12-13-2021

Origin and evolution of HIV-1 subtype A6

Syed Hani Abidi
Aga Khan University, syed.abidi@aku.edu

Lazzat Aibekova
Nazarbayev University, Nur-Sultan, Kazakhstan

Salima Davlidova
Nazarbayev University, Nur-Sultan, Kazakhstan

Aidana Amangeldiyeva
Nazarbayev University, Nur-Sultan, Kazakhstan

Brian Foley
Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

See next page for additional authors

Recommended Citation

Abidi, S. H., Aibekova, L., Davlidova, S., Amangeldiyeva, A., Foley, B., Ali, S. (2021). Origin and evolution of HIV-1 subtype A6. *PLoS One, 16*(12), e0260604.
Available at: https://ecommons.aku.edu/pakistan_fhs_mc_bbs/958

Authors

Syed Hani Abidi, Lazzat Aibekova, Salima Davlidova, Aidana Amangeldiyeva, Brian Foley, and Syed Ali

PLOS ONE

Origin and evolution of HIV-1 subtype A6

Syed Hani Abidi[1], **Lazzat Aibekova**[2], **Salima Davlidova**[3], **Aidana Amangeldiyeva**[3],
Brian Foley[4], **Syed Ali**[3]*

1 Department of Biological and Biomedical Sciences, Aga Khan University, Karachi, Pakistan, **2** Department
of Biology, School of Sciences and Humanities, Nazarbayev University, Nur-Sultan, Kazakhstan,
3 Department of Biomedical Sciences, Nazarbayev University School of Medicine, Nazarbayev University,
Nur-Sultan, Kazakhstan, **4** Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los
Alamos, New Mexico, United States of America

* syed.ali@nu.edu.kz
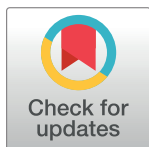
Abstract

Background

HIV outbreaks in the Former Soviet Union (FSU) countries were characterized by repeated
transmission of the HIV variant AFSU, which is now classified as a distinct subtype A sub-
subtype called A6. The current study used phylogenetic/phylodynamic and signature muta-
tion analyses to determine likely evolutionary relationship between subtype A6 and other
subtype A sub-subtypes.

Methods

For this study, an initial Maximum Likelihood phylogenetic analysis was performed using
a total of 553 full-length, publicly available, reverse transcriptase sequences, from A1, A2,
A3, A4, A5, and A6 sub-subtypes of subtype A. For phylogenetic clustering and signature
mutation analysis, a total of 5961 and 3959 pol and env sequences, respectively, were
used.

Results

Phylogenetic and signature mutation analysis showed that HIV-1 sub-subtype A6 likely orig-
inated from sub-subtype A1 of African origin. A6 and A1 pol and env genes shared several
signature mutations that indicate genetic similarity between the two subtypes. For A6,
tMRCA dated to 1975, 15 years later than that of A1.

Conclusion

The current study provides insights into the evolution and diversification of A6 in the back-
drop of FSU countries and indicates that A6 in FSU countries evolved from A1 of African ori-
gin and is getting bridged outside the FSU region.

Introduction

Since 2010, cases of new HIV infections have declined by 23% worldwide, whereas these numbers in Eastern Europe and Central Asia have escalated by 72% between 2010 and 2019, reaching 1.7 million HIV positive people by 2019 [1]. Among the Former Soviet Union (FSU) countries, Russia has the highest prevalence of HIV positive people [1]. Collapse of Soviet Union in the early 90s led to massive transmigration within the FSU countries, which in turn led to spread of infectious diseases. The epidemic in the FSU countries was characterized by repeated transmissions of a variant of HIV, subtype A6 (previously known as A_{FSU}), that initially emerged among people who inject drugs (PWID) and heterosexual populations [2, 3]. A6/A_{FSU} was originally recognized as a sister clade of A1, until it was recently re-classified as a new HIV-1 sub-subtype, A6 [4, 5]. Previous phylogeographic analyses have shown the origin of this variant to be the Demographic Republic of Congo [3]. Currently, A6 is responsible for the fastest growing epidemic in the FSU countries. While travel and trade thrives within FSU countries, and flexible immigration laws facilitate cross-border migrant mobility, risk of HIV transmission continues to grow within this region [6, 7]. Furthermore, considering the A6 epidemic has now been expanding within FSU for almost three decades, chances are that this variant has now been transmitted into countries outside the FSU region.

The present study used phylogenetic/phylodynamic and signature mutation analyses to explore the origin, transmission, and evolution of subtype A6.

Methodology

Collection of sequences, alignment and editing

For this analysis, two different datasets were collected. For initial phylogenetic analysis, a total of 841 full-length reverse transcriptase sequences, from A1, A2, A3, A4, A6 and newly described A8 [8] sub-subtypes of subtype A, spanning the corresponding nucleotides 2550→3869 in HXB2 genome, were retrieved from the Los Alamos HIV database (https://www.hiv.lanl.gov/content/index). After removing the duplicate sequences, 557 sequences were used for initial phylogenetic analysis (S1 Table).

For phylogenetic clustering and signature mutation analysis, a total of 5961 and 3959 *pol* sequences of, respectively, HIV-1 subtype A1 and subtype A6 (S2 Table), spanning 2253→3254 corresponding nucleotides in HXB2 genome were retrieved from HIV Los Alamos database (http://www.hiv.lanl.gov). From the same database, 1075 and 703 *env* sequences of, respectively, HIV-1 subtype A1 and subtype A6 (S2 Table), spanning 6213→8795 corresponding nucleotides in HXB2 genome (complete *env* CDS) were also retrieved. The said genomic regions were selected, since they were represented most in the HIV Los Alamos database. The sequences were aligned and edited using *MEGA 7.0* software. Editing involved removal of non-nucleotide characters and trimming of the ends to bring all the sequences to same length. The final length after editing was 1002 bp for *pol* and 2575 bp for *env*.

Analysis of phylogenetic relationship among different HIV-1 A sub-subtypes

To analyze phylogenetic relationship among different A sub-subtypes, a Maximum Likelihood (ML) phylogenetic tree was constructed using PhyML 3.0 software [9] using a total of 557 *pol* sequences. ML tree was constructed using the following parameters: substitution model: general time reversible (GTR) with empirical equilibrium frequencies, transition / transversion ratio = 4 (default), estimated proportion of invariable sites, estimated gamma shape parameters and number of substitution rate categories = 8 (default); starting trees = BIONJ; Tree

optimization = Tree topology and branch length; Tree improvement = Nearest Neighbor Interchange (NNI); Branch Support = approximate likelihood-ratio test, Shimodaira–Hasegawa-like (aLRT SH-like). The resulting tree was refined and visualized using *Figtree 1.4.2* software (http://tree.bio.ed.ac.uk/software/figtree/). Significant clusters were identified by SH value of $\geq$ 0.85. The final tree was color-coded using the Rainbow Tree software [10]. Phylogenetic relationship among different A sub-subtypes was also confirmed using a Neighbor-joining phylogenetic tree constructed using *TreeMaker* software (https://www.hiv.lanl.gov/components/sequence/HIV/treemaker/treemaker.html) using 1850 *gag* sequences (A1 = 1606, A2 = 56, A3 = 14, A4 = 3, A6 = 171). The genetic distance between A1 and A6 sequences were determined using Bioedit tool, Neighbor-joining distance tree algorithm [11].

