 doi:10.20944/preprints202102.0575.v1

Sequence length of HIV-1 subtype B increases over time: analyzing a cohort of patients with hemophilia over 30 years

Short title: Sequence length of HIV-1 increases over time.

Young-Keol CHO[1]*, Jung-Eun KIM[1], Brian T. FOLEY[2]

1. Department of Microbiology, Asan Medical Center, University of Ulsan College of Medicine, Seoul 05505, South Korea; kimje2000@nate.com

2. HIV Databases, Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, USA; btf@lanl.gov

*Correspondence: ykcho2@amc.seoul.kr; Tel: 82-2-3010-4283; Fax: 82-2-3010-4259

Young-Keol Cho ORCID 0000-0003-0424-8911

Jung-Eun Kim E-mail: kimje2000@nate.com

Brian T. Foley E-mail: btf@lanl.gov

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

Abstract

The objective of this study is to investigate whether the sequence length of HIV-1 increases over time. A longitudinal analysis of full-length coding region sequences (FLs) in an outbreak of HIV-1 infection among patients with hemophilia and local controls identified as infected with the Korean subclade B of HIV-1 (KSB). Genes amplified by overlapping RT-PCR or nested PCR were subjected to direct sequencing. In total, 141 FLs were sequentially determined over 30 years in 62 KSB-infected patients. Non-KSB sequences were retrieved from the Los Alamos National Laboratory HIV Database. Phylogenetic analysis indicated that within KSB, 2 FLs from plasma donors O and P comprised two clusters together with 8 and 12 patients with hemophilia, respectively. Signature pattern analysis for the KSB of HIV-1 revealed signature nucleotide residues at 1.05%, compared with local controls. Additionally, in-depth FLs sequence analysis over 30 years in KSB indicates that the KSB FL significantly increases over time before combined antiretroviral therapy (cART) and decreases on cART. Furthermore, the increase in FLs over time significantly occurred in the subtypes B, C and G, but, there was no increase in the subtypes D, A, and F1. Consequently, subtypes F1 and D had the shortest sequence length. Our analysis was extended to compare HIV-1 with HIV-2 and SIVs. Essentially, the longer the sequence length (SIVsm > HIV-2 > SIVcpz > HIV-1), the longer the survival period. The increase in the length of the HIV-1 sequence over time suggests that it might be an evolutionary direction toward attenuated pathogenicity.

Keywords: full-length coding region sequence; HIV-1; Korean subclade B; sequence length; hemophilia; evolution

 doi:10.20944/preprints202102.0575.v1

1. Introduction

We previously conducted a nationwide genetic analysis of HIV-1 with sera from individuals in the early stages of HIV-1 infection (before 1994) to identify the cause of an HIV-1 outbreak among patients with hemophilia in Korea in 1990–1994. These molecular epidemiological studies revealed that viruses from two cash-paid plasma donors were incompletely inactivated in the process of manufacturing clotting factor IX and were identified as the agents of infection among 20 HIV-1-infected patients with hemophilia and HIV-1 infection [1-6]. The viruses in 8 and 12 patients with hemophilia infected with the Korean subclade of HIV-1 subtype B (KSB) originated from plasma donors O and P, respectively. In these studies [1-7], we conducted an in-depth analysis by genes, but a few genes are yet to be studied. The sequence length significantly affects the extent of clustering in a phylogenetic tree [8-9]. KSB is a distinct, monophyletic clade within HIV-1 subtype B, and presumed to have originated from strains in the USA through a founder effect [8, 10-13]. The most recent common ancestor is estimated to have been active around 1984 [14]; however, the earliest case was diagnosed in 1988 [2-6].

In this study, we identified signature pattern residues in the full-length coding region sequence of KSB (FLs). We also performed a phylogenetic analysis at the FL level in 64 patients, including 20 patients with hemophilia B (HPs) [1-4]. We confirmed a previously postulated epidemiological link between the viruses that infected 20 HPs and 2 plasma donors and the viruses that infected local control patients [2-6]. Additionally, their longitudinal sequence analyses over about 30 years indicate that the sequence length in FL significantly increases over time before cART. It is primarily known that the progression in subtype D-infection is four fold faster than subtype B [15]. The sequence length was nearly the shortest among group M. If the estimation that the recently sampled sequences obtained in group M has a longer sequence length than historical samples and that this may be related to viral attenuation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

is true, then it poses a question whether the sequence length in HIV-2 and SIVsm is longer than HIV-1. Hence, we compared the sequence length of HIV-1 with HIV-2 and SIVsm.

To date, our study is one of the most comprehensive and longest longitudinal studies for HIV-1 subtype B evolution *in vivo* originated from a single source of HIV-1 [1-6] and its results provide novel insights into the pathogenesis of HIV-1 infection over time.

2. Materials and Methods

2.1. Ethical Statement

The institutional review board of the Asan Medical Center approved this study's conduct (Code 2012-0390, 4 June 2012). All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki.

