

UNIVERSITY OF BELGRADE
SCHOOL OF MEDICINE

Marina M. Šiljić

**PHYLOGENETIC ANALYSIS AND
MOLECULAR CHARACTERIZATION OF
HUMAN IMMUNODEFICIENCY VIRUS IN
SERBIA**

Doctoral Dissertation

Belgrade, 2017

UNIVERZITET U BEOGRADU
MEDICINSKI FAKULTET

Marina M. Šiljić

**FILOGENETSKA ANALIZA I
MOLEKULARNA KARAKTERIZACIJA
VIRUSA HUMANE
IMUNODEFICIJENCIJE U SRBIJI**

doktorska disertacija

Beograd, 2017

INFORMATION ON PROMOTER AND DISSERTATION COMMITTEE

Promoter:

Prof. dr Maja Stanojević, Institute of Microbiology and Immunology,
Associate Professor at School of Medicine, University of Belgrade

Dissertation committee:

1. Prof. dr Đorđe Jevtović, Full Professor at School of Medicine, University of Belgrade
2. Prof. dr Tanja Jovanović, Full Professor at School of Medicine, University of Belgrade
3. Prof. dr Aleksandra Knežević, Associate Professor at School of Medicine, University of Belgrade
4. Dr Sanja Glišić, Senior Research Associate, Vinča Institute for Nuclear Sciences
5. Prof. dr Anne-Mieke Vandamme, Full Professor at Katholieke Universiteit Leuven, Faculty of Medicine, Department of Microbiology and Immunology, Rega Institute for Medical Research, Leuven, Belgium

Zahvaljujem se svim zaposlenim na Institutu za mikrobiologiju i imunologiju, Medicinskog fakulteta Univerziteta u Beogradu, na prenetom bogatom naučnom iskustvu, pruženim smernicama i dragocenim savetima.

Posebno se zahvaljujem mojoj mentorki, profesorki dr Maji Stanojević na velikoj privilegiji da budem deo vrhunskog istraživačkog tima i prilici da moje istraživanje virusa humane imunodeficijencije i sveta filogenetike načinim uz brižnu podršku velikog entuzijaste, vrhunskog naučnika i predanog učitelja. Svojim dragocenim savetima, analičkim pristupom, korisnim sugestijama i sjajnim idejama učinila je da ova doktorska disertacija bude ponos mog višegodišnjeg istraživanja.

Veliku zahvalnost dugujem i mojoj kolegici i divnoj prijateljici Valentini Ćirković koja je svojim nesebičnim trudom, velikim entuzijazmom, sjajnim idejama i predanim istraživačkim radom značajno doprinela ovom istraživanju.

Zahvaljujem se asistentkinji dr Danijeli Karalić na neiscrpoj energiji, svesrdnoj pomoći, iskrenim sugestijama i divnoj kolegijalnosti.

Gabrijeli Pavlović se zahvaljujem na njenoj velikoj i nesebičnoj pomoći u laboratorijskom izvođenju istraživanja, trudu, stručnoj a pre svega prijateljskoj podršci.

Zahvaljujem se svim članovima komisije koji su svojim korisnim savetima značajno doprineli poboljšanju konačne verzije ove doktorske disertacije.

Mijn speciale dank gaat uit naar Prof. dr Annemie Vandamme voor de deelname aan de promotie commissie en haar grote bijdrage aan de kwaliteit van mijn proefschrift!!!

Zahvaljujem se lekarima i saradnicima sa odeljenja za HIV/AIDS, Klinike za Infektivne i Tropske bolesti Kliničkog centra Srbije u Beogradu: Prof dr Đorđu Jevtoviću, Doc. dr Jovanu Raninu, dr Ivani Pešić Pavlović. Posebno veliku zahvalnost dugujem dr Dubravki Salemović na izuzetnoj saradnji, nesebičnoj pomoći, entuzijazmu i velikoj posvećenosti ovom istraživanju. Čast je i privilegija sarađivati sa takvim čovekom!

Kosti Stanojević, studentu Elektrotehničkog fakulteta u Beogradu, zahvaljujem se jer je konstruisao program koji mi je omogućio da analiziram procenat različitih baza istovremeno prisutnih na istoj poziciji virusnog genoma.

Najviše se zahvaljujem mojim roditeljima na vaspitanju, безусловnoj podršci i beskrajnoj ljubavi. Mom suprugu i najboljem prijatelju Dejanu hvala na strpljenju, osloncu i beskompromisnoj ljubavi koju mi pruža. Mom sinu Mihailu hvala na najčistijoj i najiskrenijoj ljubavi koja me je oplemenila, kao i na neiscrpoj energiji koja me svakim danom čini sve jačom i odlučnijom.



POSVEĆENO

najhrabrijoj ženi, mojoj voljenoj majci

Vesni Stefanović

SUMMARY

**Phylogenetic Analysis and Molecular Characterization of Human
Immunodeficiency Virus in Serbia****Introduction**

Human immunodeficiency virus (HIV) is a retrovirus, the causative agent of Acquired immunodeficiency syndrome (AIDS). Since the beginning of the epidemic over 35 years ago, more than 78 million people have been infected so far and over 30 million have died. The high genetic variability and rapid evolution of HIV have been critical to its persistence and spread throughout the world. HIV-1 and HIV-2 comprise two distinct types of HIV. HIV-1 has diversified extensively into numerous genetic forms, including four groups (M, N, O, P), of which group M is causing the pandemic of HIV infection and AIDS. Group M viruses are further classified in multiple phylogenetically distinct subtypes (A-D, F, G, H, J and K), sub-subtypes (A1, A2, F1 and F2) and numerous recombinant forms. The global distribution of HIV-1 is complex and dynamic with regional epidemics representing only a subset of the global diversity. Molecular phylogenetic analysis, a method of reconstructing evolutionary relationships between nucleotide sequences, is one of the strategies for studying viral diversity and transmission dynamics. It is estimated that around half of HIV infected people are undiagnosed, making identification of transmission networks important for targeted public health intervention programs. In this doctoral dissertation modern phylogenetic techniques were used to identify HIV-1 sequences from Serbia, to characterize

the molecular epidemiology and transmission dynamics which is crucial to understand the actual characteristics of the Serbian HIV-1 epidemic.

The Aims of the Study

The aims of this research were to determine the current prevalence and distribution of HIV-1 subtypes in Serbia. The objectives also included identification of local transmission networks and reconstruction of the history of the introduction of founder strains. Different molecular footprints on HIV-1 sequences and their association with duration of infection and phylogenetic clustering were also investigated.

Materials and Methods

The study enrolled HIV-1 seropositive patients attending the HIV/AIDS Center, University Hospital for Infectious and Tropical Diseases, Clinical Centre of Serbia, Belgrade. Blood samples from 155 patients were collected from 2008 to 2013, whereas 162 HIV-1 sequences from Serbian isolates deposited in the NCBI database, dating from 1997 to 2007, were also included in the study. Nested-PCR (Polymerase chain reaction) method was used for amplification of the *pol* and *env* gene of HIV-1. All positive PCR products were directly sequenced and further analyzed by different phylogenetic and other bioinformatics approach. HIV-1 subtypes and circulating recombinant forms (CRF) were identified both by REGA subtyping tool and phylogenetic analysis of the *pol* gene sequences. Overall, 304 sequences were analysed with different phylogenetic software packages, depending on specific objective. Phylogenetic trees reconstruction was performed using different methods including Bayesian inference of phylogeny as well as dedicated softwares such as MEGA, PAUP and MrBayes. All analyses were performed under nucleotide substitution model that was chosen based on the likelihood scores obtained by jModeltest. In order to

SUMMARY

identify transmission clusters, defined as viral lineages derived from the same variant in the Serbian population, we applied different series of rigorous criteria sets, based on genetic distance and bootstrap support. A different approach, based on Bayesian phylogenetic inference, has been employed here to elucidate the ancestry of HIV-1 clades. A bioinformatics approach was used to estimate the duration of infection by calculating the fraction of ambiguous nucleotides in the sequence as a delimiter for more recent (less than 1 year) versus chronic infection (longer than 1 year). By analyzing complete dataset of subtype B sequences, prevalence of amino acid (aa) substitutions at 245 codon, and association with duration of infection and phylogenetic clustering, were investigated.

Results

Results of this study showed that among HIV-1 infected patients in Serbia subtype B predominated 90.8% (129/142), while the prevalence of non B subtypes was 9.2% (13/142). Phylogenetic analyses, that included 304 viral sequences, revealed a number of transmission clusters that accomplished all predefined criteria sets, along with one large transmission network. All sequences within transmission clusters and transmission network were identified as subtype B sequences. In total, 42.2% (116/275) of viral *pol* gene sequences were found within local phylogenetic clusters, while 57.8% (159/275) were found intermixed in the phylogenetic tree with sequences sampled Europe and America, indicative of multiple subtype B introductions. The majority of clustering sequences 82.7% (96/116) were from male patients living in Belgrade that predominantly reported MSM (men who have sex with men) as transmission category. The tMRCA (time of the Most Recent Common Ancestor) inferred for the local subtype B transmission network composed of 45 sequences was in 1994 (95% Higher Posterior Density (HPD): 1982–2000). Estimated temporal origin for the largest local subtype B, MSM associated

transmission cluster composed of 13 sequences was much more recent, 2004 (95% HPD: 2002–2006). The early nineties were found to be temporal origin of subtype G in Serbia, 1991 (95% HPD: 1979–2000). Results of the phylogenetic exploration of one particular transmission cluster in a forensic context have shown query samples to form a well-supported transmission chain, in support of the *a priori* hypothesis of their epidemiological linkage. However, in spite of the cluster topology with paraphyly of subject 1 sequences to those of subjects 2 and 3, this finding is not sufficient to unambiguously prove the transmission event and its direction. Based on the threshold of 0.47% ambiguous bases per sequence a total of 55.1% of samples (114/207) were classified as a recent infection, of duration of less than 1 year, whereas among subtype B samples this percentage was 54% (58/180). The predominant aa at RT codon 245 was the wild type valine (V), found in 61% (168/275), hence 36.7% (101/275) contained mutation at this position. The most common substitution at RT codon 245 was methionine (M) 22.9% (63/275), followed by glutamic acid (E) 7.2% (20/275), glutamine (Q) 5.4% (15/275) and others. Very high prevalence 93.6% (41/45) of aa substitutions at the investigated position was found among sequences from transmission network.

Conclusions

Results obtained in this research, enable drawing a more comprehensive and dynamic picture of the HIV-1 epidemic Serbia. HIV epidemics in Serbia continues to be dominated with subtype B, but with changes in distribution of non B subtypes over time, the emergence of new non B subtypes and increased genetic diversity among them. In this study we characterized a chain of ongoing subtype B HIV-1 transmission spanning the period of 16 years. Results of this research showed that the most encountered risk factor for infection with subtype B virus was the MSM transmission. Furthermore, transmission clusters were highly associated with MSM rather than other transmission risks. Our

SUMMARY

data support the need for increased public health interventions targeting MSM. This research indicated that local epidemic spread within transmission networks and outside the firstly implicated IDUs community dating from the beginning of the nineties, while epidemic spread of HIV subtype B among MSMs represents the most recent HIV-1 epidemic in Serbia. Based on transmission clusters analyses forensic application of phylogenetics was also explored. Estimation of the duration of infection, based on the ambiguous nucleotides sequence content, suggested an increasing proportion of recent infections, significantly higher in the second half of the study period. Equal prevalence of RT 245 substitutions in samples from recent and chronic HIV infection, together with high prevalence of this polymorphism in sequences within transmission network, might suggest early fixation of an HLA induced selective imprint, during intra-host viral evolution. Significantly higher prevalence of RT 245 substitution, compared to the preliminary reports of the HLA-B*57-01 allele frequency in HIV infected population in Serbia, was found.

Keywords: HIV, HIV-1 Serbian epidemic, subtypes, genetic diversity, phylogenetic analyses, transmission clusters, molecular footprints, bioinformatic analyses

Scientific field: Molecular Medicine / Virology

REZIME

Filogenetska analiza i molekularna karakterizacija virusa humane imunodeficijencije u Srbiji**Uvod**

Virus humane imunodeficijencije (HIV) je retrovirus koji uzrokuje sindrom stečene imunodeficijencije. Od početka epidemije pre 35 godina, ovim virusom je inficirano više od 78 miliona ljudi a preko 30 miliona je umrlo. Visoka genetička varijabilnost i brza evolucija HIV-a su ključni uzroci opstanka i globalnog širenja epidemije. HIV je filogenetski klasifikovan u dva tipa: HIV-1 i HIV-2. Visoki diverzitet HIV-1 ogleda u postojanju četiri grupe (M, N, O, P) od kojih su virusi grupe M uzročnici globalne HIV-1 pandemije. Grupa M virusa je podeljena u više filogenetski različitih podtipova (A-D, F-H, J i K), podtipove (A1, A2, F1 i F2) i cirkulišuće rekombinantne forme. Distribucija podtipova u svetu je složena i dinamična sa regionalnim HIV-1 epidemijama unutar globalnog diverziteta. Molekularna filogenetska analiza, metod za rekonstrukciju evolutivnih odnosa između nukleotidnih sekvenci, je tehnika za proučavanje varijabilnosti virusa i dinamike transmisije unutar regionalnih populacija. Procenjuje se da kod blizu polovine inficiranih osoba HIV infekcija nije dijagnostikovana, zbog čega je identifikacija puteva transmisije izuzetno značajna u cilju javno zdravstvenog nadzora. U ovom istraživanju primenjene su savremene filogenetske metode u analizi HIV-1 sekvenci izolata iz Srbije u cilju karakterizacije molekularne epidemiologije i dinamike transmisije, što je ključno za bolje razumevanje karakteristika aktuelne HIV-1 epidemije u Srbiji.

Ciljevi

Ciljevi ovog istraživanja bili su određivanje aktuelne distribucije podtipova HIV-1 u Srbiji. Ciljevi su obuhvatili i analizu transmisionih lanaca kao i filogenetsko datiranje epidemije za najzastupljenije HIV-1 podtipove u Srbiji. Istraživanje je takođe obuhvatilo analizu učestalosti i svojstava izmena nukleotidne sekvence genoma koje su odraz trajanja HIV-1 infekcije i interakcije virusa i domaćina.

Materijal i metode

U istraživanje su uključeni HIV-1 seropozitivni pacijenti praćeni u odeljenju za HIV/AIDS, Klinike za infektivne i tropske bolesti, Kliničkog centra Srbije, u Beogradu. Uzorci krvi od 155 pacijenata uključenih u studiju uzimani su u periodu od 2008. do 2013. godine, a dodatno su u istraživanje uključene 162 sekvence sojeva iz Srbije deponovane u NCBI bazi podataka, koje su generisane u periodu od 1997. do 2007. godine. Sekvence *pol* i *env* gena umnožene su metodom reakcije lančane polimerizacije u dva kruga (engl. „nested polymerase chain reaction“-nested PCR). DNK sekvence svih pozitivnih PCR produkata su analizirane filogenetskim i drugim bioinformatičkim metodama. Identifikacija podtipova i cirkulišućih rekombinantnih formi izvršena je pomoću REGA alata za genotipizaciju kao i filogenetskom analizom *pol* sekvenci. Ukupno, 304 virusne sekvence analizirane su različitim filogenetskim softverskim paketima, u zavisnosti od postavljenog cilja. Konstrukcija filogenetskih stabala je urađena primenom različitih filogenetskih metoda, uključujući metode Bajesove statistike, korišćenjem kompjuterskih programa MEGA, Paup i MrBayes. Filogenetska analiza urađena je na osnovu nukleotidnog substitucionog modela izabranog na osnovu skora verovatnoće pomoću Jmodeltest programa. U cilju identifikacije transmisionih lanaca, definisanih kao skup virusnih varijanti koje potiču od istog soja, primenjeni su

višestruki strogo definisani kriterijumi zasnovani na nukleotidnoj distanci i statističkoj podršci filogenetskom grupisanju. Primenjene su različite analize koje se zasnivaju na Bajesovoj statistici u cilju filogenetskog datiranja epidemije za najzastupljenije HIV-1 podtipove u Srbiji. Bioinformatički pristup, zasnovan na analizi učestalosti pojave istovremenog prisustva različitih baza na istoj poziciji u genomu, poslužio je kao molekularni marker u proceni rane (kraće od godinu dana) ili hronične (duže od godinu dana) HIV-1 infekcije. Posebno je analiziran nukleotidni sastav i izmena kodona na poziciji 245 RT produkta *pol* gena HIV-1 kao i povezanost promena sa dužinom trajanja infekcije i filogenetskim grupisanjem.

Rezultati

Rezultati ovog istraživanja su pokazali da je među HIV-1 inficiranim osobama u Srbiji dominantan podtip B virusa, sa prevalencijom od 90,8% (129/142), dok je prevalencija drugih podtipova iznosila 9,2% (13/142). Filogenetskom analizom, u koju su uključene 304 virusne sekvence, identifikovano je prisustvo većeg broja transmisionih klastera i jedna transmisiona mreža. Sve sekvence u okviru transmisionih klastera i transmisione mreže identifikovane su kao podtip B. Ukupno, 42,2% (116/275) virusnih sekvenci *pol* gena podtipa B pokazalo je lokalno filogenetsko grupisanje, dok je 57,8% (159/275) sekvenci filogenetski pomešano sa izolatima iz Evrope i Amerike, što ukazuje na višestruko poreklo HIV-1 epidemije u Srbiji. Većina sekvenci unutar filogenetski grupisanih sekvenci 82,7% (96/116), poticala je od pacijenata iz Beograda, muškog pola, kojima je najverovatniji način izlaganja HIV infekciji bio seksualni kontakt (muškarci koji imaju seksualne odnose sa muškarcima – MSM). Analizom je procenjeno da najraniji zajednički predak transmisione mreže, sačinjene od 45 virusnih sekvenci, potiče iz 1994. godine (95% IP : 1982–2000.). Procenjeno poreklo najvećeg transmission klastera sačinjenog od 11 virusnih sekvenci, dominantno izolovanih iz pacijenata koji su prijavili homoseksualni odnos kao

rizik infekcije, je znatno kasnije 2004. godina (95% IP: 2002–2006.). Analiza porekla podtip G sekvenci izolata iz Srbije pokazala je da je naraniji predak poreklom iz 1991 (95% IP: 1979–2000.). Filogenetsko istraživanje jednog transmissionog klastera u kontekstu forenzičke analize pokazalo je da ispitivani uzorci formiraju jasno izdvojen transmissioni lanac, u prilog *a priori* hipotezi o njihovoj epidemiološkoj povezanosti. Međutim, uprkos utvrđenoj topologiji filogenetskog stabla i parafiliji sekvenci subjekta 1 u odnosu na subjekte 2 i 3, dobijeni rezultati ne mogu sa sigurnošću da dokažu neposrednu transmisiju virusa između dva subjekta, kao ni smer transmisije. Analiza učestalosti pojave istovremnog prisustva različitih baza na istoj poziciji u genomu, na osnovu usvojene granice od 0,47% za ukupno 54% ispitanih virusnih sekvenci podtipa B ukazivala je na trajanje infekcije kraće od godinu dana. Na poziciji 245 produkta RT gena u najvećem procentu 61% (168/275) identifikovana je aminokiselina divljeg soja virusa, aminokiselina valin (V), dok je 36.7% (101/275) na toj poziciji imalo aminokiselinsku izmenu. Najzastupljenija promena aminokiselina na poziciji 245 produkta RT gena bila je metionin (M) 22.9% (63/275), zatim glutaminska kiselina (E) 7.7% (20/275), glutamin (Q) 5.5% (15/275), i ostale. Veoma visoka prevalenca, 93.6% (41/45), promene aminokiseline na ispitivanoj poziciji pronađena je u okviru sekvenci koje su obuhvađene transmissionom mrežom, dok je jednaka prevalence ove substitucije pronađena među sekvencama rane i hronične infekcije.

Zaključak

Dobijeni rezultati doprinose stvaranju detaljne i jasne slike kompleksne HIV-1 epidemije u Srbiji. HIV-1 epidemijom u Srbiji i dalje dominira podtip B virusa, ali sa promenom distribucije drugih podtipova kroz godine, pojavom novih podtipova i povećanjem diverziteta među njima. U ovom istraživanju okarakterisani su lanci transmisije podtip B virusa u Srbiji, dominantno povezani sa seksulanom transmisijom MSM kontaktom. Pokazano je da je

širenje HIV-1 epidemije unutar lokalnih transmisionih mreža počelo početkom devedesetih godina, dok je širenje među pacijentima mlađeg uzasta, iz Beograda, sa MSM kontaktom kao rizikom za transmisiju virusa počelo znatno kasnije i predstavlja najskoriju epidemiju. Na slučaju jednog transmisionog klastera ispitana je forenzička primena filogenetske analize. Analiza dužine trajanja infekcije, pokazala je statistički značajno veću zastupljenost rane infekcije u periodu od 2008. do 2013. godine. Ista prevalenca substitucije na kodonu 245 među sekvencama rane i hronične infekcije, uz visoku prevalencu iste substitucije u okviru sekvenci transmisione mreže, ukazuje na ranu fiksaciju ove promene nastale HLA selektivnim pritiskom. Pronađena je značajno viša prevalenca substitucije na kodonu 245 u odnosu na preliminarne rezultate frekvence alela HLA B* 57-01 među HIV-1 inficiranim pacijentima iz Srbije.

Ključne reči: HIV, HIV-1 epidemija u Srbiji, podtipovi, genetički diverzitet, filogenetska analiza, transmisioni klasteri, molekularni markeri, bioinformatička analiza

Naučna disciplina: Molekularna medicina / Virusologija

TABLE OF CONTENTS

1. Introduction.....	1
1.1. Overview of HIV infection and AIDS pandemic.....	2
1.1.1. Discovery of HIV.....	2
1.1.2. Current pandemic of HIV infection and AIDS.....	4
1.1.3. HIV-1 epidemic in Serbia.....	10
1.2. HIV-1 particle and genome organisation.....	11
1.2.1 The morphology of HIV particle.....	12
1.2.2 Organisation of HIV-1 genome.....	13
1.3. Life cycle of HIV.....	19
1.4. Phases of HIV disease progression.....	23
1.5. Diversity of HIV.....	25
1.5.1. Molecular footprints on the HIV genome	27
1.6. HIV classification	29
1.6.1. HIV types, groups and subtypes.....	29
1.6.2. HIV recombinants.....	29
1.6.3. Distribution of HIV-1 subtypes	
1.7. Analyses of HIV using phylogenetics.....	37
1.6.1 Analyses of HIV origin using phylogenetics.....	39
1.6.2. Analyses of HIV transmission chains using phylogenetics.....	46
2. The Aims of the study.....	49
3. Materials and Methods.....	51
3.1. Study design and ethical approval.....	52
3.2. Study subjects and sample collection	52
3.3. RNA extraction from plasma samples.....	54
3.4. DNA extraction from PBMCs.....	55
3.5. Nested polymerase chain reaction.....	55
3.5.1. End point limited dilution polymerase chain reaction (EPLD-PCR).....	58
3.6. Agarose gel electrophoresis.....	59
3.7. Cycle sequencing reaction.....	60
3.7.1. Purification of PCR products.....	60
3.7.2. Chain termination sequencing reaction.....	61
3.8. Sequence datasets.....	62
3.9. Phylogenetic analyses.....	64
3.9.1. HIV subtyping.....	68

3.9.2. . Phylogenetic analyses of transmission clusters.....	69
3.9.3. Estimating time of the most recent common ancestor.....	70
3.10. Molecular footprint analyses.....	75
3.11. Statitical analyses.....	76
4. Results	77
4.1. Study population.....	78
4.2. HIV subtyping.....	82
4.3. Phylogenetic identification of transmission clusters.....	84
4.3.1. Forensic application of phylogenetic analyses.....	87
4.3.2. Paraphyletic relation of qury sequences.....	88
4.4. Timing the origin of the main clade.....	94
4.5. The prevalence of molecular footprints on HIV-1 sequenc and their association with duration of infection.....	104
4.5.1. Nucleotide changes at 245 codon of HIV-1 RT gene sequences.....	105
5. Discussion	106
6. Conclusions	122
7. References	126
List of abbreviations and acronyms	154
Appendix I	156
Appendix II	160
Biography	163

CHAPTER 1. INTRODUCTION

1.1 OVERVIEW OF HIV INFECTION AND AIDS PANDEMIC

1.1.1 DISCOVERY OF HIV

Symptoms of Acquired Immunodeficiency Syndrome (AIDS) were first described in 1981 as unusually high occurrence of pneumonia caused by *Pneumocystis carinii* (now renamed as *P. jirovecii*) as reported by Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, USA. At the beginning, all disease cases were diagnosed among previously healthy men who had sex with men (MSM), living in New York and Los Angeles, USA (Galo, 2006; Gottlieb et al., 1981, Masur et al., 1981). Soon after, aggressive forms of Kaposi sarcoma, an otherwise rarely seen opportunistic tumor caused by Human Herpesvirus 8 (HHV-8), was diagnosed in eight young men in New York (Friedman-Kien et al., 1981; Hymes et al., 1981). Both of these conditions were known to occur only in severely immunocompromised patients, and none of the initial cases had any known cause of immunodeficiency. Soon after these first reports, a number of further similar cases with similar symptoms were discovered leading to conclusion that it was a new undefined disease associated with the breakdown of the immune system. This unusual clinical finding of Kaposi's sarcoma in a population much exposed to sexually transmissible diseases suggests that such exposure may play a role in its pathogenesis (Hyemes et al., 1981). Initially referred to as gay related immune deficiency (GRID), the clustering in homosexual men with the promiscuous life stile and the history of numerous sexually transmitted infections, the clustering in homosexual men and the association with high numbers of sexual partners and previous sexually transmitted infections, alerted epidemiologists to the possibility of a sexually transmitted agent being responsible (Greene, 2007). However, soon after, the disease was recognized in other population groups as

well. This included females closely associated with intravenous drug abuse either by shared tools or sexual contact with an drug abuser (Masur et al., 1982), their offsprings (Oleske et al., 1983), hemophiliacs (Bloom, 1984) and blood transfusion recipients (Peterman et al., 1984).

In 1983, scientists around Luc Montagnier and Françoise Barré-Sinoussi and their research group of at the Institute Pasteur in Paris, were the first who isolated the causative virus of AIDS, a discovery for which they were awarded with the Nobel Prize for Physiology or Medicine twenty-five years later (Barré-Sinoussi et al., 1983). The first isolated HIV, in 1983, was the prototype of what was later designated as HIV type 1 (HIV-1) group M (HIV-M) and is the virus responsible for the current pandemic. Barre-Sinoussi and co-workers recovered a retrovirus from the lymph node of an individual suffering from lymphadenopathy syndrome (LAS), an AIDS associated condition. AIDS was initially described as the appearance of very rare, dramatic and life-threatening opportunistic infections and/or opportunistic tumors, due to severe depletion of the immune system (Ammamm et al., 1983). The scientists at the Pasteur Institute in France reported the discovery of a new retrovirus called Lymphadenopathy-Associated Virus (LAV) that could be the cause of AIDS (Barré-Sinoussi et al., 1983; Ratner et al., 1985a; Ratner et al., 1985b). Almost at the same time, the National Cancer Institute announced that they had found the cause of AIDS, the retrovirus HTLV-III. In a joint conference with the Pasteur Institute they announced that LAV and HTLV-III are identical and the likely cause of AIDS. (Marx, 1984). Two years later, the International Committee on the Taxonomy of Viruses announced that the virus that causes AIDS will officially be called Human Immunodeficiency Virus (HIV) instead of HTLV-III/LAV (Case, 1986). Within a short period of time the spread of the HIV and AIDS has reached a pandemic form.

1.1.2 CURRENT PANDEMIC OF HIV INFECTION AND AIDS

At the time of HIV discovery, it was hard to imagine the proportions that the AIDS epidemic would grow to. Over the past 35 years, HIV/AIDS has evolved into a highly heterogeneous epidemic structured in multiple sub-epidemics, each of which are influenced by biological, behavioral, and cultural factors (Tebit and Arts, 2011; Beyrer et al., 2012; Brenner and Wainberg, 2013). The pandemic of AIDS has become one of the most important global health threats due to its mortality and morbidity. The incidence of HIV-1 infections reveals the need for multisectorial efforts to combat and reduce the number of new infections, increase access to healthcare services, and guarantee access to antiretroviral therapy for the general population (UNAIDS, 2015; UNAIDS, 2013). It is estimated that 36.7 million people were living with HIV in 2015 and more than half of them were in Sub-Saharan Africa, mostly infected through unprotected sexual intercourse (**Table 1.**) (<http://www.unaids.org/>). In 2015 there were 2.1 million [1.8 million–2.4 million] new HIV infections worldwide, and more than 90 thousand new infections in Western and Central Europe only (<http://www.who.int/gho/hiv/en/>).

However, since the beginning of the epidemic enormous gains have been made and even more can be achieved in the coming years. Development of highly active antiretroviral therapy (HAART) has made HIV infection a treatable chronic disease. Early HAART is associated with a reduced latent viral reservoir, reduced viral DNA, and normalisation of some immune marker. In just the last two years the number of people living with HIV on antiretroviral therapy has increased by about a third, while since 2010 this number increased more than two times, reaching 17 million (**Table 2.**).

No single prevention is effective enough on its own, and many interventions are necessary to control the global HIV epidemic. Prevention of sexual HIV transmission has been a priority since the beginning of the epidemic. However, concerns have shifted from access and availability of HIV treatment programs to issues of late presentations, quality of care, HIV drug resistance and HIV treatment failure (Apisarnthanarak et al., 2008; Weidle et al., 2002; Kitkungvan et al., 2008). In particular, treatment failure is a huge problem as it leads to transmission of HIV resistant virus, increase in treatment complexity and cost, worsening morbidity and mortality and ultimately, failure of the HIV treatment program (Gilks et al., 2006; Rajasekaran et al., 2007; Robbins et al., 2007).

New insights into the mechanisms of viral latency and the significance of reservoirs of infection might eventually lead to a cure. The importance of immune activation in the pathogenesis of non-AIDS clinical events (major causes of morbidity and mortality in people on antiretroviral therapy) is receiving increased recognition. Breakthroughs in the prevention of HIV important to public health include male medical circumcision, antiretrovirals to prevent mother-to-child transmission, antiretroviral therapy in people with HIV to prevent transmission, and antiretrovirals for pre-exposure prophylaxis. Research into other prevention interventions, notably vaccines and vaginal microbicides, is in progress.

Table 1. Latest estimates of HIV epidemic, globally and by region, 2010 and 2015. Data from: GARPR 2016; UNAIDS 2016 estimates

	People living with HIV (all ages)		New HIV infection (all ages)	
	2010	2015	2010	2015
Global	33.3 million [30.9-36.1]	36.7 million [34.0 -39.8]	2.2 million [2.0-2.5]	2.1 million [1.8 -2.4]
Asia and Pacific	4.7 million [4.1 -5.5]	5.1 million [4.4 -5.9]	310 000 [270 000-360 000]	300 000 [240 000-380 000]
Easter and Southern Africa	17.2 million [16.1 -18.5]	19.0 million [17.7 -20.5]	1.1 million [1.0 -1.2]	960 000 [830 000-1.1]
Eastern Europe and central Asia	1.0 million [950 000-1.1]	1.5 million [1.4 -1.7]	120 000 [110 000-130 000]	190 000 [170 000-200 000]
Latin America and the Caribbean	1.8 million [1.5 -2.1]	2.0 million [1.7 -2.3]	100 000 86 000-120 000]	100 000 [86 000-120 000]
Middle East and North Africa	190 000 [150 000-240 000]	230 000 [160 000-330 000]	20 000 [15 000-29 000]	21 000 [12 000-37 000]
Western and central Africa	6.3 million [5.2 -7.7]	6.5 million [5.3 -7.8]	450 000 [350 000-560 000]	410 000 [310 000-530 000]
Western and central Europe and North America	2.1 million [1.9 -2.3]	2.4 million [2.2 -2.7]	92 000 [89 000-97 000]	91 000 [89 000-97 000]

Table 2. Latest estimates of number of HIV infected people on HAART and AIDS related deaths, globally and by region, 2010 and 2015.

Data from: GARPR 2016; UNAIDS 2016 estimates

	People living with HIV on antiretroviral treatment (all ages)		AIDS-related deaths (all ages)	
	2010	2015	2010	2015
Global	7 501 100	17 025 900	1.5 million [1.3 million–1.7 million]	1.1 million [940 000–1.3 million]
Asia and Pacific	907 600	2 071 900	240 000 [200 000–270 000]	180 000 [150 000–220 000]
Easter and Southern Africa	4 087 500	10 252 400	760 000 [670 000–860 000]	470 000 [390 000–560 000]
Eastern Europe and central Asia	112 100	321 800	38 000 [33 000–45 000]	47 000 [39 000–55 000]
Latin America and the Caribbean	568 400	1 091 900	60 000 [51 000–70 000]	50 000 [41 000–59 000]
Middle East and North Africa	13 600	38 200	9500 [7400–12 000]	12 000 [8700–16 000]
Western and central Africa	905 700	1 830 700	370 000 [290 000–470 000]	330 000 [250 000–430 000]
Western and central Europe and North America	906 200	1 418 900	29 000 [27 000–31 000]	22 000 [20 000–24 000]

In the world's most affected region, eastern and southern Africa, the number of people on treatment has more than doubled since 2010. AIDS related deaths in this region have decreased by 36% since 2010. At the same time the number of AIDS deaths is also declining: there were 1.1 (1.4–1.9) million AIDS deaths in 2015 while there were 2.3 (2.1–2.6) million in 2005 and 1.6 million in 2010 (**Table 2**). However, the prevalence of HIV varies considerably between different regions over the world. Sub-Saharan Africa is still the most seriously affected region by the HIV pandemic, where the epidemic has become a growing human and economic catastrophe (**Figure 1**). In some of these countries (e.g. South Africa, Botswana and Swaziland), almost every fifth adult is infected with HIV. In contrast, most countries in Western and Central Europe have prevalence rates of about 0.1% (UNAIDS, 2015). Complex and varied social, structural and economic dynamics within countries account for the uneven geographical distribution of HIV. Furthermore, important difference between regions that account for imbalances lies in the different primary modes of transmission (**Figure 1**). Unlike the sub-Saharan Africa epidemic, in which adolescent girls and young women accounted for 25% of new HIV infections, while women accounted for 56% of new HIV infections among adults, the epidemic affecting western hemisfera is concentrated in high risk groups including people who inject drugs (PWID), men who have sex with men (MSM) and commercial sex workers, along with a clear trend towards increased transmission among MSM while the rates of new infections in other high-risk groups decreased in the last years (de Oliveira et al., 2017). Consequently Africa has also experienced large numbers of infections acquired vertically and in 2007 2.2 million children were estimated to be living with HIV in sub Saharan Africa. Vice versa, in North America and in Western and Central Europe men outnumber women in both HIV prevalence and incidence by more than 2:1.

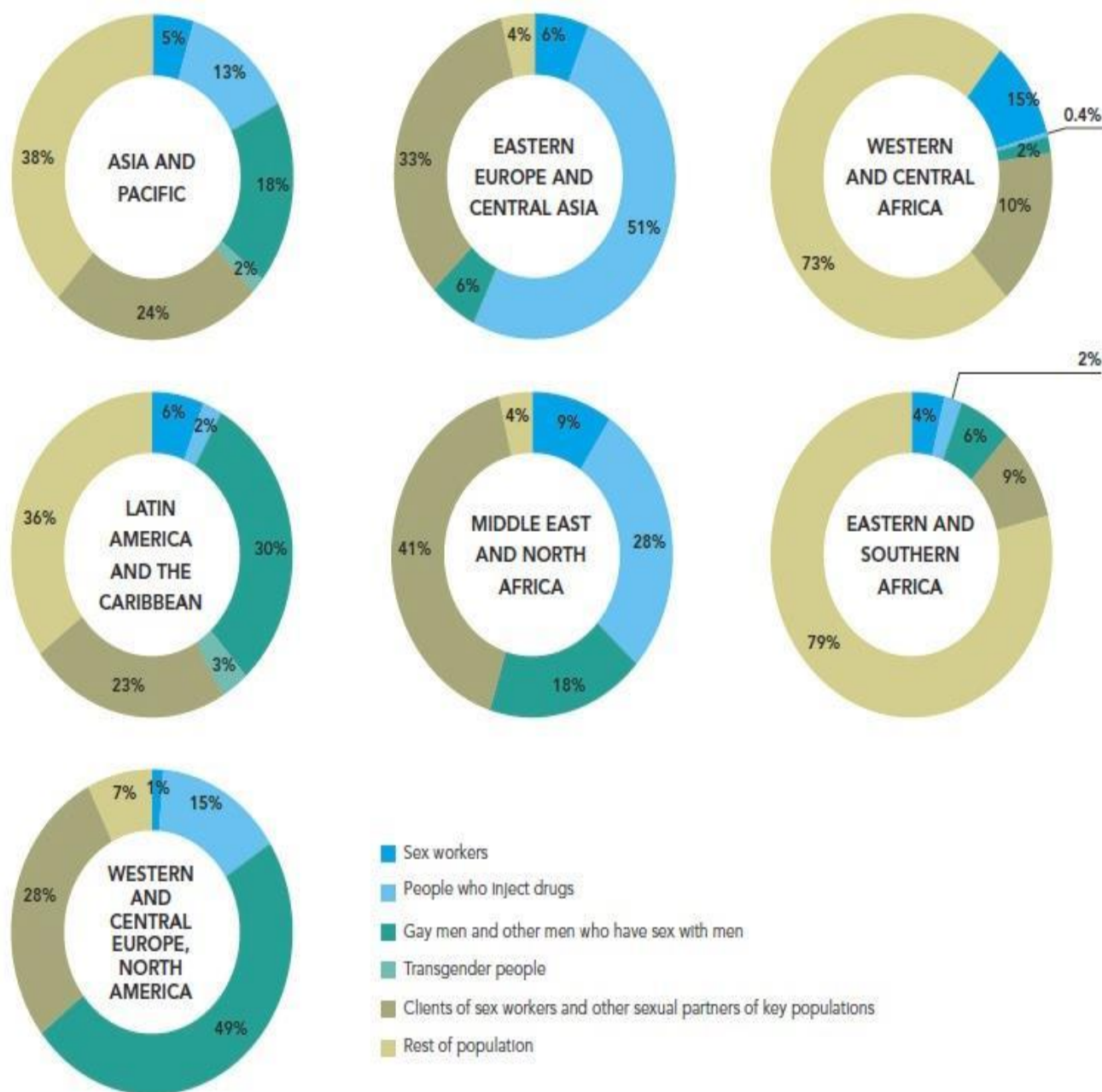


Figure 1. Distribution of new HIV infections among population groups, by region, 2015. Data from UNAIDS special analysis, 2016.

1.1.3 HIV EPIDEMIC IN SERBIA

In Serbia, the duration of HIV epidemic is similar to the one in Western European countries, with first cases registered in 1985. It was initially recognized among people who inject drugs (PWID) and this transmission route was the most prevalent over the coming years. By the end of the nineties, intravenous drug use was still the most prevalent risk among cumulative number of HIV/AIDS cases, around 48.9%, followed by sexual transmission, 33% (Stanojevic et al., 2002).

