AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 23, Number 8, 2007, pp. 1049–1054

DOI: 10.1089/aid.2007.0058

Sequence Note

Genetic Characterization of Three Newly Isolated CRF07_BC Near Full-Length Genomes in China

ZHEFENG MENG, HUI XING, XIANG HE, LIYING MA, WEISI XU, and YIMING SHAO

ABSTRACT

Though HIV-1 CRF07_BC rapidly spread in China, there have been few reports about this subtype since its first genetic characterization nearly 10 years ago. It was urgent and necessary to know the current gene variation of circulating CRF07_BC strains. Xinjiang was the main region for the CRF07_BC epidemic and also an ideal region for research on the viral gene evolution. The strains of Ulumuqi and Yili in Xinjiang were isolated, cloned, and sequenced in this study. Analyses of phylogenetic, potential CTL epitopes and N-glycosites were preformed simultaneously. New CRF07_BC isolates showed higher genetic diversity and more potential N-glycosites than old isolates. It was interesting that although the *env* and *nef* genes are highly variable, highly conserved potential CTL epitopes and N-glycosites were found in deduced gp120 V3 and Nef product of all CRF07_BC isolates. The analysis of the sequences provides some valuable information on the investigation of the epidemiology and on vaccine development.

HIV-1 CRF07_BC RECOMBINANT STRAINS have become one of the most commonly transmitted HIV-1 strains across the country since it was first reported[1–4] in the Xinjiang province of China in 1997. In several recent investigations, it was found that CRF07_BC was responsible for more than 90% of the new infections of HIV in Xinjiang province. Furthermore, in Xinjiang CRF07_BC represents the largest percentage of infections of this subtype in the entire country.[5,6] Although considered not to have originated in Xinjiang,[1–3] CRF07_BC is the most prevalent strain in the region and HIV CRF and CPX have not been reported in the past few years here. Because of the rapid increase of CRF07_BC recombinant virus infections in China, taking proper measures against this pandemic was an urgent public health priority. It is important and necessary to learn more about the genetic characterization of current frequently transmitted HIV-1 CRF07_BC strains. There were four published CRF07_BC near full-length sequences in the Los Alamos HIV database (http://hiv-web. lanl.gov/), but they were all isolated nearly 10 years ago.[1–4] Due to the extensive diversity and rapid divergence in HIV, new CRF07_BC full-length genome sequences must be evaluated to determine the current extent of variation and recombination in viral genomes. This was also essential for the design of an effective AIDS/HIV vaccine.[3] Considering the special and unique conditions of the CRF07_BC epidemic in Xinjiang province, the present research on new CRF07_BC near full-length genome sequences was focused on this region.

There are two main cities for the HIV epidemic in Xinjiang province: Ulumqi and Yili, which lie in the south and north of Xinjiang, respectively. There is no available antivirus therapy in these two regions before sampling data. New samples from these two cities were collected for subtyping and phylogenetic analysis in 2005. Among all confirmed CRF07_BC isolates by analysis of amplified small *env* gene fragments, three (XJN0302, XJN0382, and XJN0084) of 56 isolates from Ulumqi show the highest homogenicity to the CRF07_BC reference CN54, which was isolated from Ulumqi in 1997; two (XJDC6441 and XJDC6431) of 37 isolates from Yili represent the highest and lowest homogenicity with CN54, respectively. Virus from XJN0084, XJDC6441, and XJDC6431 (the geographic data are listed in Table 1) was isolated by cocultivation of healthy donor-derived phytohemagglutinin (PHA)-stim-

Division of Virology and Immunology, State Key Laboratory for Infectious Disease Control and Prevention, National Center for AIDS/STD Control and Prevention, China CDC, Beijing 100050, P.R. China.

Table 1. Epidemiological Data of the Three HIV-1 CRF07_BC Participants in Xinjiang Province, China

Participants	*Sex*	*Risk factor*	*Sampling date*	*Clinical symptom*	*Sampling only*
XJN0084	Male	IDU	2005-11-09	Symptomatic	Ulumqi
XJDC6441	Female	Sex	2005-11-09	Asymptomatic	Yili
XJDC6431-2	Male	IDU	2005-11-09	Asymptomatic	Yili

ulated peripheral blood mononuclear cells (PBMCs). The genomic DNA obtained from the infected PBMCs, using QIAmp blood extraction kits (QIAgen, Germany), was used as a template for the following polymerase chain reaction (PCR) amplification. This study was approved by the study participants with their informed consent, and also by the Committee on Human Research at the National Center for AIDS/STD Prevention and Control in Beijing, China.