2.2 Patients and samples

Four HIV-1 infected plasma donors were diagnosed during primary infection in 1990–1992. Their plasma was used to manufacture clotting factor IX. Viruses from donors O and P were incompletely inactivated. The details have been described previously [1-6, 16, 17]. In this study, FLs were sequenced from 62 KSB-infected patients, including 3 plasma donors (O, P, and R) and 20 patients with hemophilia (designated HP 1–20), 1 CRF02-AG, 1 subtype D, and 1 subtype B (Table S1). In this study, FL sequence denotes from start codon of Gag to Nef without terminal stop codon.

2.3 RNA preparation and FL gene amplification

In patients with hemophilia, sera were collected before 2000 and peripheral blood mononuclear cells (PBMC) were used after 2000. Total RNA was extracted from 300 μL of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

serum samples using a QIAamp UltraSens Viral RNA kit (Qiagen, Hilden, Germany) [2-6]. RNA was reverse-transcribed using Superscript III reverse transcriptase (Invitrogen, Carlsbad, CA, USA). Amplification was performed by five to ten overlapping PCR, followed by direct sequencing using an Applied Biosystems 3730XL DNA Analyzer (Foster City, CA, USA) [5].

2.4 Phylogenetic tree analysis

In total, 70 FL sequences were obtained from 20 patients with hemophilia. Sequences from 42 local controls and 1 subtype B infected patient were aligned against the HIV-1 subtype reference set from the HIV Sequence Database (http://hiv-web.lanl.gov/content/hiv-db/Subtype_REF/align.html). Phylogenetic trees were constructed using the IQ tree with 1,000 bootstrap replicates [18].

2.5 Viral signature pattern analysis (VESPA)

The VESPA program (http://www.hiv.lanl.gov/content/sequence/VESPA/vespa.html) was used to identify sites within each sequence group distinct from other groups [19].

2.6 Statistical analysis

Data are presented as means ± standard deviation. Statistical significance was determined using Student's two-tailed *t*-tests, paired *t*-test, chi-square tests, Fisher's exact test, and Pearson's correlation coefficient using MedCalc. Results were deemed statistically significant when *P*-value was < 0.05.

2.7 Nucleotide sequence data

The GenBank accession numbers for the sequences in this study are AF224507, AY839827, DQ054367, DQ295192-96, DQ837381, JQ316126-37, JQ341411, JQ429433,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

KF561435-43, KJ140245-66, KU869610, MK577478-81, MK871374, MG461319-22, MN237642-46, MN043576-607, MT101871, MT224125-27, MT559044-066, MT582420-24, MT679550-53, and MW405263-343. Non-KSB sequences were retrieved from the Los Alamos National Laboratory (LANL) HIV Database.

3. Results

3.1 Origin of the KSB of subtype B

A major contribution of this study is including the sequences of the earliest KSB-infected patients. We found that patient BGO diagnosed in 1988 was infected with KSB (MT559045). Hence, we obtained FLs KSB in 36, 23, and 3 patients diagnosed in 1988-1991, 1992-1993, and after 1993, respectively (Table S1 and Figure S1). These 36 patients in 1988–1991 correspond to 84% of all 43 KSB-infected patients diagnosed in 1988–1991 [4].

3.2 Molecular epidemiologic data on the FL HIV-1 gene

In 20 HPs, 71 FLs were obtained at 60 time points over 287 ± 99 months (about 24 years) from the outbreak in January 1990. We obtained 169 FLs from 65 patients. Of these, sequences were obtained from 2 or more samples collected on different dates from 30 patients, including 19 patients with hemophilia (Table S1). Phylogenetic analysis revealed that the earliest 62 FLs from 62 patients (20 HPs and 42 local controls) belonged to KSB, whereas 2 FLs from 2 patients belonged to subtypes B and D (Figure S1). The 62 KSB FLs were subdivided into several clusters, including two large clusters (“O,” which comprised 9 sequences and “P,” which comprised 13 sequences) that included 20 HPs and plasma donors O and P. The bootstrap values of the nodes for clusters O, P, and KSB were all 100% using 1,000 bootstrap replicates (Figure S1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

3.3 Korean signature pattern amino acid residues

We previously reported that the signature pattern amino acids at residues 12 and 26 in the Gag and Env proteins [5, 6]. Additionally, 8 and 9 signature pattern amino acids were determined in the Vif and Nef proteins, respectively [20, 21]. We found 31 novel Korean signature nucleotides in the *pol* gene compared with 31 sequences from 15 subtype B infected Korean patients (Table S2). Of those, 11 were nonsynonymous substitutions, and 20 were synonymous substitutions as compared to subtype B.

Overall in the FLs sequences over 8,609–8,618-bp, the signature pattern analysis indicated 91 signature nucleotides (16, 21, 6, 2, 2, 2, 35, and 7 in *gag*, *pol*, *vif*, *vpr*, *tat/rev*, *vpu*, *env*, and *nef* genes, respectively; 1.05%) that distinguish 20 HPs and 42 local controls within KSB ($P < 0.05$). In total, 48 and 43 signature nucleotides originated from clusters O and P, respectively. Of those, only six positions in *gag*, *pol*, *vif*, and *env* genes contained 100% specific nucleotide(s) positions in clusters O and P [2-6] as compared to 0% in local controls (Table S3).