According to the latest data released by the Institute for Public Health of Serbia “Dr Milan Jovanovic Batut” (the National Institution that has the mandate for surveillance and monitoring and evaluation of the national response on HIV/AIDS) the cumulative number of HIV-infected people reported until the end of 2015 was 3312, of whom 1788 developed AIDS while 1192 died (http://www.batut.org.rs/index.php?category_id=17). In 2015, 178 newly diagnosed HIV cases, 45 AIDS cases and 15 AIDS-related deaths were reported, as well as 4 non-HIV related deaths among people infected with HIV. The decreasing trend of AIDS cases and AIDS-related deaths in the last decade is the result of the introduction of HAART which is available for all people living with HIV (PLHIV) in need and fully covered by Republican Health Insurance Fund since 1997. The majority of HIV cases diagnosed in the period 2002-2015 were from the capital city Belgrade (913 cases or 54%), and Vojvodina region (341 cases or 20%), in both of which greatest HIV testing rates were achieved. The majority of HIV infected in the past presented with clinical AIDS (above 70%), but that trend has been changing recently. Moreover, in the period 2002-2015 there is a clear increasing trend of asymptomatic HIV infected persons at the time of presentation (117 cases or 66% of all newly diagnosed HIV cases in 2015 versus 14% in 2002) along with a decreasing trend of newly diagnosed AIDS cases (20% in 2015 versus 48% in 2002). However, Serbia is

still a country with a high prevalence of “late presenters” (CD4 count <350 cells/ μ l at baseline). According to current data, out of all newly diagnosed HIV infections in 2015 with reported baseline CD4 counts 56% were “late presenters” while 47% already had AIDS.

1.2. HIV-1 PARTICLE AND GENOME ORGANISATION

HIV is a member of the *Retroviridae* family, *Orthoretrovirinae* subfamily and *Lentivirus* genus (Sonigo et al., 1985, <http://ictvonline.org/virusTaxonomy.asp>). Retroviruses transmit their genomes as a single stranded linear positive sense RNA. Upon cellular infection, the positive sense RNA undergoes a process of reverse transcription creating double stranded DNA from which transcription occurs and that is also capable of integrating into the host genome. The virus therefore encodes and carries within the virion an enzyme called RNA dependent DNA polymerase or Reverse transcriptase which will transcribe the RNA genome into a double-stranded DNA intermediate.

As a group, lentiviruses have a number of common features. This group of viruses can infect non dividing cells (such as macrophages) and they express a number of regulatory genes in addition to the basic proteins required for the maturation of new infectious virions. Furthermore, this viruses cause slowly progressive infection which is a function both of their ability to integrate into the host chromosome and of their ability to evade host immunity.

1.2.1 THE MORPHOLOGY OF HIV PARTICLE

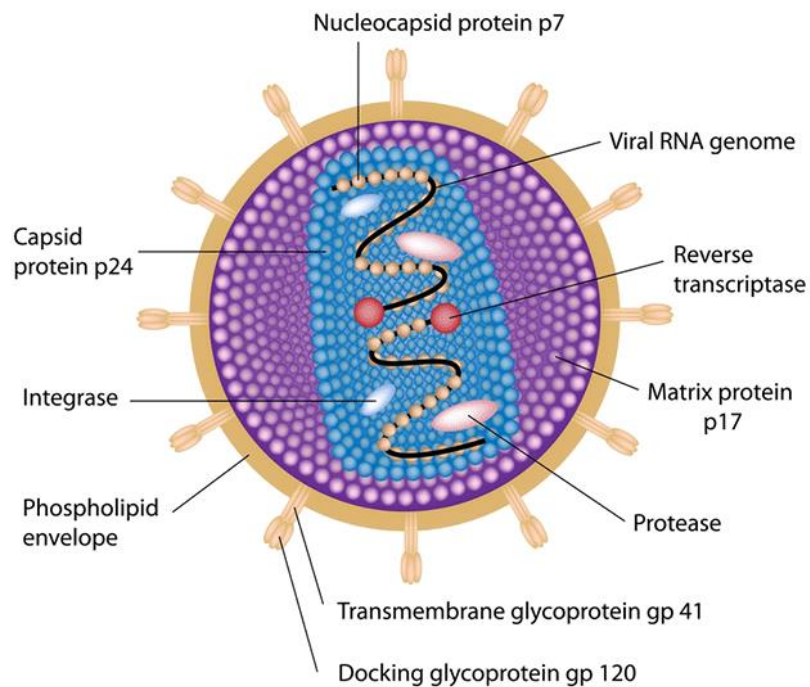
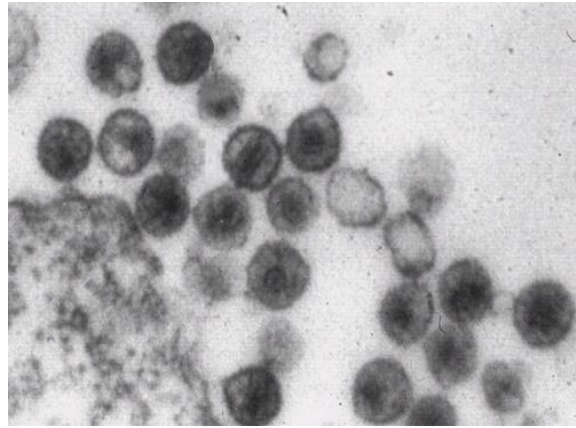


Figure 2. (a) Electron micrograph of HIV particle. Modified from <http://virology-online.com/viruses/HIV.htm> (b) Schematic representation of a single HIV virion. Modified from <https://ceufast.com/course/aids-hiv-4hr>

HIV is spherical, membrane-enveloped, pleomorphic virions, 100-120 μm in diameter, which contain two copies of single-stranded, positive-sense RNA genome (**Figure 2**). The lipid bilayer contains several cellular membrane proteins, including major histocompatibility (MHC) antigens derived from the host cell (Arthur *et al.*, 1992). Interior to the viral envelope, the virus matrix protein (MA) surrounds the capsid (CA).

The main constituents of the HIV-1 particle are Gag polyprotein (described below), which makes around 50% of the entire virion mass and the viral membrane lipids, which account for around 30% of virion mass (Carlson *et al.*, 2008). Other viral and cellular proteins together contribute an additional 20%, whereas the genomic RNA and other small RNAs amount to 2.5% of virion mass. Gag, Gag-Pro-Pol, Env, the two copies of genomic RNA, the tRNA primer, and the lipid envelope are all essential for viral replication, whereas the relevance of virion incorporation of other cellular and viral accessory proteins, small RNA molecules, and specific lipids is generally less well understood. Each viral particle contains two unspliced positive-oriented single-stranded ribonucleic acid (RNA) molecules. The RNA genome encodes three structural polyproteins (Gag, Pol and Env) and six regulatory or accessory proteins (Tat, Rev, Vif, Vpr, Vpu and Nef) flanked by noncoding long terminal repeats (LTR) at both 5' and 3' end, as described below.

1.2.2 ORGANISATION OF HIV-1 GENOME

Each single stranded RNA has a length of about 9.7kb and is bound to the nucleocapsid NC (p7) protein and surrounded by enzymes important for viral replication and maturation such as protease (PR), reverse transcriptase (RT) and integrase (IN), encoded by the viral *pol* gene. Accessory proteins (Nef, Vif, Vpr) can also be found in the viral ribonucleoprotein core structure. A

number of host molecules have also been found in the virus particle, although their importance is still unclear.

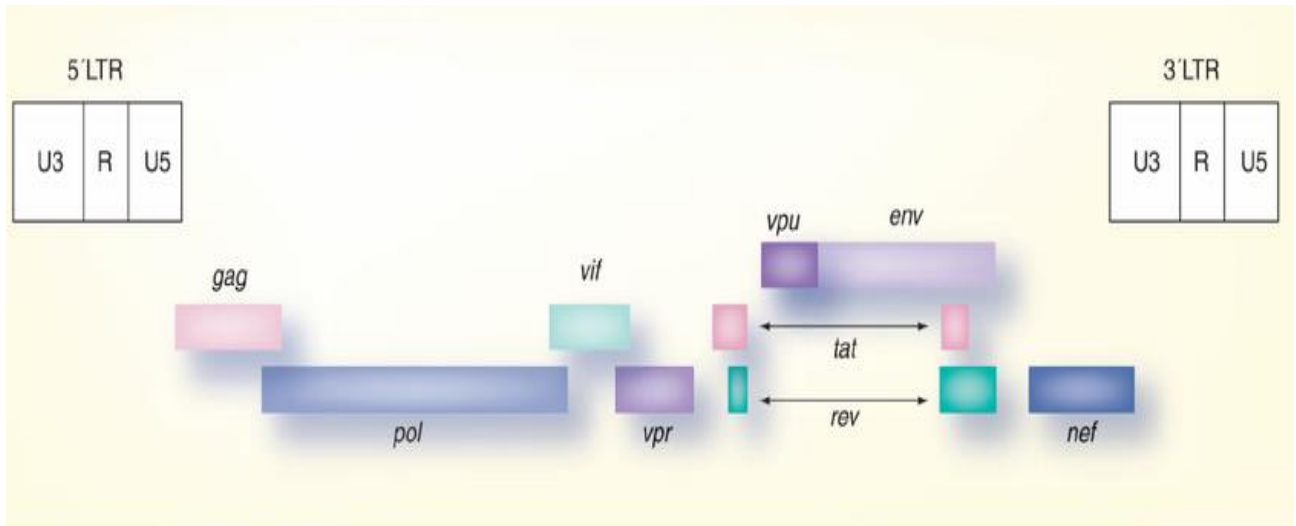


Figure 3. Diagram of the genome organization of HIV-1. Modified from Freed (2004)

Fifteen viral proteins are coded by nine genes in the HIV genome (**Figure 3; Table 3**).

The *gag* (group-specific antigen) gene provides the physical infrastructure of the virus. The *gag* gene is the precursor of four viral structural proteins: p24 (the viral capsid), p6 and p7 (the nucleocapsid proteins), and p17 (the viral matrix). The p24 CA protein is the main component of viral core containing the viral RNA genomes

The HIV *pol* gene encodes the viral enzymes protease (PR), reverse transcriptase (RT) and integrase (IN). The first viral enzyme encoded by the *pol* gene is Protease. Pol is always produced as a Gag-Pol fusion protein (protein

precursor Pr 160), even though the *pol* coding region partially overlaps and is in the -1 reading frame with respect to *gag* (Jacks et al., 1988; Ratner et al., 1985)

The Pr160 polyprotein is formed by ribosomal frameshifting, the process whereby ribosomes change the open reading frame to allow for alternative translation of mRNA (Hung et al., 1998). During viral maturation, protease cleaves the Gag-Pol polyproteins to produce viral enzymes (Protease, Reverse transcriptase, Integrase). Moreover, the activity of protease depends on the concentration of Gag-Pol polyproteins and the rate of protease-mediated autoprocessing is modulated by the adjacent p6 sequence. RT and PR proteins are encoded between the base 2258 in the *gag* gene and the base 3872 in the *pol* gene.

RT is another important enzyme encoded by the *pol* gene Reverse transcriptase heterodimer acting as an RNA-DNA-dependent DNA polymerase is required to produce viral dsDNA during reverse transcription. and is found in all retroviruses (Rodgers et al., 1995; Huang et al., 1998). The mature form of HIV-1 RT is a heterodimer that is composed of two related subunits: the larger, p66, is 560 amino acids long; the smaller, p51, contains the first 440 amino acids of p66 (Lightfoote et al. 1986). The p66 subunit consists of two domains: polymerase and RNase H; in the mature HIV-1 RT heterodimer, p66 contains the active sites for the two enzymatic activities of RT. Integrase that is a multidomain enzyme catalyzes two major reactions (3'-processing and strand transfer reactions) to insert the linear, double-stranded viral DNA into human chromosomes. In the mature viral particles, integrase is cleaved from the Gag-Pol polyprotein by viral Protease.

The *env* (envelope) gene codes for glycoprotein gp160 which acts as a precursor to the glycoproteins gp 120 and gp 41. On the virion surface, there are less than 30 envelope spikes consisting of three molecules of gp120 and gp41 each, connected by noncovalent interactions. Viral antireceptor, env protein gp120, consists of five constant (C1 to C5) and five variable regions (V1 to V5).

The variable regions are mostly found within regions encoding disulphide-constrained loops, exposed to the surface and to the host immune system (Leonard et al., 1990). The V3 region plays an important role in determining cellular tropism, allowing the virus to either use chemokine receptor type 5 (CCR5) or chemokine receptor type 4 (CXCR4) as its main co-receptor (Briggs et al., 2003). The transmembrane glycoprotein gp 41 is also encoded by the *env* gene. gp 41 contains a glycine-rich region which is essential for the membrane fusion activity.

HIV-1 Tat protein (*trans* acting activator of transcription), is one of the essential proteins, which directly enhances HIV-1 replication through interaction with HIV-1 long terminal repeat (LTR) promoter (Jeang et al., 1999). Tat is therefore a promising target for developing HIV-1 vaccines and anti-HIV-1 drugs (Ensoli et al., 2006; Hamy et al., 2000). However, Tat undergoes continuous amino acid substitutions. As a consequence, the virus escapes from host immunity indicating that genetic diversity of Tat protein in major HIV-1 subtypes is required to be continuously monitored.

The HIV-1 protein Rev facilitates the nuclear export of intron-containing viral mRNAs by recognizing a structured RNA site, the Rev-response-element (RRE), contained in an intron. After translation, Rev enters the nucleus and binds the Rev response element (RRE), a ~350 nucleotide, highly structured element embedded in the *env* gene in unspliced and singly spliced viral RNA transcripts. Once in the cytoplasm, the complexes dissociate and unspliced and singly-spliced viral RNAs are packaged into nascent virions or translated into viral structural proteins and enzymes

The *vif*, *vpr*, *vpu*, and *nef* genes encode for accessory proteins, that are not necessary for viral propagation. However, experimental observations suggest that their role *in vivo* is very important in the life cycle of HIV.

Table 3. HIV-1 genes and associated proteins

CATEGORY	NAME	ASSOCIATED PROTEIN	FUNCTION
structural	<i>gag</i>	p17, p24, p15, p7	early stages of viral replication; RNA targeting to plasma membrane; particle assembly; acts as viral cytokines
		gp120	binding to primary receptor CD4 and coreceptors (CCR5 and CXCR4)
	<i>env</i>	gp41	anchors the gp120/gp41 complex in the membrane, catalyze the membrane fusion reaction
enzymatic		protease (PR)	Gag/Pol cleavage and maturation
	<i>pol</i>	reverse transcriptase (RT)	reverse transcription
		integrase (IN)	DNA provirus integration
regulatory	<i>tat</i>	Tat	enhances HIV-1 replication
	<i>rev</i>	Rev	facilitates the nuclear export of viral mRNAs by recognizing
accessory	<i>vpu</i>	Vpu	CD4 downregulation in the Endoplasmic Reticulum extracellular release of virus particles
	<i>vif</i>	Vif	helps to counteract APOBEC3G
	<i>vpr</i>	Vpr	nuclear transport of the HIV-1 pre-integration complex (PIC)
	<i>nef</i>	Nef	establishment of high viral loads during infection and enhances viral pathogenesis

Viral infectivity factor (Vif) is an essential accessory protein for HIV-1, which critical role in replication was observed shortly after the discovery of HIV-1 (Sheehy et al., 2002). Vif protein helps to counteract Apolipoprotein B mRNA-editing enzyme-catalytic polypeptide-like 3G (APOBEC3G, A3G), a potent host restriction factor that inhibits HIV-1 replication. The APOBEC family of proteins act as editing enzymes and cause Cytosine to Uracil editing and lead to the accumulation of Guanine to Adenine mutations in the proviral sense cDNA strand (Malim, 2009). Vif also interacts with Gag polyprotein to modulate the Protease mediated proteolytic processing.

Viral protein R (Vpr) is an accessory protein which plays multiple functions to enhance viral expression in the non-dividing cells (e.g. macrophages). The protein is responsible for nuclear transport of the HIV-1 pre-integration complex (PIC) and plays an important role in the extracellular release of virus particles (Romani and Engelbrecht, 2009).

Viral protein U (Vpu) is a membrane-associated accessory protein with two major functions of CD4 downregulation in the Endoplasmic Reticulum. Furthermore this protein promotes the extracellular release of virus particles. Vpu is not incorporated in HIV particles thus is not found in the mature virus particle. Negative regulatory factor (Nef) is responsible for the establishment of high viral loads during infection and enhances viral pathogenesis, leading to faster disease progression (Kirchhoff et al., 2008).

1.3. LIFE CYCLE OF HIV

HIV can infect T-helper lymphocytes, macrophages and other cells of the same lineage, such as microglia in the brain, which express low levels of CD4 molecule. As with all members of *Retroviridae* family, the HIV replication cycle involves target cell recognition and invasion, reverse transcription, integration into the host genome and the construction and release of new virions. The primary cellular receptor for HIV entry is the CD4 molecule (Bour et al., 1995; Deng et al., 1996; Wu et al., 1996). However, the CD4 molecule is by itself not sufficient to allow HIV entry. In 1996, the chemokine receptors CCR5 and CXCR4 were found to act as coreceptors that mediated HIV infection of CD4+ immune cells (Bour et al., 1995; Deng et al., 1996; Dragic et al., 1996; Wu et al., 1996)

CD4 is a single chain class I membrane glycoprotein of the immunoglobulin superfamily and its extracellular domain is composed of 370 amino acids (aa); the amino acids at positions 40-55 are used by HIV for the binding process in the first stage of infection (Altfeld et al., 2003; Bour et al., 1995; Riminton, 2004; Sever, 1989;).

General features of the HIV replication cycle are shown in **Figure 4**. The early phase begins with the recognition of the target cell by the mature virion and involves all processes leading to integration of the genomic DNA into the chromosome of the host cell. The late phase begins with the regulated expression of the integrated proviral genome, and involves all processes up to virus budding and maturation. HIV entry involves a stepwise series of interactions with receptors that initiate conformational changes in the envelope glycoproteins. The virus attaches itself to the target cell by adsorbing its glycoproteins (the envelope gp120 protein on the surface of HIV) to two host-cell receptor proteins: the CD4 molecule receptor and CCR5 or CXCR4 co-

receptors (also known as CC chemokine receptor 5 and CXC chemokine receptor 4). The gp120 envelope protein is composed of inner and outer domains, named for their expected orientation in the oligomeric viral spike (Kwong et al., 2000; Rizzuto et al., 1998). The third variable region (V3) of the HIV-1 gp120 envelope glycoprotein is immunodominant and contains features essential for coreceptor binding. V3 typically consists of 35 amino acids (range 31 to 39) and plays a number of important biological roles (Hartley et al., 2005). Not only is it critical for coreceptor binding, but it also determines which coreceptor, CXCR4 or CCR5, will be used for entry (Hartley et al., 2005). The HIV envelope spike mediates binding to receptors and virus entry. The trimeric spike is composed of three gp120 exterior and three gp41 transmembrane envelope glycoproteins. CD4 binding to gp120 in the spike induces conformational changes that allow binding to a coreceptor, either CCR5 or CXCR4, which is required for viral entry. CCR5-using (R5) viruses represent the major transmissible HIV-1 strains, whereas CXCR4-using (X4) viruses tend to arise late in the course of disease. Binding to such a coreceptor induces further conformational change causing gp120 to move aside, exposing the trimeric gp41 structure. The gp41 protein facilitates the fusion of the viral envelope and the host-cell membrane. The hydrophobic core of gp41 embeds itself into the cell membrane and by coiling of certain domains, the viral membrane is virtually pulled towards the cell membrane. This fusion allows the release of HIV nucleocapsid into the target cell (Dev et al., 2016).

Upon release of HIV nucleocapsid containing HIV-RNA and viral enzymes reverse transcriptase (RT), integrase (IN) and protease (PR), into the host cell cytoplasmic compartment, RT converts the single-stranded viral RNA genome into double-stranded DNA (ds DNA). Immediately after viral entry, a series of processes take place to form the reverse transcriptase complex (RTC) in viral core (Klasse et al., 2012).

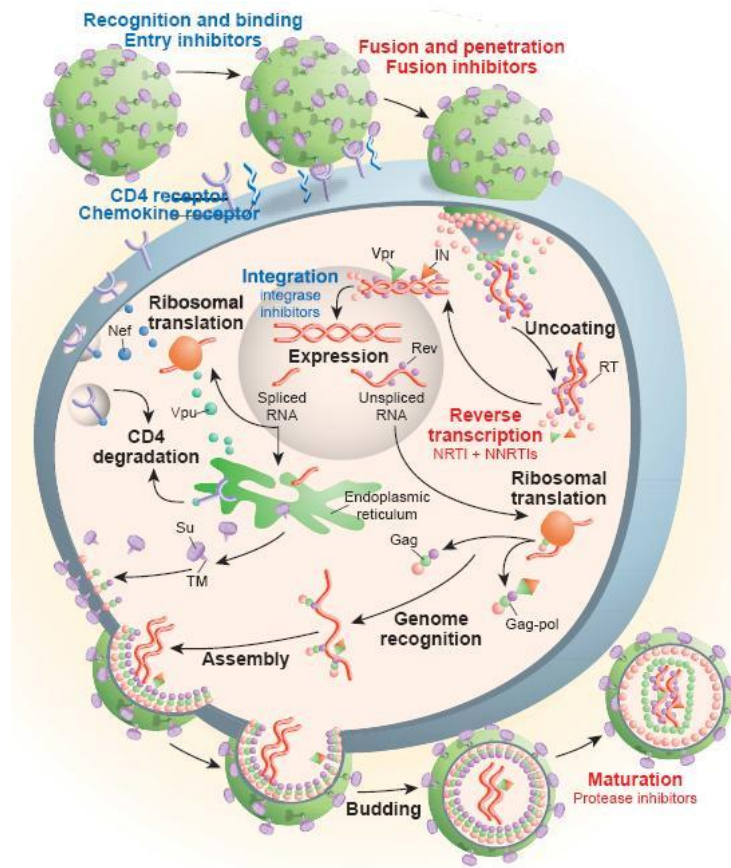


Figure 4. The HIV-1 life cycle.

Modified from Pomerantz et al., 2003.

The accessory protein Vif appears to be important during one or more of these early events, perhaps by facilitating the initial stages of reverse transcription. Recent reports suggested that reverse transcription takes place in the intact capsid core and is triggered by deoxyribonucleotides in the cytoplasm after viral entry (Hu et al., 2012; Lu et al., 2004; Burnett et al., 2007). During reverse transcription, the intact capsid core moves toward the nuclear pore on microtubules and RTC is turned into the pre-integration complex (PIC) (Hu et al., 2012). The dsDNA, together with other cellular and viral proteins, forms the PIC which is transported to the nucleus through nuclear pore complex (NPC) (Freed, 2001). Nuclear localization of the preintegration complex is directed by the accessory protein Vpr, which does not contain a nuclear localization signal but appears to function by connecting the preintegration complex to the cellular

nuclear import machinery (including importin- α and the nucleoporins (Fouchier et al., 1997; Freed et al., 1995). In the nucleus, the integrase being part of the PIC, assisted by cellular cofactors, catalyzes the insertion of the dsDNA into the host chromosome (Freed, 2001). At this stage, the virus is known as a provirus. Integration is an irreversible process that is fundamental for its replication and long-term persistence in the host, marking an important point of HIV infection (Bisgrove et al., 2005; Ciuffi et al., 2011). The provirus can then remain latent or be active, generating products for the generation of new virions.

After integration viral genes are transcribed along the host cell's genes. The late phase of the virus life cycle begins with the synthesis of unspliced and spliced mRNA transcripts, which are transported out of the nucleus for translation. Initially, short spliced RNA species that encode the regulatory proteins Tat, Rev and Nef are synthesized. Tat is an essential transcriptional activator that binds to a stemloop element of the nascent RNA transcript (TAR, for trans-activating response element) and recruits the cellular proteins cyclin T and cyclin-dependent protein kinase-9 (Cdk9; previously called TAK).

Ordinarily, unspliced cellular mRNAs are retained in the nucleus where they can be further processed or degraded. However, full length and singly spliced HIV mRNA transcripts that contain functional introns are needed in the cytoplasm for Gag and Gag-Pol synthesis and packaging, and their export is mediated by the essential HIV accessory protein Rev.

Unspliced, intron containing, viral RNA, assisted by Rev protein, is transported from nucleus to cytoplasm where it is packed into the new structures (Cochrane et al., 2014; Vercruyssen et al., 2013). HIV-1, like all retroviruses, selectively incorporates two copies of the capped and polyadenylated full-length RNA genome into the virion. The envelope polyprotein gp160 is transported to plasma membrane via the endoplasmic

reticulum and the Golgi complex where it is subjected to extensive N-glycosylation resulting in long mannose chains. It then undergoes trimerization and is cleaved by a host protease into its gp120 and gp41 subunits (Freed, 2001). Trimeric gp120-gp41 protein complexes are integrated into the host cell's membrane after expression in the ER. Here, it is essential that, assisted by Vpu in the ER and by Nef on the cell's surface, CD4 is degraded, preventing viral spikes immediate binding. The precursor proteins Gag and GagPol also associate at the plasma membrane of the host cell where the formation of new virus particles begins. During budding from the cell the newly constructed virions are enveloped by fragments of this spike-enriched cell membrane. Viral protease being part of the GagPol precursor protein gets activated when the respective domains of two GagPol precursors dimerize. PR then cleaves the Gag and Gag-Pol polyproteins into their subunits. After maturation, they are capable of infecting other target cells. Replication is estimated to take 1-2 days. HIV-1 buds at the plasma membrane of infected cells and the viral membrane is therefore derived from the cellular plasma membrane.

1.4. PHASES OF HIV-1 DISEASE PROGRESSION

After the initial cross species transmission from chimpanzee to humans, HIV continued to spread through human population via three main transmission routes: sexual contact, blood-to-blood contact (mostly intravenous drug use), and transmission from mother to child (vertical transmission). The probability of HIV transmission depends on the amount of the infectious virus particles present in the body fluid, mainly blood and genital fluid in the index patient and the extent of exposure (**Table 4.**).

Table 4. Risk contacts and the associated probability of infection. Data from <https://www.cdc.gov/hiv/risk/estimates/riskbehaviors.html>

Type of Exposure	Risk per 10,000 Exposures
Parenteral	
Blood Transfusion	9,25
Needle-Sharing During Injection Drug Use	63
Percutaneous (Needle-Stick)	23
Sexual	
Receptive Anal Intercourse	138
Insertive Anal Intercourse	11
Receptive Penile-Vaginal Intercourse	8
Insertive Penile-Vaginal Intercourse	4

HIV disease staging and classification systems are important tools for tracking and monitoring the HIV epidemic and for providing clinicians and patients with important information about HIV disease stage and clinical management. Two major classification systems currently are in use: the U.S. Centers for Disease Control and Prevention (CDC) classification system and the World Health Organization (WHO) Clinical Staging and Disease Classification System.

The CDC disease staging system (most recently revised in 1998) categorizes persons on the basis of clinical conditions associated with HIV infection and CD4+ T- lymphocyte count (www.cdc.gov). The system is based on three ranges of CD4+ T- lymphocyte counts and three clinical categories and is represented by a matrix of nine mutually exclusive categories. It provides uniform and simple criteria for categorizing conditions among adolescents and adults with HIV infection and should facilitate efforts to evaluate current and

future health-care and referral needs for persons with HIV infection. The definition of AIDS includes all HIV-infected individuals with CD4 counts of <200 cells/ μ L (or CD4 percentage <14%) as well as those with certain HIV-related conditions and symptoms. Although the fine points of the classification system rarely are used in the routine clinical management of HIV-infected patients, a working knowledge of the staging criteria (in particular, the definition of AIDS) is useful in patient care. In addition, the CDC system is used in clinical and epidemiologic research. This system replaces the classification system published in 1986, which included only clinical disease criteria and which was developed before the widespread use of CD4+ T-cell testing.

In contrast to the CDC system, the WHO Clinical Staging and Disease Classification System (revised in 2007) can be used readily in resource-constrained settings without access to CD4 cell count measurements or other diagnostic and laboratory testing methods. The WHO system classifies HIV disease on the basis of clinical manifestations that can be recognized and treated by clinicians in diverse settings, including resource-constrained settings, and by clinicians with varying levels of HIV expertise and training.

1.5. DIVERSITY OF HIV

One of the most important factors in the worldwide spread of HIV is its rapid evolution and enormous genetic variability. Its evolutionary rate is approximately one million times faster than higher organisms such as humans, meaning that the amount of changes within the HIV-1 genome in just one year corresponds to the amount of changes within the human genome in one million years.

A first investigation of HIV genome diversity indicated that the major reason for high sequence variation is the lack of proof reading and poor fidelity of HIV-1 Reverse transcriptase. Soon after, it became clear that this represents only one of the several aspects of HIV diversity. Other members of the *Retroviridae* family, which have mutation rate similar to that of HIV-1, have a lower genetic variation because these viruses replicate less frequently in their hosts. The enormous genetic variation in HIV-1 in patients is primarily due to the very rapid turnover of a relatively large population of infected cells (Butler et al., 2007).

The second evolutionary mechanism of HIV that is critical to its enormous genetic diversity is an unusually high recombination frequency. As with other retroviruses, HIV-1 recombines during reverse transcription (Butler et al., 2007). The HIV reverse transcriptase can utilize both copies of the co-packaged viral genome in a process termed retroviral recombination. In addition to the obligatory strand transfers that occur at the ends of the genome, RT has been shown to switch between co-packaged RNA templates within internal regions of the genome, leading to formation of recombinant DNA molecules (Goodrich and Duesberg, 1990). It is now apparent that template switching occurs throughout the HIV genome and is much more frequent than mutation (Schlub et al., 2010). It has been estimated that HIV-1 undergoes recombination at a rate of 2.8 crossovers per genome per cycle.

In addition, the high replication rate of the virus in combination with selective forces in the host environment further contributes to the genetic variation. As a consequence, an HIV-1 population within a person consists of a large number of genetically related but non identical viruses, a population structure that gives this pathogen an opportunity of rapid adaptation to changes in its environment. The quasispecies concept has often been used to describe the genetic variation of HIV populations, although it has been argued that not all requirements are completely fulfilled. The HIV quasispecies is

shaped at almost every stage of the virus life cycle. The clinical implications of the great genetic variability are extensive. It allows the virus to escape the host immune system survey, allowing drug resistance development and escape candidate vaccines. Therefore, knowledge about the genetic variation of HIV-1 is important for the drugs and vaccines design and for improving combination therapy. It can also help us to understand more the natural history and pathogenesis of the infection.

Together with the founder effect genetic diversity of HIV has resulted in several genetically divergent lineages that can be classified into groups, subtypes and sub-subtypes based on their phylogenetic relationships. This classification also reflects important zoonotic transmission events.

1.5.1 MOLECULAR "FOOTPRINTS" ON THE HIV GENOME

The HIV evolutionary process is driven by complex interplay between viral and host factors, leaving a measurable footprint in viral gene sequences (Lemey et al., 2006). Within hosts, the viral evolution is strongly influenced by natural selection as a result of a continuous effort to evade the immune response. HIV successively fixes mutations that allow it to evade immune response, acquiring combinations of polymorphisms in viral genome termed „molecular footprint“. For instance, previous work has shown that HLA alleles, in particular those alleles associated with effective immune control of HIV (viral evasion of CD8⁺ T-cell responses), such as HLA-B*57 and HLA-B*27, select a characteristic, predictable combination of escape mutations in the virus – termed a “footprint” (Goulder et al., 2004; Kiepiela et al., 2007; Leslie et al., 2004;

McMichael and Klenerman, 2002). Several genetic polymorphisms, gene variants and number of certain genes have been studied and persistently related to HIV pathogenesis and disease progression. The most clearly defined footprints are those associated with HLA alleles that are linked with successful control of HIV, such as HLA-B*57 (Matthews et al., 2009; Kiepiela et al., 2007; Leslie et al., 2004).

It is of paramount importance to investigate the extent to which HLA footprints impact on HIV phylogeny within a clade and these studies is still lacking, Evidence of favorable and unfavorable relationships of certain HLA-1 markers to virologic and immunologic outcomes of HIV infection is conclusive. HIV infection is one of only a few infectious diseases showing a clear-cut and consistent HLA association which has been confirmed by several genome wide association study (GWAS).

Additionally, other viral features may have important influences on viral evolution, pathogenesis and disease outcomes, often through indirectly disturbing replicative capacity of virus, including resistance mutations to antiretroviral drugs and other mutations in different HIV-1 genes. The impact of positive selective pressure on viral diversity inter-host, as opposed to spatial and temporal factors, is not equally clear.

A variety of statistical models and inference techniques have been developed to reconstruct the HIV evolutionary history and to investigate the population genetic processes that shape viral diversity.

Investigation of the level of HIV genetic diversity observed within viral sequences could be complementary approach to antibody assays and surrogate epidemiological definitions of recent infection. Viral diversity in individual patients is reflected in the proportion of ambiguous bases observed in HIV-1 pol sequences obtained from population based genotyping with increasing diversity seen with increased duration of infection (Kouyos et al., 2011).

Ambiguous bases appear in sequences derived by population based sequencing due to the simultaneous detection of more than one nucleotide at the same position of the genome. Viral diversity has been used to estimate time since infection in HIV-1, mainly subtype B infection (Kouyos et al., 2011)

1.6. HIV CLASSIFICATION

1.6.1 HIV TYPES, GROUPS AND SUBTYPES

Accurate HIV-1 subtyping information is vital for all kinds of HIV research. It gives insights into molecular epidemiology, viral evolution, and facilitates subtype-specific vaccine antigens and testing reagents.

To date, two types of HIV have been identified; HIV type 1 (HIV-1) in 1983 and HIV type 2 (HIV-2) in 1986 (**Figure 5**). While HIV-1 is widely distributed throughout the world, HIV-2 infection is predominantly found in West African nations, such as Guinea-Bissau, Gambia, Senegal, Cape Verde, Cote d'Ivoire, Mali, Sierra Leone, and Nigeria. However, an increasing number of cases have been recognized in Europe (mainly in Portugal and France), India, and the United States, and other countries especially in those with historical and political ties to West Africa (Campbell-Yesufu and Gandhi, 2011).

HIV-1 is the pandemic type and consists of 4 groups, groups M (main or major), non-M, non-O (N), outlier (O) and pending (P), each of which represents independent cross-species transmission events of Simian immunodeficiency virus (SIV) from chimpanzees (*Pan troglodytes troglodytes*) and gorillas (D'arc et al., 2015; Gao et al., 1999; Keele et al., 2006). This, together with human adaptation, accounts for their genomic, phylogenetic, and virological specificities (Mourez et al., 2013).

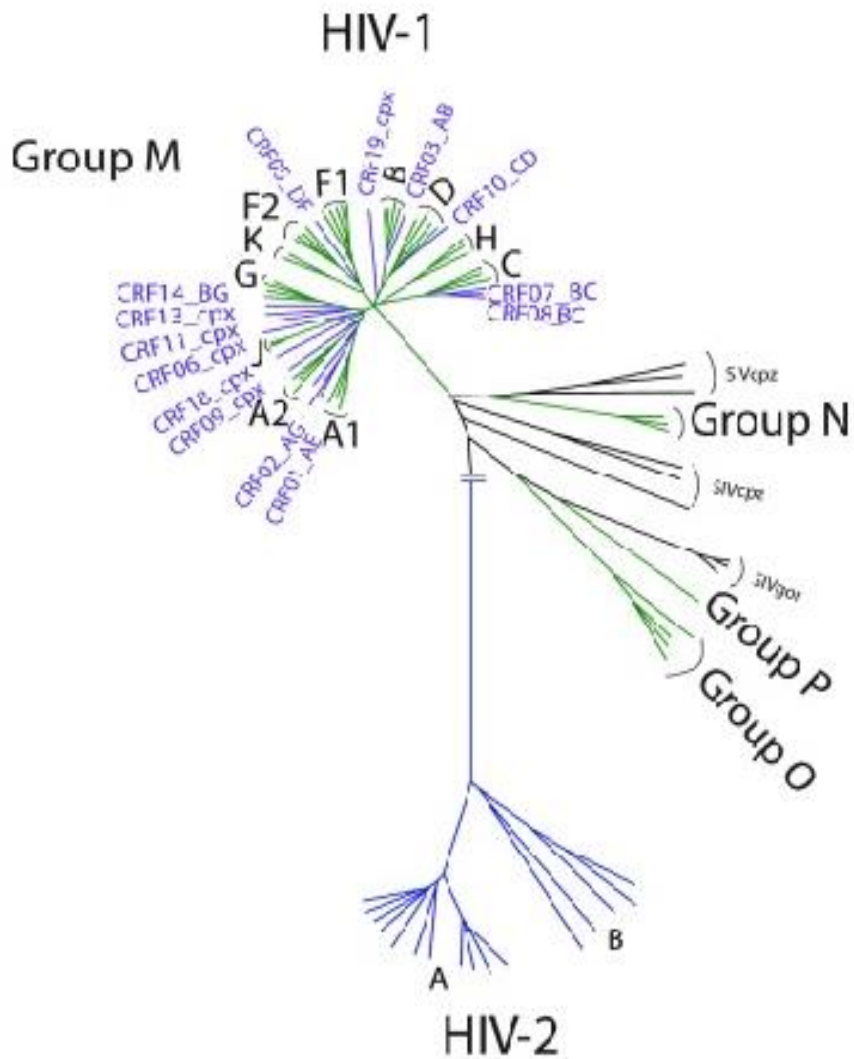


Figure 5. Classification of HIV in types 1 (green) and 2 (dark blue). HIV-1 is divided in groups M, N, O and P. The group M includes nine subtypes, subtype A and F are divided in subsubtypes e.g. A1 or F1. Some examples of circulating recombinant forms (CRFs) are shown in violet. HIV-2 is classified in eight groups (A-H), but only the most frequent groups A and B are shown. Modified from Tebit et al., 2011.

When represented on a phylogenetic tree strains within group M form well defined clusters (**Figure 5**). In the early 1990s the sequencing and alignment of viral genes *env* and *gag* from different strains of HIV-1 allowed for first time to establish the presence of well-defined HIV-1M genetic clades. Based on this information, subtypes A, B, C, D, E and F were recognized in 1993. In 1994, subtypes G and H were characterized in Central Africa, and later subtypes I (1995), J (1999) and K (2000) were described (Janssens et al., 1994; Kostrikis et al., 1995; Laukkanen et al., 1999; Triques et al., 2000; Peeters et al. 2013; Tongo et al., 2015). All HIV strains previously classified as subtype E based on *env* gene phylogeny had divergent subtype A classification in *gag* and *pol*, revealing that HIV-1 can generate inter-subtype recombinant strains. Indeed, a “pure” subtype E has not been found to date (Gao et al., 1996).

