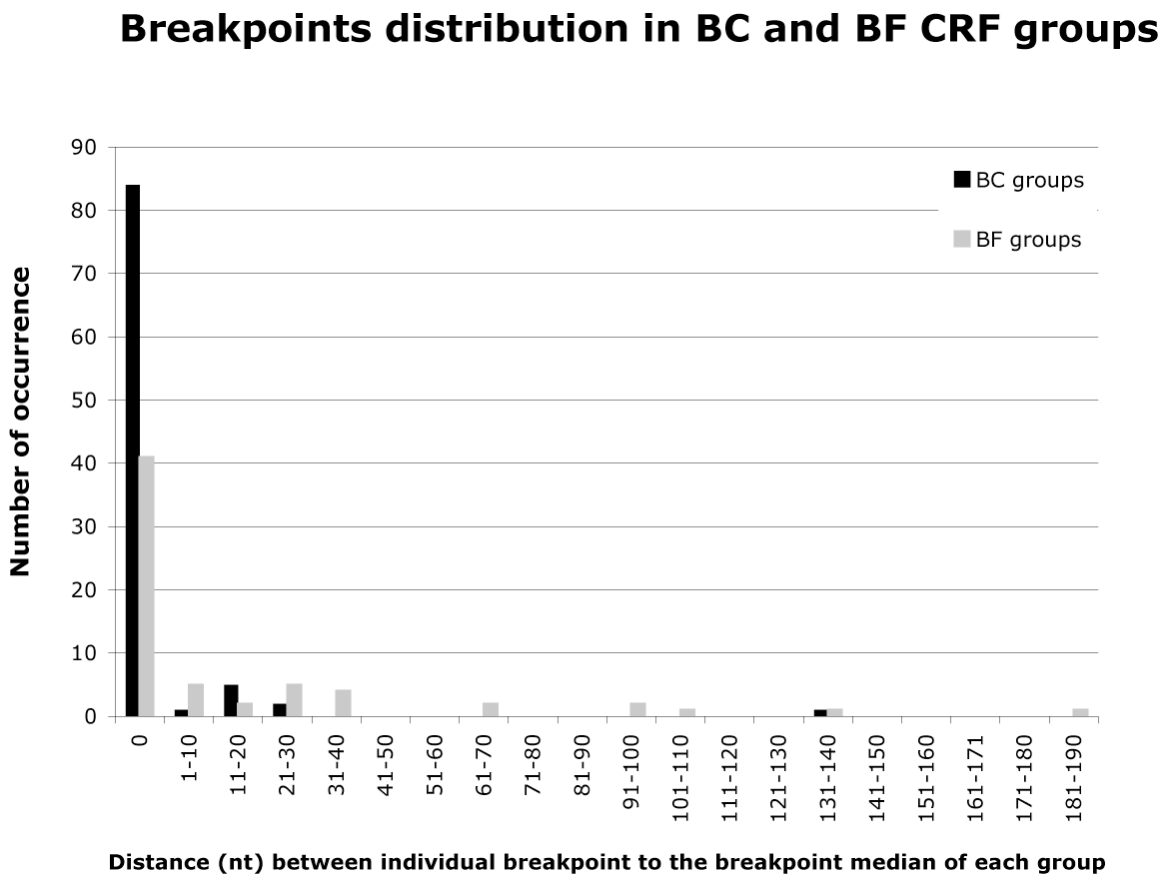


Defining CRF02 family using family plot.

Figure at the top: A cartoon shows the difference between the AG recombinants that were formed by contemporary subtypes A and G, and the AG recombinants (CRF02) that were generated between old A and G. Dots inside the CRF02 circle: CRF02 sequences. Dot in red: central strain in CRF02 family.

Figure at the bottom: Measuing a sequence's F84 distance to subtype A references, to subtype G references, and to IBNG, which was found to be the central sequence of the CRF02 family. By comparing these three distances, the last sequence that is closer to IBNG than to subtype A1, or G, was identified. This seqeunce defines the radius of the CRF02 family circle. All CRF02 sequences identified by this method were consistently found in the CRF02 clusters of all phylogenetic trees in figure 2, and 3.

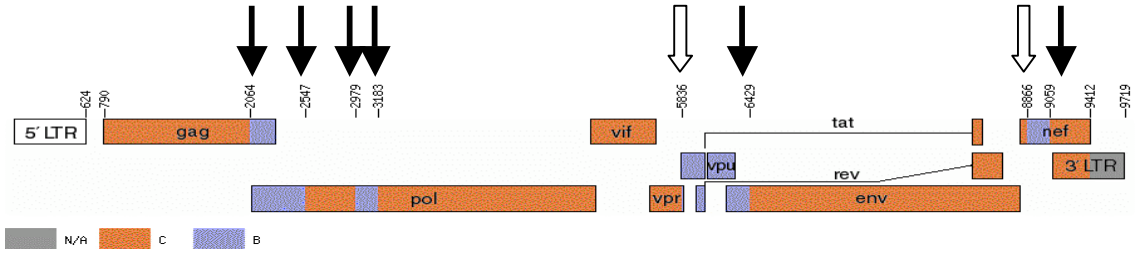
Figure 9



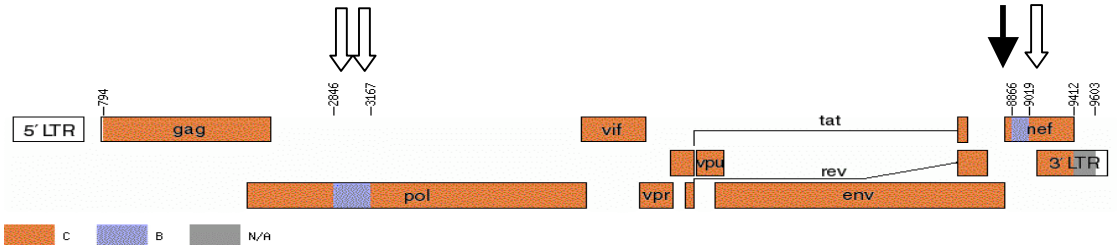
Defining the certainty region of breakpoints in BC and BF CRF groups.
The distance (nt) between each jpHMM-predicted breakpoint and the median of each breakpoint window was measured. In BC CRFs, > 95% breakpoints are off breakpoint median by 16nt. In BF CRFs, > 95% breakpoints are 98nt off breakpoint median. These two numbers (16nt and 98nt) were used as breakpoint certainty regions in BC and BF CRFs.

Figure 10

CRF07 group

Breakpoint number	1	2	3	4	5	6	7	8
Genome graph								
Breakpoint locations Of this CRF								
Breakpoint range	2064	2547	2979	3183	5836-5851	6429	8840-8866	9059
Breakpoint median	2064	2547	2979	3183	5836	6429	8866	9059
Interquartile range (mean)	const. (2064)	const (2547)	Const (2979)	const. (3183)	5836-5844 (5840)	const. (6429)	8854-8866 (8859)	const (9059)
# of complete BC recombinants (world)	Sequence total: 20 (China: 17)							
Frequency of full-length BC recombinants that bear bk within median ± 16nt (world)	7/20	7/20	8/20	9/20	7/20	7/20	12/20	7/20
Frequency of full-length BC recombinants that bear bk within median ± 16nt (China)	7/17	7/17	8/17	9/17	7/17	7/17	12/17	7/17
# of Asian BC fragments	Sequence total: 103 (China: 98)							
Frequency of fragmental BC recombinants that bear bk within median ± 16nt (Asia)	9/90	17/94	16/58	12/52	2/3	0/3	0/0	0/0
Frequency of fragmental BC recombinants that bear bk within median ± 16nt (China)	9/90	17/94	16/58	12/52	0/0	0/2	0/0	0/0

CRF08 group

Breakpoint number	1	2	3	4
Genome graph				
Breakpoint locations Of this CRF				
Breakpoint range	2846-2979	3167-3183	8866	9004-9028
Breakpoint median	2846	3167	8866	9019
Interquartile range (mean)	2846-2846 (2868)	3167-3179 (3172)	const. (8866)	9019-9019 (9018)
of complete BC recombinants (world)	Sequence total: 20 (China: 17)			
Frequency of full-length BC recombinants that bear bk within median ± 16nt (world)	5/20	13/20	12/20	6/20
Frequency of full-length BC recombinants that bear bk within median ± 16nt (China)	5/17	13/17	12/17	6/17
# of Asian BC fragments	Sequence total: 103 (China: 98)			
Frequency of fragmental BC recombinants that bear bk within median ± 16nt (Asia)	67/94	43/52	0/0	0/0
Frequency of fragmental BC recombinants that bear bk within median ± 16nt (China)	67/94	43/52	0/0	0/0

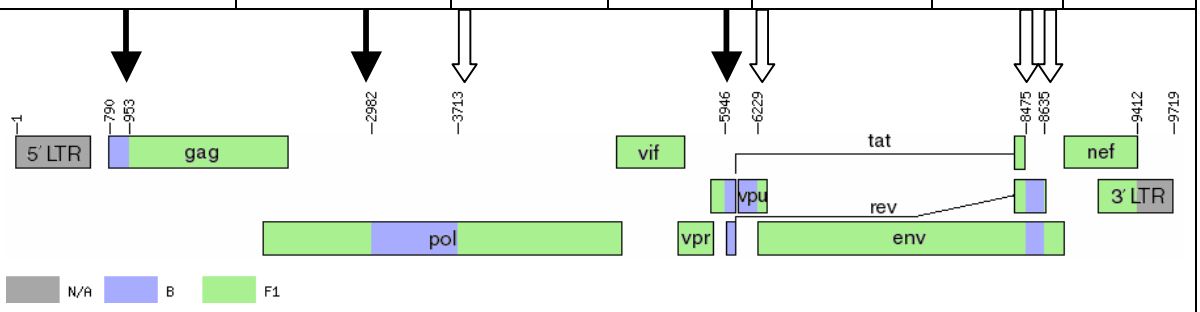
Frequency of CRF07 and CRF08 breakpoints.

Breakpoint frequency = N/M, where N=total sequence number of BC recombinants that have breakpoints within breakpoint median ± 16nt, and have the same subtypes flanking the breakpoint as they are in the CRF group; M= total sequence number of BC recombinants that span this genomic region.

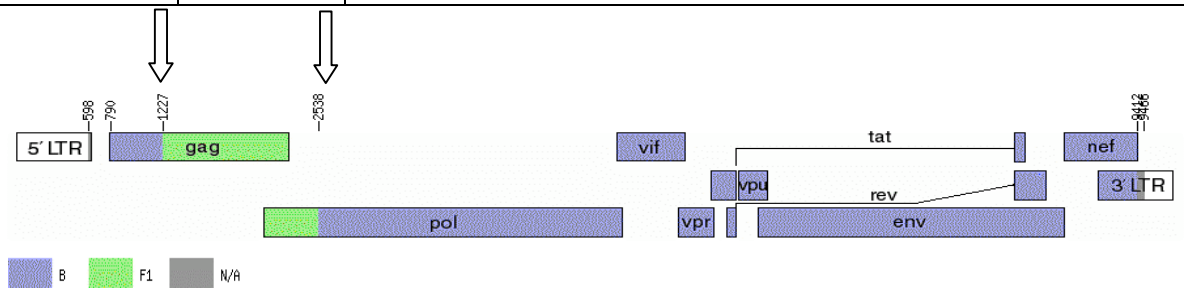
Black arrow: a breakpoint exists at a fixed position; Hollow arrow: breakpoint region.

Figure 11

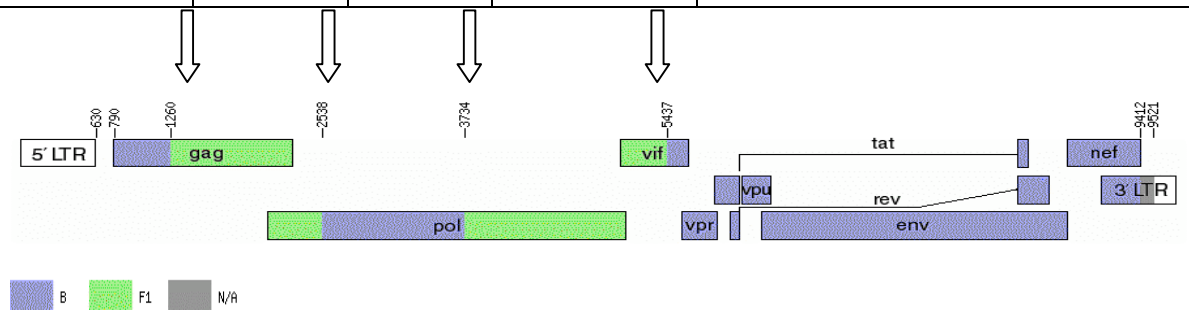
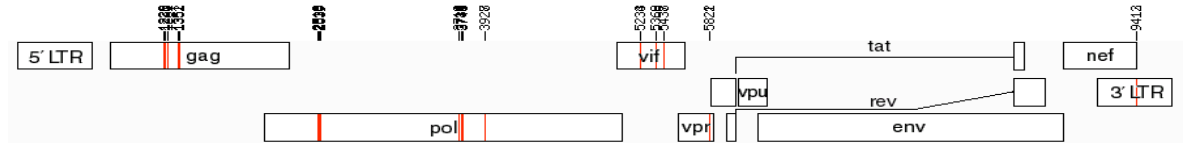
CRF12 group

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 (world)	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 American 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	4/5	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

CRF28 group:

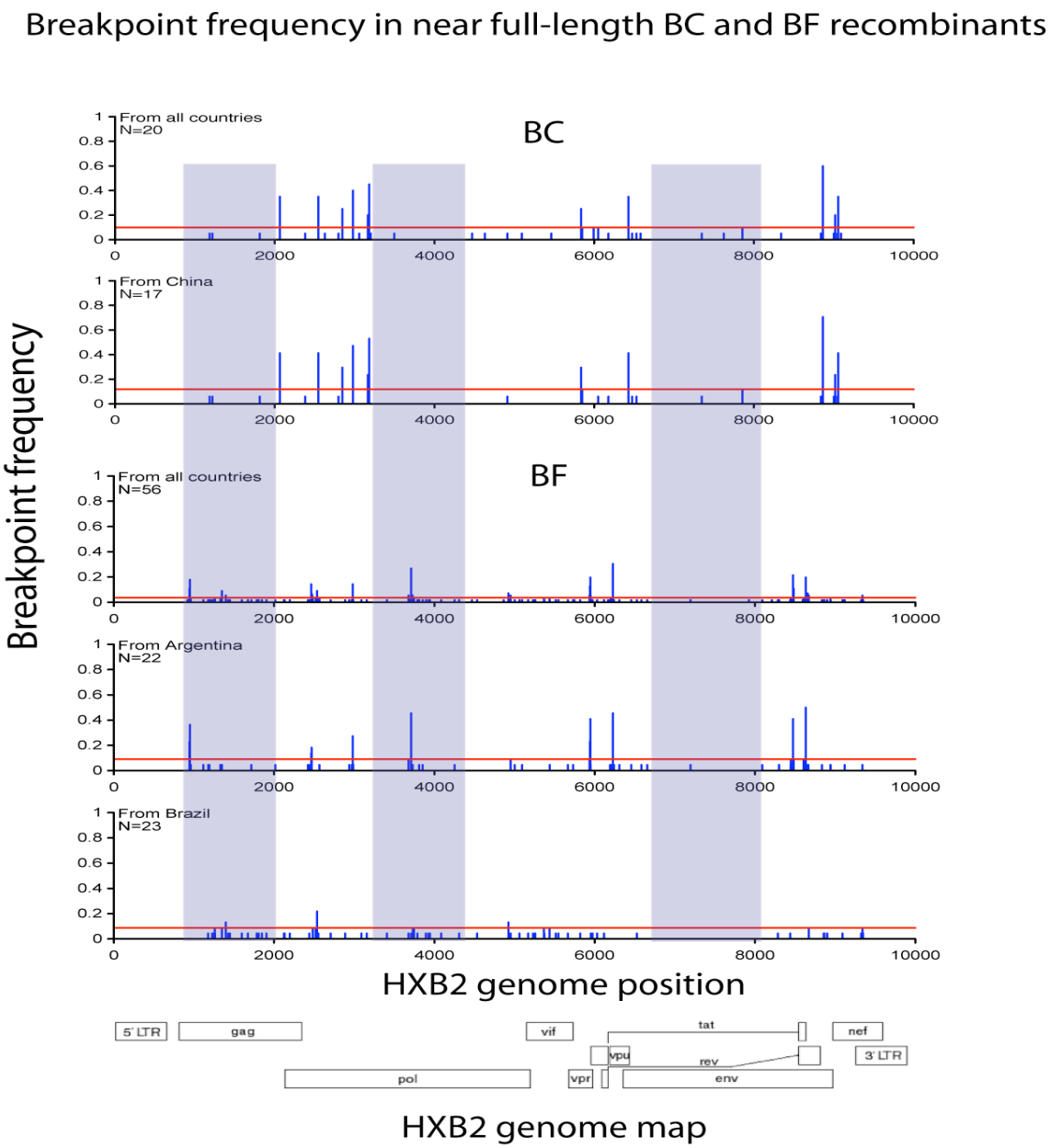
Breakpoint number	1	2	
Genome graph			
Breakpoint locations Of this CRF			
Breakpoint range	1227-1398	2538-2565	
Breakpoint median	1329	2538	
Interquartile (mean)	1278-1364 (1318)	2538-2552 (2547)	
# of complete BF recombinants (world)	Sequence total: 56 (Argentina: 22, Brazil: 23)		
Frequency of full-length 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 SouthAmerica 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	

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 (world)	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)	1/4	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	

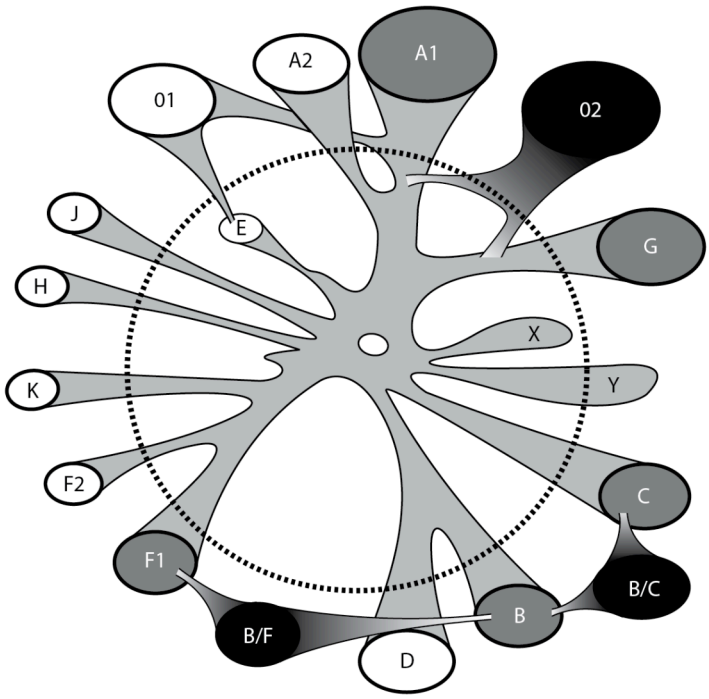
Frequency of CRF12, CRF28, and CRF29 breakpoints.
The breakpoint frequency = N/M, where N=total sequence number of BF recombinants that have breakpoints within breakpoint median ± 98nt, and have the same subtypes flanking the breakpoint as they are in the CRF group; M= total sequence number of BF recombinants that span this genomic region.
Black arrow: a breakpoint exists at a fixed position; Hollow arrow: breakpoint region.

Figure 12



Breakpoint frequency in near full-length BC and BF recombinants.
The breakpoint positions are based on HXB2-numbering.
Highlighted grey regions. Left and middle: BC recombinants have less breakpoints than BF recombinants. Right: both BC and BF recombinants have few breakpoints in portion of gp120.
Red line: breakpoints present in 3 sequences. Above red line: breakpoints shared in > 3 sequences. Below red line: breakpoints shared in < 3 sequences.

Figure 13



Contemporary sequences co-exist with some old sequences in the current HIV-1 epidemic.
The dashed circle differentiates the old and contemporary sequences. Inside the circle, the old sequences, like subtype E strains, may be no longer exist in the epidemic. We can only deduce its old presence from CRF01_AE, a recombinant between subtype A and E. "X" represents an extinct strain, "Y" represents an old strain still circulating in the current epidemic, but it hasn't been identified. CRF02 is an old recombinant derived from old A and old G. BF and BC recombinants are rather new. Their parental sequences are contemporary sequences.

Glycobiology vol. 14 no. 12 pp. 1229–1246, 2004
doi:10.1093/glycob/cwh106
Advance Access publication on June 22, 2004

Tracking global patterns of *N*-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin

Ming Zhang[2], Brian Gaschen[2], Wendy Blay[3,4], Brian Foley[2], Nancy Haigwood[3,4], Carla Kuiken[2], and Bette Korber[1,5]

[2]Theoretical Biology Group, Los Alamos National Laboratory, Los Alamos, NM 87544; [3]Seattle Biomedical Research Institute, Seattle, WA 98109; [4]Pathobiology Department, University of Washington, Seattle, WA 98195; and [5]The Santa Fe Institute, Santa Fe, NM 87501

Received on February 13, 2004; revised on May 26, 2004; accepted on May 28, 2004

Human and simian immunodeficiency viruses (HIV and SIV), influenza virus, and hepatitis C virus (HCV) have heavily glycosylated, highly variable surface proteins. Here we explore *N*-linked glycosylation site (sequon) variation at the population level in these viruses, using a new Web-based program developed to facilitate the sequon tracking and to define patterns (www.hiv.lanl.gov). This tool allowed rapid visualization of the two distinctive patterns of sequon variation found in HIV-1, HIV-2, and SIV CPZ. The first pattern (fixed) describes readily aligned sites that are either simply present or absent. These sites tend to be occupied by high-mannose glycans. The second pattern (shifting) refers to sites embedded in regions of extreme local length variation and is characterized by shifts in terms of the relative position and local density of sequons; these sites tend to be populated by complex carbohydrates. HIV, with its extreme variation in number and precise location of sequons, does not have a net increase in the number of sites over time at the population level. Primate lentiviral lineages have host species–dependent levels of sequon shifting, with HIV-1 in humans the most extreme. HCV E1 and E2 proteins, despite evolving extremely rapidly through point mutation, show limited sequon variation, although two shifting sites were identified. Human influenza A hemagglutinin H3 HA1 is accumulating sequons over time, but this trend is not evident in any other avian or human influenza A serotypes.

Key words: immune escape/*N*-linked glycosylation/neutralization antibody/variability/virus

Introduction

N-*linked glycosylation sites and their role in viral immune escape*

N-linked glycosylation sites are also referred to as sequons. For an asparagine (N) to be glycosylated, it requires the context of the amino acid pattern N-X-[S or T] (Marshall, 1974), where X can be any amino acid, followed by a Serine (S) or Threonine (T). A sequon will not be glycosylated if it contains or is followed by a Proline (Gavel and von Heijne, 1990), and glycosylation may be inhibited by certain combinations of N-X-S or when followed by specific amino acids (Kasturi *et al*., 1995; Mellquist *et al*., 1998; Shakin-Eshleman *et al*., 1996).

Alteration of a glycosylation site can have dramatic consequences for a virus. It can impact protein folding (Hebert *et al*., 1997; Land and Braakman, 2001; Slater-Handshy *et al*., 2004) and conformation (Meunier *et al*., 1999) and affect distant parts of a protein through masking or conformational changes. Although gain of a carbohydrate can sterically mask epitopes, the loss of one could result in tighter packing of glycoprotein regions involved in neutralization epitopes, reduce accessibility, and so also facilitate immune escape (Ye *et al*., 2000). The loss of sequons can even impact immunogenicity of noncovalently associated proteins, for example a change in sequons in the human immunodeficiency virus type 1 (HIV-1) transmembrane envelope (Env) protein gp41 induces conformational changes in the associated Env gp120 surface protein that dramatically diminish the binding of many gp120- specific antibodies (Si *et al*., 2001). Altered patterns of glycosylation in viral proteins can also contribute to escape from T cell responses (Botarelli *et al*., 1991; Ferris *et al*., 1999; Selby *et al*., 1999) and influence receptor binding and phenotypic properties of viruses (Kaverin *et al*., 2002; Koito *et al*., 1995; Matrosovich *et al*., 1999; Ogert *et al*., 2001; Pollakis *et al*., 2001).

Influenza, glycosylation, and antigenic drift

Some of the earliest studies on the biological and immunological consequences of glycosylation site variation were conducted in influenza proteins (Alexander and Elder, 1984). The number of sequons in the heavily glycosylated influenza A hemagglutinin 1 (HA1) of the pandemic H3 virus has increased from 6 to 10 since it entered the human population in 1968 (Skehel and Wiley, 2000), and the increase is assumed to make HA1 generally more refractive to antibodies. For example, the amino acids around the *N*-linked glycosylations site at position N165 of HA stopped participating in antigenic drift (Skehel and Wiley, 2000; Wiley and Skehel, 1987).

HIV Env and glycosylation site variation

HIV-1 is highly variable, and variants are grouped through phylogenetic analysis into major clades or subtypes (A–K). Recombinant forms of HIV are very common (Robertson *et al*., 2000), and when a lineage based on recombination

[1]To whom correspondence should be addressed; e-mail: btk@lanl.gov

between two subtypes becomes an important epidemic lineage, it is called a circulating recombinant form (CRF). HIV-1 varies dramatically within a clade and within infected individuals. HIV Env gp120 is among the most heavily glycosylated proteins in nature (Myers *et al.*, 1992), far more heavily glycosylated than envelopes of other retroviruses of similar size (e.g., HTLV-1, MuLV) (Polonoff *et al.*, 1982). The influence of sequon variation on HIV antibody recognition and viral phenotype has been studied in the context of HIV and SIV Env proteins (Chackerian *et al.*, 1997; Cheng-Mayer *et al.*, 1999; Losman *et al.*, 2001; Ly and Stamatatos, 2000; Malenbaum *et al.*, 2000; Matthews *et al.*, 1987; Ratner 1992; Ye *et al.*, 2000). Sigvard Olofsson and colleagues first showed that glycosylation of gp120 could change exposure of neutralizing antibody epitopes (Bolmstedt *et al.*, 1996). In the HIV-1 CRF01, the absence of a sequon near the base of the V3 loop, an important antigenic doman of Env gp120, was correlated with rapid amino acid substitutions and positive selection (Kalish *et al.*, 2002), suggesting a similar situation to the influenza A HA protein, where a glycosylation site may provide regional protection from antibodies (Skehel and Wiley 2000). The number of Env sequons in both HIV and SIV infections varies extensively within infected individuals (Overbaugh and Rudensey, 1992; Wolinsky *et al.*, 1992), and this variation constitutes a mechanism of immune escape during the course of an HIV or SIV infection (Cheng-Mayer *et al.*, 1999; Davis *et al.*, 1990; Simmonds *et al.*, 1991).

A heavily glycosylated face of the 3D structure of Env gp120 has been called the immunologically silent face (Moore and Sodroski, 1996). Generally carbohydrate moieties appear as self to the immune system, so this face reduces the antigenicity of a large region on the gp120 surface. Glycosylation of variable loops restricts access to conserved receptor binding sites and limits their exposure to the immune system (Wyatt and Sodroski, 1998), and HIV Env has been described as having a glycan shield (Wei *et al.*, 2003).

A recent survey of the global collection of HIV-1 Env gp120 surface protein M group sequences in the Los Alamos HIV database showed that gp120 varied in length between 484 and 543 amino acids (Korber *et al.*, 2001). The number of potential sequons in gp120 ranges between 18 and 33 with a median of 25 (Korber *et al.*, 2001) (Figure 1) (this range is often ignored and HIV is frequently reported to have 25 *N*-linked glycosylation sites). The dramatic variation in number of sequons partly results from gp120 length variation frequently involving insertions and deletions of potential glycosylation sites inside HIV hypervariable domains, however, evolutionary propensities in base substitution patterns in HIV may also contribute to the rapid flux in numbers of sequons in gp120 (Bosch *et al.*, 1994; Kuiken *et al.*, 1999).

Not all potential glycosylation sites on a given HIV-1 Env protein are fully occupied (Zhu *et al.*, 2000). For example, NNTT is a common pattern among HIV sequences, and steric occlusion may prevent carbohydrate addition to both asparagines, and the protein context of sequons does not always favor glycosylation (Kasturi *et al.*, 1995; Mellquist *et al.*, 1998; Shakin-Eshleman *et al.*, 1996).

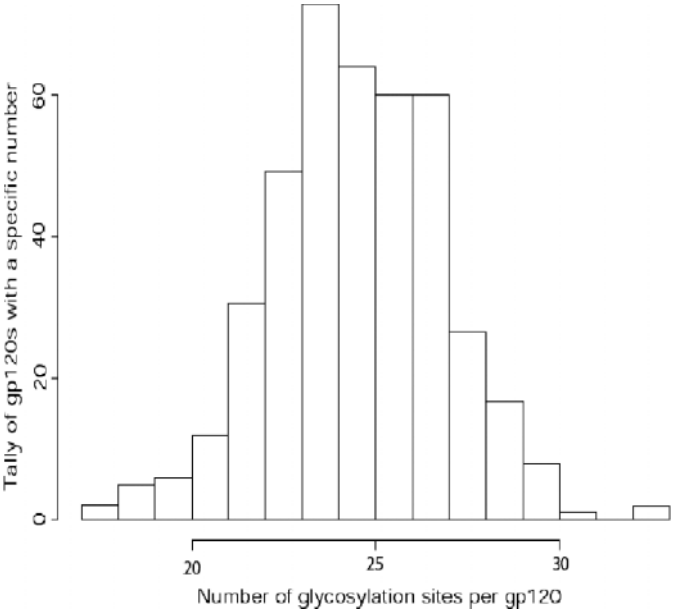


Fig. 1. Histogram showing the relative frequency of envelope gp120 proteins with different numbers of N-linked glycosylation sites found in the HIV database (www.hiv.lanl.gov, 2002 listing). Only one sequence obtained from a given individual and complete gp120 Env sequences were included in this comparison ($n = 386$).

Glycosylation and hepatitis C Env diversity

Hepatitis C virus (HCV) belongs to the Hepacivirus genus in the Flaviviridae family (Rice, 1996). Like HIV-1, it is highly variable; HCV establishes a chronic infection in most hosts and so is subject to continuous immune pressure and rapid accumulation of mutations. The virus has been classified into six different genotypes, which have each been subdivided into a large number of subtypes. The Env E1 and E2 proteins of HCV form heterodimers on the virion surface, and glycosylation is essential for this dimerization (Meunier *et al.*, 1999). The variability of both proteins is comparable, from 88% nucleotide (90% amino acid) identity between strains from the same subtype to 55% nucleotide (59% amino acid) identity between different genotypes. In both E1 and E2, *N*-glycosylation sites are limited to the amino terminus of the proteins; the carboxy-terminal region of these proteins is the transmembrane portion. The efficient glycosylation of E1 is dependent on the presence of E2 in a polyprotein (Deleersnyder *et al.*, 1997), although it does not appear to depend on the specific sequence of E2 (Dubuisson *et al.*, 2000), and the noncovalent association of E1 and E2 depends on the first and fourth glycosylation sites of E1 (Meunier *et al.*, 1999). It has been shown that the glycans attached to the E1E2 heterodimer prior to budding of the virus are exclusively high-mannose (Deleersnyder *et al.*, 1997), although E1E2 found circulating on HCV virions also have complex carbohydrates (Sato *et al.*, 1993). As in HIV, sequon changes have been shown to change antibody exposure in HCV (Fournillier *et al.*, 2001).

In this study, we characterize patterns in viral glycosylation site variation at the population level for influenza, HIV, and HCV and describe a Web-based program that

was used to facilitate tracking sequons in protein alignments for these comparisons.

Results

Influenza proteins and N*-linked glycosylation patterns over time*

Using alignments and sampling dates obtained from the 2003 Los Alamos Influenza Sequence Database (www.flu.lanl.gov), we confirm (Figure 2A) the H3 HA1 protein gradually increases in the median number of *N*-linked glycosylation sites over time (from 6 to 10), with a small amount of within-year variation. The acquisition of some of these sites can be directly related to antigenic drift (Skehel and Wiley, 2000), the change in the antigenic profile of influenza from year to year that necessitates annual review and frequent updates of the vaccine strain.

However this net increase in the number of sequons is not a general feature of influenza evolution and is only found in the human H3 serotype of influenza A, not in other serotypes of influenza A (Figure 2), nor in the avian H3 over time, or in any other avian serotype with adequate sampling over time, nor in the human influenza B HA1 (Figure 2, plus summary in legend). There was also no trend for increasing numbers of sequons over time in the avian or human neuraminidase (N2) serotype proteins, although this protein is also heavily glycosylated and shows variable numbers of sites from year to year (Figure 2D).

Global trends in glycosylation patterns in HIV Env

There is no net tendency for the sequon number to increase or decrease over time in the HIV-1 M group, or within subtypes or CRFs (Figure 3). Two kinds of sequons are evident in HIV-1 and HIV-2: those embedded in readily aligned positions and those embedded in hypervariable regions that shift in relative position and regional density by base mutation and by insertions and deletions. Here we refer to these two classes of sequons as *fixed* and *shifting*, respectively. Although most of the shifting sequons are found in variable loops, some are also present in the C4 region (conserved, or C, domains in the HIV envelope were called conserved because they are relatively conserved when compared to the variable regions, however, they can also span insertions and deletions that can result in shifting sequon locations). The location and frequency of sequons in gp120 in each major subtype of the HIV-1 M group (Gao *et al*., 1996; Robertson *et al*., 2000) are illustrated in Figure 4. Despite the extreme variability between isolates, there is an essentially conserved pattern of variation in each of the HIV-1 M group subtypes, and even in the genetically very distant HIV-1 group O. (There are insufficient full-length HIV-1 group N sequences for a comparison). Protein regions show the same frequencies for most sequons in each HIV-1 lineage and subtype, suggesting that selective pressures on sequons in these diverse lineages are consistent (Figure 4).

HXB2 and SF2 are common HIV-1 reference strains for which the *N*-linked carbohydrates additions to Env gp120 have been biochemically defined (Leonard *et al*., 1990; Zhu *et al*., 2000). We aligned the SF2 protein sequence to HXB2 strain and noted the types of carbohydrate found: mannose, complex carbohydrate, or a mixture of both (hybrid carbohydrates). We also found that different types of carbohydrates were associated with fixed and shifting sequons defined at the population level (Figure 4). There are 24 *N*-linked glycosylation sites in HXB2, and 26 in SF2, and they differ in location in seven cases. In HXB2, all seven shifting sequons in Env gp120 are complex carbohydrates (7/13 complex carbohydrate additions, 54% are in the shifting regions), whereas all high-mannose or hybrid carbohydrates are in fixed positions (11/11, 100%, Fisher's exact test p-value = 0.006). When an additional sequon in the shifting V1 region of the HXB2-related variant LAI was later characterized, it also carried a complex carbohydrate (Gram *et al*., 2002), consistent with this pattern. For SF2, some of the sequon positions contained hybrid carbohydrates, and some were only partially glycosylated. Though the high-mannose residues were not significantly correlated the fixed sequons of SF2, the trend from HXB2 was evident: 8/11 (72%) of high mannose glycosylation sites were in fixed sequons, whereas 8/14 (57%) complex carbohydrates were in shifting sequons, and one was unknown. A structural model of the fully glycosylated SF2 gp120 suggests that the high-mannose glycans are clustered on one surface of the protein and the complex carbohydrates are localized on distinct region of the protein surface (Zhu *et al*., 2000).

The 2G12 monoclonal antibody is one of the few broadly cross-reactive human HIV-neutralizing antibodies, and it has an unusual epitope involving several mannose carbohydrates that project from the generally immunorefractive glycosylated face of HIV-1 Env. Not surprisingly, cross-competition studies have shown that B cell responses to 2G12 epitope are very unusual in HIV-1-infected individuals (Moore and Sodroski, 1996). The sites that comprise the epitope, indicated in Figure 4, require mannose at positions of N295, N332, and N392 for 2G12 binding (Sanders *et al*., 2002; Scanlan *et al*., 2002). (See the HXB2 sequence locator tool in the HIV database [www.hiv.lanl.gov] to determine the specific positions referred to in this text.) These sites are well conserved in most subtypes, and tend to show comparable levels of glycosylation. The exception is subtype C, which only rarely has a glycosylation site immediately next to the Cys at the base of the V3 loop (N295). It has been suggested that the 2G12 epitope may be conserved because its structure enhances gp120–mannose interactions with the human protein dendritic cell–specific HIV-1-binding protein (DC-SIGN), an interaction that faciliatates efficient HIV-1 infection (Sanders *et al*., 2002).