3.4 Sequence identities of HPs compared to plasma donors O and P

The earliest FLs from two donors O and P were 8,606 bp and 8,618 bp, respectively. The sequence identity between the earliest sequences in October 1991 and the last sequences in January 2002 (8,627 bp) from donor O was 96.48%. The sequence similarity in cluster O between the earliest sequences from donor O and each HP's sequence averaged 97.73% ± 0.76%. In cluster P, the sequence similarity between the earliest sequences from donor P and each HP was 97.38% ± 1.36%.

We determined the correlation between the sampling intervals after the outbreak (January 1990) and the number of nucleotide differences observed, relative to those of the corresponding plasma donor. Four patients who were first sampled in 2002 displayed the lowest sequence identity (Figure 1A). The sequence identity dropped significantly over time from the outbreak.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

The lowest sequence identity was 89.8% at 153 months in October 2002 in HP-20. In 19 HPs, sequencing was repeatedly performed with about a 128-month interval. The overall correlation coefficient, γ, was estimated to be −0.85 ($P < 0.001$) (Figure 1A).

When we analyzed 40 FLs before cART, the γ was 0.77 ($P < 0.001$) (Figure 1B), whereas γ was -0.01 on cART compared to the sequences just before cART (Figure 1C). In the same context, the correlations was also significant between sampling year and sequence length before cART in 20 HPs ($r = 0.79$; $P < 0.001$) (Figure 1D).

In addition, we found that this phenomenon occurs in 42 local controls patients without hemophilia ($r = 0.38$; $P < 0.01$) (Figure 1E) and the correlation was more higher in 12 local controls with 2 time points' sequences before cART ($r = 0.61$; $P = 0.001$).

3.5 Sequence length of HIV-1 KSB significantly increases over time before cART

Among the earliest 20 FLs in 20 HPs, we excluded 4 HPs because they were obtained after ten years since the outbreak in January 1990. The earliest sequences over 32 ± 5 months from the outbreak revealed 8,614 ± 17 nucleotides (n = 16) with an increase of 4.1 ± 16 nucleotides than the corresponding donor's sequences. The second sequences obtained more than 5 years since the outbreak were available in 17 HPs before cART (n = 23). This revealed 8,650 ± 22 nucleotides with a significant increase of 39 ± 22 nucleotides over 143 ± 41 months from the outbreak ($P < 0.0001$); however, donors O and P's first sequence obtained in 1991 and 1993 were used.

We previously found that sequence length significantly increased over time before cART (Figure 1B). The sequence length was analyzed by genes to confirm which gene had increased sequence length. We found that the increase occurred due to a significant increase in *env* ($P < 0.0001$) and *nef* ($P < 0.01$) genes **(**Table 1). The increase originated in the variable regions of both genes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

In contrast, we analyzed the change of sequence length within each HP. The sequence length increased by 40 ± 30 nucleotides over an interval of 110 ± 43 months between first and last sequences in 16 HPs with ≥ 2 time point sequences before cART ($P < 0.0001$) (Table 1). This corresponds to annual increase of 4.4 nucleotides and means 47 and 32 years to reach the mean length of SIVsm and HIV-2, respectively.

The sequence length in *env* gene increased by 30 ± 29 nucleotides for the same period ($P< 0.001$).

In contrast, the sequence length decreased by 2.6 ± 39 nucleotides over 138 ± 63 months on cART in 16 HPs than the sequence length just before cART (Figure S2).

To confirm whether this phenomenon occurs in patients without hemophilia, we analyzed sequences in 12 KSB infected patients with ≥2 time point sequences among 42 local controls. There was also a significant increase in sequence length between the 2 FLs over an interval of 104 ± 43 months (33 ± 29 nucleotides) ($P < 0.01$). Of the increase of 33 ± 29 nucleotides, the sequence length in the *env* gene increased by 24 ± 25 nucleotides over 104 ± 43 months ($P < 0.05$), whereas the increase in the *nef* gene was not significant.

Combining 20 HPs with 42 local controls, there were also significant correlations between 100 FLs and sampling year ($P<0.001$) (Figure 1F) and sequence length in the *env* and sampling year ($\boldsymbol{r} = 0.45$, $P < 0.001$), whereas no significant correlation on cART.

Additionally, we found that there was also significant correlation even in individual patients with long-term slow progression (Figure 2).

3.6 Sequence length in HIV-1 subtypes B, C and G significantly increased over time

To confirm whether the sequence length increases in subtypes other than KSB, we randomly selected 64 FLs subtype B from the LANL HIV Database from 1983 to 1998. The correlation between sampling year and sequence length was significant for FLs ($r = 0.43$, $P <$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

0.001) (Figure 3A), *env* (r = 0.26, P < 0.05), and *nef* (r = 0.30, P < 0.05) genes, respectively. Interestingly, there was a significant correlation between sampling year and sequence length in subtype G (P < 0.01) (Figure 3E), whereas no such correlation was found in subtype C under the same condition with subtype B (n = 64). However, there was a significant correlation when subtype C selection was extended to 2013 (P < 0.05).