The Expand High Fidelity PCR System (Roche Molecular Biochemicals, Mannheim, Germany) was used to amplify the HIV-1 proviral genome from the genomic DNA. The primers were designed to obtain 9.0 kb near full-length sequences as follows: sense primer MSF12: 5′-AAATCTCTAgCAgTggCgCCCgAACAg-3′ and antisense primer MZF-1: 5′-ggTTCgCgAgATAgCCAgAgAgCTCCCAggC-3′. Thermal cycling conditions were as follows: started at 94°C for 2 min, followed by 32 cycles of denaturation at 96°C for 15 sec, an extension at 68°C for 10 min, and a final extension for 15 min at 68°C. The PCR product was purified using the gene purification kit (QIAgen, Germany) and was then ligated with pCR-XL-TOPO vector (Invitrogen Life Technologies). Then 2 μl of the ligation reaction was transformed into *Escherichia coli* Top10 competent cells. Bacterial colonies were grown at 30°C overnight, were then screened for the insert by small fragment PCR, and were confirmed by restriction enzyme analysis of plasmid DNA. Positive clones were sequenced using an ABI 310 Genetic Analyzer (Applied Biosystems) by a walking-primer approach.

The three near full-length nucleotide sequences, designated as XJN0084, XJDC6431-2, and XJDC6441, were about 9.0 kb in length. Analysis of their genomic organization revealed the presence of nine intact potential open reading frames corresponding to the *gag*, *pol*, *vif*, *vpr*, *tat*, *rev*, *vpu*, *env*, and *nef* genes. A deletion of 21 bp was found in *gag* of XJN0084, while multiple insertions appear in the *env* and *nef* genes of XJDC6431-2. Several in-frame stop codons were found in *env* of XJN0084 as well as XJDC6441. There was no shared model and the length of the deletion or insertion varied between three and nine amino acids among these new isolates, which suggested an expanding diversity for circulating CRF07_BC in this region.

The three sequences were aligned with reference sequences (from the Los Alamos HIV database) using Bio-Edit, version 7.0. After managed adjustments were finished, the alignments were used to construct a neighbor-joining tree with 1000 bootstrap replicates, using the Mega3.1 software.[7,8] The "predicted" parental sequences (B′_RL42, C_95 IN21068) for boot scanning and informative site analysis were selected based on the data obtained from analysis of the phylogenetic tree. Boot scanning analysis[9] was performed by the SIMPLOT 2.5 program on neighbor-joining trees for a window of 200 nucleotides moving along the alignment in increments of 20 nucleotides, using the following reference sequences: A_92UG037, B′_RL42, C_95IN21068, D.UG.94UG114, 08_BC.CN.97CNGX, and 07_BC.CN.97.CN54.

The neighbor-joining tree in Fig. 1 shows these three new sequences all clustered with reference sequences of HIV-1 CRF07_BC, supported by 100% of bootstrap trees. Furthermore, XJDC6441 and XJDC6431-2 are on a different branch, whereas XJN0084 and the CRF07_BC reference strains formed one branch supported by 96% of bootstrap. These data suggest these three sequences belonged to the HIV-1 CRF07_BC subtype and were different strains. Bootscan analysis did not show any new breakpoint in three new sequences. The results also confirmed that the three sequences were from the CRF07_BC strain (Fig. 2A–C). They all shared a C/B′ mosaic structure similar to that of the prototype CRF07_BC reference strain CN54 (Fig. 2D); the majority of its genome was from subtype C, while the internal portion of its *gag*, *pol*, *env*, and *nef* gene as well as the first exon of the *tat* gene derived from subtype B′ (Thai-B).