Analysis of phylogenetic relationships between A1 and A6 sequences

HIV-1 subtype A1 and A6 *pol* (5961 and 3959, respectively) and *env* (1075 and 703, respectively) sequences were used to construct two separate phylogenetic trees using the *TreeMaker* software, using Neighbor-Joining algorithm. The final tree was visualized using the *Figtree 1.4.2* software (http://tree.bio.ed.ac.uk/software/figtree/).

Effective population size and the time to the most recent common ancestor (tMRCA) for A1 and A6 sequences

Bayesian Markov Chain Monte Carlo (MCMC) inference was applied to estimate the effective population size and tMRCA of HIV subtypes A1 and A6, using BEAST v1.10.4 software [12]. For this analysis, a total of 231 A1 and A6 *pol* sequences were used that were sampled during 1985 to 2017. Sequences were selected using the approach described previously [13, 14], where a limit was set to 5–6 sequences per country and per year. If 5 or less sequences were represented in a country all of them were included in the analysis. Prior to Bayesian analysis, best model for the representative dataset was determined using model selection employed in MEGA 7 software. Using this analysis, Hasegawa–Kishono–Yano (HKY) nucleotide + Gamma + invariant sites (HKY+G+I) was found to best describe the substitution patterns based on its low Bayesian Information Criterion (BIC) scores. Hence, Bayesian analysis was performed using the following parameters: uncorrelated lognormal relaxed molecular clock; Hasegawa–Kishono–Yano (HKY) nucleotide substitution model; estimated base frequencies, and gamma distribution model + invariant sites for heterogeneity among nucleotide sites. The analysis was performed using Bayesian skygrid plot as tree prior [15]. The MCMC chain length was set at 2×10^8, which gave an effective sample size (ESS) of >200. MCMC sample analysis and Bayesian skygrid plot construction was performed using Tracer v1.7.1.

Analysis of A1 and A6 amino acid signatures

For signature mutation analysis, a total of 5961 *pol* and 1075 *env* sequences for HIV-1 subtype A1; and 3959 *pol* and 703 *env* sequences for HIV-1 subtype A6 were used.

The sequences were translated into protein sequences using the *MEGA 7.0* software, and the translated sequences were used to perform signature analysis using the *VESPA* (Viral Epidemiology Signature pattern analysis) tool (https://www.hiv.lanl.gov/content/sequence/VESPA/vespa.html) [16]. The software calculates the frequency of each amino acid at each position in an alignment for the query and background set and selects the positions (signature positions) for which the most common character in the query set differs from that in the background set. For this analysis, A6 and A1 sequences were used, respectively, as query and background. Additionally, both *pol* and *env* A1 and A6 sequences sharing nodes were also analyzed separately for signature mutations.

Results

Analysis of phylogenetic relationships between different HIV-1 A sub-subtypes

An ML phylogenetic tree was constructed to study the relationship among HIV-1 A sub-subtypes. A6 sequences formed a discrete cluster with a strong (>0.90) node support value (Fig 1, blue colored sequences). Ancestral node of A6 cluster was shared by an A1 sequence from Switzerland (accession number JQ403028.1) with a strong node support value (>0.90)

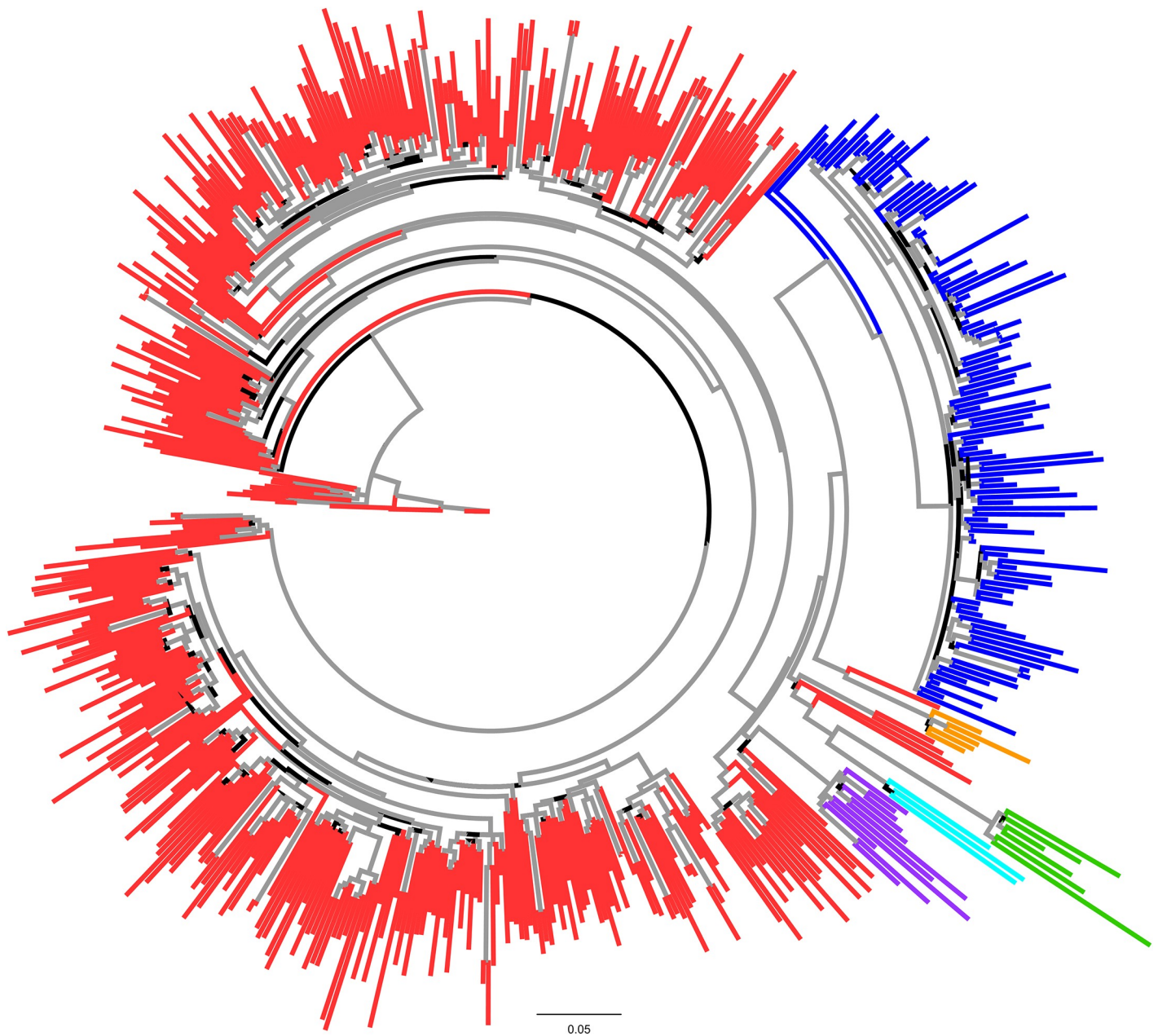**Fig 1. Maximum Likelihood phylogenetic tree of full-length RT HIV sub-subtype A sequences: The tree was constructed using full-length RT sequences.** HIV-1A sub-subtypes are shown in red (A1), green (A2), purple (A3), turquoise (A4), blue (A6) and orange (A8) colors. Grey colored branches signify ≥ 0.9 support value.