3.7 Clinical significance of the increase in sequence length in HIV-1 subtype B

There are several reports on that V1-V2 envelope loops and *env* sequences length increase by ~1% per year in the early phases of typical infections [22-24]. There had been several reports on the elongation of V2 region in long-term nonprogressor (LTNPs) [25-28]. We anlyzed the correlation between the number of amino acids in the V2 region and the duration since the diagnosis of KSB and subtype B infection. Consistent with the increase in sequence length over time, there were significant correlations between two abovementioned factors (n = 213, r = 0.60, P < 0.001) as well as between CD4+ T cell count and the number of amino acids in the V2 region in 75 patients (n = 213, r = -0.17, P < 0.05). Additionally, there were significant inverse correlations between FLs and CD4+ T cell count >100/μL in all 65 patients (r = -0.30, P = 0.001) and between sequence length in the *env* gene and CD4+ T cell counts >100/μL before cART (r = -0.34, P < 0.001) (Table S1).

3.8 Sequence length in HIV-1 subtypes A, D, and F1 did not significantly increase over time

In the same way, there was no correlation between sampling year and sequence length in subtype D (Figure 3C). When we compared the sequence length in the *env* gene among thress subtypes, sequence length in subtypes B (n = 64) and C (n = 95) was significantly longer (2,575 ± 21 and 2,575 ± 31) than 2,553 ± 25 in subtype D (P < 0.0001). FLs in subtypes B (8,639 ±

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

28) were significantly longer than in subtype D (n = 64, 8,618 ± 33, P < 0.01). The FL of subtype F1 was 8,598 ± 31 (n = 64), which was significantly shorter than in subtype D (P < 0.01). There was no correlation between sampling year and sequence length in subtypes A, D, , and F1 (Figure 3C, 3D, and 3F) as well as CRF01-AE (n = 63, *r* = -0.04) and CRF02_AG (n = 62, *r* = 0.05).

3.9 Sequence length of HIV-1 is significantly shorter than in HIV-2 and SIVsm.

Persons with HIV-2 infection slowly progress to AIDS as compared to persons with HIV-1 infection. We compared the length of HIV-2 in the LANL database with HIV-1 to investigate whether there is a relationship between this slow progression in HIV-2 infected patients and sequence length. Surprisingly, the length of HIV-2 (n = 34, 8,765 ± 35) was significantly longer than those in SIVcpz and HIV-1 Group M (P<0.0001) (Figure 4). Interestingly, the length of *env* gene of HIV-2 was similar to HIV-1 subtypes B and C except subtype D (P<0.0001). In contrast, the length of *nef* gene of HIV-2 (759 ± 20) was significantly longer than that of HIV-1 (P<0.0001) (Figure S3). Taken together, these data suggest that the difference of the length of *nef* gene might be important for virulence.

Sequence length was significantly longer in SIVsm than in HIV-2 (8,831 ± 32 versus 8,765 ± 35). The increase resulted from the increase of 87-bp and 36-bp in *env* and *nef* genes, respectviely. In conclusion, data suggest that the longer sequence length (SIVsm > HIV-2 > SIVcpz > HIV-1 subtypes G > A/B/KSB/C > D > F1) (Figure 4) corresponds to a longer survival. When we translate the longevity of chimpanzees and sooty mangabeys into human beings (60 years), there was a significant correlation between the sequence length and survival duration (Figure S4).

4. Discussion

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

This study provides an evidence suggesting that sequence length increases over time through extensive sequence data in 20 patients with hemophilia with a well-known natural history and KSB-infected local controls. Additionally, we extended our analysis to subtypes B and C, and the fastest progressing HIV-1 subtype D and extrapolated to the slowest progressing HIV-2 and SIVsm. Due to well-documented primary HIV-1 infection [1-7] and extensive sequence data over 30 years in the 20 HPs with a common source of HIV-1, we were able to confirm that FL significantly increases over time, and the strength of correlation was the stronger in the 20 HPs (Figure 1D) than in local controls (Figure 1E). Otherwise stated, the more homogeneous cohort, the higher strength of correlation.

To our knowledge, this is the first report on the association between sequence length in FLs level and duration of infection. At any given time, viral populations will be dominated by those strains that are most fit at that time [25]. However, there are several reports focusing on V1 and V2 elongation in an elite controller [26-29], and consistent usage of CCR5 coreceptor [30]. As a virulence gene, sequence length in *nef* gene compared to other genes is significantly shortened in HIV-1 and SIVcpz than in HIV-2 and SIVsm [31] (Figure S3), suggesting great implication for pathogenicity [32].

It has been reported that the decreasing rate of CD4+ T cell counts was faster in subtype D than in subtype B [15, 33]. FL in subtype D was significantly shorter than in subtypes A, B, C and G except F1 (Figure 4). That might be the shorter survival time for subtype D viruses to adapt *in vivo* than in other subtypes. However, subtype F1 infected individuals did not show rapid progression as shown in subtype D [34]. Probably, the difference might result from observation that subtype D is the most divergent among HIV-1 group M viruses relative to human genome, whereas subtype F1 is the least divergent [35] despite the highest replication capacity [36] and higher RNA copy at baseline [37] than subtype B.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

It is important to reemphasize that viruses take a symbiotic strategy rather than killing the host or evolving toward attenuated pathogenicity [25]. In fact, the replicative capacity of the HIV-1 in the 2000s was significantly lower than the virus in the 1980s [38]. Consequently, the survival period was actually longer in the infected people in the 2000s even although various factors might be involved. It is also known that replicative fitness by subtype is D > A > C [39]. In this respect, it is possible to understand the increase in sequence length as a strategy or evolutionary direction for the virus to adapt under the immunogical pressure of host and coexist with the host.