According to the current classification, HIV-1 group M is divided into nine different “pure” subtypes or non-recombinant forms (A-D, F-H, J and K) (Robertson et al., 2000). Viral strains belonging to subtype A and F cluster distinctly into two different sub-lineages and therefore further divided into the sub-subtypes A1/A2 and F1/F2, respectively. For historical reasons the B and D clades are called subtypes, but in fact the genetic distance between these two clades corresponds to a sub-subtype distance. Additional taxonomic units include regional subepidemics (monophyletic clades), such as the B epidemic in Southeastern Asia, or the subtype A former Soviet Union, (variant Afsu) epidemic in PWIDs in former Soviet Union countries (FSU) (Bobkov et al., 1997; Bobkov et al., 2001; Deng et al., 2008). These taxonomic units correspond to sub-epidemics in specific geographic areas forming monophyletic clusters within major HIV-1 clades (subtypes or CRFs).

HIV-1 genomic diversity is the lowest within single patients and increases in the following order when different patients are considered: within subtypes, between subtypes, between groups and between HIV types (Li et al.,

2015). A nucleotide genomic diversity was quantified to be 48.3% between HIV-1 and HIV-2, 37.5% between HIV-1 groups, 14.7% between HIV-1 subtypes, 8.2% within HIV-1 subtypes, and 0.6% within single patients infected with HIV-1. (Li et al., 2015). The highest variation within the genome is seen within the *env* gene, whereas the *pol* gene, encoding for important viral enzymes are the most conserved (Gaschen et al., 2002; Lee et al., 2009).

Group N has been identified in very few cases since its first description in 1998. With the exception of 1 case, all are documented in Cameroon (Sharp and Hahn, 2011; Vallari et al., 2010; Mourez et al., 2013).

Group O is limited to Cameroon and other adjacent countries representing less than 1% of HIV-1 infections (Bush et al., 2015; D'arc et al., 2015; Leoy et al., 2015; Roques et al., 2002; Sharp and Hahn, 2011). Sporadic cases of HIV-O infection have been described in east and West Africa, Europe and the United States but always in Cameroonians or in patients with a link to Cameroon.

Group P was discovered in 2009 and, despite extensive screening, has been detected insofar in only two persons (D'arc et al., 2015; Mourez et al., 2013).

For HIV-2, the groups A-H are known but only groups A and B have infected substantial number of people in West Africa (Sharp et al., 2011; Ayouba et al., 2013; Visseaux et al., 2016). To date, no subtypes have been formally described but some preliminary data suggest that HIV-2 group A may be divided in two distinct subtypes with distinct geographical origins.

To date only two recombinant forms have been described: one circulating recombinant form (CRF01_AB) and one unique recombinant form. Overall, HIV-2 group A predominates. HIV-2 group B is less prevalent and co-circulates with HIV-2 A mainly in Ivory Coast and Ghana (Visseaux et al.,

2016). Nowadays, in West Africa, between 1 and 2 million people live with HIV-2, but both its incidence and prevalence show decreasing trends probably because the transmission efficiency of HIV-2 is lower than that of HIV-1 (Campbell-Yesufu and Gandhi, 2011; Tienen et al., 2010).

Dual infections with HIV-1 and HIV-2 have been frequently observed in areas where both viruses co-circulate, but today no recombinant virus between HIV-1 and 2 has been documented yet.

1.6.2 HIV RECOMBINANTS

Other significant clusters are formed by Circulating Recombinant Forms (CRFs) and unique recombinant forms (URFs). Currently, more than 80 CRFs and numerous URFs have been reported (<https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html>). Analyses of multiple genome regions and in particular full-length genome sequencing have revealed that recombination between strains is a frequent occurrence.

Recombinants between different HIV-1 group M subtypes are designated as either circulating recombinant forms (CRFs) if fully sequenced and found in three or more epidemiologically unlinked individuals or as unique recombinant forms (URFs) if not meeting these criteria. These have arisen as a result of recombination events between divergent HIV strains within individual hosts (<http://www.hiv.lanl.gov>). If a patient gets infected multiply with viruses from different subtypes, it can occasionally happen that two viruses recombine to a new viral strain with a mosaic genome composed of regions from both subtypes. Certain CRFs, like CRF01_AE and CRF02_AG, were already present early in the epidemic but many other CRFs emerged more recently.

1.6.3 DISTRIBUTION OF HIV-1 SUBTYPES

Geographical distribution of HIV-1 subtypes worldwide is complex and influenced by transmission risk behaviors and the patterns of human mobility. Among the four HIV-1 groups, only HIV-1 group M has spread worldwide. As the AIDS pandemic progresses, an increasingly broad range of genetic diversity is being reported within the M group of HIV-1 viruses (Gifford et al., 2007; Tongo et al., 2015)

On a global scale, subtype C predominates representing almost half of HIV-1 infections (47%), followed in decreasing order by subtype A (27.2%), B (12.3%), and D (5.3%) (Hemelaar et al., 2006; Hemelaar et al., 2011; Osmanov et al., 2002). Other subtypes (F, H, J and K) and all other CRFs represented about 5% of infections in the world (**Figure 6.**) (Hemelaar et al., 2011). However, the global proportion of all CRFs increased by 4.5% in the recent years.

The first HIV-1 subtype identified was subtype B, which is predominant in many resource-rich settings including North America, western and central Europe and Australia, along with other regions like the Caribbean and parts of Latin America, while subtype C is predominant in more resource-limited settings like Ethiopia, South Africa and India. Subtype B is most prevalent in the developed, industrialized regions of the world, and therefore a representative of this clade was used for development of antiretroviral drugs.

When we look at the subtype distribution worldwide, we see that subtype C is almost exclusively responsible for all infections in Southern Africa, India and Ethiopia, responsible for 30, 13 and 4% of global infections respectively (**Figure 5**). Central Africa is thought to be the origin of all subtypes of HIV-1 (Hemelaar et al., 2011; Vidal *et al.*, 2000) The initial diversification of group M viruses may have occurred within or near the territory of the Democratic Republic of Congo (DRC), which is considered as the epicenter

where the highest diversity of group M has been reported and from where the different HIV-1 M variants started to spread across Africa and subsequently to other continents in the world. A high genetic diversity is also seen in the surrounding countries like Cameroon, Angola, Central African Republic, Gabon and Equatorial Guinea.

In West Africa, which harbors 16% of the global HIV-1 infections, subtype A, G and CRF02_AG are the most predominant. Subtype A, C, D and unique recombinant forms make up most of the HIV-1 infections in East Africa, which has 10% of global infections. Subtype A dominates the subtype diversity in Kenya, while subtypes D, C and G are responsible for most of the remaining infections. In Southern Africa, subtype C accounts for 92% of urban HIV cases while subtype B plays a minor role, with about 7% of cases (Williamson et al., 2003). Subtype C is primarily transmitted heterosexually, and subtype B is associated with homosexual transmission (Bredell et al., 2000; Pillay et al., 2002). The majority of other subtypes in South Africa occur in immigrant populations from elsewhere in the continent (Bredell et al., 2002).

CRF01_AE is responsible for most infections in South and South-East Asia (excluding India where subtype C is predominant), while in East Asia the burden of infections is caused by subtype B, CRF01_AE and other recombinants (Aldrich et al., 2012; Hemelaar et al., 2011).

In Europe there are clear differences in geographical distribution and prevalence of HIV subtypes. After the first introduction of subtype B in early 80s, this clade is still predominant one in most Western and Central European countries, whilst in Eastern Europe the epidemic has been dominated by subtype A (variant Afsu) (Eastern-type epidemic) (Abecasis et al., 2013; Hemelaar et al., 2011).

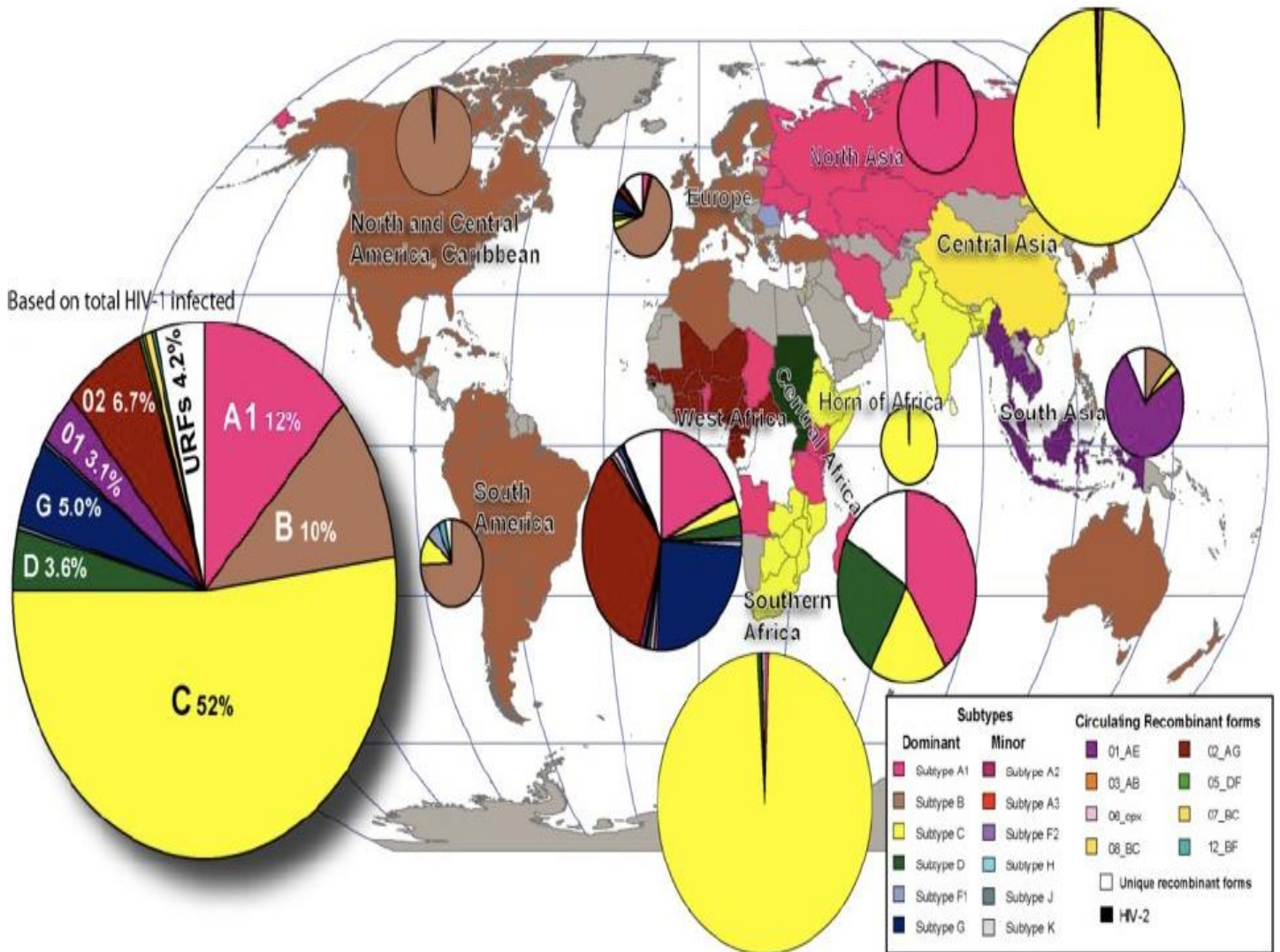


Figure 6. Global distribution of HIV-1 subtypes and recombinants.

Modified from (<http://www.hivviralload.com>)

However, in Europe the complexity of HIV-1 epidemic has been increasing during recent years. The cumulative number of diagnosed HIV-1 infections in the European continent (i.e. European Union, European Economic Area (EU/EEA), Russia and FSU countries) reached 1,840,136 by the end of 2014 with 49% of these diagnosed in Russia, as reported by The European Surveillance System (TESSy), a joint ECDC/WHO database for HIV/AIDS surveillance ((ECDC), 2015) (Beloukas et al., 2016). Each country has a unique pattern of HIV-1 epidemic shaped by different regional circumstances and high prevalent local transmission networks with patterns varying between individual countries. Non-B epidemics are mainly associated with immigrants, heterosexuals and females but more recently, non-B clades have also spread amongst groups where non-B strains were previously absent - non-immigrant European populations and amongst men having sex with men (MSM).

The subtype distribution in some of the Balkan countries (e.g. Croatia, Slovenia, Montenegro, Hungary, and Serbia) is similar to that in many of the Western and other European countries, with the predomination of subtype B. On the other hand, subtype distribution of HIV-1 in Greece, Albania, and Bulgaria is marked by predominance of non B subtype (**Figure 7**) (Stanojevic et al., 2012). In Turkey, the prevalence of non-B clades is very high and include many different subtypes and CRFs. Unfortunately, the origin of these transmissions and/or local epidemics remains unclear. In Albania, local HIV-1 epidemic is characterized by a high prevalence of non-B infections (65.2%) (Ciccozzi et al., 2005). Specifically, spread of subtype A in Albania is a result of a founder effect from the A clade epidemic in neighboring Greece (Paraskevis et al., 2007; Salemi et al., 2008a). In Bulgaria, there are several HIV-1 subtypes circulating and, as it has been shown, clades B and A1 were introduced by at least three or four independent sources in last 25 years (Salemi et al., 2008b). Finally, in Romania, the HIV-1 epidemic is unique as the globally-rare subtype F1 predominates and any non-F1 subtypes are referred to as divergent strains.

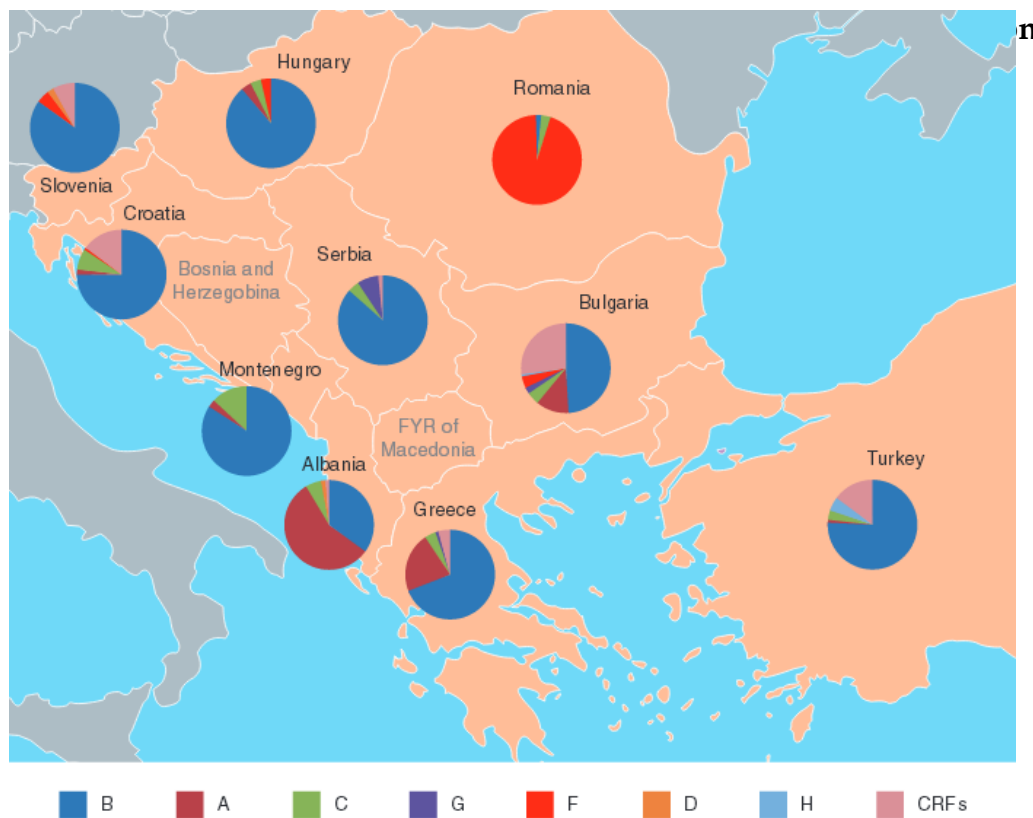


Figure 7. General distribution of HIV- 1 subtypes and recombinants in the Balkan region. Modified from Stanojevic et al., 2012.

Here, significant numbers of mainly institutionalized children were infected in the late 1980s via transfusion of infected blood products or unsafe parenteral treatments (Apetrei et al., 1997). Since 2010, an increasing trend of HIV-1 infections among people who inject drugs (PWID) has been observed, largely centered in Bucharest. Although F1 clade still predominates in the PWID epidemic, other clades, including CRF14_BG, have also been found (Niculescu et al., 2015).

The time of the initial identification of the HIV-1 epidemics in the Balkans is comparable to the Western European countries. The first HIV-1/AIDS cases in the majority of countries were diagnosed in the beginning or mid-eighties, with the exception of Albania where the first case was registered in 1992.

In Serbia, the predominance of subtype B was reported (in 87.5-91.2%) with other subtypes also appearing namely G (7.7%), C (4.4%) and CRF (1.8-2.2%) (Stanojevic et al., 2012).

1.6.4 THE SIGNIFICANCE OF HIV-1 SUBTYPE DIVERSITY

HIV-1 subtypes differ in biological characteristics that may affect pathogenicity and disease progression. Monitoring of subtypes distribution can give valuable information to a health care services with regards measures to sustain the epidemics and reduce onward transmission. One of the major challenges of the subtype diversity might be on therapeutic and preventive management of the disease. With ongoing generation of viral genetic diversity and potential for the emergence of novel genetic variants, it has become increasingly important to establish potential clinical implications of subtype variation. Some studies gave evidence that different subtypes may respond differently to drug therapy and found different rates of disease progression within HIV subtypes, while others suggested limited significance of differences among subtypes with regards development and clinical interpretation of antiviral resistance (Camacho and Vandamme,2007; Easterbrook et al., 2010; Kantor, 2006; Vijver et al., 2006).

An important problem extensively discussed in the recent years relies on the impact of ARV drugs in different HIV-1 subtypes. This concern is based on the fact that the great majority of the investigations conducted on HIV drug design, acquisition of drug resistance mutations, genotyping for drug resistance evaluation, and the phenotypic impact of drug resistance mutation (DRM) on HAART have been performed primarily for subtype B strains.

Various non-B HIV-1 group M subtypes present genetic signatures and polymorphisms in their protease gene which are considered as compensatory DRM in subtype B. These differences have triggered discussions as to whether those non-B subtypes are naturally less susceptible to PIs, which could in turn compromise the use of PI-containing HAART regimens. Upon

antiretroviral treatment, such differences in baseline polymorphisms among subtypes may result in the evolution of drug resistance along distinct mutational pathways, or in differences in the incidence of these specific pathways. These genetic differences may be clinically relevant when considering long-term treatment strategies for patients infected with different subtypes.

1.7. ANALYSES OF HIV USING PHYLOGENETICS

Phylogenies are important for addressing various biological questions such as relationships among species or genes, the origin and spread of viral infection and the demographic changes and migration patterns of species.

Molecular phylogeny is the science of estimating evolutionary histories using DNA and amino acid sequences. Before the advent of DNA sequencing technologies, phylogenetic trees were used almost exclusively to describe relationships among species in systematics and taxonomy. The advancement of sequencing technologies has taken phylogenetic analysis to a new height. Today, phylogenies are used in almost every branch of biology.

A phylogeny is an evolutionary tree in which the leaves of the tree are the sampled sequences or taxa, branches are the genetic distance between taxa, and the nodes denote estimated separation events. Reconstruction of phylogenies can be accomplished through a variety of methodological approaches including neighbor joining (NJ), maximum likelihood (ML), and Bayesian methods that are described in detail in Chapter 3.

In *On the Origin of Species* in 1859, Charles Darwin first proposed the phylogenetic tree as a structure to describe the evolution of different organisms.

The tree structure is currently the accepted paradigm to represent evolutionary relationships between organisms, species or other taxa. A phylogenetic tree is a geographical representation of the evolutionary relationships that exist between aligned sequences (Maddison, 2000; Whelan et al., 2017). Once a phylogenetic tree has been constructed, any characteristic of the related organisms can be analyzed in the evolutionary framework specified by the tree. Changes in such characteristics (“characters”) can be inferred between descendent organisms and their ancestors—even if those ancestors were never observed directly (Maddison and Maddison, 2000). The first step in any phylogenetic investigation of evolutionary history is the identification of homologous sequences. Once a tree has been created there are many programs available for viewing and analyzing the tree topology (described in detail in Chapter 3.) (Whelan et al., 2017).

Historically, the case of the dentist and his patients who have been infected with HIV-1 in Florida was one of the first with regard to transmission investigation (Abele and DeBry, 1992; Palca, 1992a, 1992b). Shortly after, phylogenetics became an important tool for HIV research, that offers opportunities to understand critical aspects of the HIV epidemic. The application of molecular phylogenetics has become a common practice of many HIV/AIDS research studies, due mainly to the many insights these analyses can provide and the novel questions they can address over a variety of topics related to HIV biology. Together with population genetics and epidemiological principles, nowadays scientists are using viral phylogenetic to improve our understanding of HIV diversity, generating an unprecedented knowledge of viral dynamics to improve strategies of HIV prevention and treatment of HIV infected persons.

Phylogenetic analyses have been used extensively in the studies of HIV-1, and have proven to be an invaluable tool, revealing important aspects of evolution, origins, disease progression, and transmission of this virus. To date,

the application of phylogenetic analysis to the study of HIV has elucidated the origin of HIV-1 and HIV-2, the relationships of HIV to other simian lentiviruses, for characterizing HIV-1 genetic heterogeneity (i.e. molecular surveillance, phylogenetic classification into groups, subtypes, sub-subtypes, recombinant forms) and spatiotemporal characteristics (Castro-Nallar et al., 2012; Robbins et al., 2003). Phylogenetic tools brought to light dimensions of HIV evolution such as “where and when” and even “how” infections are spreading worldwide that are impossible to assess with other approaches. Phylogenetic analyses have also been key to the identification of drug-resistance mutational transmission chains. (Lemey et al., 2005; Machado et al., 2009).

1.7.1 ANALYSES OF HIV ORIGIN USING PHYLOGENETIC

Nowadays, it is still not known how many people became infected with HIV before the initial recognition of AIDS in the USA in the early 1980s. However, it is clear, that before its discovery the virus silently spread to at least five continents of the world and it has been suggested that at least 100 000-300 000 persons were infected. Molecular analyses of stored serum sample and lymph node biopsy specimen collected in 1959 and 1960, respectively, in Democratic Republic of Congo (DRC) showed seropositive results for HIV (Zhu et al., 1998).

A year after the first identification of HIV-1 as the cause of AIDS, the first Simian lentivirus, SIVmac, was isolated from captive rhesus macaques (*Macaca mulatta*) at the New England Primate Research Center (NEPRC) (Chakrabarti et al., 1987; Daniel et al., 1988; Li et al., 1989). These infected macaques developed symptoms very similar to AIDS in humans and soon the simian origin of AIDS was suspected. Thereafter, it became clear that macaques are not natural hosts

of virus but became infected with SIVsmm from captive sooty mangabeys that are naturally infected with this virus. (Fultz et al., 1986; Apetrei et al., 2005).

Today, more than 40 SIVs have been identified in different non human primates (NHP) species from Africa. SIVs are named according to the host species, and a three letter code refers to the common name of the corresponding NHP species.

Phylogenetic investigations have shown that HIV originated from multiple crosspieces transmissions from NHP that had taken place long before the first diagnoses (**Figure 8.**) (Keele et al., 2006; Sharp and Hahn, 2011; Van Heuverswyn et al., 2006). The most recent ancestor of the pandemic strain was probably circulating in human populations in Central Africa early in the 20th century (Faria et al., 2014; Korber et al., 2000; Worobey et al., 2008). Already thirteen transmissions involving three different NHP species to humans have been documented, four for HIV-1 and nine for HIV-2.

How the virus crossed from NHP to humans is still debated, but the most accepted theory is the so called „hunter theory“. These zoonotic transmissions probably happened during the hunting and butchering of primates for bush meat, as well as the capture, trade and keeping of monkeys as pets (Hahn et al., 2000). Through studies of phylogenetics, it was found that HIV-1 is most closely related to Simian immunodeficiency virus (SIVcpz) which is found in the chimpanzee sub-species *Pan troglodytes troglodytes* (Gao et al., 1999, Korber et al., 200, Yusim et al., 2001). Even though SIVcpz was also found in the eastern chimpanzee subspecies (*Pan troglodytes schweinfurthii*), virus from this group does not appear to have jumped successfully into humans (Santiago et al., 2002). Extensive analyses of fecal samples of wild chimpanzees suggested to the origin of HIV-1 group M in southeastern Cameroon and HIV-1 group N in south central Cameroon (Keele et al., 2006). HIV-1 groups O and P originated from SIVgor identified in Western lowland gorillas living in Cameroon.

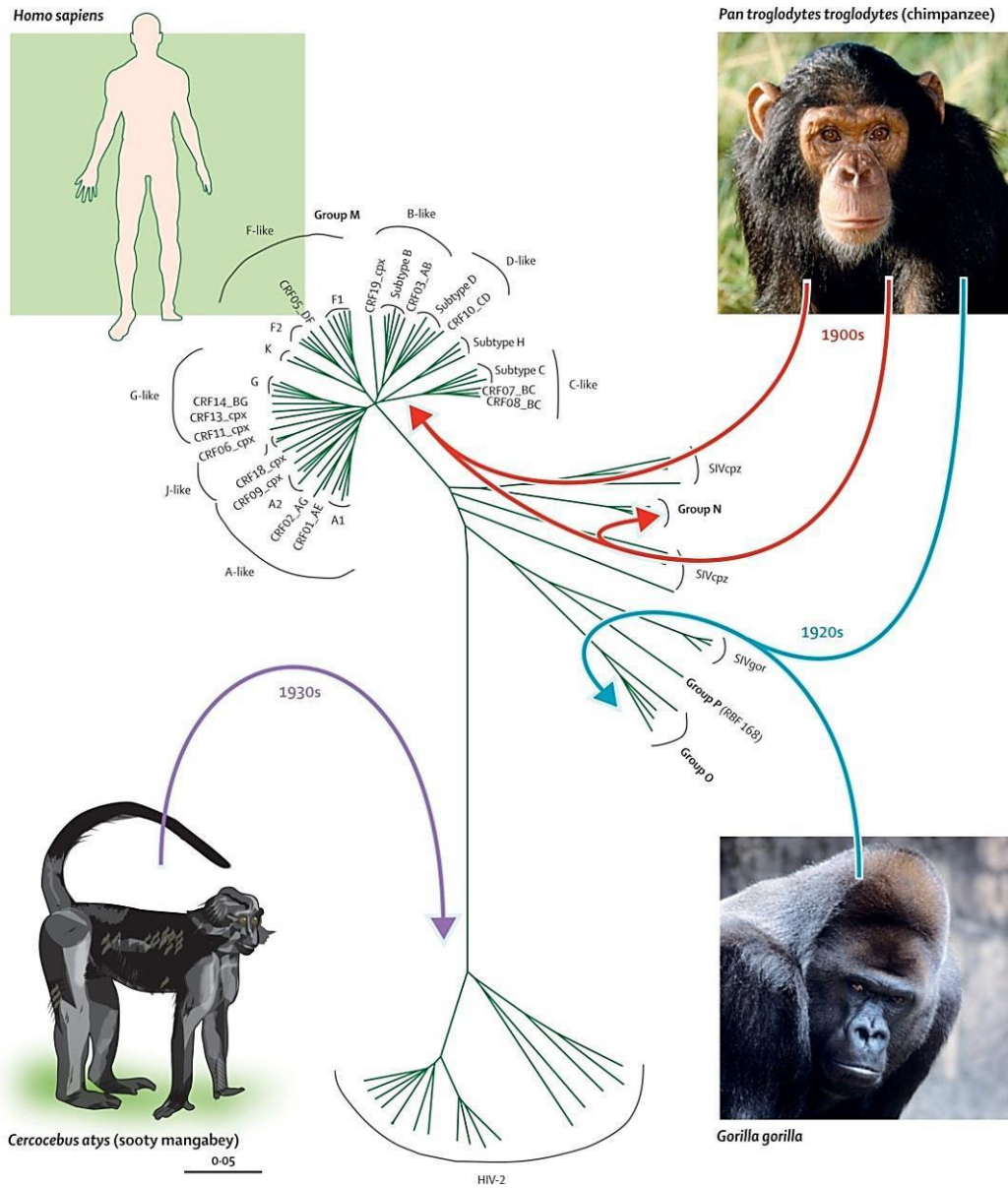


Figure 8. Cross-species transmission of HIV and distribution of subtypes and recombinant forms.

Modified from Tebit et al., 2011

Using the molecular clock dating analyses, the time of the Most Recent Common Ancestor (tMRCA) of the pandemic HIV-1 group M was initially dated back to 1931 (1915-1941) (Korber et al., 2000). However, after the identification of the oldest HIV-1 viruses ZR59 and DRC60 from the DRC and inclusion of these viral sequences in the analysis, the tMRCA (using phylogenetic method Bayesian skyline plot tree prior) has been pushed back slightly to 1908 (1884-1924) (Worobey M et al., 2008; Zhu et al., 1998). Both HIV-1 group M and O dates were inferred using a relaxed molecular clock, which allows the rate of evolution to vary along different branches of the tree. The tMRCA of O and N group was estimated at 1920 (1890- 1940) and 1963 (1948-1977) respectively (Lemey et al., 2004; Wertheim et al., 2009_HIV-2 group A and B tMRCAs were estimated to be 1940 (1924-1956) and 1945 (1931-1959), respectively (Lemey et al., 2003). These dates were estimated using a strict molecular clock, (i.e., a single, constant evolutionary rate along all branches) (Lemey et al., 2003).

By analyzing viral sequences obtained over several decades, Faria et al., used Bayesian skyline plot analyses and estimated that between 1920 and 1960, group M underwent an early phase of relatively slow exponential growth. Further analyses showed that, around 1960 (95% BCI: 1952-1968) group M transitioned to a second, faster phase of exponential growth. Although the tMRCA of group O is similar to that of group M and both grew at similar rates until ~1960, group O exhibits no subsequent increase in growth rate and remains largely confined to Cameroon and surrounding countries (**Figure 9**).

After the initial cross-species transmission, that took place in southern Cameroon, presumably resulting from the hunting of primates, group M viruses most likely entered Kinshasa, the capital of the Democratic Republic of Congo, around 1920 (1909-1930) (Keele et al., 2006; Faria et al., 2014). In that time, Cameroon was under German colonization and fluvial connections

between southern Cameroon and Kinshasa were frequent due to the exploitation of rubber and ivory. After localized transmission in Kinshasa, the virus probably reached large neighboring cities, such as Brazzaville, Lubumbashi, and Mbuji-Mayi within about 20 years (**Figure 10.**) (Faria et al., 2014).

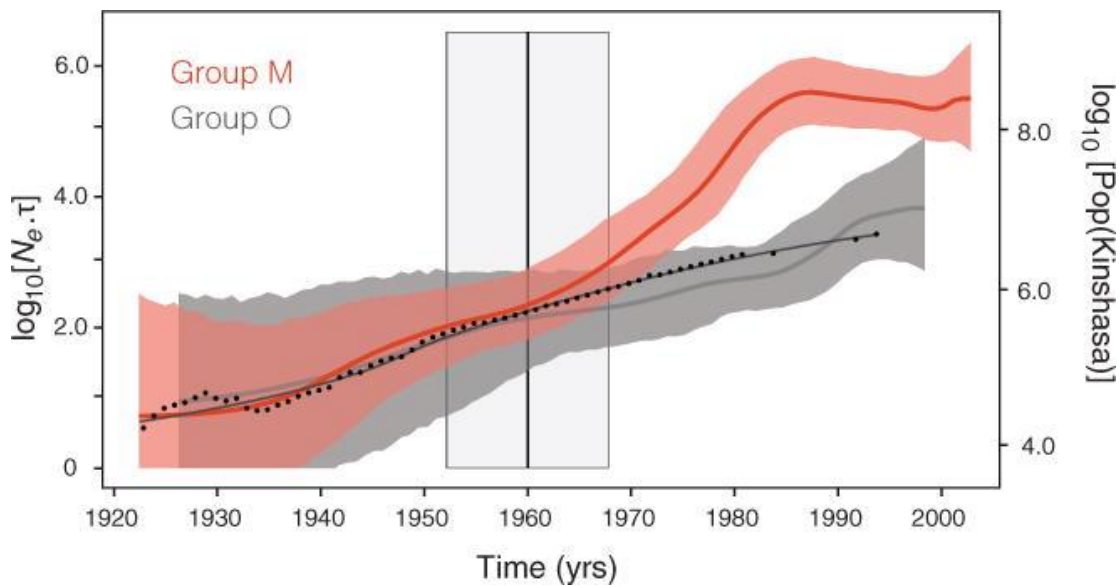


Figure 9. Population dynamics of HIV-1 groups M and O obtained by Bayesian skyline plot analyses. Modified from Faria et al., 2014.

The changing behavior of sex workers, urban growth and human mobility by railway connections play fundamental roles in the early spread of the virus and may explain the linkage among these cities (Faria et al., 2014; Hahn et al., 2000; Worobey et al., 2008). Crucial factor responsible for dissemination of both HIV-1 and HIV-2 worldwide was the intense migration of individuals, from rural to urban centers with subsequent return migration and internationally due to civil wars, tourism, business purposes, and the drug trade. The migration of poor, rural, sexually active young people to urban centers in the Third World clearly played a role in the dissemination of HIV and other infectious diseases

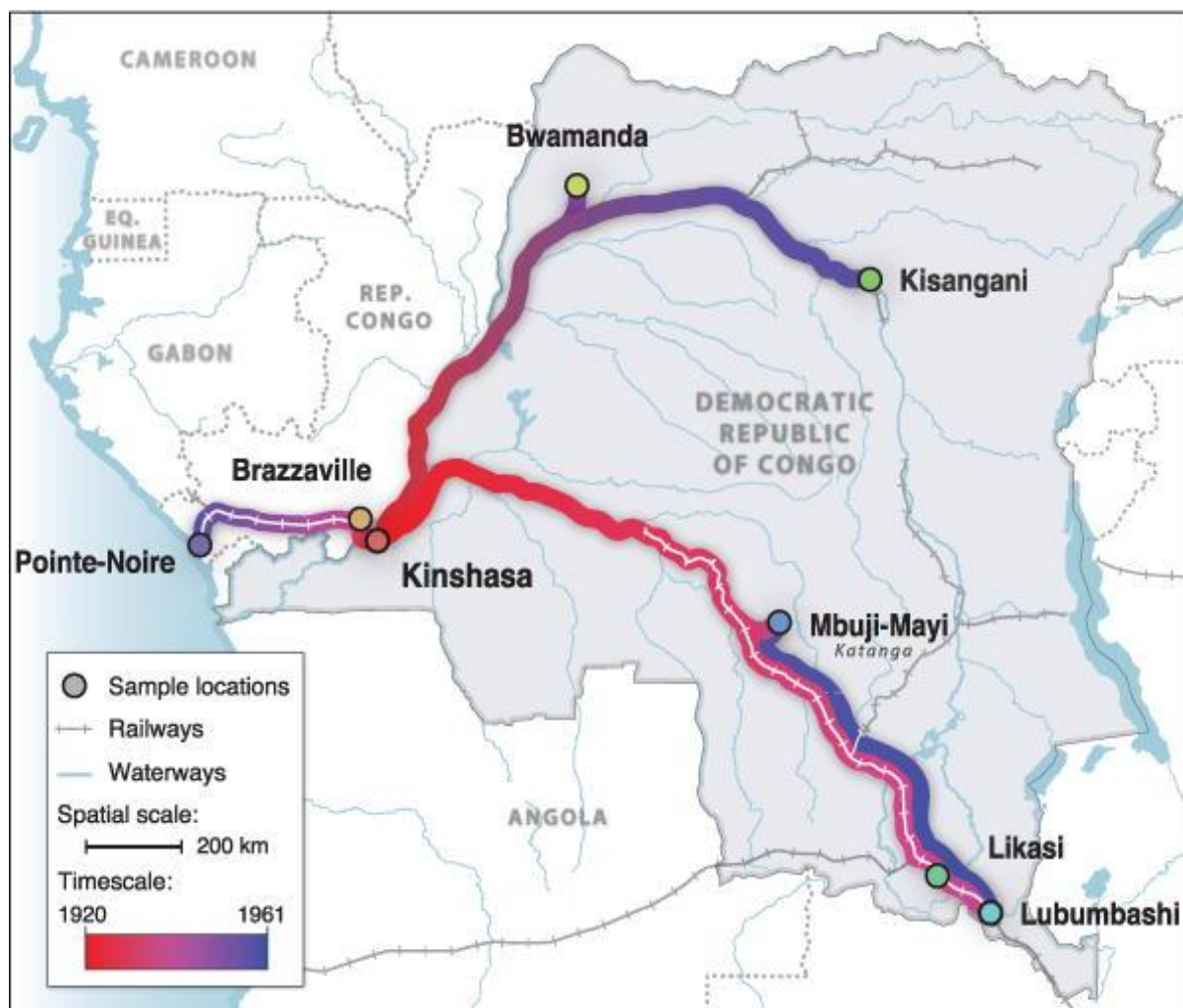


Figure 10. Spatial spread of HIV-1 group M. Modified from Faria et al., 2014.

Many studies, dealing with a phylogenetic origin of HIV, were based on a specific mechanism which facilitate the first few serial transmissions of the virus in humans, while largely limit the problem of initial adaptation (Gisselquist, 2003; Marx et al., 2001). Recent research proposed a major role of genital ulcer diseases (GUD), such as syphilis, chancroid, and lymphogranuloma venereum (LGV), and possibly also a relatively low frequency of male circumcision, in initial HIV transmission and adaptation to

humans (de Sousa et al., 2010). GUD induced genital ulcers can dramatically increase HIV transmission, and this research suggested that the high prevalence of these diseases might have provided crucial “help” for the initial spread of the virus. GUD and especially syphilis induces a potent inflammatory response, and tumor-necrosis-factor (TNF)- α production, which is a major enhancer of HIV replication (de Sousa et al., 2010; de Sousa et al., 2012).

1.7.2 ANALYSES OF HIV TRANSMISSION CHAINS

DNA sequencing and molecular phylogenetics are increasingly being used in virology laboratories to study the transmission of viruses. In the field of molecular epidemiology it is considered that epidemiologically related sequences should group together in a phylogenetic tree, forming transmission clusters, because they all share a common, recent ancestor (Hue et al., 2004; Hue et al., 2005). HIV phylogenetics requires the generation of viral sequences, which in most cases are derived from virions isolated from the peripheral blood of infected persons with unknown duration of disease. Analysis of HIV nucleotide sequence data can be used to identify individuals with highly similar HIV strains and understand transmission. As HIV-1 is a fast evolving organism, it will accumulate mutations within reasonable amount of time, and simultaneously preserve some common signature patterns with its ancestors, such that genetic relationships between viruses within and among patients can be reconstructed in phylogenetic trees. For instance, several studies have demonstrated the presence of phylogenetic clusters of highly related HIV-1 sequences, particularly among recently HIV-infected individuals, which have been used to argue for a high transmission rate during acute infection.