https://doi.org/10.1371/journal.pone.0260604.g001

indicating the likely evolution of A6 sub-subtype from A1. Rest of the A sub-subtypes, especially sub-subtype A3 and A8 were found to cluster with A1 sequences at different discrete nodes (Fig 1).

The observation was further corroborated by ML phylogenetic tree based on *gag* sequences, where A6 sequences formed a discrete cluster with a strong (>0.90) node support value. The ancestral node of A6 cluster was shared by an A1 sequence of African (Kenyan; accession numbers GQ430197 and GQ430049) origin (S1 Fig, blue colored sequences) with a strong node support value (>0.90). Furthermore, several Kenyan A1 *gag* sequences exhibited mixing with A6 sequences (S1 Fig, blue colored sequences) indicating close phylogenetic relationship between A6 sub-subtype from A1. Here also, rest of the A sub-subtypes, especially sub-subtype A3 were found to cluster with A1 sequences (S1 Fig), however, the node did not exhibit significant statistical support value.

A1 and A6 clustering was confirmed by a distance tree (S2 Fig, black arrow), where ancestral A1 sequence from Switzerland (accession number JQ403028.1) shared ancestral node of A6 cluster, where genetic distance between A6 and A1 was found to be 0.0715 (S2 Fig).

Effective population size and time of the most recent common ancestor (tMRCA) for A1 and A6 sequences

Since the previous analysis showed A6 sub-subtype to possibly have evolved from A1 sequences (Fig 1), we used Bayesian Skygrid analysis to determine tMRCA of A1 and A6 sub-types. Using this analysis, tMRCA of HIV-1 subtype A1 and A6 was estimated to be around 1950 and 1975, respectively (Fig 2, red boxed arrows). Compared to the origin of tMRCA, the Bayesian Skygrid plot identified a growth in viral effective population size (correlating with effective number of infections and/or transmission opportunities [17]) especially from 1975 to 1985 (Fig 2, pink shaded area). Effective population size declined from 2005 onwards (Fig 2, light blue shaded area) and became stationary from 2010 onwards (Fig 2, green shaded area). The mean substitution rate from this plot was determined to be 2.3×10^{-3}.

Analysis of the A1 and A6 phylogenetic clustering

Phylogenetic relationships among the HIV subtype A1 and A6 *pol* sequences was inferred using a NJ tree. Analysis of the tree revealed that HIV-1 subtype A1 and A6 *pol* sequences from different countries separated into distinct clusters (Fig 3A, magenta and black branches). The A1 cluster primarily comprised of sequences from non-FSU countries, with few FSU sequences (from Belarus, Ukraine, and Russian Federation). In contrast, A6 cluster primarily comprised of sequences from FSU counties, mainly from Russia and Ukraine, while few non-FSU sequences (from Bulgaria, Cyprus, Czech Republic, Mongolia, Poland, Sweden, Slovenia, Turkey, and United States) were also present (Fig 3B). Interestingly, A1 nodes conjoining with A6 cluster were occupied by sequences from Cameroon (accession numbers AY444189 and AY444189) and the Democratic Republic of the Congo (accession number FR666640), both of African origin (Fig 3A, black arrow)—likely indicating African transmission networks. This observation was confirmed by earlier tree (Fig 1), where ancestral nodes of A1 was occupied by sequences from Cameroon (accession numbers AM279375, AM279372, AM279371 and GU201516) and the Democratic Republic of the Congo (accession numbers MH705163, AF286240, AF286238, AM000053).

Similar analysis of A1 and A6 *env* sequences also revealed two completely distinct A1 and A6 clusters, where A1 nodes conjoining with A6 cluster were occupied by sequences also of African origin, namely from the Democratic Republic of the Congo (Fig 3C, black arrow;
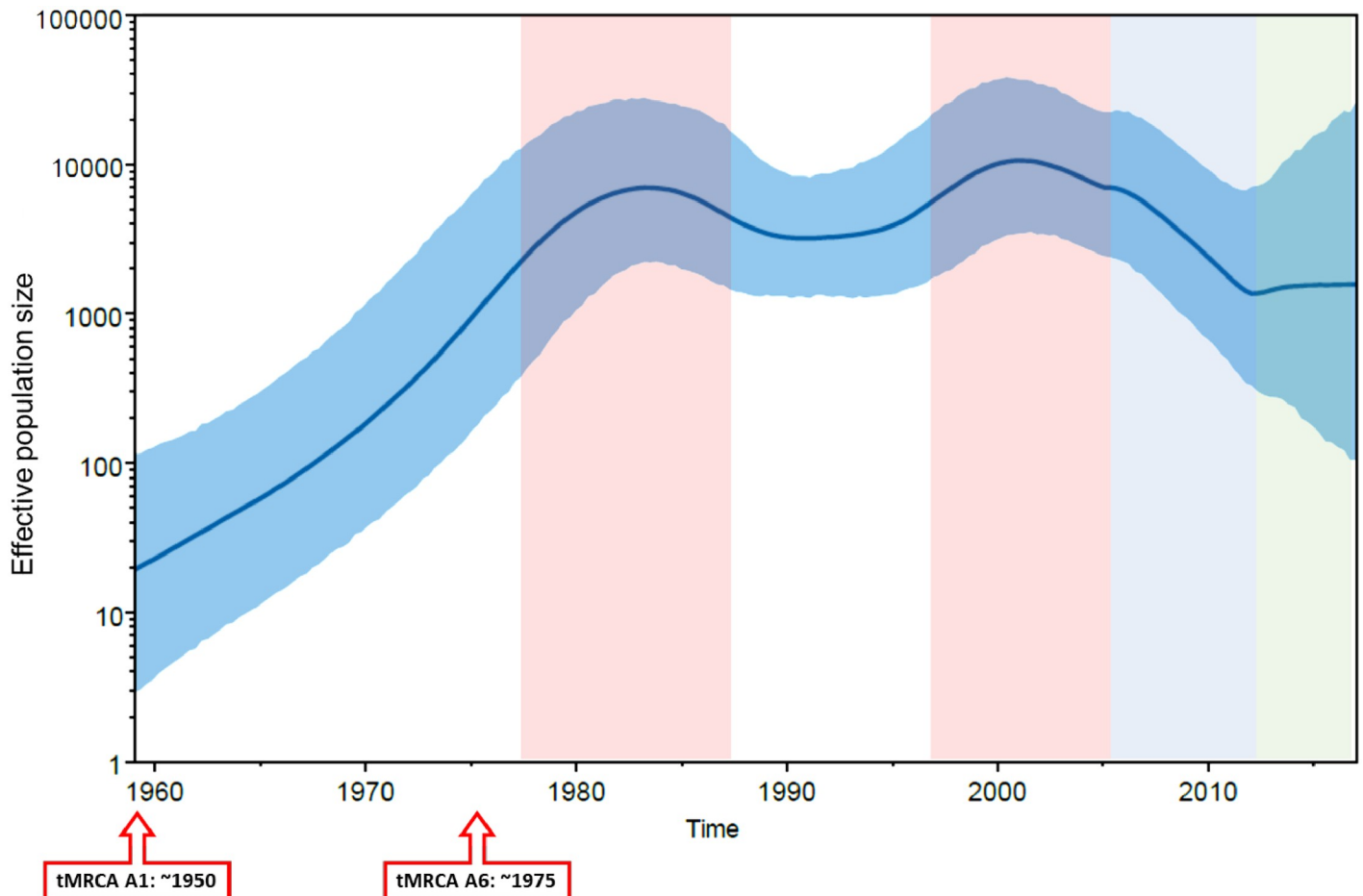
Fig 2. Effective population size and time to most recent common ancestor for HIV-1 subtype A1 and A6. Bayesian Skyline plot, based on a 'relaxed clock' coalescent framework analysis, was constructed using 231 sequences (representing all years and countries). X-axis represents time in years, while Y-axis shows the effective population size. The thick blue line and solid grid (blue) represents mean 95% lower posterior density (HPD) intervals. The tMRCA of HIV-1 subtype A1 and A6 are indicated with red boxed arrows. Pink, light blue and green shaded areas represent the period of increase in viral effective population size, decline and plateau phase, respectively.