The meaning of the increase in sequence length is that the longer the coding sequence (CDS) length, the lower the density of ribosomes, resulting in less efficient protein synthesis [40] and fewer virus copies. This may be the reason why an increase in sequence length such as V1 or V2 elongation occurs in elite controller or LTNP [25-27] as well as in this study, but furthermore, the prognosis in SIVsm and HIV-2 with longer CDS is better than that of HIV-1. Probably, the increase in CDS in HIV-1 over time might be related to attenuated pathogenicity and evolutionary direction.

This study had a few limitations. First, sera sampled from different individuals at various time points were used, and the time lag between primary infection and sampling was particularly long over 10 years in four HPs. Second, sequence data on both LTR regions were not analyzed. Third, sequences before 2000 and after 2000 were originated from serum and PBMC, respectively.

In this study, the sequence length of HIV-1 increased by 4.4 nucleotides per year before cART. When viruses were transmitted to another patient, while considering, the possibility that among quasispecies, HIV-1 with shorter sequence will infect might be greater than HIV-1 with long sequences. Thus, in the population level, the accumulation effect of the increase in

 doi:10.20944/preprints202102.0575.v1

sequence length might be slower than in the individual level because of bottleneck effect. Our novel data suggest that the increase in CDS over time might be evolutionary direction and a path toward attenuated virulence.

Author Contributions: Y.-K.C. designed the experiments. Y.-K.C. and J.-E.K. performed the experiments.Y.-K.C., and B.T.F analyzed the data and wrote the paper.

Funding: This work was supported by a grant from the Korean Society of Ginseng (2012-2020).

Institutional Review Board Statement: The institutional review board of the Asan Medical Center approved this study's conduct (Code 2012-0390, 4 June 2012). All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki.

Acknowledgments: We thank the patients with hemophilia for their cooperation. The authors would like to thank Enago (www.enago.co.kr) for the English language review.

Conflicts of Interest: The authors declare no conflict of interest.

 doi:10.20944/preprints202102.0575.v1

References

1. Cho, Y.K.; Foley, B.T.; Sung, H; Kim, Y.B.; Kim, J.H. Molecular epidemiologic study of a human immunodeficiency virus 1 outbreak in Hemophiliacs B infected through clotting factor 9 after 1990. *Vox Sang.* **2007**, *92*, 113-120.
2. Cho, Y.K.; Jung, Y.S.; Foley, B.T.; Phylogenetic analysis of full-length *pol* gene from Korean Hemophiliacs and plasma donors infected with Korean Subclade B of HIV-1. *AIDS Res. Hum. Retrovir*. **2011**, *27,* 613-621.
3. Cho, Y.K.; Jung, Y.; Lee, J.S.; Foley, B.T. Molecular evidence of HIV-1 transmission in 20 Korean Hemophiliacs; phylogenetic analysis of vif gene. *Haemophilia* **2012**, *18*, 291-299.
4. Cho, Y.K.; Kim, J.E.; Foley, B.T. Phylogenetic analysis of the earliest *nef* gene from Hemophiliacs and local controls in Korea. *BioRes. Open Access* **2012**, *1*, 41-49.
5. Cho, Y.K.; Kim, J.E.; Jeong, D.; Foley, B.T. Signature pattern analysis for the full-length *env* gene of the earliest Korean subclade B of HIV-1: outbreak among Korean Hemophiliacs. *Virus Genes* **2017**, *53*, 789-796.
6. Cho, Y.K.; Kim, J.E.; Foley, B.T. Genetic analysis of the full-length gag gene from the earliest Korean subclade B of HIV-1: an outbreak among Korean Hemophiliacs. *Viruses* **2019**, *11*:545.
7. Cho, Y.K.; Kim, J.E.; WOO, J.H.; Korean Red Ginseng increases defective *pol* gene in peripheral blood mononuclear cells of HIV-1-infected patients; inhibition of its detection during ginseng-based combination therapy. *J Ginseng Res.* **2019**, *43*, 684-691.
8. Leitner, T.; Escanilla, D.; Franzén, C.; Uhlén, M.; Albert, J. Accurate reconstruction of

 doi:10.20944/preprints202102.0575.v1

a known HIV-1 transmission history by phylogenetic tree analysis. *Proc. Natl. Acad. Sci. U S A* **1996,** *93*, 10864-10869.

9. Novitsky, V.; Moyo, S.; Lei, Q.; DeGruttola, V.; Essex, M. Importance of viral sequence length and number of variable and informative sites in analysis of HIV clustering. *AIDS Res. Hum. Retrovir*. **2015**, *31*, 531-542.

10. Daniels, R.S.; Kang, C.; Patel, D.; Xiang, Z.; Douglas, N.W.; Zheng, N.N.; Cho, H.W.; Lee, J.S. An HIV type 1 Subtype B founder effect in Korea: gp160 signature patterns infer circulation of CTL-escape strains at the population level. *AIDS Res. Hum. Retrovir*. **2003**, *19*, 631-641.