Genetic distances were calculated by Mega3.1 as shown in Table 2. The old CRF07_BC near full-length viral sequences were obtained as CN54, CNGL179, 97CN001, and 98CN009 in 1997 or 1998. These old isolates were obtained in different provinces at different times and these new isolates were obtained in the same province at the same time. However, a higher degree of interisolate diversity was found in the genes among the three new isolates than among the old isolates. As expected, the highest degrees of interisolate diversity on a subgenomic level of the three new isolates were found in the *env* gene. Moreover, the highest divergence between the new and old isolates on a subgenomic level was also found in the *env* gene.

gp120 is considered the most variable gene in the HIV genome, and also the most valuable antigen for vaccine development.[10] As described above, gp120 showed the highest diversity at the subgenomic level of new and old CRF07_BC isolates. As a glycoprotein, glycon contributed approximately 50% of the molecular mass of gp120.[10] Glycosylation of the HIV envelope protein could limit its immunogenicity[11,12] and influence viral replication levels.[12] N-linked glycosylation sites of gp120 of all isolated CRF07_BC were analyzed using the N-GLYCOSITE tools in the HIV database (http://hiv-web.lanl.gov/ content/hiv-db/GLYCOSITE/glycosite.html). In gp120, the old CRF07_BC isolates had 23 common glycosylation sites. XJN0084, XJDC6441, and XJDC6431-2 had 25, 26, and 30 glycosylation sites, respectively. Obviously, new isolates had more potential *N*-glycosites than old isolates. These