The global collection of HIV Env gp120 molecules shows small but significant variations in the distributions of sequon frequencies for each subtype (Kruskal-Wallis p-value = 0.0002). Occasionally a site will be completely lost or added in a subtype (Figure 4), like the loss of a shifting sequon in CRF01 (subtype E in Env) in the V4 region or the additional *N*-terminal fixed site in O group. These differences may simply reflect a founder effect in the lineage, or may confer a critical change to the conformation of Env in the context of a particular lineage. Although the sequon frequencies in different HIV-1 subtypes are

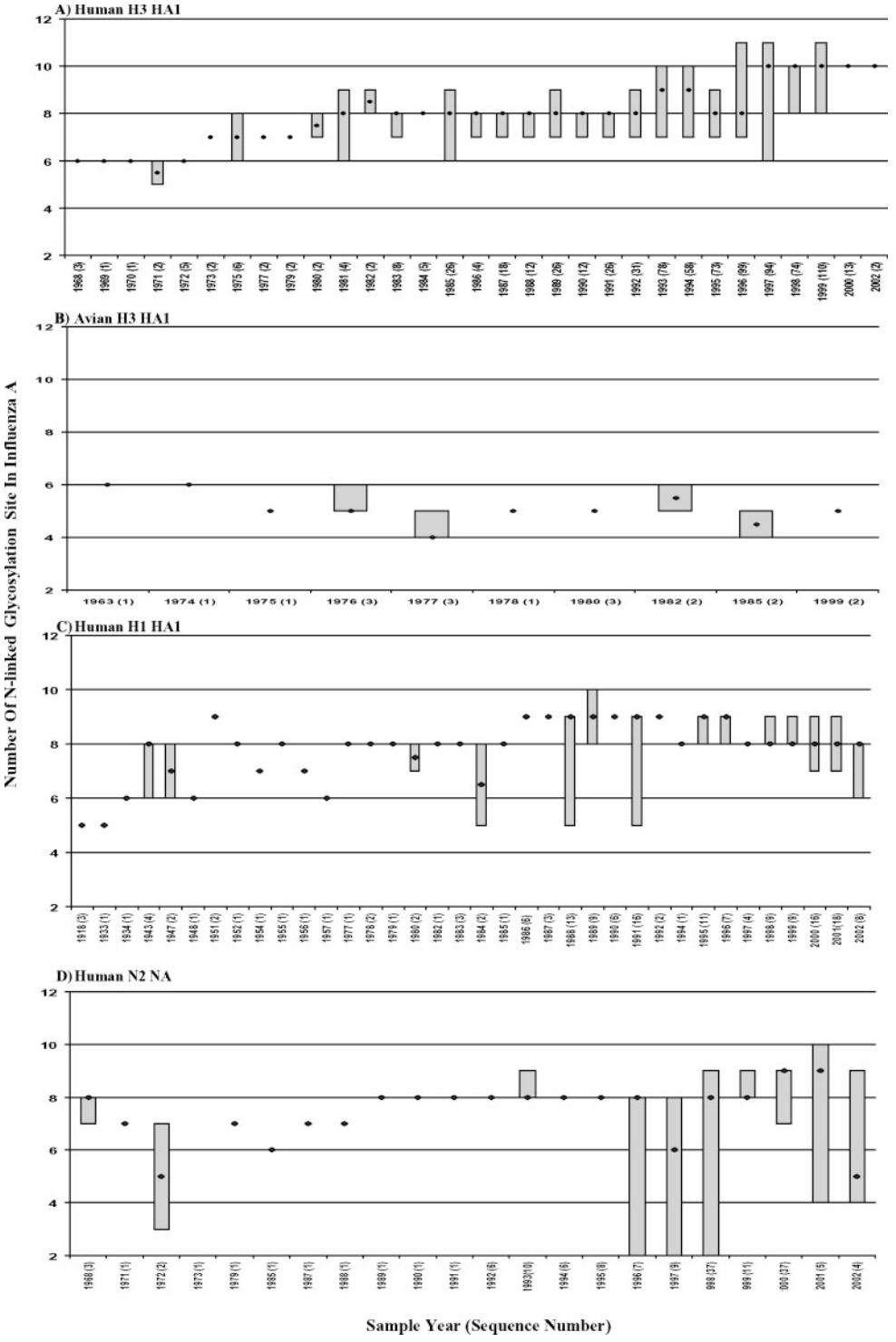


Fig. 2. Changes in the number of N-linked glycosylation sites in different influenza A proteins over time. (**A**) Human H3 HA1 proteins. This figure shows the median value and the range of number of sequons per sequence by year of sampling, with the number of sequences from each year in parentheses and the year noted on the *x*-axis. Samples were collected between 1968 and 2002. The increase in number of sites each year is evident; the *p*-value indicating this accumulation is not due to chance alone is very low ($p \ll 10^{-5}$), however, this is misleading because the epidemic strains in any one year are not independent (Korber *et al.*, 2001). (**B**) Avian H3 HA1 proteins. The H3 HA1 avian samples collected between 1963 and 1999 oscillate between four and six glycosylation sites. (**C**) Human H1 HA1 proteins. The H1 HA1 human isolates obtained between 1934 and 2002 oscillate between five and nine sites, with a lone samples from 1918 and 1933 having only five sites. (**D**) Human neuraminidase N2 proteins. The N2 proteins from human isolates do not show accumulation of glycosylation sites over time, although there is variation (between two and nine sites). Other influenza HA1 molecules with adequate data for testing are not shown because no clear trends over time were apparent. These include the avian H5 HA1, oscillating between five and eight sites between 1959 and 2001 ($n = 94$); avian H7 HA1, oscillating between three and five sites between 1927 and 2000 ($n = 122$); avian H9 HA1, staying steady with a median of six sites between 1966 and 2001 ($n = 63$), and the human influenza B hemagglutinin HA1, which maintained a stable median of seven sites with a range of six to eight from 1940 to 2002 ($n = 420$).

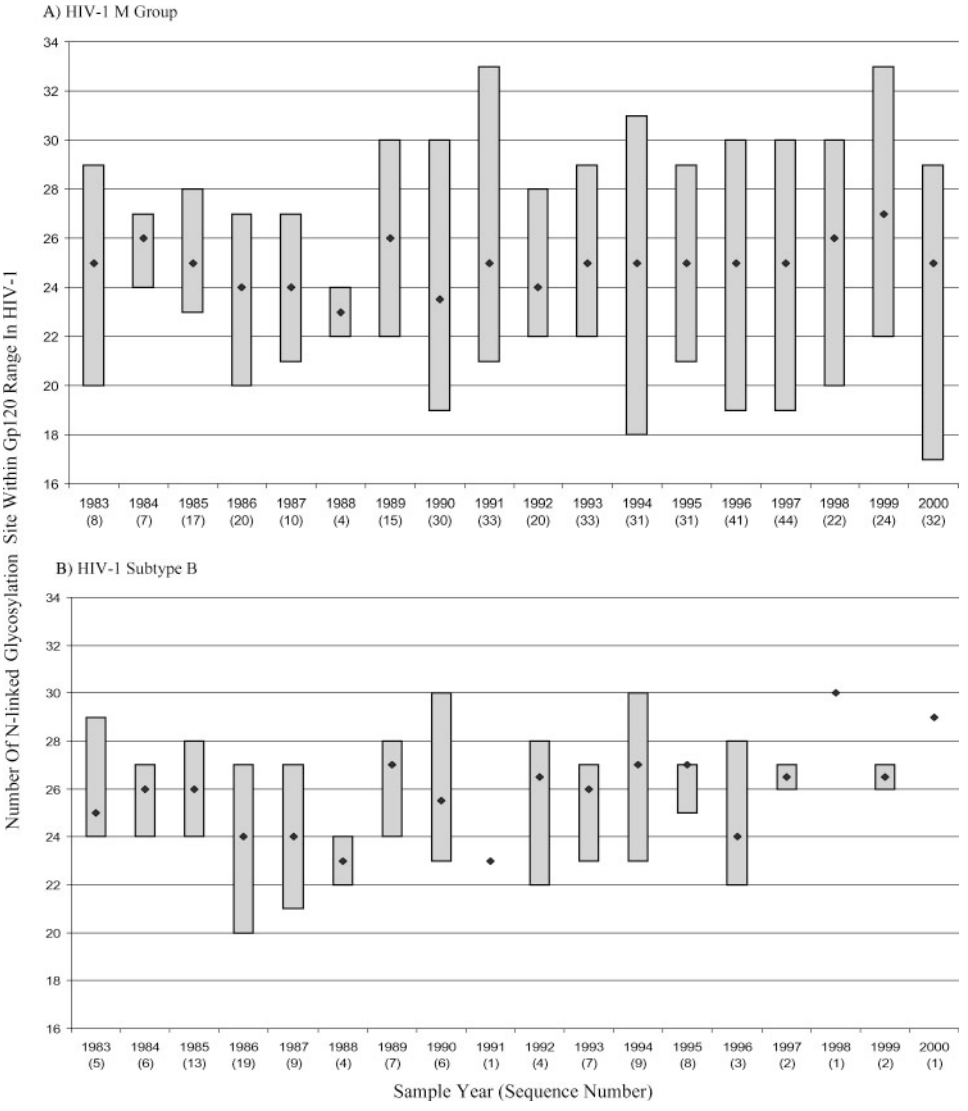


Fig. 3. The median and range of number of *N*-linked glycosylation sites per Env gp120 in the HIV-1 M group and subtype B between 1983 and 2000. (**A**) No correlation was observed in terms of increasing or decreasing numbers of sequons in gp120 at the population level over a 20-year period of sampling of M group sequences in the database. (**B**) Distributions of the number of *N*-linked glycosylation sites for each of the subtypes and CRFs in **A** did not reveal any trends in terms of accumulation or decline in number of sequons over time; subtype B is shown as a representative set.

significantly different, perhaps more important are the similarities between subtypes: The range of sequon frequencies is broad and basically overlapping within all subtypes (Gao *et al*., 1996) (Figure 5), and the conserved sites as well as frequencies of variable sites tend to be comparable between subtypes (Figure 4).

N-*linked glycosylation patterns related to coreceptor usage and phenotype*

HIV-1 Env enters CD4-positive T cells through a series of steps involving binding to both the CD4 protein and a chemokine receptor protein, usually CCR5 (R5 viruses) or CXCR4 (X4 viruses); some viruses can use either chemokine receptor (R5X4 viruses) (Cho *et al*., 1998). Sets of Env protein sequences derived from R5, X4, or R5X4 viruses were obtained from in the Los Alamos HIV database (www.hiv.lanl.gov). The total number of sequons in gp120, and in subregions of gp120 including the variable domains V1 and V2, which play a role in coreceptor usage in certain contexts (Cho *et al*., 1998; Hoffman *et al*., 1998; Pollakis *et al*., 2001), were nearly identical among R5, X4, and R5X4 viruses. However, two sequons in fixed sites show distinctive frequencies (Figure 6). The most striking was the sequon located within the V3 loop at position N300 of HXB2, which was present in all 44 R5 isolates (100%), but only in 4/11 (36%) X4 isolates and 7/11 (64%) R5X4 isolates (Fisher's exact test p-value $= 4 \times 10^{-7}$). Site-directed mutagenesis has shown this site to be associated with coreceptor usage (Ogert *et al*., 2001; Pollakis *et al*., 2001). The other potentially interesting site was at position N230 of HXB2, which was less common in the CCR5-utilizing group relative to the CXCR4 and R5X4 viruses, although this difference was just a trend and not statistically significant (11/44 [25%] had the site among R5 viruses, versus 6/11 [55%] in X4, and 5/11 [45%] in R5X4).

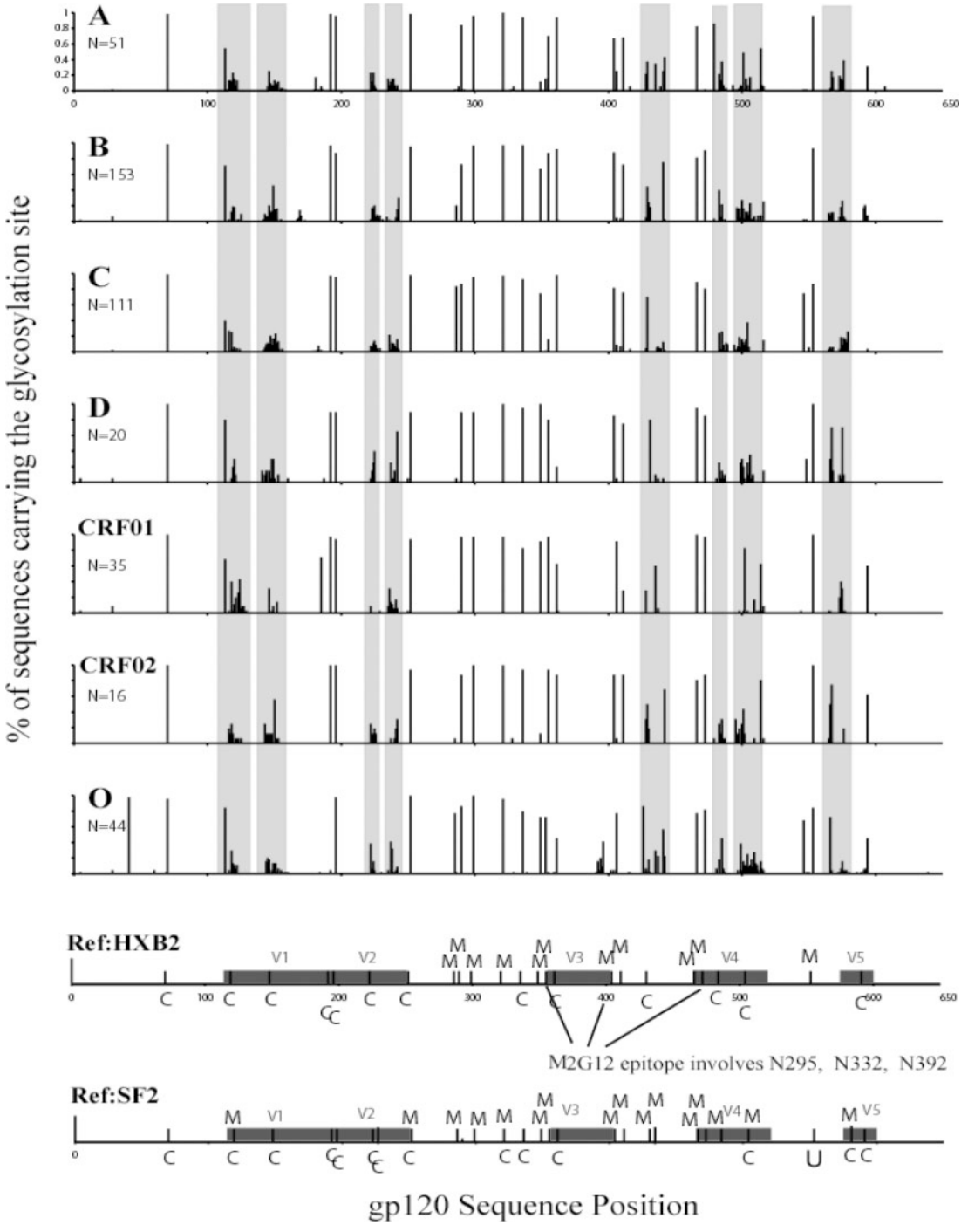


Fig. 4. Frequency of *N*-linked glycosylation sites in HIV Env gp120 based on an alignment of HIV-1 M group subtypes and circulating recombinant forms. Sequons in HIV-1 subtypes and CRFs with the most available full-length gp120 sequences are shown here. The sequences were aligned, and the fraction of sequons in every position in the alignment of each subtype is indicated. The sequons we refer to as shifting are highlighted in gray. The alignment is numbered according to positions in the HXB2 reference strain. The sequons in particularly positions in two HIV-1 reference strains are indicated, HXB2 and SF2. The biochemistry of the oligosaccharide additions has been determined for these two strains (Leonard *et al*., 1990; Zhu *et al*., 2000), and high mannose is indicated at the top of appropriate sequons with an M, complex carbohydrates are labeled under each sequon with a C. Some mixtures of both were identified in SF2 (Zhu *et al*., 2000). The locations of variable loops in gp120 are indicated in black box on the reference strains. The high-mannose carbohydrates critical for 2G12 binding are indicated (Sanders *et al*., 2002; Scanlan *et al*., 2002).

A virus's ability to form multinucleated syncytia in indicator cell lines distinguishes R5 and X4 viruses. Viruses isolated from newly infected people tend to be nonsyncytium-inducing (NSI) and R5, whereas X4 and syncytium-inducing (SI) viruses tend to appear later in infection, and the SI/NSI designations are available for many viral sequences that do not have defined coreceptor usage. So to further compare the glycosylation patterns

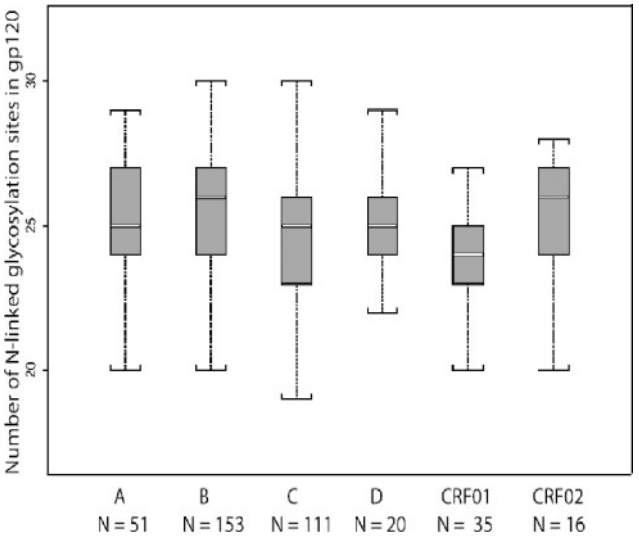


Fig. 5. Distributions of number of *N*-linked glycosylation sites in HIV-1 Env gp120 proteins by HIV-1 subtype. A box plot indicating median and interquartile range of the number of sequons in the most commonly sequenced subtypes and circulating recombinant forms of HIV-1 M group is shown. The number of full-length sequences included in the summary of each clade is noted. A nonparametric Kruskal-Wallace test indicated that there were distinction between the distribution of numbers of sequons in the M group subtypes and circulating recombinant forms (*p*-value of 0.0002), despite the overlap in the distributions.

with viral phenotypes, we analyzed the number and position of sequons in full length Env, including gp41 (Figure 7) as well as gp120, from SI and NSI patients. Two significant changes were observed: One was the loss of a sequon at position N136 in SI viruses, located in the V1 loop, and the other was a gain of a sequon in gp41 at position N674.

We then determined if the sequon alterations associated with phenotypic change at the population level are mirrored in an individual patient (Hu *et al*., 2000). In the patient studied, there was a one-to-one correspondence between NSI and R5 usage and SI and X4 usage. All significant changes of sequons were found in regions V1, V2, V4, and V5, primarily in the *N*-terminal regions of V1 and V4. In gp120, there are 19 highly conserved sites, 4 additional sequons from NSI to SI, and 5 losses from NSI to SI. Three of these changes were in shifting sites, simply moving the sequon by one position relative to context of the protein (in V1, N136 to N136+, N141 to N142, and in V4, N401 to N402). The NSI-to-SI phenotypic switch was accompanied by these three shifts in sequon position, a loss of sequon N188 in V2 and N405 in V4, and a gain of a sequon at position N463 in V5. In this patient, the sequon at position N300 was unchanged and not related to coreceptor usage, and gp41 sequons were invariant, emphasizing that there are exceptions to the statistical correlations that can be detected at the population level.

N-*linked glycosylation patterns in other primate lentiviruses*

Among the primate lentiviruses, Env sequon variation in M group viruses is the most extreme relative to the genetic distances based on base substitution (Wills *et al*., 1996). However, SIVs do vary extensively and also show some interesting glycosylation patterns (Chackerian *et al*., 1997; Nyambi *et al*., 1997; Overbaugh and Rudensey, 1992). Using phylogenetic analysis, we selected HIV-2 and SIV sequences representing the spectrum of natural diversity in each lineage. Shifting sites, analogous to those found in HIV-1, were apparent in HIV-2 and in chimpanzee viral Envs (SIV_{CPZ}). The human HIV-1 epidemic is thought to have resulted from a cross-species transmission from chimpanzee, whereas human HIV-2 and macaque SIV_{MAC} result from cross-species transmission of SIV_{SMM} virus from the sooty mangabey, its natural host (Gao *et al*., 1996; Hirsch *et al*., 1989). In SIV_{SMM} and African green monkey SIV_{AGM}, shifting sites tended to resolve into fixed sites (Figure 8A), with the exception of the V4 region in SIV_{AGM}. Table I highlights the sequons in the V1 loops from sequences used to generate the frequencies shown in Figure 8. Although there was length variation in SIV_{SMM} and SIV_{AGM} V1 sequences, the relative placement of the sequons and the Cys involved in the disulfide bridge that forms the base of the loop was conserved (Table I). In contrast, shifting sequons are frequently seen in infected macaques (SIV_{MAC}) followed over time through progression to simian AIDS (Chackerian *et al*., 1994, 1997; Overbaugh *et al*., 1991) and in HIV-2 sequences in human (Figure 8A, Table I). Because HIV-2 and SIV_{MAC} infections result from cross-species transmission of SIV_{SMM} virus from sooty mangabeys (Gao *et al*., 1996; Hirsch *et al*., 1989), the degree of variation in shifting sites is host-specific for viruses of this lineage. SIV_{SMM} does not cause disease in sooty mangabeys, but it does in macaques, as does HIV-2 in humans. Like SIV_{SMM}, SIV_{AGM} does not cause disease in its natural host species, African green monkeys, and is readily found in animals in the wild (Norley *et al*., 1999; Ohta *et al*., 1988).

HCV E1 and E2 protein N-*linked glycosylation site patterns*

The sequons in HCV Env E1 and E2 proteins are far less variable in HIV-1 envelopes, despite HCV otherwise being an extraordinarily variable virus. After creating alignments of the available HCV E1 and E2 sequences in GenBank, the sequences were checked by phylogenetic analysis, and very similar sequences were removed, leaving 294 distinct E1 protein sequences and 130 E2 sequences for comparisons. Numbering of the E1 and E2 positions in this section is based on the Los Alamos Hepatitis C Database HCV Sequence Locator tool (http://hcv.lanl.gov/content/hcv-db/LOCATE/locate.html).

There are typically 5–7 sequons in E1 over a stretch of 194 amino acids. Four of these sites are highly conserved among all genotypes. The site at position N61 in the alignment was present in genotype 6 and 1b but absent in all other genotypes (Figure 9, top). A pair of adjacent sequons at positions N43/N44 also showed subtype-specific variation, the site at position N43 being present in subtype 2b, and the one at N44 in 2a and 2c, as well as in most other sequences. In 35% of the subtype 1b sequences, the N43 and N44 sites were both present; this occurred in 1 sequence of another genotype, a 3b. The sequences with two sequons did not

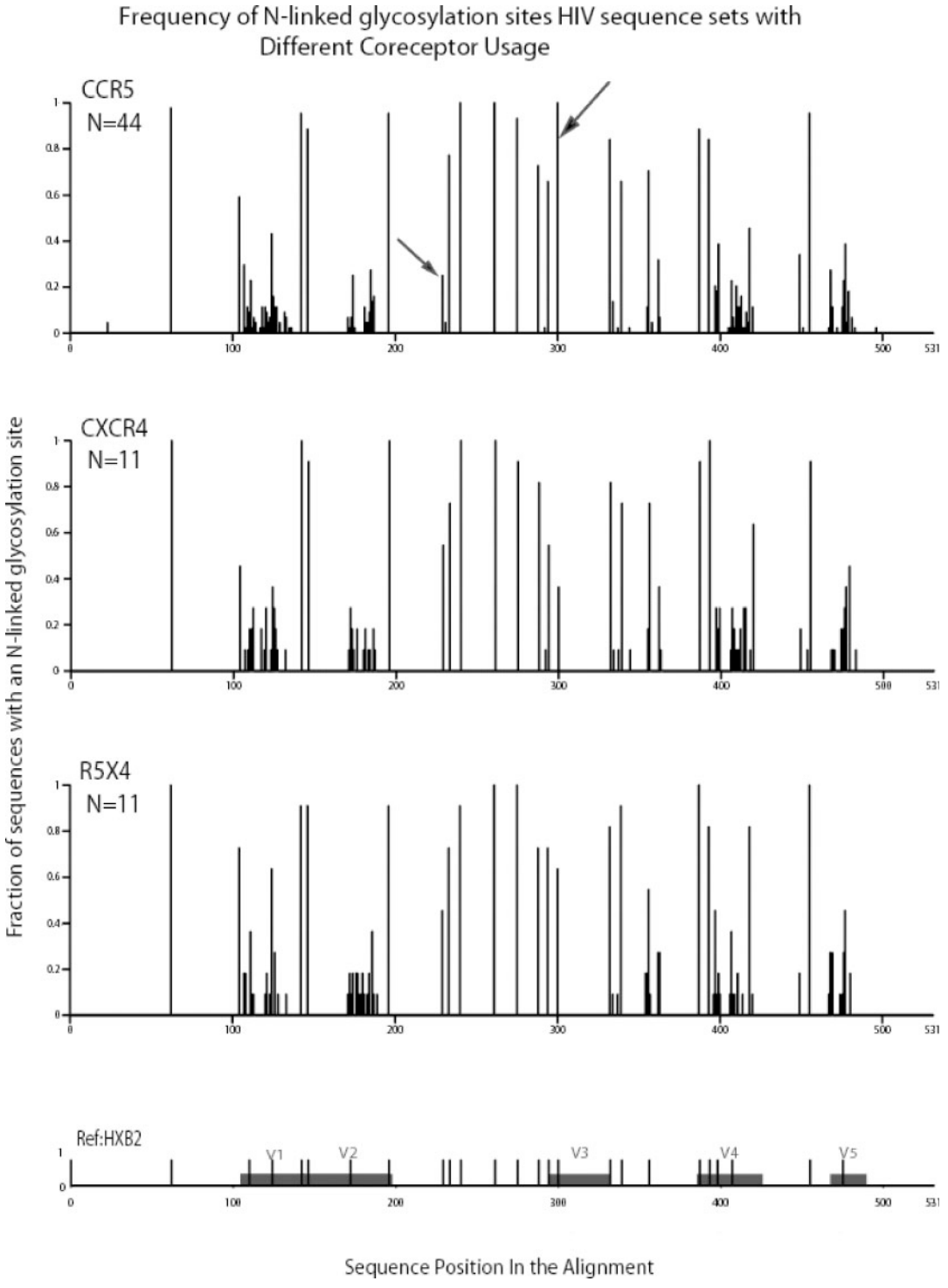


Fig. 6. The relationship between coreceptor usage and *N*-linked glycosylation sites in the HIV-1 M group. The sequences that made up the sets associated with particular coreceptor usage were very diverse M group sequences from a variety of subtypes. The set of 44 R5 viruses included 8 subtype A, 12 B, 13 C, and 11 others; the 11 X4 viruses included 1 subtype A, 5 B, 2 C, and 3 D; the 11 R5X4 viruses included 3 subtype A, 6 B, 1 C, and 1 D. Arrows in the R5 graph indicate the sequons with potentially distinctive frequencies.

appear to come from a specific geographic region and probably arose independently. The sequons at positions N5 and N136 of E1 have been shown to be essential for formation of E1–E2 complexes on the virion surface (Meunier *et al*., 1999); N136 also is important in reducing the antigenicity of the virus (Fournillier *et al*., 2001). Little is known about genotype- or subtype-specific differences in viral phenotype.

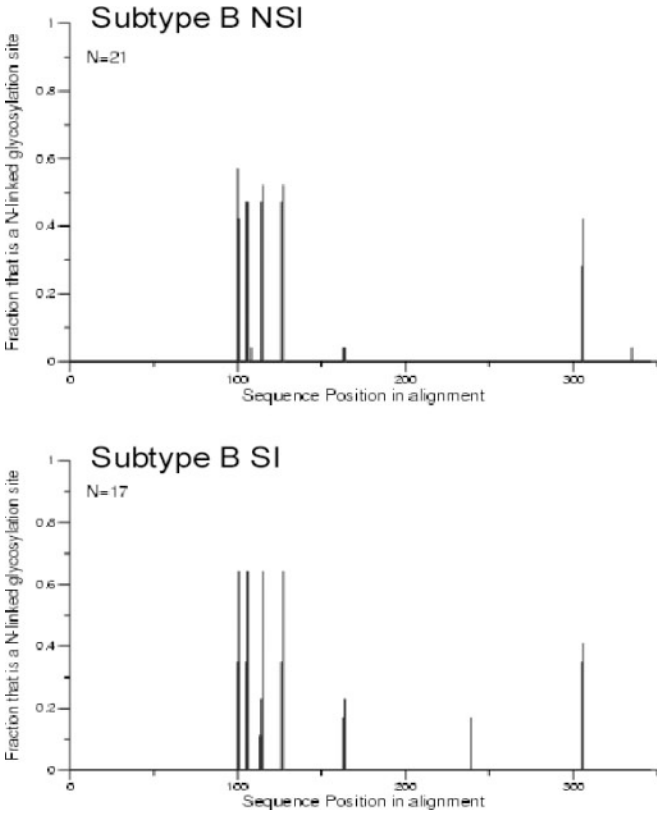


Fig. 7. *N*-linked glycosylation site changes in Env gp41 in NSI isolates and SI isolates from patients infected with HIV-1 subtype B. HIV-1 B gp41 sequences were aligned and analyzed for sequon differences between 21 NSI and 17 SI sequences. Protein sequences were obtained from the HIV database; only one sequence per patient was used. Sequences were analyzed for patterns of glycosylation using the LANL *N*-glycosite program (http://hiv-web.lanl.gov/content/hiv-db/GLYCOSITE/glycosite.html).

Sequons in the E2 protein were mostly limited to the *N*-terminal part of the protein, with 10–11 sites all located before amino acid 272 (the total length of the E2 alignment was 436 amino acids). Some type-specific sequon patterns were also found in E2, as all sequences of genotypes 3 and 6 missed the site at position N160. Only one or two sequences of genotypes 4 and 5 were available, so variation analysis was not possible for these genotypes. Within genotype 1, the sequon at position N94 was highly conserved in subtype 1a but present in only 17/101 sequences of subtype 1b. This site was one of the two sequons in E2 that shifted in terms of its relative position, that is, it could be found starting anywhere between positions 92 and 95 (Figure 9, bottom), suggesting that either there is pressure for this subtype to incorporate a glycosylation site in this region of the protein or that there is a selective advantage in retaining the site in the region but modifying its precise location. Another shifting site was found in E2, specifically in genotype 6. This site was embedded in the only region with multiple insertions and deletions in HCV, and it shifts relative location on the basis of insertions and deletions, and is lost in one of the sequences (Figure 9, bottom). The presence of shifting glycosylation sites in the E2 protein, like HIV-1 Env, is probably related to its profound immunogenicity and escape-related variability.

Discussion

The potential of changing *N*-linked glycosylation sites as a mechanism for immune evasion has long been appreciated (Alexander and Elder, 1984). In influenza, the steady increase in the number of sequons in H3 HA1 over time (Skehel and Wiley, 2000) is the exception, as the number of sequons in human H1 HA1 and neuraminidase N2 proteins oscillates but does not show a net increase. The sequon frequency in avian influenza hemagglutinin proteins oscillates at roughly comparable levels to the human proteins, despite having a very different ecology: In aquatic avian species flu is asymptomatic, and the viruses are in evolutionary stasis (Webby and Webster, 2001), whereas in human epidemics, the virus causes disease and generates new variants that are able to infect hosts who have become immune to earlier influenza strains (Bush *et al*., 1999; Ferguson *et al*., 2003). It seems possible that the increase in sequon frequency in H3 HA1 in humans is a local fluctuation over time, and it too will move to random oscillation in future years. Although year of sampling was not readily available for the HCV sequences, HCV shows much less variation in sequons, and this variation seems to be due to lineage-specific patterns that would be unlikely to vary by sampling year.

There is a striking recapitulation of sequon patterns in different levels of HIV evolution. The level of sequon variation that can be seen in an individual is mirrored at the level of variation seen within a subtype, which is further reflected at the level of variation seen within the HIV-1 M group global epidemic. Thus the boundaries of the potential to add and eliminate sequons to facilitate immune escape and infect a range of cell types, within what is tolerated in terms of fitness costs, may be explored anew during the course of each HIV-1 infection of a single individual. Although the phylogenetic tree of HIV-1 is clearly expanding over the time period during which we have followed HIV-1 in the human population (Korber *et al*., 2000; Robbins *et al*., 2003; Yusim *et al*., 2001), the sequon variations in gp120 and gp41 seem instead to be rapidly fluctuating within inherent boundaries.

There is probably an upper bound to the number of sequons that can be maintained on Env gp120, because carbohydrate structures are quite large and their presence or absence must influence the protein's conformation. For example, there are two sequons next to cysteines that form the base of the V3 loop. A typical glycan is about 2000 Da, and the V3 loop is only about 3000 Da, so the presence or absence of these sequons would logically impact the orientation with which the V3 loop projects from the protein surface. Glycosylation sites may be preserved *in vivo* to mask neutralizing antibody sites, but replication efficiency and access to receptor binding sites may provide a counterbalancing force—multiple glycosylation sites can be removed from gp120 without loss of infectivity (Ohgimoto *et al*., 1998).

The observation that the complex carbohydrates tend to be localized in the shifting positions while high mannose is

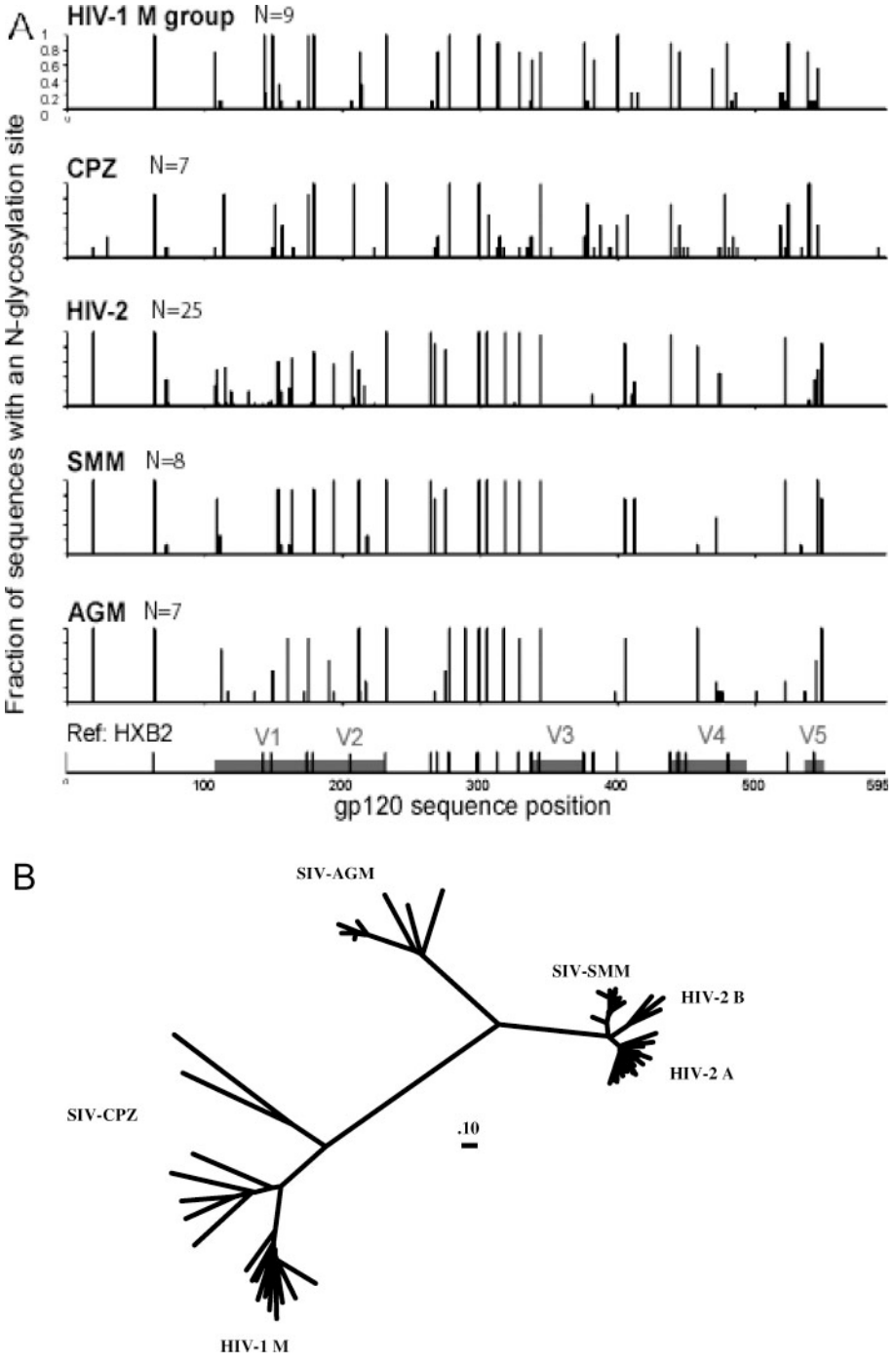


Fig. 8. (**A**) Sequon frequencies in different primate lentiviral lineages. Both shifting and fixed glycosylation site patterns are readily identified in the HIV-1 M group, CPZ, and HIV-2 sequences. Although only limited numbers of complete gp120s that represent distinct lineages of SIV_{SMM} and SIV_{AGM} are available, the shifting sites are not readily apparent in these viruses when infecting their natural hosts. Two exceptions are that shifting sequons are not evident V4 in the HIV-2, and are indicated in the SIV_{AGM} set. (**B**) Phylip PROTDIST tree of gp120 sequences (http://evolution.genetics.washington.edu/phylip.html) used for the analysis shown in **A**. This tree is simply meant to illustrate the extent of diversity in the population of protein sequences of these viruses in different hosts. Most of the specific sequences used to generate **A** and **B** of this figure are listed in Figure 4, although only a subset of the HIV-2 sequences are included in Table I to save space (18 are included here).

favored in the fixed positions in HIV-1 Env gp120 could result from multiple contributing factors. The interactions of DC-SIGN and other C-type lectins with gp120 high-mannose oligosaccharides have an important role in HIV infectivity (Lin *et al.*, 2003), suggesting there may be a fitness cost in disrupting their precise orientation. If the high-mannose oligosaccharides are important for successful sexual transmission, there would be selection pressure to

Table I. Relative spacing of unaligned V1 sequons to illustrate the level of diversity found in shifting sites

Sequence name	No. *N*-linked glycosylation sites in V1	No. sites in full gp120	Spacing of sequons in V1 loops
A) HIV-1: One patient over time			
1988	NA	NA	C....N..N.............NC
1994	NA	NA	C.....N..N.............NC
1994	NA	NA	C.....N..N..N..N..N.................NC
1994	NA	NA	C....N..N.............NC
1995	NA	NA	C....N.....................NC
1995	NA	NA	C....N.......N..N...........NC
1995	NA	NA	C.....N......N..........NC
B) HIV-1: A subtype			
HIV1_A_MA246	4	25	NC.....N.....N.........NC
HIV1_A_K89	3	24	NC...N..............NC
HIV1_A_KIG93	4	24	.C..N...N.........N..........NC
HIV1_A_SE8131	5	27	NC...N.........N....N.........NC
HIV1_A_SE6594	4	24	NC..N.....N........NC
HIV1_A_92UG037	4	27	.C..N...N.....N.........NC
HIV1_A_U455	5	23	.C.N...N...N..N.........NC.
C) HIV-1: M group			
HIV1_A_U455	5	23	.C.N...N...N..N.........NC.
HIV1_B_HXB2R	3	24	.C....N....N.............NC.
HIV1_C_ETH2220	3	20	.C....N.....N..........NC.
HIV1_D_84ZR085	5	25	NC.....N...NN...........NC.
HIV1_01_CM240	6	24	NC...N..N.....N......N......NC.
HIV1_F_VI850	4	25	NC.N..N.........NC.
HIV1_G_SE6165	7	29	NC.........NN...N....NN.........NC.
HIV1_H_90CF056	5	28	NC....NN..N...........NC.
HIV1_J_SE92809	5	24	NC.N.....N...N.........NC.
HIV1_K_MP535C	4	26	NC......N...N...........NC.
D) SIV_{CPZ}:			
CPZ_US	3	20	.C.....N....N......NC.
CPZ_GAB	4	22	.C...N.....N..N.........NC.
CPZ_CAM3	3	25	.C.....N.....N.................NC.
CPZ_CAM5	3	23	.C....N......N...........NC.
CPZ_GAB2	4	22	.C...N...N..N.............NC.
CPZ_ANT	2	20	NC...N.........................C.
CPZ_TAN1	4	25	.C...N..N.............N.........NC.
E) HIV-2			
HIV2_A_GH1	5	24	.CN...NN...........N........NC.
HIV2_A_CBL21	3	22	NC....N.................N..C.
HIV2_A_CAM3	2	26	.CN.................N....C.