11. Korber, B.; Myers, G.; Signature pattern analysis: a method for assessing viral sequence relatedness. *AIDS Res. Hum. Retrovir*. **1992**, *8*, 1549-1560.

12. Korber, B.T.; Foley, B.; Gaschen, B.; Kuiken, C. Epidemiological and immunological implications of the global variability of HIV-1. In: Pataleo, G.; Walker, B.D. (eds), Humana Press, Totowa, NJ. 2001; pp. 1-32, ISBN 978-1-61737-128-8.

13. Junqueira, D.M.; Almeida, S.E. HIV-1 subtype B: traces of a pandemic. *Virology* **2016**, *495*, 173-184.

14. Kim, M.S.; Jang, S.Y.; Park, C.S.; Lee, K.M.; Lee, D.H.; Lee, C.H. Timing and evolution of the most recent common ancestor of the Korean clade HIV subtype B based on *nef* and *vif* sequences. *J. Microbiol.* **2009**, *47*, 85-90.

15. Easterbrook, P.J.; Smith, M.; Mullen, J.; O'Shea, S.; Chrystie, I.; de Ruiter, A.; Tatt, L.D.; Geretti, A.M.; Zuckerman, M. Impact of HIV-1 viral subtype on disease progression and response to antiretroviral therapy. *J. Int. AIDS Soc.* **2010**, *3*, 13:4, doi: 10.1186/1758-2652-13-4.

16. Cho, Y.K.; Kim, J.E.; Foley, B.T. Phylogenetic analysis of near full-length HIV-1 genomic sequences from 21 Korean individuals. *AIDS Res. Hum. Retrovir.* **2013**, *29*, 738-743.

17. Cho, Y.K.; Sung, H.; Bae, I.G.; Oh, H.B.; Kim, N.J.;Woo, J.H.; Kim, Y.B. Full sequence of HIV type 1 Korean subtype B in an AIDS case with atypical Seroconversion: TAAAA at TATA box. *AIDS Res. Hum. Retrovir.* **2005**, *21,* 961-964.

18. Nguyen, L.T.; Schmidt, H.A.; Haeseler, A.; Minh, B.Q. IQ-TREE: A Fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Bio. Evol.* **2015**, *32*, 268-274.

19. Ou, C.Y.; Ciesielski, C.A.; Myers, G.; Bandea, C.I.; Luo, C.C.; Korber, B.T.; Mullins, J.I.; Schochetman, G.; Berkelman, R.L.; Economou, A.N.; et al. Molecular epidemiology of HIV transmission in a dental practice. *Science* **1992**, *256*, 1165-1171.

20. Park, C.S.; Kim, M.S.; Lee, S.D.; Kim, S.S.; Lee, K.M.; Lee, C.H. Molecular phylogenetic analysis of HIV-1 *vif* gene from Korean isolates. *J. Microbiol.* **2006**. *44*, 655-659.

21. Park, C.S.; Lee, D.H.; Lee, K.M.; Lee, C.H. Characterization and signature pattern analysis of Korean clade HIV-1 using *nef* gene sequences. *J. Microbiol.* **2008**, *46*, 88-94.

22. Shankarappa, R.; Margolick, J.B.; Gange, S.J.; Rodrigo, A.G.; Upchurch, D, Farzadegan H, Gupta, P.; Rinaldo, C.R.; Learn, G.H.; et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **1999**, *73*, 10489-10502.

23. Sagar, M.; Wu, X.; Lee, S.; Overbaugh, J. Human immunodeficiency virus type 1 V1-

 doi:10.20944/preprints202102.0575.v1

V2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. *J. Virol.* **2006**, *80*, 9586-9598.

24. Bunnik, E.M.; Euler, Z.; Welkers, M.R.; Boeser-Nunnink, B.D.; Grijsen, M.L.; Prins, J.M.; Schuitemaker, H. Adaptation of HIV-1 envelope gp120 to humoral immunity at a population level. *Nat. Med.* **2010**, *16*, 995-997.

25. Malim, M.H.; Emerman, M.; HIV-1 sequence variation, drift, shift, and attenuation. *Cell* **2001**, *104*, 469-472.

26. Silver, Z.A.; Dickinson, G.M.; Seaman, M.S.; Desrosiers, R.C. A highly unusual V1 region of Env in an elite controller of HIV infection. *J. Virol.* **2019**, 93, e00094-19, doi: 10.1128/JVI.00094-19.

27. Shioda, T.; Oka, S.; Xin, X.; Liu, H.; Harukuni, R.; Kurotani, A.; Fukushima, M.; Shiino, T.; Takebe, Y.; Lwamoto, A.; et al. In vivo sequence variability of Human Immunodeficiency Virus Type 1 envelope gp120: association of V2 extension with slow disease progression. *J. Virol.* **1997**, *71*, 4871-4881.

28. Wang, B.; Spira, T.J.; Owen, S.; Lal, R.B.; Saksena, N.K. HIV-1 strains from a cohort of American subjects reveal the presence of a V2 region extension unique to slow progressors and non-progressors. *AIDS* **2000**, *14*, 213-223.