Identification HIV clusters could be useful in tracking the leading edge of HIV transmission in epidemics. Analysis of HIV clusters could provide

valuable information for understanding the structure and dynamics of HIV transmission networks. If the virus is phylogenetically tightly clustered, it is an indication that the hosts are connected by a short chain of transmissions. Definitions of phylogenetic transmission clusters vary widely in the HIV literature, though many studies use both clade support and genetic distance cutoffs.

In addition, phylogenetic analyses have also been used to supplement evidence in criminal investigations of HIV transmission (de Oliveira et al., 2006; Saludes et al., 2013; Scaduto et al., 2010; González-Candelas et al., 2003; González-Candelas et al., 2013). Transmission of HIV between individuals may have important legal implications, therefore requiring forensic investigation, as for example when it occurs as implication of inappropriate medical practice or unprotected sex with a person aware of his/her infection status (Abecasis et al., 2011)).

DNA fingerprinting is a widely-accepted technique employed by forensic scientists as a means of matching individuals' respective DNA profiles (Jobling et al., 2004). Unlike DNA fingerprinting, phylogenetic analysis is a technique to compare how closely DNA sequences from different sources are related (Baxevanis et al., 2001). When using HIV phylogenetic analysis for forensic purpose, investigators are often warned to be cautious (Bernard et al., 2007). In fact, even if the strains carried by two subjects are more related to each other than the control strains, the following circumstances could not be excluded: both subjects were infected by one or more third parties with similar viruses, or a third party mediated HIV transmission from one subject to the other. Thus, phylogenetic analysis can and does include a certain degree of approximation and error. Phylogenetic evidence, in the context of other clinical and epidemiological evidence, can provide support for linkage between cases, but cannot be proof of transmission by itself (Bernard et al., 2007).

For forensic investigation based on phylogenetic analyses, at least two independent samples need to be blindly tested at two different time points, and the results between the time points should be consistent (Bernard et al., 2007). In order to give strong forensic evidence regarding transmission, phylogenetic analysis needs to be enhanced with application of several methods and conducted under strictly controlled conditions (Bernard et al., 2007; Leitner et al., 2000). Firstly, it is vitally important for phylogenetic analysis to include adequate local controls that are viral sequences from infected individuals sharing the same transmission risk, from similar geographic location, of the same genetic clades (subtype or recombinant form), and diagnosed in the same time period as the query subjects, but who are not believed to be a part of the investigated outbreak. The use of inappropriate controls could overemphasize the relatedness between the viruses under study as being uncommonly unique (Leitner et al., 2000). However, it is never possible to sample all infected individuals in a population. Nevertheless, if sufficient number of local control sequences is included in the analyses, significant clustering of sequences under investigation can indicate that they do belong to a transmission chain. Secondly, apart from appropriate local controls, additional precondition in applying forensic phylogenetic is inclusion into the analysis at least two genetic regions of reasonable length, depending on the gene under investigation. There are many different phylogenetic tree building methods that could be used within forensic research, and the choice is based on their reliability. Current techniques are not reliable enough to estimate the direction of transmission. Several studies is being done in this area, but for this purpose, multiple samples would need to be obtained very soon after the presumed transmission event between subjects included in the forensic investigation. In this research, end point limiting dilution (EPLD) assay was performed to access different subpopulations of HIV quasispecies, thus, improving the range of the quasispecies detection.

Molecular and phylogenetic analyses have inherent limitations, and are subject to issues that are associated with sampling and methods. More broadly, limitations include the paucity of molecular sequences or inadequate data (related not only to primary sequence data but also to demographics and additional parameters), which are common in resource-limited countries and settings (which often represent the greater burden of HIV disease), the non-representative nature of data, partial or incomplete phylogenetic signal in nucleotide sequences, and non-existence of reference datasets. HIV sequencing technology is rapidly changing, and it is critical to understand the advantages and limitations of the different sequencing technologies available prior to performing phylogenetic studies. Methodologically, the underlying assumptions, specific methods, and models used, coupled with computation time, and size of datasets (analyses on large datasets using computationally intensive methods cannot converge in reasonable time sometime) serve as limitations. Lastly, ethical issues around the confidentiality, particularly in the context of identifying networks of HIV transmission, have to be carefully considered by public health officials and researchers.

CHAPTER 2. THE AIMS OF THE STUDY

The general aim of this research was to investigate genetic diversity and molecular phylogeny of circulating HIV strains in Serbia based on samples collected from 2008 to 2013 and compares it to previous data in order to better understand local HIV epidemic spread.

Specific aims of the study included:

1. To describe and characterize current prevalence and distribution of subtypes and circulating recombinant forms in Serbia.
2. To investigate and characterize local transmission networks as well as to phylogenetically estimate the time of the most recent common ancestor (tMRCA) of the main HIV-1 clades in Serbia.
3. To analyse the prevalence and patterns of viral DNA sequence changes as markers of duration of infection and host-virus interaction.

CHAPTER 3. MATERIALS AND METHODS

3.1. STUDY DESIGN AND ETHICAL APPROVAL

The study was designed to be cross-sectional, molecular and phylogenetic investigation of HIV-1 diversity. It was conducted at the Institute of Microbiology and Immunology, of the School of Medicine, University of Belgrade. Before the start of the research, the protocol was approved by the Ethical Comitee of the School of Medicine, University of Belgrade. Patients were informed of the objectives of the study and participants were included upon informed consent. The proposed eligibility criteria for participant selection included: HIV-1 seropositive adults (men and non-pregnant women) aged 18 years or above, both treatment naive and treatment experienced (patients on HAART) and viral load higher than 1000 copies per ml of blood.

Epidemiological, clinical and virological data were collected using a comprehensive standardized questionnaire. Transmission risk was categorized as men who have sex with men (MSM), heterosexual, intravenous drug use (IDU), transfusion, vertical transmission or unknown. For patients reporting IDU in addition to another risk, the former was considered as main risk for acquiring HIV infection.

3.2. STUDY SUBJECTS AND SAMPLE COLLECTION

The study included 155 consenting, HIV infected patients, both male and female, who fulfilled the inclusion criteria of the study. Blood samples used in the study were collected from 2008 to 2013 from patients who were referred to the Center for HIV/AIDS, University Hospital for Infectious and Tropical Diseases, Clinical Centre of Sebia, Belgrade. In addition, 162 HIV-1 sequences from Serbian patients, retrieved from

the NCBI database, sampled and deposited in the period from 1997 to 2007, were also included.

Using a sterile syringe and a needle, ten-milliliter of venous blood samples were collected in Ethylenediaminetetraacetic acid (EDTA) tubes Vacutainer (BD) and immediately transported at 4°C. These samples were transported to the Institute of Microbiology and Immunology, Virology Department, for processing. Upon arrival, peripheral blood mononuclear cells (PBMC) and plasma was separated by centrifugation at 5000 rpm (~ 370 × g) for 15 minutes and preserved at -80 °C for further molecular analyses.

Within this study a forensic analysis was performed in order to infer genetic relatedness and explore suspected epidemiological linkage between HIV genome sequences isolated from three HIV-1 infected patients. Three patients, fulfilling the inclusion criteria: a man (subject 1) and two women (subject 2 and subject 3), all coming from the capital city of Belgrade, were HIV diagnosed at the Center for HIV/AIDS, University Hospital for Infectious and Tropical Diseases in Belgrade, Serbia, in August and September 2011. *A priori* information were as follows: (i) subject 1 and subject 2 had been married for over 15 years with two kids aged 10 and 14, both found to be HIV negative; (ii) subject 3 was sexual partner of subject 1 and together they had several prolonged sojourns in Thailand in years preceding HIV infection diagnosis of all three patients; (iii) finally, subject 1 sued subject 3 for knowingly infecting him, without disclosing her HIV status.

3.3. RNA EXTRACTION FROM PLASMA SAMPLES

RNA was extracted from plasma samples using QIAamp Viral RNA Mini Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions.

Prior to RNA extraction a total of 1.5 mL of each plasma sample was centrifuged for 1.30 h at 4 C at 22,000 g. The supernatant was carefully removed to the volume of 140 µl, and the pellet was resuspended, and used for RNA-extraction.

Briefly, 140 µl of plasma was put into 560 µl of RNA viral lysis buffer (Buffer AVL) containing carrier RNA and vortexed for 15 seconds and incubated for 10 minutes at room temperature. This mixture was briefly centrifuged and 560 µl of ethanol was added, pulse vortexed and centrifuged. Approximately 630 µl of the solution was applied to QIAamp spin column in 2 ml collection tube without wetting the rim and centrifuged for 1 minute at 8000x rpm. The collection tube containing filtrate was discarded. To this column, 500 µl of wash buffer 1 (Buffer AW1) was then added and centrifuged for 1 minute at 8000x rpm. Further, the column was placed in a clean 2 ml collection tube and filtrate discarded. About 500 µl of wash buffer 2 (Buffer AW2) was added to the collecting tube, cap was closed, and centrifuged at maximum speed (14500x rpm) for 3 minutes. This procedure was repeated at the same speed for 1 min in order to remove any wash buffer carryovers. The QIAamp spin column was placed in a clean 1.5 ml microcentrifuge tube and 60 µl of elution buffer was added. This was incubated at room temperature for 1 minute and finally centrifuged at 8000x rpm for 1 minute to elute the RNA. The yielded viral RNA was stored at -80°C for later use in Reverse transcriptase polymerase chain reaction (RT-PCR).

3.4. DNA EXTRACTION FROM PBMCs

After the separation of PBMCs, DNA extraction was done using QIAamp DNA Blood Mini kit (Qiagen, Hidden, Germany), as per manufacturer's

protocol. Briefly, 20 µl of QIAGEN Protease (proteinase K) was pipetted into a sterile microfuge tube and 200 µl of samples was added to the same tube. Further 200 µl of the lysis buffer (Buffer AL) was added to each sample and mixed thoroughly by pulse-vortexing for 15 sec. The tube was then incubated at 56°C for 10 minutes to lyse the cells. In the following step 96% ethanol was added to the sample, and mixed again by pulse-vortexing for 15 sec. This mixture was further carefully applied on the top of the QIAamp Mini spin column (in a 2 ml collection tube), and centrifuged at 6000 x g (8000 rpm) for 1 min. To this column, 500 µl of wash buffer 1 (Buffer AW1) was then added and centrifuged for 1 minute at the same at the same speed as in the preceding step. Further, the column was placed in a clean 2 ml collection tube and filtrate was discarded. About 500 µl of wash buffer 2 (Buffer AW2) was added on the top of the QIAamp Mini spin column, cap was closed and centrifuged at maximum speed (14500x rpm) for 3 minutes. The QIAamp Mini spin column was placed in sterile 1.5 ml microfuge tube and the collection tube containing the filtrate was discarded. The QIAamp spin column was opened carefully and 50 µl of elution buffer (Buffer AE) was added. The tube was incubated at room temperature for 1 min, and then centrifuged at 6000 x g (8000 rpm) for 1 min. Extracted DNA was stored in AE buffer (provided by QIAamp DNA Mini Kit) at - 20°C until further usage.

3.5. NESTED POLYMERASE CHAIN REACTION

The starting template for genotypic reverse transcriptase polymerase chain reaction (RT-PCR) was RNA extracted from plasma, for both *pol* and *env* regions. All reagents were thawed and put on ice. The list of PCR primers used in this study are indicated in the **Table 5**. For the first round of the PCR assays (outer PCR), RNA was reverse transcribed using a One Step RNA PCR Kit

(Qiagen, Hilden, Germany). Reactions were set up on ice in a UV-treated laminar flow hood. Reaction mix and cycling parameters were typically as per manufacturer's instructions although optimization of annealing temperature was occasionally necessary.

Standard conditions were:

- 8 μ l of 5x OneStep RT-PCR Buffer
- 1,6 μ l of dNTP mix
- 1.6 μ l of Fw primer in 0.6 μ M final concentration (Invitrogen by Life Technologies, Carlsbad, California, USA) (**Table 5.**)
- 1.6 μ l of Rev primer in 0.6 μ M final concentration (Invitrogen by Life Technologies, Carlsbad, California, USA) (**Table 5.**)
- 15.6 μ l of Rnase-free water
- 10 μ l of isolated RNA template.

Final reaction volume was 40 μ l.

Prior to PCR lyophilized primers were resuspended in sterile distilled water to a stock concentration of 100 pmol/ μ l. Primers used for amplifications were equilibrated to working concentrations of 20 pmol/ μ l. Primers were stored at -20 °C until required.

In the thermal cycler "Eppendorf Mastercycler ep gradient S" samples were initially reverse transcribed at 50°C for 40 minutes and then denatured at 94°C for 15 minutes. After an initial denaturation step the cycling begins with a short denaturation at 94°C for 30 sec the primers were annealed for 1 minute at an appropriate temperature, according to the melting temperature of the primers and the template, followed by an extension time of 3 minute

at 72 °C. This cycle was repeated 35-40 times and followed by a final elongation step of 10 minutes at 72 °C.

First round of one step RT-PCR assay was followed by second round (inner PCR), using Thermo scientific dream taq PCR master mix (2X) (Applied Biosystem, Foster City, California). In summary, 5 µl product from first round PCR was used as a starting template for a second round.

General protocol was as follows:

- 25 µl of DreamTaq PCR Master Mix (2X)
- 1 µl of Fw primer in 0.6 µM final concentration (Invitrogen by Life Technologies, Carlsbad, California, USA) (**Table 5.**)
- 1. µl of Rev primer in 0.6 µM final concentration (Invitrogen by Life Technologies, Carlsbad, California, USA) (**Table 5.**)
- 18 µl of nuclease free water
- 5 µl of outer DNA template

Final reaction volume was 50µl.

In the thermal cycler “Eppendorf Mastercycler ep gradient S” after an initial denaturation step on 95°C for 3 min, the cycling begins with a short denaturation at 95°C for 30 sec, the primers were annealed for 45 sec at an appropriate temperature, according to the melting temperature of the primers and the template (**Table 5.**), followed by an extension time of 3 minute at 72 °C. This cycle was repeated 40 times and followed by a final elongation step of 10 minutes at 72 °C.

Table 5. Table of primer pairs used in nested PCR and cycle sequencing reaction with melting temperature (T_m) of each primer calculated by T_m calculator (Snoeck et al., 2005; McGovern et al., 2010)

	<i>pol</i> gene	T _m	<i>env</i> gene	T _m
OUTER PRIMERS (5'-3' direction)	GGTAAATAAAATAGTAAG	55°C	GAGCCAATCCCATAACATTATGT	58°C
	ATGTTTTACATCATTAGTGTG	53°C	GCCCATGTGTTCTGCTGCTCCCAAGAACC	60°C
INNER PRIMERS (5'-3' direction)	GTACTACTAGAAGAAATGATGAC	58°C	TGTCCCCAGCTGGTTTGGCAT	60°C
	CCGATAAATTGATATGICCATIG	56°C	TATAATCACTTCTCCAATGTCC	57°C
SEQUENCING PRIMERS (5'-3' direction)	GGCAAATACTGGAGTATTGTATGCA	55°C	AATGTCAGYACAGTACAATGTACAC	55°C
	CCTGTCAACATAAT	55°C		
	TACTAGGTATGGTAAATGCAGT	55°C		
	CAGTACTGGATGTGGGTGATG	55°C	GAAAAATCCCTTCCACAATTTAAA	53°C

3.5.1 END POINT LIMITING DILUTION POLYMERASE CHAIN REACTION (EPLD - PCR)

Phylogenetic analyses of query viral sequences for forensic purpose included viral RNA extracted from 140 µL of plasma using the QIAamp Viral RNA Kit (Qiagen, Hilden, Germany), as described in detail above. Two step RT PCR was performed on a 20-µL volume containing 10 µL of eluted RNA using Qiagen Long range 2-step RT PCR kit (Qiagen, Hilden, Germany). EPLD-PCR is an assay that is based on serial dilutions of nucleic acid (template for PCR) to the point that precedes the dilution that gives a negative PCR result. Conditions can

be manipulated to access different subpopulations of HIV quasispecies, thus, improving the range of the quasispecies detection.

After reverse transcription of viral RNA, complementary DNA (cDNA) was diluted by end point limiting-dilution (EPLD) procedure and further PCR amplified using primer pairs as described in (Table 5). For three query samples a minimum of five repeats of end point three-fold limiting-dilution PCR reactions (EPLD-PCR) were performed, starting from reverse transcription obtained cDNA. Extraction of proviral DNA, was also included of this part of research based on manufacturer's protocol as described above. Prior to nested PCR assay, DNA eluate was diluted by EPLD procedure.

3.6. AGAROSE GEL ELECTROPHORESIS

Conventional agarose gel electrophoresis was used to analyze all PCR products. In order to analyze PCR products of *pol* and *env* gen, 2 % agarose was prepared. All agarose gels were stained with ethidium bromide stock solution (1 µg/ml) for visualization of the DNA bands. Ethidium bromide is a powerful mutagen and should be handled carefully. TAE (Tris-acetate-EDTA) buffer was used as both a running buffer and in agarose gel (Sambrook et al., 1989). TAE buffer, is commonly prepared as a 50X stock solution for laboratory use. This stock solution can be diluted 50:1 with water to make a 1X working solution. This 1X solution will contain 40mM Tris, 20mM acetic acid, and 1mM EDTA. Samples were mixed with 2 µl of the gel loading dye (0.125% Bromophenol blue, 40% Sucrose) and approximately 8 µl were the loaded per lane on agarose gel. The Gene Ruler™ 1 kb DNA ladder (DNA Standard 100bp - Serva Electrophoresis GmbH, Heidelberg,) was used as a molecular length size marker. Electrophoresis reactions were run at 120 V. The DNA bands were

visualized on a UV transilluminator under a UV light with a wavelength between 280 and 320 nm.

3.7. CYCLE SEQUENCING REACTION

3.7.1 PURIFICATION OF PCR PRODUCTS

To dispose of excess dNTPs, enzymes and buffers, amplified PCR products, which showed a clear band on the agarose gel, were purified with MinElute Purification Kit (Qiagen, Hilden, Germany) kit, using the manufacturer's instructions. The purification protocol based on silica membrane spin was used, which allows binding of nucleic acids to a silica membrane inside a spin column (Sambrook *et al.*, 1989).

General protocol was as follows

- Firstly, 200 μ l of Buffer PB and 40 μ l product of nested PCR reaction were roughly mixed and applied to MinElute column and centrifuge for 1 min at 18000 x g. Flow-through was discarded and the MinElute column was put back into the same collection tube.
- Secondly, 750 μ l Buffer PE was to the MinElute column and centrifuge for 1 min at 18000 x g. Once again . Flow-through was discarded and the MinElute column was put back into the same collection tube. The column was centrifuged for an additional 1 min at maximum speed.
- Lastly, MinElute column was placed in a clean 1.5 ml microcentrifuge tube. 10 μ l Buffer EB (10 mM Tris \cdot Cl, pH 8.5) was added to the center of the membrane, and let the column stand for 1 min, and then centrifuge for 1 min at 18000 x g.

3.7.2 CHAIN TERMINATION SEQUENCING REACTION

Purified PCR products were subjected to direct sequencing of both the sense (forward) and antisense strands (reverse). DNA sequencing was carried out using the Sanger cycle sequencing method. Sanger sequencing is based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during *in vitro* DNA replication (Sanger et al, 1977). Chain termination sequencing reactions were set up in house, using BigDye Terminator v3.1. Cycle Sequencing kit (Applied Biosystems Incorporated, Foster City, CA) according to manufacturer's instructions. Fluorescently labeled dyes are attached to ACGT extension products in DNA sequencing reactions. The dyes come in four colors red (labels Thymidine base), blue (Cytosine), black (Guanine) and green (Adenine). The dyes are incorporated using either 5'-dye label primers or 3'-dye label dideoxynucleotide terminators. For each sample six separate sequencing reactions were performed using respectively the two inner PCR primers and four additional internal pol primers described in the **Table 5**.

General protocol was as follows:

- 1µl each primer 5 µM working concentration
- 1µl of purified RT-PCR product,
- 2µl 5× Cycle sequencing dilution buffer
- 2µl BigDye
- 4µl ultrapure water.

Final reaction volume was 10µl

The sequencing PCR included 40 cycles of 96 °C for 30 s, 50 °C for 7 s and 60°C for 4 min. After running the sequencing reaction, non-incorporated dideoxynucleoside triphosphates were removed by 75% isopropyl alcohol precipitation and the pellet was resuspended in 20µl High Density Formamide (Applied Biosystems Incorporated, Foster City, CA) for denaturation and detected in an ABI Prism 310- Genetic Analyzer capillary electrophoresis system (Applied Biosystem, Foster City, CA, USA). Results were analyzed with the Sequencing analysis software v.5.2 (Applied Biosystem, Foster City, USA).

3.8. SEQUENCE DATASETS

Analyses were performed on HIV sequences obtained from blood samples of 155 patients included in the study, collected from 2008 to 2013. Additionally, in this doctoral dissertation, 162 HIV-1 sequences from Serbia retrieved from the NCBI database, sampled in the period from 1997 to 2007, were also included. Accession numbers of the local sequences retrieved from NCBI database are listed in **Table AI, Appendix 1**. Furthermore, a total of 250 HIV-1 sequences sampled worldwide and downloaded from the NCBI database (<http://www.ncbi.nlm.nih.gov/nuccore>), was also included in the phylogenetic analyses within this research. Criteria for inclusion of sequences were clearly identified subtype for each sequence, clearly established origin of each corresponding patient. Furthermore, for the tMRCA phylogenetic analyses clearly defined collection data of samples was also included, as an inclusion criteria. The country distribution and the accession numbers of all foreign sequences are listed in the **Table AI** shown in the **Appendix 1**.

Within the scope of transmission clusters analyses, three viral sequences obtained from a man and two women, included in this study as previously

described, were phylogenetically analyzed for forensic investigation of suspected HIV-1 transmission case. In addition to viral sequences obtained from plasma samples, these three query viral sequences that were under forensic investigation were obtained from peripheral blood mononuclear cells (PBMC). From proviral DNA as starting template nested PCR was performed for both *pol* and *env* gene.

For this purpose, 34 sequences of *env* gene region sampled from patients diagnosed in the time frame of two years preceding and two years after the time of diagnosis of the three query patients were also included in phylogenetic analysis as local controls.

Furthermore, query sequences were used in the NCBI blast analysis (<http://blast.ncbi.nlm.nih.gov>) to search for the best-matching HIV-1 sequences according to score significance. The purpose was to identify additional highly similar sequences not included in the sampling. Ten sequences with the highest similarity scores to each query sequence were retrieved. To maximize the number of unrelated but high-scoring HIV-1 sequences from GenBank, those repeatedly selected were included only once, resulting in 30 GenBank sequences, from different geographical regions, as background controls. Intermixing of the best-matching HIV-1 foreign sequences downloaded using BLAST, with the ones under investigation would increase the probability of refuting the a priori hypothesis that case samples were involved in an alleged HIV-1 transmission chain. In view of possible epidemiological linkage to Thailand, 20 subtype B viral sequences in both *pol* and *env*, sampled from infected individuals in Thailand were included in the study, also retrieved through the BLAST search with query sequences using "Thailand" as an Entrez query to limit search. The final dataset analyzed consisted of 101 local and 50 foreign sequences as controls for *pol* and 34 local and 50 foreign controls for *env*. Complete list of control sequences with countries of origin and NCBI accession numbers is found in the **Table AII** shown in the **Appendix 2**.

3.9. PHYLOGENETIC ANALYSES

HIV sequences obtained during this study together with sequences downloaded from NCBI database were converted to appropriate format (FASTA, NEXUS) and analyzed using various phylogenetic software packages, depending on specific research questions.

The multiple sequence alignment is the starting point for any phylogenetic analyses and it has to be done before a phylogenetic tree can be inferred. The accuracy of the alignment generated will directly affect the quality of any inference of phylogenetic history. The most frequently used software, implementing the algorithm for multiple alignment is Clustal series of programs.

Herein, Clustal W implemented in Molecular Evolutionary Genetics Analyses (MEGA version 6.0) software, was used for alignment of all nucleotide sequences obtained during the study (Tamura et al., 2011). It is phylogenetic based sequences alignment program that use progressive alignment methods. Firstly, a pairwise distance matrix for all the sequences to be aligned has to be generated, as well as „guided tree“ using the neighbor-joining algorithm (described in detail below). The guide tree serves as a rough template for clades that tend to share insertion and deletion features. Then, each of the most closely related pairs of sequences the outermost branches of the tree is aligned to each other using dynamic programming.

When editing the alignment, we kept the coding information for each sequence, using the HXB2 coding sequence for the same genomic region as a reference. Insertions relative to the HXB2 reference, as well as ambiguous and gapped regions, were removed from the alignment

Once an alignment has been constructed the relationship between the sequences can be derived from their evolutionary distances to each other. First, JmodelTest was used to carry out statistical selection of best-fit models of nucleotide substitution. This software implements five different model selection strategies: hierarchical and dynamical likelihood ratio tests (hLRT and dLRT), Akaike and Bayesian information criteria (AIC and BIC), and a decision theory method (DT). Herein, the best fitting nucleotide-substitution model was general time reversible model of nucleotide-substitution with a proportion of invariant sites (i) and gamma distribution of rates (C) (GTR + I + G), selected according to the Akaike Information Criterion (AIC) using all 88 proposed models. Once a model of nucleotide evolution has been selected it can be used in the inference of a tree with increased accuracy in relation to the genetic distances derived from the alignment. The evolutionary distances were computed using the Maximum Composite Likelihood method and are in the units of the number of base substitutions per site (Tamura et al., 2011).

After an alignment has been generated, appropriate model of sequence evolution has been selected and evolutionary distances were computed a phylogenetic tree can be inferred. Phylogenetic tree building methods can broadly be classified as distance based and criterion-based or algorithmic.

During the course of this research three different tree building methods, neighbor joining, maximum likelihood and Bayesian methods were used (**Table 6**). Neighbor joining is algorithmic distance matrix method that cluster taxa according to a pre-defined set of rules. Maximum likelihood and Bayesian methods are criterion-based methods search for the best tree according to some criterion. Algorithmic methods will return a single tree based upon a series of operations, whereas the search for an optimal tree in criterion-based methods means that multiple trees are considered and the 'best' tree is reported.

Table 6. Phylogenetic tree building methods. In bold are the ones that were used in this study

	TREE BUILDING METHOD	DESCRIPTION
DISTANCE- BASED METHODS	Unweighted Pair Group Method with Arithmetic Means (UPGMA)	Phylogenetic tree is built based on pairwise distance matrix. Method assumes that the rate of evolution is linear
	Neighbour Joining (NJ)	Pairwise distances can be adjusted through application of appropriate evolutionary model and tree is constructed based on pairwise matrix
CRITERION- BASED METHODS	Maximum likelihood (ML)	Based on specific model of nucleotide substitution the method calculates the likelihood of all possible trees for the specified alignment, and selects the one associated with the maximum likelihood.
	Bayesian	Evolutionary model is defined by user in order to calculate posterior probabilities of all possible trees for the specified alignment, and selects the one with the highest probability.
	Parsimony	Tree constructed on the basis that the minimum number of evolutionary differences between sequences is the most likely.

In this thesis, the MEGA software was used to conduct distance-based (neighbor-joining) phylogenetic analyses, and PAUP software was used to construct ML trees. MrBayes and BEAST were used for Bayesian inference of phylogenies (Guindon et al., 2010; Tamura et al., 2011; Drummond and Rambaut, 2007; Ronquist and Huelsenbeck, 2003). Bootstrapped replicates of phylogenies were sampled to assess support for clades. The method of bootstrapping has been applied in phylogenetic analysis to allow uncertainty in phylogenetic reconstructions to be assessed and it is performed in order to infer the reliability of branch orders (Efron et al., 1996; Felsenstein, 1985). For an alignment of sequences where the rows represent different taxa and the columns are sites along the genome, columns of sites are randomly sampled with replacement to create a new alignment of the same size as the original. For the random sample a tree is then inferred. This resampling process is usually repeated many times (often 1000 times). The branch order on the correct tree is then compared to the random trees. In general the more random samples that support a given branch order on the tree the more reliable that branch order is. The minimum cut off for reliability is normally about 90%. Obtained phylogenetic trees were visualized and edited using the FigTree program version 1.3.1 (<http://www.tree.bio.ed.ac.uk/software/figtree/>)

3.9.1. HIV SUBTYPING

Obtained sequences were visually inspected, manually edited and then assembled with SeqScape HIV-1 Genotyping System Software v 2.5 (Applied Biosystem, Foster City, CA, USA). Firstly, analysis of viral sequences was performed using the REGA HIV-1 subtyping tool version 3 (REGAv3) (<http://regatools.med.kuleuven.be/typing/v3/hiv/typingtool/>) and the Los Alamos HIV database (<https://www.hiv.lanl.gov/content/index>). The Rega subtyping tool is based on phylogenetic analysis in order to take into account

the epidemiological and evolutionary relationships among subtypes (Robertson et al., 2000).

Subtyping of all *pol* gene sequences collected during the course of this research was further performed by construction ML and NJ phylogenetic tree, under appropriate models, using the PAUP and MEGA software, with reference sequences of different subtypes, downloaded from HIV-1 Los Alamos Database (LANL, www.hiv.lanl.gov). Together with the *pol* sequences obtained in this research, subtype identification was also performed for 162 NCBI retrieved sequences dating from 1997-2007. Sequences that were not unambiguously classified by REGA subtyping tool or that were assigned to a different subtype from the one specified in the Los Alamos database were removed from the data set.

3.9.2. PHYLOGENETIC ANALYSES OF TRANSMISSION CLUSTER

The identification of transmission clusters, defined as viral lineages derived from the same variant, was accomplished through analysis of a series of criteria sets (**Figure 11**). In this part of research total of 304 *pol* gene sequences together with 250 background sequences were included.

Before starting this investigation, from all full codon alignment of sequences, 40 codons associated with major resistance in PR (30, 32, 46, 47, 48, 50, 54, 58, 74, 76, 82, 83, 84, 88, 90) and RT (41, 62, 65, 67, 69, 70, 74, 75, 77, 100, 101, 103, 106, 108, 115, 116, 151, 181, 184, 188, 190, 210, 215, 219 y 225) were removed (Johnson et al., 2013; Wensing et al., 2014).

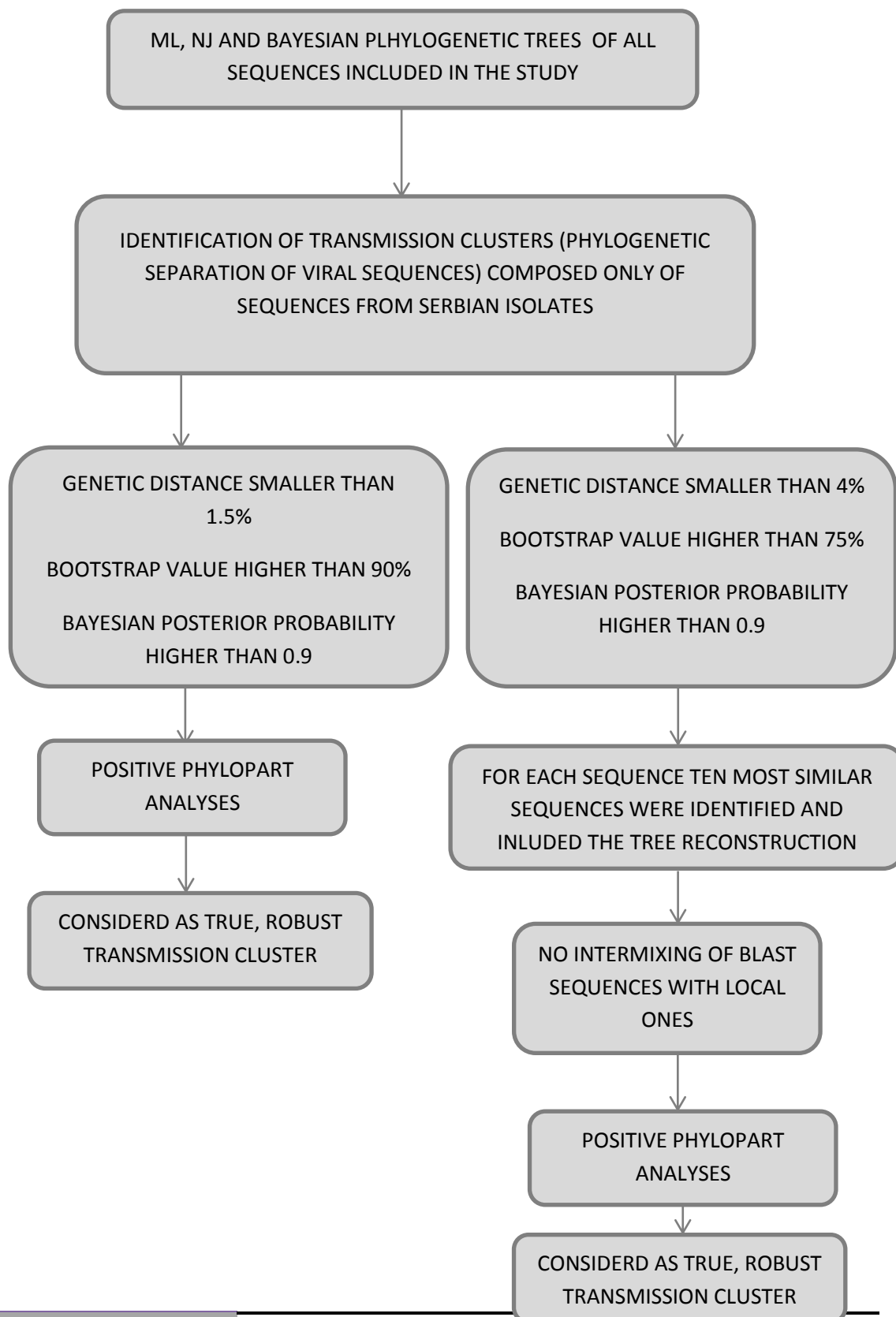
According to the first set of criteria, transmission clusters were assigned as those phylogenetic clades consisting of three or more sequences, fulfilling the conditions of genetic distance of 1.5% or less, with minimal bootstrap support

of 90%. Additionally, Bayesian inference were performed in MrBayes v3.1.2 with the GTR + G + I model of nucleotide substitution (Ronquist F, and Huelsenbeck J, 2003). In this analysis, two Monte Carlo Markov chains (MCMCs) are run simultaneously for the number of generations needed for a stationary distribution to be maintained long enough after convergence. Typically, the number of generations was 4×10^6 to 5×10^6 , with the initial 10% of these generations discarded as burn-in. Only those clades with statistical support of posterior probability higher 90% in the Bayesian analysis were considered as transmission cluster.

Second criteria sets included those clades with bootstrap support over 75% and with genetic distance of less than 4% were further analyzed. For every sequence belonging to such cluster, ten most similar sequences were identified using BLAST analysis. (<http://blast.ncbi.nlm.nih.gov>) and included in tree reconstruction as previously described. In the posterior analyses, only those clades with no intermixing of foreign sequences downloaded using BLAST, with the local ones, were considered as transmission clusters (Yebra et al., 2013).

Lastly, the presence of transmission clusters was assessed using PhyloPart, a novel phylogeny based software for the identification of transmission clusters, under following parameters: threshold 0.05 and sample distance limit 4 (Prosperi et al., 2011). Only those clades which accomplish all conditions from the first or second set of criteria and were further confirmed by PhyloPart were considered as transmission clusters and selected for further analysis.

Figure 11. Phylogenetic analyses and sets of criteria for identification of transmission clusters



3.9.3. ESTIMATING TIME OF THE MOST RECENT COMMON ANCESTOR

Molecular clock analysis was performed using a Bayesian MCMC coalescent method, as implemented in BEAST v1.8.1 (Drummond et al., 2003; Drummond and Rambaut, 2007). We used GTR nucleotide substitution model with six category gamma distributed rate variation among sites and two partitions in the codon positions (the best fitting model for all three data sets according to jModelTest). As the different HIV-1 genes showed different rates of evolution. Preliminary analysis included combinations of two relaxed molecular clock models, 'Relaxed: exponential' and 'Relaxed: log-normal' in combination with four different coalescent tree priors: constant size, exponential growth, logistic growth and Bayesian skyline tree priors. Bayes factor analysis, performed in Tracer v.1.5, showed that the lognormal relaxed clock with the Bayesian Skyline was the best model, as indicated by a log Bayes Factor >10 according to Tracer v1.5 (<http://beast.bio.ed.ac.uk/Tracer>) – evidence of very strong statistical support (Suchard et al., 2001). The Bayes Factor represents the ratio of the marginal likelihoods (approximated using the harmonic mean estimator) of the two models. Therefore in subsequent analyses we used uncorrelated log normal relaxed clock model and Bayesian skyline with 10 numbers of groups (Drummond and Rambaut, 2007). Rates were estimated taking into account the known sampling time of the sequences. A log-normal prior was placed on the ucl.d.mean parameter (Hue et al., 2005; Zehender et al., 2010). A tree search was carried out running a Markov chain Monte Carlo (MCMC) sampler for 5×10^6 generations, sampling every 5000th generation. Convergence of the Markov chain was assessed by program Tracer v 1.5 (<http://tree.bio.ed.ac.uk/software/tracer/>) calculating the effective sample size (ESS) for each parameter. Tracer software analyses posterior samples of continuous parameters from Bayesian MCMCs to allow visual inspection of the chain behavior, estimating of the ESS of parameters and the

plotting of marginal posterior densities. Herein, ESS values higher than 100 were considered robust. The ESS is the number of independent samples that would be the equivalent to the autocorrelated samples produced by the MCMC. This provides a measure of whether the chain has been run for an adequate length (for example, if the effective sample sizes of all continuous parameters are greater than 200). Tree samples were used to generate a maximum clade credibility tree (MCCT) after a 10% burn-in, using TreeAnnotator v 1.6.1 (<http://beast.bio.ed.ac.uk/TreeAnnotator>).

3.10. MOLECULAR FOOTPRINTS ANALYSES

A bioinformatic approach was used to estimate the duration of infection by calculating the fraction of ambiguous nucleotides in the sequence as a delimiter for more recent (less than 1 year) versus chronic infection (longer than 1 year). Ambiguous mutations, representing a mixed nucleotide signal, were identified when the sequencing signal intensity of the minor base in both directions was $\pm 20\%$ of the major base signal at a particular position. The method was applied using HIV sequences of newly diagnosed patients, to both the complete dataset and subtype B sequences only. In estimating duration of infection three different ambiguity cutoff values were evaluated (0.45%, 0.47%, and 0.5%) for recent infection.

Amino acid (aa) substitutions at position 245 of HIV-1 reverse transcriptase (RT) have been described to be associated to the presence of human leukocyte antigen HLA-B*5701 allele in the host, in particular in subtype B infection. Therefore, this aa substitution can be used as a marker to identify individuals who may be safely treated with Abacavir. Furthermore, viral diversity in individual patients is reflected in the proportion of ambiguous

bases observed in HIV-1 pol sequences, including RT codon 245, obtained from population based genotyping with increasing diversity seen with increased duration of infection.