Table I. Continued

Sequence name	No. *N*-linked glycosylation sites in V1	No. sites in full gp120	Spacing of sequons in V1 loops
HIV2_A_CBL23	3	24	.CN.....N.....N..C.
HIV2_A_CAM4	3	25	.C.....N.....N.....N....C.
HIV2_A_MDS	5	30	NC.....N.....N...N.....N.....C.
HIV2_A_FG	5	24	NC..N.....N....N.....NC.
HIV2_A_UC2	4	24	.C.....N..N.....N.....N....C.
HIV2_A_ALI	2	24	.C.N.....N....C.
HIV2_B_EHO	1	19	.CN.....C.
HIV2_B_D205	3	25	NC.....N.....N....C.
HIV2_B_UC1	3	23	.CN....N.....N....C.
HIV2_G_ABT96	2	23	.CN.....N....C.
F) SIV_{SMM}			
SMM_STM	2	22	.C..N.....N....C.
SMM_SL92B	2	23	.C..N.....N....C.
SMM_MM142	2	23	.CN.....N....C.
SMM_MNE027	2	22	.CN.....N..C.
SMM_MAC251	2	20	.CN.....N....C.
SMM_DeltaB670	2	23	.CN.....N....C.
SMM_F236	2	21	.CN.....N....C.
SMM_H9	2	21	.CN.....N....C.
G) SIV_{AGM}			
AGM_VER155	1	19	.C.....C.
AGM_VERAGM3	2	22	.C...N.....N.....C.
AGM_VER9063	2	22	.C...N....N.....C.
AGM_VERTYO	1	21	.C...N.....C.
AGM_SAB1C	2	20	.C.....N.....NC.
AGM_GRI677	1	21	.C...N.....C.
AGM_TAN1	1	20	NC.....C.

Only *N*-linked glycosylation sites (N) and Cys (C), which close the base of the V1 loop, are marked, all other amino acids are indicated by a period (.), to highlight the relative change of position of the *N*-linked sites and the length variation of the V1 loop. Parts C–G correspond to the sequences used in Figure 8A. Part A includes examples of V1 sequences taken from a single patient over time in a study of long-term survival (patient 20 from Shioda *et al.*, 1997), who was randomly selected from among longitudinal sequences sets spanning V1 available in the HIV database, and in our experience is representative. Part B includes seven randomly selected subtype A sequences from the HIV Los Alamos database alignment to represent intraclade diversity, respectively. Part C includes one randomly selected sequence from each clade to represent HIV-1 M group diversity. Part D includes available viral sequences isolated from chimpanzees, thought to be the source of the human HIV-1 epidemic. Part E lists viruses from HIV-2 clades A and B, and part F from sooty mangabey, which are thought to be the source of the HIV-1 epitdemic. Part G represents viral sequences taken from African green monkeys.

preserve the high-mannose forms. It has been suggested that the conservation of the high-mannose epitope for the 2G12 antibody might be related to the preservation of mannose structure that facilitates DC-SIGN interaction (Sanders *et al.*, 2002). *N*-linked glycosylation patterns in HIV Env are further complicated by the fact that different cell types, H9 and Chinese hamster ovary (Mizuochi *et al.*, 1988, 1990), and primary T cells and macrophages have different patterns of glycan modification (Liedtke *et al.*, 1997; Lin *et al.*, 2003; Willey *et al.*, 1996). DC-SIGN preferentially binds to HIV Env gp120 enriched for high-mannose oligosaccharides, which are typically produced

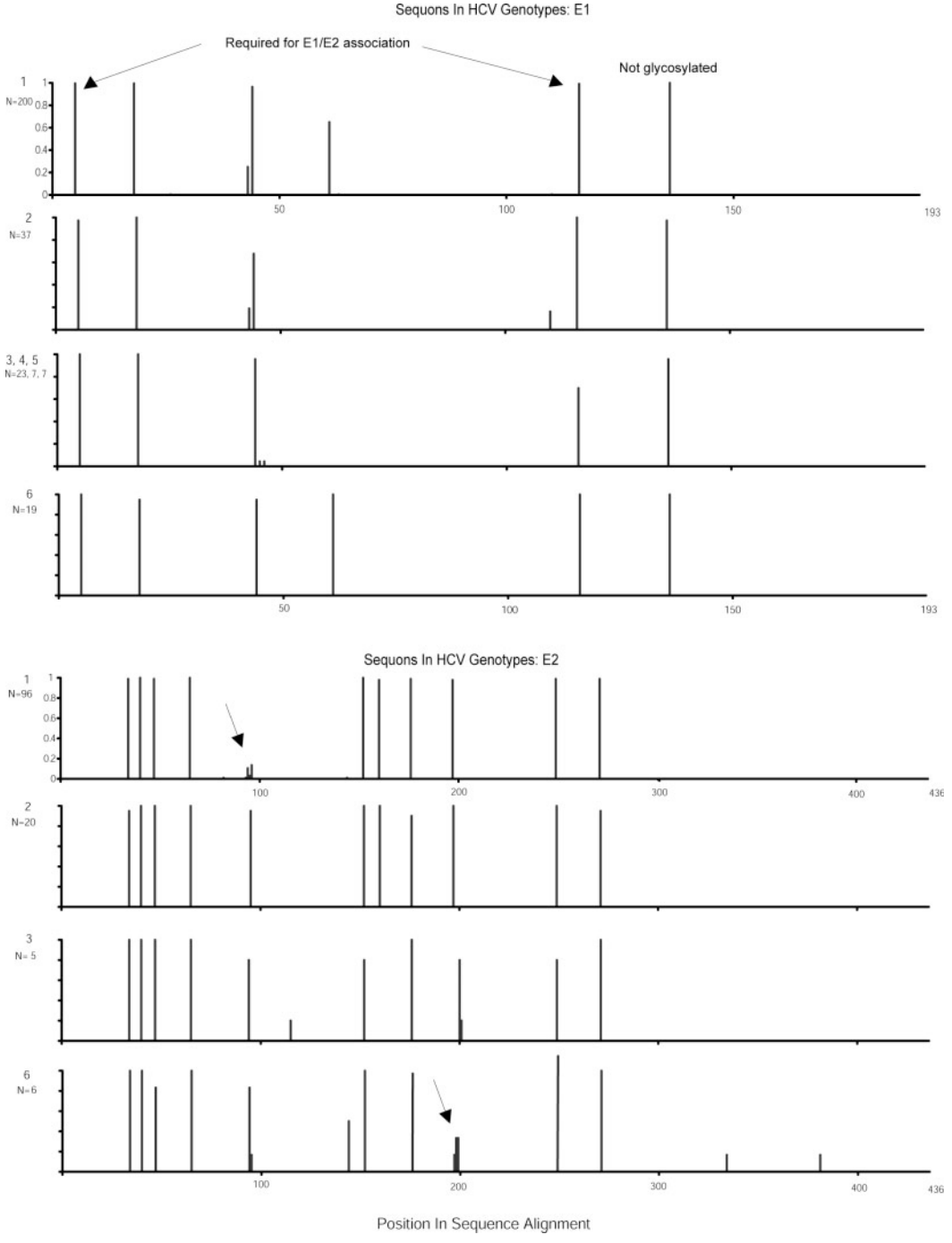


Fig. 9. *N*-linked glycosylation patterns in the E1 and E2 proteins of HCV. Sequon frequencies in HCV genotypes with adequate sampling are shown. The plot was nearly identical for E1 genotypes 3, 4, and 5, so only the genotype 3 plot is shown. Potential shifting sites are seen in E2 near position 90 in genotype 1, and near position 200 in genotype 6. The most C-terminal sequon in E1 is not glycosylated, and the sequon second from the C-terminal end and the most *N*-terminal together are critical for the formation of noncovalent E1E2 complexes (Meunier *et al*., 1999).

by peripheral blood mononuclear cells and T cells, compared to macrophage-produced gp120, which contains more complex carbohydrates (Lin *et al*., 2003). Also, macrophage-derived gp120 carbohydrates are modified by lactosaminoglycans, whereas peripheral blood mononuclear cells–derived gp120s are not. Interestingly, macrophage-derived virus tends to be more neutralization resistant (Willey *et al*., 1996). Bisecting *N*-glycans have been implicated in the suppression of natural killer (NK) cell–mediated responses (Yoshimura *et al*., 1996), the suppression of innate immune responses involving NK cells that may be responsible for initiating AIDS pathogenesis (Kottilil *et al*., 2003). Thus cell-specific glycosylation profiles may affect immune susceptibility, and AIDS progression may be related to the glycobiology of the virus (Clark *et al*., 1997).

Positions that accept complex carbohydrate additions may also be determined simply by being more exposed during passage through the Golgi apparatus, where carbohydrate modifications occur. Such exposure may be related to accessibility in the folded functional protein, and immune evasion and neutralizing antibody escape are likely to be the driving force for the variation in shifting sites. So the complex carbohydrate additions may occur in more exposed sites, and those sites, due to the exposure, may be under greater pressure to shift to escape from antibody recognition.

We did not observe variation in the net number of sequons in the variable loop domains of HIV-1 Env gp120 associated with any particular pattern of coreceptor usage, although a particular sequon in the V3 loop was highly conserved in R5 viruses but not in X4 or R5X4 viruses. Potential involvement of sequon changes in gp41 in cell tropism was also noted, as some sites are rarely found in NSI variants. Glycosylation in gp41 may be more influential than previously thought, given new evidence that additional loops of the transmembrane protein are located extracellularly (Cleveland *et al*., 2003). The role of alterations of specific sequons in HIV-1 Env associated with changes in coreceptor usage on a population basis can be subtle and may be context-specific. For example, in our analysis of one patient for changes associated with phenotypic variation (Figure 7), significant differences were found at positions that were different from those identified in the cross-sectional population analyses.

Glycosylation in the V1V2 region can be important for coreceptor usage (Ogert *et al*., 2001), and additional glycosylation sites in V1V2 may in some circumstances potentiate the use of CXCR4 (Pollakis *et al*., 2001). V2 elongation and sequon changes have been associated with slow disease progression (Shioda *et al*., 1997). The effects may be subtle; for example, removal of three sequons in V1 increased the affinity between gp120 and the CXCR4 receptor but did not alter the infectivity of the virus (Losman *et al*., 2001). Limited V1V2 length variation and sequon shifts in rapid progressors (Masciotra *et al*., 2002; Shioda *et al*., 1997) versus long-term survivors may be a consequence of a poor immune response resulting in weak selection pressure (Delwart *et al*., 1997; Wolinsky *et al*., 1996) and not related directly to the coreceptor usage. This is particularly plausible because changes in V1 V2 sequons often influence antigenic domains in other regions, for example, they can alter antibody recognition of both the V3 loop (Losman *et al*., 2001; Ly and Stamatatos, 2000; Ye *et al*., 2000) and CD4 binding site (Ly and Stamatatos, 2000). It is intriguing that the capacity for shifting sequons is found not only in HIV but also in two HCV E2 locations in two lineages in genotypes 1 and 6, suggesting they may give a selective advantage in rapidly evolving viruses.

Although the SIV_{SMM} and SIV_{AGM} viruses have some degree of shifting sequons in their natural hosts (particularly in the Env V4 region in SIV_{AGM}), the shifting Env sequon characteristics appear far less pronounced in these lineages than in HIV-1 and CPZ lineages (Figure 8). HIV-2, which stems from cross-species transmission of SIV_{SMM}, also has more extreme levels of shifting sites, suggesting that selective forces in the new host bring out the greater levels of position diversity seen in sequons. Neutralizing antibody responses to SIV_{SMM} and SIV_{AGM} infections in their natural hosts are present but may be relatively reduced (Fultz *et al*., 1990; Gicheru *et al*., 1999; Kaur *et al*., 1998; Norley *et al*., 1990). However, despite the lack of shifting sequons, both SIV_{SMM} and SIV_{AGM} diversify rapidly *in vivo* (Broussard *et al*., 2001). At least some of this diversification may be due to cytotoxic T lymphocyte escape (Kaur *et al*., 2001), which may not select for patterns of shifting sequons. Neutralizing activity of sera from HIV-infected individuals and SIV-infected primates generally lags behind, so serum from one time point can neutralize earlier but not contemporary virus (Albert *et al*., 1990; Arendrup *et al*., 1992; Bradney *et al*., 1999; Montefiori *et al*., 1991; Nyambi *et al*., 1997). Rapid cycles of response and escape may be related to the gain and loss of sequons (Richman *et al*., 2003; Wei *et al*., 2003). Thus it is possible that the strength and neutralizing antibody response in the host dictates the extent of the shifting antibody sites in different primate lentiviruses.

N-linked glycosylation sites are a critical component of the external proteins of primate lentiviruses, influenza, and hepatitis C viruses, and their modification can be important for evolution of the immune response. The gain or loss of such sites can play a key role in viral infectivity, antigen conformation, and immune escape. The mechanisms that generate shifting sites and tolerance of such shifting sequons in viral proteins provides a unique evolutionary avenue for immune evasion.

Materials and methods

A Web-based tool was developed for tracking and quickly assessing patterns in *N*-linked glycosylation sites in protein alignments (www.hiv.lanl.gov/content/hiv-db/GLYCOSITE/glycosite.html). This facilitates comparing glycosylation patterns by providing five different summaries:

i. A simple tally of the number of sequons in each protein in an alignment;
ii. Red highlighted *N*-linked glycosylation sites in the alignment, and a downloadable fasta-format text file that leaves the N in sequons uppercase while reducing all other amino acids to lowercase;
iii. Plots of the fraction of sequences in each alignment that carry an *N*-linked glycosylation site at each position;
iv. A summary of the average number of sites within a user-specified window size, and a break down for each sequence;
v. A list of sequons and their context in each sequence, providing the amino acid string of the sequons, as not all sequons have the same capacity to be glycosylated (for example, *N*-P-[ST] is not glycosylated), and such sites were not observed among the viral sequences studied here. We also include links to references defining glycosylation propensities of different local combinations of amino acids.

All features of the program were used to analyze the sequence sets included in this study, but most of the figures were made using feature iii.

The protein alignments used for this study were retrieved from the Los Alamos HIV sequence database (www.hiv.lanl.gov), Influenza database (www.flu.lanl.gov), and the Hepatitis C database (www.hcv.lanl.gov). All alignments used in this study are available on request. All hepatitis, SIV, and HIV alignments were restricted so that only sequence from a single individual was included. Thus the sample sets were not biased by including multiple sequences from one person or from closely related infections. The one exception is the sequon change analysis of single patient in Figure 7 (Hu *et al*., 2000).

Acknowledgments

We thank Drs. John Moore and Julie Overbaugh for helpful discussions and Dennis Bruton for asking the questions lead to this study. This work was supported by the NIH-DOE interagency agreement for the HIV-1 sequence and immunology database project YI-AI-1500-03 (M.Z., B.G., B.F., C.K., B.K.), a Los Alamos National Laboratory research development award (B.K. and B.G.), and NIH-PHS grants AI-26503 (N.H.) and AI-34251 (N.H.). W.B. is supported by a University of Washington Center for AIDS Research fellowship (AI-27757).

Abbreviations

DC-SIGN, dendritic cell–specific HIV-1-binding protein; HA1, influenza A hemagglutinin 1; HCV, hepatitis C virus; HIV human immunodeficiency virus; NK: natural killer cell; NSI, nonsyncytium-inducing; CRF, circulating recombinant form; SI, syncytium-inducing; SIV, simian immunodeficiency virus.

References

Albert, J., Abrahamsson, B., Nagy, K., Aurelius, E., Gaines, H., Nystrom, G., and Fenyo, E.M. (1990) Rapid development of isolate-specific neutralizing antibodies after primary HIV-1 infection and consequent emergence of virus variants which resist neutralization by autologous sera. *AIDS*, **4**, 107–112.

Alexander, S. and Elder, J.H. (1984) Carbohydrate dramatically influences immune reactivity of antisera to viral glycoprotein antigens. *Science*, **226**, 1328–1330.

Arendrup, M., Nielsen, C., Hansen, J.E., Pedersen, C., Mathiesen, L., and Nielsen, J.O. (1992) Autologous HIV-1 neutralizing antibodies: emergence of neutralization-resistant escape virus and subsequent development of escape virus neutralizing antibodies. *J. AIDS*, **5**, 303–307.

Bolmstedt, A., Sjolander, S., Hansen, J.E., Akerblom, L., Hemming, A., Hu, S.L., Morein, B., and Olofsson, S. (1996) Influence of *N*-linked glycans in V4–V5 region of human immunodeficiency virus type 1 glycoprotein gp160 on induction of a virus-neutralizing humoral response. *J. AIDS Hum. Retrovirol.*, **12**, 213–220.

Bosch, M.L., Andeweg, A.C., Schipper, R., and Kenter, M. (1994) Insertion of *N*-linked glycosylation sites in the variable regions of the human immunodeficiency virus type 1 surface glycoprotein through AAT triplet reiteration. *J. Virol.*, **68**, 7566–7569.

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

Bradney, A.P., Scheer, S., Crawford, J.M., Buchbinder, S.P., and Montefiori, D.C. (1999) Neutralization escape in human immunodeficiency virus type 1–infected long-term nonprogressors. *J. Infect. Dis.*, **179**, 1264–1267.

Broussard, S.R., Staprans, S.I., White, R., Whitehead, E.M., Feinberg, M.B., and Allan, J.S. (2001) Simian immunodeficiency virus replicates to high levels in naturally infected African green monkeys without inducing immunologic or neurologic disease. *J. Virol.*, **75**, 2262–2275.

Bush, R.M., Fitch, W.M., Bender, C.A., and Cox, N.J. (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.*, **16**, 1457–1465.

Chackerian, B., Morton, W.R., and Overbaugh, J. (1994) Persistence of simian immunodeficiency virus Mne variants upon transmission. *J. Virol.*, **68**, 4080–4085.

Chackerian, B., Rudensey, L.M., and Overbaugh, J. (1997) Specific *N*-linked and O-linked glycosylation modifications in the envelope V1 domain of simian immunodeficiency virus variants that evolve in the host alter recognition by neutralizing antibodies. *J. Virol.*, **71**, 7719–7727.

Cheng-Mayer, C., Brown, A., Harouse, J., Luciw, P.A., and Mayer, A.J. (1999) Selection for neutralization resistance of the simian/human immunodeficiency virus SHIVSF33A variant *in vivo* by virtue of sequence changes in the extracellular envelope glycoprotein that modify *N*-linked glycosylation. *J. Virol.*, **73**, 5294–5300.

Cho, M.W., Lee, M.K., Carney, M.C., Berson, J.F., Doms, R.W., and Martin, M.A. (1998) Identification of determinants on a dualtropic human immunodeficiency virus type 1 envelope glycoprotein that confer usage of CXCR4. *J. Virol.*, **72**, 2509–2515.

Clark, G.F., Dell, A., Morris, H.R., Patankar, M., Oehninger,S., and Seppala, M. (1997) Viewing AIDS from a glycobiological perspective: potential linkages to the human fetoembryonic defence system hypothesis. *Mol. Hum. Reprod.*, **3**, 5–13.

Cleveland, S.M., McLain, L., Cheung, L., Jones, T.D., Hollier, M., and Dimmock, N.J. (2003) A region of the C-terminal tail of the gp41 envelope glycoprotein of human immunodeficiency virus type 1 contains a neutralizing epitope: evidence for its exposure on the surface of the virion. *J. Gen. Virol.*, **84**, 591–602.

Davis, D., Stephens, D.M., Willers, C., and Lachmann, P.J. (1990) Glycosylation governs the binding of antipeptide antibodies to regions of hypervariable amino acid sequence within recombinant gp120 of human immunodeficiency virus type 1. *J. Gen. Virol.*, **71**(12), 2889–2898.

Deleersnyder, J., Pillez, A., Wychowski, C., Blight, K., Xu, J., Hahn, Y.S., Rice, C.M., and Dubuission, J. (1997) Formation of native hepatitis C virus glycoprotein complexes. *J. Virol.*, **71**, 697–704.

Delwart, E.L., Pan, H., Sheppard, H.W., Wolpert, D., Neumann, A.U., Korber, B., and Mullins, J.I. (1997) Slower evolution of human immunodeficiency virus type 1 quasispecies during progression to AIDS. *J. Virol.*, **71**, 7498–7508.

Dubuisson, J., Duvet, S., Meunier, J.C., De Beeck, O., Cacan, R., Wychowski, C., and Cocquerel, L. (2000) Glycosylation of the hepatitis C virus envelope protein E1 is dependent on the presence of a downstream sequence on the viral polyprotein. *J. Biol. Chem.*, **275**, 30605–30609.

Ferguson, N.M., Galvani, A.P., and Bush, R.M. (2003) Ecological and immunological determinants of influenza evolution. *Nature*, **422**, 428–433.

Ferris, R.L., Hall, C., Sipsas, N.V., Safrit, J.T., Trocha, A., Koup, R.A., Johnson, R.P., and Siliciano, R.F. (1999) Processing of HIV-1 envelope glycoprotein for class I–restricted recognition: dependence on TAP1/2 and mechanisms for cytosolic localization. *J. Immunol.*, **162**, 1324–1332.

Fournillier, A., Wychowski, C., Boucreux, D., Baumert, T.F., Meunier, J.C., Jacobs, D., Muguet, S., Depla, E., and Inchauspe, G. (2001) Induction of hepatitis C virus E1 envelope protein–specific immune response can be enhanced by mutation of *N*-glycosylation sites. *J. Virol.*, **75**, 12088–12097.

Fultz, P.N., Stricker, R.B., McClure, H.M., Anderson, D.C., Switzer, W.M., and Horaist, C. (1990) Humoral response to SIV/SMM infection in macaque and mangabey monkeys. *J. AIDS*, **3**, 319–329.

Gao, F., Morrison, S.G., Robertson, D.L., Thornton, C.L., Craig, S., Karlsson, G., Sodroski, J., Morgado, M., Galvao-Castro, B.,

von Briesen, H. and others. (1996) Molecular cloning and analysis of functional envelope genes from human immunodeficiency virus type 1 sequence subtypes A through G. The WHO and NIAID networks for HIV isolation and characterization. *J. Virol.*, **70**, 1651–1667.

Gavel, Y. and von Heijne, G. (1990) Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. *Protein Eng.*, **3**, 433–442.

Gicheru, M.M., Otsyula, M., Spearman, P., Graham, B.S., Miller, C.J., Robinson, H.L., Haigwood, N.L., and Montefiori, D.C. (1999) Neutralizing antibody responses in Africa green monkeys naturally infected with simian immunodeficiency virus (SIVagm). *J. Med. Primatol.*, **28**, 97–104.

Gram, G.J., Bolmstead, A., Schonning, K., Biller, M., Hansen, J.E., Olofasson, S. (2002) Detection of orientation-specific anti-gp120 antibodies by a new *N*-glycanase protection assay. *APMIS*, **110**, 123–31.

Hebert, D.N., Zhang, J.X., Chen, W., Foellmer, B., and Helenius, A. (1997) The number and location of glycans on influenza hemagglutinin determine folding and association with calnexin and calreticulin. *J. Cell Bio.*, **139**, 613–623.

Hirsch, V.M., Olmsted, R.A., Murphey-Corb, M., Purcell, R.H., and Johnson, P.R. (1989) An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature*, **339**, 389–392.

Hoffman, T.L., Stephens, E.B., Narayan, O., and Doms, R.W. (1998) HIV type I envelope determinants for use of the CCR2b, CCR3, STRL33, and APJ coreceptors. *Proc. Natl Acad. Sci. USA*, **95**, 11360–11365.

Hu, Q.X., Barry, A.P., Wang, Z.X., Connolly, S.M., Peiper, S.C., and Greenberg, M.L. (2000) Evolution of the human immunodeficiency virus type 1 envelope during infection reveals molecular corollaries of specificity for coreceptor itilization and AIDS pathogenesis. *J. Virol.*, **74**, 11858–11872.

Kalish, M.L., Korber, B.T., Pillai, S., Robbins, K.E., Leo, Y.S., Saekhou, A., Verghese, I., Gerrish, P., Goh, C.L., Lupo, D., Tan, B.H., Brown, T.M., and Chan, R. (2002) The sequential introduction of HIV-1 subtype B and CRF01AE in Singapore by sexual transmission: accelerated V3 region evolution in a subpopulation of Asian CRF01 viruses. *Virology*, **304**, 311–329.

Kasturi, L., Eshleman, J.R., Wunner, W.H., and Shakin-Eshleman, S.H. (1995) The hydroxy amino acid in an Asn-X-Ser/Thr sequon can influence *N*-linked core glycosylation efficiency and the level of expression of a cell surface glycoprotein. *J. Biol. Chem.*, **270**, 14756–14761.

Kaur, A., Grant, R.M., Means, R.E., McClure, H., Feinberg, M., and Johnson, R.P. (1998) Diverse host responses and outcomes following simian immunodeficiency virus SIVmac239 infection in sooty mangabeys and Rhesus macaques. *J. Virol.*, **72**, 9597–9611.

Kaur, A., Alexander, L., Staprans, S.I., Denekamp, L., Hale, C.L., McClure, H.M., Feinberg, M.B., Desrosiers, R.C., and Johnson, R.P. (2001) Emergence of cytotoxic T lymphocyte escape mutations in nonpathogenic simian immunodeficiency virus infection. *Eur. J. Immunol.*, **31**, 3207–3217.

Kaverin, N.V., Rudneva, I.A., Ilyushina, N.A., Varich, N.L., Lipatov, A.S., Smirnov, Y.A., Govorkova, E.A., Gitelman, A.K., Lvov, D.K., and Webster, R.G. (2002) Structure of antigenic sites on the haemagglutinin molecule of H5 avian influenza virus and phenotypic variation of escape mutants. *J. Gen. Virol.*, **83**, 2497–2505.

Koito, A., Stamatatos, L., and Cheng-Mayer, C. (1995) Small amino acid sequence changes within the V2 domain can affect the function of a T-cell line-tropic human immunodeficiency virus type 1 envelope gp120. *Virology*, **206**, 878–884.

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B.H., Wolinsky, S., and Bhattacharya, T. (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science*, **288**, 1789–1796.

Korber, B., Gaschen, B., Yusim, K., Thakallapally, R., Kesmir, C., and Detours, V. (2001) Evolutionary and immunological implications of gontemporary HIV-1 variation. *Br. Med. Bull.*, **58**, 19–42.

Kottilil, S., Chun, T.W., Moir, S., Liu, S., McLaughlin, M., Hallahan, C.W., Maldarelli, F., Corey, L., and Fauci, A.S. (2003) Innate immunity in human immunodeficiency virus infection: effect of viremia on natural killer cell function. *J. Infect. Dis.*, **187**, 1038–1045.

Kuiken, C., Foley, B., Guzman E., and Korber, B. (1999) Determinants of HIV-1 protein evolution. In Crandall, K. (Ed.), *Molecular evolution of HIV*. Johns Hopkins University Press, Baltimore, MD.

Land, A. and Braakman, I. (2001) Folding of the human immunodeficiency virus type 1 envelope glycoprotein in the endoplasmic reticulum. *Biochimie*, **83**, 783–790.

Leonard, C.K., Spellman, M.W., Riddle, L., Harris, R.J., Thomas, J.N., and Gregory, T.J. (1990) Assignment of intrachain disulfide bonds and characterization of potential glycosylation sites of the type 1 recombinant human immunodeficiency virus envelope glycoprotein (gp120) expressed in Chinese hamster ovary cells. *J. Biol. Chem.*, **265**, 10373–10382.

Liedtke, S., Geyer,R., and Geyer, H. (1997) Host-cell-specific glycosylation of HIV-2 envelope glycoprotein. *Glycoconj. J.*, **14**, 785–793.

Lin, G., Simmons, G., Pohlmann, S., Baribaud, F., Ni, H., Leslie, G.J., Haggarty, B.S., Bates, P., Weissman, D., Hoxie, J.A., and Doms, R.W. (2003) Differential *N*-linked glycosylation of human immunodeficiency virus and Ebola virus envelope glycoproteins modulates interactions with DC-SIGN and DC-SIGNR. *J. Virol.*, **77**, 1337–1346.

Losman, B., Bolmstedt, A., Schonning, K., Bjorndal, A., Westin, C., Fenyo, E.M., and Olofsson, S. (2001) Protection of neutralization epitopes in the V3 loop of oligomeric human immunodeficiency virus type 1 glycoprotein 120 by *N*-linked oligosaccharides in the V1 region. *AIDS Res. Hum. Retroviruses*, **17**, 1067–1076.

Ly, A. and Stamatatos, L. (2000) V2 loop glycosylation of the human immunodeficiency virus type 1 SF162 envelope facilitates interaction of this protein with CD4 and CCR5 receptors and protects the virus from neutralization by anti-V3 loop and anti-CD4 binding site antibodies. *J. Virol.*, **74**, 6769–6776.

Malenbaum, S.E., Yang, D., Cavacini, L., Posner, M., Robinson, J., and Cheng-Mayer, C. (2000) The *N*-terminal V3 loop glycan modulates the interaction of clade A and B human immunodeficiency virus type 1 envelopes with CD4 and chemokine receptors. *J. Virol.*, **74**, 11008–11016.

Marshall, R.D. (1974) The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. *Biochem. Soc. Symp.*, 17–26.

Masciotra, S., Owen, S.M., Rudolph, D., Yang, C., Wang, B., Saksena, N., Spira, T., Dhawan, S., and Lal, R.B. (2002) Temporal relationship between V1V2 variation, macrophage replication, and coreceptor adaptation during HIV-1 disease progression. *AIDS*, **16**, 1887–1898.

Matrosovich, M., Zhou, N., Kawaoka, Y., and Webster, R. (1999) The surface glycoproteins of H5 influenza viruses isolated from humans, chickens, and wild aquatic birds have distinguishable properties. *J. Virol.*, **73**, 1146–1155.

Matthews, T.J., Weinhold, K.J., Lyerly, H.K., Langlois, A.J., Wigzell, H., and Bolognesi, D.P. (1987) Interaction between the human T-cell lymphotropic virus type IIIB envelope glycoprotein gp120 and the surface antigen CD4: role of carbohydrate in binding and cell fusion. *Proc. Natl Acad. Sci. USA*, **84**, 5424–5428.

Mellquist, J.L., Kasturi, L., Spitalnik, S.L., and Shakin-Eshleman, S.H. (1998) The amino acid following an Asn-X-Ser/Thr sequon is an important determinant of *N*-linked core glycosylation efficiency. *Biochemistry*, **37**, 6833–6837.

Meunier, J.C., Fournillier, A., Choukhi, A., Cahour, A., Cocquerel, L., Dubuisson, J., and Wychowski, C. (1999b) Analysis of the glycosylation sites of hepatitis C virus (HCV) glycoprotein E1 and the influence of E1 glycans on the formation of the HCV glycoprotein complex. *J. Gen. Virol.*, **80**(4), 887–896.

Mizuochi, T., Spellman, M.W., Larkin, M., Solomon, J., Basa, L.J., and Feizi, T. (1988) Carbohydrate structures of the human-immunodeficiency-virus (HIV) recombinant envelope glycoprotein gp120 produced in Chinese-hamster ovary cells. *Biochem. J.*, **254**, 599–603.

Mizuochi, T., Matthews, T.J., Kato, M., Hamako, J., Titani, K., Solomon, J., and Feizi, T. (1990) Diversity of oligosaccharide structures on the envelope gycoprotein gp 120 of human immunodeficiency virus 1 from the lymphoblastoid cell line H9. Presence of complex-type oligosaccharides with bisecting *N*-acetylglucosamine residues. *J. Biol. Chem.*, **265**, 8519–8524.

Montefiori, D.C., Zhou, I.Y., Barnes, B., Lake, D., Hersh, E.M., Masuho, Y., and Lefkowitz, L.B. Jr. (1991) Homotypic antibody

responses to fresh clinical isolates of human immunodeficiency virus. *Virology*, **182**, 635–643.

Moore, J.P. and Sodroski, J. (1996) Antibody cross-competition analysis of the human immunodeficiency virus type 1 gp120 exterior envelope glycoprotein. *J. Virol.*, **70**, 1863–1872.

Myers, G., MacInnes, K., and Korber, B. (1992) The emergence of simian/human immunodeficiency viruses. *AIDS Res. Hum. Retroviruses*, **8**, 373–386.

Norley, S.G., Kraus, G., Ennen, J., Bonilla, J., Konig, H., and Kurth, R. (1990) Immunological studies of the basis for the apathogenicity of simian immunodeficiency virus from African green monkeys. *Proc. Natl Acad. Sci. USA*, **87**, 9067–9071.

Norley, S., Beer, B., Holzammer, S., zur Megede, J., and Kurth, R. (1999) Why are the natural hosts of SIV resistant to AIDS? *Immunol. Lett.*, **66**, 47–52.

Nyambi, P.N., Lewi, P., Peeters, M., Janssens, W., Heyndrickx, L., Fransen, K., Andries, K., Vanden Haesevelde, M., Heeney, J., Piot, P., and van der Groen, G. (1997) Study of the dynamics of neutralization escape mutants in a chimpanzee naturally infected with the simian immunodeficiency virus SIVcpz-Ant. *J. Virol.*, **71**, 2320–2330.

Ogert, R.A., Lee, M.K., Ross, W., Buckler-White, A., Martin, M.A., and Cho, M.W. (2001) *N*-linked glycosylation sites adjacent to and within the V1/V2 and the V3 loops of dualtropic human immunodeficiency virus type 1 isolate DH12 gp120 affect coreceptor usage and cellular tropism. *J. Virol.*, **75**, 5998–6006.

Ohgimoto, S., Shioda, T., Mori, K., Nakayama, E.E., Hu, H., and Nagai, Y. (1998) Location-specific, unequal contribution of the N glycans in simian immunodeficiency virus gp120 to viral infectivity and removal of multiple glycans without disturbing infectivity. *J. Virol.*, **72**, 8365–8370.

Ohta, Y., Masuda, T., Tsujimoto, H., Ishikawa, K., Kodama, T., Morikawa, S., Nakai, M., Honjo, S., and Hayami, M. (1988) Isolation of simian immunodeficiency virus from African green monkeys and seroepidemiologic survey of the virus in various non-human primates. *Int. J. Cancer*, **41**, 115–122.

Overbaugh, J. and Rudensey, L.M. (1992) Alterations in potential sites for glycosylation predominate during evolution of the simian immunodeficiency virus envelope gene in macaques. *J. Virol.*, **66**, 5937–5948.

Overbaugh, J., Rudensey, L.M., Papenhausen, M.D., Benveniste, R.E., and Morton, W.R. (1991) Variation in simian immunodeficiency virus Env is confined to V1 and V4 during progression to simian AIDS. *J. Virol.*, **65**, 7025–7031.

Pollakis, G., S., Kang, Kliphuis, A., Chalaby, M.I., Goudsmit, J., and Paxton, W.A. (2001) *N*-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization. *J. Biol. Chem.*, **276**, 13433–13441.

Polonoff, E., Machida, C.A., and Kabat, D. (1982) Glycosylation and intracellular transport of membrane glycoproteins encoded by murine leukemia viruses. Inhibition by amino acid analogues and by tunicamycin. *J. Biol. Chem.*, **257**, 14023–14028.

Ratner, L. (1992) Glucosidase inhibitors for treatment of HIV-1 infection. *AIDS Res. Hum. Retroviruses*, **8**, 165–173.

Rice, C.M. (1996) Flaviviridae: the viruses and their replication. In Fields, B.N., Knipe, D.M., and Howley, P.M. (Eds.), *Field's virology*, Lippincott-Raven Philidelphia, pp. 931–959.

Richman, D.D., Wrin, T., Little, S.J., and Petropoulos, C.J. (2003) Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proc. Natl Acad. Sci. USA*, **100**, 4144–4149.

Robbins, K.E., Lemey, P., Pybus, O.G., Jaffe, H.W., Youngpairoj, A.S., Brown, T.M., Salemi, M., Vandamme, A.M., and Kalish, M.L. (2003) U.S. human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J. Virol.*, **77**, 6359–6366.

Robertson, D.L., Anderson, J.P., Bradac, J.A., Carr, J.K., Foley, B., Funkhouser, R.K., Gao, F., Hahn, B.H., Kalish, M.L., Kuiken, C., and others. (2000) HIV-1 nomenclature proposal. *Science*, **288**, 55–56.

Sanders, R.W., Venturi, M., Schiffner, L., Kalyanaraman, R., Katinger, H., Lloyd, K.O., Kwong, P.D., and Moore, J.P. (2002) The mannose-dependent epitope for neutralizing antibody 2G12 on human immunodeficiency virus type 1 glycoprotein gp120. *J. Virol.*, **76**, 7293–7305.

Sato, K., Okamoto, H., Aihara, S., Hoshi, Y., Tanaka, T., and Mishiro, S. (1993) Demonstration of sugar moiety on the surface of hepatitis C virions recovered from the circulation of infected humans. *Virology*, **196**, 354–357.

Scanlan, C.N., Pantophlet, R., Wormald, M.R., Ollmann, S.E., Stanfield, R., Wilson, I.A., Katinger, H., Dwek, R.A., Rudd, P.M., and Burton, D.R. (2002) The broadly neutralizing anti-human immunodeficiency virus type 1 antibody 2G12 recognizes a cluster of alpha1 $\rightarrow$ 2 mannose residues on the outer face of gp120. *J. Virol.*, **76**, 7306–7321.

Selby, M., Erickson, A., Dong, C., Cooper, S., Parham, P., Houghton, M., and Walker, C.M. (1999) Hepatitis C virus envelope glycoprotein E1 originates in the endoplasmic reticulum and requires cytoplasmic processing for presentation by class I MHC molecules. *J. Immunol.*, **162**, 669–676.

Shakin-Eshleman, S.H., Spitalnik, S.L., and Kasturi, L. (1996) The amino acid at the X position of an Asn-X-Ser sequon is an important determinant of *N*-linked core-glycosylation efficiency. *J. Biol. Chem.*, **271**, 6363–6366.

Shioda, T., Oka, S., Xin, X., Liu, H., Harukuni, R., Kurotani, A., Fukushima, M., Hasan, M.K., Shiino, T., Takebe, Y. and others. (1997) *In vivo* sequence variability of human immunodeficiency virus type 1 envelope gp120: association of V2 extension with slow disease progression. *J. Virol.*, **71**, 4871–4881.

Si, Z., Cayabyab, M., and Sodroski, J. (2001) Envelope glycoprotein determinants of neutralization resistance in a simian-human immunodeficiency virus (SHIV-HXBc2P 3.2) derived by passage in monkeys. *J. Virol.*, **75**, 4208–4218.

Simmonds, P., Zhang, L.Q., McOmish, F., Balfe, P., Ludlam, C.A., and Brown, A.J. (1991) Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 Env sequences in plasma viral and lymphocyte-associated proviral populations *in vivo*: implications for models of HIV pathogenesis. *J. Virol.*, **65**, 6266–6276.

Skehel, J.J. and Wiley, D.C. (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu. Rev. Biochem.*, **69**, 531–569.

Slater-Handshy, T., Droll, D.A., Fan, X., Di Bisceglie, A.M., and Chambers, T.J. (2004) HCV E2 glycoprotein: mutagenesis of *N*-linked glycosylation sites and its effects on E2 expression and processing. *Virology*, **319**, 36–48.

Webby, R.J. and Webster, R.G. (2001) Emergence of influenza A viruses. *Phil. Trans. R. Soc. Lond B Biol. Sci.*, **356**, 1817–1828.

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

Wiley, D.C. and Skehel, J.J. (1987) The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annu. Rev. Biochem.*, **56**, 365–394.