29. Masciotra, S.; Owen, S.M.; Rudolph, D.; Yang, C.; Wang, B.; Saksena, N.; Spira, T.; Dhawan, S.; Lal, R.B. Temporal relationship between V1V2 variation, macrophage replication, and coreceptor adaptation during HIV-1 disease progression. *AIDS* **2000**, *16*, 1887-1898.

30. Daniels, R.S.; Wilson, P.; Patel, D.; Longhurst, H.; Pattersom, S. Analysis of full-

length HIV type 1 env genes indicates differences between the virus infecting T cells and dendritic cells in peripheral blood of infected patients. *AIDS Res. Hum. Retrovir*. **2004**, *20*, 409-413.

31. Hirao, K.; Andrews, S.; Kuroki, K.; Kusaka, H.; Tadokoro, T.; Kita, S.; Ose, T.; Rowland-Jones, S.L.; Maenaka, K. Structure of HIV-2 Nef reveals features distinct from HIV-1 involved in immune regulation. *iScience* **2020**, *24*, doi:10.1016/j.isci.2019.100758.

32. Schindler, M.; Münch, J.; Kutsch, O.; Li, H.; Santiago, M.L.; Bibollet-Ruche, F.; Müller-Trutwin, M.C.; Novembre, F.J.; Peeters, M.; Courgnaud, V.; et al. Nef-mediated suppression of T cell activation was lost in a lentiviral lineage that gave rise to HIV-1. *Cell* **2006**, *125*, 1055-1067.

33. Cho, Y.K.; Kim, J.E.; Lee, S.U.; Foley, B.T.; Choi, B.S. Impact of HIV-1 subtype and Korean Red Ginseng on AIDS progression: comparison of subtype B and subtype D. *J. Ginseng Res* **2019**, *43*, 312-319.

34. Leite, T.C.; Campos, D.P.; Coelho, A.B.; Teixeira, S.L.; Veloso, V.; Morgado, M.G.; Guimarães, M.L. Impact of HIV-1 subtypes on AIDS progression in a Brazilian cohort. *AIDS Res. Hum Retrovir.* **2017**, *33*, 41-48.

35. Vabret, N.; Bailly-Bechet, M.; Najburg, V.; Müller-Trutwin, M.; Verrier, B.; Tangy, F. The biased nucleotide composition of HIV-1 triggers type I interferon response and correlates with subtype D increased pathogenicity. *PLoS One* **2012**; *7*, doi: 10.1371/journal.pone.0033502.

36. Poveda, E.; de Mendoza, C.; Parkin, N.; Choe, S.; Garcı´a-Gasco, P.; Corral, A.; Soriano, V. Evidence for different susceptibility to tipranavir and darunavir in patients

 doi:10.20944/preprints202102.0575.v1

infected with distinct HIV-1 subtypes. *AIDS* **2008**, *22*, 611–616.

37. Pernas, B.; Grandal, M.; Mena, A.; Castro-Iglesias, A.; Cañizares, A.; Wyles, D.L.; López-Calvo, S.; Pértega, S.; Rodríguez-Osorio, I.; Pedreira, J.D.; et al. High prevalence of subtype F in newly diagnosed HIV-1 persons in northwest Spain and evidence for impaired treatment response. *AIDS* **2014**, *28*, 1837-1840.
38. Ariën, K.K.; Troyer, R.M.; Gali, Y.; Colebunders, R.L.; Arts, E.J.; Vanham, G. Replicative fitness of historical and recent HIV-1 isolates suggests HIV-1 attenuation over time. *AIDS* **2005**, *19*, 1555-1564.
39. Venner, C.M.; Nankya, I.; Kyeyune, F.; Demers, K.; Kwok, C.; Chen, P.L.; Rwambuya, S.; Munjoma, M.; Chipato, T.; Byamugisha, J.; et al. Infecting HIV-1 subtype predicts disease progression in women of sub-Saharan Africa. *EBioMedicine* **2016**, *13*, 305-314.
40. Fernandes, L.D.; de Moura, A.P.S.; Ciandrini, L. Gene length as a regulator for ribosome recruitment and protein synthesis: theoretical insights. *Sci. Rep.* **2017**, *7*, 17409, doi: 10.1038/s41598-017-17618-1.
41. Cho, Y.K.; Kim, J.E.; Woo, J.H. Genetic defects in the nef gene are associated with Korean Red Ginseng intake: monitoring of nef sequence polymorphisms over 20 years. *J Ginseng Res*. **2017**, *41*, 144-150.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

Figure 1. Sequence length increases over time and sequence identity decreases in KSB-infected patients. (A - C): Sequence identity and full-length (FL) sequences according to the interval since the outbreak in January 1990 in 20 HPs. (A) Correlation between the intervals from the outbreak to sampling and the sequence identity of each earliest FL sequence before cART in the 20 patients with hemophilia, compared with the corresponding plasma donor's earliest sequence, respectively. (B) The correlation coefficient (CC), γ, was 0.77 for the 39 FLs before cART ($P < 0.001$). (C) CC, γ was -0.01 on cART. (D - F), Significant increase in sequence length by sampling year in patients with hemophilia (D), 42 KSB-infected local controls (n = 60) (E), and all 62 KSB-infected patients (F).