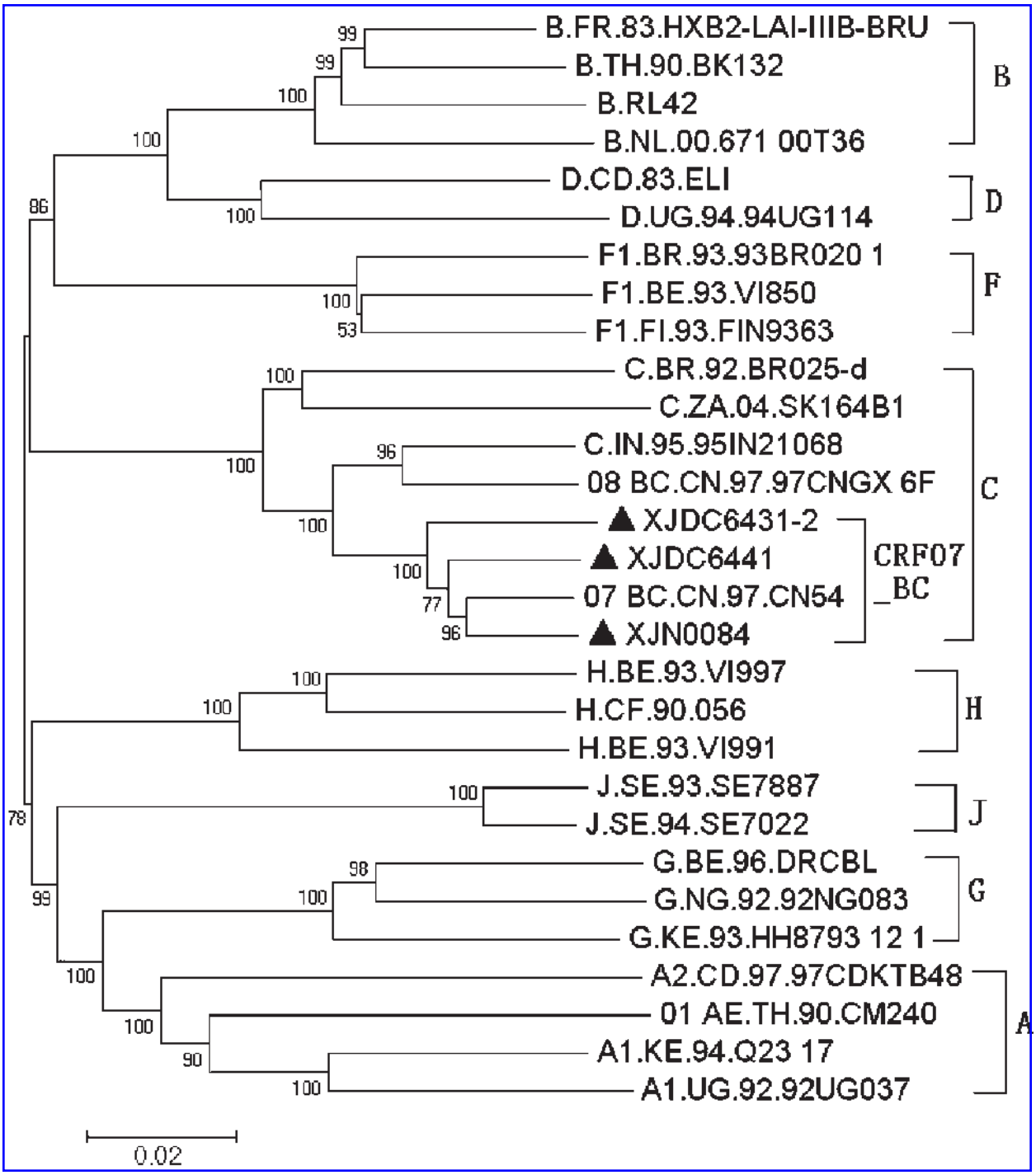

FIG. 1. Phylogenetic tree analysis. A phylogenetic tree was created by neighbor-joining analysis of the three near full-length sequences (XJN0084, XJDC6441, and XJDC6431-2) and the representative HIV-1 clones. Capital letters identify the subtypes and circulating recombinant forms. Values along the branches indicate the bootstrap probability (%) that supports branching. The new isolates (▲) were confirmed as CRF07_BC recombinant strains by the phylogenetic tree.

data were consistent with a recently proposed model of an "evolving glycan shield" to explain HIV escape from antibody-mediated neutralization.[11] Particularly for XJDC6431-2, its gp120 had 30 predicted *N*-glycosites and it also showed the greatest genetic distance from CN54. It is interesting to study the neutralizing activity of antibody to gp120 and the viral biological properties of the XJDC6431-2 strain.

The gp120 glycoprotein contains five conserved regions (C1 to C5) and five variable regions (V1 to V5).[13] The third variable region (the V3 loop) contained the principal neutralizing