Willey, R.L., Shibata, R., Freed, E.O., Cho, M.W., and Martin, M.A. (1996) Differential glycosylation, virion incorporation, and sensitivity to neutralizing antibodies of human immunodeficiency virus type 1 envelope produced from infected primary T-lymphocyte and macrophage cultures. *J. Virol.*, **70**, 6431–6436.

Wills, C., Farmer, A., and Myers, G. (1996) Rapid sequon evolution in human immunodeficiency virus type 1 relative to human immunodeficiency virus type 2. *AIDS Res. Hum. Retroviruses*, **12**, 1383–1384.

Wolinsky, S.M., Wike, C.M., Korber, B.T., Hutto, C., Parks, W.P., Rosenblum, L.L., Kunstman, K.J., Furtado, M.R., and Munoz, J.L. (1992) Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. *Science*, **255**, 1134–1137.

Wolinsky, S.M., Korber, B.T., Neumann, A.U., Daniels, M., Kunstman, K.J., Whetsell, A.J., Furtado, M.R., Cao, Y., Ho, D.D., and Safrit, J.T. (1996) Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. *Science*, **272**, 537–542.

Wyatt, R. and Sodroski, J. (1998) The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science*, **280**, 1884–1888.

Ye, Y., Si, Z.H., Moore, J.P., and Sodroski, J. (2000) Association of structural changes in the V2 and V3 loops of the gp120 envelope glycoprotein with acquisition of neutralization resistance in a simian-human immunodeficiency virus passaged *in vivo*. *J. Virol.*, **74**, 11955–11962.

Yoshimura, M., Ihara, Y., Ohnishi, A., Ijuhin, N., Nishiura, T., Kanakura, Y., Matsuzawa, Y., and Taniguchi, N. (1996) Bisecting *N*-acetylglucosamine on K562 cells suppresses natural killer cytotoxicity and promotes spleen colonization. *Cancer Res.*, **56**, 412–418.

Yusim, K., Peeters, M., Pybus, O.G., Bhattacharya, T., Delaporte, E., Mulanga, C., Muldoon, M., Theiler, J., and Korber, B. (2001) Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. *Phil. Trans. R. Soc. Lond. B Biol. Sci.*, **356**, 855–866.

Zhu, X., Borchers, C., Bienstock, R.J., and Tomer, K.B. (2000) Mass spectrometric characterization of the glycosylation pattern of HIV-gp120 expressed in CHO cells. *Biochemistry*, **39**, 11194–11204.

APPENDIX

Nucleic Acids Research, 2006, Vol. 34, Web Server issue ***W463–W465***
doi:10.1093/nar/gkl255

jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1

Ming Zhang[1,2], Anne-Kathrin Schultz[1], Charles Calef[2], Carla Kuiken[2], Thomas Leitner[2], Bette Korber[2,3], Burkhard Morgenstern[1] and Mario Stanke[1,*]

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

Received February 14, 2006; Revised March 2, 2006; Accepted March 30, 2006

ABSTRACT

Detecting recombinations in the genome sequence of human immunodeficiency virus (HIV-1) is crucial for epidemiological studies and for vaccine development. Herein, we present a web server for subtyping and localization of phylogenetic breakpoints in HIV-1. Our software is based on a jumping profile Hidden Markov Model (jpHMM), a probabilistic generalization of the jumping-alignment approach proposed by Spang *et al.* The input data for our server is a partial or complete genome sequence from HIV-1; our tool assigns regions of the input sequence to known subtypes of HIV-1 and predicts phylogenetic breakpoints. jpHMM is available online at http://jphmm.gobics.de/.

INTRODUCTION

Currently, more than 150 000 partial or complete HIV genome sequences are available in the central HIV database at *Los Alamos National Laboratory* (1); these data are crucial for the development of drugs against AIDS. Analysis of HIV sequence data is challenging, however, since HIV is among the most genetically variable organisms known and recombinations of different HIV subtypes are very common (2). HIV-1 is divided into three major phylogenetic groups, one of which—the M group—is responsible for the AIDS pandemic (3,4). This group is classified into ten subtypes, some of which are further divided into sub-subtypes. Accurate classification of HIV-1 subtypes and recombinants is of crucial importance for epidemiological monitoring and drug development. Therefore, a number of software tools have been developed to classify HIV genome sequences and to identify phylogenetic breakpoints and subtypes in recombinant strains (5,6).

We recently developed a HMM-based method to compare nucleic acid sequences to a given multiple alignment A of a sequence family S for which a classification into subclasses is available (7). We called this method jumping profile Hidden Markov Model (jpHMM) since our approach is a probabilistic generalization of the jumping-alignment (JALI) algorithm proposed by Spang *et al.* (8,9). In JALI, a query sequence s is aligned to a multiple alignment A of a sequence family $S = \{s_1, \ldots, s_n\}$—but s is not aligned to the alignment A as a whole, but different parts of s can be aligned to different individual sequences s_i from A.

Within an alignment of the query s to the sequence family S, 'jumps' are allowed between different sequences from S depending on where the strongest degree of similarity is found. For a jump between two sequences s_i and s_j, a penalty is imposed, similar to the familiar gap penalty used in standard sequence alignment. This approach is particularly useful if the query sequence s is a result of phylogenetic recombinations such that different parts of s are related to different sequences from the family S. JALI has been shown to perform well if an alignment A is to be searched against a sequence database (9).

In our jpHMM approach, we assume that a partition of the sequences from the family S into subclasses is given. Each subclass is modeled as a profile Hidden Markov Model (10). *Within* a subclass, the usual transitions between match, insert and delete states are possible, as in standard profile HMM theory—but in addition, our model allows transitions between profile HMMs corresponding to different subclasses, so a path through our model can switch back and forth between different subclasses. Jumps between subclasses are associated with so-called jump probabilities. A detailed description of this approach is given in Schultz *et al.* (7).

PREDICTION OF PHYLOGENETIC RECOMBINATION POINTS IN HIV-1 AT GOBICS

In (7), we found that jpHMM is a useful tool to predict phylogenetic breakpoints and subtypes in recombinant HIV and hepatitis C sequences (11). For HIV subtyping, we start with a pre-calculated multiple alignment of HIV-1 genome sequences consisting of all major subtypes and sub-subtypes; these (sub-)subtypes are modeled as profile HMMs in our

*To whom correspondence should be addressed. Tel: +1 831 459 5232; Fax: +1 831 459 1809

jpHMM approach. It turned out that 'jumps' between these (sub-)subtypes correspond quite well to known phylogenetic breakpoints and (sub-)subtypes to which a query sequence *s* is aligned, reliably indicate the real (sub-)subtypes in recombinant HIV sequences. To evaluate our tool and to compare its prediction accuracy to competing methods such as Simplot (12) and *RDP* (13), we used a large set of real and simulated data from HIV-1 and hepatitis C. These test runs demonstrated that jpHMM is far more accurate than existing tools for phylogenetic breakpoint detection. Details of this program evaluation are described in (7).

To make jpHMM available to the HIV research community, we set up an easy-to-use WWW interface at *Göttingen Bioinformatics Compute Server (GOBICS)*: http://jphmm.gobics.de/ At our server, the user can paste or upload up to 5 full-length HIV-1 genome sequences that is to be searched for phylogenetic breakpoints and subtypes. Our server uses a pre-calculated multiple alignment of 309 HIV sequences from the major HIV (sub-)subtypes obtained from the HIV database at http://hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html. These sequences include nine subtypes *A–D, F, G, H, J, K*, and a persumed recombinant 01_*AE*. Subtype *A* has two sub-subtypes, *A*1 and *A*2; similarly *F* has two sub-subtypes, *F*1 and *F*2. *B* and *D* could be regarded as sub-subtypes because their relative distance and relation are similar to *A*1 and *A*2, *F*1 and *F*2, respectively. But we still consider *B* and *D* as subtypes, not sub-subtypes because of historical reasons (14).

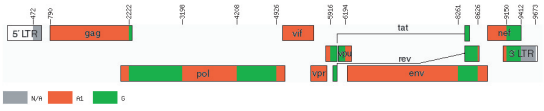
jpHMM result:

Sequence #1: 02_AG.NG.x.IBNG

This sequence is related to subtype: **A1 G** .

Fragment Start Position	Fragment End Position	Fragment Subtype
Position in the original sequence [text]		
1	329	N/A
330	1743	A1
1744	2722	G
2723	3732	A1
3733	4450	G
4451	5439	A1
5440	5726	G
5727	7766	A1
7767	8152	G
8153	8676	A1
8677	8938	G
8939	9201	N/A
Position based on HXB2 numbering [text]		
472	789	N/A
790	2221	A1
2222	3197	G
3198	4207	A1
4208	4925	G
4926	5915	A1
5916	6193	G
6194	8260	A1
8261	8625	G
8626	9149	A1
9150	9411	G
9412	9673	N/A

Genome map (based on HXB2 numbering)



Note:

- Numbers in the figure are breakpoint positions based on HXB2 numbering.
- The uncolored regions denote missing information due to input fragment sequence.
- The gray regions denote missing infomation due to uninformative subtype models (subtype: N/A).
- The sequence regions of less than 10 nucleotides long are too short to be mapped onto the genome map.

Figure 1. Sample output from our jpHMM web server. The output file contains a list of fragments from the input HIV-1 sequences that are assigned to different HIV subtypes, including predicted breakpoints. At the bottom of the file, a graphical representation of the input sequence is given where recombinant subtypes are color coded. Gray regions denote missing subtype information due to uninformative subtype models.

01_*AE*, though being called recombinant, contains the only information of subtype *E*. Thus we include 01_*AE* in the alignment. The alignment of these sequences has been carried out using HMMER (15) and subsequent manual improvement.

A hyperlink to the results of the program run is returned to the user by e-mail. The result file contains a list of fragments of the input sequence that are assigned to different subtypes and sub-subtypes, including predicted breakpoints between these fragments. In addition, the output file contains a graphical representation of the predicted recombinant fragments within the HIV-1 genome. A sample output file is shown in Figure 1. The predicted breakpoint positions are provided in two ways. One is based on the original sequence position, and the other is based on HXB2 numbering. HXB2 (GenBank accession number K03455) is the most commonly used reference strain for many different kinds of HIV-1 functional studies. The HXB2 numbering provided for the output breakpoints is especially useful to facilitate the identification of the precise location of interest in HIV sequences.

PROGRAM LIMITATIONS AND FUTURE WORK

It should be mentioned that our tool is sometimes not sensitive to detect HIV-1 subtypes H, J, K, as only few full-length genome sequences of these subtypes are available to train our model. For these subtypes, we recommend to compare the results of jpHMM with those of other HIV-1 subtyping tools, for example, RIP (http://hiv-web.lanl.gov/content/hiv-db/RIPPER/RIP.html).

As shown in (7), the overall prediction accuracy of our method is high compared with alternative approaches. Nevertheless, it would be useful for the user to assess the relative reliability of individual predicted breakpoints. In principle, this is possible by using posterior probabilities that can be calculated using the Forward and Backward algorithms as explained in (16). We are currently implementing these algorithms to estimate the (local) reliability of our predictions. This feature will be available on our web site in the near future.

For predicted recombinants, users of our software may want to know putative parental sequences. Our method cannot provide this information directly, since jpHMM compares input sequences to a model derived from a pre-calculated alignment of representative sequences. It is possible, however, to search predicted recombinant segments of input sequences against the HIV-1 database to retrieve potential parent sequences. We are planning to add this functionality to our web server soon.

ACKNOWLEDGEMENTS

This project was funded in part by grant NIH Y1-AI-1500-01, the NIH-DOE interagency agreement, the HIV Immunology and Sequence Database and by BMBF grant 01AK803G (Medigrid) and DFG grant 1048/1-1 to BM. We thank Rasmus Steinkamp and Maike Tech for helping us with the web server at GOBICS. Two anonymous referees made useful comments on the manuscript. Funding to pay the Open Access publication charges for this article was provided by the annual budget of BM's research group.

Conflict of interest statement. None declared.

REFERENCES

1. Leitner,T., Foley,B., Hahn,B., Marx,P., ,McCutchan,F., Mellors,J., Wolinsky,S. and Korber,B. (eds) (2005) *HIV Sequence Compendium 2005*. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM.
2. Hoelscher,M., Dowling,W.E., Sanders-Buell,E., Carr,J.K., Harris,M.E., Thomschke,A., Robb,M.L., Birx,D.L. and McCutchan,F.E. (2002) Detection of HIV-1 subtypes, recombinants, and dual infections in East Africa by a multi-region hybridization assay. *AIDS*, **16**, 2055–2064.
3. Sharp,P.M., Shaw,G.M. and Hahn,B.H. (2005) Simian immunodeficiency virus infection of chimpanzees. *J. Virol.*, **79**, 3891–3902.
4. Robertson,D.L., Anderson,J.P., Bradac,J.A., Carr,J.K., Foley,B., Funkhouser,R.K., Gao,F., Hahn,B.H., Kalish,M.L., Kuiken,C. *et al.* (2000) HIV-1 nomenclature proposal. *Science*, **288**, 55–57.
5. Martin,D. and Rybicki,E. (2000) RDP: detection of recombination amongst aligned sequences. *Bioinformatics*, **16**, 3535–3540.
6. Siepel,A.C., Halpern,A.L., Macken,C. and Korber,B.T. (1995) A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retroviruses*, **11**, 1413–1416.
7. Schultz,A.-K., Zhang,M., Leitner,T., Kuiken,C., Korber,B., Morgenstern,B. and Stanke,M. A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC Bioinformatics*, **7**, 265.
8. Spang,R., Rehmsmeier,M. and Stoye,J. (2000) Sequence database search using jumping alignments. In Bourne,P., Gribskov,M., Altman,R., Jensen,N., Hope,D., Lengauer,T., Mitchell,J., Scheeff,E., Smith,C., Strande,S. and Weissig,H. (eds), *Proceedings of Intelligent Systems for Molecular Biology 2000*. AAAI Press.
9. Spang,R., Rehmsmeier,M. and Stoye,J. (2002) A novel approach to remote homology detection: Jumping alignments. *J. Comput. Biol.*, **9**, 747–760.
10. Krogh,A., Brown,M., Mian,I., Sjolander,K. and Haussler,D. (1994) Hidden markov models in computational biology: applications to protein modelling. *J. Mol. Biol.*, **235**, 1501–1531.
11. Zhang,M., Schultz,A.-K., Morgenstern,B., Stanke,M., Korber,B. and Leitner,T. (2005) Greater HIV genome diversities inferred from re-subtyping of HIV database sequences. In *Proceedings of German Conference on Bioinformatics (GCB'05), Discovery Notes, Poster Abstracts*, pp. 5–7.
12. Lole,K.S., Bollinger,R.C., Paranjape,R.S., Gadkari,D., Kulkarni,S.S., Novak,N.G., Ingersoll,R., Sheppard,H.W. and Ray,S.C. (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.*, **73**, 152–160.
13. Martin,D.P., Williamson,C. and Posada,D. (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics*, **21**, 260–262.
14. Kuiken,C. and Leitner,T. (2005) Chapter HIV-1 Subtyping. *Computational and Evolutionary Analysis of HIV Molecular Sequences*. Kluwer Academic Publishers, pp. 27–53.
15. Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein families based on seed alignments. *Proteins*, **28**, 405–420.
16. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

BMC Bioinformatics

Methodology article

A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes

Anne-Kathrin Schultz[1], Ming Zhang[1,2], Thomas Leitner[2], Carla Kuiken[2], Bette Korber[2,3], Burkhard Morgenstern[1] and Mario Stanke*[1]

Address: [1]Institute of Microbiology and Genetics, University of Göttingen, Goldschmidtstr. 1, 37077 Göttingen, Germany, [2]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA and [3]The Santa Fe Institute, Santa Fe, NM 87501, USA

Email: Anne-Kathrin Schultz - aschult2@gwdg.de; Ming Zhang - mingzh@lanl.gov; Thomas Leitner - tkl@lanl.gov; Carla Kuiken - kuiken@lanl.gov; Bette Korber - btk@lanl.gov; Burkhard Morgenstern - bmorgen@gwdg.de; Mario Stanke* - mstanke@gwdg.de

* Corresponding author

Published: 22 May 2006

BMC Bioinformatics 2006, **7**:265 doi:10.1186/1471-2105-7-265

Received: 16 December 2005
Accepted: 22 May 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/265

Abstract

Background: Jumping alignments have recently been proposed as a strategy to search a given multiple sequence alignment *A* against a database. Instead of comparing a database sequence *S* to the multiple alignment or profile as a whole, *S* is compared and aligned to individual sequences from *A*. Within this alignment, *S* can *jump* between different sequences from *A*, so different parts of *S* can be aligned to different sequences from the input multiple alignment. This approach is particularly useful for dealing with *recombination* events.

Results: We developed a *jumping profile Hidden Markov Model* (jpHMM), a probabilistic generalization of the jumping-alignment approach. Given a partition of the aligned input sequence family into known sequence *subtypes*, our model can jump between states corresponding to these different subtypes, depending on which subtype is locally most similar to a database sequence. *Jumps* between different subtypes are indicative of intersubtype recombinations. We applied our method to a large set of genome sequences from *human immunodeficiency virus* (HIV) and *hepatitis C virus* (HCV) as well as to simulated recombined genome sequences.

Conclusion: Our results demonstrate that jumps in our jumping profile HMM often correspond to recombination breakpoints; our approach can therefore be used to detect recombinations in genomic sequences. The recombination breakpoints identified by jpHMM were found to be significantly more accurate than breakpoints defined by traditional methods based on comparing single representative sequences.

Background

Profile Hidden Markov Models [1] are a popular way of modelling nucleic-acid or protein sequence families for database searching, see [2] for a review. Like other Hidden Markov Models (HMMs), profile HMMs consist of so-called *states* that can *emit* symbols of the underlying alphabet, i.e. nucleotides or amino acids [3]. *Transitions* are possible between these states, and a DNA or protein sequence is thought to be generated by a *path Q* through the model beginning with a special *begin* state and ending with an *end* state. There are probabilities (*a*) for possible transitions from one state to another and (*b*) for the emission of

symbols at a given state. The states together with the possible transitions between them are called the *topology* of the model while the corresponding transition and emission probabilities are called its *parameters*. A sequence S is generated by the model with a certain probability $P(S)$. In general, a sequence S can be generated by more than one path Q through the model. For a given sequence S, the well-known *Viterbi Algorithm* [4] finds the most probable path that generates S. More precisely, the algorithm finds a path Q^* that maximizes the conditional probability $P(Q|S)$ which is equivalent to maximizing the joint probability $P(Q, S)$. For a general introduction to HMMs, see [5].

The starting point for a profile HMM is a multiple alignment of a sequence family. Columns of the alignment are modeled as states of the HMM. These states are called *match states* and are denoted by M_i; the indexing is such that the alignment column associated with a match state M_i is to the left of the column associated with M_j whenever $i < j$. Emission probabilities for a match state M_i depend on nucleotide or amino acid frequencies in the respective alignment column. In general, not every column of the input multiple alignment corresponds to a match state, but only those columns that have a certain minimum number of non-gap characters are modeled as match states. Columns that correspond to match states are called *consensus columns*. State transitions are possible from one match state M_i to the next match state M_{i+1}. To account for insertions and deletions, additional states are defined. *Insert states* I_i can emit additional symbols while *delete states* D_i can be used to omit one or more match states in the model. As the *Begin* and *End* states, delete states are *silent*, i.e. they do not emit any symbols. Figure 1 shows the topology of a profile HMM. An insert state I_i is located between the match states M_i and M_{i+1} and there are possible transitions from M_i to I_i and from I_i to M_{i+1} such that an additional character can be inserted between M_i and M_{i+1}. By contrast, a deletion state D_i is in the same column as a match state M_i. There are possible transitions from M_{i-1} to D_i and from D_i to M_{i+1} to circumvent match state M_i. Profile HMMs are frequently used tools for database searching. They are slower but more accurate than standard local-alignment approaches such as BLAST [6]. The best known implementation of profile HMMs is Sean Eddy's software program *HMMer* [2,7].

Jumping alignments have been proposed by Spang *et al*. as a new approach to database searching [8,9]. Like profile HMMs, jumping alignments start with a *multiple alignment* A of a sequence family, and database sequences S are compared to A. But unlike in standard methods, the database sequence is not aligned to the query multiple alignment A or to the corresponding profile as a whole, but the program aligns segments from the database sequence S to *single* sequences from the multiple alignment A. Each position of S is aligned with only one sequence from A, the so-called *reference sequence* for this position. Within one alignment, the program can *jump* between the reference sequences. For such jumps a *penalty* is imposed similar to the gap penalties that are used in standard alignment algorithms.

In the present paper, we describe a novel approach to compare a single nucleic acid or protein sequence to a multiple alignment of a sequence family. Our approach combines the above outlined methods and can be seen as a probabilistic generalization of the jumping-alignment approach. We therefore call our method *jumping profile Hidden Markov Model (jpHMM)*. The proposed tool is a

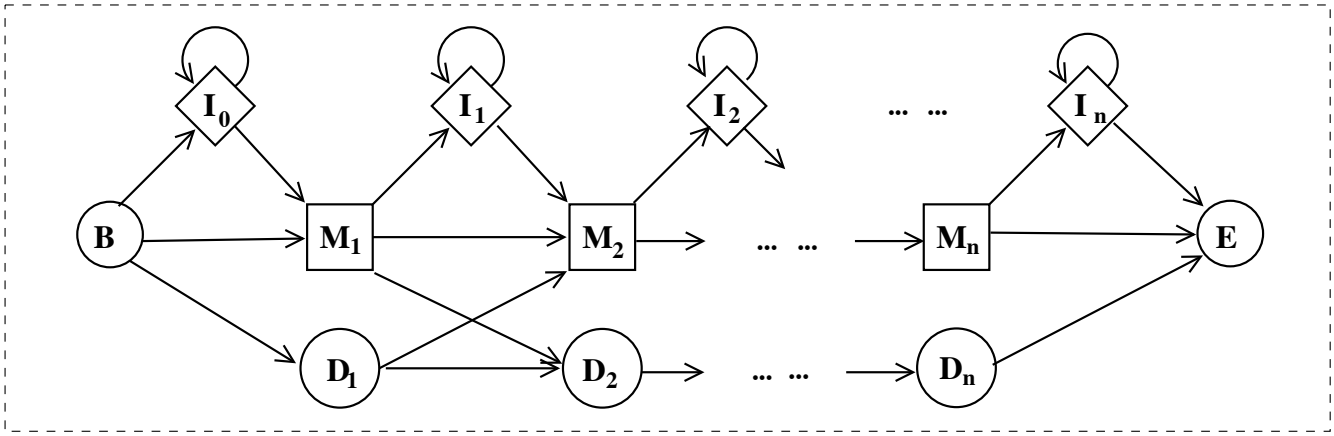


Figure 1
A profile hidden Markov model as introduced by Krogh et al. [1]. Squares indicate match states (M_i), diamonds insert states (I_i) and circles delete states (D_i). Possible transitions are shown as arrows. *Begin* state (*B*), *End* state (*E*) and delete states (D_i) are *silent* states, *i.e.* they do not emit symbols of the alphabet.

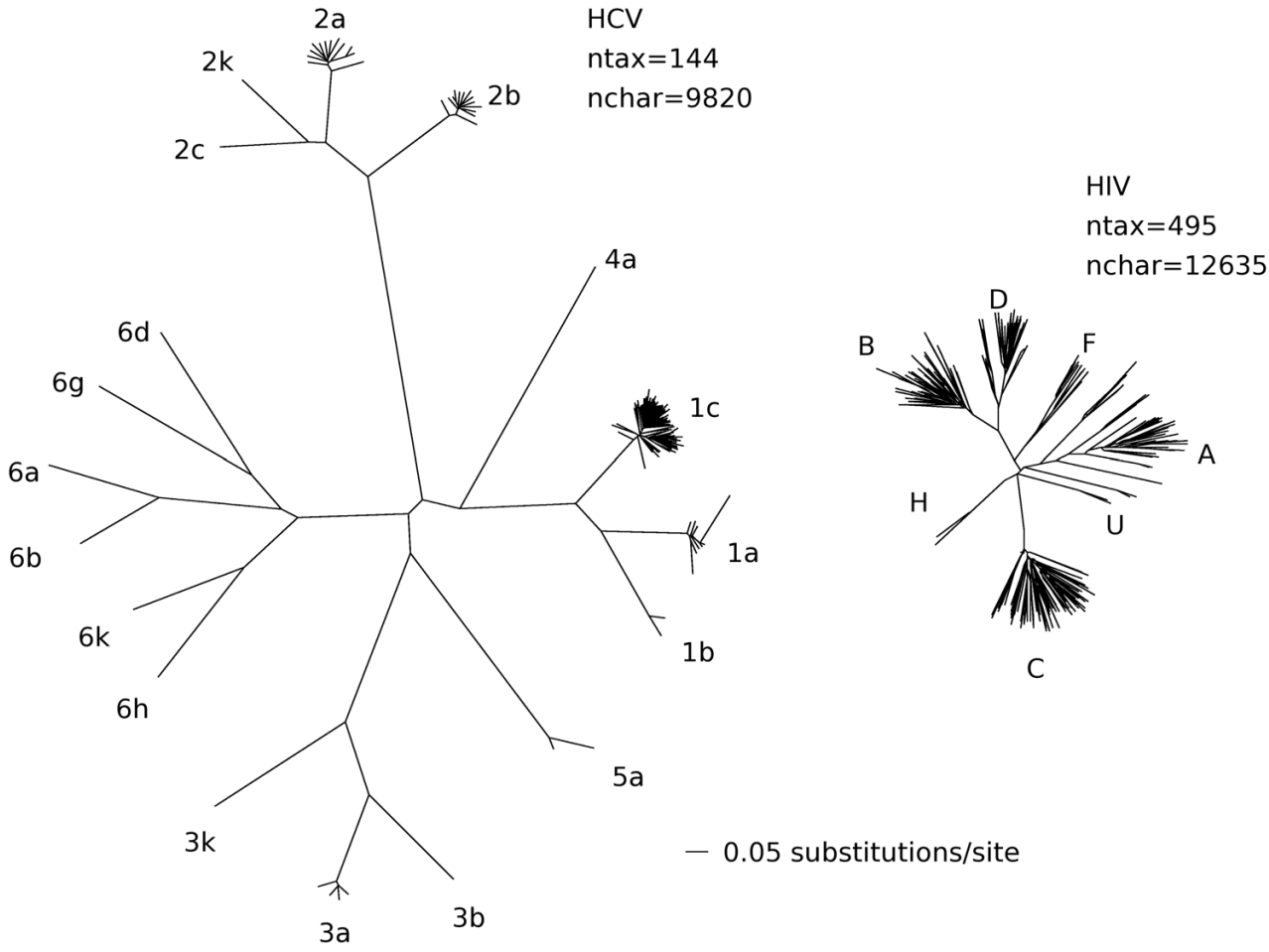


Figure 2
Complete genome trees of the hepatitis C and HIV-1 (M group) viruses. The trees are drawn on the same scale. Only non-recombinant complete genomes have been included. The trees are based on manually curated alignments containing only one sequence per patient. The optimalization method was maximum-likelihood, as implemented in the GAML/Garli program (Zwickl et al. 2005, in preparation).

flexible method for database searching; it is particularly useful if sequence recombinations have to be taken into account. In the present study, we apply our method to localize phylogenetic breakpoints in viral genome sequences. We applied our approach to identify genomic recombinations in HIV and HCV and to classify subtypes. Accurate classification of HIV and HCV sequences is of crucial importance for epi-demiological monitoring as well as for the design of molecular detection systems and potential vaccines. HIV and HCV are among the most genetically variable organisms known. Based on phylogenetic clustering, these viruses have been classified into clades (Figure 2). The classification is not always trivial, because genetic forms that do not cluster within the phylogenetic clusters exist, and for both viruses recombinants have been discovered that make the classification more obscure. Furthermore, some genes and genome segments do not contain enough information to resolve the subtypes, especially when DNA sequences become too short.

Most classification methods depend on an accurate sequence alignment, and those that do not, still depend on pair-wise comparisons between a query sequence and some set of reference sequences. Since the subtypes are phylogenetically defined, tree building is the gold standard. Reconstructing accurate phylogenetic trees is, however, neither trivial nor easy to incorporate in automatic screening procedures. In recent years, many methods have been developed to detect genomic recombinations based on phylogenetic trees, sequence patterns and population

genetics. In the virology field, the most popular methods have been based on pair-wise genetic distance calculations. Generally, HIV recombination is detected based on pairwise distances [10] and breakpoint locations are defined based on a method called "informative sites analysis" [11,12] prior to generating phylogenies, which are then used to validate the level of support for recombination in different genomic regions. Alternatively a *bootscan* is performed to determine whether phylogenetic branching patterns differ in trees constructed based on a sliding window approach [13]. Informative sites analysis divides the sequence at the midpoint between the substitutions that mark the dividing point that gives the most support for the recombination events using a chi square test; single representatives of the two recombinant clades are compared to the query sequence. In contrast, the jpHMM approach that we propose in this paper enables defining breakpoints based on a model derived from the full alignment of a particular clade; this is particularly useful when the precise parental strain is not known, rather a parental lineage is defined.

HIV-1 is classified into three major phylogenetic groups, called M, N and O, that arose due to separate introductions of SIVs from chimpanzees into humans [14]. The M group, which is responsible for the HIV pandemic, is further divided into ten subtypes, some of which have been even further subdivided into sub-subtypes [15]. Inter-subtype recombination is extremely common among HIV-1 subtypes [16]. Identifying intersubtype recombinants is important from many perspectives, giving insights into such issues as molecular epidemiology, viral evolution, and indirectly, the frequency of dual infections.

For hepatitis C, the picture is even more complex; there are currently 6 major genotypes, each subdivided into subtypes, of which there can be dozens. To curb the explosion in new subtypes, it was recently decided that new subtypes will only be named when there are at least three unrelated samples for them [17]. Recombinants that have been epidemiologically successful exist for both viruses, and are called *CRFs* in HIV and *RFs* in HCV, for *(circulating) recombinant forms*. HIV CRFs are common, and they often emerge as the dominant clade in a regional epidemic [18]. More precise breakpoint definitions will help in identifying and tracking HIV CRFs.

Currently only a small number of naturally occurring recombinants have been identified for the hepatitis C virus, despite frequent dual infection [19,20,20-24]. Until 2005, only one circulating recombinant form had been described, from St. Petersburg [25]. However, the discovery of new recombinants does appear to be speeding up, with two recent publications describing new circulating recombinants between genotypes 2a and 2c from Peru [26], and between genotypes 2i and 6h from Vietnam (S. Noppornpanth, unpublished results). It is very likely that more recombinants will be discovered in the near future. Discovery and accurate characterization of new recombinants is hampered by the scarcity of complete genome sequences for most of the less frequent HCV genotypes.

Recombinants found for hepatitis C so far are simple in structure, none of them appear to combine fragments from more than two genotypes and all appear to contain only one breakpoint. Thus, characterizing HCV recombinants found to date is a simpler task than characterizing the complex recombinants that are often found in HIV (for review, see [15]), for a specific example of the complexity, see [27]. However, given the improved sampling and sequencing capacity in HCV and the associated growing frequency of detection of recombinants, it will be increasingly important to have a tool that can reliably and efficiently identify recombinants and locate their breakpoints.

Results

Jumping profile Hidden Markov Model

A jumping profile Hidden Markov Model (jpHMM) as introduced in this section combines profile HMMs with the jumping alignment (JALI) approach introduced by Spang et al. [9]. The data that a jpHMM models are a sequence family $\mathcal{S} = \{S_1,...,S_n\}$ together with a multiple sequence alignment A of $\mathcal{S}$. In addition, we assume that we have a partition of $\mathcal{S}$ into k *subtypes* $\mathcal{S}_1,...,\mathcal{S}_k$ such that each sequence S_i belongs to exactly one of the subtypes. Our jpHMM approach can be seen as a generalization of the 'jumping alignment' algorithm. In JALI, each position of a database sequence is aligned to one reference sequence $S_i \in \mathcal{S}$. By contrast, jpHMM aligns parts of the input sequence to entire *subtypes* of the input multiple alignment. Thus, JALI corresponds to the special case in our approach where each subtype $\mathcal{S}_i$ consists of exactly one sequence. As with standard profile HMMs, each match state in our model is derived from one column of the input alignment A. However, in our model we define match states specific for the subtypes. Thus, a column in the query alignment A may correspond to up to k match states, and a match state is specified by two indices, the corresponding column of the multiple alignment A and the subtype it belongs to.

For a given subtype $\mathcal{S}_i$, a column is modeled as a match state only if it is a consensus column for that subtype. Consequently, a column i in the alignment A may be

modeled as a match state for some of the subtypes but not for other subtypes. In addition to the match states, we have distinct insert and delete states for each subtype just as in standard profile HMMs. In our notation, match state $M_{i,j}$ is the *j*-th match state within the *i*-th subtype, and $I_{i,j}$ and $D_{i,j}$ are the insert and delete states corresponding to $M_{i,j}$. There is a single *Begin* state and a single *End* state, respectively, for the entire model. Further, there are general, not subtype-specific insert and delete states just after the *Begin* state and just before the *End* state. From the delete state immediately after the *Begin* state, each match state from each subtype can be reached. Similarly, from each match state, the delete state directly before the *End* state can be reached. These states have been introduced to deal with the fact that the sequences are often incompletely sequenced and are missing the initial or terminal part.

Note that the sub-model associated with a subtype $\mathcal{S}_i$ in our jpHMM corresponds to a standard profile HMM for $\mathcal{S}_i$. Thus, our model can be seen as the union of *k* standard profile HMMs with additional transitions between these standard HMMs. The underlying multiple alignment *A* induces a *quasi partial order relation* on the set of all states of our jpHMM. We say that a match or delete state *T* is (strictly) to the left of a match or delete state *R* if the alignment column associated to *T* is (strictly) to the left of the column associated with *R*. This ordering is related to the quasi partial order relation $\circ_A$ defined on the set of all *sites* of a multiple alignment introduced in [28] in the context of *consistency* of alignments. As for standard HMMs, the states of a jpHMM are connected by transitions to which transition probabilities are assigned. Transitions are possible *within* one subtype $\mathcal{S}_i$ as in standard profile HMMs, e.g. from one match state a transition is possible to the next match state or to the corresponding insert and delete states of $\mathcal{S}_i$. In addition, our model allows transitions *between* different subtypes as shown in Figures 3 and 4. Transitions between subtypes are called *jumps*.

Transitions *between* subtypes are more complicated than *within* subtypes since not every alignment column is represented in every subtype as a match state. Thus, it is not obvious from which state in one subtype we can jump to which state in another subtype, so we need to specify which jumps between subtypes are allowed. Generally, there are two reasons to limit the number of possible jumps between states. (*a*) To reduce the computer resources required by our algorithm, we need to limit the number of possible transitions between states. (*b*) More importantly, we need to make sure that a path through our model cannot jump to the left or too far to the right. A jump to the left would have the biological meaning of a tandem repeat of a certain part of the sequence, which we do not allow. A jump to the right that overjumps consensus columns in one of the two subtypes involved in the jump means that some part of one of those subtypes is deleted with respect to the query sequence. This is possible but should be punished as in a standard profile HMM by using the alternative path, a chain of delete states. This exclusion of forward jumps similarly reduces the number of transitions as done in [29]. In our approach, we imposed the following rules:

r1 For two subtypes $\mathcal{S}_i$ and $\mathcal{S}_j$, the algorithm can jump from a match state of $\mathcal{S}_i$ only to a match state or a delete state of $\mathcal{S}_j$, and from an insert state or delete state of $\mathcal{S}_i$ a jump is possible only to a match state of $\mathcal{S}_j$.

r2 A jump from a state *T* in $\mathcal{S}_i$ is possible only to the *leftmost* state in $\mathcal{S}_j$ that is *strictly to the right* of *T*.

r3 A jump from a state $M_{i,k}$, $D_{i,k}$ or $I_{i,k}$ to a state in $\mathcal{S}_j$ that is to the right of $M_{i,k+1}$ is not possible.

Rule r1 reduces the number of possible transitions in our model. Rules r2 and r3 ensure that there are no insertions or deletions introduced during a jump without using insert or delete states.

Parameter estimation

A jpHMM has a large number of parameters that need to be specified, namely the emission probabilities of match and insert states, the transition probabilities *within* subtypes and the probability of the jumps *between* different subtypes. With the exception of the probabilities of jumps, which is discussed below, the above probabilities can be estimated based on observed frequencies. Given the topology of the jpHMM, each of the sequences in the given multiple sequence alignment defines a unique path through the states, and gives rise to observed emissions and transitions. For example, a particular residue that is aligned in a consensus column is emitted from the corresponding match state of the subtype the respective sequence belongs to. To give another example, an insert region of length *l* gives rise to one transition from the preceding match state to the corresponding insert state, *l* - 1 transitions from that insert state to itself and one transition from the insert state to the next match state.