https://doi.org/10.1371/journal.pone.0260604.g002

accession numbers AJ401040 and HM623589) supporting the observation of African transmission networks in A6 prevalent countries (Fig 3A and 3C).

Analysis of A1 and A6 amino acid signatures

We employed VESPA software to analyze Pol and Env amino acid sequences to identify amino acid signatures that distinguish A6 from A1 subtype (Figs 4 and 5 and S3 Table). For effective representation, only the dominant amino acids residues at each position are displayed (Figs 4 and 5). Analysis of the Pol sequences revealed 8 sites carrying mutations distinguishing between the subtypes A6 and A1 sequences (Fig 4A). As two A1 sequences from Cameron, and 1 A1 sequence from the Democratic Republic of the Congo occupied the A6 ancestral nodes (Fig 3A, black arrows) A6 signatures were specifically compared with these three sequences. Signature analysis of these A1 Pol sequences revealed presence of A6 signatures at 6/8 sites in sequences from Cameroon, and 5/8 sites in sequence from the Democratic Republic of the Congo (Fig 4B, yellow highlights).
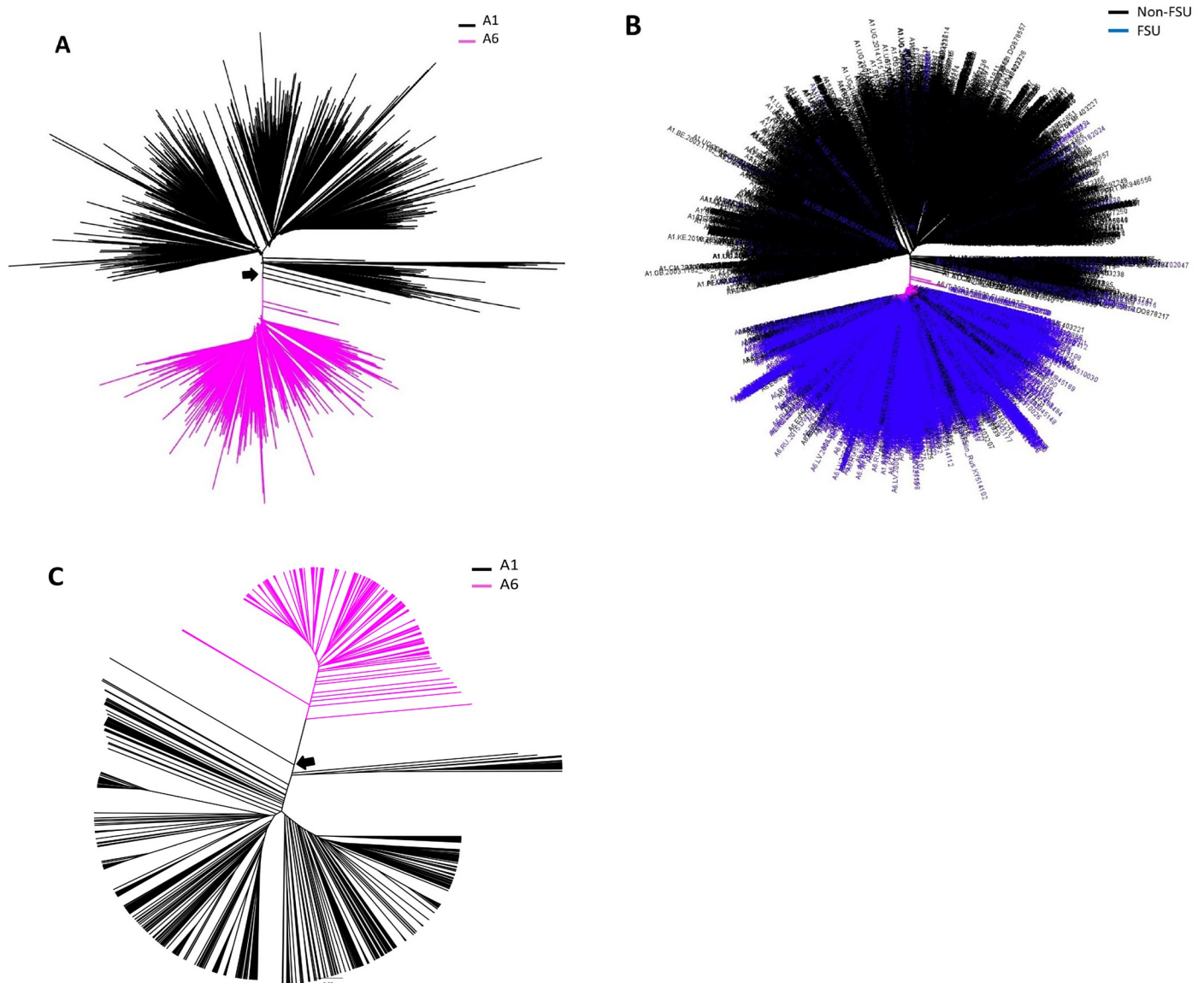
Fig 3. Phylogenetic analysis of HIV subtype A1 and A6 *pol* and *env* sequences: Using NJ method, HIV A and B) *pol* and C) *env* sequences, representing sub-subtypes A1 and A6, were analyzed. A) A1 and A6 sequences are shown in black and magenta, respectively. **B)** *pol* sequences from FSU and non-FSU countries are shown in blue and black colors, respectively. **A** and **C)** Black arrow points to sequences at the node shared (defined as nodes where A1 and A6 sequence(s) converge) between A1 and A6 clusters. Significant nodes had *aLRT* support value of ≥ 0.85.