3.11. STATISTICAL ANALYSES

The results were analyzed by standard statistical analysis. Categorical data were compared using the chi-square test and Fisher's exact test. To better understand the temporal trends of the epidemiology of HIV-1 infection in Serbia we analyzed trends in the major epidemiological parameters across two time periods: 1997-2007 (162 sequences existing in the GenBank database,) in comparison to 2008-2013 (sequences generated in this study, isolated from 155 patients).

CHAPTER 4. RESULTS

4.1. STUDY POPULATION

The majority of patients included in the research, were newly diagnosed 62.5% (97/155), partly within the European project for monitoring of primary HIV resistance SPREAD/EuropeHIVResistance, while 37.5% (58/155) patients were on treatment. The overall population was 76% men, with over half of the total number of patients living in the urban areas and majority of patients reporting to be infected in Serbia. Information on transmission route was available for 139 patients, no risk factor was known for 16 patients. The majority of patients in the study were infected through sexual contact (97.4%). Regarding overall study population, MSM was the highest reported risk for infection in 69.6% of all study patients (108/155), followed by heterosexual contact in 18% (28/155). Intravenous drug use was reported by a very small proportion of the population, 1.9% (3/155). Regarding transmission risk among patient in this research there was a significant difference in the prevalence of all STDs among MSMs compared to patients infected heterosexually (31% versus 11%, $p = 0.0287$).

General demographic and clinical data of the study population are shown in **Table 7**.

Table 7. Basic demographic and epidemiological characteristic of the study population

Table 7.	
Characteristics of the study population	
Gender	Patients
Male	118 (76%)
Female	37 (24%)
Transmission route	
MSM	80 (51.6%)
Heterosexual	55 (35.4%)
IVDU	4 (2.5%)
	16 (10.3%)
Place of residence	
Urban area	132 (59.7%)
Rural	66 (29.9%)
Unknown	23 (10.4%)
CDC disease stage	
A	60 (38.7%)
B	25 (16.1%)
C	35 (22.5%)
Unknown	35 (22.5%)

The prevalence of transmission by men having sex with men (MSM) was much higher in the second half of the study period (39% in 1997–2007 compared to 77% in 2008–2013, $p < 0.0001$), whereas the prevalence of heterosexual transmission and the risk of transmission by intravenous drug use declined. The percentage of diagnosis in the earlier disease stage (CDC stage A) tended to increase in the second half of the study period and this increase was shown to be statistically significant (34% vs. 61%, $p = 0.0012$). The prevalence of sexually transmitted diseases (STDs), taking into account hepatitis B virus (HBV) as a sexually transmitted infection, was significantly increased in the second half of the study period (17% vs. 38%, $p < 0.001$).

General demographic, epidemiological and clinical characteristic of the patients included in this study compared to data on 162 sequences retrieved from NCBI database are shown in **Table 8**.

Table 8. General demographic, epidemiological and clinical characteristic of the patients included in this study compared to data on 162 sequences retrieved from NCBI database, sampled in the period 1997 to 2007. Marked in red are characteristics with higher prevalence found in the second period (2008-2013, found to be statistically significant). Marked in blue are features with statistically higher prevalence found in the first period (1997-2007).

Table 8. General epidemiological and clinical data of the overall population		
	1997-2007	2008-2013
NUMBER (percentage) OF PATIENTS		
	162(51.1%)	155 (48.9%)
GENDER		
<i>Male</i>	101 (62.3%)	118 (76.1%)
<i>Female</i>	61 (37.6%)	37 (23.8%)
TRANSMISSION ROUTE		
<i>MSM</i>	45 (27.7%)	108 (69.6%)
<i>Heterosexual</i>	69 (42.5%)	28 (18%)
<i>IVDU</i>	17 (10.4%)	3 (1.9%)
<i>Transfusion</i>	9 (5.5%)	0
<i>Unknown</i>	22 (13.5%)	16 (7.6%)
MEDIAN AGE		
<i>Male</i>	35.19±11.07	33.19±8
<i>Female</i>	34.19±9.7	30.1±8.7
PLACE OF RESIDENCE		
<i>Urban</i>	98 (60.4%)	112 (72.2%)
<i>Rural</i>	64 (39.5%)	43 (27.8%)
CDC DISEASE STAGE		
<i>A</i>	42 (25.9%)	85 (54.8%)
<i>B</i>	20 (12.2%)	25 (16.1%)
<i>C</i>	72 (44.4%)	35 (22.5%)
<i>Unknown</i>	28 (17.2%)	10 (6.4%)

4.2. HIV SUBTYPING

During the course this research, from 155 samples included in research, 142 *pol* HIV sequences were successfully obtained, yielding full-length protease (PR) and more than 250 reverse transcriptase (RT) codons and 34 *env* HIV sequences, as well.

The obtained sequences analyzed in this research were deposited in the GenBank database under the accession numbers: KF157435-KF157549, KX944763-KX944796.

Rega subtyping tool and phylogenetic analysis, performed with Paup and MEGA software, of all 142 HIV-1 *pol* gene sequences obtained with a full length during the research gave congruent results in terms of subtype assignment.

Results of this research showed that among HIV-1 infected patients in Serbia subtype B predominates 90.8% (129/142), while the prevalence of non B subtypes was 9.2% (13/142). Among non B subtypes, subtype C was found with the highest prevalence, in 3.4% (5/142) samples, followed by subtype A that was found in 2% (4/142). Circulating recombinant forms (CRFs) were detected in 2.8% (4/142) of the collected samples (**Figure 12.** and **Table 8**).

In addition, results on prevalence of HIV-1 subtypes obtained in this research was compared to the one obtained from 1997 to 2007, in order to analyse statistically significant differences between to time period (**Table 8**.)

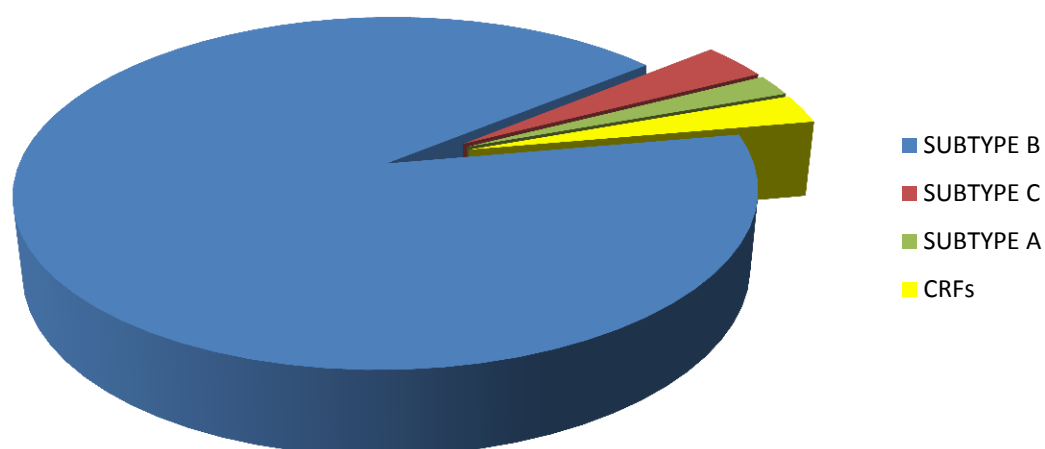


Figure 12. HIV-1 subtypes distribution from 2008-2013.

Table 8. Changes in the distribution of subtypes in the two periods with similar viral sequences that were generated (1997-2007 vs 2008-2013). Statistically significant differences are marked in red

	PREVALENCE OF SUBTYPES FROM 1997 - 2007	PREVALENCE OF SUBTYPES FROM 2008 - 2013
SUBTYPE B	90.1%	90.8%
SUBTYPE C	3.1%	3.5%
SUBTYPE G	3.7%	/
SUBTYPE A	/	2.1%
CRFs	3.5%	2.8%

4.3. PHYLOGENETIC IDENTIFICATION OF TRANSMISSION CLUSTERS

Phylogenetic analyses of transmission clusters included all 142 *pol* gene sequences of full length successfully obtained during the course of this research, 162 *pol* viral sequences from local isolates obtained from 1997 to 2007 and a number of sequences from foreign isolates sampled worldwide, retrieved from NCBI database.

Phylogenetic analysis of total 304 *pol* sequences revealed the presence of 14 transmission clusters and 12 transmission pairs within the subtype B sequences (accounted for 275 sequences) included in the study (**Figure 13**). Considerable proportion, 57.4% (158/275) of local sequences were found intermixed in the phylogenetic tree with sequences sampled across Europe and North America, indicative of multiple subtype B introductions. However, 24.7% of analyzed subtype B sequences (68/275), were found grouped in 14 clusters that accomplished predefined sets of criteria (**Figure 13.**, marked in dark blue). Eight clusters composed of 48 sequences in total were defined by the first, most stringent criteria set only, while 5 clusters were further added by second criteria set, including one well defined cluster, included in the forensic investigation. Furthermore, additional 17.5% (48/275) sequences were found to form a well-supported separate phylogenetic clade (bootstrap support of 97%), albeit with the maximum pairwise genetic distance of 7.9% (range 0–7.9% SD 1.2 average 2.3%) (**Figure 13.**, marked in blue). The corresponding patients were diagnosed in a twelve year period (2002–2013) and the majority of them were male 96% (46/48) and MSM 83% (40/48). Despite the high bootstrap value, maximal genetic distance of this clade exceeded the predefined ranges for transmission cluster, however, in view of the time-span of sampling within the clade and uniformity of transmission route, we considered it as a transmission network. Three of fourteen identified transmission clusters were part of this large

network. Thus, in total, 42.2% (116/275) of analyzed subtype B sequences may be considered to be part of transmission clusters/network.

Transmission clusters ranged from 3 to 13 sequences with the mean number of patients per cluster 3.7, with collection dates within the span of 1–9 years (mean 3.8, SD 2.2) and median pairwise genetic distance of 0.9% (range 0–3.5%). Majority of clustering sequences were from patients living Belgrade 66.1% (45/68), while 11/14 identified transmission clusters, containing 53 sequences, were composed of viral sequences isolated from male patients (except for single sequence from a female patient) that predominantly reported MSM risk behavior (75.4%, 40/53). Remaining 3/14 transmission clusters included sequences from patients reporting heterosexual contact as risk for acquiring HIV infection, that were all living in Belgrade. Of note, these three heterosexual clusters contained significantly higher number of sequences isolated from women compared to predominantly MSM clusters ($p < 0.001$). We did not identify any transmission cluster encompassing sequences isolated from IDUs.

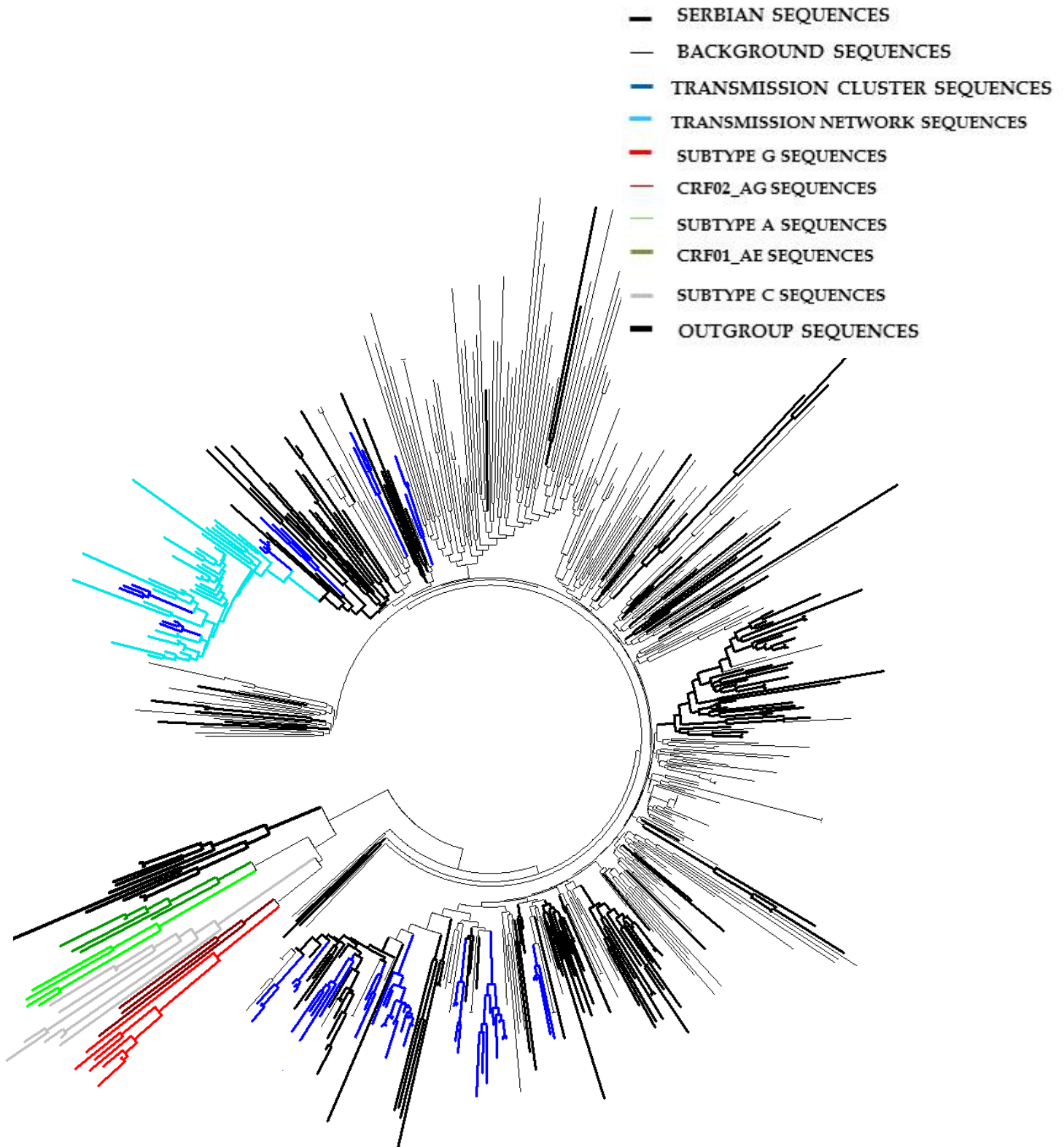


Figure 13. ML phylogenetic tree obtained by MEGA software v 6.0 under GTR + G + I nucleotide substitution model. Local *pol* gene sequences generated in this research, together with Serbian *pol* sequences generated from 1997 to 2007, and background sequences retrieved from NCBI were included.

4.3.1. FORENSIC APPLICATION OF PHYLOGENETIC ANALYSES

HIV sequences of both *pol* and *env* genes were successfully obtained from all three query subjects, originating from both plasma RNA and PBMCs obtained DNA. A total of twelve sequences was deposited in the GenBank under the accession numbers: KF157495.1, KF157498.1, KF157497.1, KX944760, KX944761, KX944762, KX944797 - KX944802.

The results of phylogenetic analysis of the two analyzed genomic regions (*pol* and *env*) were consistent, by all the applied methods. The three viral sequences isolated from patients whose epidemiologically linkage was under investigation were shown to form a well-supported separate transmission cluster that accomplished all predefined sets of criteria (**Figure 14. a**, for *pol* and **b** for *env* gene). Phylogenetic tree showed no intermixing of local or foreign sequences into the transmission cluster. The cluster contained no viral sequences from Thailand. Phylogenetic analyses of HIV-1 *pol* and *env* sequences isolated from the three patients and a local population sample of HIV-1-positive individuals showed that sequences under investigation were more closely related to one another, than to any control sequences. The mean pairwise nucleotide divergence among three *pol* sequences included in the transmission cluster was 0.8% (0.3-1.1%), that is almost ten times less than mean pairwise nucleotide divergence among observed local controls *pol* sequences (7.1%). Regarding the *env* gene, known for its higher genetic variability, mean nucleotide divergence among three query sequences was 1.3% with maximum nucleotide divergence among them of 1.5% - in comparison to all observed local control sequences with mean nucleotide divergence of 16.1% and maximum nucleotide divergence of 26.3%.

Each of the above described sequences from query samples was used as a starting point for EPLD-PCR analysis. A total of twelve sequences per patient per genetic region were generated and included in the phylogenetic analysis.

The end point limiting-dilution that resulted in positivity varied between the samples, from 81-fold to 243-fold. Phylogenetic trees based on the EPLD-PCR obtained viral sequences also contained the cluster specific for the three query patients with no intermixing of other viral sequences (**Figure 15.** and **Figure 16**).

4.3.2. PARAPHYLETIC RELATION BETWEEN QUERY SEQUENCES

In this research, paraphyletic relationship between sequences of subjects 2 and 3 and those of subject 1 was found, suggestive of the hypothesis of subject 1 being the source of infection for both subjects 2 and 3. This finding would be in opposition of the *a priori* hypothesis of HIV transmission from subject 3 to subject 1, but would still correlate to the epidemiological *a priori* information. This relation was consistent when separately analyzing query subjects in pairs, as described previously (Romero-Severson et al., 2016). According to this approach, tree topologies of the EPLD-PCR based phylogenetic trees analyzing pairs of query subjects, revealed paraphyletic-monophyletic (PM) topology when sequences of Subject 1 were analyzed with either sequences of Subject 2 and 3, in both *env* and *pol* genes (**Figure 17** and **Figure 18**) . This result is consistent with the transmission scenario of Subject 1 being the source for both Subjects 2 and 3 (Romero-Severson et al., 2016). When analyzing sequences of Subjects 2 and 3, the obtained topology corresponded to dually monophyletic (MM) in the *env* region, implying the common source transmissions, whereas in the *pol* region the obtained topology corresponded to a combination of paraphyletic and polyphyletic (PP), which might be the consequence of a short time period between the two transmission events.

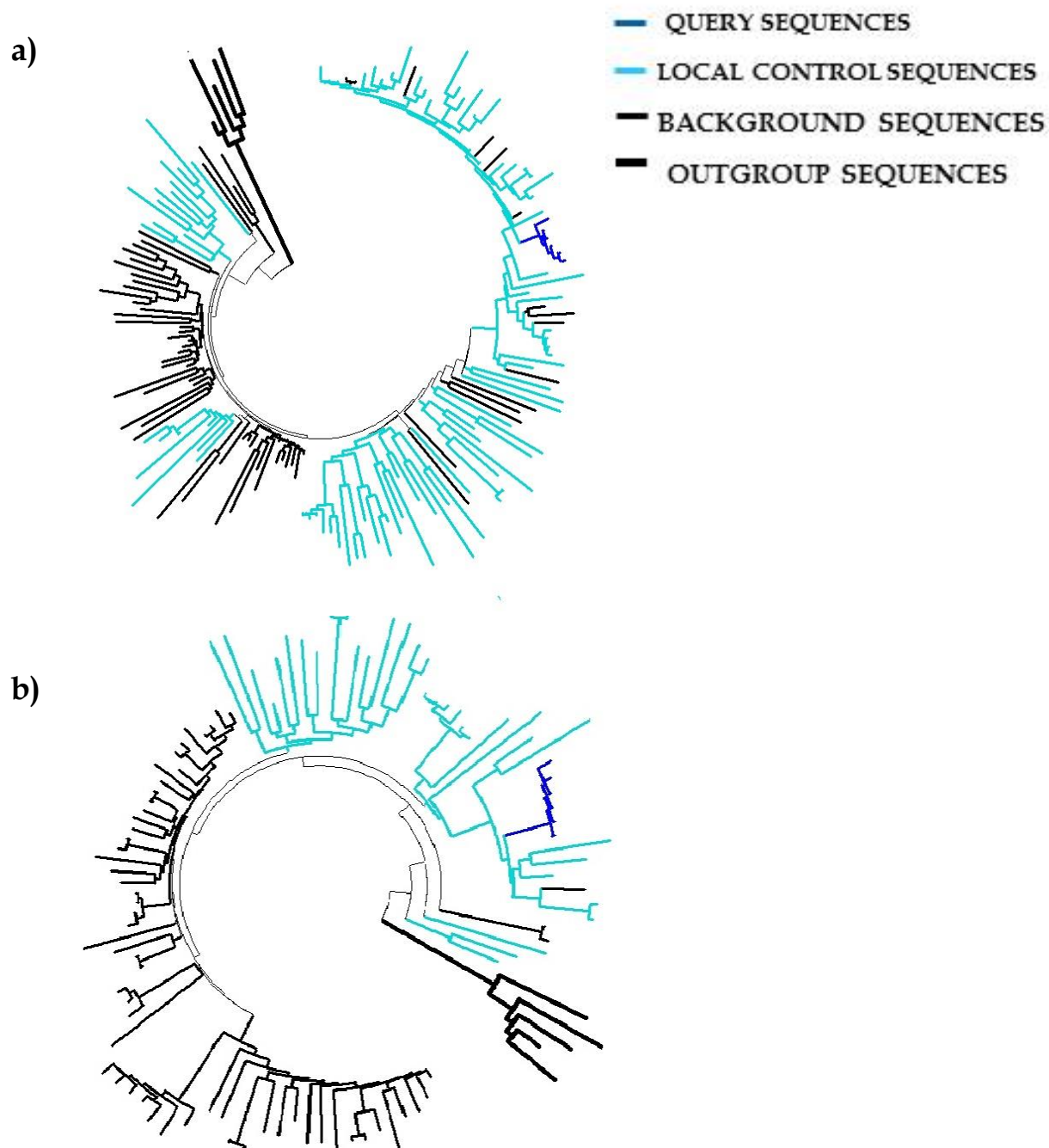


Figure 14. ML phylogenetic tree of

a) *pol* gene sequences

b) *env* gene sequences; under GTR+G+I substitution model obtained by MEGA software v. 6.0. Query sequences (from proviral DNA and RNA), 34 local control sequences together with background sequences (foreign sequences) obtained by BLAST were included.

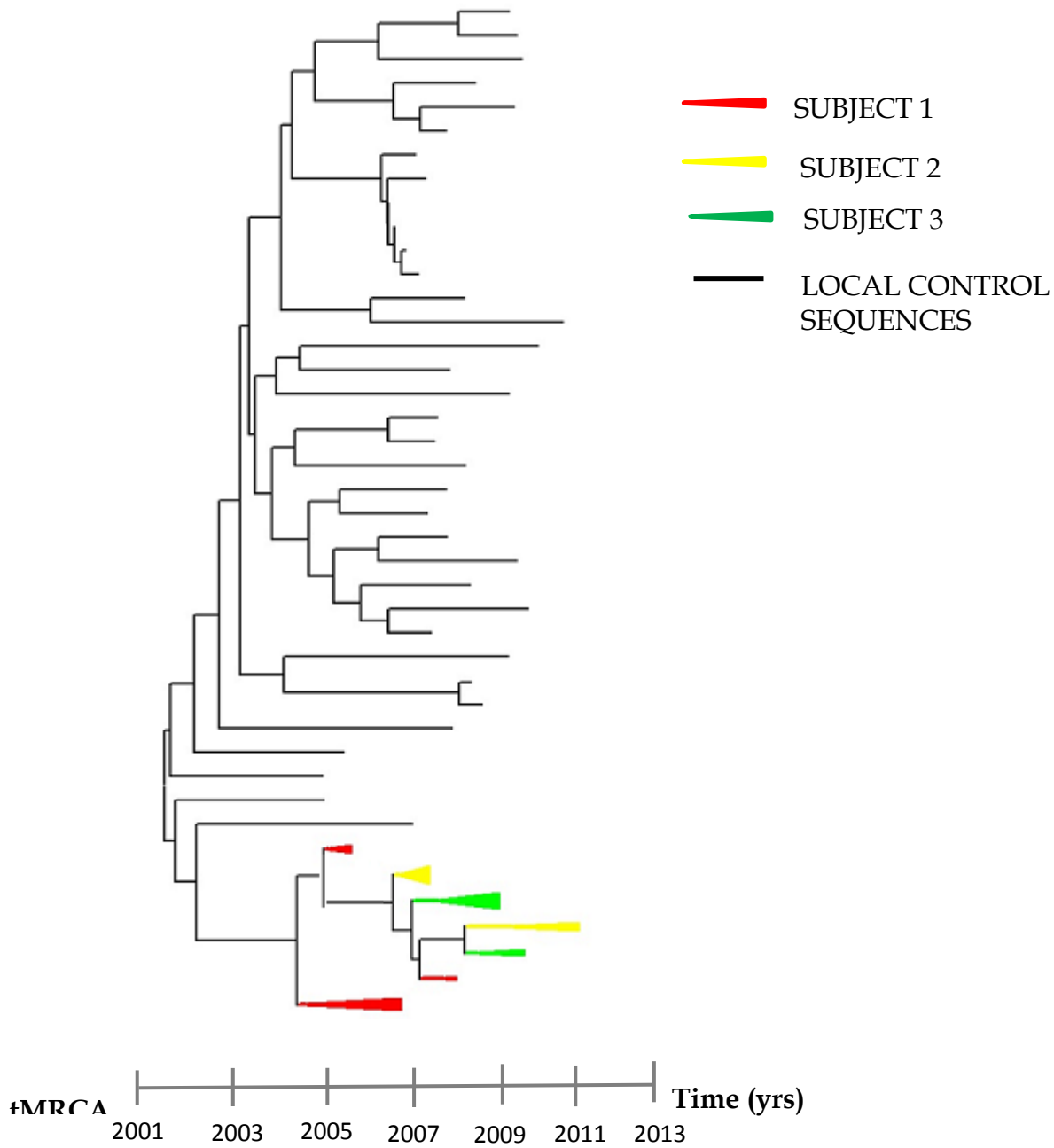


Figure 15. ML phylogenetic tree obtained by MEGA software v 6.0 including EPLD query sequences of *pol* gene and local control *pol* gene sequences.

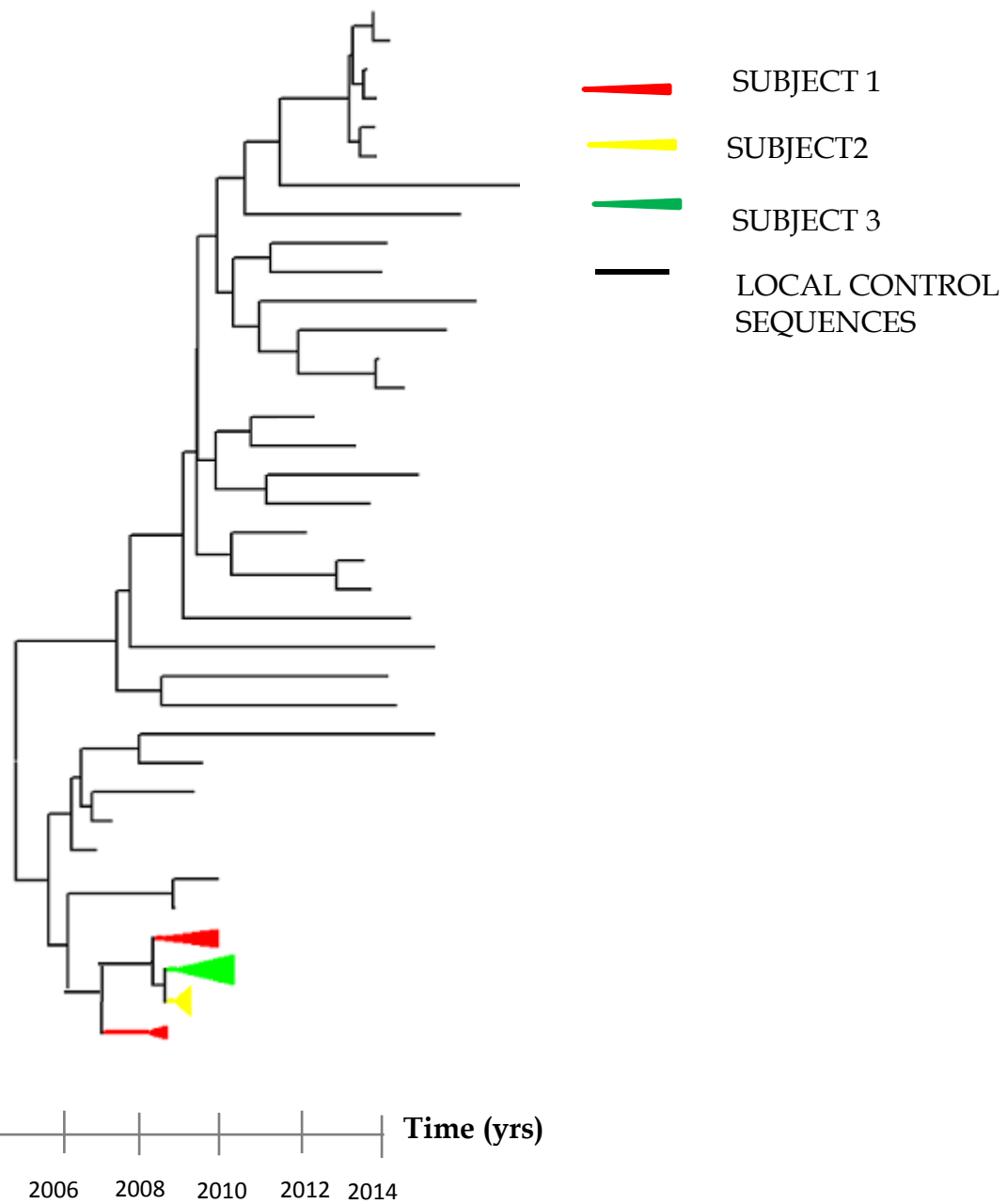


Figure 16. ML phylogenetic tree obtained by MEGA software v 6.0 including EPLD query sequences of *env* gene and local control *env* gene sequences.

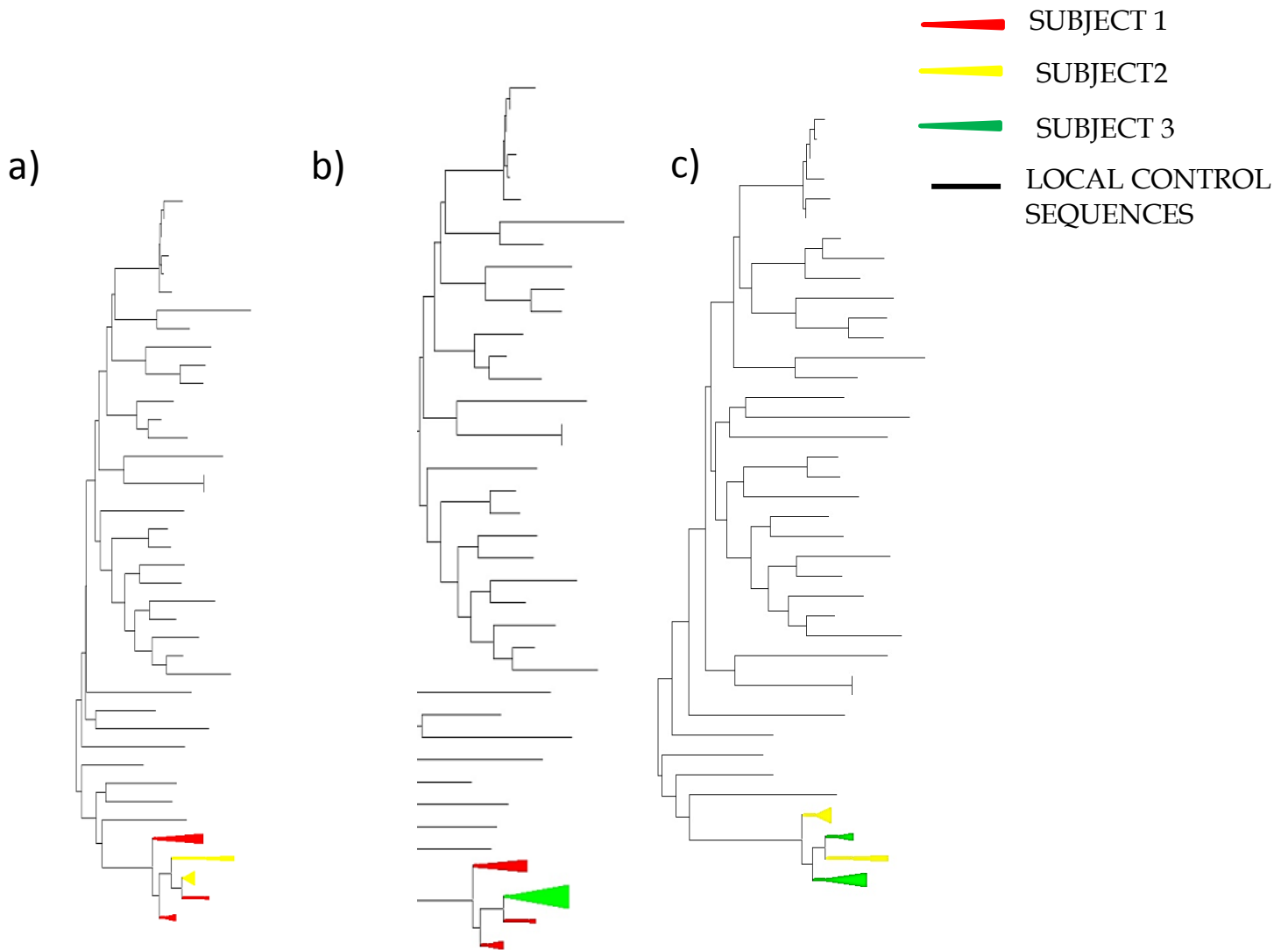


Figure 17. ML phylogenetic trees based on the EPLD-PCR obtained *pol* sequences of pairs of subjects: a) subject 1 and subject 2; b) subject 1 and subject 3; c) subject 2 and subject 3; together with local control sequences. Clusters of query subjects' sequences are shown in color.

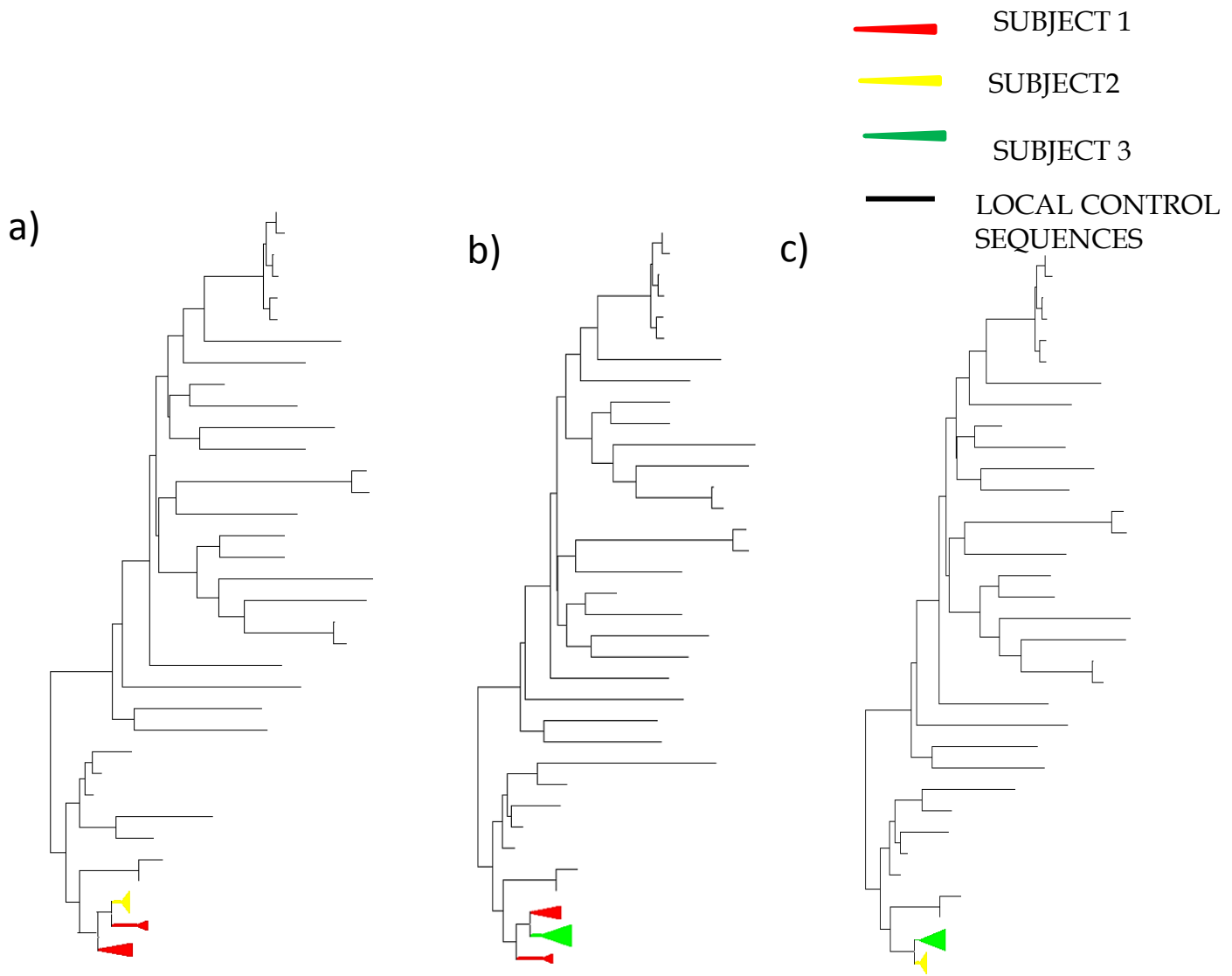


Figure 18. ML phylogenetic trees based on the EPLD-PCR obtained env sequences of pairs of subjects: a) subject 1 and subject 2; b) subject 1 and subject 3; c) subject 2 and subject 3; together with local control sequences. Clusters of query subjects' sequences are shown in color.

4.4. TIMING THE ORIGIN OF THE MAIN CLADE

The time of the Most Recent Common Ancestor (tMRCA) was estimated, using the Bayesian Markov chain Monte Carlo (MCMC) method implemented in BEAST v1.8.2., for all local subtype B sequences, for the transmission network of subtype B sequences and for the most extended transmission cluster, identified in the previously explained phylogenetic analysis (13 sequences) (**Figure 19.**, **Figure 21.**, respectively).

The median tMRCA of the all local subtype B sequences was dated to 1983 (95% HPD: 1977-1989) (**Figure 19.**).

The tMRCA inferred for local transmission network composed of 48 viral sequences was in 1994 (95% Higher Posterior Density HPD: 1982–2000). Estimated temporal origin for the local subtype B transmission cluster, composed of 13 viral sequences, with bootstrap support of 100%, was much more recent, 2004 (95% HPD: 2002–2006) (**Figure 22.**).

Furthermore, temporal origin was also estimated for G clade and subtype C clade of Serbian sequences, as the most prevalent non B subtypes in the overall period (**Figure 22.**, **Figure 24.**, respectively). Considering estimated temporal origin of subtype G in Serbia it was found to be in the early nineties, 1991 (95% Higher Posterior Density HPD: 1983–1999). For all local subtype C sequences, median tMRCA was found to be similar as for subtype G sequence with similar lower and upper credibility limits, 1990 (95% Higher Posterior Density HPD: 1984–1995).

Bayesian skyline plots analyses showed an increase of subtype B sequences over the study period, contrary to the trend observed for subtype G and subtype C which showed limited dispersal over the study period (**Figure 20.**, **Figure 22.**, and **Figure 23.**, in comparison to **Figure 25.** and **Figure 27.**).