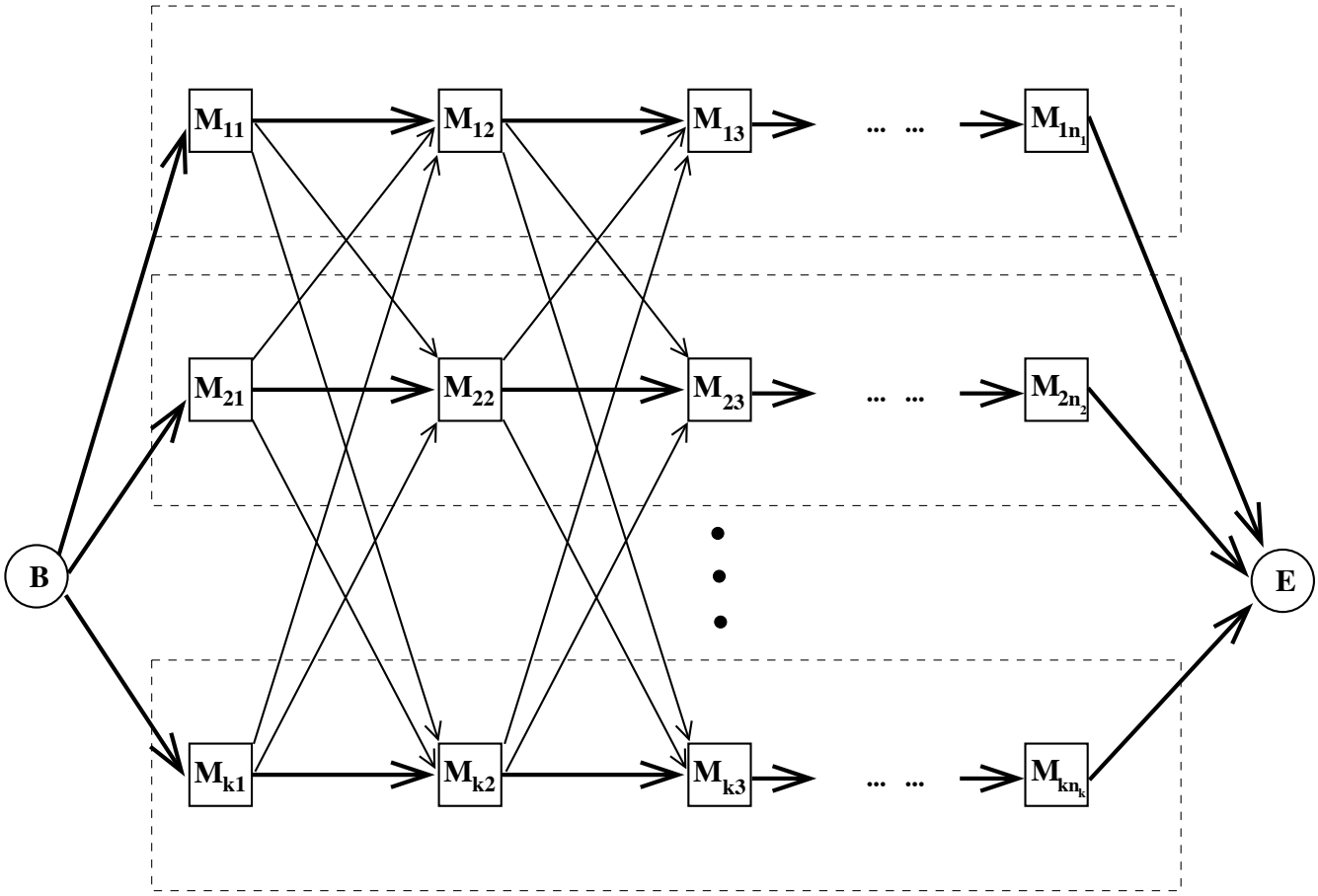


Figure 3

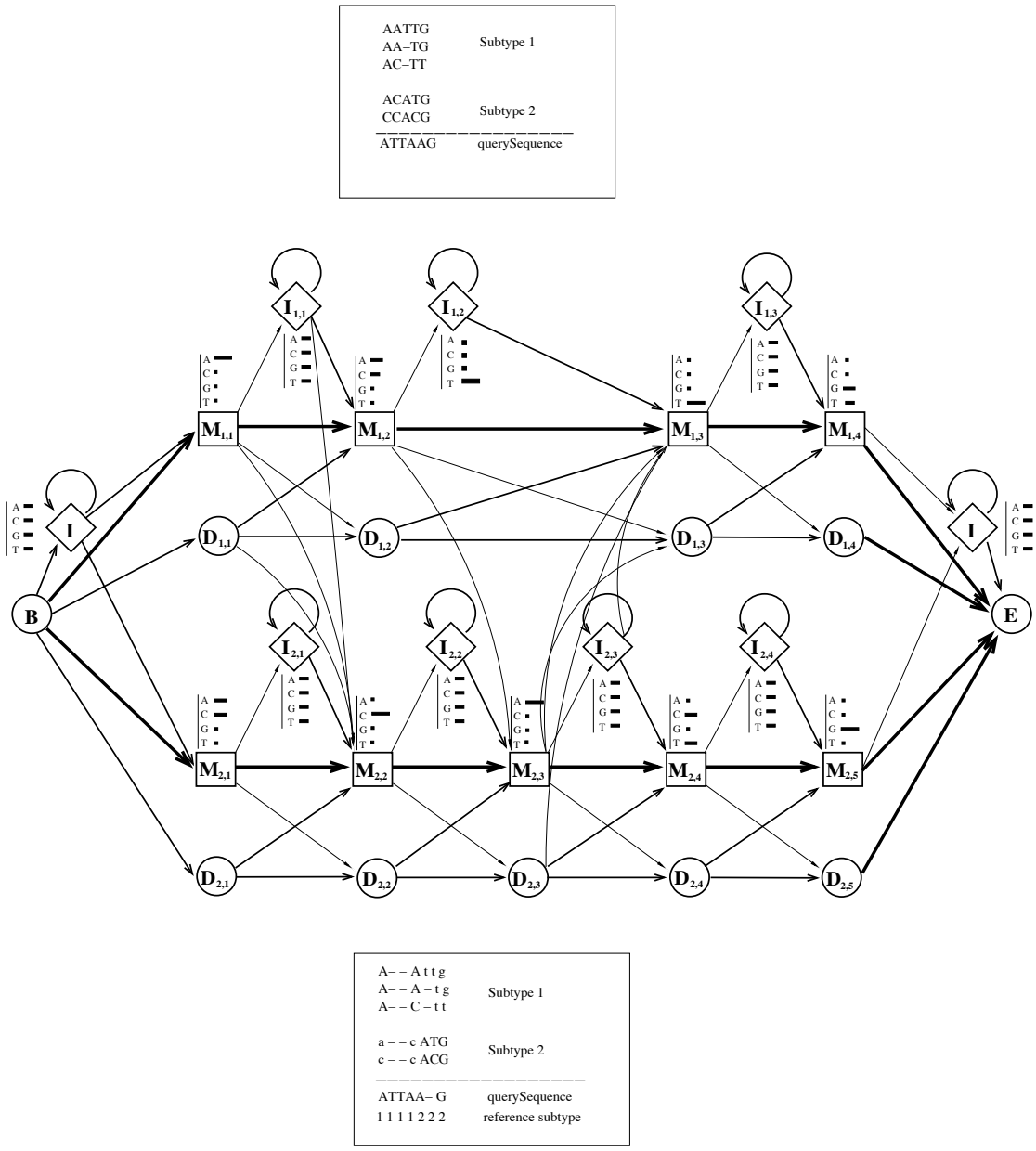
Simplified topology of a jumping profile HMM (jpHMM). The sequence family $\mathcal{S}$ is partitioned into *k* subtypes $\mathcal{S}_1,...,\mathcal{S}_k$. Each subtype is modeled by a profile HMM here pictured as a dashed box. Arrows indicate possible transitions between states within the same subtypes and transitions between different subtypes, so-called *jumps*. For clarity we omit insert and delete states of the profile HMMs and sketch a case where the first three columns are consensus columns.

The generalized problem for estimating each emission distribution and each distribution of possible transitions out of a state is the following. We are given a count vector $\vec{n} = (n_1,...,n_s)$, where *s* is the number of emissions (or transitions, respectively) out of this state. For example, we have $s = 4$ in case of nucleotide emissions. n_i is the number of times the *i*th emission (or transition) is observed. These observed frequencies $\vec{n}$ are distributed according to a multinomial distribution with parameters $\vec{p} = (p_1,...,p_s)$, where p_i is the probability of observing option *i*. For this problem of estimating $\vec{p}$ given $\vec{n}$, we chose a Bayesian approach as in [30]. This means we assume a prior distribution on the set of all possible $\vec{p}$, and then estimate $\vec{p}$ by the following conditional expectation

$$E(\vec{p} \mid \vec{n}).$$

We model this prior knowledge using a Dirichlet distribution [30] which has parameters $\vec{\alpha} = (\alpha_1,...,\alpha_s)$. These parameters can be interpreted as pseudocounts that are added to the observed counts. For the emission probabilities we estimated the parameters $\vec{\alpha}$ of the prior distribution with a Maximum Likelihood approach [30] based on the input multiple alignment. For the transition probabilities we used the parameters $\vec{\alpha}$ of the prior distribution taken from [31]. Those were shown to perform better than the parameters derived by Maximum Likelihood.

In contrast to the transitions *within* a subtype of the jpHMM, jumps *between* subtypes cannot be observed in the input alignment data. Since we cannot estimate the



viterbi path (gives us the alignment of the query to the alignment):

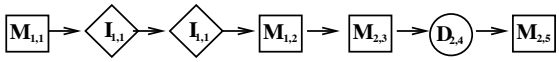


Figure 4

A toy example of a jumping profile Hidden Markov Model. This is build from a multiple sequence alignment of nucleotides with two subtypes. The first subtype consists of three sequences with four consensus columns, the second subtype consists of two sequences with five consensus columns. With each match and insert state a vector is associated for the emission probability values corresponding to the nucleotides A, C, G and T. For clarity, some transitions are omitted. Also, the figure does not show the delete states immediately after *B* from which each match state can be reached nor the delete state immediately before *E* that can be reached from each match state. Fat lines indicate high transition probabilities, thin lines correspond to low probabilities.

corresponding jump probabilities from observed frequencies, we use a fixed, empirically derived value P_j for the probability of observing a jump. If in a given state of the jpHMM, there are several possibilities for a jump, this probability is evenly distributed between the possible jumps. In other words, if we have K options to jump into another subtype, the probability of each of these jumps is given by P_j/K. In the application of our program to the identification of HIV and HCV recombinants we use a jump probability of $P_j = 10^{-9}$ which we derived by optimizing the results by comparing them to published HIV intersubtype recombination breakpoints. Taking into account that the transition and jump probabilities out of each state must sum up to 1, we scale the non-jump transition probabilities, i.e. the probabilities for transitions within the same subtype by multiplying them by $(1 - P_j)$, if jumps are possible out of this state.

Alignment algorithm and efficiency

The jumping alignment of a query sequence $S = s_1, s_2,...,s_n$ and a given multiple alignment is determined by searching the most probable path Q^* through the jpHMM that emits S as described above. This is done with a dynamic programming algorithm, the Viterbi algorithm. For each position $i = 1,..., n$ of the query sequence S and for each state q of the jpHMM we calculate the probability $\delta_i(q)$ of the prefix $s_1,...,s_i$ of the query sequence and the most probable path through the jpHMM ending in state q and emitting $s_1,...,s_i$. These probabilities are called Viterbi values and the following recursion holds.

$$\delta_{i+1}(q) = \begin{cases} \max_{q'} \left\{ \delta_{i+1}(q') t_{q'q} \right\}, & \text{if } q \text{ is a delete state;} \\ \max_{q'} \left\{ \delta_i(q') t_{q'q} \right\} e_{q,s_{i+1}}, & \text{otherwise.} \end{cases}$$

Here, q' ranges over all states of the model, $t_{q'q}$ is the probability of the transition from state q' to state q and $e_{q,s_{i+1}}$ is the probability of emitting nucleotide s_{i+1} out of state q. The Viterbi values can be computed by increasing i with the states sorted from left to right. By backtracking we can construct the most probable path Q^* (see Figure 4) and thus the jumping alignment. This algorithm has a complexity of $O(n\ell k)$ in time and $O(n\ell)$ in space where n is the length of the query sequence, k the number of subtypes in the alignment and ℓ the number of states in the jpHMM.

In the case of very long alignments this may require too much time and memory for current computer hardware. For example, genomes of the HIV-1 group M have a length of roughly 10, 000 nucleotides. Thus, given a multiple alignment of 14 (sub-)subtypes of such sequences and a query sequence of length ~10, 000 in a straightforward implementation we would need to calculate and store roughly $10,000 \cdot 10,000 \cdot 14 \cdot 3 = 4 \cdot 2 \cdot 10^9$ floating point numbers.

To accelerate the computation and to save memory we bound the number of considered states in each step i by using the beam-search algorithm [32,33]. The idea behind this algorithm is to exclude possible irrelevant paths in each step and to restrict the search space to 'promising' paths. If an alignment of an initial part is *much* worse than another alignment of that part then we do not try to extend that low quality alignment. This is achieved by computing and storing in each step i a modified Viterbi value $\delta'_i(q) \le \delta_i(q)$ only for a subset $\mathcal{A}_i$ of the states, namely those states q whose modified Viterbi value is not much lower than the optimal local solution $\delta_i^* = \max_q \delta'_i(q)$ These states are called 'active' states and the set $\mathcal{A}_i$ of active states of step i is determined by

$$\mathcal{A}_i = \{q \mid \delta'_i(q) \ge \mathcal{B}\delta_j^*\}, \quad 0 < \mathcal{B} \ll 1.$$

The modified Viterbi value of the inactive states is set to 0, and does not need to be stored. In the next step $i + 1$ of the recursion the modified Viterbi value $\delta'_{i+1}(q)$ needs only be computed for states, which can be reached from a state in the subset of active states $\mathcal{A}_i$ through a path with one emission. This speeds up the computation of the recursion.

In the tradeoff between memory efficiency and speed (large $\mathcal{B}$) against accuracy (small $\mathcal{B}$)) we chose a beam in the order that allows maximal accuracy within the limits of currenct PC hardware: $\mathcal{B} = 10^{-20}$. In Figure 5 and 6 we sketch the set of columns of activated states in the multiple alignment for an example with HIV sequences. Using this beam search heuristic very rarely affects the output of the computation but the time and memory savings are immense. The average number of active states per input sequence position in this example is 1,690 which compares to roughly $10,000 \cdot 14 \cdot 3 = 4 \cdot 2 \cdot 10^5$ states if the beam search heuristic was not used. For the HIV-1 sequences that we tested, the average number of active states was between 1,620 (CRF 12, length = 8,760 nt) and 2,862 (CRF 11, length = 9,768 nt). The CPU time for those sequences using the beam search heuristics is 7.2 min (CRF 12) and 13.6 min (CRF 11) on a Linux PC with 3 GB RAM and 3.2 GHz. This includes model building as well

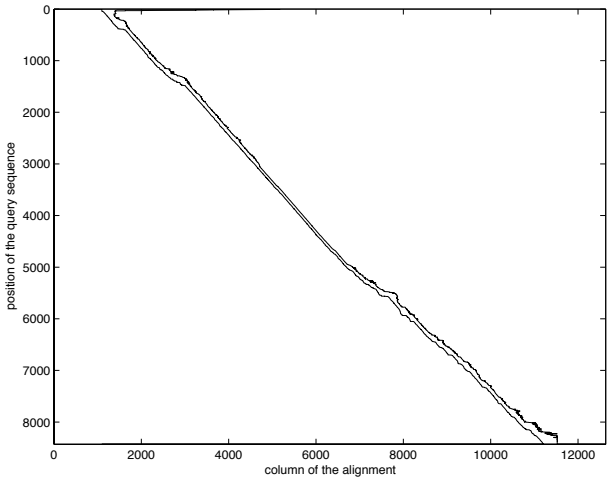


Figure 5
Reduction of active states for a set of HIV test sequences using the beam-search heuristics [32, 33]. In this example, we have 14 (sub-)subtypes each of which has three states per alignment column (match, delete and insert). Thus, a column corresponds to 14 × 3 = 42 states. The beam-search algorithm reduces the number of active states considerably; the figure indicates for each position in the query sequence those columns that contain active states. Thus, instead of considering the entire dynamic-programming matrix, our algorithm needs to consider only the small strip between the two lines. We used a beam width of $\mathcal{B} = 10^{-20}$ and a jump probability of 10^{-9}.

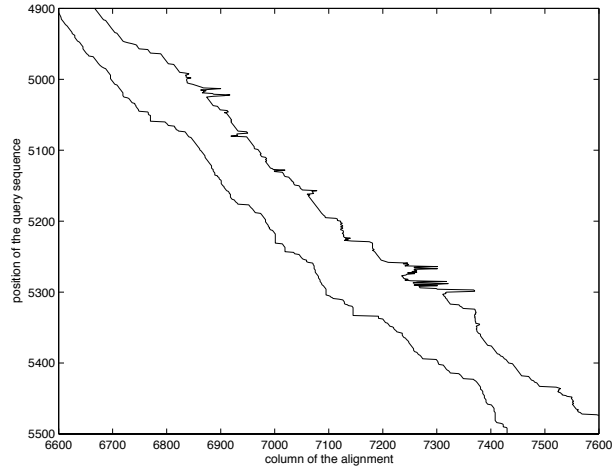


Figure 6
Detailed section of Figure 5.

as a search of one sequence against the model, but most of the CPU time was consumed by the second step.

Test results

Results on HIV genomic sequences

To evaluate the accuracy of our jpHMM approach on HIV-1 sequences, we used two different types of test data including simulated as well as real-world sequence data.

First, we wanted to know to what extend our method is able to recognize subtypes in artificial sequences produced by the underlying probabilistic model itself. This test can be considered as a minimal check of our model. We sampled 800 random sequences according to the transition and emission probabilities of the HMM built for HIV-1 subtyping. Since the "jump probability" in our model is rather small, each of our 800 artificial sequences consisted of one subtype only without any recombinations. For each of these sequences our method correctly predicted the underlying single subtype, and no jumps between different subtypes were predicted. Moreover, for 752 of our 800 sequences, 100 % of the individual sequence positions were assigned to the correct subtype. In the remaining 48 sequences, the only differences between the sampled and predicted paths were in the lengths of short unclassified regions at the ends of the sequence. All in all, 99.99 % of the sequence positions in our test sequences were assigned to the correct subtype.

Further, we used *simulated* inter-subtype recombinant sequences with known breakpoints. Artificial recombinant sequences were created in the following way: for each simulated sequence, two real-world 'parent' sequences were taken from two different clades of HIV-1. For a fixed value X, these sequences were split at every X-th nucleotide, and a simulated recombinant sequence was composed of alternating segments of length X from these two parent sequences. Thus, for two parent sequences $P1$ and $P2$ and, for example, $X = 1000$, the first 1000 nucleotides of the artificial recombinant are from $P1$, nucleotides 1001 to 2000 are from $P2$, residues 2001 to 3000 are from $P1$ etc. In the present study, we used values of $X = 500$, 1000, 1500. This way, known breakpoints were introduced into both relatively conserved regions, and highly variable regions, and the performance of the jpHMM method could be assessed in both contexts. Here, we distinguish between recombinant sequences with parents from different *subtypes* which we call *inter-subtype* recombinants and recombinants with parents from the same subtype but different *sub-subtypes*, which we refer to as *inter sub-subtype* recombinants. Sub-subtypes are clearly distinguishable, established lineages that occur within the subtypes, but do not have the minimal genetic distances required to be considered an independent subtype. For historical reasons the B and D clades are called subtypes, but in fact the distances between these two clades correspond to a sub-subtype distance [15].

The sequences used in the inter-subtype recombinants have been created using parent sequences from the following subtypes (GenBank accession numbers of the parent sequences are in parentheses): A1 and C (A1: AF193275, C:AY463217); A1 and D (A1: AF193275, D:AF133821); A1 and G (A1: AF193275, G: AF450098); B and C (B: AF042101, C: AY463217); B and F1 (B: AF042101, F1:AY173958); B and 01 (B: AF042101, 01:AB032741). The sequences used in the inter sub-subtype artificial recombinants have been created from the following sub-subtypes and parents, respectively: A1 and A2 (A1: AF413987, A2:AF286240); A2 and A1 (A2:AF286241, A1:AF539405); B and D (B: AF538302, D:AJ320484); D and B (D: AJ488926, B:AY352275); F1 and F2 (F1:AY173957, F2:AF377956); F2 and F1 (F2: AY371158, F1:AY173958). We selected the above combinations of subtypes for the parent sequences, because they correspond to known real-world recombinants. From each subtype, we selected parent sequences for which breakpoints are assumed to be reliably annotated. Figure 7A illustrates the creation of these artificial recombinants, Figure 7B shows the evolutionary relations of the subtypes and inter sub-subtypes used for our simulated recombinants.

Based on these artificial recombinants, we evaluated the performance of our jpHMM tool and compared it with *Simplot* [12], a widely used HIV subtyping tool. We measured the distances between the predicted and the real breakpoints, and assessed the differences in prediction accuracy using non-parametric statistics, namely calculating the (*a*) the *median* value of the distances and (*b*) the *interquartile range*, and comparing distributions using the Wilcoxon signed-rank test implemented with R http://www.r-project.org. As shown in Figure 8, our method consistently showed much better predictions of the artificial recombinant breakpoints than Simplot. In the inter-subtype sequences, jpHMM's median value for the distances between the predicted positions and the real breakpoints is 10, with an interquartile range from 4 to 15. By contrast, Simplot's median is 54, while its interquartile range is from 19 to 72. The difference between the predictions of jpHMM and Simplot is significant with $p < 10^{-9}$ in the Wilcoxon signed-rank test.

The inter sub-subtype simulated sequences are more similar to each other, and so it becomes a more difficult problem to distinguish breakpoints. Here, jpHMM's median value for the distances between the predicted and the actual breakpoints is 9, the interquartile range is from 3.5 to 19, while Simplot's median value is 84 and its interquartile range is 19.5 to 122. The accuracy difference between jpHMM and Simplot's predictions are significant with $p < 10^{-7}$ in the Wilcoxon signed-rank test. Finally, Figure 8 also shows there were no particular breakpoints that were consistently hard to define, rather whether or not a particular breakpoint (for example, position 1000) was accurately resolved depended on the particular combination of sequences; the artificial breakpoints we introduced were embedded in both conserved and variable regions of HIV. Introducing breakpoints at intervals of 500 and 1500 gave comparable results (data not shown) to the 1000 base intervals included in Figures 7 and 8. Finally, the breakpoint definition methods in Simplot uses a chi squared statistic for resolution [34].

We have tried to compare jphmm and Simplot chi square results to the suite of programs available in the RDP package [35], including RDP, GENECONV, MaxChi, Chimaera, Siscan. While Simplot and jphmm readily recognized the artificially generated breakpoints in our recombinants, shown in Figure 7, and could distinguish the parental subtypes, the other methods missed many of the breakpoints and often assigned incorrect subtype designations. The LARD [36] program appears not to be designed for recognition of multiple breakpoints. Bootscanning works well for correct identification of the subtypes of parental fragments in a recombinant genome, whether using the Simplot or RDP implementation, but does not attempt to optimally resolve breakpoints. Another algorithm to detect recombination events has been described in [37,38]. However, this method is limited to detect chimeric sequences that are recombinations of only *two* sequences with only one breakpoint. While the jumping alignment program JALI [8,9] can be adapted to run on DNA sequences, its computer memory requirements are far too high for applying it to the test data in our study. Thus we used the chi squared method [11,34] implemented in Simplot [12] for Figure 8, as it gave the best results among existing methods.

In addition to simulated recombinants, we used real-world *circulating recombinant forms (CRFs)* for which the recombination breakpoints have been very carefully defined, and published in the literature. Here, we compared only our jpHMM predictions to the published data but not predictions by Simplot, since the published breakpoints already mainly rely on predictions by *Simplot* or similar methods. Thus, it would be redundant to compare CRFs to our own revised Simplot predictions. We tested reference sequences from 12 different CRFs, namely CRF02 to CRF08 and CRF10 to CRF14. These recombinants are known to be well annotated. Figure 9 shows the published genome map of CRF02 together with the subtypes as predicted by our jpHMM software. For the 12 CRFs that we analyzed, subtypes predicted by jpHMM roughly correspond to the previously published subtypes: for 70% of the CRFs with breakpoints, the breakpoints predicted by jpHMM are located in a distance of <150*nt* from the corresponding published breakpoints (with an

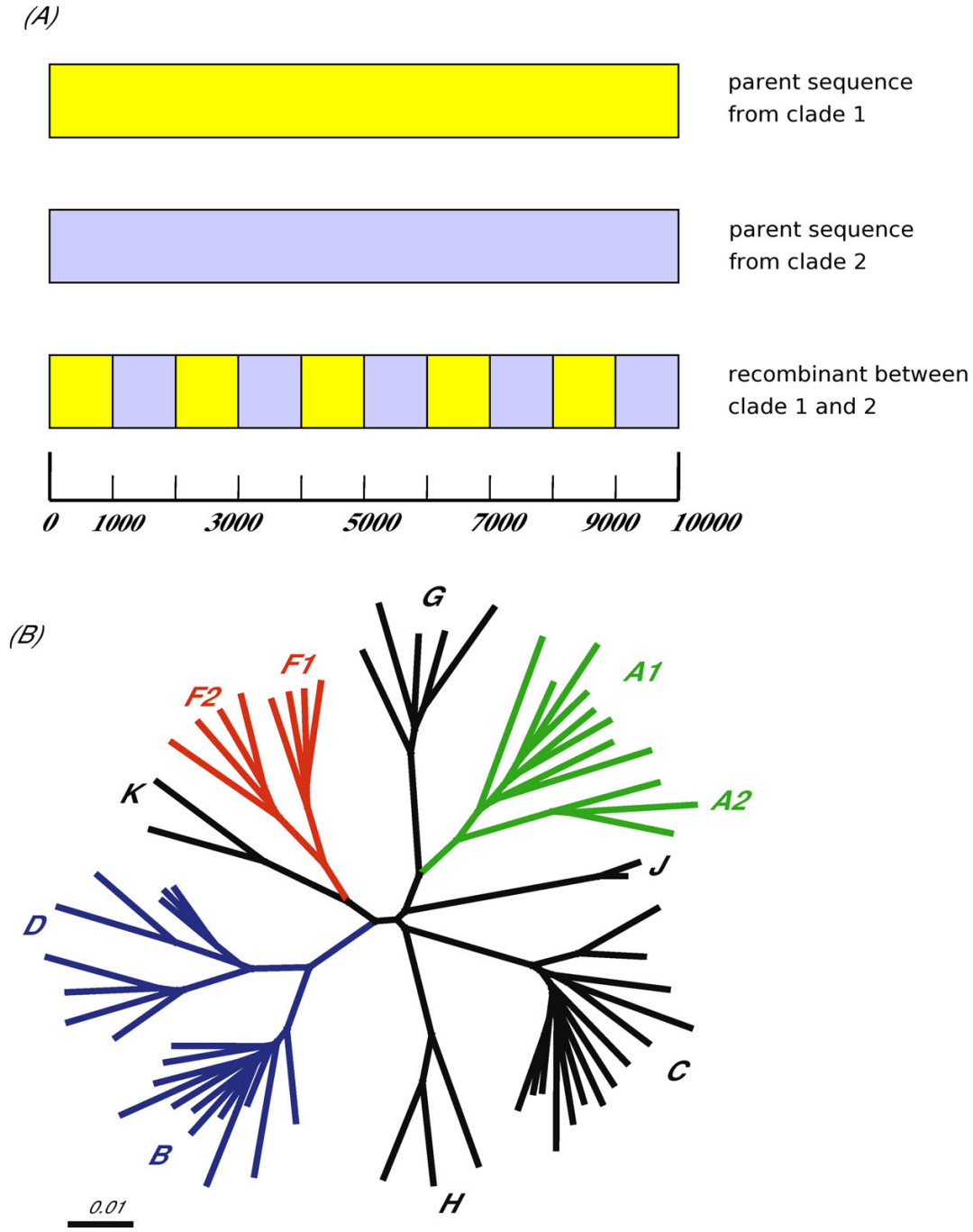


Figure 7
Creation of artificial recombinants with known breakpoints to test jpHMM and Simplot accuracy. (A) The artificial recombinant, constructed from two different clades, has the actual breakpoints at every X-th nucleotide for X = 500, 1000, 1500. Only the construction with X = 1000 is shown here. (B) The phylogenetic tree demonstrates the relations and relative distances between the clades used in the artificial recombinants' construction.

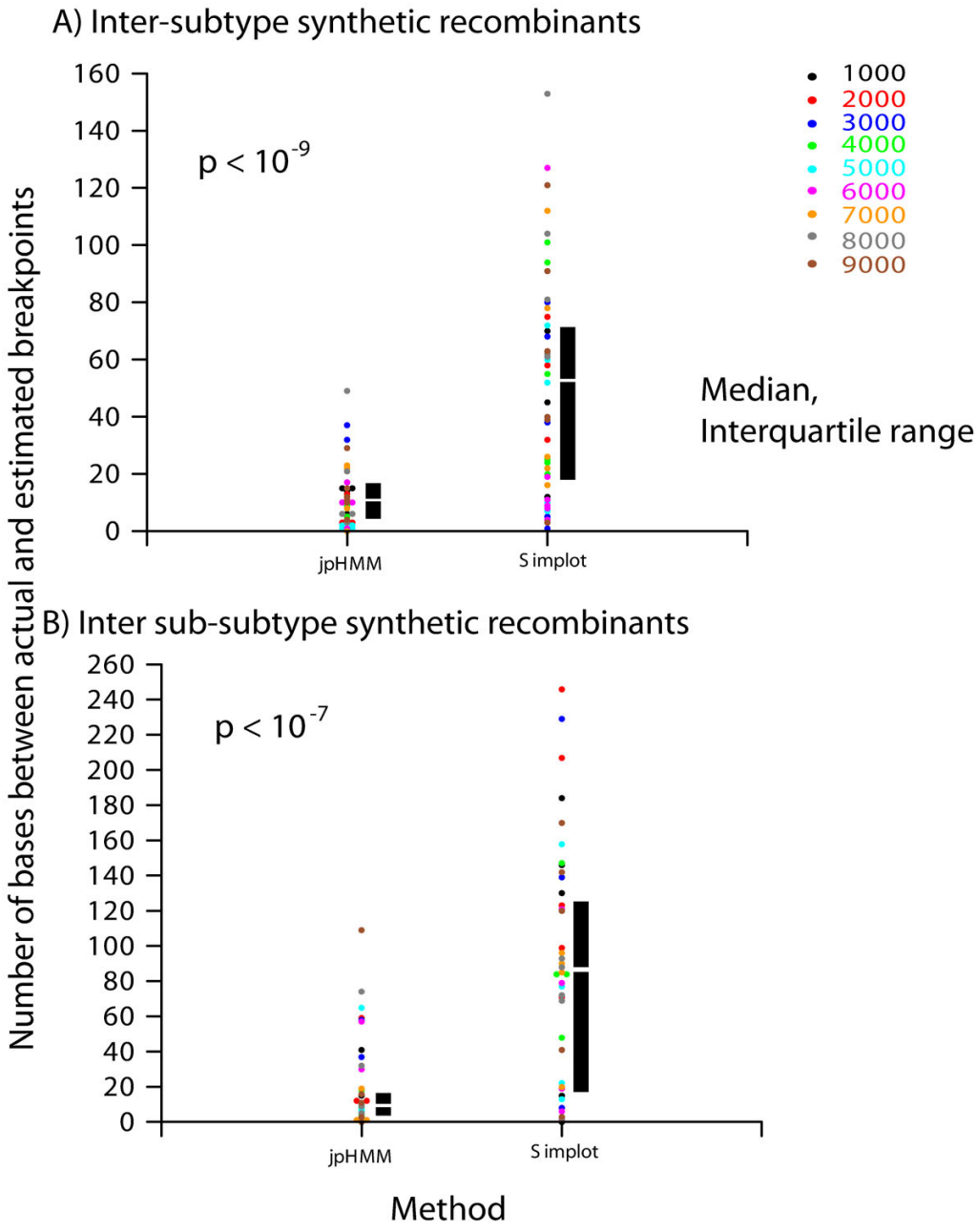


Figure 8
Evaluation of breakpoints predicted by jpHMM and Simplot. We measured the distance between the predicted breakpoints and the real breakpoints as described in the literature. For these distances, we calculated the median value and the interquartile range. A) Inter-subtype artificial recombinants: jpHMM's median is 10, with the interquartile range of 4–15; Simplot's median is 54, with the interquartile range of 19–72 and $p < 10^{-9}$ in the Wilcoxon signed-rank test. B) Inter sub-subtype artificial recombinants: jpHMM's median values 9, interquartile ranges 3.5–19; Simplot's median values 84, interquartile ranges 19.5–122 and $p < 10^{-7}$ in the Wilcoxon signed-rank test. In both plots (A and B), the *Y* -axis represents the number of bases that predicted breakpoints were away from the actual ones. The median values are shown here as the white bars, and the interquartile ranges are shown as the black bars. Numbers in color represent the actual breakpoint positions (every 1000-th nucleotide) in all synthetic recombinants. The duplicated data positions were spread apart in order to show every individual breakpoint here.

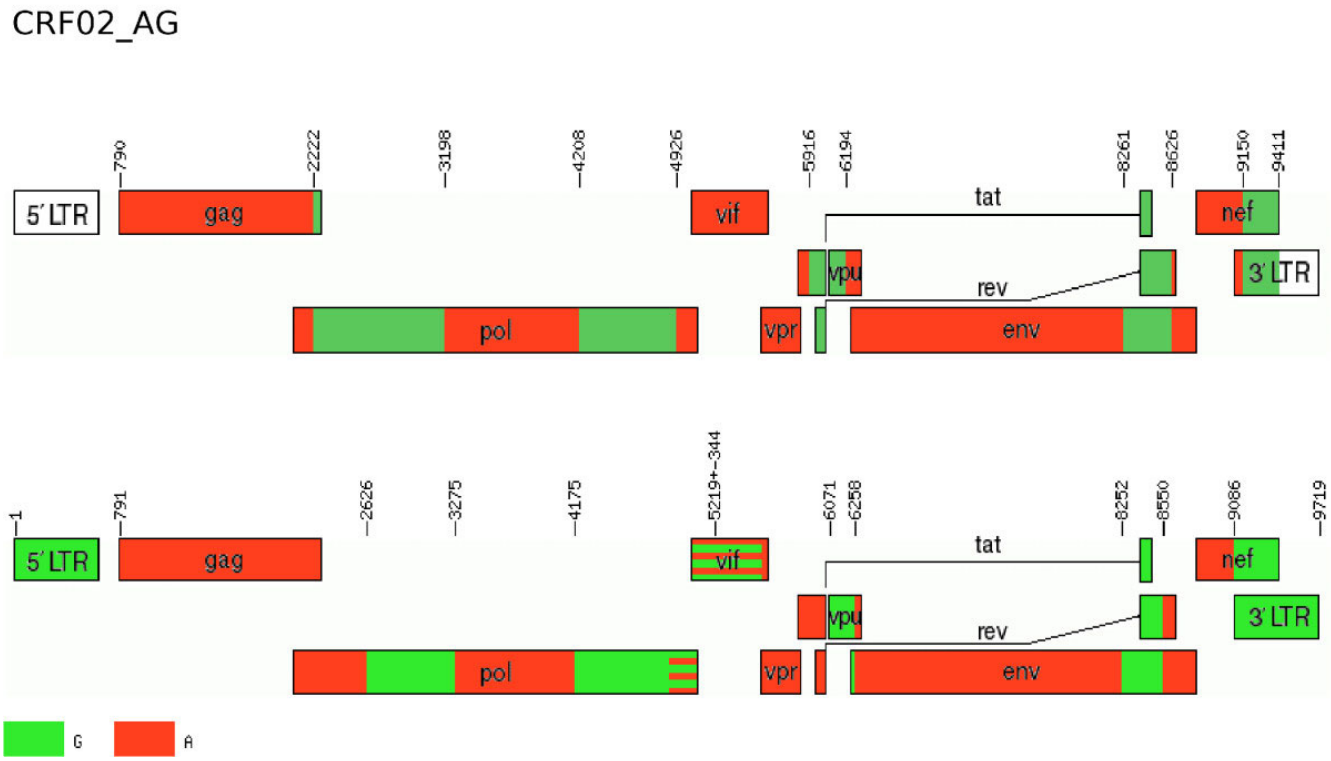


Figure 9
Comparison of genomic recombinations predicted by our jpHMM tool and reported in the literature. In all reference sequences from CRF02 to CRF08 and CRF10 to CRF14, 70% of the jpHMM's predictions were consistent with the published data. As an example, the figure shows the CRF02 recombinant that consists of subtypes A (shown in red) and G (shown in green). Above is the prediction by jpHMM, below the recombination as reported in the literature (see [42]).

average distance of 27*nt*). Discrepancies between the published and jpHMM predicted recombinant sequences were found in the H, J and K-containing CRFs.

Results for the hepatitis C virus

We analyzed the two recombinant strains of HCV that were available until mid-2005, the 1b/2k St Petersburg recombinants (AY587845, [22]) and an artificial 1a/2a recombinant (AF177037, [39]). In both cases, the jpHMM method accurately reconstructed the recombinant. jpHMM located the breakpoint for the 1b/2k St Petersburg recombinant between nucleotide 3186 and 3187 (in HCV-H77 numbering). The original authors manually pinpointed this breakpoint to the exact same nucleotide, based on a Simplot graphic and a sequence alignment.

The location of the breakpoint of the artificial 1a/2a recombinant was estimated to be at position 2759/2760 (HCV-H77 numbering), while the cross-over point in the actual artificial recombinant was reported to be between the fourth and fifth nucleotide of a mutated restriction site Nde 1 located at position 2761–66 of their reference sequence (the boundary of p7 and NS2). However, in the same reference sequence (AF177037) the only site (CATATG) was located at positions 2773–2778, which corresponds to 2762–2767 of HCV-H77. This would place the breakpoint at position 2765/2766, so in this case the prediction is off by 6 nucleotides.

In both recombinants, the genotype of the 5'UTR part of the sequence was misidentified, as 1a for the 1a/2a recombinant and as 2a for the 1b/2k recombinant. Thus, a spurious breakpoint was postulated for both sequences, at position 238/239 for the 1b/2k recombinant and at position 349/350 for the 1a/2a recombinant. This region of the HCV genome, around 350 nucleotides long, is known to be too highly conserved to contain a good phylogenetic signal, and often cannot even be used to phylogenetically distinguish different genotypes, let alone subtypes (CK, unpublished results), so it is not surprising that the jpHMM method is unable to make an accurate determination. As a consequence, for any automatic recombination

detection method to work accurately, we expect that this region will have to be excluded a priori; and conversely, because this region does contain little phylogenetic information, detection of recombination will be almost impossible.

Discussion and conclusions

We developed *jpHMM*, a novel probabilistic approach to compare DNA or protein sequences to a family of aligned sequences. In this study, we applied this tool to sequence subtyping and classification to enhance viral sequence quality control in the rapidly expanding HIV/HCV sequence databases. jpHMM combines the idea of a profile HMM with the jumping-alignment approach that has been previously proposed by Spang *et al*. [9]. For HIV and HCV genome sequences, we constructed profile HMMs for each subtype of the respective virus; these models were then connected by subtype transitions (jumps). These jumps make it possible to detect whether a query sequence is an inter-subtype recombinant by finding a best reference subtype match at each position along the entire query sequence. The results presented in this paper demonstrate jpHMM sensitively recognizes recombinants and gives more accurate breakpoint predictions than Simplot, a widely used subtyping tool in HIV-1 sequence analysis.

As every probabilistic model, jpHMM depends on a sufficiently large set of input data; our approach is therefore limited by the subtype background sets which are used as the model-building sequences. Our method performs best with large input data sets to inform the model, but it may fail to identify breakpoints if the input data set is too small. In the present study, for example, jpHMM failed on H, J, and K-containing CRFs. The difficulties with these sequences could be due to the following reasons: (1) jpHMM underestimates H, J, and K subtypes due to the fact that they have very rarely been sampled and sequenced, and so there are inadequate complete genome sequences from these three subtypes to develop a good model. (2) The current H, J, and K subtype reference sequences probably are not good representatives for these three subtypes, thus our jpHMM, as other subtyping tools, can be biased predicting these particular subtypes.

Thus, in its current form, while jpHMM provides clearly superior accuracy in terms of breakpoint definitions when large subtype data sets were available for input, to resolve rare subtypes it would be best to use this tool in conjunction with RIP [40] or Simplot for *de novo* HIV classification of unknown sequences. In this way, recombinant fragments from rare subtypes could be detected if present, and more accurate breakpoint definitions between common subtypes would be possible. In addition, jpHMM, like many other subtyping tools, fails on sequence classifications in the situation where a new or unknown sequence is discovered because there is no reference sequence available. We are currently developing another method to solve this problem.

In the present study, we applied jpHMM to viruses; we tested it on HIV/HCV sequences. The method, however, has been developed as a generally applicable tool, so its application should not be considered only in viral genomes. It could be successfully used to DNA and protein sequences from other organisms with individual subtype's sequences available, and be used as one important part in understanding the role of recombination in evolution and molecular epidemiology, and ultimately for integration into sequence quality control pipelines as a standard step in sequence analysis.