https://doi.org/10.1371/journal.pone.0260604.g003

Signature mutation analysis of full-length A6 an A1 Env sequences revealed 88 sites to be unique between the two sub-subtypes (Fig 5A). As two A1 sequences from the Democratic Republic of the Congo occupied the A6 ancestral nodes (Fig 3C, black arrows), A6 signatures were compared with these two sequences. Signature analysis of A1 Env sequences from these two countries revealed A6 signatures at 24 (out of 88) different sites (Fig 5B, yellow highlights).

Discussion

Here we performed phylogenetic/phylodynamic and signature mutation analysis to show that FSU-indigenous HIV-1 sub-subtype A6 likely originated from sub-subtype A1. A6 carried

A)

A6 signatures	R	R	L	D	K	I	S	V
A1 signatures	K	K	I	E	T	T	A	I
Positions in alignment	14	57	93	135	138	234	257	278
Positions relative to HXB2 genome	2266	2309	2345	2387	2390	2486	2509	2530

B)

A6 signatures	R	R	L	D	K	I	S	V
A1 signatures	K	K	I	E	T	T	A	I
A1.CM. AY204310	K	R	L	E	K	I	S	V
A1.CM.AY444189	K	R	L	E	R	I	S	V
A1.CD.FR666640	X	X	L	E	K	I	S	V
Positions in alignment	14	57	93	135	138	234	257	278
Positions according to HBX genome	2266	2309	2345	2387	2390	2486	2509	2530

Fig 4. Distinguishing Pol signature mutations in subtypes A1 and A6. The A1 and A6 sequences were compared using VESPA. A1 and A6 sequences were used as background and query, respectively. The atypical (signature) amino acids are shown in **A)** A1 and A6 sequences, and **B)** A1 sequences from two countries (Cameroon and the Democratic Republic of the Congo) sharing ancestral node with the A6 cluster (Fig 3A). Yellow highlights show A6 signature amino acids in A1 sequences.

https://doi.org/10.1371/journal.pone.0260604.g004

signature mutations that distinguished it from A1, while tMRCA for A6 dated to 1975, 15 years later than that of A1.

Phylogenetic analysis of HIV-1 sub-subtype A RT and *gag* sequences revealed that A6 formed a discrete cluster whose ancestral node was shared by A1 sequences, indicating that although A6 has diverged into a distinct sub-subtype, phylogenetically it is related most closely to (and most likely originated from) sub-subtype A1. These observations corroborate the findings in previously published reports on subtype A6, where it has been suggested that A6 was an indigenously, and predominantly, circulating strain in Former Soviet Union (FSU) countries, mainly representing the initial HIV epidemics in this region [4, 6, 18]. For the same reason, A6, originally perceived as a variant of subtype A, was referred to as A$_{FSU}$ because of its homogeneous distribution in the FSU population [4, 6].

Phylogenetic analysis of A1 and A6 sequences only revealed that few A1 sequences, of African origin, occupied the A6 ancestral nodes—likely indicating involvement of African transmission networks in the evolution of A6. These observations are supported by previous reports on A1 and A6 global transmission and phylogenetics, where most of the subtype A1 transmissions were attributed to African strains [3, 13]. In the HIV Los Alamos Database, the high-risk behavior associated with HIV transmission is reported mainly for FSU sequences from Russia and Ukraine. The main high-risk groups are injection drug users, men who have sex with men and heterosexual contact. However, no information exists about sequences from non-FSU countries, except Mongolia, where high-risk behavior was heterosexual contact. In this scenario, the HIV transmission routes can only be speculated.

Back in the 1950s Soviet Union established political relationship with African countries, such as Angola, Algeria, Ghana, Guinea, Mali, Sudan, Morocco and Libya, with aims to implement socialism [19, 20]. Subsequent to these political alliances, back and forth migrant mobility might have led to introduction of the founder strains of A6. Additionally, A6 cluster had sequences from several non-FSU countries, such as Bulgaria, Cyprus, Czech Republic, Mongolia, Poland, Slovenia, and Turkey. A6 sequences from most of these non-FSU countries, except Czech Republic (where A6 sequences were from 1998–2000), were found to be relatively recent, deposited between 2002–2017 (https://www.hiv.lanl.gov/content/index). Analysis of the papers where these samples were originally reported gave insights into A6 transmissions between FSU and non-FSU countries. For example, Bulgaria, Cyprus, Mongolia, Czech

A)

A6 signatures	K	A	R	M	Y	W	L	F	W	M	K	R	K	D	Q	A/T	E	S	S	M
A1 signatures	R	V	M	I	C	L	I	L	G	I	S	K	T	N	H	N	N	N	N/T	I
Positions in alignment	2	3	4	6	10	14	20	21	22	26	28	45	62	93	104	131	132	137	138	153
Positons according to HBX genome	6226	6227	6228	6230	6234	6238	6244	6245	6246	6250	6252	6269	6286	6317	6328	6355	6356	6361	6362	6377

A6 signatures	K	T	K	S	N	G	Q	T	M	T	N	K	D	G	I	T	N	T	I	G
A1 signatures	Q	K	R	E/N/K	S	S	E	A	I	S	E	G	N	A	N	N	E/N/R/K	N	T	N
Positions in alignment	169	170	177	185	186	191	192	202	203	211	234	270	281	283	285	291	295	297	299	302
Positons according to HBX genome	6393	6394	6401	6409	6410	6415	6416	6426	6427	6435	6458	6494	6505	6507	6509	6515	6519	6521	6523	6526

A6 signatures	T	T	V	Y	A	A	I	S	Y	K	V	T	D	D	E	V	T	Q	A	K
A1 signatures	K	A	I	H	S	E	V	A	H	T/N	I	S	G	N	N	E	S	E	S	Q
Positions in alignment	307	316	323	331	337	338	346	347	353	362	371	388	389	396	397	401	402	406	407	437
Positons according to HBX genome	6531	6540	6547	6555	6561	6562	6570	6571	6577	6586	6595	6612	6613	6620	6621	6625	6626	6630	6631	6661

A6 signatures	S	L	G	I	I	K	I	L	A	K	M	D	K	L	H	E	P	K	R	G
A1 signatures	V	L	N	G	L	R	V	I	V	R	L	N/S/R	E	F	N	R	L	E	H	W
Positions in alignment	439	450	458	469	492	498	511	513	516	586	627	642	646	715	721	723	725	732	785	802
Positons according to HBX genome	6663	6674	6682	6693	6716	6722	6735	6737	6740	6810	6851	6866	6870	6939	6945	6947	6949	6956	7009	7026