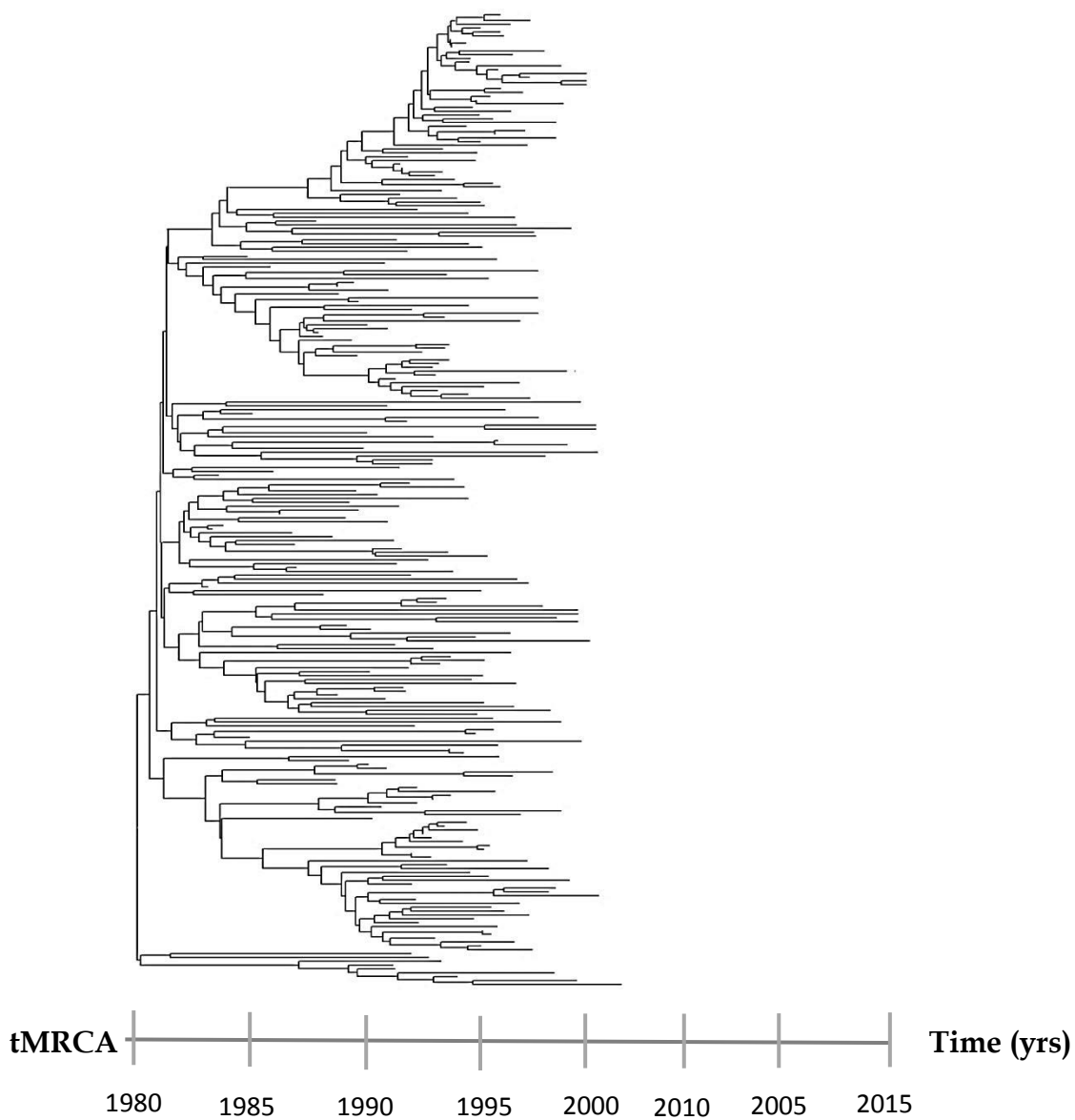


Figure 19. Time-scaled maximum clade credibility tree of all local subtype B pol sequences analyzed with BEAST, as obtained with tip dating. Branch lengths represent years before the last sampling time.

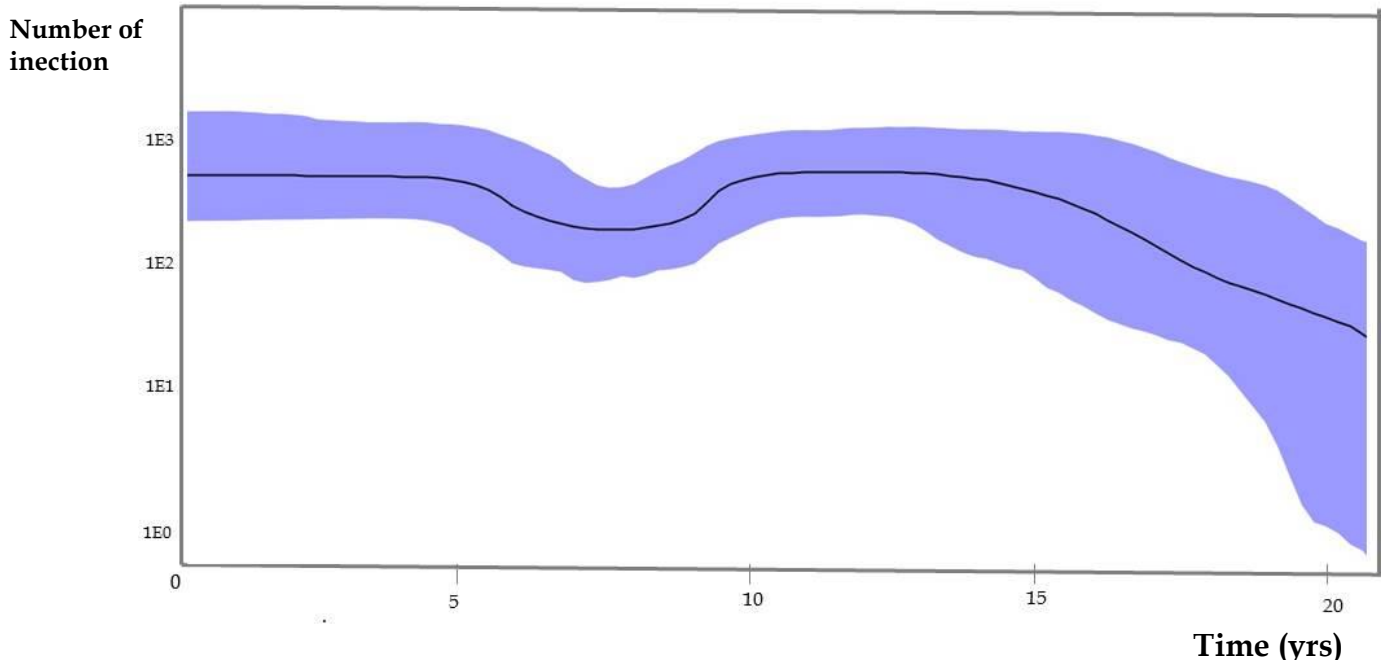


Figure 20. Bayesian skyline plots representing the estimates of the effective number of subtype B sequences in the studied population. The y-axis measures the effective number of infections in log10 scale while the x-axis represents time in years with 0 (zero point) indicating year when the last sequence of subtype B was sampled (2013) followed with years indicated period in past. The line showing the median estimate of effective number of infections over time and blue coloured areas limiting the 95% HPD interval.

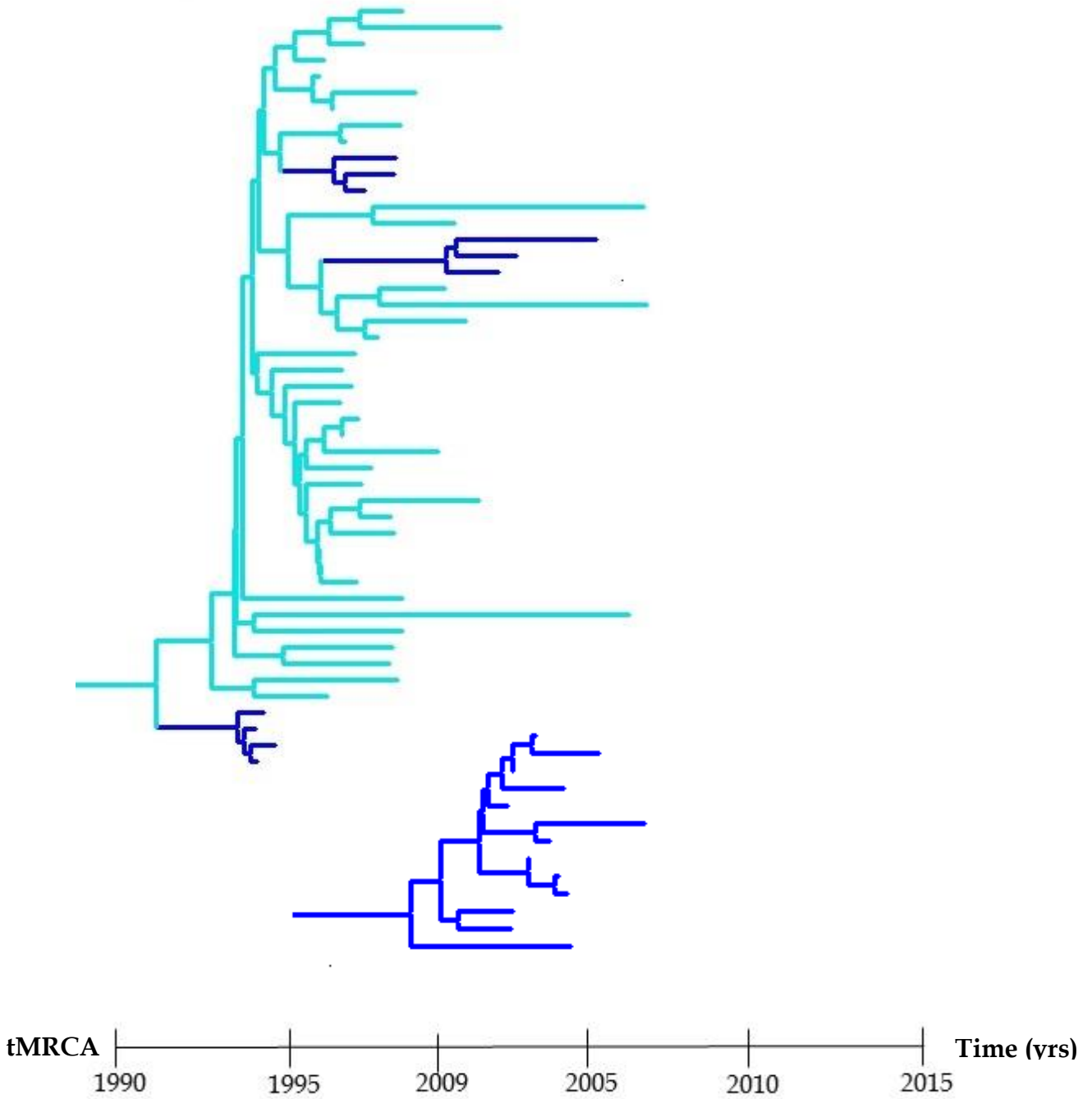


Figure 21. Estimation of the tMRCA, obtained with BEAST v 1.4 using tip dating, of subtype B sequences within transmission network marked in blue (sub-clusters are marked in dark blue), and the most extended transmission cluster (with 13 sequences), marked in dark blue. Branch lengths represent years before last sampling time.

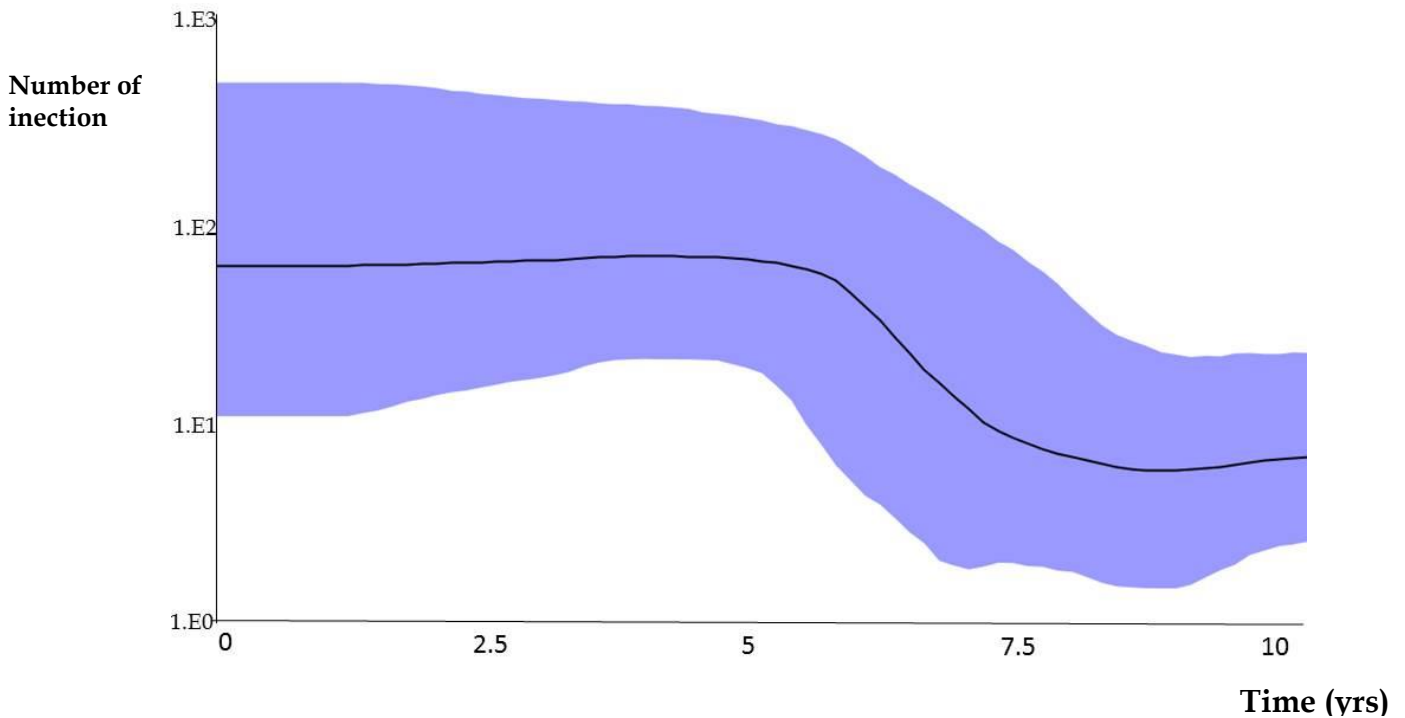


Figure 22. Bayesian skyline plots representing the estimates of the effective number of sequences within well defined transmission network in the studied population. The y-axis measures the effective number of infections in log₁₀ scale while the x-axis represents time in years with 0 (zero point) indicating year when the last sequence of subtype B was sampled (2013) followed with years indicated period in past. The line showing the median estimate of effective number of infections over time and blue coloured areas limiting the 95% HPD interval.

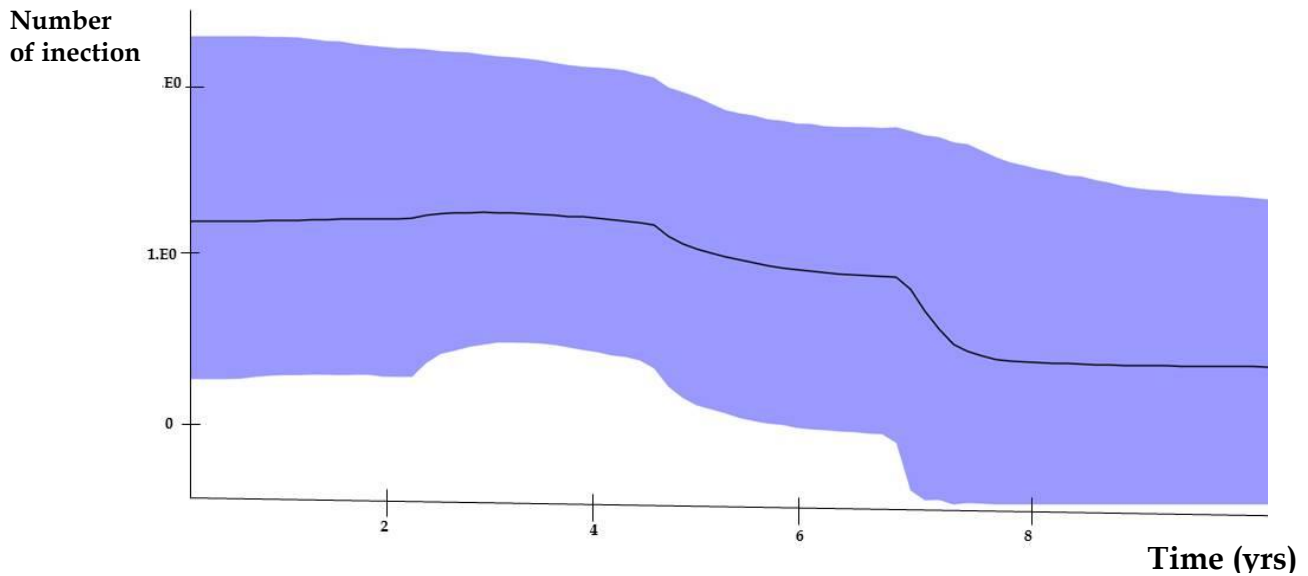


Figure 23. Bayesian skyline plots representing the estimates of the effective number of 13 sequences within transmission cluster, supported by bootstrap of 100%, in the studied population. The y-axis measures the effective number of infections in log10 scale while the x-axis represents time in years with 0 (zero point) indicating year when the last sequence of subtype B was sampled (2013) followed with years indicated period in past. The line showing the median estimate of effective number of infections over time and blue coloured areas limiting the 95% HPD interval.

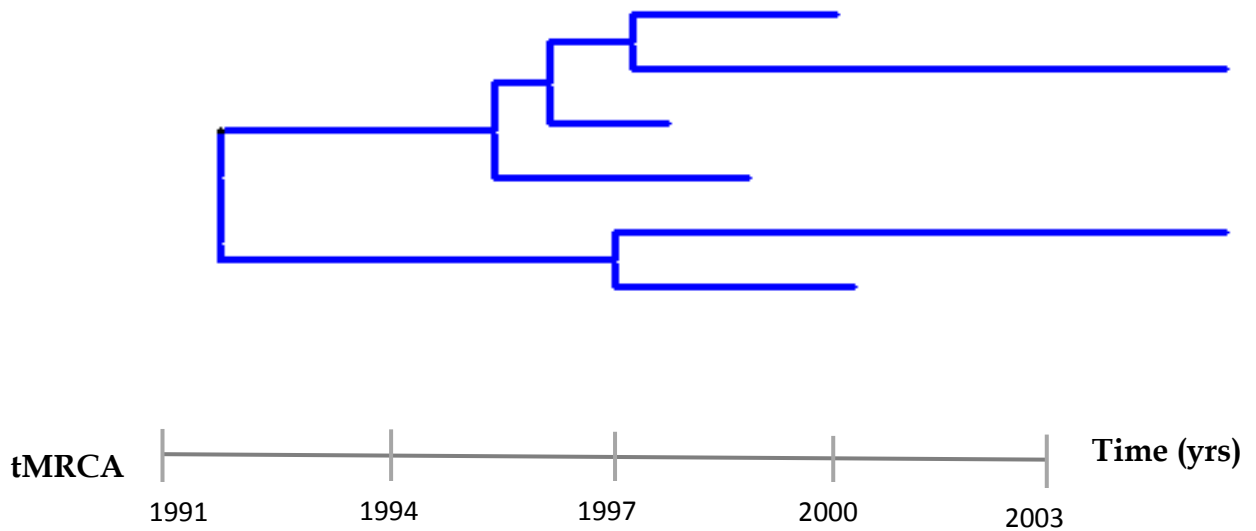


Figure 24. Time scaled maximum clade credibility tree estimated using all Serbian subtype G sequences of pol region, analyzed with BEAST v1.8.1, as obtained with tip dating. The scale at the bottom of the tree represents the time in years before the last sampling time.

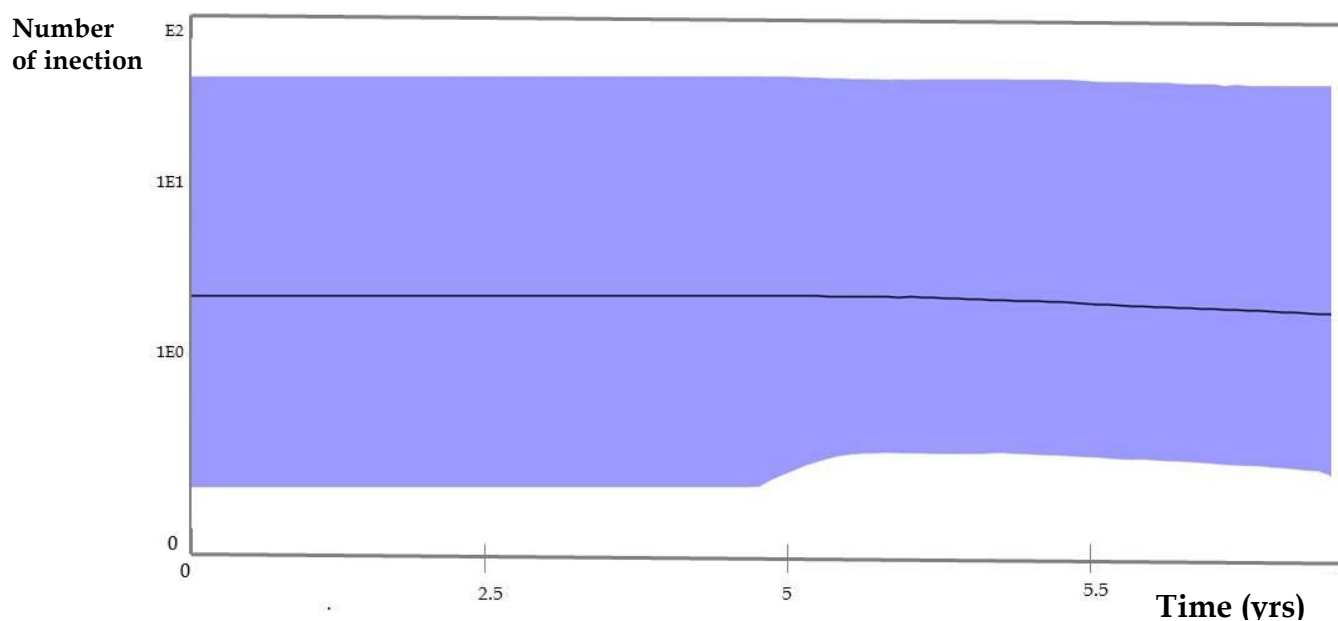


Figure 25. Bayesian skyline plots representing the estimates of the effective number of subtype G sequences in the studied population. The y-axis measures the effective number of infections in log₁₀ scale while the x-axis represents time in years with 0 (zero point) indicating year when the last sequence of subtype G was sampled (2004) followed with years indicated period in past. The line showing the median estimate of effective number of infections over time and blue coloured areas limiting the 95% HPD interval.

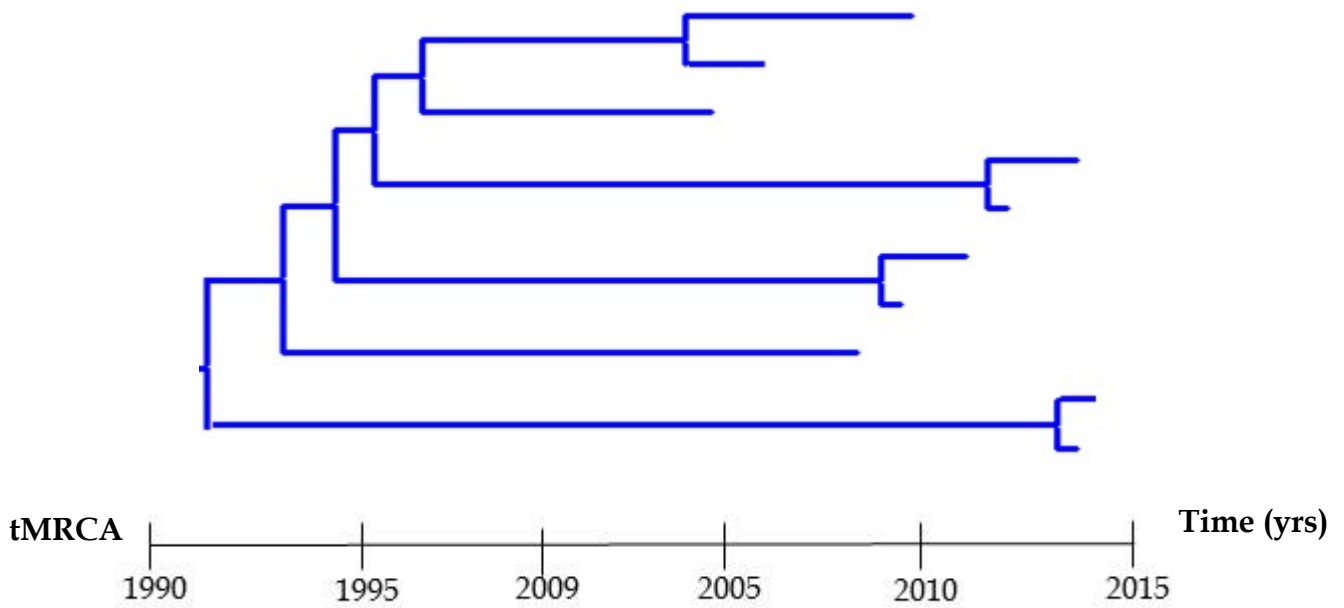


Figure 26. Time scaled Bayesian MCMC tree estimated using all Serbian subtype C sequences of pol region, analyzed with BEAST v1.8.1, as obtained with tip dating. The scale at the bottom of the tree represents the time in years.

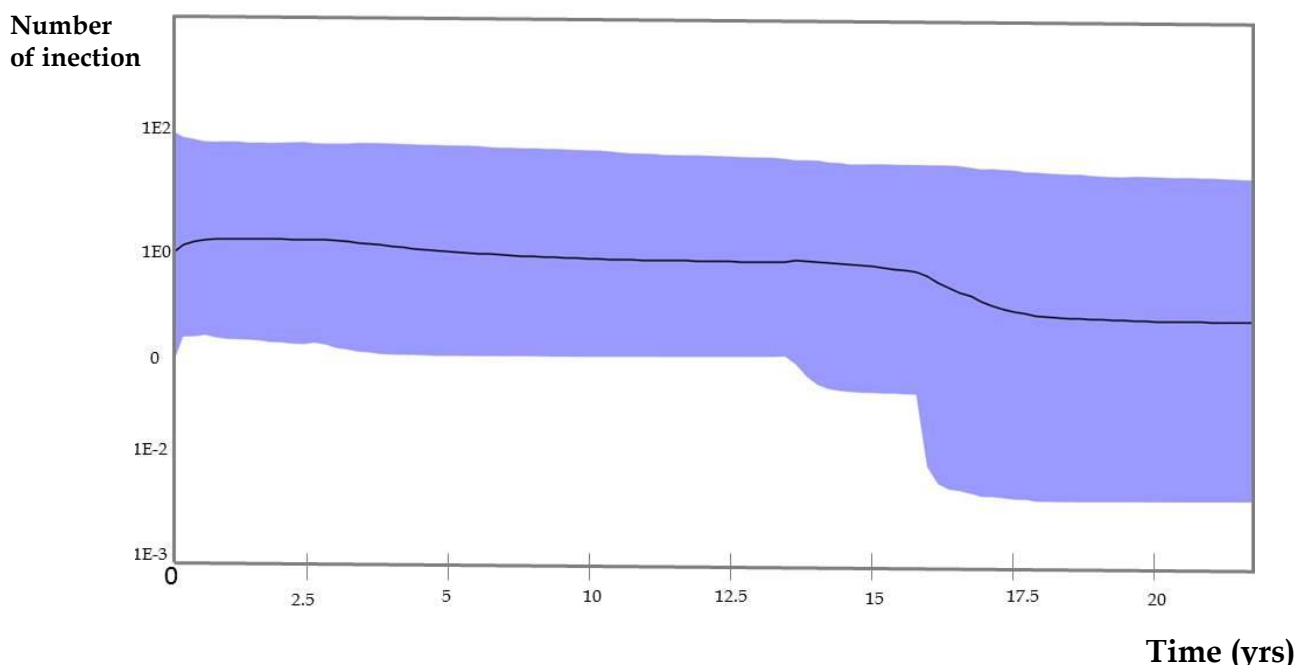


Figure 27. Bayesian skyline plots representing the estimates of the effective number of subtype C sequences in the studied population. The y-axis measures the effective number of infections in log₁₀ scale while the x-axis represents time in years with 0 (zero point) indicating year when the last sequence of subtype C was sampled (2013) followed with years indicated period in past. The line showing the median estimate of effective number of infections over time and blue colored areas limiting the 95% HPD interval.

4.5. THE PREVALENCE OF MOLECULAR FOOTPRINTS ON HIV-1 SEQUENCES ON AND THEIR ASSOCIATION WITH DURATON OF INFECTION

In this part of research, all *pol* viral sequences obtained only from therapy naive patients, included in this study were included. together with 110 sequences sampled from 1997 to 2007, were analyzed.

Based on the threshold of 0.47% ambiguous bases per sequence, a total of 55.1% of samples (114/207) were classified as a recent infection, of duration of less than 1 year, whereas among subtype B samples this percentage was 54% (58/180). Lowering the threshold to 0.45% did not influence the result, nor did raising it to 0.5% (except in the subtype B dataset, where two additional samples were identified as recent infection with the least stringent cutoff of 0.5%). In the first half of the study period (1997–2007) the percentage of recent infection was 35.4% (39/110) in the whole dataset and 36.6% (37/98) in subtype B samples only, while in the second part (2008–2013) this percentage was found to be 71% (69/97) and 70.9% (61/87), respectively. In both datasets analyzed (including non-B subtypes and subtype B samples only) this difference between the first and the second part of the study period was found to be statistically highly significant at an equal level ($p = 0.0003$). Moreover, a comparison of the mean CD4 cell count between recently infected patients, based on a low percentage of sequence ambiguity, and those designated as patients with established infection, with the percentage of sequence ambiguity above the cutoff level, revealed a significantly higher CD4 cell count in the former group ($p = 0.0022$).

4.5.1. NUCLEOTIDE CHANGES AT 245 CODON OF HIV-1 RT GENE SEQUENCES

The predominant aa at RT codon 245 was the wild type valine (V), found in 61% (168/275), hence 36.7% (101/275) contained mutation at this position. The most common substitution at RT codon 245 was methionine (M) 22.9% (63/275), followed by glutamic acid (E) 7.3% (20/275), glutamine (Q) 5.4% (15/275) and others. We found significantly higher prevalence of RT 245 substitution, compared to the preliminary reports of the HLA-B*5701 allele frequency in HIV infected population in Serbia, $p < 0.0001$.

All transmission clusters associated sequences had wt aa at the position 245. Contrary, 93.6% (44/48) sequences from transmission network had polymorphism at the investigated position. Based on the percentage of ambiguous basecalls, a total of 57% of naive samples (94/165) were classified as recent infection, while among these, 55.3% (52/94) had V at position 245. A total of 43% (71/165) were classified as chronic infection, with the presence of V at RT codon 245 found in 66.2% (47/71). We did not find significant association between polymorphisms at codon 245 and duration of infection ($p = 0.2107$).

CHAPTER 5. DISCUSSION

HIV is responsible for one of the largest viral pandemics in human history. Despite intensive efforts and considerable improvements that have been made for prevention and treatment, this pandemic still persists. In Serbia, the HIV/AIDS epidemic continues to expand, with over a hundred of new HIV diagnoses every year, lately affecting predominantly young MSM. Understanding the dynamics of HIV epidemic in Serbia, by utilizing novel methodologies, may result in major contribution to target public health measures to prevent future transmission events.

The number of public health applications for molecular epidemiology and transmission network analysis has increased rapidly since the improvement in computational capacities and the development of new sequencing techniques (Vasylyeva et al., 2016). With the recent advances in nucleotide sequencing (i.e. high throughput sequencing technologies) which allow faster and more affordable sequencing of pathogens, vast amounts of genetic data can be produced faster, cheaper and more efficiently than ever (Grada and Weinbrecht, 2013). Due to global efforts in monitoring HIV drug resistance, nowadays HIV is one of the most heavily sequenced human pathogen.

So far, this study, including 162 Serbian viral sequences from 1997 to 2007 together with sequences generated in this research is the largest molecular epidemiologic investigation of HIV-1 epidemics in Serbia. This research encompassed investigation of current distribution of HIV-1 subtypes, changing epidemiology, transmission pathways, risk groups and way of dispersal, by means of phylogenetic analysis. Furthermore, herein we described the genetic diversity of HIV-1 strains circulating among infected patients in Serbia and investigated the level of intermixing of HIV-1 strains from Serbian isolates with those from other geographical areas.

Within the present study, we have obtained and analyzed 142 HIV-1 sequences from different patients treated at the major HIV/AIDS health center

in Belgrade, Serbia, between 2008 and 2013. The demographics of the study population were comparable to the national HIV-1 epidemic, characterized predominantly by sexual transmission of infection (80%) and a high male-to-female ratio (http://www.batut.org.rs/index.php?category_id=173; ECDC). In view of the size of the epidemic in Serbia, the size of the dataset and the time-span in which these sequences were obtained provide enough confidence to consider the results obtained in this work as representative of the epidemic scenario of HIV-1 in this region. The results obtained from this research suggest that subtype B remains the predominant one, but with changes in distribution of non B subtypes over the years, the emergence of new non B subtypes and increased genetic diversity among them. The observed proportion of non-B clades was 9.1%, with the most common non-B subtype in Serbia that were found to be subtype C.

The predominance of HIV-1 subtype B found in Serbia is concordant with reports derived from West and Central European countries. HIV-1 subtype B has been responsible for what is often called the 'Western epidemic' in Europe and has remained the predominant clade despite the introduction of non-B clades from later migrating populations, whilst in Eastern Europe the epidemic has been dominated by subtype A (Afsu) (Eastern-type epidemic) (Abecasis et al., 2013; Hemelaar et al., 2011). However, the complexity of the European HIV-1 epidemic has been increasing in Western and Central Europe during recent years, reflected in increasing prevalence of non-B subtypes, linked to migration and later dispersal through transmission networks with patterns varying between individual countries within the region (Beloukas et al., 2016). Population movements including migration from the African and Asian continents have changed HIV epidemic in European countries over the past two decades and have been linked to several infectious disease outbreaks, including local HIV-1 epidemics (Kentikelenis et al., 2015). Based on the fact that immigrants have mostly been infected with non-B strains, a valid

hypothesis is that they were infected, at least at some proportions, before migrating and therefore they could provide the main source of divergent strains in Europe (Beloukas et al., 2016). In Central Europe non-B clades are mainly linked with heterosexual route of transmission but not dominantly with non-European origin. On the other hand, in the Eastern European sub-continent, non-B subtypes are predominant and have been spread through a large PWID-epidemic in FSU countries and in Russian Federation with heterosexual transmissions within local immigrants sexual networks. The SPREAD cohort database reveals B clade as predominant (70.2%) in newly HIV-1 diagnosed patients, after adjusting for oversampling in some countries, followed by C, CRF02_AG, G and A, with 5.0%, 4.9%, 4.8% and 3.6%, respectively (Beloukas et al., 2016). However, there are countries, such as Portugal, Cyprus, Sweden and Greece, where subtype B viruses are less prevalent in new infections ($\leq 50\%$), whilst in the Czech Republic, Germany, Spain, Slovenia and Poland the prevalence of B clade exceeds 80%, with the highest prevalence of subtype B observed in Poland (96.2%) and Slovenia (93.6%) (Beloukas et al., 2016; Paraskevis et al., 2009; Abecasis et al., 2013). Non-B and CRF clades have mainly been associated with immigrants, heterosexual transmission and male gender. Distribution of HIV subtypes in countries neighboring Serbia is characterized by significant diversity, with high prevalence of non B subtypes in some of them (Stanojevic et al., 2012). In Greece among non-B genetic forms subtype A is the predominant one with very high prevalence and, furthermore, it circulates among the long-term residents in this country (Paraskevis et al., 2007). Distribution of subtypes in Albania is marked by very high prevalence of subtype A, which is the predominant one, with the low prevalence of subtype B genetic form (Ciccozzi et al., 2005). Romania, which is the eastern neighbor of Serbia, is a unique case in Europe where HIV-1 epidemic started with subtype F1 which is still the most prevalent one (Abecasis et al., 2013; Paraschiv et al., 2012). In Croatia subtype B predominates, albeit to a lesser extent compared to Serbia (Grgic et al., 2013). The most common non-B subtype in Serbia is found

to be subtype C, accounting for almost one third of all non-B infections. Worldwide, HIV-1 subtype C is currently the most prevalent subtype accounting for approximately 55% of all infections. Among the countries of WHO European region, subtype C is the most prevalent in Israel, where it accounts for 58% of all HIV infections. Here, non-B clades were introduced via two major routes; C clade viruses originated from Ethiopia and infected mainly heterosexuals and Afsu clade was introduced from FSU countries and circulated and expanded mostly among PWID transmission networks (Grossman et al., 2015). Subtype G, that was in Serbia, found to be the most prevalent one from 1997 to 2007, is the most prevalent one in Portugal (34.8%), where is well established in native population (Abecasis et al., 2013).

Changes in the distribution of subtypes can be observed by comparing two periods with similar viral sequences that were generated (1997-2007 vs 2008-2013). In two analyzed period subtype B was the predominate one with the prevalence of 90.1% and 90.8%, respectively for first and second period. However, difference was found among non B subtypes distribution. In period from 1997 to 2007 subtype G was found in almost half of the total number of non B sequences accounted for 6/162 (3.7%). In comparison, from 2008 to 2013 none of 142 analyzed sequences was found to be of subtype G. On the other hand, prevalence of subtype C in the first period, that was found to be 5/162 (3%) when compared to the second period, was the same. Furthermore, subtype A sequences were only identified in the period from 2008-2013 while none of sequences from first period were identified as subtype A. In contrast only in the first period one patient was identified to carry subtype F virus, while this subtype was not identified in the second one.

Results of this research highlighted the benefits of using detailed molecular epidemiological methods for studying epidemics. Monitoring the prevalence of different clades is useful and can provide information about the

genetic diversity of circulating strains as well clues about the potential origin of infections.