Availability and requirements

The jpHMM program was written in C++ and the source code is available free of charge from the authors on request. We set up a user-friendly WWW interface for the program at [41] which is described [42]. The circulating recombinant forms of HIV are listed on a web page [43] of the HIV Sequence Database.

Authors' contributions

AKS developed, implemented and tested the jpHMM algorithm, MZ and CK evaluated the program on HIV and HCV data, TL, BK and BM guided the project, MS conceived the jpHMM approach and supervised the program development. Each author wrote a part of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This project was funded in part by grant *NIH Y1-AI-1500-01*, the *NIH-DOE interagency agreement*, the *HIV Immunology and Sequence Database* and by grant MO 1048/1-1 of the *Deutsche Forschungsgemeinschaft*. We thank Stephan Waack for discussions and suggestions that lead to the design of a HMM that models jumping alignments and the anonymous referees for their suggestions.

References

1. Krogh A, Brown M, Mian I, Sjolander K, Haussler D: **Hidden Markov Models in Computational Biology: Applications to protein modelling.** *J Mol Biology* 1994, **235:**1501-1531.
2. Eddy S: **Profile hidden Markov models.** *Bioinformatics* 1998, **14(9):**755-763.
3. Eddy S: **Hidden Markov Models.** *Current Opinion in Structural Biology* 1996, **6:**361-365.
4. Viterbi A: **Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.** *IEEE Trans Inform Theory* 1967, **IT-13:**260-269.
5. Durbin R, Eddy SR, Krogh A, Mitchison G: *Biological sequence analysis* Cambridge, UK: Cambridge University Press; 1998.
6. Altschul SF, Gish W, Miller W, Myers EM, Lipman DJ: **Basic Local Alignment Search Tool.** *J Mol Biol* 1990, **215:**403-410.
7. **HMMER web page** [http://hmmer.wustl.edu]
8. Spang R, Rehmsmeier M, Stoye J: **Sequence Database Search Using Jumping Alignments.** *Proceedings of ISMB 2000* 2000.

9. Spang R, Rehmsmeier M, Stoye J: **A Novel Approach to Remote Homology Detection: Jumping Alignments.** *Journal of Computational Biology* 2002, **9:**747-760.
10. Siepel AC, Halpern AL, Macken C, Korber BT: **A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences.** *AIDS Res Hum Retroviruses* 1995, **11:**1413-1416.
11. Robertson DL, Sharp PM, McCutchan FE, Hahn BH: **Recombination in HIV-1.** *Nature* 1995, **374:**124-126.
12. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, Ingersoll R, Sheppard HW, Ray SC: **Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination.** *J Virology* 1999, **73:**152-160.
13. Salminen MO, Carr JK, Burke DS, McCutchan FE: **Identification of Breakpoints in In-Tergenotypic Recombinants of HIV Type-1 by Bootscanning.** *AIDS Res and Human Retroviruses* 1995, **11:**1423-1425.
14. Sharp PM, Shaw GM, Hahn BH: **Simian Immunodeficiency Virus Infection of Chimpanzees.** *J Virol* 2005, **79(7):**3891-3902.
15. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leitner T, McCutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S, Korber B: **HIV-1 nomenclature proposal.** *Science* 2000, **288:**55-57.
16. Hoelscher M, Dowling WE, Sanders-Buell E, Carr JK, Harris ME, Thomschke A, Robb ML, Birx DL, McCutchan FE: **Detection of HIV-1 subtypes, recombinants, and dual infections in East Africa by a multi-region hybridization assay.** *AIDS* 2002, **16:**2055-2064.
17. Simmonds P, Bukh J, Combet C, Deleage G, Enomoto N, Feinstone S, Halfon P, Inchauspe G, Kuiken C, Maertens G, Mizokami M, Murphy DG, Okamoto H, Pawlotsky JM, Penin F, Sablon E, Shin-I T, Stuyver LJ, Thiel HJ, Viazov S, Weiner AJ, Widell A: **Consensus Proposals for a Unified System of Nomenclature of Hepatitis C Virus Genotypes.** *Hepatology* in press.
18. Kijak GH, Sanders-Buell E, Wolfe ND, Mpoudi-Ngole E, Kim B, Brown B, Robb ML, Birx DL, Burke DS, Carr JK, McCutchan FE: **Development and application of a high-throughput HIV type 1 genotyping assay to identify CRF02 AG in West/West Central Africa.** *AIDS Res and Human Retroviruses* 2004, **20:**521-530.
19. Radkowski M, Wang LF, Vargas H, Wilkinson J, Rakela J, Laskus T: **Changes in hepatitis C virus population in serum and peripheral blood mononuclear cells in chronically infected patients receiving liver graft from infected donors.** *Transplantation* 2001, **72:**833-838.
20. Laskus T, Wang LF, Radkowski M, Vargas H, Nowicki M, Wilkinson J, Rakela J: **Exposure of hepatitis C virus (HCV) RNA-positive recipients to HCV RNA-positive blood donors results in rapid predominance of a single donor strain and exclusion and/or suppression of the recipient strain.** *J Virology* 2001, **75:**2059-2066.
21. Eyster ME, Sherman KE, Goedert JJ, Katsoulidou A, Hatzakis A: **Prevalence and changes in hepatitis C virus genotypes among multitransfused persons with hemophilia. The Multicenter Hemophilia Cohort Study.** *J Infect Dis* 1999, **179:**1062-1069.
22. Kao JH, Chen PJ, Wang JT, Yang PM, Lai MY, Wang TH, Chen DS: **Superinfection by homotypic virus in hepatitis C virus carriers: studies on patients with post-transfusion hepatitis.** *J Med Virol* 1996, **50:**303-308.
23. Zhang S, Hui Z, Li H, Qi Z, Widell A: **Dynamic changes in hepatitis C virus genotypes and sequence patterns in plasma donors exposed to reinfection.** *J Med Virol* 2001, **63:**228-236.
24. Widell A, Mansson S, Persson NH, Thysell H, Hermodsson S, Blohme I: **Hepatitis C superinfection in hepatitis C virus (HCV)-infected patients transplanted with an HCV-infected kidney.** *Transplantation* 1995, **60:**642-647.
25. Kalinina O, Norder H, Mukomolov S, Magnius LO: **A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg.** *J Virology* 2002, **76:**4034-4043.
26. Colina R, Casane D, Vasquez S, Garcia-Aguirre L, Chunga A, Romero H, Khan B, Cristina J: **Evidence of intratypic recombination in natural populations of hepatitis C virus.** *J General Virology* 2004, **85:**31-37.
27. McCutchan FE, Sankale JL, MBoup S, Kim B, Tovanabutra S, Hamel DJ, Brodine SK, Kanki PJ, Birx DL: **HIV type 1 circulating recombinant form CRF09 cpx from West Africa combines subtypes A, F, G, and may share ancestors with CRF02 AG and Z321.** *AIDS Res Hum Retroviruses* 2004, **20:**819-826.
28. Morgenstern B, Dress A, Werner T: **Multiple DNA and protein sequence alignment based on segment-to-segment comparison.** *Proc Natl Acad Sci USA* 1996, **93:**12098-12103.
29. Gelfand MS, Mironov AA, Pevzner PA: **Gene recognition via spliced sequence alignment.** *Proc Natl Acad Sci USA* 1996, **93(17):**9061-9066.
30. Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian I, Haussler D: **Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology.** *Comput Appl Biosci* 1996, **12(4):**327-345.
31. Wistrand M, Sonnhammer E: **Transition Priors for Protein Hidden Markov Models: An Empirical Study towards Maximum Discrimination.** *J Comp Biol* 2002, **11:**181-193.
32. Lowerre B: **The Harpy Speech Recognition System.** In *Tech rep* Carnegie-Mellon University; 1976.
33. Plötz T, Fink GA: **Accelerating the Evaluation of Profile HMMs by Pruning Techniques.** In *Tech rep* University of Bielefeld, Faculty of Technology; 2004. [Report 2004–03]
34. Smith JM: **Analyzing the mosaic structure of genes.** *J Mol Evol* 1992, **34:**126-129.
35. Martin DP, Williamson C, Posada D: **RDP2: recombination detection and analysis from sequence alignments.** *Bioinformatics* 2005, **21:**260-262.
36. Holmes EC, Worobey M, Rambaut A: **Phylogenetic Evidence for Recombination in Dengue Virus.** *Mol Biol Evol* 1999, **16:**405-409.
37. Komatsoulis GA, Waterman MS: **Chimeric alignment by dynamic programming: algorithm and biological uses.** In *RECOMB '97: Proceedings of the first annual international conference on Computational molecular biology* New York, NY, USA: ACM Press; 1997:174-180.
38. Komatsoulis GA, Waterman MS: **A new computational method for detection of chimeric 16S rRNA artifacts generated by PCR amplification from mixed bacterial populations.** *Appl Envir Microbiol* 1997, **63:**2338-2346.
39. Yanagi M, Purcell RH, Emerson SU, Bukh J: **Hepatitis C virus: An infectious molecular clone of a second major genotype (2a) and lack of viability of intertypic 1a and 2a chimeras.** *Virology* 1999, **262:**250-263.
40. **RIP web page** [http://hiv-web.lanl.gov/content/hiv-db/RIPPER/RIP.html]
41. **jpHMM web server** [http://jphmm.gobics.de]
42. Zhang Ming, Schultz Anne-Kathrin, Calef Charles, Kuiken Carla, Leitner Thomas, Korber Bette, Morgenstern Burkhard, Stanke Mario: **jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1.** Nucleic Acids research.
43. **circulating recombinant forms of HIV** [http://hiv-web.lanl.gov/content/hiv-db/CRFs/CRFs.html]

JOURNAL OF VIROLOGY, Dec. 2006, p. 11776–11790 Vol. 80, No. 23
0022-538X/06/$08.00+0 doi:10.1128/JVI.01730-06

Genetic and Neutralization Properties of Subtype C Human Immunodeficiency Virus Type 1 Molecular *env* Clones from Acute and Early Heterosexually Acquired Infections in Southern Africa▽

Ming Li,[1] Jesus F. Salazar-Gonzalez,[2] Cynthia A. Derdeyn,[3] Lynn Morris,[4] Carolyn Williamson,[5] James E. Robinson,[6] Julie M. Decker,[2] Yingying Li,[2] Maria G. Salazar,[2] Victoria R. Polonis,[7] Koleka Mlisana,[8] Salim Abdool Karim,[8] Kunxue Hong,[9] Kelli M. Greene,[1] Miroslawa Bilska,[1] Jintao Zhou,[1] Susan Allen,[10] Elwyn Chomba,[11] Joseph Mulenga,[12] Cheswa Vwalika,[13] Feng Gao,[14] Ming Zhang,[15] Bette T. M. Korber,[15] Eric Hunter,[3] Beatrice H. Hahn,[2] and David C. Montefiori[1]*

Departments of Surgery[1] and Medicine,[14] Duke University Medical Center, Durham, North Carolina 27710; Department of Medicine, University of Alabama, Birmingham, Alabama 35294[2]; Department of Pathology and Laboratory Medicine, Emory University, Atlanta, Georgia 30329[3]; National Institute for Communicable Diseases, Johannesburg, South Africa[4]; Institute of Infectious Diseases and Molecular Medicine, University of Cape Town, Cape Town, South Africa[5]; Department of Pediatrics, Tulane University Medical Center, New Orleans, Louisiana 70112[6]; Walter Reed Army Institute of Research, Rockville, Maryland 20850[7]; CAPRISA, University of KwaZulu-Natal, Durban, South Africa[8]; Division of Virology and Immunology, National Center for AIDS, Beijing, China[9]; Department of Global Health, Emory University, Atlanta, Georgia[10]; University Teaching Hospital,[11] Zambia National Blood Transfusion Service,[12] and Zambia-Emory HIV Research Program,[13] Lusaka, Zambia; and Los Alamos National Laboratory, Los Alamos, New Mexico 87545[15]

Received 10 August 2006/Accepted 5 September 2006

A standard panel of subtype C human immunodeficiency virus type 1 (HIV-1) Env-pseudotyped viruses was created by cloning, sequencing, and characterizing functional gp160 genes from 18 acute and early heterosexually acquired infections in South Africa and Zambia. In general, the gp120 region of these clones was shorter (most evident in V1 and V4) and less glycosylated compared to newly transmitted subtype B viruses, and it was underglycosylated but no different in length compared to chronic subtype C viruses. The gp120s also exhibited low amino acid sequence variability (12%) in V3 and high variability (39%) immediately downstream of V3, a feature shared with newly transmitted subtype B viruses and chronic viruses of both subtypes. When tested as Env-pseudotyped viruses in a luciferase reporter gene assay, all clones possessed an R5 phenotype and resembled primary isolates in their sensitivity to neutralization by HIV-1-positive plasmas. Results obtained with a multisubtype plasma panel suggested partial subtype preference in the neutralizing antibody response to infection. The clones were typical of subtype C in that all were resistant to 2G12 (associated with loss of N-glycosylation at position 295) and most were resistant to 2F5, but all were sensitive to 4E10 and many were sensitive to immunoglobulin G1b12. Finally, conserved neutralization epitopes in the CD4-induced coreceptor binding domain of gp120 were poorly accessible and were difficult to induce and stabilize with soluble CD4 on Env-pseudotyped viruses. These results illustrate key genetic and antigenic properties of subtype C HIV-1 that might impact the design and testing of candidate vaccines. A subset of these gp160 clones are suitable for use as reference reagents to facilitate standardized assessments of vaccine-elicited neutralizing antibody responses.

Neutralizing antibody (NAb) responses are associated with the clinical success of many approved vaccines (65) and are a high priority for human immunodeficiency virus type 1 (HIV-1) vaccine development (25, 43, 46). To be effective against HIV-1, NAbs will need to overcome the extensive genetic diversity and complex escape mechanisms that typify the surface gp120 and transmembrane gp41 envelope glycoproteins (Env) of the virus (40, 84, 89). At least nine different genetic subtypes and a growing number of circulating recombinant forms (CRFs) of group M HIV-1 account for the majority of infections worldwide (42). Unfortunately, little progress has been made in designing a vaccine immunogen that elicits NAbs against multiple variants within a single genetic subtype, let alone cross-subtype NAbs (4, 5, 10, 49). A variety of new approaches aim to solve this difficult problem by acquiring new knowledge about Env structure, function, and immunobiology and using this knowledge to design better immunogens (12, 34). As new immunogens undergo preclinical and clinical testing, it will be important to compare them to prototypic immunogens and to each other with respect to the magnitude and breadth of the NAb response each generates.

To facilitate these comparative studies, it has been recom-

* Corresponding author. Mailing address: Department of Surgery, Laboratory for AIDS Vaccine Research and Development, P.O. Box 2926, Duke University Medical Center, Durham, NC 27710. Phone: (919) 684-5278. Fax: (919) 684-4288. E-mail: monte@acpub.duke.edu.
▽ Published ahead of print on 13 September 2006.

mended that separate panels of HIV-1 reference strains be devised for each major genetic subtype and CRF; these panels are needed to acquire standardized data sets that may be used to rank vaccine potency and to identify promising candidates for further development (47). The degree of accuracy in predicting vaccine potency with standard reference strains could depend on the particular genetic, antigenic, and biologic properties of these viruses (54). Deciding which viral properties are most suitable is a complicated task that would best be guided by information on an HIV-1 vaccine that is at least partially effective. Because no such vaccine is currently available, the process of selecting appropriate reference strains has instead relied on an alternate set of scientific judgments (47, 54). These judgments place a heavy emphasis on the use of early "transmitted" strains from sexually acquired infections with the rationale that sexual contact is the major route of HIV-1 transmission in the world (31, 79) and that a bottleneck occurs at sexual transmission that selects a subset of viral variants (18, 23, 29, 82, 86, 93, 94). In principle, these sexually transmissible variants are the major targets for vaccination and also represent suitable reference strains for immune monitoring assays.

Recently, a panel of 12 subtype B gp160 reference clones from acute/early sexually acquired infections was described that may be used as Env-pseudotyped viruses for standardized assessments of NAb responses (45). There is an urgent need to develop a separate panel for subtype C, as this is the most abundant subtype in countries that carry the heaviest burden of infections (50). Previous studies of HIV-1 neutralization have focused on the less-prevalent subtype B. Moreover, what little is known about subtype C is derived mostly from studies of chronic infection. In general, subtype C viruses have been shown to be sensitive to neutralization by subtype C and non-subtype C serum samples where only occasional subtype-specific neutralization has been observed (11, 39, 55). Subtype C viruses also are unusually resistant to the gp120 glycan-specific monoclonal Ab 2G12 and to the gp41-specific monoclonal Ab 2F5 (7, 11, 33). Another general feature of subtype C is a low level of amino acid diversity in the V3 loop of gp120 and a high level of diversity in the region immediately downstream of V3 (24, 59).

A recent study of heterosexual subtype C HIV-1 transmission in Zambia showed that gp120 from newly transmitted viruses were shorter, less glycosylated, and more sensitive to neutralization by plasma from the chronically infected partner compared to the transmitted virus (23). Similar features have been associated with the transmission of subtype A viruses in Kenya (18) but were not seen in subtype B transmission (18, 29). Another recent study found a more robust autologous neutralizing antibody response during early subtype C virus infection in Zambia compared to early subtype B virus infection in the United States, where the robust response to subtype C infection was associated with viruses that contained shorter and less-glycosylated gp120s (44). Additional work is needed to confirm and extend these early observations, as they have important implications for vaccine design and testing. Here we describe the cloning, sequencing, biologic phenotype, and neutralization profile of functional gp160 genes from acute/early, heterosexually acquired subtype C HIV-1 infections in South Africa and Zambia.

MATERIALS AND METHODS

Plasma samples, soluble CD4 (sCD4), and monoclonal Abs. Plasma samples from chronically HIV-1-infected blood donors in South Africa who were presumed to be treatment naive were obtained from the South African National Blood Services. These individual plasma samples were designated BB8, BB12, BB28, BB55, BB70, and BB106 and were selected from a larger set of plasma samples based on their greater neutralizing activities in an initial screen against three subtype C Env-pseudotyped viruses (Du151.2, Du156.12, and Du172.17). All donors were infected with HIV-1 subtype C viruses as determined by *gag* and *env* sequences. Plasma pools for subtypes A, B, C, and D were described previously (45). These latter plasma pools were each comprised of plasma from 6 to 10 subjects who had pure subtype infections as verified by full HIV-1 genome sequencing of DNA from cryopreserved peripheral blood mononuclear cells (PBMC). The plasma samples and plasma pools were derived from chronically infected individuals and did not include the HIV-1-positive subjects who served as a source of *env* clones. A normal plasma pool was prepared from Leukopaks from four HIV-1-negative subjects (BRT Laboratories, Inc., Baltimore, MD). All plasma samples were heat inactivated at 56°C for 1 h prior to use. Informed consent was obtained from all study participants as approved by local institutional review boards and biosafety committees; the blood bank samples were obtained from anonymous donors, with all sample codes delinked from subject identifiers.

Recombinant sCD4 comprising the full-length extracellular domain of human CD4 and produced in Chinese hamster ovary cells was obtained from Progenics Pharmaceuticals, Inc. (Tarrytown, NY). Human monoclonal Ab immunoglobulin G1b12 (IgG1b12) was kindly provided by Dennis Burton (The Scripps Research Institute, La Jolla, CA). Human monoclonal Abs 2G12, 2F5, and 4E10 were obtained from the NIH AIDS Research and Reference Reagent Program (NIH ARRRP; Rockville, MD) as contributed by Hermann Katinger. Monoclonal Abs 17b, 23e, 31H, 21c, E51, 48d, 112d, 412d, and ED10 against CD4-induced (CD4i) epitopes on gp120 were isolated from Epstein-Barr virus-transformed B-cell lines established from subtype B HIV-1-infected individuals as described previously (66, 90). TriMab is a mixture of three monoclonal Abs (IgG1b12, 2G12, and 2F5) prepared as a 1-mg/ml stock solution containing 333 μg of each monoclonal Ab/ml in phosphate-buffered saline, pH 7.4.

Cells. TZM-bl (JC53-bl) cells were obtained from the NIH ARRRP as contributed by John Kappes and Xiaoyun Wu. This is a genetically engineered HeLa cell clone that expresses CD4, CXCR4, and CCR5 and contains Tat-responsive reporter genes for firefly luciferase (Luc) and *Escherichia coli* β-galactosidase under regulatory control of an HIV-1 long terminal repeat (60, 83). 293T/17 cells were obtained from the American Type Culture Collection (catalog no. 11268). Both cell lines were maintained in Dulbecco's modified Eagle's medium (Gibco BRL Life Technologies) containing 10% heat-inactivated fetal bovine serum and 50 μg gentamicin/ml in vented T-75 culture flasks (Corning-Costar). Cultures were incubated at 37°C in a humidified 5% CO2–95% air environment. Cell monolayers were split 1:10 at confluence by treatment with 0.25% trypsin, 1 mM EDTA (Invitrogen).

Amplification and cloning of *env*/*rev* DNA cassettes. Specimens for *env*/*rev* gene cloning were obtained from 18 HIV-1-infected individuals living in either South Africa or Zambia and who were judged to be in acute/early stages of infection as determined by either their last known seronegative clinic visit, time of onset of acute retroviral syndrome, or a combination of these two clinical parameters. All infections occurred through heterosexual contact. *env-rev* cassettes were cloned from either plasma viral RNA, uncultured PBMC DNA, or cultured PBMC DNA. In the latter case, fresh phytohemagglutinin-stimulated PBMC from healthy HIV-1-negative donors were inoculated with low-passage (<5 passages) primary isolates, and the infected PBMC were used as the source of DNA for PCR amplification of HIV-1 *env-rev* cassettes as described elsewhere (45). Virus-containing culture supernatants were clarified from cells by 0.45-μm filtration and stored at −80°C in 1-ml aliquots as archives of the immediate uncloned parent of the corresponding molecularly cloned Env-pseudotyped viruses. In some cases, uncultured PBMC direct from study subjects were used for DNA extraction and PCR amplification. In other cases, virion-associated plasma RNA was purified and subjected to cDNA synthesis prior to PCR amplification as described elsewhere (23, 44, 45). The inner antisense primer EnvNF (5′-GG CACCTGAGGTCTGACTGGAAAGCC-3′) replaced antisense primer EnvN in the second-round PCR and was used in conjunction with inner sense primer EnvA (23) to amplify full-length *env* from ZM197M. PCR products from cultured PBMC DNA were inserted directly into pcDNA 3.1D/V5-His-TOPO (Invitrogen Corp., Carlsbad, CA). PCR products from uncultured PBMC DNA and from plasma viral RNA were inserted into either pcDNA 3.1/V5-His-TOPO or pCR3.1 (Invitrogen Corp., Carlsbad, CA) by TA cloning.

TABLE 1. Demographic and biologic properties of molecularly cloned, Env-pseudotyped strains of subtype C HIV-1

Env clone[a]	Panel designation	Subtype	CoR	Gender	Mode of transmission	Mo/yr isolated	Wks after est. infection date	Location	Plasma VL	CD4 count	Source[b]	Accession no.
Du123.6	SVPC1	C	R5	F	M-F	11/1998	12	Durban, South Africa	19,331	841	ccPBMC	DQ411850
Du151.2	SVPC2	C	R5	F	M-F	11/1998	6	Durban, South Africa	>500,000	367	ccPBMC	DQ411851
Du156.12*	SVPC3*	C	R5	F	M-F	02/1999	<4	Durban, South Africa	22,122	404	ccPBMC	DQ411852
Du172.17*	SVPC4*	C	R5	F	M-F	11/1998	12	Durban, South Africa	1,916	793	ccPBMC	DQ411853
Du422.1*	SVPC5*	C	R5	F	M-F	11/1998	8	Durban, South Africa	17,118	409	ccPBMC	DQ411854
ZM197M.PB7*	SVPC6*	C	R5	M	F-M	10/2002	<15	Lusaka, Zambia	NA[c]	NA	ucPBMC	DQ388515
ZM214M.PL15*	SVPC7*	C	R5	M	F-M	07/2003	<13	Lusaka, Zambia	198,800	NA	Plasma	DQ388516
ZM215F.PB8	SVPC8	C	R5	F	M-F	10/2002	<15	Lusaka, Zambia	608,560	NA	ucPBMC	DQ422948
ZM233M.PB6*	SVPC9*	C	R5	M	F-M	12/2002	<15	Lusaka, Zambia	NA	NA	ucPBMC	DQ388517
ZM249M.PL1*	SVPC10*	C	R5	M	F-M	08/2003	<1	Lusaka, Zambia	1,143,760	NA	Plasma	DQ388514
ZM53M.PB12*	SVPC11*	C	R5	M	F-M	02/2000	<14	Lusaka, Zambia	26,643	NA	ucPBMC	AY423984
ZM55F.PB28a	SVPC12	C	R5	F	M-F	08/1998	<13	Lusaka, Zambia	88,544	NA	ucPBMC	AY423971
ZM109F.PB4*	SVPC13*	C	R5	F	M-F	03/2000	<14	Lusaka, Zambia	887,586	NA	ucPBMC	AY424138
ZM106F.PB9	SVPC14	C	R5	F	M-F	06/1998	<18	Lusaka, Zambia	48,442	NA	ucPBMC	AY424163
ZM135M.PL10a*	SVPC15*	C	R5	M	F-M	06/1998	<15	Lusaka, Zambia	202,999	NA	Plasma	AY424079
CAP45.2.00.G3*	SVPC16*	C	R5	F	M-F	05/2005	5	Durban, South Africa	236,000	974	Plasma	DQ435682
CAP210.2.00.E8*	SVPC17*	C	R5	F	M-F	05/2005	5	Durban, South Africa	127,000	461	Plasma	DQ435683
CAP244.2.00.D3	SVPC18	C	R5	F	M-F	05/2005	8	Durban, South Africa	19,200	557	Plasma	DQ435684

[a] Clones selected as standard reference strains are marked with an asterisk.
[b] ccPBMC, cocultured PBMC; ucPBMC, uncultured PBMC.
[c] NA, not available.

Plasmid minipreps from multiple colonies of transformed JM109 cells were screened by restriction enzyme digestion for full-length inserts. Clones with inserts in the correct orientation were screened by cotransfection with an *env*-deficient HIV-1 (SG3Δenv) backbone in 293T cells to produce Env-pseudotyped viruses that were subsequently titrated for infectivity in TZM-bl cells as described previously (45). Env clones conferring the highest infectivity were selected for further characterization. Twelve of these molecularly cloned gp160 genes were chosen to comprise a new panel of standard reference reagents and have been donated to the NIH ARRRP (item no. 11326).

Other viruses. A functional gp160 clone for the R5 subtype B virus SF162.LS (17, 73) was obtained from Leonidas Stamatatos. T-cell line-adapted HIV-1_{MN} (30) was obtained from Robert Gallo and was propagated in H9 cells as described elsewhere (53). Molecularly cloned gp160 genes from a standard panel of subtype B reference strains was described previously (45). The CD4-independent strain NL-ADArs (38) was obtained from Joseph Sodroski and was propagated in PBMC.

DNA sequence and phylogenetic analysis. Sequence analysis was performed by cycle sequencing and BigDye terminator chemistry with automated DNA Sequenators (models 3100 and 3730; Applied Biosystems, Inc.) as recommended by the manufacturer. Individual sequence fragments for each *env* clone were assembled and edited using the Sequencher program 4.2 (Gene Codes Corp., Ann Arbor, MI). Nucleotide and deduced Env amino acid sequences were initially aligned using CLUSTAL W (35, 76) and manually adjusted for an optimal alignment using MASE (27). Pair-wise evolutionary distances were estimated using Kimura's two-parameter method (37) to correct for superimposed substitutions; sequence gaps and ambiguous areas within the alignment were excluded from all comparisons. Phylogenetic trees were constructed by the neighbor-joining method (67), and the reliability of branching orders was assessed by bootstrap analysis using 1,000 replicates (28). The sequences of all new gp160 genes are available from GenBank and the Los Alamos HIV Sequence Database (see Table 1 for accession numbers).

Analysis of coreceptor usage. Coreceptor usage of 15 Env-pseudotyped viruses was determined in TZM-bl cells by measuring reductions in infectivity in the presence of the CXCR4 antagonist AMD 3100 and the CCR5 antagonist TAK-779 as described elsewhere (45). For the three CAP clones, coreceptor usage was determined by green fluorescence protein (GFP) reporter gene activation in GHOST cells that expressed either CCR5 or CXCR4 (19).

Neutralization assay. Neutralization was measured as a reduction in Luc reporter gene expression after a single round of virus infection in TZM-bl cells as described previously (52). This assay is a modified version of the assay used by Wei et al. (83, 84). The 50% inhibitory dose (ID_{50}) was defined as either the plasma dilution or sample concentration (in the case of sCD4 and monoclonal Abs) at which relative luminescence units (RLU) were reduced 50% compared to virus control wells after subtraction of background RLU in cell control wells.

Statistical analyses. Differences in the neutralizing potencies of subtype C plasma samples (six samples with a BB prefix) were determined by comparing their geometric mean titer (GMT) against each virus. Differences in neutralization sensitivity between subtype B and C viruses were compared using a two-sided Wilcoxon matched pairs test with a 95% confidence interval. Differences were considered significant if P was <0.05. Differences in envelope lengths and number of potential N-linked glycans (PNLG) between subtype B and C viruses were compared using the Wilcoxon two-sided rank test. The two-sided Fisher's exact test was used to determine the relative loss of specific PNLG in one HIV-1 subtype compared to another. Both of these latter tests were performed using the R Project for Statistical Computing (www.r-project.org).

Nucleotide sequence accession numbers. GenBank accession numbers for the sequences identified in this study are DQ411850 to DQ411854, DQ388514 to DQ388517, DQ422948, DQ435682 to DQ435684, AY423971, AY423984, AY424079, AY424138, and AY424163.

RESULTS

Demographics and biologic properties of the molecularly cloned gp160 genes. Candidate gp160 reference clones were obtained from studies of acute and early sexually acquired subtype C infections in South Africa and Zambia (Table 1). Five clones with a Du prefix were obtained in 1998 from commercial sex workers recruited from truck stops between Durban and Johannesburg (11); these women were participating in a multicenter clinical trial of a potential vaginal microbicide (81). Clones with a ZM prefix were obtained between June 1998 and July 2003 from a study of HIV-1-discordant couples in Lusaka, Zambia (1, 23). Clones with a CAP prefix were obtained in 2005 from a study of acute/early heterosexual HIV-1 transmission organized by the Center for the AIDS Program of Research in South Africa (CAPRISA). Twelve strains arose from male-female transmission and six strains arose from female-male transmission. All molecularly cloned Env-pseudotyped viruses used CCR5 but not CXCR4 for cell entry.

Sequence analysis. Phylogenetic analysis of full-length gp160 nucleotide sequences confirmed that all 18 functional clones grouped within subtype C (Fig. 1). The clones comprised a wide spectrum of genetic diversity with only two clones (Du151.2 and Du422.1) that clustered with a high bootstrap value.

Deduced amino acid sequences showed that all 18 clones

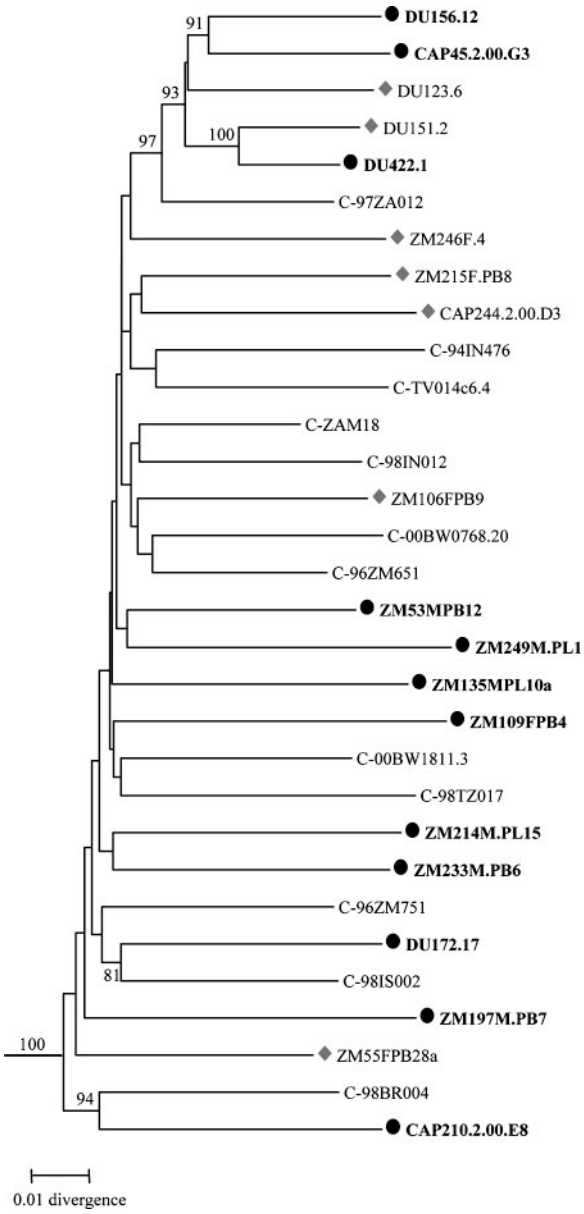


FIG. 1. Phylogenetic relationships of subtype C reference *env* clones. The newly characterized sequences are indicated by solid circles and shaded diamonds, and the 12 *env* clones that are recommended as standard reference reagents are represented by the solid circles and are bolded. Horizontal branch lengths are drawn to scale (the scale bar represents 0.01 nucleotide substitutions per site), but vertical separation is for clarity only. Values at nodes indicate the percentage of bootstraps in which the cluster to the right was found; only values of 80% or greater are shown. The phylogenetic tree was rooted with subtype D *env* sequences (NDK, Z2Z3, and 94UG114).

had an uninterrupted *env* open reading frame and 100% conservation of the cysteine residues that form the V1, V2, V3, and V4 loops of gp120 (Fig. 2). Considerable amino acid sequence variability was seen in V1, V2, V4, V5, and a region of approximately 34 amino acids immediately downstream of V3. V1, V2, V4, and V5 also exhibited substantial length variation (Table 2). V3 did not vary in length (35 amino acids in all clones) and exhibited only minor sequence variability that included position 320 (numbered as in Fig. 2) flanking the tip of the loop; this same position also shows greatest variability in the V3 loop of subtype B gp160 reference clones (45). Compared to a subtype C consensus sequence, amino acids in the V3 loop of these clones exhibited an average of 12% variability, whereas the 34 amino acids immediately downstream varied by an average of 39%.

Moderate variation was seen in the number and position of PNLG, where V1, V2, and V4 of gp120 were the most heavily glycosylated regions (Fig. 2; Table 3). All clones contained at least one PNLG in the relatively short stretch of amino acids comprising V5. Six PNLG in gp120 and three in gp41 were 100% conserved in all 18 clones (Fig. 2). Another three PNLG were conserved in all but one clone (Fig. 2). One highly conserved PNLG in the N terminus of V3 has been shown to mask neutralization epitopes in the V3 loop of subtype B viruses (2, 71).

Neutralization phenotype. Neutralization phenotypes were characterized with six subtype C plasma samples, four subtype-specific plasma pools, sCD4, and the broadly neutralizing monoclonal Abs IgG1b12, 2G12, 2F5, and 4E10 (Table 4). They were also characterized with a series of monoclonal Abs against CD4i epitopes on gp120 (Table 5). Each subtype C Env-pseudotyped virus was broadly sensitive to neutralization by individual subtype C plasma samples and by subtype-specific plasma pools. Some subtype C viruses were clearly more sensitive than others, but the level of sensitivity in all cases was at least 10-fold lower than that of MN and SF162.LS (Fig. 3A). This diminished sensitivity relative to MN and SF162.LS distinguishes the subtype C viruses from easily neutralized "tier 1" viruses (47). Another interesting property of the subtype C Env-pseudotyped viruses was their greater sensitivity to neutralization by a subtype C plasma pool compared to plasma pools of subtypes A, B, and D ($P = 0.0002$). The overall potencies of these plasma pools were ranked in the following order: subtype C > subtype A > subtype D > subtype B pool. Significant differences also were seen for the subtype A pool compared to the subtype B ($P = 0.0078$) and subtype D ($P = 0.0443$) pools but not when the subtype B pool was compared to the subtype D pool ($P = 0.107$).

All subtype C Env-pseudotyped viruses were sensitive to inhibition by sCD4. ID_{50} doses ranged from 0.2 μg/ml to 26 μg/ml, suggesting a broad spectrum of epitope exposure in and around the CD4 binding site of gp120. IgG1b12, which targets a complex epitope on gp120 that affects CD4 binding (13, 58), neutralized 10 of the 17 subtype C Env clones. No clear association was seen between the neutralizing activity of this monoclonal Ab and sensitivity to inhibition by sCD4.