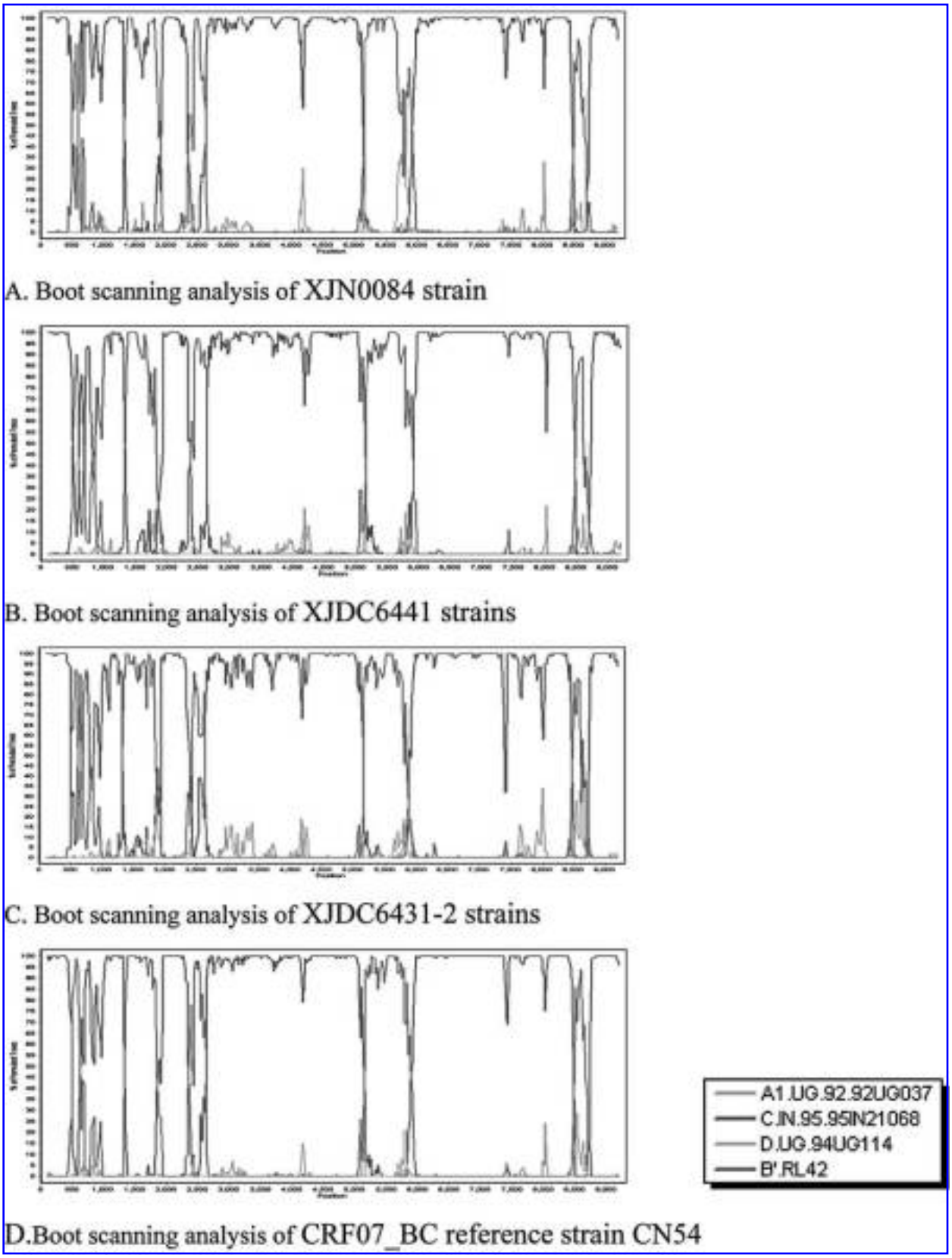

FIG. 2. Boot scanning analysis of near full-length HIV-1 nucleotide sequence. The bootstrap values were plotted for a window of 200 bases moving in increments of 20 bases along the alignment. The *x*-axis is the nucleotide position in the multiple alignments of the full-length HIV-1 sequences, and the *y*-axis is the bootstrap value (percent). Boot scanning plots depicting the relationship of XJN0084 (**A**), XJDC6441 (**B**), XJDC6431-2 (**C**), and CRF07_BC reference strain CN54 (**D**) to the HIV-1 reference strains of the respective subtype, which were indicated in the box.

TABLE 2. GENETIC DISTANCES OF HIV-1 CRF07_BC AMONG OR BETWEEN THE NEW AND OLD ISOLATIONS (%)[a]

Types	*Diversity within the new isolations (%)*	*Diversity within the old isolations (%)*	*Divergence between the two groups (%)*
Genome	4.7 ± 0.2	1.4 ± 0.1	3.6 ± 0.1
gag gene	3.5 ± 0.4	1.1 ± 0.2	2.9 ± 0.3
pol gene	3.0 ± 0.3	1.0 ± 0.1	2.3 ± 0.2
env gene	6.8 ± 0.4	1.8 ± 0.2	5.0 ± 0.3
nef gene	5.7 ± 0.8	2.3 ± 0.4	4.7 ± 0.6

[a]New isolations: XJN0084, XJDC6441, and XJDC6431-2; old isolations: 07_BC.CN.97.CN54, 07_BC.CN-CNGL179, 07_BC,CN.97CN001, and 07_BC.CN98CN009.

determinant (PND) of HIV-1 and was the main target of subtype-specific neutralizing antibodies.[14] Many HIV subtypes have shown a highly variable amino acid in the N- or C-terminal of the V3 loop,[15] which was a huge obstacle to the rational design of a subunit vaccine. However, the V3 loops were highly conserved in sequence among these old CRF07_BC isolates. Even in three new CRF07_BC isolates, there was low amino acid variation among the gp120 V3 loop as follows: a 24G deletion in XJDC6441 and a G6N mutation in XJDC6441 and XJDC6431-2 (Fig. 3). The potential epitopes of all CRF07_BC V3 loops were mapped using the Epitope Location Finder (http://hiv-web.lanl.gov/ content/hiv–db/ELF/epitope_analyzer.html). Initial analysis has demonstrated that probably the cytotoxic T lymphocyte (CTL) epitope in the gp120 V3 loop was completely shared by old CRF07_BC isolates and new isolates; 15 shared epitopes were mapped as follows: A3, A2; B7; B*0702; B*07; B*4201; A11; A2, A3; A*0201; A*2402; A11; A*3002; A2; A*0201; B27; and B*2705. Interestingly, the potential glycosylation site of the gp120 V3 loop remains conserved in both old and new isolates.[16] As initially analyzed using the Epitope Location Finder in this study, these observations clearly predicted a considerable and stable CTL reactivity, suggesting that the functionally and immunologically conserved HIV-1 proteins are strong potential candidates for future vaccine construction.

It became increasing clear that Nef plays an important role in virus infectious activity and host immune protection.[17] It was also noteworthy that the *nef* gene showed high diversity and great divergence in CRF07_BC isolates. For the deduced Nef protein of all CRF07_BC isolates, some amino acid variations were observed in several important motifs [such as the acidic region (EEEE) and V-ATPase recruitment (EE) motifs] in Nef of other HIV-1 subtypes.[17] However, most Nef motifs are well conserved among the new and old CRF07_BC isolates. However, although a few substitutions that involve single amino acid changes were observed in these domains (V143I and H201R), the putative CTL epitopes located on the Nef protein were well conserved, as was shown by utilization of the Epitope Location Finder.