Our findings reflect some particular aspects of the HIV-1 subtype distribution in Serbia. The proportion of MSM was significantly higher in subtype B infected patients than in those infected with other subtypes. Regarding subtype B epidemics, our findings are suggestive of multiple subtype B introductions in Serbia. A previous phylogeographic study has shown important level of “mixing” of HIV-1 subtype B sequences across Europe, where sequences from Serbia were found highly interconnected to those from other European countries (Paraskevis et al., 2009). However, a number of subtype B introductions have resulted in further local dispersal, giving rise to local transmission chains. The majority of patients carrying non B subtype strains were infected through heterosexual contact and this association was found to be statistically significant. Numerous studies have shown that non-B infections in Europe are mainly associated with heterosexual infection among immigrants or persons epidemiologically linked to sub-Saharan Africa (Thomson and Nájera, 2001). In contrast to that finding, the vast majority of patients in this study reported to be infected locally, with no epidemiological links abroad. The recent study based on HIV isolates of newly diagnosed patients throughout Europe has shown that transmission risk for HIV is among the main determinants of subtype distribution (Abecasis et al., 2013). Contrary to the majority of western countries where subtype B epidemic was firstly described in the MSM population and subsequently found associated with other transmission routes, in particular PWID, epidemiological data indicate that the epidemic of HIV in Serbia was first established in PWID and was almost exclusively caused by the subtype B genetic form, (Stanojevic et al., 2002). In this respect, the HIV epidemic in Serbia is similar to the one in Italy, where the most common route of transmission, at the beginning of the epidemics, was also found to be needle sharing among intravenous drug users (Callegaro et al., 2011).

In Serbia, as in other countries, the HIV epidemic is evolving with changing patterns of the HIV subtypes present and prevailing transmission risk. In this research, the changing epidemiology of HIV-1 disease among main transmission groups was described. HIV epidemic in Serbia has entered a new phase with the number of infections acquired through MSM contact, rising each year. Our findings are in line with epidemiological data indicating that, unlike previously found, MSM contact currently represents the driving force of the HIV epidemic in Serbia, followed by heterosexual contact, while infections among injecting drug users are at low levels.

The remarkable finding of this study was identification and characterization of HIV-1 transmission clusters in Serbia by analyzing large scale of 304 sequences from isolates sampled between 1997 and 2013, that were for the first time described in this research. This study demonstrates a unique view into the structure of local transmission in Serbia through the integration of molecular, clinical, and demographic data. Molecular epidemiologic evaluation of HIV-1 transmission networks can elucidate behavioral components of transmission that can be targets for intervention. Understanding HIV transmission patterns is important in the design and implementation of prevention interventions. One of the major challenges in transmission clusters analysis is adoption of adequate phylogenetic criteria for cluster identification. There is no consensus on phylogenetic criteria that should be used to establish a putative transmission cluster in phylogenetic based studies of HIV transmission network. Those are generally based on defined cut-off bootstrap support and/or genetic distance values, but also include Bayesian probability and reference sequence analysis. Appropriate phylogenetic criteria may depend on the underlying epidemiological and evolutionary dynamics in a given research settings.

Notably, this study highlighted the need for consensus criteria on defining transmission clusters, by showing that using strict criteria for

clustering (high bootstrap values and low genetic distances) can result in underestimation of number of true transmission clusters. Using a tight definition for transmission clustering in which inclusion in a cluster required each sequence to have another sequence within a genetic distance of $\leq 1.5\%$ can likely select viral sequences isolated only from patients with early HIV infection. Instead of using only this strict criteria, applied up to now in several research, herein we used an additional and more stringent criteria in order not to exclude from actual clusters sequences from related chronically infected patients. Criteria sets used in our analysis were aimed to be sufficiently strict, in order to maintain specificity in identifying true transmission clusters, yet to avoid underestimation of the incidence of transmission chains, in particular in view of the rather extended ten years timeframe of sampling. On the other hand, the range of identified transmission events and the level of their epidemiological linkage are largely influenced by sampling strategy and the level of coverage of target population. Our sample corresponds to around 10% of the total number of registered cases from the beginning of the epidemics, however, it contains around 30% of all newly diagnosed patients in the years of the study period. The proportion of patients involved into clustered events in our study, of 24.5% within characterized transmission clusters and 16.7% within larger transmission network, is in the range of 12.7% to 64% as found in reports from others geographical regions, although both the populations and the methods of these phylogenetic studies are heterogeneous (Brenner et al., 2007, 2011; Chalmet et al., 2010; Frange et al., 2012; Hué et al., 2005; Recordon-Pinson et al., 2009; Yerly et al., 2009). Transmission clusters identified in our study are highly associated with male, in particular MSM patients, rather than with people infected heterosexually or IDU. Compared to other studies we found that local transmission structure in Serbia is similar to the regions where MSM transmission predominates. Depending on the studied population, some studies revealed transmission structure of equal contribution to HIV epidemic through HIV spread among both heterosexual and MSM risk groups (Dennis et

al., 2012). However, increase in the proportion of HIV-1 transmission through MSM networks has been reported in a number of studies throughout the world (Cuevas et al., 2009; Kouyos et al., 2010; Lewis et al., 2008; Mitsch et al., 2008; Paraschiv et al., 2012). The latest molecular epidemiological data show that clustering of HIV infections in high-transmission bursts and faster spread in networks are associated with HIV epidemics in MSM, much more than related to heterosexual transmission (Beyrer et al., 2012a). Substantial clustering of HIV infections in MSM networks underscores the need of reinforcing prevention measures within this group on the local level (Beyrer et al., 2012b). This notion is further emphasized by the analyses of the temporal origin of subtype B epidemics in Serbia. Higher rates of clustering in MSM, suggest that public health interventions should target this key population at risk, including prevention, testing and linkage to care strategies, to reduce HIV-1 transmission in Serbia.

In comparison to human DNA, genome of RNA viruses is much less stable, undergoing dynamic evolution (Rambaut et al., 2004). Expansion of multiple viral lineages in infected individual is driven by high rates of mutation, large population sizes and short generation times. Hence, unlike common practice of DNA fingerprinting, analyzing the transmission of fast evolving viruses such as HIV is difficult since finding full identity between two HIV samples is highly improbable (Abecasis et al., 2011). Nevertheless, virus sequences sampled from epidemiologically linked individuals are likely to be more closely related compared to unlinked ones. This phenomenon can be exploited to give forensic evidence whether or not there is a genetic relatedness between viral sequences from alleged donor and an infected individual (de Oliveira et al., 2006; Saludes et al., 2013; Scaduto et al., 2012; Metyker et al., 2002; González-Candelas et al., 2003; González-Candelas et al., 2013)

In Serbia, as in many countries worldwide, deliberate exposure to and transmission of HIV are criminalized (Weait, 2011). In this study, initiated by

the investigative process within the private lawsuit regarding HIV transmission, partial polymerase (*pol*) and envelope (*env*) genetic segments were analyzed in order to infer genetic relatedness and explore suspected epidemiological linkage between HIV genome sequences isolated from three HIV-1 infected patients.

Phylogenetic analysis of transmission clusters as a forensic evidence in criminal HIV transmission prosecutions were first used in court of law in Sweden in 1992 (Albert et al., 1994). In order to give strong forensic evidence regarding transmission, phylogenetic analysis needs to be enhanced with application of several methods and conducted under strictly controlled conditions (Bernard et al., 2007; Leitner and Albert, 2000). Firstly, it is vitally important to include adequate local controls comprising viral sequences from infected individuals sharing the same transmission risk, from similar geographic location, of the same genetic clades (subtype or recombinant form), and diagnosed in the same time period as the query subjects, but who are not believed to be a part of the investigated outbreak. The use of inappropriate controls could overemphasize the relatedness between the viruses under study as being uncommonly unique (Leitner and Albert, 2000). All infected individuals in a population can never be sampled, nevertheless, if sufficient number of local control sequences is included in the analyses, significant clustering of sequences under investigation can indicate that they do belong to a transmission chain. Secondly, apart from appropriate local controls, additional precondition in forensic phylogenetics is inclusion into the analysis of at least two genetic regions of reasonable length, depending on the gene under investigation. To our best knowledge, this is the first report of phylogenetic analyses utilized for forensic exploration of suspected HIV transmission in Serbia. The current Criminal Code of Republic of Serbia recognizes deliberate or inadvertent spread of HIV, and it is punishable by 1–15 years sentence (http://www.paragraf.rs/propisi/krivicni_zakonik.htm, in

Serbian). In this study, phylogenetic analyses indicated that the three HIV sequences under investigation were all of subtype B and more genetically related to each other than to any control sequence. Notably, pairwise dissimilarity between the 3 query sequences was much lower than average between controls. However, due to an unknown number of unsampled HIV infected individuals from the same transmission network, accurate transmission history cannot be unambiguously reconstructed. Further, restricting genetic distance indicative of transmission cluster to a threshold of 1.5% across the studied genome part has not been universally accepted, since transmission events may have been separated in time, allowing further evolution. Hence, the level of similarity and phylogenetic clustering of the examined sequences does support their epidemiological relatedness, however cannot be considered proof of their direct epidemiological linkage. Moreover phylogenetic analysis cannot be used to unambiguously infer the directionality of transmissions (who infected whom). An approach to address this issue is based on the topology of the phylogenetic tree, if a paraphyletic relationship is observed (i.e., a subset of source viral sequences is more closely related to all recipient sequences than to other source sequences) (Scaduto et al., 2010). This assumption is based on the fact that HIV-1 infection is associated with a transmission bottleneck: newly acquired HIV infection (in particular when sexually transmitted) is established by a limited number or a single viral strain (Keele et al., 2007). In this research, paraphyletic relationship between sequences of subjects 2 and 3 and those of subject 1 is found, suggestive of the hypothesis of subject 1 being the source of infection for both subjects 2 and 3. This finding would be in opposition of the *a priori* hypothesis of HIV transmission from subject 3 to subject 1, but would still correlate to the epidemiological *a priori* information. This relation was consistent when separately analyzing query subjects in pairs, as described by Romero-Severson et al. (Romero-Severson et al., 2016). According to the analysis approach by Romero-Severson et al. (Romero-Severson et al., 2016), tree topologies of the EPLD-PCR based phylogenetic trees analyzing pairs of

query subjects, revealed paraphyletic-monophyletic (PM) topology when sequences of Subject 1 were analyzed with either sequences of Subject 2 and 3, in both *env* and *pol* genes . This result is consistent with the transmission scenario of Subject 1 being the source for both Subjects 2 and 3 (Romero-Severson et al., 2016). When analyzing sequences of Subjects 2 and 3, the obtained topology corresponded to dually monophyletic (MM) in the *env* region, implying the common source transmissions, whereas in the *pol* region the obtained topology corresponded to a combination of paraphyletic and polyphyletic (PP), which might be the consequence of a short time period between the two transmission events. Still, caution is needed with regards potential interpretation of findings in a forensic case in court. Even with phylogenetic analysis performed using state of the art phylogenetic methods together with the strictest conditions for cluster identifications, indirect link between viral strains can never be ruled out; hence even demonstrating direction of transmission is not identical to demonstrating direct transmission from one to another (Abecasis et al., 2011). Furthermore, at the molecular level sequences also contain temporal information about the date of origin and age of epidemics. Therefore, using Bayesian Markov chain Monte Carlo (MCMC) approach we can estimate the date of the most recent common ancestor of each cluster, even including clusters formed from viral variants of one infected person. Neither of these methods should be used separately but together with other epidemiological, clinical and behavioral data as well as with other forensic evidence and only in the context of hypothesis testing. The dated tree in our analysis confirmed the clustering pattern of query sequences. Obtained tMRCA estimates for the time of infection for subject 1 preceded the estimates for the time of infection for subjects 2 and 3, consistent with the dates of diagnosis and also suggestive of the hypothesis of subject 1 being the source of infection for both subjects 2 and 3, opposite to the a priori hypothesis. Besides the limitations inherent to phylogenetic analyses of HIV transmission, our study has some additional limitations. We had access only one sample per

subject, whereas it is preferable to have two, from different time points. It is recommended by the relevant guides to analyze preferably two or more samples of the same subject (Bernard et al., 2007). However, as stated under the limitations of the study, we had access to only single sample per subject. In this research, follow-up sampling was not possible, the main reason being the fact that all the subjects meanwhile started the successful antiretroviral therapy leading to undetectable viral load in further samples. However, using an EPLD-PCR approach to assess viral quasispecies allowed to sufficiently explore viral diversity at the given time point. Further on, the identities of the parties were not blinded to us throughout the analysis. However, the analysis was performed in compliance to all other recommendations for forensic application of phylogenetic analysis, whereas, circumstantially, the case has been settled out of court

The availability of greater numbers of sampled viral sequences combined with increasing power of phylogenetic techniques makes it possible to retrieve valuable and unique information about the course of viral epidemics from molecular data (Lemey et al., 2003; Robbins et al., 2003; Travers et al., 2004; Deng et al., 2008).

Many studies were based on partial *pol* gene phylogeny, which has shown to be adequate to infer transmission events and to characterize epidemiological patterns of public health relevance (Hué et al., 2004; Lemey et al., 2005). The *pol* region of the HIV genome is generally used for phylogenetic analyses out of convenience, since it is a by-product of the routine HIV drug resistance testing. Lemey et al., showed that *pol* gen, analysed also in this research to infer transmission events, is adequate and very good region for reconstructing transmission chains and characterisation epidemiological patterns of public health relevance (Lemey et al., 2005). However, there are concerns that the *pol* region is too conserved since it codes for regulatory genes involved in viral replication and consequently has insufficient genetic

variability (Palmer et al. 2002). This has led to debate about its suitability for phylogenetic reconstructions (Palmer et al. 2002; Sturmer, Preiser et al. 2004). While *gag* and *env* genes have been preferred due to their greater genetic variability, *pol* remains attractive due to its greater accessibility through routine drug resistance testing. Hue et al (Hue et al. 2004)

Studies of the evolutionary history of HIV enable us to explore the past and to estimate the age of an epidemic. Phylogenetic analysis has successfully been used to investigate HIV-1 transmission and epidemiological linkage, however, the direction and timing of HIV-1 transmission are far more difficult to assess (Rachinger et al., 2011). Standard phylogenetic reconstruction only establishes evolutionary relatedness but not evolutionary direction. The integration of molecular clock models in phylogenetic inference has enabled the reconstruction of rooted, time-measured phylogenies (Drummond et al., 2002; Rachinger et al., 2011).

In this doctoral dissertation phylogenetic analyses and a Bayesian coalescent-based framework was used, to investigate the origin and to estimate with more precision the initial introduction of the HIV-1 subtype B in Serbia. We based this analysis on datasets of *pol* gene sequences sampled over a period of 17 years (1997-2013). Molecular clock analysis dated the initial introduction of both the founder strain of IDUs and MSM in a similar period to the early 1980s. Our findings suggest that IDU populations might have played a major role in the introduction and initial dissemination HIV-1 in Serbia. In contrast to IDU regional spread among MSM population that gave rise to local transmission chains happened years after initial introduction of founder strain. Molecular clock analysis dated the tMRCA of the large network of Serbian subtype B sequences to 1994 (95% Higher Posterior Density HPD: 1982–2000), whilst for the most extended local subtype B MSM transmission cluster tMRCA was estimated at 2004 (95% HPD: 2002–2006)

Estimated origin of the subtype C epidemics in Serbia is much more recent, however, similar to the time found for subtype G in the previous study

By reconstructing the evolutionary history of viral genomes the behavior of viral populations can be modelled, and the future of epidemics may be forecast. Together with epidemiological data these findings illustrate that the epidemic spread of HIV in Serbia within the local network and outside the firstly implicated IDUs community probably started in the beginning of 90s, while epidemic spread of HIV subtype B among MSMs represents the most recent HIV-1 epidemic in Serbia. Evolutionary studies based on a single gene analysis may suffer from limitations of evolutionary rate varying between genes due to numerous host and virus related factors, giving rise to either overestimation or underestimation of the timeframe of viral introduction (Abecasis et al., 2009; Neogi et al., 2012). However, many newly reported studies, pol gene based or comparing tree genetic regions (gag, pol, env), have shown pol-gene analysis to provide the most reliable estimate of tMRCA (Ciccozzi et al., 2012; Han et al., 2013; Neogi et al., 2012; Paraschiv et al., 2012; Yebra et al., 2013). We estimated the origin of the subtype G epidemics in Serbia to the beginning of the 90s, however the subsequent dispersal within the local population remained limited. Decline in the prevalence of heterosexual HIV transmission in recent years as well as proven relationship between this risk group and non B subtypes may be the reason for the absence of further spread of the subtype G epidemic in Serbia.

HIV-1 infection is associated with a transmission bottleneck: newly acquired HIV infection (in particular when sexually transmitted – which is the case in almost 90% of our study population) is established by a limited number or a single viral strain (Keele et al., 2008). Further inpatient viral evolution over time leads to increasing HIV viral diversity within individuals by initial accumulation of mixed bases, as new mutations emerge in existing quasispecies. Hence, a threshold of percentage of ambiguous bases per

sequence has been proposed to identify non-recent vs. recent HIV infection (the one lasting less than 6 months to 1 year). Essential for this method's performance is standardized mixed bases calling, variability whereupon lays ground for intrasample and intersample interpretation variability, using either automated or manual sequence editing. However, it has been shown that duration of HIV infection can be inferred from the proportion of mixed bases identified during population-based sequencing of the *pol* region, with similar accuracy using any mixed base threshold between 15% and 25%. The threshold we used falls within the proposed range. Furthermore, a cut-off of 0.47% of ambiguous nucleotides for distinguishing recent from established infection has been proposed for HIV subtype B *pol* gene sequences, whereas differing percentage has been described for some of the non-B subtypes (Zheng et al., 2013). We found high prevalence of HLA B*57 associated polymorphism that is almost twice as reported prevalence of HLA B*57-01 allele in Serbia (8%). This may be related to the presence of other, similar HLA alleles, limiting specificity of the correlation between HLA-B*5701 and RT codon 245 variation. Furthermore, our well-supported separate phylogenetic network was consisted of 51% (25/54) sequences with HLA B*57 associated polymorphism. In addition, no statistically significant difference was found in the prevalence of RT 245 substitutions between recent and chronic infection. In view of the time-span of sampling within the clade and uniformity of transmission route (mainly MSM) this may be related to early fixation of an HLA induced selective imprint, during viral evolution in an infected drug naive patient and its onward spread. On the other hand, these results may imply that mutation at RT codon 245 in patients infected with HIV strain carrying mutation on this position tend to reverse into a wild type form during the evolution of a virus in infected individual in order to incise adaptive value (W) and higher fitness effect of the virus.

CHAPTER 6. CONCLUSIONS

According to the defined objectives and based on the obtained results the following conclusions can be reached:

- HIV epidemics in Serbia continues to be dominated with subtype B, but with changes in distribution of non B subtypes over time, the emergence of new non B subtypes and increased genetic diversity among them.
- The majority of patients carrying non B subtype strains were infected through heterosexual contact and this association was found to be statistically significant.
- An important proportion of the local subtype B spread is clustered, with 39.6% of analyzed sequences may be considered to be part of transmission clusters/network.
- Transmission clusters are highly associated with MSM rather than with other risk categories.

- Transmission cluster analysis in the context of forensic investigation revealed a well-supported transmission chain, in line of the *a priori* hypothesis of their epidemiological linkage with tree topology possibly implying the direction of transmission contrary to the *a priori* hypothesis. It is important to emphasize that the results of forensic phylogenetics need to be interpreted in the context of hypothesis testing and in view of all the limitations of this approach.
- tMRCA for local HIV epidemic in Serbia, regardless of the transmission risk, dates back to the early eighties of the 20st century.
- Local epidemic spread within transmission networks and outside the firstly implicated IDUs community is dating from the beginning of the nineties, while epidemic spread of HIV subtype B among MSMs represents the most recent HIV-1 epidemic in Serbia.
- Non-B epidemics in Serbia, although introduced at a similar time (early 90s), resulted in limited dispersal, and exemplified by subtype G clade, possibly due to the prevailing transmission route.
- Ambiguous nucleotides analyses, used as a method for distinguishing recent from established infection suggested an increasing proportion of recent infections, significantly higher in the second half of the study period.

- Equal prevalence of RT 245 substitutions in samples from recent and chronic HIV infection, together with high prevalence of this polymorphism in sequence within transmission network, suggested early fixation of an HLA induced selective imprint, during intra-host viral evolution and raises the question of onward transmission of HLA induced mutation and their inter-host persistence.

CHAPTER 7. REFERENCES

Abecasis A, Vandamme AM, Lemey P. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J Virol.* 2009; 83(24):12917-12924.

Abecasis A, Wensing A, Paraskevis D, Vercauteren J, Theys K, Van de Vijver et al. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology.* 2013; 10:7.

Abecasis A, Geretti A, Albert J, Power L, Weait M, Vandamme AM. Science incourt: the myth of HIV fingerprinting. *Lancet Infect Dis.* 2011; 11(2):78-79.

Abele LG, DeBry RW. Florida dentist case: research affiliation and ethics. *Science.* 1992; 255(5047):903.

Abram E, Ferris L, Shao W, Alvord WG, Hughes SH. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J Virol.* 2010; 84(19):9864-9878.

Aghokeng AF, Ayouba A, Mpoudi-Ngole E, Loul S, Liegeois F, Delaporte E, Peeters M. Extensive survey on the prevalence and genetic diversity of SIVs in primate bushmeat provides insights into risks for potential new cross-species transmissions. *Infect Genet Evol.* 2010; 10(3):386-396.

Albert J, Wahlberg J, Leitner T, Escanilla D, Uhlén M. Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes. *J Virol.* 1994; 68(9):5918-5924.

Aldrich C, Hemelaar J. Global HIV-1 diversity surveillance. *Trends Mol Med.* 2012; 18:691-694.

- Altfeld M, Behrens G, Ostrowski M, Rubbert A, Schieferstein C, R. E. Schmidt R.E, Walker B.D., and E. Wolf, *HIV Medicine* 2003. Paris, Cagliari,Wuppertal, Sevilla: Flying Publisher, 2003.
- Ammann A, Cowan M, Wara D, Weintrub P, Dritz S, Goldman H, Perkins HA. Acquired immunodeficiency in an infant: possible transmission by means of blood products. *Lancet*. 1983; **1**(8331):956–958.
- An P, Winkler CA. Host genes associated with HIV/AIDS: advances in gene discovery. *Trends Genet*. 2010; **26**(3):119-131.
- Andersson E, Shao W, Bontell I, Cham F, Cuong do D, Wondwossen A, Morris L, Hunt G, Sönnnerborg A, Bertagnolio S, Maldarelli F, Jordan MR. Evaluation of sequence ambiguities of the HIV-1 pol gene as a method to identify recent HIV-1 infection in transmitted drug resistance surveys. *Infect Genet Evol*. 2013; **18**:125-31.
- Apetrei C, Kaur A, Lerche NW, et al. Molecular epidemiology of simian immunodeficiency virus SIVsm in U.S. primate centers unravels the origin of SIVmac and SIVstm. *J Virol*. 2005; **79**:8991–9005.
- Apisarnthanarak A, Jirayasethpong T, Sa-nguansilp C, et al. Antiretroviral drug resistance among antiretroviral-naïve persons with recent HIV infection in Thailand. *HIV Med*. 2008; **9**(5):322-325.
- Arthur LO, Peeters M, Shaw GM, Sharp PM and Hahn BH. (1999). Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. *Nature* **397**: 4360-4441.
- Ayouba A, Akoua-Koffi C, Calvignac-Spencer S, Esteban A, Locatelli S, et al. Evidence for continuing cross-species transmission of SIVsmm to humans: characterization of a new HIV-2 lineage in rural Côte d'Ivoire. *AIDS*. 2013; **27**(15):2488-2491.

Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, et al. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*. 1983; 220(4599):868-71.

Beloukas A, Psarris A, Giannelou P, Kostaki E, Hatzakis A, Paraskevis D. Molecular epidemiology of HIV-1 infection in Europe: An overview. *Infect Genet Evol*. 2016; 46:180-189.

Bernard J, Azad Y, Vandamme M, Weait M, Geretti M. HIV forensics: pitfalls and acceptable standards in the use of phylogenetic analysis as evidence in criminal investigations of HIV transmission. *HIV Med*. 2007; 8(6):382-327.

Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, Wirtz AL, Brookmeyer R. Global epidemiology of HIV infection in men who have sex with men. *Lancet*. 2012; 380(9839):367-377.

Beyrer C, Sullivan PS, Sanchez J, Dowdy D, Altman D, Trapence G, et al. A call to action for comprehensive HIV services for men who have sex with men. *Lancet*. 2012; 380(9839):424-38.

Bisgrove D, Lewinski M, Bushman F, Verdin E. Molecular mechanisms of HIV-1 proviral latency. *Expert Rev Anti Infect Ther*. 2005; 3(5):805-814.

Bloom A. Acquired immunodeficiency syndrome and other possible immunological disorders in European haemophiliacs. *Lancet*. 1984; 1(8392):1452-1455.

Bobkov A, Kazennova E, Khanina T, Bobkova M, Selimova L, Kravchenko A, Pokrovsky V, Weber J. 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):257-261.

Bobkov F, Kazennova V, Selimova M, Khanina A, Ryabov S, Bobkova R, et al. Temporal trends in the HIV-1 epidemic in Russia: predominance of subtype A. *J Med Virol*. 2004a;74(2):191-196.

Bour S, Geleziunas R, Wainberg MA. The human immunodeficiency virus type 1(HIV-1) CD4 receptor and its central role in promotion of HIV-1 infection. *Microbiol Rev*. 1995 ; 59(1):63-93.

Bredell H, Hunt G, Casteling A, Cilliers T, Rademeyer C, Coetzer M, et al. HIV-1 Subtype A, D,G, AG and unclassified sequences identified in South Africa. *AIDS Res Hum Retroviruses*. 2002; 18(9):681-683.

Bredell H, Hunt G, Morgan B, Tiemessen CT, Martin DJ, Morris L. Identification of HIV type 1 intersubtype recombinants in South Africa using env and gag heteroduplex mobility assays. *AIDS Res Hum Retroviruses*. 2000; 16(5):493-497.

Brenner BG, Roger M, Stephens D, Moisi D, Hardy I, Weinberg J, et al. Montreal PHI Cohort Study Group. Transmission clustering drives the onward spread of the HIV epidemic among men who have sex with men in Quebec. *J Infect Dis*. 2011;204(7):1115-1119.

Buendia P, Cadwallader B, DeGruttola V. A phylogenetic and Markov model approach for the reconstruction of mutational pathways of drug resistance. *Bioinformatics*. 2009; 25:2522-2529.

Bush S, Tebit DM. HIV-1 Group O Origin, Evolution, Pathogenesis, and Treatment: Unraveling the Complexity of an Outlier 25 Years Later. *AIDS Rev*. 2015; 17(3):147-158.

Burnett A, Spearman P. APOBEC3G multimers are recruited to the plasma membrane for packaging into human immunodeficiency virus type 1 virus-like particles in anRNA-dependent process requiring the NC basic linker. *J Virol.* 2007; 81(10):5000-5013.

Butler IF, Pandrea I, Marx PA, Apetrei C. HIV genetic diversity: biological and public health consequences. *Curr HIV Res.* 2007; 5(1):23-45.

Callegaro A, Svicher V, Alteri C, Lo Presti A, Valenti D, Goglio A, et al. Epidemiological network analysis in HIV-1 B infected patients diagnosed in Italy between 2000 and 2008. *Infect Genet Evol.* 2011; 11(3):624-632.

Campbell-Yesufu OT, Gandhi RT. Update on human immunodeficiency virus (HIV)-2 infection. *Clin Infect Dis.* 2011; 52(6):780-787.

Carvajal-Rodríguez A, Posada D, Pérez-Losada M, Keller E, Abrams EJ, Viscidi R et al.. Disease progression and evolution of the HIV-1 env gene in 24 infected infants. *Infect Genet Evol.* 2008; 8:110-120.

Case K. Nomenclature: Human Immunodeficiency Virus. *Annals of Internal Medicine* 1986; 105(1): 133.

Castro-Nallar E, Pérez-Losada M, Burton GF, Crandall KA. The evolution of HIV: inferences using phylogenetics. *Mol Phylogenet Evol.* 2012; 62(2):777-92.

Cazein F, Pillonel J, Le Strat Y, Pinget R, Le Vu S, Brunet S, et al. New HIV and AIDS diagnoses, France, 2003-2013. *Bulletin épidémiologique hebdomadaire*, 2015; pp. 152-161

Cen S, Guo F, Niu M, Saadatmand J, Deflassieux J, Kleiman L. The interaction between HIV-1 Gag and APOBEC3G. *J Biol Chem.* 2004; 279:33177-33184.

Ciccozzi M, Gori C, Boros S, Ruiz-Alvarez MJ, Harxhi A, Dervishi M, et al. Molecular diversity of HIV in Albania. *J Infect Dis.* 2005; 192(3):475-479.

Ciuffi A, Barr SD. Identification of HIV integration sites in infected host genomic DNA. *Methods.* 2011; 53(1):39-46.

Chalmet K, Staelens D, Blot S, Dinakis S, Pelgrom J, Plum J, et al. Epidemiological study of phylogenetic transmission clusters in a local HIV-1 epidemic reveals distinct differences between subtype B and non-B infections. *BMC Infect Dis.* 2010; 10:262.

Chakrabarti L, Guyader M, Alizon M, Daniel MD, Desrosiers RC, Tiollais P, Sonigo P. Sequence of simian immunodeficiency virus from macaque and its relationship to other human and simian retroviruses. *Nature.* 1987; 328(6130):543-547.

Cochrane A. HIV-1 Rev function and RNA nuclear-cytoplasmic export. *Methods Mol Biol.* 2014; 1087:103-114.

Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP. Parallel evolution of drug resistance in HIV: Failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol Biol Evol.* 1999; 16:372-382.

Cuevas MT, Muñoz-Nieto M, Thomson MM, Delgado E, Iribarren JA, Cilla G, et al. Spanish Group of HIV-1 Antiretroviral Resistance Studies in the Basque Country. HIV-1 transmission cluster with T215D revertant mutation among newly diagnosed patients from the Basque Country, Spain. *J Acquir Immune Defic Syndr.* 2009; 51(1):99-103.

Daniel MD, Letvin NL, King NW, et al. Isolation of T-cell tropic HTLV-III-like retrovirus from macaques. *Science.* 1985;228:1201-1204.

Daniel MD, Li Y, Naidu YM, Durda PJ, Schmidt DK, Troup CD, Silva DP, et al. Simian immunodeficiency virus from African green monkeys. *J Virol.* 1988; 62(11):4123-418.

D'arc M, Ayoub A, Esteban A, Learn G, Boué V, Liegeois, F, et al. Origin of the HIV-1 group O epidemic in western lowland gorillas. *Proc. Natl. Acad. Sci. U. S. A.* 2015; 112, E1343–E1352

Deng X, Liu H, Shao Y, Rayner S, Yang R. The epidemic origin and molecular properties of B': a founder strain of the HIV-1 transmission in Asia. *AIDS.* 2008; 22(14):1851-1858.

Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhart M, et al. Identification of a major co-receptor for primary isolates of HIV-1. *Nature.* 1996; 381(6584):661-666.

Dennis AM, Hué S, Hurt CB, Napravnik S, Sebastian J, Pillay D, Eron JJ. Phylogenetic insights into regional HIV transmission. *AIDS.* 2012; 26(14):1813-1822.

de Oliveira T, Kharsany AB, Gräf T, Cawood C, Khanyile D, Grobler A, Puren A, Madurai S, Baxter C, Karim QA, Karim SS. Transmission networks and risk of HIV infection in KwaZulu-Natal, South Africa: a community-wide phylogenetic study. *Lancet HIV.* 2017 Jan;4(1):e41-e50.

de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, Snoeck J, van Rensburg EJ, Wensing AM, van de Vijver DA, Boucher CA, Camacho R, Vandamme AM. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics.* 2005;21(19):3797-800.

de Oliveira T, Pybus OG, Rambaut A, Salemi M, Cassol S, Ciccozzi M, Rezza G, Gattinara GC, D'Arrigo R, Amicosante M, Perrin L, Colizzi V, Perno CF.

Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature* 2006; 444: 836-837.

de Sousa JD, Müller V, Lemey P, Vandamme AM. High GUD incidence in the early 20 century created a particularly permissive time window for the origin and initial spread of epidemic HIV strains. *PLoS One*. 2010; 5(4):e9936.

de Sousa JD, Alvarez C, Vandamme AM, Müller V. Enhanced heterosexual transmission hypothesis for the origin of pandemic HIV-1. *Viruses*. 2012;4(10):1950-1983.

Dev J, Park D, Fu Q, Chen J, Ha HJ, Ghantous F, Herrmann T, Chang W, Liu Z, Frey G, Seaman MS, Chen B, Chou JJ. Structural basis for membrane anchoring of HIV-1 envelope spike. *Science*. 2016; 353(6295):172-175.

Dragic T, Litwin V, Allaway GP, Martin SR, Huang Y, Nagashima KA, et al. HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature*. 1996; 381(6584):667-673.

Drummond, A., Oliver, G., Rambaut, A., 2003. Inference of viral evolutionary rates from molecular sequences. *Adv. Parasitol.* 54, 331-358.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7, 214.

Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A*. 1996; 93(14):7085-7090.

Ensoli B, Fiorelli V, Ensoli F, Cafaro A, Titti F, Buttò S, et al. Candidate HIV-1 Tat vaccine development: from basic science to clinical trials. *AIDS*. 2006; 20(18):2245-2261.

European Centre for Disease Prevention and Control/WHO Regional Office for Europe. HIV/AIDS surveillance in Europe 2015. Stockholm: ECDC; 2016.

Faria N, Hodges-Mameletzis I, Silva J, Rodés B, Erasmus S, Paolucci S, et al. Phylogeographical footprint of colonial history in the global dispersal of human immunodeficiency virus type 2 group A. *J Gen Virol.* 2012; 93(Pt 4):889-899.

Faria, N.R., Rambaut, A., Suchard, M.A., Baele, G., Bedford, T., Ward, M.J., et al., 2014. The early spread and epidemic ignition of HIV-1 in human populations. *Science* 346, 56–61.

Felsenstein J. Phylogenies and the Comparative Method 1985 *The American Naturalist*.

Fouchier RA, Meyer BE, Simon JH, Fischer U, Malim MH. HIV-1 infection of non-dividing cells: evidence that the amino-terminal basic region of the viral matrix protein is important for Gag processing but not for post-entry nuclear import. *EMBO J.* 1997; 16(15):4531-4539.

Frange P, Meyer L, Deveau C, Tran L, Goujard C, et al. French ANRS CO6 PRIMO Cohort Study Group.. Recent HIV-1 infection contributes to the viral diffusion over the French territory with a recent increasing frequency. *PLoS One.* 2012;7(2):e31695.

Freed EO. HIV-1 replication. *Somat Cell Mol Genet.* 2001; 26(1-6):13-33.

Freed O. HIV-1 and the host cell: an intimate association. *Trends Microbiol.* 2004; 12(4):170-177.

Freed E, and Martin A. Virion incorporation of envelope glycoproteins with long but not short cytoplasmic tails is blocked by specific, single amino acid substitutions in the human immunodeficiency virus type 1 matrix. *J. Virol.* 1995; 69, 1984-1989.

Friedman-Kien A., Laubenstein L, Marmor M, Hymes K, Green J, Ragaz A, et al. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men - New York City and California. *MMWR* 1981; 30(25):305-308.

Fouchier, R. A. M., Meyer, B. E., Simon, J. H. M., Fischer, U. & Malim, M. H. (1997). HIV-1 infection of non-dividing cells: evidence that the amino-terminal basic region of the viral matrix protein is important for Gag processing but not for post-entry nuclear import. *EMBO J.* 16, 4531-4539.

Fultz PN, McClure HM, Anderson DC, Swenson RB, Anand R, Srinivasan A. Isolation of a T-lymphotropic retrovirus from naturally infected sooty mangabey monkeys (*Cercocebus atys*) *Proc Natl Acad Sci U S A.* 1986;83:5286-5290.

Gao F, Robertson DL, Morrison SG, Hui H, Craig S, Decker J, et al. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *J Virol.* 1996 Oct;70(10):7013-29.

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF, Cummins LB, Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S and Bhattacharya T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science.* ;288: 1789- 1795

Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, et al. Diversity considerations in HIV-1 vaccine selection. *Science.* 2002; 296(5577):2354-2360.

Galo, R.C: A reflection on HIV/AIDS research after 25 years *Retrovirology*, 3:72.

Gilks CF, Crowley S, Ekpini R, et al. The WHO public-health approach to antiretroviral treatment against HIV in resource-limited settings. *Lancet*. Aug 5 2006; 368(9534):505-510.

Gifford RJ, de Oliveira T, Rambaut A, Pybus OG, Dunn D, Vandamme AM, et al. UK Collaborative Group on HIV Drug Resistance.. Phylogenetic surveillance of viral genetic diversity and the evolving molecular epidemiology of human immunodeficiency virus type 1. *J Virol*. 2007; 81(23):13050-13056.

Gisselquist D. Emergence of the HIV type 1 epidemic in the twentieth century: comparing hypotheses to evidence. *AIDS Res Hum Retroviruses*. 2003; 19(12):1071-1078.

Gottlieb S, Schroff R, Schanker M, Weisman D, Fan T, Wolf A. *Pneumocystis carinii* pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *N Engl J Med*. 1981; 305(24):1425-1431.

Goodrich DW, Duesberg PH. Retroviral recombination during reverse transcription. *Proc Natl Acad Sci U S A*. 1990; 87(6):2052-2056.

González-Candelas F, Bracho MA, Moya A. Molecular epidemiology and forensic genetics: application to a hepatitis C virus transmission event at a hemodialysis unit. *J Infect Dis*. 2003; 187: 352-358.

González-Candelas F, Bracho MA, Wróbel B, Moya A. Molecular evolution in court: analysis of a large hepatitis C virus outbreak from an evolving source. *BMC Biol*. 2013 ;11: 76.

Goulder PJ and Watkins DI. HIV and SIV CTL escape: implications for vaccine design. *Nat Rev Immunol*. 2004;4(8):630-640.

Grada A, Weinbrecht K. Next-generation sequencing: methodology and application. *J Invest Dermatol*. 2013 A; 133(8):e11. Greene WC. A history of AIDS: looking back to see ahead. *Eur J Immunol* 2007;37 Suppl 1:S94-102.

Grgic I, Lepej Z, Lunar M, Poljak M, Vince A, Vrakela B, et al. The prevalence of transmitted drug resistance in newly diagnosed HIV-infected individuals in Croatia: the role of transmission clusters of men who have sex with men carrying the T215S surveillance drug resistance mutation. *AIDS Res Hum Retroviruses*. 2013; 29(2):329-336.

Grossman, Z., Avidor, B., Mor, Z., Chowers, M., Levy, I., Shahar, E., et al. A population-structured HIV epidemic in Israel: roles of risk and ethnicity. *PLoS One* 2015; 10, e0135061.

Hahn, B.H. et al. (2000) AIDS as a zoonosis: scientific and public health implications. *Science* 287, 607-614.

Hall BG. Building phylogenetic tree from molecular data with MEGA. *Mol Biol Evol*. 2013;30: 1229-1235.

Halvas EK, Svarovskaia ES, Pathak VK 2000. Role of murine leukemia virus reverse transcriptase deoxyribonucleoside triphosphate-binding site in retroviral replication and in vivo fidelity. *J Virol* 74: 10349-10358.

Hartley O, Klasse PJ, Sattentau QJ, Moore JP. V3: HIV's switch-hitter. *AIDS Res Hum Retroviruses*. 2005; 21(2):171-189.