All 18 gp160 clones were highly resistant to neutralization by the glycan-specific monoclonal Ab 2G12. Similar broad resistance to 2G12 has been reported for other subtype C viruses (7, 11, 33) and is associated with an absence of PNLG at critical positions that affect the 2G12 epitope (7, 14, 33, 69, 70). Each of our clones lacked a PNLG at one or more of these critical positions; most often this was position 295 at the N-terminal base of V3. Loss of a PNLG at position 295 has been associated with the general 2G12-resistant phenotype of subtype C viruses from chronically infected individuals (7) and pediatric infections (33). Our results indicate that the same is

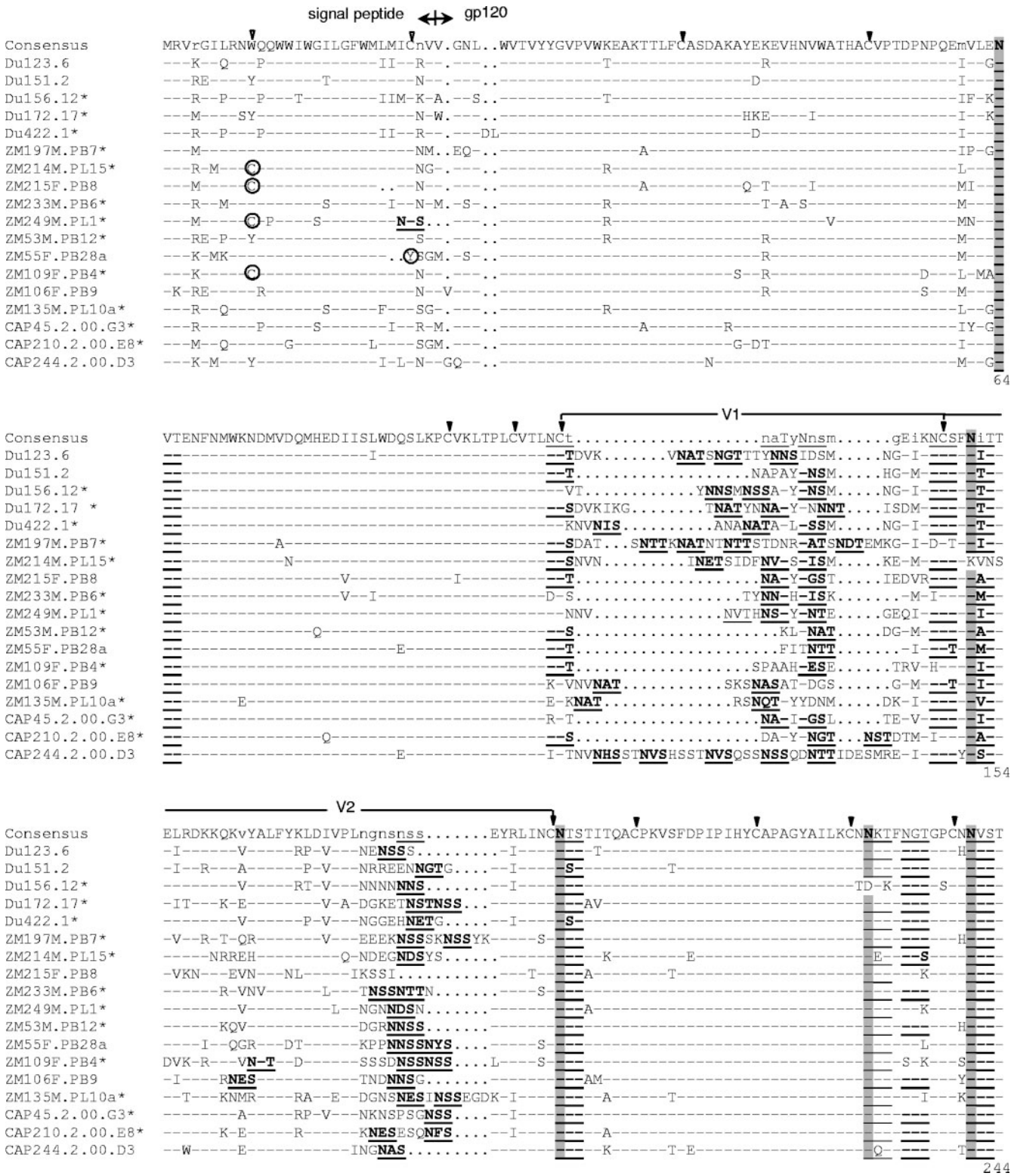


FIG. 2. Alignment of deduced amino acid sequences from acute/early subtype C HIV-1 *env* genes. Nucleotide sequences of newly derived *env* genes were translated, aligned, and compared with a consensus sequence generated by MASE. Numbering of amino acid residues begins with the first residue of gp120 and does not include the signal peptide. Dashes denote sequence identity, while dots represent gaps introduced to optimize alignments. Small letters in the consensus sequence indicate sites at which fewer than 50% of the viruses share the same amino acid residue. Triangles above the consensus sequence denote cysteine residues (solid triangles indicate sequence identity, while open triangles indicate sequence variation). V1, V2, V3, V4, and V5 regions designate hypervariable HIV-1 gp120 domains as previously described. The signal peptide and Env precursor cleavage sites are indicated; msd denotes the membrane-spanning domain in gp41; asterisks mark in-frame stop codons. Open circles highlight altered cysteine residues. Potential N-linked glycosylation sites (NXYX motif, where X is any amino acid other than proline and Y is either serine or threonine) are bolded and underlined. Highly conserved sites of potential N-linked glycosylation are shaded. The solid diamond denotes a highly variable amino acid position in the V3 loop.

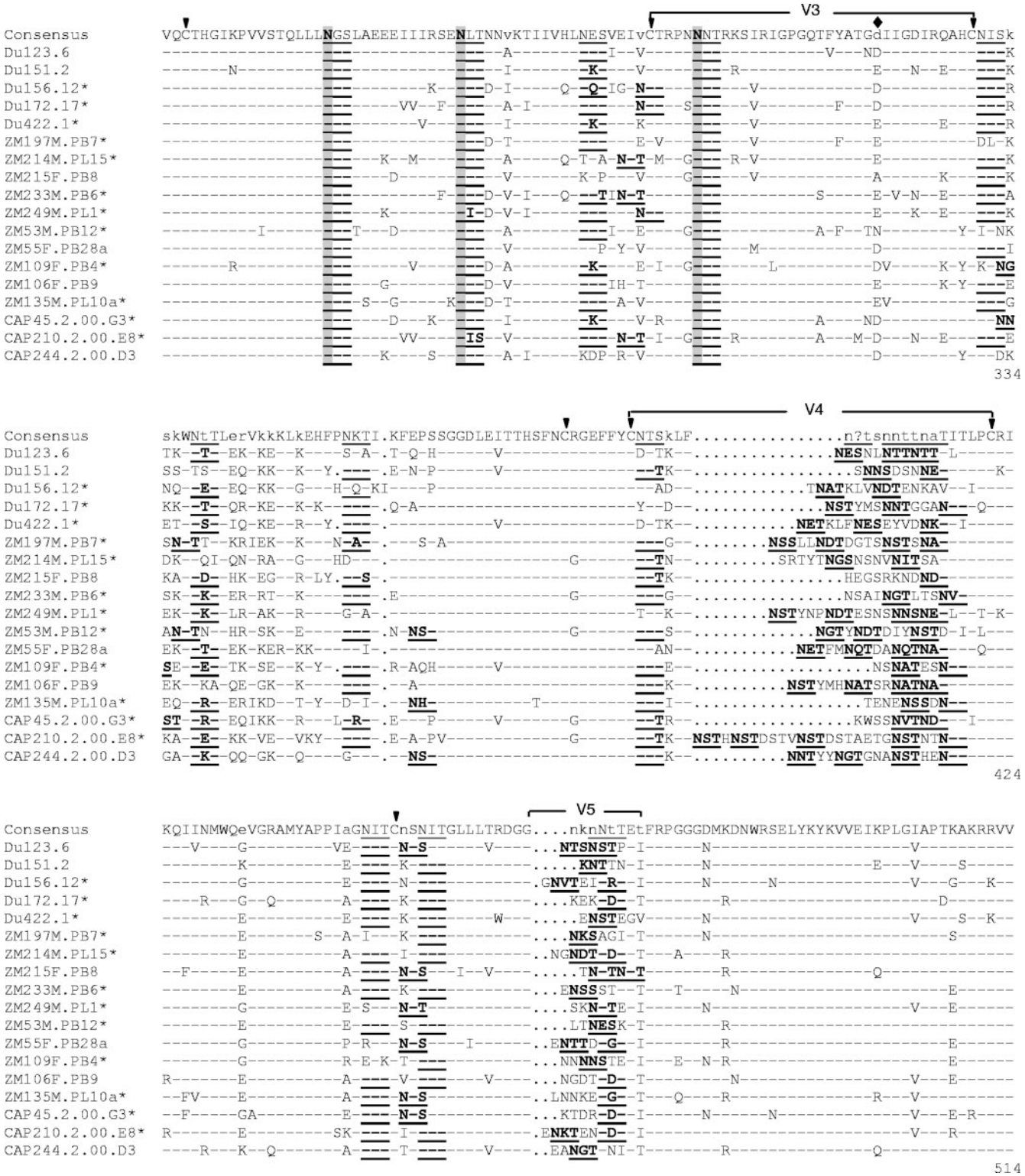


FIG. 2—*Continued*.

true for newly transmitted, heterosexually acquired subtype C viruses.

Monoclonal Abs 2F5 and 4E10, which recognize adjacent epitopes in the membrane-proximal external region of gp41 (3, 56, 57, 61, 74, 95), were quite different from one another in their ability to neutralize these subtype C viruses: 2F5 neutralized only 2 viruses, whereas 4E10 neutralized all 18 viruses. Neutralization by these two monoclonal Abs was highly predicted by amino acid sequence. Thus, both 2F5-sensitive viruses contained a DKW motif that has been reported to be a

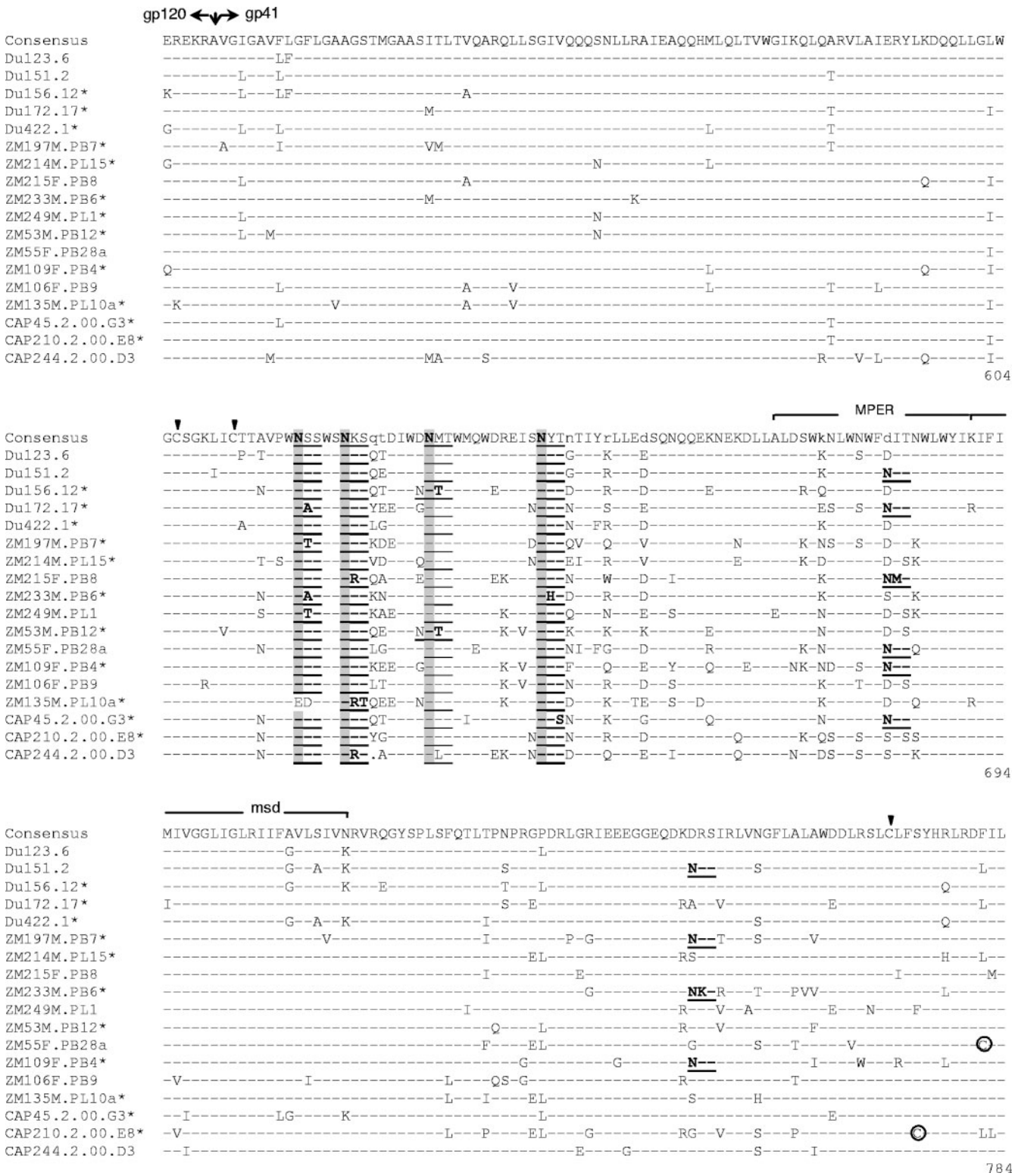


FIG. 2—*Continued*.

minimum requirement for 2F5 recognition (7, 45). Two additional clones (ZM214M.PL15 and CAP210.2.00.E8) contained this motif but had changes elsewhere in the 2F5 epitope. All clones except ZM215F.PB8 contained the WFXI motif that has been reported to be important for 4E10 recognition (7, 45). Clone ZM215F.PB8 contained a different motif (WFNM) yet it was highly susceptible to neutralization by 4E10, suggesting additional flexibility in the 4E10 core epitope. Six clones con-

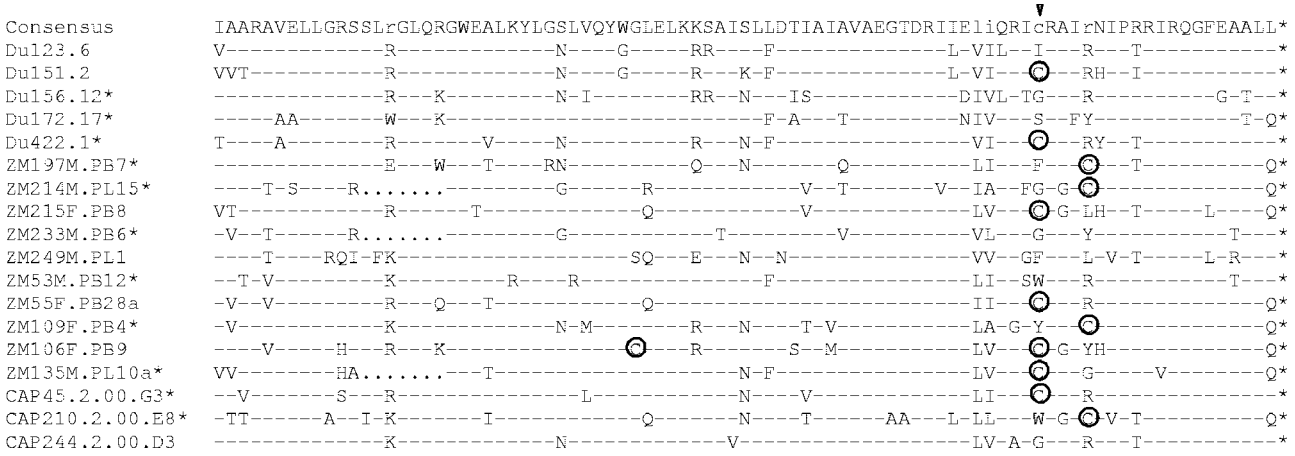


FIG. 2—*Continued*.

tained a potential PNLG in this core epitope that might be expected to mask the virus from 4E10, but biochemical evidence suggests this site is not glycosylated (41).

Consistent with the general resistance of these viruses to 2G12 and 2F5, their ability to be neutralized by TriMab (mixture of IgG1b12, 2G12, and 2F5) tracked with their sensitivity to IgG1b12. We saw no clear evidence of synergism with this triple combination of monoclonal Abs.

Neutralization phenotypes were further assessed by using CD4i monoclonal Abs to probe epitopes in the coreceptor binding domain of gp120. These are some of the most highly conserved epitopes on the gp120 molecule (22), but they are often concealed from Abs unless sCD4 is present to induce their exposure prior to virus-cell engagement (22, 68, 75). The CD4-independent NL-ADArs virus on which CD4i epitopes are spontaneously exposed (38) was highly sensitive to all of these monoclonal Abs, confirming that they possessed potential neutralizing activity (Table 5). Most of these CD4i monoclonal Abs also neutralized SF162.LS and, to a lesser degree, MN, although much higher concentrations were usually needed in these cases compared to NL-ADArs. In general, the subtype C Env-pseudotyped viruses were resistant to these CD4i monoclonal Abs regardless of whether subinhibitory doses of sCD4 were present (Table 5). An exception was E51, which neutralized 5 of 12 subtype C viruses but only at high concentrations. These latter results agree with a previous study in which E51 was superior to other CD4i antibodies (91). We observed two cases where sCD4 augmented the neutralizing activity of a CD4i monoclonal Ab: one case produced at least a fivefold increase in sensitivity (neutralization of Du172.17 by E51), whereas the other case was much less dramatic (neutralization of ZM135M.PL10a by 412d). Notably, this latter virus was the only new strain that was neutralized by four CD4i

TABLE 2. Lengths of gp120 variable regions

Env clone[a]	No. of amino acid residues				
	V1	V2	V3	V4	V5
Du123.6	30	40	35	24	10
Du151.2	18	45	35	22	8
Du156.12*	25	42	35	27	12
Du172.17*	30	46	35	25	9
Du422.1*	27	44	35	28	8
ZM197M.PB7*	39	49	35	31	9
ZM214M.PL15*	28	43	35	30	11
ZM215F.PB8	17	38	35	23	8
ZM233M.PB6*	17	42	35	23	10
ZM249M.PL1*	23	42	35	31	9
ZM53M.PB12*	15	42	35	26	9
ZM55F.PB28a	13	45	35	28	11
ZM109F.PB4*	17	45	35	20	10
ZM106F.PB9	25	42	35	29	10
ZM135M.PL10a*	23	49	35	21	11
CAP45.2.00.G3*	17	45	35	22	10
CAP210.2.00.E8*	19	45	35	39	12
CAP244.2.00.D3	42	40	35	29	11
Mean	23.6	43.6	35	26.6	9.9

[a] Clones selected as standard reference strains are marked with an asterisk.

TABLE 3. Potential N-linked glycosylation sites on gp120 and gp41

Env clone[a]	No. of N-linked glycosylation sites		
	gp120	gp41	gp41 ectodomain
Du123.6	25	4	4
Du151.2	22	6	5
Du156.12*	24	5	5
Du172.17*	27	5	5
Du422.1*	24	4	4
ZM197M.PB7*	27	5	4
ZM214M.PL15*	23	4	4
ZM215F.PB8	21	5	5
ZM233M.PB6*	24	5	4
ZM249M.PL1*	24	4	4
ZM53M.PB12*	25	5	5
ZM55F.PB28a	23	5	5
ZM109F.PB4*	23	6	5
ZM106F.PB9	25	4	4
ZM135M.PL10a*	23	3	3
CAP45.2.00.G3*	24	5	5
CAP210.2.00.E8*	29	4	4
CAP244.2.00.D3	26	4	4
Mean	24.4	4.6	4.4

[a] Clones selected as standard reference strains are marked with an asterisk.

TABLE 4. Neutralization phenotypes as determined with serum samples from HIV-1-infected individuals, sCD4, and monoclonal Abs

Virus[a]	ID_{50} in TZM-bl cells based on[b]:																
	Reciprocal serum dilution											Concn (μg/ml)					
	BB8	BB12	BB28	BB55	BB70	BB106	Pool A	Pool B	Pool C	Pool D	Normal Pool	sCD4	IgG1b12	2G12	2F5	4E10	TriMab
MN	5,107	4,573	537	2,964	43,740	203	9,013	43,740	5,550	16,952	<20						
SF162.LS	15,548	5,480	2,629	20,302	43,740	552	5,597	43,264	8,066	2,438	<20						
Du123.6	207	352	165	84	147	198	306	182	18,845	220	90	0.3	0.2	>50	>50	0.1	1.6
Du151.2	196	1,555	2,529	818	241	487	347	518	527	123	20	3.0	1.4	>50	>50	0.8	13.5
Du156.12*	369	426	238	336	406	258	810	231	2,368	228	52	13.4	0.8	>50	>50	0.2	1.2
Du172.17*	429	884	499	196	549	550	462	315	629	562	91	1.7	1.0	>50	>50	0.3	2.6
Du422.1*	134	354	193	91	51	63	335	114	2,138	65	90	9.1	0.2	>50	>50	0.7	0.8
ZM197M.PB7*	79	103	76	117	348	68	163	185	332	176	<20	3.9	19.9	>50	12.3	0.5	24.8
ZM214M.PL15*	259	366	104	90	206	96	127	<20	162	64	<20	8.0	3.0	>50	>50	4.0	9.7
ZM215F.PB8	231	694	122	384	90	<20	165	51	706	108	32	14.0	>50	>50	>50	0.4	>25
ZM233M.PB6*	295	153	214	413	2,899	80	127	108	640	218	<20	2.9	>50	>50	>50	1.2	>25
ZM249M.PL1*	160	105	287	158	112	117	193	63	383	156	33	9.7	3.2	>50	>50	2.1	7.9
ZM53M.PB12*	76	73	411	86	<20	98	57	91	566	65	<20	8.3	25.9	>50	>50	7.0	>25
ZM55F.PB28a	96	206	106	169	47	110	99	56	250	59	27	24.0	>50	>50	33.6	8.0	>25
ZM109F.PB4*	160	188	89	160	238	61	185	82	365	119	<20	0.2	>50	>50	>50	0.6	>25
ZM106F.PB9	249	378	287	262	68	71	101	76	433	<20	<20	23.0	>50	>50	>50	7.2	>25
ZM135M.PL10a*	138	179	93	190	50	177	108	57	202	109	29	6.1	>50	>50	>50	0.6	>25
CAP45.2.00.G3*	61	437	351	44	170	52	142	47	453	60	<20	26.0	0.7	>50	>50	2.6	1.5
CAP210.2.00.E8*	111	75	82	113	272	75	101	29	293	51	<20	3.4	20.4	>50	>50	1.2	20.8
CAP244.2.00.D3	53	47	32	51	91	30	38	27	94	40	<20	8.9	>50	>50	>50	1.9	>25
GMT[c]	151	227	177	151	142	97	158	79	514	98		5.3				1.1	

[a] Clones selected as standard reference strains are marked with an asterisk.
[b] Values are the dilution or concentration at which RLU were reduced 50% compared to virus control wells. BB8, BB12, BB28, BB70, and BB106 are plasma samples from individuals infected with subtype C HIV-1.
[c] Geometric mean titer against the 18 pseudoviruses containing cloned primary isolate Env (excludes MN and SF162.LS). Serum/plasma titers of <20 were assigned a value of 10 for calculations. Because multiple ID_{50} values were >50, the GMT was not determined for IgG1b12, 2G12, 2F5, or TriMab.

monoclonal Abs, including two that did not neutralize the other subtype C viruses. Although the potency of neutralization against ZM135M.PL10a was relatively weak, there appears to be greater exposure of CD4i epitopes on this virus compared to other subtype C viruses. Overall, the results suggest that epitopes in the coreceptor binding domain on newly transmitted subtype C viruses are mostly concealed from Abs.

Noting that our gp160 genes were cloned from either cocultured PBMC virus, uncultured PBMC, or directly from plasma, we performed an analysis to determine whether the source of gp160 influenced the general neutralization sensitivity of the corresponding Env-pseudotyped viruses. For this analysis, combined neutralization data from the subtype C plasma samples (BB samples) and subtype-specific plasma pools in Table 4 were compared between the three categories of clones. The results showed a trend in neutralization sensitivity such that cocultured PBMC clones (GMT, 344) > uncultured PBMC clones (GMT, 129) > plasma clones (GMT, 101), but this trend was not significant ($P > 0.05$). Also, greater resistance to IgG1b12 was seen in the uncultured PBMC group of clones,

TABLE 5. Neutralization with CD4i monoclonal Abs in the presence and absence of sCD4

Virus[a]	ID_{50} in TZM-bl cells[b]								
	17b	23e	31H	21c	E51	48d	112d	412d	ED10
NL-ADArs	0.002	0.002	0.002	0.002	<0.0001	0.2	6.1	0.002	2.9
MN	3.5	4.4	—	14.4	0.3	0.4	—	16.5	—
SF162.LS	0.9/2.3	1.3/3.7	1.0/2.9	1.1/3.5	0.1/0.6	4.6/12.6	—/—	0.2/1.4	3.0/5.9
Du156.12*	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—
Du172.17*	—/—	—/—	—/—	—/—	—/4.7	—/—	—/—	—/—	—/—
Du422.1*	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—
ZM197M.PB7*	—/—	—/—	—/—	—/—	19.5/21.6	—/—	—/—	16.7/—	—/—
ZM214M.PL15*	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—
ZM233M.PB6*	—/—	—/—	—/—	—/—	6.5/10.2	—/—	—/—	—/—	—/—
ZM249M.PL1*	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—
ZM53M.PB12*	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—
ZM109F.PB4*	—/—	—/—	—/—	—/—	9.9/16.8	—/—	—/—	19.3/19.1	—/—
ZM135M.PL10a*	17.0/—	5.9/—	—/—	—/—	17.6/—	—/—	—/—	—/20.7	—/—
CAP45.2.00.G3*	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—
CAP210.2.00.E8*	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—	—/—

[a] Clones selected as standard reference strains are marked with an asterisk. NL-ADArs and MN were uncloned viruses.
[b] Values are the monoclonal Ab concentration at which RLU were reduced 50% compared to virus control wells. Neutralization was measured in the absence/presence of sCD4. sCD4 was present in all wells (including cell control and virus control wells) at a concentration equal to the ID_{50} for each virus. Dashes signify no neutralization at the highest concentration of monoclonal Ab tested (25 μg/ml). NL-ADArs and MN were assayed in the absence of sCD4 only.

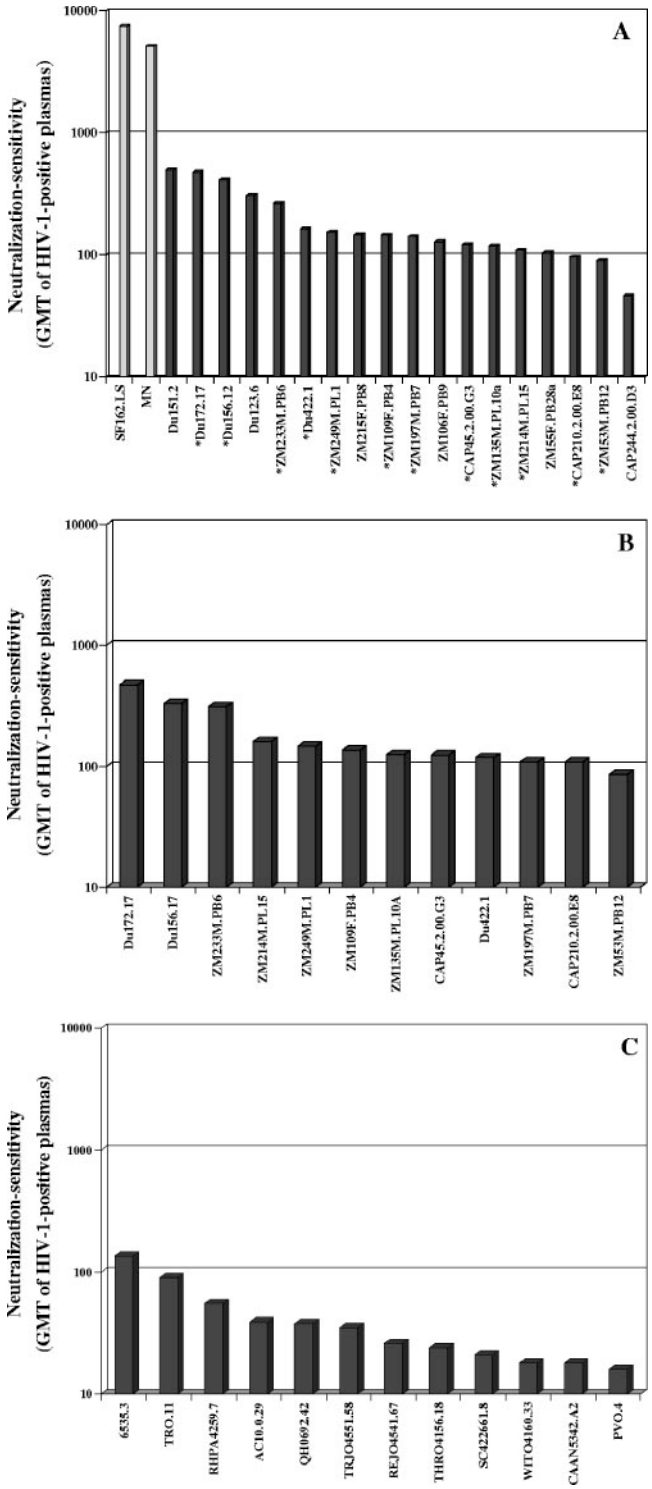


FIG. 3. Neutralization sensitivity of acute/early subtype C and subtype B HIV-1 Env-pseudotyped viruses as determined with plasma samples from HIV-1-infected individuals. Bar height represents the GMT of neutralizing Abs against the indicated Env-pseudotyped viruses. A. GMT for all HIV-1-positive plasma samples shown in Table 4 (BB8, BB12, BB28, BB55, BB70, and BB106 and pool A, pool B, pool C, and pool D). Clones selected as standard subtype C reference strains are marked with an asterisk. B. GMT of neutralizing Abs for the individual subtype C plasma samples BB8, BB12, BB28, BB55, BB70, and BB106 assayed against the 12 subtype C reference strains. C. GMT of neutralizing Abs in the individual subtype C plasma samples BB8, BB12, BB28, BB55, BB70, and BB106 assayed against 12 subtype B reference strains.

especially when compared to the cocultured PBMC clones. We caution that a larger number of clones will be needed to confirm this latter observation.

Selection of standard reference strains. The demographic, biologic, genetic, and neutralization properties of many of the Env-pseudotyped viruses described in this report appear suitable for standardized assessments of vaccine-elicited NAb responses. For adequate statistical power, it has been recommended that each standard panel be comprised of 12 reference strains (47). Among the 18 candidate reference strains described here, 12 were selected to comprise a panel that represents the greatest genetic and antigenic diversity while avoiding strains that are unusually sensitive or resistant to neutralization. None of the 18 strains would be considered unusually sensitive to neutralization; however, 1 strain was omitted for being relatively insensitive to neutralization by HIV-1-positive plasma samples (CAP244.2.00.D3). Another strain (ZM215F.PB8) was omitted because careful analysis showed that all bulk-derived *env* genes from this patient (who had a heterogenous early infection) were in vitro recombinants. Also, the Du123.6 strain was excluded because it was easily neutralized by pool C plasma, with titers higher than those seen for MN and SF162.LS viruses. On the other hand, two strains (Du422.1 and Du151.2) were phylogenetically closer to each other than any other pair of strains, clustering together with a high bootstrap value (Fig. 1); to maximize the genetic diversity among members in the panel, only Du422.1 was selected while Du151.2 was excluded. Of the 12 selected strains, 7 originated in Zambia and 5 in South Africa (Table 1). Also, six arose by male-female transmission, while the remaining six arose by female-male transmission (Table 1).

Genetic comparisons between subtype B and C reference strains. The variable loops and number and position of PNLG in gp120 have been shown to influence the neutralization phenotype of HIV-1 (8, 26, 48, 56, 62, 84). Recent results also suggest that acute/early subtype B and C strains differ from one another in their neutralization properties (23, 29). For these reasons, amino acid sequences of the 12 subtype C gp160 reference genes were compared to the sequences of 12 subtype B gp160 reference genes described previously (45). Comparisons also were made between subtype B and C sequences in the LANL database. These comparisons focused on the number of amino acids that comprise gp120 and its variable regions. They also focused on the number and position of PNLG in gp120.

LANL database sequence comparisons showed that subtype C gp120 was generally shorter and contained fewer PNLG compared to subtype B (Table 6). Differences in gp120 length were evident by shorter V1, V3, and V4 regions despite a slightly longer V2 region in subtype C. Fewer PNLG were present in V1, V3, V4, and V5 of subtype C gp120, whereas the number of PNLG in V2 remained constant in both subtypes. Many PNLG sites were common in both subtypes, but four were prevalent in only one subtype. Subtype-specific PNLG

TABLE 6. Comparison of sequence lengths and PNLG between HIV-1 subtypes B and C gp120

Genomic region	Analysis group[a]	Sequence length interquartile	Glycan no. interquartile	Comparison	*P* value	
					Sequence length comparison[b]	Glycan no. comparison[b]
gp120	B-ref	509–519	26–28	B-ref/C-ref	0.01	0.01
	C-ref	492–509	24–26	B-db/C-db	**8.44×10^{-15}**	0.004
	B-db	507–515	24–27	B-ref/B-db	0.44	0.02
	C-db	499–510	23–26	C-ref/C-db	0.45	0.86
V1	B-ref	25–31	4–5	B-ref/C-ref	0.03	0.06
	C-ref	17–27	3–4	B-db/C-db	**3.37×10^{-13}**	**0.0008**
	B-db	26–32	4–5	B-ref/B-db	0.75	0.43
	C-db	22–28	3–5	C-ref/C-db	0.31	0.45
V2	B-ref	40–46	4–5	B-ref/C-ref	0.52	0.85
	C-ref	42–45	4–5	B-db/C-db	**3.33×10^{-5}**	0.50
	B-db	40–44	4–5	B-ref/B-db	0.23	0.52
	C-db	41–45	4–5	C-ref/C-db	0.16	0.88
V3	B-ref	35–35	3–3	B-ref/C-ref	0.36	0.0036
	C-ref	35–35	1–2	B-db/C-db	0.02	**$<2.2 \times 10^{-16}$**
	B-db	35–35	2–3	B-ref/B-db	0.61	0.4259
	C-db	35–35	2–2	C-ref/C-db	0.60	0.7318
V4	B-ref	31–33	4–5	B-ref/C-ref	0.005	0.02
	C-ref	23–30	3–4	B-db/C-db	**1.74×10^{-12}**	0.02
	B-db	30–33	4–5	B-ref/B-db	0.50	0.75
	C-db	27–31	4–5	C-ref/C-db	0.06	0.01
V5	B-ref	9–10	2–2	B-ref/C-ref	0.57	0.09
	C-ref	9–11	1–2	B-db/C-db	0.13	0.004
	B-db	9–11	1–2	B-ref/B-db	0.26	0.12
	C-db	9–11	1–2	C-ref/C-db	0.83	0.71

[a] B-ref, 12 selected sequences from the subtype B panel of reference strains; B-db, 176 well-characterized subtype B database sequences; C-ref, 12 selected sequences from the C panel of reference strains; C-db, 122 well-characterized subtype C database sequences.

[b] Bolded *P* values ($P < 0.002$) are significant differences after Bonferroni correction. Underlined *P* values ($0.002 < P < 0.05$) are significantly different trends after Bonferroni correction.

included positions 230 (between V2 and V3) and 442 (between V4 and V5) that were diminished in subtype B (HXB2 numbering). They also included positions 295 (proximal to N terminus of V3) and 362 (between V3 and V4) that were diminished in subtype C (HXB2 numbering).

No significant differences were seen in the lengths of V2 and V3 and the number of PNLG in V1 and V5 when sequences of the subtype B and C reference clones were compared. Thus, overall the subtype C panel of reference clones contained shorter gp120 regions (V1 and V4) and fewer PNLG in gp120 (V1, V3, and V4) than the subtype B panel of reference clones. Differences observed only in database (mostly chronic virus) sequences might be an indication that gp120 of certain subtypes evolved over time in infected individuals to give rise to viral variants that are rarely transmitted. In this regard, our intrasubtype comparisons suggest that transmission of subtype C favors variants with fewer PNLG in V4 compared to chronic subtype C variants ($P = 0.01$). The comparisons also suggest that transmission of subtype B favors variants with a greater number of PNLG in gp120 compared to chronic subtype B variants ($P = 0.02$).

Neutralization phenotype comparisons between subtype B and C reference strains. Results obtained with the subtype C plasma samples from South Africa, subtype-specific plasma pools, sCD4, and monoclonal Abs were used to compare the neutralization phenotypes of subtype B and C reference strains. Overall, the subtype C reference strains were more sensitive to neutralization by subtype C plasma samples ($P = 0.0005$) (Fig. 3). Additional comparisons were made with published data on a panel of 12 subtype B reference strains (45). In general, the subtype C panel was less sensitive to neutralization by a subtype B plasma pool ($P = 0.009$), but no significant difference was found between the two panels of reference strains when compared in terms of their overall sensitivity to CD4, IgG1b12, 4E10, and the subtype A, C, and D plasma pools. The most dramatic difference between the two panels of reference strains was the uniform resistance of subtype C to neutralization by 2G12 and their infrequent neutralization by 2F5.

DISCUSSION

We describe here the first extensive characterization of the neutralization properties of newly transmitted subtype C HIV-1 variants from heterosexually acquired infections. The viruses were characterized as molecularly cloned Env-pseudotyped viruses by using a wide range of antibody specificities and in the context of genetic features that are known to influence epitope exposure on the complex gp120 molecule.

Comparisons were made to subtype B, the most extensively studied HIV-1 subtype.

The R5 biologic phenotype (6) of all 18 subtype C gp160 clones in this report is typical of newly transmitted subtype B viruses (21, 93, 94) and appears to be common for subtype C regardless of the stage of infection (16, 59, 85). A general distinguishing feature was the shorter and less glycosylated gp120 on these subtype C viruses compared to newly transmitted subtype B viruses. A similar observation was made previously for both subtypes C and A where, unlike subtype B (29), the gp120 of subtypes C and A might expand and add PNLG over the course of infection (18, 23, 44). Li et al. recently reported that early autologous NAb responses in subtype C-infected individuals might be more potent than the early responses in subtype B-infected individuals (44). These authors suggested that increased exposure of epitopes on the shorter, less-glycosylated gp120 of newly transmitted subtype C viruses could have resulted in enhanced immunogenicity, greater antigenicity, or both. This interpretation is based on evidence that the variable loops and number and position of PNLG on gp120 are used by the virus to mask epitopes as a very effective means to evade NAbs (15, 26, 62, 84, 88). It has been suggested that the shorter, less-glycosylated gp120 of newly transmitted subtype C viruses might enhance the accessibility of certain epitopes to vaccine-elicited NAbs (23). The subtype C gp160 clones described here should be useful in future studies that address this possibility in greater detail.

It has been reported that the gp120 region of envelopes from newly transmitted subtype C viruses is shorter and less glycosylated compared to subtype C variants from chronically infected individuals (23) and that these differences are not seen in subtype B (18, 29). We performed a similar analysis by comparing the newly derived subtype B and C reference *env* sequences to subtype-matched gp120 controls in the LANL database, which is known to mostly contain sequences from chronic viruses. Results of this comparison showed a significant trend toward fewer PNLG in the V4 of newly transmitted subtype C viruses, but we found no evidence that these viruses contained shorter gp120 regions than chronic subtype C viruses. We also found a significant trend toward a greater number of PNLG on the gp120 of newly transmitted subtype B viruses compared to chronic subtype B viruses. We caution that our comparisons did not have strong statistical support because of the small sample size. Thus, a larger number of gp120 sequences from acutely and chronically infected individuals will be needed to resolve whether or not newly transmitted viruses have unique genetic features.

A general feature of the newly transmitted subtype C viruses was their relatively conserved V3 loop that was followed immediately downstream by a region of approximately 34 amino acids that was highly variable. A similar observation has been made for subtype C viruses from chronically infected individuals (24, 59) and for the mostly chronic subtype C viruses in the Los Alamos sequence database (32). Moreover, Milich et al. reported that the V3 loop of R5 but not X4 subtype B viruses also exhibits low sequence variability (51). In this regard, the V3 loop of our subtype C viruses was no more variable than the V3 loop of newly transmitted R5 subtype B reference clones (45). Notably, the region immediately downstream of V3 in the subtype B reference clones was highly variable (45). Thus, low amino acid sequence variability in V3 and high sequence variability immediately downstream of V3 appear to be features that are shared between newly transmitted and chronic R5 viruses of both subtypes.