Early investigation of the genetic characterization and outbreak of 07BC recombinant strains was performed in Xinjiang and Guangxi in 1997 and 1998.[1–4] The 07BC isolates from these two regions had very high homogenicity in the genome and share common ancestral strains.[1–4] Based on epidemiological evidence, 07BC could be transmitted from Yunnan to Xinjiang and Guangxi by the heroine routine.[1–3] In fact, many subtype B and C recombinant forms including 07BC were previously reported in Yunnan province.[1,3,4] Unlike conditions in Xinjiang, where rare HIV-1 subtype except 07BC was reported, 08BC not 07BC was identified as the main prevalence strain in the Guangxi region.[3] Therefore, it is reasonable to identify the genetic characterization of current 07BC isolates from Xinjiang for the evaluation of viral evolution and gene variation. In the 10 years from the first report of 07BC in Xinjiang to now, the CRF has experienced extensive variation in the genome accompanied by a rapidly spreading pandemic. Compared to the old CRF07_BC isolate, new isolates show higher gene variation at both the genomic and subgenomic level. However, so

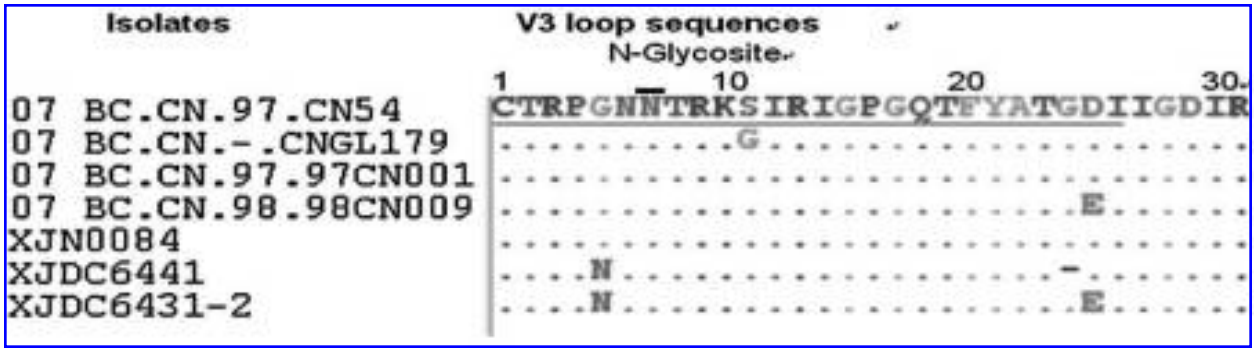

FIG. 3. Alignment of the V3 loop sequences of the CRF07_BC new and old isolates. A dot represents identical residues and a dash represents a gap. Predicted *N*-glycosite is labeled and the best defined CTL epitopes are underlined.

far, no new putative breakpoints were found among new CRF07_BC isolates, which could be related to the special conditions of the HIV epidemic in Xinjiang. Moreover, it is possible that intrasubtype recombination of CRF07_BC is underestimated because of limitations in current technology.

Old 07BC isolates from Ulumuqi of Xinjiang have been used to identify genetic characterization. In previous documents, the same HIV subtype could indicate different epidemiological conditions in different geographic environments. Three new 07BC isolates were reported from Ulumuqi and Yili in Xinjiang: XJN0084 was isolated in Ulumuqi and XJDC6441 and XJDC6431-2 showed the highest and lowest homogeneity with CN54 in all local 07BC isolates. By phylogenetic analysis, new isolates and CN54 share common ancestral strains. Among the new isolates, XJN0084 has the least genetic distance from CN54, XJDC6441 is second, and XJDC6441 is farthest. A similar condition is found in the analysis of the N-linked glycosylation site: the *env* gp120 of XJDC6431-2 has the most potential N-glycosites. As is well known, variation and recombination of the genome present great difficulties in the war against the HIV-1 pandemic. Gene variation and glycan shield are important factors contributing to the rapid spread of CRF07_BC in Xinjiang. However, although the new CRF07_BC isolates showed higher genetic diversity and more glycosylation at the subgenomic level than old isolates nearly 10 years ago, some important CTL epitopes in the products of the most variable *env* and *nef* genes are well conserved, and related N-glycosylation sites with these CTL epitopes are also high conserved. In summary, although further cytological and immunological *in vitro* studies are needed, analysis of HIV near full-length genome sequences provides valuable information for epidemiological investigations and vaccine development.