Hemelaar, J., Gouws, E., Ghys, P.D., Osmanov, S., Isolation, WHO-UNAIDS Network for HIV Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* 2011; 25, 679-689.

Hemelaar J, Gouws E, Ghys PD, Osmanov S; WHO-UNAIDS Network for HIV Isolation and Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* 2011; 25(5):679-689.

Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS* 2006; 20(16):W13-23.

Henriet S, Mercenne G, Bernacchi S, Paillart JC, Marquet R. Tumultuous relationship between the human immunodeficiency virus type 1 viral infectivity factor (Vif) and the human APOBEC-3G and APOBEC-3F restriction factors. *Microbiol Mol Biol Rev* 2009; 73:211-232.

Hu W, Hughes S. HIV-1 reverse transcription. *Cold Spring Harb Perspect Med*. 2012 ; 2(10).

Hue, S., Clewley, J.P., Cane, P.A., Pillay, D., 2004. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* 18, 719-728.

Hué S, Pillay D, Clewley JP, Pybus OG. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A*. 2005; 102(12):4425-9.

Huelsenbeck JP, Ronquist F. MRBAYES. Bayesian inference of phylogenetic tree. 2001;17: 754-755.

Hung M, Patel P, Davis S, Green SR. Importance of ribosomal frameshifting for human immunodeficiency virus type 1 particle assembly and replication. *J Virol*. 1998; 72(6):4819-4824.

Hymes, K. B., Cheung, T., Greene, J. B., Prose, N. S., Marcus, A., Ballard, H., William, D. C., and Laubenstein, L. J. (1981). Kaposi's sarcoma in homosexual men—a report of eight cases. *Lancet*, 2(8247):598–600.

Jacks T, Power M D, Masiarz F R, Luciw P A, Barr P J, Varmus H E. Characterization of ribosomal frameshifting in HIV-1 *gag-pol* expression. *Nature*. 1988; 331:280–283.

Janssens W, Heyndrickx L, Franssen K, Motte J, Peeters M, Nkengasong JN, Ndumbe PM, Delaporte E, Perret JL, Atende C, et al. Genetic and phylogenetic analysis of env subtypes G and H in central Africa. *AIDS Res Hum Retroviruses*. 1994;10(7):877-9.

Johnson VA, Calvez V, Gunthard HF, Paredes R, Pillay D, Shafer RW, Wensing AM, Richman DD. Update of the drug resistance mutations in HIV-1: 2013. *Top Antivir Med*. 2013 Feb-Mar;21(1):6-614.

Jeang KT, Xiao H, Rich EA. Multifaceted activities of the HIV-1 transactivator of transcription, Tat. *J Biol Chem*. 1999 ;274(41):28837-28840.

Kanki PJ, McLane MF, King NW Jr, Letvin NL, Hunt RD, Sehgal P, et al. Serologic identification and characterization of a macaque T lymphotropic retrovirus closely related to HTLV-III. *Science*. 1985 ;228(4704):1199-1201.

Kiepiela P, Ngumbela K, Thobakgale C, Ramduth D, Honeyborne I, Moodley E, et al. CD8⁺ T-cell responses to different HIV proteins have discordant associations with viral load. *Nat Med*. 2007;13(1):46-53.

Kitkungvan D, Apisarnthanarak A, Laowansiri P, Mundy LM. Secure antiretroviral therapy delivery in a resource-limited setting: streamlined to minimize drug resistance and expense. *HIV Med*. 2008;9(8):636-641.

Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, et al. (2006) Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 313: 523-526.

Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A.* 2008; 105(21):7552-7557.

Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science.* 2006; 313(5786):523-526.

Kentikelenis, A., Karanikolos, M., Williams, G., Mladovsky, P., King, L., Pharris, A., et al. How do economic crises affect migrants' risk of infectious disease? A systematic-narrative review. *Eur. J. Pub. Health.* 2015; 25, 937-944.

Klasse PJ. The molecular basis of HIV entry. *Cell Microbiol.* 2012;14(8):1183-1192.

Kouyos RD, von Wyl V, Yerly S, Böni J, Rieder P, Joos B, Taffé P. Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. *Clin Infect Dis.* 2011;52(4):532-539.

Kouyos RD, von Wyl V, Yerly S, Böni J, Taffé P, Shah C, et al. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J Infect Dis.* 2010; 201(10):1488-1497.

Kostrikis LG, Bagdades E, Cao Y, Zhang L, Dimitriou D, Ho DD. Genetic analysis of human immunodeficiency virus type 1 strains from patients in

Cyprus: identification of a new subtype designated subtype I. *J Virol.* 1995;69(10):6122-30.

Kwong PD, Wyatt R, Majeed S, Robinson J, Sweet RW, Sodroski J, Hendrickson WA. Structures of HIV-1 gp120 envelope glycoproteins from laboratory-adapted and primary isolates. *Structure.* 2000 ; 8(12):1329-1339.

Laukkanen T, Albert J, Liitsola K, Green SD, Carr JK, Leitner T, McCutchan FE, Salminen MO. Virtually full-length sequences of HIV type 1 subtype J reference strains. *AIDS Res Hum Retroviruses.* 1999 Feb 10;15(3):293-7

Lee HY, Giorgi EE, Keele BF, Gaschen B, Athreya GS, Salazar-Gonzalez JF, et al. Modeling sequence evolution in acute HIV-1 infection. *J Theor Biol.* 2009; 261(2):341-60.

Leitner T., Albert J. Reconstruction of HIV-1 transmission chains for forensic purposes. *AIDS Rev.* 2000; 2(4): 241–251.

Lemey P, Vandamme AM. Exploring full-genome sequences for phylogenetic support of HIV-1 transmission events. *AIDS.* 2005;19(14):1551-1552.

Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, Roques P, et al. The molecular population genetics of HIV-1 group O. *Genetics* 2004,167:1059-1068.

Lemey P, Pybus O, Wang B, Saksena NK, Salemi M, Vandamme AM. Tracing the origin and history of the HIV-2 epidemic. *Proc Natl Acad Sci U S A.* 2003; 100(11):6588-6592.

Lemey P, Salemi M, Vandamme AM, editors. *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing.* Cambridge: 2009.

Lemey P, Derdelinckx I, Rambaut A, Van Laethem K, Dumont S, Vermeulen S, et al. Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J Virol.* 2005; 79(18):11981-11989.

Leoz M, Feyertag F, Kfutwah A, Mauclère P, Lachenal G, Damond F et al. Emergence of Non Pandemic HIV-1 Group O in Cameroon. *PLoS Pathog.* 2015; 11(8):e1005029.

Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* 2008; 5(3):e50.

Li G, Piampongsant S, Faria NR, Voet A, Pineda-Peña AC, Khouri R, Lemey P, Vandamme AM, Theys K. An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology.* 2015 ;12:18.

Li Y, Naidu YM, Daniel MD, Desrosiers RC. Extensive genetic variability of simian immunodeficiency virus from African green monkeys. *J Virol.* 1989; 63(4):1800-1802.

Lightfoote M, Coligan J, Folks T, Fauci A, Martin M, Venkatesan S. Structural characterization of reverse transcriptase and endonuclease polypeptides of the acquired immunodeficiency syndrome retrovirus. *J Virol.* 1986;60(2):771-775.

Luo K, Liu B, Xiao Z, Yu Y, Yu X, Gorelick R, *et al.* Amino-terminal region of the human immunodeficiency virus type 1 nucleocapsid is required for human APOBEC3G packaging. *J Virol.* 2004; 78:11841-11852.

Lu X, Yu Q, Binder GK, Chen Z, Slepushkina T, Rossi J, Dropulic B. Antisense-mediated inhibition of human immunodeficiency virus (HIV) replication by use

of an HIV type 1-based vector results in severely attenuated mutants incapable of developing resistance. *J Virol.* 2004; 78(13):7079-7088.

Machado ES, Afonso AO, Nissley DV, Lemey P, Cunha SM, Oliveira RH, et al. Emergence of primary NNRTI resistance mutations without antiretroviral selective pressure in a HAART-treated child. *PLoS One.* 2009; 4(3):e4806.

Malim MH. APOBEC proteins and intrinsic resistance to HIV-1 infection. *Philos Trans R Soc Lond B Biol Sci.* 2009;364:675-677.

Marx J. Strong new candidate for AIDS agent. *Science* 1984; 224(4648): 475-477.

Masur H, Michelis MA, Greene JB, Onorato I, Stouwe RA, Holzman RS et al. An outbreak of community-acquired *Pneumocystis carinii* pneumonia: initial manifestation of cellular immune dysfunction. *N Engl J Med.* 1981; 305(24):1431-1438.

Marx PA, Alcabes PG, Drucker E. Serial human passage of simian immunodeficiency virus by unsterile injections and the emergence of epidemic human immunodeficiency virus in Africa. *Philos Trans R Soc Lond B Biol Sci.* 2001;356(1410):911-920.

Masur H, Michelis MA, Wormser GP, Lewin S, Gold J, Tapper ML. Opportunistic infection in previously healthy women. Initial manifestations of a community-acquired cellular immunodeficiency. *Ann Intern Med.* 1982; 97(4):533-539.

Maddison WP. Testing character correlation using pairwise comparisons on a phylogeny. *J Theor Biol.* 2000; 202(3):195-204.

McGovern RA, Harrigan PR, Swenson LC. Genotypic inference of HIV-1 tropism using population-based sequencing of V3. *J Vis Exp.* 2010: 27;

McMichael A, Klenerman P. HIV/AIDS. HLA leaves its footprints on HIV. *Science.* 2002;296(5572):1410-1411.

Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM. Molecular evidence of HIV-1 transmission in a criminal case. *Proc Natl Acad Sci U S A.* 2002; 99: 14292-14297.

Mourez T, Simon F, Plantier JC. Non-M variants of human immunodeficiency virus type 1. *Clin Microbiol Rev.* 2013; 26(3):448-461.

Narrative Progress Report on HIV/AIDS Response of Republic of Serbia in 2015. <http://www.batut.org.rs/index.php?content=1396>

Neogi U, Bontell I, Shet A, De Costa A, Gupta S, Diwan V, et al. Molecular epidemiology of HIV-1 subtypes in India: origin and evolutionary history of the predominant subtype C. *PLoS One.* 2012;7(6):e39819.

Oleske J, Minnefor A, Cooper R J, Thomas K, dela Cruz A, Ahdieh H. Immune deficiency syndrome in children. *JAMA.* 1983; 249(17):2345-2359.

Organization, W.H., 2007. AIDS epidemic update, December 2006. World Health Organization

Osmanov S, Pattou C, Walker N, Schwardländer B, Esparza J; WHO-UNAIDS Network for HIV Isolation and Characterization.. Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. *J Acquir Immune Defic Syndr.* 2002; 29(2):1841-90.

Palca, J., 1992b. The case of the Florida dentist. *Science* 255, 392-394.

Palmer S, Vuitton D, Gonzales MJ, Bassignot A, Shafer RW. Reverse transcriptase and protease sequence evolution in two HIV-1-infected couples. *J Acquir Immune Defic Syndr*. 2002; 31(3):285-290.

Paraskevis D, Pybus O, Magiorkinis G, Hatzakis A, Wensing AM, van de Vijver M et al. SPREAD Programme. Tracing the HIV-1 subtype B mobility in Europe: a phylogeographic approach. *Retrovirology*. 2009;6:49.

Paraschiv S, Otelea D, Batan I, Baicus C, Magiorkinis G, Paraskevis D. Molecular typing of the recently expanding subtype B HIV-1 epidemic in Romania: evidence for local spread among MSMs in Bucharest area. *Infect Genet Evol*. 2012;12(5):1052-1057.

Pineda-Peña A, Faria N, Imbrechts S, Libin P, Abecasis A, Deforche K et al.. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect Genet Evol*. 2013; 19:337-348.

Peeters M, Jung M, Ayouba A. The origin and molecular epidemiology of HIV. *Expert Rev Anti Infect Ther*. 2013; 11(9):885-896.

Peterman T, Jaffe H, Feorino P, Getchell J, Warfield D, Haverkos H. Transfusion-associated acquired immunodeficiency syndrome in the United States. *JAMA*. 1998; 254(20):2913-2917.

Posada D. JModelTest: phylogenetic model averaging. *Mol Biol Evol*. 2008; 25:251253-251256.

Poulsen AG, Aaby P, Jensen H, Dias F. Risk factors for HIV-2 seropositivity among older people in Guinea-Bissau. A search for the early history of HIV-2 infection. *Scand J Infect Dis*. 2000; 32(2):169-175.

Ratner L, Haseltine W, Patarca R, Livak K J, Starcich B, Josephs S F, et al. Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature*. 1985; 313:277-284.

Rajasekaran S, Jeyaseelan L, Vijila S, Gomathi C, Raja K. Predictors of failure of first-line antiretroviral therapy in HIV-infected adults: Indian experience. *AIDS*. 2007;21 Suppl 4:S47-53.

Ramachandran S, Xia GL, Ganova-Raeva LM, Nainan OV, Khudyakov Y. End-point limiting-dilution real-time PCR assay for evaluation of hepatitis C virus quasispecies in serum: performance under optimal and suboptimal conditions. *Virology Methods*. 2008; 151: 217-224.

Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. *Nat Rev Genet*. 2004; 5: 52-61.

Ratner L, Gallo C, Wong-Staal, F. HTLV-III, LAV, ARV are variants of same AIDS virus. *Nature* 1985a; 313, 636-637.

Recordon-Pinson P, Anies G, Bruyand M, Neau D, Morlat P, Pellegrin JL, et al. ANRS CO3 Aquitaine Cohort. HIV type-1 transmission dynamics in recent seroconverters: relationship with transmission of drug resistance and viral diversity. *Antivir Ther*. 2009; 14(4):551-556.

Riminton S. "Basic Science & Pathogenesis," in *Distance Learning for HIV Medicine*. Australia: ASHM, 2004.

Rizzuto CD, Wyatt R, Hernández-Ramos N, Sun Y, Kwong PD, Hendrickson WA, et al. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science*. 1998;280(5371):1949-1953.

Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, et al. HIV-1 nomenclature proposal. *Science*. 2000;288(5463):55-6.

Robbins GK, Daniels B, Zheng H, Chueh H, Meigs JB, Freedberg KA. Predictors of antiretroviral treatment failure in an urban HIV clinic. *J Acquir Immune Defic Syndr*. Jan 1 2007; 44(1):30-37.

Robbins KE, Lemey P, Pybus OG, Jaffe HW, Youngpairoj AS, Brown TM, et al. Human immunodeficiency virus type 1 epidemic: date of origin, population history, and characterization of early strains. *J Virol*. 2003; 77(11):6359-6366.

Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. *Proc Natl Acad Sci U S A*. 2016; 113(10):2690-2695.

Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*. 2003; 19(12):1572-1574.

Roques P, Robertson DL, Souquière S, Damond F, Ayoub A, Farfara I et al. Phylogenetic analysis of 49 newly derived HIV-1 group O strains: high viral diversity but no group M-like subtype structure. *Virology*. 2002; 302(2):259-273.

Saludes V, Esteve M, Casas I, Ausina V, Martró E. Hepatitis C virus transmission during colonoscopy evidenced by phylogenetic analysis. *J Clin Virol*. 2013; 57: 263-266.

Santiago ML, Rodenburg CM, Kamenya S, Bibollet-Ruche F, Gao F, et al. (2002) SIVcpz in wild chimpanzees. *Science* 295: 465.

Santoro MM, Perno CF. HIV-1 Genetic Variability and Clinical Implications.

ISRN Microbiol. 2013;2013:481314. Saxon, A., 1981. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men:

evidence of a new acquired cellular immunodeficiency. *N. Engl. J. Med.* 305, 1425-1431.

Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM, Metzker ML. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proc Natl Acad Sci U S A.* 2010; 107(50):21242-21247.

Schlub TE, Smyth RP, Grimm AJ, Mak J, Davenport MP. Accurately measuring recombination between closely related HIV-1 genomes. *PLoS Comput Biol.* 2010;6(4)

Sharp, P.M., Hahn, B.H., 2011. Origins of HIV and the AIDS pandemic. *Cold Spring Harb. Perspect. Med.* 1, a006841.

Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* 2002; 418:646-650.

Simon F, Maucière P, Roques P, Loussert-Ajaka I, Müller-Trutwin MC, Saragosti

Simon F, Maucière P, Roques P, Loussert-Ajaka I, Müller-Trutwin MC, Saragosti S, et al. Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. *Nat Med.*1998; 4(9):1032-1037.

Sever J., "HIV: biology and immunology," *Clin Obstet Gynecol*, vol. 32, pp. 423-8, 1989.

Snoeck J, Riva C, Steegen K, Schrooten Y, Maes B, Vergne L, Van Laethem K, Peeters M, Vandamme AM. Optimization of a genotypic assay applicable to all human immunodeficiency virus type 1 protease and reverse transcriptase subtypes. *J Virol Methods.* 2005; 128(1-2):47-53.

Soriano V, Gomes P, Heneine W, Holguín A, Doruana M, Antunes R., et al. Human immunodeficiency virus type 2 (HIV-2) in Portugal: clinical spectrum, circulating subtypes, virus isolation, and plasma viral load. *J. Med. Virol.* 2000; 61, 111-116.

Sonigo P, Alizon M., Staskus K., Klatzmann D, Cole S, Danos O, Retzel E, Tiollais P, Haase A., Wain-Hobson, S. Nucleotide sequence of the visna lentivirus: relationship to the AIDS virus. *Cell.* 1985; 42(1):369-382.

Stürmer M, Berger A, Preiser W. HIV-1 genotyping: comparison of two commercially available assays. *Expert Rev Mol Diagn.* 2004; 4(3):281-291.

Swofford DL (2003) PAUP*: phylogenetic analysis using parsimony (*and other methods), version 4 beta. Sinauer Associates, Sunderland, MA.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 2011; 28(10):2731-2739.

Tienen Cv, van der Loeff MS, Zaman SM, Vincent T, Sarge-Njie R, Peterson I, et al. Two distinct epidemics: the rise of HIV-1 and decline of HIV-2 infection between 1990 and 2007 in rural Guinea-Bissau. *J Acquir Immune Defic Syndr.* 2010; 53(5):640-647.

Tebit DM, Arts EJ. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *Lancet Infect Dis* 2011,11:45-56.

Thomson MM, Nájera R. Travel and the introduction of human immunodeficiency virus type 1 non-B subtype genetic forms into Western countries. *Clin Infect Dis.* 2001; 32(12):1732-1737.

Triques K, Bourgeois A, Vidal N, Mpoudi-Ngole E, Mulanga-Kabeya C, Nzilambi N, Torimiro N, Saman E, Delaporte E, Peeters M. Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K. *AIDS Res Hum Retroviruses*. 2000;16(2):139-151.

Vallari A, Bodelle P, Ngansop C, Makamche F, Ndembi N, Mbanya D, Kaptué L, et al. Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. *AIDS Res Hum Retroviruses*. 2010;26(1):109-115.

Vallari A, Holzmayer V, Harris B, Yamaguchi J, Ngansop C, Makamche F, et al. Confirmation of putative HIV-1 group P in Cameroon. *J Virol*. 2011;85(3):1403-1407.

Vallari A, Bodelle P, Ngansop C, Makamche F, Ndembi N, Mbanya D, et al. Four new HIV-1 group N isolates from Cameroon: Prevalence continues to be low. *AIDS Res Hum Retroviruses*. 2010; 26(1):109-115.

Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, Liu W, Loul S, Butel C, Liegeois F, Bienvenue Y, Ngolle EM, Sharp PM, Shaw GM, Delaporte E, Hahn BH, Peeters M. Human immunodeficiency viruses: SIV infection in wild gorillas. *Nature* 2006; 444(7116):164.

Visseaux B, Damond F, Matheron S, Descamps D, Charpentier C. Hiv-2 molecular epidemiology. *Infect Genet Evol*. 2016; 46:233-240.

Yerly S, Junier T, Gayet-Ageron A, Amari EB, von Wyl V, Günthard HF, Hirschel B, Zdobnov E, Kaiser L; Swiss HIV Cohort Study.. The impact of transmission clusters on primary drug resistance in newly diagnosed HIV-1 infection. *AIDS*. 2009; 23(11):1415-1423.

Yusim K, Peeters M, Pybus OG, Bhattacharya T, Delaporte E, Mulanga C, et al. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. *Philos Trans R Soc Lond B Biol Sci.* 2001;356(1410):855-866.

Weait, M. The Criminalisation of HIV Exposure and Transmission: A Global Review. Working Paper prepared for the Third Meeting of the Technical Advisory Group, Global Commission on HIV and the Law, 7-9 July, 2011.

Weidle PJ, Malamba S, Mwebaze R, Sozi C, Rukundo G, Downing R, et al. Assessment of a pilot antiretroviral drug therapy programme in Uganda: patients' response, survival, and drug resistance. *Lancet* 2002; 360(9326):34-40.

Wensing A, Calvez V, Günthard H, Johnson V, Paredes R, Pillay D, Shafer RW, Richman DD. 2014 Update of the drug resistance mutations in HIV-1. *Top Antivir Med.* 2014; 22(3):642-650.

Whelan S, Morrison DA. Inferring Trees. *Methods Mol Biol.* 2017; 1525:349-377.

Williamson C, Morris L, Maughan MF, Ping LH, Dryga SA, Thomas R, Reap EA, Korber B, Muldoon M, Theiler J, Gao F, Gupta R, et al. (2000) Timing the ancestor of the HIV-1 pandemic strains. *Science* 288: 1789-1796.

Wertheim JO, Worobey M. Dating the Age of the SIV Lineages That Gave Rise to HIV-1 and HIV-2. *Plos Comput Biol* 2009,5.

Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M, et al. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature.* 2008;455(7213):661-664.

Wu L, Gerard NP, Wyatt R, Choe H, Parolin C, Ruffing N. CD4-induced interaction of primary HIV-1 gp120 glycoproteins with the chemokine receptor CCR-5. *Nature*. 1996; 384(6605):179-183.

Zehender, G., Ebranati, E., Lai, A., Santoro, M.M., Alteri, C., Giuliani, M., Palamara, G., Perno, C.F., Galli, M., Lo, P.A., Ciccozzi, M., 2010. Population dynamics of HIV-1 subtype B in a cohort of men-having-sex-with-men in Rome, Italy. *J. Acquir. Immune Defic. Syndr.* 55, 156-160.

Zheng D-P, Rodrigues M, Bile E, et al.: Molecular characterization of ambiguous mutations in HIV-1 polymerase gene: Implications for monitoring HIV infection status and drug resistance. *PLoS One* 2013;8(10):e77649.

Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM, Ho DD. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* 1998; 391(6667):594-597.

LIST OF ABBREVIATIONS AND ACRONYMS

AIDS	Acquired Immunodeficiency Syndrome
Afsu	Subtype A Former Soviet Union
aa	Amino Acid
BLAST	Basic Local Alignment Search Tool
CDC	Centers for Disease Control and Prevention
CD4	Cluster of Differentiation 4
CCR5	C-C chemokine receptor type 5
CXCR4	C-X-C chemokine receptor type 4
CRF	Circulating Recombinant Forms
DNA	Deoxyribonucleic Acid
DRC	Democratic Republic of Congo
dNTP	deoxyribonucleotide triphosphate
<i>env</i>	Envelope
EPLD	End Point Limited Dilution
FASTA	text-based format for representing nucleotide sequences
<i>gag</i>	group specific antigen gene
<i>gp</i>	glycoprotein
GRID	Gay Related Immune Deficiency
GTR	Generalised time reversible model

GTR+G+I	GTR plus gamma (Γ) and proportion invariant model
HAART	Highly Active Antiretroviral Therapy
HHV-8	Human Herpesvirus 8
HIV	Human Immunodeficiency Virus
HIV-1	Human Immunodeficiency Virus type 1
HIV-2	Human Immunodeficiency Virus type 2
HIVdb	Stanford HIV Drug Resistance Database
HLA	Human Leukocyte Antigen
HPD	Highest posterior density
HTLV-III	Human T-Lymphocyte Virus III
IN	Integrase
LAS	Lymphadenopathy Syndrome
LAV	Lymphadenopathy-Associated Virus
LANL	Los Alamos National Laboratories
LTR	Long Terminal Repeats
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood
MTCT	Mother to Child Transmission
MSM	Men who have Sex with Men
NCBI	National Center for Biotechnology Information
PBMC	Peripheral Blood Mononuclear Cells

PCR	Polymerase chain reaction
PIs	Protease Inhibitors
pol	Polymerase gene
PR	Protease
PWID	People Who Inject Drugs
REGAv3	REGA automated subtyping tool version 3
RNA	Ribonucleic Acid
RT	Reverse Transcriptase
SIV	Simian Immunodeficiency Virus
STI	Sexually Transmitted Infection
tMRCA	time to Most Recent Common Ancestor
UNAIDS	Joint United Nations Programme on HIV/AIDS
WHO	World Health Organisation

APPENDIX 1. Table AI. Accession numbers and country of origins of *pol* gene sequences from Serbian isolates sampled from 1997 to 2007 and foreign isolates obtained by NCBI database.

POL GENE					COUNTRY OF ORIGIN	
ACCESSION NUMBER						
AY433800;	AY433798;	AY433797;	AY433796;	AY433795;	Serbia	
AY433794;	AY433793;	AY433792;	GQ399763;	GQ399551;		
GQ400327;	GQ400482;	GQ398972;	GQ399179;	GQ399012;		
GQ399888;	GQ399221;	GQ399018;	GQ480327;	GQ399263;		
GQ395533;	GQ400459;	GQ400490;	GQ400092;	GQ399955;		
GQ400505;	GQ39934;	GQ399810;	GQ398855;	GQ400529;		
GQ400303;	GQ400203;	GQ400169;	GQ399328;	GQ398698;		
GQ399770;	GQ399505;	GQ399605;	GQ399463;	GQ399262;		
GQ400380;	GQ399684;	GQ399526;	GQ400192;	GQ399335;		
GQ400562;	GQ400664;	GQ400943;	GQ400568;	GQ400867;		
GQ399293;	GQ400985;	GQ400636;	GQ400934;	GQ400637;		
GQ400971;	GQ400727;	GQ400860;	GQ400842;	GQ400711;		
GQ400847;	GQ400975;	GQ401005;	GQ400863;	GQ399151;		
GQ400634;	GQ400696;	GQ400623;	GQ400576;	GQ400698;		
GQ400683;	JX299860;	JX300595;	JX300466;	JX301157;		
JX300670;	JX300934;	JX300963;	JX301026;	JX299883;		
JX299967;	JX300342;	JX300698;	JX301113;	JX299941;		
JX300732;	KF157408-	KF157434				
AY611666;	AY611672;	AY611684;	AY611688			Albania
AF347214;	AF347518;	DQ878531;	DQ878532			Austria
DQ177230;	DQ177232;	DQ877759;	FJ653084;	EU248460	Belgium	
DQ177224;	DQ177231;	DQ177227;	DQ177234.1;	DQ177218.1;		
KF301720.;	DQ177231.1					
EF517439;	EF517457;	EF517462;	EF517464;	EF517488;	Bulgaria	
EF517472;	EF517439;	EF517410				
FN424300;	FN424301				Croatia	
EU673375;	EU673382;	EU673408			Cyprus	
AY694218;	AY694233;	AY694364			Czech Republic	
AJ419453;	AJ582147;	AM490879;	DQ108366;	DQ87779	Denmark	

AF487122; DQ878075; DQ877953; DQ877930	France
AF347190; AF347288; AF347140; AY878668; AY878677; DQ878276; DQ878304 ; FJ030769; GQ400800; KC340462.1; KX467180.1; KX467175.1; KX467167.1; KX467170.1; KX467149.1	Germany
DQ878544; DQ878548; DQ878559; DQ878569; DQ878595; EF563173	Greece
AM285220; AM285242; AM285267; AM937019; AM937024	Greenland
DQ877830; DQ877832	Ireland
AY375051; AY362443; DQ348057; DQ348033; DQ345139; DQ345123; DQ345246; DQ345123; DQ345262; DQ345233; DQ345221; AF251947; AF252026; AF376547; AF493371; AF517266; AF517471; AY672455; AY352444; AY855419; AY855724; AY994341; AY995503; DQ345170; DQ345265; DQ369253; DQ672623; DQ878603; EF526205; EU019810; EU496146; FJ228037; FJ228081; FJ228038; FJ209055; FJ209061; FJ228131; FJ228123; FJ228127; HM990501.1; HM990500.1; HM990485.1; HM990469.1; HM990443.1; HM990432.1; FJ228038.1 FJ228036.1; GU969575.1; GU969555.1; GU969546.1; GU969500.1	Italy
DQ877749; EF563190	Luxembourg
AY423387; AY423383; AY877314; DQ877839; U34604; GQ399672; DQ877848	Netherlands
DQ663718; DQ663758; DQ666409; DQ666416; DQ877854; DQ877866; DQ877875; EF563195	Portugal
JN982165; JN982159, JN982157; JN982198; JN982190; JN982162; JN982158	Romania
JX290245; JX290243; JX290239; JX290244; JX290242; KX517406; KX517405; KX517394; KX517388; KX46528; KX465287; KX465276; KX465274	Russia
AJ971111; AJ971144	Slovenia
AF256207; AJ006287; AY188561; AY315950; AY316009; AY541992; AY542131; AY833593; DQ103904; DQ878901; DQ878956; EF397445; EF583196; EF583217; EF583285; EU255372; EU255417; EU255511; EU252228; EU786674; JQ828989; JX140659; JX140674	Spain
AF368317; AF378391; AF394468; AY165277; AY165231; DQ877891	Sweden

AF077679; AF316851; DQ877898; EF449825; EF449848; GU344240; GU344255	Switzerland
EF517434; DQ013272.1; DQ013269	South Africa
AF181126; AF494109; AY362127; DQ879066; EU236465; EU817055	United Kingdom
HQ115072; HQ115066, DQ055312; DQ055292; DQ055331; DQ055325; DQ055321	Ukraine
JX161447; JX161313; JX161215; JX161335; JX161079; JX161023; JX160755; JX160671; JX160659; JX161424; JX161084; GU331235; GU331229	USA

APPENDIX 2. **Table AII.** Accession numbers and country of origins of 60 pol gene and 60 env gene background control sequences obtained by NCBI database BLAST.

POL GENE		ENV GENE	
ACCESSION NUMBER	COUNTRY OF ORIGIN	ACCESSION NUMBER	COUNTRY OF ORIGIN
HQ655137.1	Montenegro	HQ595795.1	United Kingdom
KT168102.1	USA	DQ410258.1	USA
KT168100.1	USA	AY614970.1	USA
KT168110.1	USA	KX129198.1	USA
KT168108.1	USA	FJ222419.1	USA
AY835771.1	USA	AY614940.1	USA
KT168168.1	USA	JQ609969.1	USA
KT168111.1	USA	JQ609965.1	USA
KT168097.1	USA	FJ222418.1	USA
AY835770.1	USA	AY614939.1	USA
AY835755.1	USA	AY247218.1	USA
KT168124.1	USA	GQ118605.1	USA
KT168109.1	USA	KX129207.1	USA
KT167927.1	USA	EU184348.1	USA
KT168104.1	USA	AY614905.1	USA
KT168124.1	USA	KX129197.1	USA
JQ650584.1	Netherlands	JQ609968.1	USA
JX299701.1	Germany	HQ231110.1	Netherlands
KX465993.1	Germany	HQ231101.1	Netherlands
KX465994.1	Germany	HQ231099.1	Netherlands
KJ771209.1	Germany	EU744029.1	Netherlands
KJ770207.1	Germany	EU744020.1	Netherlands

JX299701.1	Germany	HQ644937.1	Netherlands
JX300212.1	Austria	EU744034.1	Netherlands
KP013647.1	Slovenia	EU744028.1	Netherlands
GU807516.1	Venezuela	EU744023.1	Netherlands
JQ698863.1	Sweden	HQ644910.1	Netherlands
AF251949.1	Italy	GU455498.1	Netherlands
KT167848.1	Canada	EU744043.1	Netherlands
GU344259.1	Switzerland	EU744019.1	Netherlands
AF447821.1	Thailand	AY669729.1	Thailand
KF745481.1	Thailand	U08801.1	Thailand
AF447830.1	Thailand	DQ354116.1	Thailand
KF745264.1	Thailand	DQ354120.1	Thailand
KF745563.1	Thailand	DQ354114.1	Thailand
JN248343	Thailand	JN248351.1	Thailand
AF362994.1	Thailand	JN248350.1	Thailand
KF745691.1	Thailand	KC749047.1	Thailand
DQ354119.1	Thailand	HM215396.1	Thailand
AY713408.1	Thailand	AY945710.1	Thailand
KF745662.1	Thailand	KJ953227.1	Thailand
KF745673.1	Thailand	KJ953201.1	Thailand
AF191194.1	Thailand	DQ354119.1	Thailand
AF345971.1	Thailand	U08802.1	Thailand
AF191205.1	Thailand	KJ952538.1	Thailand
KF745619.1	Thailand	KC749034.1	Thailand
DQ354118.1	Thailand	DQ354118.1	Thailand
KF745670.1	Thailand	AF209201.1	Thailand
KF745355.1	Thailand	JQ715413.1	Thailand
KP109514.1_	Thailand	JX446822.1	Thailand
JX448097.1	Thailand	JX446817.1	Thailand

HQ898636.1	Thailand	JN248325.1	Thailand
DQ354116.1	Thailand	KC749040.1	Thailand
DQ354112.1	Thailand	KC749036.1	Thailand
KF745513.1	Thailand	KC749033.1	Thailand
KF745225.1	Thailand	KC749054.1	Thailand
KF745501.1	Thailand	KC749048.1	Thailand
KF745379.1	Thailand	AF209204.1	Thailand
AF240383.1	Thailand	AF209203.1	Thailand
AF447821.1	Thailand	KJ953200	Thailand

BIOGRAPHY

Marina Šiljić was born on the 10th of November, 1982 in Belgrade, where she finished Elementary and High School. She has graduated at the Faculty of Biology, University of Belgrade, study program Biology of Microorganisms, with average mark of 9.41/10.0. She defended her Master's thesis that included investigation of the genotoxicity of PET bottles. During the last years of her study and after graduation, she volunteered at the Institute for Biological Research "Siniša Stanković" in the laboratory of Genetic research.

After graduation, Marina Šiljić has enrolled the PhD studies, study program "Molecular Medicine" at the School of Medicine, University of Belgrade in 2010. From 2011 she has been employed as a research assistant on the scientific project no 175024 entitled "Phylogenetic analysis and molecular evolution in highly variable viruses: coinfections, host-pathogene interaction" funded by Ministry of Education, Science and Technological Development of the Republic of Serbia. Throughout her work, she has had a continuous interest in research, particularly in the field of the phylogenetic analysis and molecular evolution of highly variable viruses.

Marina Šiljić is the coauthor of 12 articles indexed in Current Contents-u (CC) or Science Citation Index (SCI). In four articles she is the first author.

BIOGRAFIJA

Marina Šiljić rođena je 10. novembra 1982. godine u Beogradu, gde je završila osnovnu i srednju školu. Diplomirala je na Biološkom fakultetu, Univerziteta u Beogradu, na studijskom program biologija mikroorganizama, sa prosečnom ocenom 9.41/10.0. Odbranila je diplomski rad koji je obuhvatio istraživanje genotoksičnosti PET ambalaže. Tokom poslednje godine njenih studija kao i nakon diplomiranja volontirala je u Institutu za Biološka istraživanja "Siniša Stanković", u laboratoriji za genetičko istraživanje.

Nakon diplomiranja, Marina Šiljić upisala je 2010. godine doktorske studije na Medicinskom fakultetu Univerziteta u Beogradu, na studijskom programu "Molekularna medicina". Od 2011. godine zaposlena je na Institutu za mikrobiologiju i imunologiju, Medicinskog fakulteta Univerziteta u Beogradu kao istraživač saradnik na naučno istraživačkom projektu broj 175024 pod nazivom "Filogenetski pristup analizi molekularne evolucije visoko varijabilnih virusa – koinfekcije, interakcija virusa i domaćina" koji je finansiran od strane Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije.

Marina Šiljić autor je koautor 12 radova koji su koji je indeksirani u Current Contents-u (CC) ili Science Citation Index-u (SCI). Prvi je autor u četiri rada.

Prilog 1.

Izjava o autorstvu

Potpisani-a Marina Šiljić

broj upisa MM04/10

Izjavljujem

da je doktorska disertacija pod naslovom:

„Filogenetska analiza i molekularna karakterizacija virusa humane imunodeficijencije u Srbiji“

- rezultat sopstvenog istraživačkog rada,
- da predložena disertacija u celini ni u delovima nije bila predložena za dobijanje bilo koje diplome prema studijskim programima drugih visokoškolskih ustanova,
- da su rezultati korektno navedeni i
- da nisam kršio/la autorska prava i koristio intelektualnu svojinu drugih lica.

Potpis doktoranda

U Beogradu, 01.03.2017. godine



Handwritten signature of Marina Šiljić, written in black ink over a horizontal line.

Izjava o istovetnosti štampane i elektronske verzije doktorskog rada

Ime i prezime autora Marina Šiljić

Broj upisa MM04/10

Studijski program Molekularna medicina

Naslov rada „Filogenetska analiza i molekularna karakterizacija virusa humane imunodeficijencije u Srbiji“

Mentor Prof. dr Maja Stanojević

Potpisani Marina Šiljić


izjavljujem da je štampana verzija mog doktorskog rada istovetna elektronskoj verziji koju sam predao/la za objavljivanje na portalu Digitalnog repozitorijuma Univerziteta u Beogradu.

Dozvoljavam da se objave moji lični podaci vezani za dobijanje akademskog zvanja doktora nauka, kao što su ime i prezime, godina i mesto rođenja i datum odbrane rada.

Ovi lični podaci mogu se objaviti na mrežnim stranicama digitalne biblioteke, u elektronskom katalogu i u publikacijama Univerziteta u Beogradu.

Potpis doktoranda

U Beogradu, 01.03.2017. godine



Prilog 3.

Izjava o korišćenju

Ovlašćujem Univerzitetsku biblioteku „Svetozar Marković“ da u Digitalni repozitorijum Univerziteta u Beogradu unese moju doktorsku disertaciju pod naslovom:

„Filogenetska analiza i molekularna karakterizacija virusa humane imunodeficijencije u Srbiji“

koja je moje autorsko delo.

Disertaciju sa svim priložima predao/la sam u elektronskom formatu pogodnom za trajno arhiviranje.

Moju doktorsku disertaciju pohranjenu u Digitalni repozitorijum Univerziteta u Beogradu mogu da koriste svi koji poštuju odredbe sadržane u odabranom tipu licence Kreativne zajednice (Creative Commons) za koju sam se odlučio/la.

1. Autorstvo
2. Autorstvo - nekomercijalno
3. Autorstvo – nekomercijalno – bez prerade
4. Autorstvo – nekomercijalno – deliti pod istim uslovima
5. Autorstvo – bez prerade
6. Autorstvo – deliti pod istim uslovima

(Molimo da zaokružite samo jednu od šest ponuđenih licenci, kratak opis licenci dat je na poleđini lista).

Potpis doktoranda

U Beogradu, 01.03.2017. godine