A6 signatures	G	S	I	F	C	N	K	Q
A1 signatures	L	I	L	I	G	H	R	Q
Positions in alignment	806	814	820	841	842	847	858	861
Positons according to HBX genome	7030	7038	7044	7065	7066	7071	7082	7085

B)

A6 signatures	K	A	R	M	Y	W	L	F	W	M	K	R	K	D	Q	A/T	E	S	S	M
A1 signatures	R	V	M	I	C	L	I	L	G	I	S	K	T	N	H	N	N	N/T	I	
A1_CD_1997_KMST91_AJ401040	R	A	R	I	Y	W	I	F	G	I	N	R	Q	N	H	T	N	S	N/T	M
A1_ZA_2009_707PKE02N3_HM623589	R	V	R	T	C	W	I	I	G	I	R	K	T	N	H	T	N	I	N/T	I
Positions in alignment	2	3	4	6	10	14	20	21	22	26	28	45	62	93	104	131	132	137	138	153
Positons according to HBX genome	6226	6227	6228	6230	6234	6238	6244	6245	6246	6250	6252	6269	6286	6317	6328	6355	6356	6361	6362	6377

A6 signatures	K	T	K	S	N	G	Q	T	M	T	N	K	D	G	I	T	N	T	I	G
A1 signatures	Q	K	R	E/N/K	S	S	E	A	I	S	E	G	N	A	N	N	E/N/R/K	N	T	N
A1_CD_1997_KMST91_AJ401040	K	K	R	G	S	K	Q	T	I	S	Q	G	N	A	T	D	R	N	T	N
A1_ZA_2009_707PKE02N3_HM623589	R	R	R	E	K	R	T	T	I	S	D	G	N	A	T	D	K	N	S	N
Positions in alignment	169	170	177	185	186	191	192	202	203	211	234	270	281	283	285	291	295	297	299	302
Positons according to HBX genome	6393	6394	6401	6409	6410	6415	6416	6426	6427	6435	6458	6494	6505	6507	6509	6515	6519	6521	6523	6526

A6 signatures	T	T	V	Y	A	A	I	S	Y	K	V	T	D	D	E	V	T	Q	A	K
A1 signatures	K	A	I	H	S	E	V	A	H	T/N	I	S	G	N	N	E	S	E	S	Q
A1_CD_1997_KMST91_AJ401040	K	A	V	W	T	A	V	V	H	Q	I	S	G	P	G	E	S	N	P	
A1_ZA_2009_707PKE02N3_HM623589	D	A	I	H	S	I	V	A	Q	T	I	S	S	P	A	Y	N	Q	G	E
Positions in alignment	307	316	323	331	337	338	346	347	353	362	371	388	389	396	397	401	402	406	407	437
Positons according to HBX genome	6531	6540	6547	6555	6561	6562	6570	6571	6577	6586	6595	6612	6613	6620	6621	6625	6626	6630	6631	6661

A6 signatures	S	L	G	I	I	K	I	L	A	K	M	D	K	L	H	E	P	K	R	G
A1 signatures	V	L	N	G	L	R	V	I	V	R	L	N/S/R	E	F	N	R	L	E	H	W
A1_CD_1997_KMST91_AJ401040	I	L	N	G	L	K	V	I	V	K	L	N	E	N	G	P	K	R	W	
A1_ZA_2009_707PKE02N3_HM623589	V	L	E	E	L	R	V	L	V	V	L	L	V	E	N	E	L		W	
Positions in alignment	439	450	458	469	492	498	511	513	516	586	627	642	646	715	721	723	725	732	785	802
Positons according to HBX genome	6663	6674	6682	6693	6716	6722	6735	6737	6740	6810	6851	6866	6870	6939	6945	6947	6949	6956	7009	7026

A6 signatures	G	S	I	F	C	N	K	Q
A1 signatures	L	I	L	I	G	H	R	L
A1_CD_1997_KMST91_AJ401040	Q	N	L	I	C	N	R	L
A1_ZA_2009_707PKE02N3_HM623589	L	T	F	I	C	N	T	L
Positions in alignment	806	814	820	841	842	847	858	861
Positons according to HBX genome	7030	7038	7044	7065	7066	7071	7082	7085

Fig 5. Distinguishing Env signature mutations between subtypes A1 and A6. A1 and A6 sequences were compared using VESPA. A1 and A6 sequences were used as background and query, respectively. The atypical (signature) amino acids are shown in **A)** A1 and A6 sequences, and **B)** A1 sequences from two countries sharing ancestral node with the A6 cluster (Fig 3B). Yellow highlights show A6 signature amino acids in A1 sequences.

Republic and Turkey all reported A6 transmission involving heterosexual and homosexual contact with individuals from Russia and Ukraine [21–25]. Another study reported transmission in Bulgaria and Eastern Europe through PWID networks in Russia and Ukraine [26]. These transmissions might be responsible for the bridging of A6 into Eastern and Central European countries.

Next, we constructed a Bayesian Skygrid plot to deduce tMRCA and transmission dynamics of A1 and A6. The tMRCA of HIV-1 subtype A1 and A6 was estimated to be around 1950 and 1975, respectively, aligning with the hypothesis that A6 originated from A1. The tMRCA for A1 is in agreement with the one previously estimated [13]. *Tongo et al*, suggested that A1 epidemic, initiating within the Congo Basin in late 1950s subtype, expanded to other countries and also diversified to give rise to recombinant forms. The high frequency of subtype A1 in Congo Basin might have contributed to the spread of A1 and A1-derived sequences during the period of 1960–70 [27]. Whereas tMRCA for A6 differed from the tMRCA reported for A$_{FSU}$

(1984) but matched the tMRCA for African ancestor of A_{FSU} (1970) reported by Diez-Fuertes *et al.* [28]. The difference might be due to calculation of tMRCA for A1 and A6 using a single dataset (with A1 and A6 defined as separate taxa) which might have considered A1 ancestral traits. Nonetheless, the tMRCA calculated in this study aligns with the time when initial cases of HIV in the FSU region were reported [12, 18, 20, 29, 30]. Bayesian Skygrid plot identified a growth in viral effective population size (correlating with effective number of infections and/ or transmission opportunities [17]) from 1975 to 1985, and then from late 90s to 2005 –corresponding to the collapse of the Soviet Union followed by increased trans-border migrant mobility within FSU region. The effective population size declined from 2005 onwards and became stationary after 2010. This trend explains the previously reported HIV outbreaks in USSR/FSU countries, where initial HIV cases were reported in late 1980s, followed by a massive increase in HIV cases during mid-1990s stabilizing from 2011 onwards [18].