SEQUENCE DATA

The new sequences generated in this study have been deposited in the GenBank under accession numbers EF368370, EF368371, and EF368372, respectively.

ACKNOWLEDGMENTS

The authors thank the workers at the Xinjiang Provincial Center for Disease Control and Prevention for their help with sample collection. We also thank Dr. Zhong Ping (Shanghai Municipal Center for Disease Control and Prevention, China) for his assistance in the preparation of the manuscript. This work was supported by China 973 National Key Project (2005 CB522903). It was also supported by the National Natural Science Foundation of China (30671847).

REFERENCES

1. Shao Y, Zhao F, Yang W, *et al.*: The identification of recombinant HIV-1 strains in IDUS in southeast and northwest China. Chinese J Exp Clin Virol 1999;13:109–112.
2. Ling Su,Marcus Graf,Hagen von Briesen, Hui Xing *et al.*: Characterization of a virtually full length HIV-1 genome of a prevalent intersubtype (C/B) recombinant strain in China. J Virol 2000;74: 1136–1137.
3. Piyasirisilp S, McCutchan FE, Carr K, *et al.*: A recent outbreak of human immunodeficiency virus type infection in southern China was initiated by two highly homogeneous, geographically separated strains, circulating recombinant form AE and a novel BC recombinant. J Virol 2000;74:11286–11295.
4. Gao F, Robertson DL, Carruthers CD, Morrison SG, *et al.*: Comprehensive panel of near-full length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J Virol 1998;72:5680–5698.
5. Shao Y, Xing H, Pang P, *et al.*: The evolution of subtype C HIV-1 and its recombinant forms among IDUs in China. XIII International AIDS Congress, 167, Durban, South Africa, July 9–14, 2000.
6. Xing H, Chen Z, Liang H, *et al.*: The evolution and rapid spread of B′/C recombinant HIV-1 strains in western China. XIV International AIDS Congress, TuOrC1191, Barcelona, Spain, July 7–12, 2002.
7. Thompson JD, Higgins DG, and Gibson TJ: CLUSTAL W. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
8. Saitou N and Nei M: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–425.
9. Ray.SC. Simplot for Windows, version 2.5. http://www.welch.jhu.edu/~sray/download. Johns Hopkins Medical Institutions, Baltimore, MD, 1999.
10. Wyatt R and Sodroski J: The HIV-1 envelope glycoproteins: Fusogens, antigens, and immunogens. Science 1998; 280:1884–1888.
11. Reitter J, Means R, and Desrosiers R: A role for carbohydrates in immune evasion in AIDS. Nat Med 1998;4:679–684.
12. Wolk T and Schreiber M: N-Glycans in the gp120 V1/V2 domain of the HIV-1 strain NL4-3 are indispensable for viral infectivity and resistance against antibody neutralization. Med Microbiol Immunol (Berl) 2006;195:165–172.
13. Wyatt R, Kwong P, Desjardin E, Sweet R, *et al.*: The antigenic structure of the HIV gp120 envelope glycoprotein. Nature 1998;393:705–711.
14. Zolla-Pazner S: Improving on nature: Focusing the immune response on the V3 loop. Humantibodies 2005;14:69–72.
15. Moore JP, Trkola A, Korber B, Boots LJ, *et al.*: A human monoclonal antibody to a complex epitope in the V3 region of gp120 of human immunodeficiency virus type 1 has broad reactivity within and outside clade B. J Virol 1995;69:122–130.
16. Back NK, Smit L, De Jong JJ, Keulen W, Schutten M, *et al.*: An N-glycan within the human immunodeficiency virus type 1 gp120 V3 loop affects virus neutralization. Virology 1994;199: 431–438.
17. Ndjomou J, Zekeng L, Kaptue L, Daumer M, *et al.*: Functional domains of the human immunodeficiency virus type 1 Nef protein are conserved among different clades in Cameroon. AIDS Res Hum Retroviruses 2006;22:936–944.

Address reprint requests to:
YiMing Shao
National Center for AIDS/STD Control and Prevention
China CDC
No. 27 Nanwei Road
Xuanwu District
Beijing 100050, P.R. China

E-mail: yshao@bbn.cn