Figure 2. FLs increased over time in 3 long-term non progressors (LTNP), whereas it decreased in subtype D infected patient. The correlation coefficient was significant between sequence length and duration of infection in 3 LTNPs over 25 years (up to 294, 307 and 286 months). Patients infected with CRF02_AG (A), KSB (B), and subtype B (C) were diagnosed with HIV-1 infection in 1988, 1992, and 1987, respectively. Interestingly, the sequence length in patient subtype B significantly decreased from 8,697 ± 26 (n = 6) just before 286 months (March 2011) to 8,623 ± 24 (n = 6; marked as black star) after 286 months when plasma RNA copy significantly increased ($P<0.001$) [41]. We also found that the sequence length significantly decreased after 307 months. In contrast, it was inversely significant in patient 93-01 who was diagnosed in December 1992 (D). The patient's first sexual contact and diagnosis of pulmonary tuberculosis were done in 1988 and 1989, respectively. He had been treated with Korean Red Ginseng (12,720 g) from April 1993 to August 2004 [33]. Thus, despite most rapidly progressing subtype D infection, he had been remained healthy at least for 12 years [33]. The correlation analysis did not include the data marked as star and white circle on cART. The sequence length on cART was decreased in patients 88-17 and 87-05.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

Figure 3. Sequence length also increases over time in non-KSB infected patients. (A) 64 FLs of Western subtype B were randomly selected from Los Alamos National Laboratory Database over 16 years from 1983 to 1998 (evenly 4 per year). (B - D, and F) There was no such correlation in subtypes A (21 years), C (over 13 years), D (over 28 years), and F1 (over 31 years). Sequences obtained after 1998 were not included in subtypes B and C because cART affects on sequence length. However, there was a significant correlation in subtype C when 108 FLs were selected up to 2013 over 28 years ($r = 0.19$, $P < 0.05$), whereas the significance was lost when FLs were selected up to 2019. (E) There was significant correlation in subtype G over 30 years ($P < 0.01$). Data on subtypes H, K, J, SIVcpz, and SIVsm at the LANL database were not sufficient for the analysis.

Figure 4. The order of length in the near full-length sequence of HIV-1 (from subtype F1 to G), SIVcpz, HIV-2, and SIVsm. Among the HIV-1, the sequence length of subtype F1 is significantly shorter than all others, and the sequence length of subtype D is significantly shorter than the subtypes KSB, A, B, and G. There were significant differences between two neighboring strains of HIV-1 group M, SIVcpz, HIV-2, and SIVsm. To exclude the effect of HAART, we put a limit on the sequence data before 1999 if possible (Figure 3).

Figure S1. Phylogenetic tree analysis of the earliest near full-length sequences (about 8,615 bp from gag to *nef* gene) of 64 Korean patients with HIV-1 using the IQ-tree with 1,000 bootstrap replicates: twenty patients with hemophilia (HPs); 3 plasma donors (O, P, and R); 39 local controls infected with the Korean subclade of HIV-1 subtype B (KSB) and 2 non-KSB-infected

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 February 2021 doi:10.20944/preprints202102.0575.v1

patients. The upper 103 sequences belonged to KSB, and 2 sequences (05CSR3 and 04KBH8) belonged to subtypes B and D, respectively. In total, 9 and 13 patients, including donors O (Cluster O: donor O, 1–4, 6, 8, 10, and 18 as designated by red taxa) and P (Cluster P as designated by red taxa), strongly clustered within the KSB-infected Local controls. The two digits before patient IDs and the one or two digits after patient IDs denote the year and month of sampling, respectively. The bootstrap values of the nodes for clusters O, P, and KSB (designated by bold in red) were 100% by 1,000 bootstrap replicates. Furthermore, each HP's sequences over 30 years also revealed 100% bootstrap values without dispersing except a HP-12 with G-to-A hypermutations in *pol* gene.

Figure S2. Comparison of the increase in the sequence length of the full-length HIV-1 sequence before cART and upon cART in 16 patients with hemophilia. The sequence length significantly increased over 110 ± 43 months (39.4 ± 30 nucleotides) before cART ($P < 0.0001$), whereas it dcreased a little (2.6 ± 39 nucleotides) over 138 ± 63 months on cART.

Figure S3. Comparison of the length in the Nef proteins of HIV-1, SIVcpz, HIV-2, and SIVsm. Among the KSB strain, all three with the longest Nef proteins of 651-bp reveal that 10 amino acids have been added at the variable region (FJ201816 in HP-4, KJ140260 in HP-15 and JQ316130 in patient KJS). In patient KBH infected with subtype D, all Nef proteins were the shortest of 609-bp although the longest Nef (657-bp) was observed in another patient with subtype D (See also Table S1).

 doi:10.20944/preprints202102.0575.v1

Figure S4. Correlation between the sequence length and survival duration. We translated the longevity of sooty mangabeys (sm, about 20 years) into 60-80 years of human beings, and there was a significant correlation between the sequence length and survival duration ($r = 0.90 \sim r = 0.88$, $P < 0.05$). In the same way, in the case of inclusion of chimpanzees (cpz, about 40 - 45 years), there was a significant correlation ($r = 0.84$, $P < 0.05$).