Signature mutation analysis of Pol and Env sequences revealed, respectively, 8 and 88 sites that carried mutations distinguishing between sub-subtypes A6 and A1 sequences. Aptly, A1 *pol* and *env* sequences conjoining with A6 ancestral nodes (Fig 3A and 3C) exhibited A6 signatures at several sites (Figs 4 and 5). These observations further strengthened the hypothesis that A6 diversified into a sub-subtype distinct from A1. To the best of our knowledge, signatures between HIV subtype A sub-subtypes have not been previously explored. In an amino acid or nucleic acid alignment, signature sites are the variables distinctly represented in query sequences when compared to a background set [31]. Signature mutations can be often unique to one or a small number of viral genetic subtypes/variants, as in the case of a diversified variant, or they may be dominant characters present in the majority of sequences, as in the case of epidemic strains [31]. It is being increasingly recognized that viral subtype-specific genetic variation or signature polymorphisms/mutations can have significant impact on evolution and diversification [32, 33]. For example, signature amino acid residues identified in HIV-1 Tat protein are evolve in a subtype-specific manner (unique for Indian HIV-1C) and can be used to design tailored vaccines against the Indian HIV-1C [34].

Taken together, the current study provides insights into origin, evolution and diversification of A6 in the backdrop of FSU countries. Our observations indicate that A6 in FSU countries likely originated from A1 strains of African origin, and is now bridging into non-FSU countries.

Supporting information

S1 Fig. Maximum Likelihood phylogenetic tree of full-length *gag* HIV sub-subtype A sequences: HIV-1A sub-subtypes are shown in red (A1), green (A2), purple (A3), turquoise (A4), blue (A6) colors. Grey-colored branches signify ≥0.9 support value. For clarity, unrelated subtype A1 clusters have been collapsed. The black arrow indicates the A6 ancestral node shared by A1 sequences. SIV *gag* sequence from chimpanzee was used as an outlier to root the tree.
(TIFF)

S2 Fig. Distance tree of full-length RT HIV sub-subtype A sequences: The tree was constructed using full-length RT sequences. HIV-1A sub-subtypes are shown in red (A1), green (A2), purple (A3), turquoise (A4), blue (A6) colors. The black arrow indicates the A6 ancestral node shared by A1 sequences. Values on the nodes show genetic distances.
(TIFF)

S1 Table. HIV-1 A1, A2, A3, A4 and A6 RT sequences used in the study.
(DOCX)

S2 Table. Accession numbers of HIV-1 A1 and A6 sequences used in the study. HIV-1 A1 pol sequences were from countries located in South America, the Middle East, Asia, Caribbean, Oceania, Former USSR, Europe, North America, Africa, Sub-Saharan Africa, while HIV A6 pol sequences were from countries located in the Middle East, Asia, Oceania, Former USSR, Europe, and North America. Similarly, HIV-1 A1 env sequences were from countries located in Asia, Oceania, Europe, North America, Africa, and Sub-Saharan Africa, while A6 sequences were from countries located in Former USSR and Europe.
(DOCX)

S3 Table. Frequency of all Pol and Env residues, including signature mutations between subtypes A1 and A6. A1 and A6 sequences were compared using VESPA. A1 and A6 sequences were used as background and query, respectively.
(DOCX)

Author Contributions

Conceptualization: Syed Hani Abidi, Syed Ali.

Data curation: Lazzat Aibekova, Salima Davlidova, Aidana Amangeldiyeva, Brian Foley.

Formal analysis: Syed Hani Abidi, Lazzat Aibekova, Salima Davlidova, Aidana Amangeldiyeva, Brian Foley.

Funding acquisition: Syed Ali.

Methodology: Syed Hani Abidi.

Supervision: Syed Ali.

Writing – original draft: Syed Hani Abidi.

Writing – review & editing: Brian Foley, Syed Ali.

References

1. UNAIDS. UNAIDS Data 2020 2020. https://www.unaids.org/sites/default/files/media_asset/2020_aids-data-book_en.pdf.

2. Bobkov A, Kazennova E, Khanina T, Bobkova M, Selimova L, Kravchenko A, et al. An HIV type 1 subtype A strain of low genetic diversity continues to spread among injecting drug users in Russia: study of the new local outbreaks in Moscow and Irkutsk. AIDS Res Hum Retroviruses. 2001; 17(3):257–61. https://doi.org/10.1089/088922201750063188 PMID: 11177409.

3. Díez-Fuertes F, Cabello M, Thomson MM. Bayesian phylogeographic analyses clarify the origin of the HIV-1 subtype A variant circulating in former Soviet Union's countries. Infect Genet Evol. 2015; 33:197–205. https://doi.org/10.1016/j.meegid.2015.05.003 PMID: 25952568.

4. Desire N, Cerutti L, Le Hingrat Q, Perrier M, Emler S, Calvez V, et al. Characterization update of HIV-1 M subtypes diversity and proposal for subtypes A and D sub-subtypes reclassification. Retrovirology. 2018; 15(1):80. Epub 2018/12/24. https://doi.org/10.1186/s12977-018-0461-y PMID: 30577842.

5. Kirichenko A, Lapovok I, Baryshev P, van de Vijver D, van Kampen JJA, Boucher CAB, et al. Genetic Features of HIV-1 Integrase Sub-Subtype A6 Predominant in Russia and Predicted Susceptibility to INSTIs. Viruses. 2020; 12(8). Epub 2020/08/06. https://doi.org/10.3390/v12080838 PMID: 32752001.

6. Aibekova L, Foley B, Hortelano G, Raees M, Abdraimov S, Toichuev R, et al. Molecular epidemiology of HIV-1 subtype A in former Soviet Union countries. PLoS One. 2018; 13(2):e0191891. Epub 2018/02/02. https://doi.org/10.1371/journal.pone.0191891 PMID: 29389951.

7. Amangaldiyeva A, Davlidova S, Baiserkin B, Dzissyuk N, DeHovitz J, Ali S. Implementation of antiretroviral therapy (ART) in former Soviet Union countries. AIDS Res Ther. 2019; 16(1):35. Epub 2019/11/21. https://doi.org/10.1186/s12981-019-0251-1 PMID: 31744530.

8. Mendes Da Silva RK, Monteiro de Pina A II, Venegas Maciera K, Goncalves Morgado M, Lindenmeyer Guimaraes M. Genetic Characterization of a New HIV-1 Sub-Subtype A in Cabo Verde, Denominated A8. Viruses. 2021; 13(6). Epub 2021/07/03. https://doi.org/10.3390/v13061093 PMID: 34201179.

9. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59(3):307–21. https://doi.org/10.1093/sysbio/syq010 PMID: 20525638.

10. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004; 20(2):289–90. https://doi.org/10.1093/bioinformatics/btg412 PMID: 14734327.

11. TA H. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 1999; 41:95–8.

12. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018; 4(1):vey016. https://doi.org/10.1093/ve/vey016 PMID: 29942656.

13. Abidi SH, Kalish ML, Abbas F, Rowland-Jones S, Ali S. HIV-1 subtype A gag variability and epitope evolution. PLoS One. 2014; 9(6):e93415. https://doi.org/10.1371/journal.pone.0093415 PMID: 24892852.