The V3 loop is of considerable interest because it plays a central role in receptor and coreceptor interactions that determine viral tropism and entry, making this region an attractive target for NAb-based vaccines. Unfortunately, the cryptic nature of V3 (8) makes it extremely difficult for Abs to gain access to this region. Thus, HIV-1 has evolved to escape V3-specific NAbs by masking epitopes through strategic placement of PNLG (2, 71) and by the conformation of its V2 loop (15). We observed fewer PNLG on the V3 loop of newly transmitted subtype C viruses compared to subtype B, making it possible that enhanced epitope exposure on subtype C viruses could result in a stronger V3-directed NAb response compared to subtype B. An exposed V3 loop can in fact be a major target for NAbs (36, 80). Although we did not examine V3-specific neutralization here, we found no evidence in a previous study that newly transmitted subtype C viruses, including some of the same viruses used here, were unusually sensitive to V3-specific Abs (11). All of our subtype C viruses retain a highly conserved PNLG at a key site (position 301, HXB2 numbering) that is known to mask V3 epitopes on subtype B viruses (2, 71). They also possessed a V2 loop that on average was the same size as newly transmitted subtype B reference clones. These structural features might contribute to the effective masking of V3 epitopes on newly transmitted subtype C viruses.

Another interesting and potentially important neutralization target on the gp120 molecule is the coreceptor binding domain that lies in a region within and around the bridging sheet and might also involve a portion of the V3 loop (20, 63, 64, 72, 77, 78, 87). This region contains some of the most highly conserved neutralization epitopes on the virus; however, exposure of these epitopes in most cases requires conformational changes in gp120 that occur upon CD4 ligation (22, 68, 75). Viruses that possess a spontaneously exposed coreceptor binding domain have been observed in vivo (22, 92), making it possible that conditions exist where this property affords a fitness advantage. One example would be to provide a selective advantage for transmission and early virus replication before CD4i antibodies are made. We tested this hypothesis by using a series of monoclonal Abs to probe CD4i epitopes on our newly transmitted subtype C viruses and found very few cases of positive neutralization regardless of whether subinhibitory doses of sCD4 were present. This outcome is discordant with a previous study in which CD4i epitopes were shown to be potent targets for neutralization on a subset of HIV-2 viruses (22). Based on our results, CD4i epitopes on newly transmitted subtype C viruses appear to be effectively masked and difficult to induce and stabilize with sCD4. We acknowledge that, because our viruses were obtained several weeks after infection, it remains possible that initial transmission involves viruses on which CD4i epitopes are spontaneously exposed.

We found it interesting that SF162.LS and MN, two strains of HIV-1 that are highly sensitive to neutralization by HIV-1-positive sera, were both sensitive to neutralization by CD4i monoclonal Abs. T-cell-line-adapted strains, including MN, are thought to be highly sensitive to HIV-1-positive sera because the V3 loop on these viruses is exposed (36, 80). Our

results suggest that exposure of CD4i epitopes is another reason why certain strains of HIV-1 are highly sensitive to HIV-1-positive sera. Indeed, serum samples from most HIV-1-infected individuals have been shown to contain high titers of CD4i antibodies (22).

A major goal of this study was to create a well-characterized panel of subtype C gp160 reference clones to facilitate standardized assessments of vaccine-elicited NAb responses. A similar panel of gp160 reference clones for subtype B was described recently (47). We considered it important that the envelope glycoproteins of reference strains exhibit general antigenic features that will not overestimate the potential value of Abs against cryptic epitopes, such as those in the V3 loop and coreceptor binding domain, that tend to be poor targets on primary isolates (45). The fact that our subtype C Env-pseudotyped viruses were a great deal less sensitive to neutralization by HIV-1-positive plasma samples than MN and SF162.LS suggests they possess appropriate antigenic properties. These antigenic properties were characterized in greater detail by using a set of broadly neutralizing monoclonal Abs whose epitopes are of major interest for vaccine development. As expected, the 2G12 epitope was completely absent and the 2F5 epitope was rarely detected on these subtype C Env-pseudotyped viruses. Similar results have been reported previously for newly transmitted subtype C viruses (11) and for subtype C viruses from chronically infected individuals (7) and pediatric infections (33). Also in agreement with these previous reports, all of our subtype C Env-pseudotyped viruses contained the 4E10 epitope and many contained the IgG1b12 epitope.

The overall neutralization phenotype and inferred structural properties of these HIV-1 subtype C Env clones appear to be suitably representative of subtype C to support their use as reference reagents. For uniformity, we have designated a subset of 12 of these gp160 clones to recommend as standards for assessing the NAb responses generated by the current pipeline of candidate vaccines. This panel complements an existing panel of subtype B gp160 reference clones (45) and a multisubtype panel of viruses from chronically infected individuals (9), all of which may need to be modified in the future as new information emerges on how to improve their correlative value (47). Additional panels of newly transmitted gp160 reference clones are needed for other major subtypes of HIV-1, including A, D, CRF01_AE, and CRF02_AG. The subtype preferences of HIV-1-positive plasma samples seen here, combined with genetic differences between subtypes that might affect neutralization, support the need for separate panels. We also encourage the development of multiple panels of gp160 clones within each major subtype to more completely represent intrasubtype diversity at the geographic level (11). Considering the prevalence and global distribution of subtype C, it will be especially important to develop additional panels of reference gp160 clones from other parts of the world where this subtype dominates the epidemic, including China, India, and other African countries. These additional panels should facilitate vaccine discovery by enhancing the standardized assessment of vaccine-elicited NAb responses on a global scale.

ACKNOWLEDGMENTS

We thank Gita Ramjee and the CAPRISA clinical and laboratory staff for sample collection. We also thank Florette Treunicht, Isaac Choge, and Penny Moore for generating functional *env* clones from the CAP isolates and Natasha Taylor-Meyer, Eleanor Cave, and Isaac Choge for characterizing the BB plasma samples. Finally, we thank Opendra Sharma and his staff for their assistance in making the reference clones available through the NIH ARRRP.

This work was supported by National Institutes of Health grants AI30034 and AI46705 (D.C.M.), AI055386 (F.G.), AI51231 and AI64060 (E.H.), and AI54497, AI85338, AI41530, and AI27767 (B.H.H.). CAPRISA is supported by NIH grant AI51794.

REFERENCES

1. **Allen, S., J. Meinzen-Derr, M. Kautzman, I. Zulu, S. Trask, U. Fideli, R. Musonda, F. Kasolo, F. Gao, and A. Haworth.** 2003. Sexual behavior of HIV discordant couples after HIV counseling and testing. AIDS **17:**733–740.
2. **Back, N. K. T., L. Smit, J.-J. de Jong, W. Keulen, M. Schutten, J. Goudsmit, and M. Tersmette.** 1994. An N-glycan within the human immunodeficiency virus type 1 gp120 V3 loop affects virus neutralization. Virology **199:**431–438.
3. **Barbato, G., E. Bianchi, P. Ingallinella, W. H. Hurn, M. D. Miller, G. Ciliberto, R. Cortese, R. Bazzo, J. W. Shiver, and A. Pessi.** 2003. Structural analysis of the epitope of the anti-HIV antibody 2F5 sheds light into its mechanism of neutralization and HIV fusion. J. Mol. Biol. **330:**1101–1115.
4. **Beddows, S., S. Lister, R. Cheingsong, C. Bruck, and J. Weber.** 1999. Comparison of the antibody repertoire generated in healthy volunteers following immunization with a monomeric recombinant gp120 construct derived from a CCR5/CXCR4-using human immunodeficiency virus type 1 isolate with sera from naturally infected individuals. J. Virol. **73:**1740–1745.
5. **Belshe, R. B., G. J. Gorse, M. J. Mulligan, T. G. Evans, M. C. Keefer, J.-L. Excler, A.-M. Duliege, J. Tartaglia, W. I. Cox, J. McNamara, K.-L. Hwang, A. Bradney, D. Montefiori, and K. J. Weinhold.** 1998. Induction of immune responses to HIV-1 canarypox virus (ALVAC) HIV-1 and gp120 SF-2 recombinant vaccines in uninfected volunteers. AIDS **12:**2407–2415.
6. **Berger, E. A., R. W. Doms, E.-M. Fenyö, B. T. M. Korber, D. R. Littman, J. P. Moore, Q. J. Sattentau, H. Schuitemaker, J. Sodrsoki, and R. A. Weiss.** 1998. HIV-1 phenotypes classified by co-receptor usage. Nature **391:**240.
7. **Binley, J., T. Wrin, B. Korber, M. Zwick, M. Wang, C. Chappey, G. Stiegler, R. Kunert, S. Zolla-Pazner, H. Katinger, C. Petropoulos, and D. Burton.** 2004. Comprehensive cross-subtype neutralization analysis of a panel of anti-human immunodeficiency virus type 1 monoclonal antibodies. J. Virol. **78:**13232–13252.
8. **Bou-Habib, D. C., G. Roderiquez, T. Oravecz, P. W. Berman, P. Lusso, and M. A. Norcross.** 1994. Cryptic nature of envelope V3 region epitopes protects primary monocytotropic human immunodeficiency virus type 1 from antibody neutralization. J. Virol. **68:**6006–6013.
9. **Brown, B. K., J. M. Darden, S. Tovanabutra, T. Oblander, J. Frost, E. Sanders-Buell, M. S. DeSouza, D. L. Birx, F. E. McCutchan, and V. R. Polonis.** 2005. Biologic and genetic characterization of a panel of 60 human immunodeficiency virus type 1 (HIV-1) isolates, representing clades A, B, C, D, CRF01_AE, and CRF02_AG, for the development and assessment of candidate vaccines. J. Virol. **79:**6089–6101.
10. **Bures, R., A. Gaitan, T. Zhu, C. Graziosi, K. M. McGrath, J. Tartaglia, P. Caudrelier, R. El Habib, M. Klein, A. Lazzarin, D. M. Stablein, M. Deers, L. Corey, M. L. Greenberg, D. H. Schwartz, and D. C. Montefiori.** 2000. Immunization with recombinant canarypox vectors expressing membrane-anchored gp120 followed by gp160 protein boosting fails to generate antibodies that neutralize R5 primary isolates of human immunodeficiency virus type 1. AIDS Res. Hum. Retrovir. **16:**2019–2035.
11. **Bures, R., L. Morris, C. Williamson, G. Ramjee, M. Deers, S. A. Fiscus, S. A. Karim, and D. C. Montefiori.** 2002. Regional clustering of shared neutralization determinants on primary isolates of subtype C human immunodeficiency virus type 1 from South Africa. J. Virol. **76:**2233–2244.
12. **Burton, D., R. Desrosiers, R. Doms, W. Koff, P. Kwong, J. Moore, G. Nabel, J. Sodroski, I. Wilson, and R. Wyatt.** 2004. HIV vaccine design and the neutralizing antibody problem. Nat. Immunol. **5:**233–236.
13. **Burton, D. R., J. Pyati, R. Koduri, S. J. Sharp, G. B. Thornton, P. W. Parren, L. S. Sawyer, R. M. Hendry, N. Dunlop, P. L. Nara, et al.** 1994. Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. Science **266:**1024–1027.
14. **Calarese, D. A., C. N. Scanlan, M. B. Zwick, S. Deechongkit, Y. Mimura, R. Kunert, P. Zhu, M. R. Wormald, R. L. Stanfield, K. H. Roux, J. W. Kelly, P. M. Rudd, R. A. Dwek, H. Katinger, D. R. Burton, and I. A. Wilson.** 2003. Antibody domain exchange is an immunologic solution to carbohydrate cluster recognition. Science **300:**2065–2071.
15. **Cao, J., N. Sullivan, E. Desjardin, C. Parolin, J. Robinson, R. Wyatt, and J. Sodroski.** 1997. Replication and neutralization of human immunodeficiency virus type 1 lacking the V1 and V2 variable loops of the gp120 envelope glycoprotein. J. Virol. **71:**9808–9812.
16. **Cecilia, D., S. S. Kulkarni, S. P. Tripathy, R. R. Gangakhedkar, R. S. Paranjape, and D. A. Gadkari.** 2000. Absence of coreceptor switch with disease progression in human immunodeficiency virus infections in India. Virology **271:**253–258.
17. **Cheng-Mayer, C., R. Liu, N. R. Landau, and L. Stamatatos.** 1997. Macro-

phage tropism of human immunodeficiency virus type 1 and utilization of the CC-CKR5 coreceptor. J. Virol. **71:**1657–1661.
18. **Chohan, B., D. Lang, M. Sagar, B. Korber, L. Lavreys, B. Richardson, and J. Overbaugh.** 2005. Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter V1-V2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral RNA levels. J. Virol. **79:**6528–6531.
19. **Cilliers, T., J. Nhlapo, M. Coetzer, D. Orlovic, T. Ketas, W. C. Olson, J. P. Moore, A. Trkola, and L. Morris.** 2003. The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C. J. Virol. **77:**4449–4456.
20. **Cocchi, F., A. L. DeVico, A. Garzino-Demo, A. Cara, R. C. Gallo, and P. Lusso.** 1996. The V3 domain of the HIV-1 gp120 envelope glycoprotein is critical for chemokine-mediated blockade of infection. Nat. Med. **2:**1244–1247.
21. **Conner, R. I., K. E. Sheridan, D. Ceradini, S. Choe, and N. R. Landau.** 1997. Change in coreceptor use correlates with disease progression in HIV-1-infected individuals. J. Exp. Med. **185:**621–628.
22. **Decker, J., F. Bibollet-Ruche, X. Wei, S. Wang, D. N. Levy, W. Wang, E. Delaporte, M. Peeters, C. A. Derdeyn, S. Allen, E. Hunter, M. S. Saag, J. A. Hoxie, B. H. Hahn, P. D. Kwong, J. E. Robinson, and G. M. Shaw.** 2005. Antigenic conservation and immunogenicity of the HIV coreceptor binding site. J. Exp. Med. **201:**1407–1419.
23. **Derdeyn, C. A., J. M. Decker, F. Bibollet-Ruche, J. L. Mokili, M. Muldoon, S. A. Denham, M. L. Heil, F. Kasolo, R. Musonda, B. H. Hahn, G. M. Shaw, B. T. Korber, S. Allen, and E. Hunter.** 2004. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. Science **303:**2019–2022.
24. **Engelbrecht, S., T. de Villiers, C. C. Sampson, J. Z. Megede, S. W. Barnett, and E. J. Van Rensburg.** 2001. Genetic analysis of the complete *gag* and *env* genes of HIV type 1 subtype C primary isolates from South Africa. AIDS Res. Hum. Retrovir. **17:**1533–1547.
25. **Esparza, J., R. Klausner, and the Coordinating Committee of the Global HIV/AIDS Vaccine Enterprise.** 2005. The Global HIV/AIDS Vaccine Enterprise: scientific strategic plan. Policy Forum, vol. 2. [Online.] doi: 10.1371 /journal.pmed.0020025.
26. **Etemad-Moghadam, B., G. B. Karlsson, M. Halloran, Y. Sun, D. Schenten, M. Fernandes, N. L. Letvin, and J. Sodroski.** 1998. Characterization of simian-human immunodeficiency virus envelope glycoprotein epitopes recognized by neutralizing antibodies from infected macaques. J. Virol. **72:** 8437–8445.
27. **Faulkner, D. M., and J. Jurka.** 1988. Multiple aligned sequence editor (MASE). Trends Biochem. Sci. **13:**321–322.
28. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the boostrap. Evolution **39:**783–791.
29. **Frost, S. D. W., Y. Liu, S. L. Kosakovsky Pond, C. Chappey, T. Wrin, C. J. Petropoulos, S. J. Little, and D. D. Richman.** 2005. Characterization of human immunodeficiency virus type 1 (HIV-1) envelope variation and neutralizing antibody responses during transmission of HIV-1 subtype B. J. Virol. **79:**6523–6527.
30. **Gallo, R. C., S. Z. Salahuddin, M. Popovic, G. M. Shearer, M. Kaplan, B. F. Haynes, T. J. Palker, R. Redfield, J. Oleske, B. Safai, G. White, P. Foster, and P. D. Markham.** 1984. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. Science **224:**500–503.
31. **Galvin, S. R., and M. S. Cohen.** 2006. Genital tract reservoirs. Curr. Opin. HIV AIDS **1:**162–166.
32. **Gaschen, B., J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber.** 2003. Diversity of considerations in HIV-1 vaccine selection. Science **296:**2354–2360.
33. **Gray, E. S., T. Meyers, G. Gray, D. C. Montefiori, and L. Morris.** 2006. Insensitivity of pediatric HIV-1 subtype C viruses to broadly neutralizing monoclonal antibodies raised against subtype B. PloS Med. **3:**1023–1031.
34. **Haynes, B. F., and D. C. Montefiori.** 2006. Aiming to induce broadly reactive neutralizing antibody responses with HIV-1 vaccine candidates. Exp. Rev. Vaccines **5:**347–363.
35. **Higgins, D. G., and P. M. Sharp.** 1989. Fast and sensitive multiple sequence alignments on a microcomputer. Comput. Appl. Biosci. **5:**151–153.
36. **Javaherian, K., A. J. Langlois, C. McDanal, K. L. Ross, L. I. Echkler, C. L. Jellis, A. T. Profy, J. R. Rusche, D. P. Bolognesi, S. D. Putney, and T. J. Matthews.** 1989. Principal neutralizing domain of the human immunodeficiency virus type 1 envelope protein. Proc. Natl. Acad. Sci. USA **86:**6768–6772.
37. **Kimura, M.** 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:**111–120.
38. **Kolchinsky, P., E. Kiprilov, and J. Sodsroski.** 2001. Increased neutralization sensitivity of CD4-independent human immunodeficiency virus variants. J. Virol. **75:**2041–2050.
39. **Kostrikis, L. G., Y. Cao, H. Ngai, J. P. Moore, and D. D. Ho.** 1996. Quantitative analysis of serum neutralization of human immunodeficiency virus type 1 from subtypes A, B, C, D, E, F, and I: lack of direct correlation between neutralization serotypes and genetic subtypes and evidence for prevalent serum-dependent infectivity enhancement. J. Virol. **70:**445–458.
40. **Kwong, P. D., M. L. Doyle, D. J. Casper, C. Cicala, S. A. Leavitt, S. Majeed, T. D. Steenbeke, M. Venturi, I. Chaiken, M. Fung, H. Katinger, P. W. H. I. Parren, J. Robinson, D. Van Ryk, L. Wang, D. R. Burton, E. Freire, R. Wyatt, J. Sodroski, W. A. Hendrickson, and J. Arthos.** 2002. HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. Nature **420:**678–682.
41. **Lee, W. R., X. F. Yu, W. J. Syu, M. Essex, and T. H. Lee.** 1992. Mutational analysis of conserved N-linked glycosylation sites of human immunodeficiency virus type 1 gp41. J. Virol. **66:**1799–1803.
42. **Leitner, T., B. Korber, M. Daniels, C. Calef, and B. Foley.** 2005. HIV-1 subtype and circulating recombinant form (CRF) reference sequences, 2005, p. 41–48. *In* T. Leitner, B. Foley, B. Hahn, P. Marx, F. McCutchan, J. W. Mellors, S. Wolinski, and B. Korber (ed.), HIV Sequence Compendium 2005. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.Mex.
43. **Letvin, N. L., D. H. Barouch, and D. C. Montefiori.** 2002. Prospects for vaccine protection against HIV-1 infection and AIDS. Annu. Rev. Immunol. **20:**73–99.
44. **Li, B., J. M. Decker, R. W. Johnson, F. Bibollet-Ruche, X. Wei, J. Mulenga, S. Allen, E. Hunter, B. H. Hahn, G. M. Shaw, J. L. Blackwell, and C. A. Derdeyn.** 2006. Evidence for potent autologous neutralizing antibody titers and compact envelopes in early infection with subtype C human immunodeficiency virus type 1. J. Virol. **80:**5211–5218.
45. **Li, M., F. Gao, J. R. Mascola, L. Stamatatos, V. R. Polonis, M. Koutsoukos, G. Voss, P. Goepfert, P. Gilbert, K. M. Greene, M. Bilska, D. L. Kothe, J. F. Salazar-Gonzalez, X. Wei, J. M. Decker, B. H. Hahn, and D. C. Montefiori.** 2005. Human immunodeficiency virus type 1 *env* clones from acute and early subtype B infections for standardized assessments of vaccine-elicited neutralizing antibodies. J. Virol. **79:**10108–10125.
46. **Mascola, J. R.** 2003. Defining the protective antibody response for HIV-1. Curr. Mol. Med. **3:**211–218.
47. **Mascola, J. R., P. D'Souza, P. Gilbert, B. Hahn, N. L. Haigwood, L. Morris, C. J. Petropoulos, V. R. Polonis, M. Sarzotti-Kelsoe, and D. C. Montefiori.** 2005. Recommendations for the design and use of standard virus panels to assess the neutralizing antibody response elicited by candidate human immunodeficiency virus type 1 vaccines. J. Virol. **79:**10103–10107.
48. **Mascola, J. R., and D. C. Montefiori.** 2003. HIV: nature's master of disguise. Nat. Med. **9:**393–394.
49. **Mascola, J. R., S. W. Snyder, O. S. Weislow, S. M. Belay, R. B. Belshe, D. H. Schwartz, M. L. Clements, R. Dolin, B. S. Graham, G. J. Gorse, M. C. Keefer, M. J. McElrath, M. C. Walker, K. F. Wagner, J. G. McNeil, F. E. McCutchan, and D. S. Burke.** 1996. Immunization with envelope subunit vaccine products elicits neutralizing antibodies against laboratory-adapted but not primary isolates of human immunodeficiency virus type 1. J. Infect. Dis. **173:**340–348.
50. **McCutchan, F. E.** 2000. Understanding the genetic diversity of HIV-1. AIDS **14**(Suppl. 3)**:**S31–S44.
51. **Milich, L., B. H. Margolin, and R. Swanstrom.** 1997. Patterns of amino acid variability in NSI-like and SI-like V3 sequences and a linked change in the CD4-binding domain of the HIV-1 Env protein. Virology **239:**108–118.
52. **Montefiori, D. C.** 2004. Evaluating neutralizing antibodies against HIV, SIV and SHIV in luciferase reporter gene assays, p. 12.11.1–12.11.15. *In* J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, and R. Coico (ed.), Current protocols in immunology. John Wiley & Sons, New York, N.Y.
53. **Montefiori, D. C., W. E. Robinson, Jr., S. S. Schuffman, and W. M. Mitchell.** 1988. Evaluation of antiviral drugs and neutralizing antibodies to human immunodeficiency virus by a rapid and sensitive microtiter infection assay. J. Clin. Microbiol. **26:**231–235.
54. **Moore, J. P., and D. R. Burton.** 2004. Urgently needed: a filter for the HIV-1 vaccine pipeline. Nat. Med. **10:**769–771.
55. **Moore, J. P., Y. Cao, J. Leu, L. Qin, B. Korber, and D. D. Ho.** 1996. Inter- and intrasubtype neutralization of human immunodeficiency virus type 1: genetic subtypes do not correspond to neutralization serotypes but partially correspond to gp120 antigenic serotypes. J. Virol. **70:**427–444.
56. **Moore, J. P., and J. Sodroski.** 1996. Antibody cross-competition analysis of the human immunodeficiency virus type 1 gp120 exterior envelope glycoprotein. J. Virol. **70:**1863–1872.
57. **Muster, T., F. Steindl, M. Purtscher, A. Trkola, A. Klima, G. Himmler, F. Ruker, and H. Katinger.** 1993. A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. J. Virol. **67:**6642–6647.
58. **Pantophlet, R., E. O. Saphire, P. Poignard, P. W. H. I. Parren, I. A. Wilson, and D. R. Burton.** 2003. Fine mapping of the interaction of neutralizing and nonneutralizing monoclonal antibodies with the CD4 binding site of human immunodeficiency virus type 1 gp120. J. Virol. **77:**642–658.
59. **Ping, L.-H., J. A. E. Nelson, I. F. Hoffman, J. Schock, S. L. Lamers, M. Goodman, P. Vernazza, P. Kazembe, M. Maida, D. Zimba, M. M. Goodenow, J. J. Eron, Jr., S. A. Fiscus, M. S. Cohen, and R. Swanstrom.** 1999. Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. J. Virol. **73:**6271–6281.
60. **Platt, E. J., K. Wehrly, S. E. Kuhmann, B. Chesebro, and D. Kabat.** 1998. Effects of CCR5 and CD4 cell surface concentrations on infection by macro-

phage tropic isolates of human immunodeficiency virus type 1. J. Virol. **72:**2855–2864.
61. **Purtscher, M., A. Trkola, G. Gruber, A. Buchacher, R. Predl, F. Steindl, C. Tauer, R. Berger, N. Barrett, A. Jungbauer, and H. Katinger.** 1994. A broadly neutralizing human monoclonal antibody against gp41 of human immunodeficiency virus. AIDS Res. Hum. Retrovir. **10:**1651–1658.
62. **Reitter, J. N., R. E. Means, and R. C. Desrosiers.** 1998. A role for carbohydrates in immune evasion in AIDS. Nat. Med. **4:**679–684.
63. **Rizzuto, C., and J. Sodroski.** 2000. Fine definition of a conserved CCR5-binding region on the human immunodeficiency virus type 1 glycoprotein gp120. AIDS Res. Hum. Retrovir. **16:**741–749.
64. **Rizzuto, C. D., R. Wyatt, N. Hernández-Ramos, Y. Sun, P. D. Kwong, W. A. Hendrickson, and J. Sodroski.** 1998. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science **280:**1949–1953.
65. **Robbins, J. B., R. Schneerson, and S. C. Szu.** 1995. Perspective: serum IgG antibody is sufficient to confer protection against infectious diseases by inactivating the inoculum. J. Infect. Dis. **171:**1387–1398.
66. **Robinson, J. E., D. Holton, J. Liu, H. McMurdo, A. Murciano, and R. Gohd.** 1990. A novel enzyme-linked immunosorbent assay (ELISA) for the detection of antibodies to HIV-1 envelope glycoproteins based on immobilization of viral glycoproteins in microtiter wells coated with concanavalin A. J. Immunol. Methods **132:**63–71.
67. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
68. **Salzwedel, K., E. D. Smith, B. Dey, and E. A. Berger.** 2000. Sequential CD4-coreceptor interactions in human immunodeficiency virus type 1 Env function: soluble CD4 activates Env for coreceptor-dependent fusion and reveals blocking activities of antibodies against cryptic conserved epitopes on gp120. J. Virol. **74:**326–333.
69. **Sanders, R. W., M. Venturi, L. Schiffer, R. Kalyanaraman, H. Katinger, K. O. Lloyd, P. D. Kwong, and J. P. Moore.** 2002. The mannose-dependent epitope for neutralizing antibody 2G12 on human immunodeficiency virus type 1 glycoprotein gp120. J. Virol. **76:**7293–7305.
70. **Scanlan, C. N., R. Pantophlet, M. R. Wormald, E. O. Saphire, R. Stanfield, I. A. Wilson, H. Katinger, R. A. Dwek, P. M. Rudd, and D. R. Burton.** 2002. The broadly neutralizing anti-HIV-1 antibody 2G12 recognizes a cluster of α1-2 mannose residues on the outer face of gp120. J. Virol. **76:**7306–7321.
71. **Schønning, K., B. Jansson, S. Olofsson, J. O. Neilsen, and J.-E. S. Hansen.** 1996. Resistance to V3-directed neutralization caused by an N-linked oligosaccharide depends on the quaternary structure of the HIV-1 envelope oligomer. Virology **218:**134–140.
72. **Speck, R. F., K. Wehrly, E. J. Platt, R. E. Atchison, I. F. Charo, D. Kabat, B. Chesebro, and M. A. Goldsmith.** 1997. Selective employment of chemonike receptors as human immunodeficiency virus type 1 coreceptors determined by individual amino acids within the envelope V3 loop. J. Virol. **71:**7136–7139.
73. **Stamatatos, L., M. Wiskerchen, and C. Cheng-Mayer.** 1998. Effect of major deletions in the V1 and V2 loops of a macrophage-tropic HIV-1 isolate on viral envelope structure, cell-entry and replication. AIDS Res. and Hum. Retrovir. **14:**1129–1139.
74. **Stiegler, G., R. Kunert, M. Purtscher, S. Wolbank, R. Voglauer, F. Steindl, and H. Katinger.** 2001. A potent cross-subtype neutralizing human monoclonal antibody against a novel epitope on gp41 of human immunodeficiency virus type 1. AIDS Res. Hum. Retrovir. **17:**1757–1765.
75. **Sullivan, N., Y. Sun, Q. Sattentau, M. Thali, D. Wu, G. Denisova, J. Gershoni, J. Robinson, J. Moore, and J. Sodroski.** 1998. CD4-induced conformational changes in the human immunodeficiency virus type 1 gp120 glycoprotein: consequences for virus entry and neutralization. J. Virol. **72:**4694–4703.
76. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
77. **Trkola, A., T. Dragic, J. Arthos, J. M. Binley, W. C. Olson, G. P. Allaway, C. Cheng-Mayer, J. Robinson, P. J. Maddon, and J. P. Moore.** 1996. CD4-dependent, antibody-sensitive interactions between HIV-1 and its co-receptor CCR-5. Nature **384:**184–187.
78. **Ugolini, S., I. Mondor, P. W. H. I. Parren, D. R. Burton, S. A. Tilley, P. J. Klasse, and Q. J. Sattentau.** 1997. Inhibition of virus attachment to $CD4^+$ target cells is a major mechanism of T cell line-adapted HIV-1 neutralization. J. Exp. Med. **186:**1287–1298.
79. **UNAIDS.** 2005. UNAIDS/WHO AIDS epidemic update: December 2005. [Online.] http://www.unaids.org/epi/2005/.
80. **Vancott, T. C., V. R. Polonis, L. D. Loomis, N. L. Michael, P. L. Nara, and D. L. Birx.** 1995. Differential role of V3-specific antibodies in neutralization assays involving primary and laboratory-adapted isolates of HIV type 1. AIDS Res. Hum. Retrovir. **11:**1379–1390.
81. **van Damme, L., C. Chandeying, G. Ramjee, H. Rees, P. Sirivongrangson, M. Laga, and J. Perriens.** 2000. Safety of multiple daily applications of COL-1492, a nonoxynol-9 vaginal gel, among female sex workers. AIDS **14:**85–88.
82. **van't Wout, A. B., N. A. Kootstra, G. A. Mulder-Kampinga, N. Albrecht-van Lent, H. J. Scherpbier, J. Veenstra, K. Boer, R. A. Coutinho, F. Miedema, and H. Schuitemaker.** 1994. Macrophage-tropic variants initiate human immunodeficiency virus type 1 infection after sexual, parenteral, and vertical transmission. J. Clin. Investig **94:**2060–20607.
83. **Wei, X., J. M. Decker, H. Liu, Z. Zhang, R. B. Arani, J. M. Kilby, M. S. Saag, X. Wu, G. M. Shaw, and J. C. Kappes.** 2002. Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. Antimicrob. Agents Chemother. **46:**1896–1905.
84. **Wei, X., J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw.** 2003. Antibody neutralization and escape. Nature **422:**307–312.
85. **Williamson, C., L. Morris, M. F. Maughan, L. H. Ping, S. A. Dryga, R. Thomas, E. A. Reap, T. Cilliers, J. van Harmelen, A. Pascual, G. Ramjee, G. Gray, R. Johnston, S. A. Karim, and R. Swanstrom.** 2003. Characterization and selection of HIV-1 subtype C isolates for use in vaccine development. AIDS Res. Hum. Retrovir. **19:**133–144.
86. **Wolfs, T. F., G. Zwart, M. Bakker, and J. Goudsmit.** 1992. HIV-1 genomic RNA diversification following sexual and parenteral virus transmission. Virology **189:**103–110.
87. **Wu, L., N. P. Gerard, R. Wyatt, H. Choe, C. Parolin, N. Ruffing, A. Borsetti, A. A. Cardoso, E. Desjardin, W. Newman, C. Gerard, and J. Sodroski.** 1996. CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. Nature **384:**179–183.
88. **Wyatt, R., J. Moore, M. Accola, E. Desjardin, J. Robinson, and J. Sodroski.** 1995. Involvement of the V1/V2 variable loop structure in the exposure of human immunodeficiency virus type 1 gp120 epitopes induced by receptor binding. J. Virol. **69:**5723–5733.
89. **Wyatt, R., and J. Sodroski.** 1998. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. Science **280:**1884–1888.
90. **Xiang, S.-H., N. Doka, R. K. Choudhary, J. Sodroski, and J. E. Robinson.** 2002. Characterization of CD4-induced epitopes on the HIV type 1 gp120 envelope glycoprotein recognized by neutralizing human monoclonal antibodies. AIDS Res. Hum. Retrovir. **18:**1207–1217.
91. **Xiang, S.-H., L. Wang, M. Abreu, C.-C. Huang, P. D. Kwong, E. Rosenberg, J. E. Robinson, and J. Sodroski.** 2003. Epitope mapping and characterization of a novel CD4-induced human monoclonal antibody capable of neutralizing primary HIV-1 strains. Virology **315:**124–134.
92. **Zhang, P. F., P. Bouma, E. J. Park, J. B. Margolick, J. E. Robinson, S. Zolla-Pazner, M. N. Flora, and G. V. Quinnan, Jr.** 2002. A variable region 3 (V3) mutation determines a global neutralization phenotype and CD4-independent infectivity of a human immunodeficiency virus type 1 envelope associated with a broadly cross-reactive, primary virus-neutralizing antibody response. J. Virol. **76:**644–655.
93. **Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Brown, and P. Simmonds.** 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J. Virol. **67:**3345–3356.
94. **Zhu, T., H. Mo, N. Wang, D. S. Nam, Y. Cao, R. A. Koup, and D. D. Ho.** 1993. Genotypic and phenotypic characterization of HIV-1 patients with primary infection. Science **261:**1179–1181.
95. **Zwick, M. B., A. F. Labrijn, M. Wang, C. Spenlehauer, E. O. Saphire, J. M. Binley, J. P. Moore, G. Stiegler, H. Katinger, D. R. Burton, and P. W. H. I. Parren.** 2001. Broadly neutralizing antibodies targeted to the membrane-proximal external region of human immunodeficiency virus type 1 glycoprotein gp41. J. Virol. **75:**10892–10905.

Ming Zhang

Education

- Ph.D. candidate in Bioinformatics, Faculty of Mathematics, Universität Göttingen, Germany
  Degree expected: summer semester, 2007.
- M.S. in Bioinformatics, 2001, Georgia Institute of Technology, USA
- M.S. in Molecular & Cellular Biology, 1999, Shandong Normal University, China
- B.S. in Biological Sciences, 1996, Shandong Normal University, China

Bioinformatics research experience

- 07/2003-present Ph.D. candidate in Bioinformatics. Universität Göttingen, Germany.
  Main research: Molecular evolutionary study; Comparative genome analysis.
- 07/2002-present Graduate Research Assistant in Bioinformatics, HIV Sequence Database Group, Los Alamos National Lab, USA.
  Main research: HIV-1 evolutionary study; HIV-1 subtyping; Sequence database and analysis tools development; Sequence database annotation. Participated in the HIV research in CHAVI (http://www.chavi.org) projects.
- 05/2001-07/2002 Bioinformatics Intern. Centers for Disease Control & Prevention (CDC), USA.
  Main research: Smallpox virus evolutionary analysis; Optimized sequence alignment; Bioinformatics softwares evaluation and development.

Peer-reviewed publications during PhD study

- G. Didier, L. Debomy, M. Pupin, **M. Zhang**, A.Grossmann, C. Devauchelle, I. Laprevotte. Comparing sequences without using alignment scores: application to HIV/SIV subtypings. *BMC Bioinformatics*. 8:1. 2007.
- J. Xie, **M. Zhang**, T. Zhou, X. Hua, L. Tang, W. Wu. Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs. *Nucleic Acids Res.* 35:D183-D187. 2007.
- M. Li, JF.Salazar-Gonzalez, CA.Derdeyn, L.Morris, C.Williamson, JE.Robinson, JM.Decker, Y.Li, MG.Salazar, VR.Polonis, K.Mlisana, S.Karim, K.Hong, KM.Greene, M.Bilska, J.Zhou, S.Allen, E.Chomba, J.Mulenga, C.Vwalika, F.Gao, **M. Zhang**, B.Korber, E.Hunter, BH.Hahn, DC.Montefiori. Genetic and neutralization properties of acute and early subtype C human immunodeficiency virus type 1 molecular env clones from heterosexually acquired infections in Southern Africa. *J Virol*. 80(23):11776-90. 2006.
- J. Esposito, S. Sammons, M. Frace, J. Osborne, M. Rasmussen, **M. Zhang**, D. Govil, I. Damon, R. Kline, M. Laker, Y. Li, G. Smith, H. Meyer, J. LeDuc, R. Wohlhueter. Genome sequence diversity and clues to the evolution of variola virus. *Science*. 2006.
- **M. Zhang**, A. Schultz, C. Calef, C. Kuiken, T. Leitner, B. Korber, B. Morgenstern, M. Stanke. JpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. *Nucleic Acids Res*. 34: W463-5. 2006.
- A. Schultz, **M. Zhang**, T. Leitner, C. Kuiken, B. Korber, B. Morgenstern, M. Stanke.A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. *BMC Bioinformatics*. 7:265. 2006.
- C. Calef, C. Kuiken, J. Szinger, B. Gaschen, W. Abfalterer, **M. Zhang**, N. Tao, R. Funkhouser, K. Yusim, M. Flynn, A. Dalwani, B. Foley, W. Bruno, T. Leitner, B. Korber. Gateway to tools of the HIV and HCV databases In *HIV Molecular Immunology*. p33-56. 2005. Korber et al, editors. Publisher: Los Alamos Natl Lab, Los Alamos, New Mexico.
- **M. Zhang**, K. Wilbe, N. Wolfe, B. Gaschen, J. Carr, T. Leitner. HIV-1 CRF13-cpx revisited: identification of a new sequence from Cameroon and signal for sub-subtype J2. *AIDS Res and Hum Retroviruses*. 21(11):955-960. 2005.
- **M. Zhang**, B. Gaschen, W. Blay, B. Foley, N. Haigwood , C. Kuiken, B. Korber. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes, and influenza hemagglutinin. *Glycobiology*. 14 (12): 1229-1246. 2004.
- Contributed publications as a group member:
  - HIV Sequence Compendium. 2002, 2003, 2004, 2005, 2007. Korber B, et al Eds. Publisher: Los Alamos Natl Lab, Los Alamos, New Mexico.

Technical talks during PhD study

- "N-linked glycosylation sites and sequence lengths characterization in acute, early and chronic HIV-1 infected patients". CHAVI Sequence Analysis Working Group Meeting. Los Alamos, NM. 2006.
- "Something old, something new, and some families: A/G, B/C and B/F recombinants revisited". T-10, Los Alamos, NM. 2006.
- "Reclassification of sequences in HIV database". 13th International Meeting on HIV Dynamics and Evolution. Woods Hole, MA. 2006.
- "N-glycosylation sites variation in HIV-1". Algorithmics group, Max Planck Institute for Molecular Genetics. Berlin, Germany. 2005.
- "Greater HIV genome diversities inferred from re-subtyping of HIV database sequences". German Conference on Bioinformatics 2005. Hamburg, Germany. 2005.
- "N-glycosylation sites variation in HIV-1". The Laboratoire Genome et Informatique (CNRS). Evry, France. 2004.
- "Evolutionary and immunological implications of N-linked glycosylation sites in HIV Envelope". AIDS Vaccine 2004 International Conference. Lausanne, Switzerland. 2004.