14. Novitsky V, Wang R, Lagakos S, Essex M. HIV-1 Subtype C Phylodynamics in the Global Epidemic. Viruses. 2010; 2(1):33–54. Epub 2010/01/01. https://doi.org/10.3390/v2010033 PMID: 21994599.

15. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. Mol Biol Evol. 2013; 30(3):713–24. https://doi.org/10.1093/molbev/mss265 PMID: 23180580.

16. Korber B, Myers G. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res Hum Retroviruses. 1992; 8(9):1549–60. Epub 1992/09/01. https://doi.org/10.1089/aid.1992.8.1549 PMID: 1457200.

17. Frost SD, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. Philos Trans R Soc Lond B Biol Sci. 2010; 365(1548):1879–90. https://doi.org/10.1098/rstb.2010.0060 PMID: 20478883.

18. Bobkova M. Current status of HIV-1 diversity and drug resistance monitoring in the former USSR. AIDS Rev. 2013; 15(4):204–12. PMID: 24192601.

19. Bervoets J. The Soviet Union in Angola: Soviet and African Perspectives on the Failed Socialist Transformation. Vestnik, The Journal Of Russian And Asian Studies. 2011.

20. Thomson MM, de Parga EV, Vinogradova A, Sierra M, Yakovlev A, Rakhmanova A, et al. New insights into the origin of the HIV type 1 subtype A epidemic in former Soviet Union's countries derived from sequence analyses of preepidemically transmitted viruses. AIDS Res Hum Retroviruses. 2007; 23 (12):1599–604. Epub 2007/12/28. https://doi.org/10.1089/aid.2007.0166 PMID: 18160020.

21. Yilmaz G, Midilli K, Turkoglu S, Bayraktaroglu Z, Kuskucu AM, Ozkan E, et al. Genetic subtypes of human immunodeficiency virus type 1 (HIV-1) in Istanbul, Turkey. Int J Infect Dis. 2006; 10(4):286–90. https://doi.org/10.1016/j.ijid.2005.06.011 PMID: 16516519.

22. Kousiappa I, Van De Vijver DA, Kostrikis LG. Near full-length genetic analysis of HIV sequences derived from Cyprus: evidence of a highly polyphyletic and evolving infection. AIDS Res Hum Retroviruses. 2009; 25(8):727–40. https://doi.org/10.1089/aid.2008.0239 PMID: 19619035.

23. Kousiappa I, van de Vijver DA, Demetriades I, Kostrikis LG. Genetic analysis of HIV type 1 strains from newly infected untreated patients in cyprus: high genetic diversity and low prevalence of drug resistance. AIDS Res Hum Retroviruses. 2009; 25(1):23–35. https://doi.org/10.1089/aid.2008.0168 PMID: 19182918.

24. Reinis M, Bruckova M, Graham RR, Vandasova J, Stankova M, Carr JK. Genetic subtypes of HIV type 1 viruses circulating in the Czech Republic. AIDS Res Hum Retroviruses. 2001; 17(13):1305–10. https://doi.org/10.1089/088922201750461384 PMID: 11559432.

25. Jagdagsuren D, Hayashida T, Takano M, Gombo E, Zayasaikhan S, Kanayama N, et al. The second molecular epidemiological study of HIV infection in Mongolia between 2010 and 2016. PLoS One. 2017; 12(12):e0189605. https://doi.org/10.1371/journal.pone.0189605 PMID: 29244859.

26. Salemi M, Goodenow MM, Montieri S, de Oliveira T, Santoro MM, Beshkov D, et al. The HIV type 1 epidemic in Bulgaria involves multiple subtypes and is sustained by continuous viral inflow from West and East European countries. AIDS Res Hum Retroviruses. 2008; 24(6):771–9. https://doi.org/10.1089/aid.2007.0181 PMID: 18544022.

27. Tongo M, Harkins GW, Dorfman JR, Billings E, Tovanabutra S, de Oliveira T, et al. Unravelling the complicated evolutionary and dissemination history of HIV-1M subtype A lineages. Virus Evol. 2018; 4(1): vey003. Epub 2018/02/28. https://doi.org/10.1093/ve/vey003 PMID: 29484203.

28. Diez-Fuertes F, Cabello M, Thomson MM. Bayesian phylogeographic analyses clarify the origin of the HIV-1 subtype A variant circulating in former Soviet Union's countries. Infect Genet Evol. 2015; 33:197–205. Epub 2015/05/09. https://doi.org/10.1016/j.meegid.2015.05.003 PMID: 25952568.

29. Riva C, Romano L, Saladini F, Lai A, Carr JK, Francisci D, et al. Identification of a possible ancestor of the subtype A1 HIV Type 1 variant circulating in the former Soviet Union. AIDS Res Hum Retroviruses. 2008; 24(10):1319–25. Epub 2008/10/15. https://doi.org/10.1089/aid.2008.0119 PMID: 18851681.

30. Lukashov VV, Cornelissen MT, Goudsmit J, Papuashvilli MN, Rytik PG, Khaitov RM, et al. Simultaneous introduction of distinct HIV-1 subtypes into different risk groups in Russia, Byelorussia and Lithuania. AIDS. 1995; 9(5):435–9. Epub 1995/05/01. PMID: 7639968

31. Aralaguppe SP, Sharma S, Menon M, Prasad VR, Saravanan S, Murugavel KG, et al. The Evolving Profile of the Signature Amino Acid Residues in HIV-1 Subtype C Tat. AIDS Res Hum Retroviruses. 2016; 32(5):503–14. Epub 2015/12/19. https://doi.org/10.1089/AID.2015.0208 PMID: 26678403.

32. McCormick-Davis C, Dalton SB, Singh DK, Stephens EB. Comparison of Vpu sequences from diverse geographical isolates of HIV type 1 identifies the presence of highly variable domains, additional invariant amino acids, and a signature sequence motif common to subtype C isolates. AIDS Res Hum Retroviruses. 2000; 16(11):1089–95. Epub 2000/08/10. https://doi.org/10.1089/08892220050075363 PMID: 10933625.

33. Shankarappa R, Chatterjee R, Learn GH, Neogi D, Ding M, Roy P, et al. Human immunodeficiency virus type 1 env sequences from Calcutta in eastern India: identification of features that distinguish subtype C sequences in India from other subtype C sequences. Journal of virology. 2001; 75(21):10479–87. Epub 2001/10/03. https://doi.org/10.1128/JVI.75.21.10479-10487.2001 PMID: 11581417.

34. Neogi U, Gupta S, Sahoo PN, Shet A, Rao SD, Ranga U, et al. Genetic characterization of HIV type 1 Tat exon 1 from a southern Indian clinical cohort: identification of unique epidemiological signature residues. AIDS Res Hum Retroviruses. 2012; 28(9):1152–6. Epub 2012/01/13. https://doi.org/10.1089/aid.2011.0380 PMID: 22236201.