Department of Microbiology, Tumor and Cell Biology
Karolinska Institute, Stockholm, Sweden

HIV EVOLUTION:

THEORETICAL FRAMEWORK AND PRACTICAL APPLICATIONS

Helena Skar

Stockholm 2010

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet.

ISBN 978-91-7457-090-8

Printed by
REPROPRINT AB
Stockholm 2010
www.reproprint.se
Gårdsvägen 4, 169 70 Solna

To my family

ABSTRACT

The human immunodeficiency virus (HIV) is one of the most important and interesting organisms to study today. This pathogen causes life-long infection that presently cannot be cured and the infection leads to development of opportunistic diseases and death if not treated. Finding the answers to the questions still remaining about the evolutionary dynamics of the virus may be crucial in order to develop new therapeutics and functional vaccines, as well as to achieve efficient prevention and surveillance of HIV spread. In terms of evolution, the virus has a remarkable ability to accumulate new mutations over short time. Hence, theoretical models can be applied to HIV data from which parameter estimations can be done directly, and consequently detailed inference of the evolutionary history of HIV can be done. In this thesis the evolution of HIV was studied from several different aspects, and both existing as well as newly developed methods were used.

The spread dynamics of HIV-1 among injecting drug users (IDUs) in Sweden were studied using genetic viral material from newly diagnosed patients and by comparing clinical and demographic data. We found several old lineages of subtype B that had been present at least since the 1990s and that have continued to spread up until late 2007, and we estimated the rate of spread in these lineages to have been generally slow. There have been additional introductions of subtype B into Sweden but these introductions appear to have caused no or limited spread. An introduction of CRF01_AE from Helsinki to Stockholm caused an outbreak in 2006-2007, probably in a standing social network of IDUs. We estimated the incidence rate to increase with a factor of 12 at the outbreak onset, but time from infection to diagnosis during the outbreak was estimated to be short, indicating a rapid discovery of infected individuals. However, both before and after the outbreak, newly HIV-1 infected individuals seem to have remained undiagnosed for longer time periods than during the outbreak.

Within-patient evolutionary rates of HIV were studied in HIV-2 and HIV-1 patients, matched according to viral load, CD4 count, antiretroviral treatment and sampling times. We found that the envelope gene evolved at a faster rate in HIV-2 than in HIV-1 in patients at similar disease stage. The faster rate was more pronounced at synonymous sites, probably a result of factors influencing the replication or mutation rate of the virus.

Finally, we investigated the evolutionary dynamics of HIV-1 in an asymptomatic patient during chronic infection. Through high-frequency sampling it was possible to perform detailed analyses of the processes influencing the short-time evolution of HIV-1 (up to months). We found that several subpopulations were present over time, whose fluctuations over longer time periods (~1.5 years) were consistent with a neutral model of evolution. However, signatures of positive selection were observed on the branches connecting the subpopulations. Thus, non-neutral evolution had likely influenced the formation of these subpopulations and is probably acting over longer time periods in chronic infection of HIV-1.

LIST OF PUBLICATIONS

This thesis is based on the following papers, referred to in the text by their Roman numerials:

I. **Skar H**, Sylvan S, Hansson H-B, Gustavsson O, Boman H, Albert J, Leitner T. 2008. Multiple HIV-1 introductions into the Swedish intravenous drug user population. MEEGID. 8: 545-52

II. **Skar H**, Axelsson M, Berggren I, Thalme A, Gyllensten G, Liitsola K, Brummer-Korvenkontio H, Kivelä P, Spångberg E, Leitner T, Albert J. 2010. The dynamics of two separate but linked CRF01_AE outbreaks among IDUs in Stockholm and Helsinki. J Virol. E-pub ahead of print, 20 October 2010.

III. **Skar H**, Borrego P, Wallstrom T, Mild M, Marcelino JM, Barosso H, Taveira N, Leitner T, Albert J. 2010. HIV-2 genetic evolution in patients with advance disease is faster than that in matched HIV-1 patients. J Virol. 84: 7212-15

IV. **Skar H**, Gutenkunst R, Wilbe-Ramsay K, Alaeus A, Albert J, Leitner T. 2010. Persistence of multiple *env* subpopulations is consistent with neutrality during high-frequency sampling of a chronic HIV-1 patient. Submitted.

LIST OF ABBREVIATIONS

AIC	Akaike information criterion
AIDS	Acquired Immunodeficiency Syndrome
APOBEC	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
cART	combination antiretroviral therapy
CCR	CC-chemokine receptor
CD4	cluster of differentiation 4
CRF	circulating recombinant form
CTL	cytotoxic T-lymphocyte
CXCR	CXC-chemokine receptor
DNA	deoxyribonucleic acid
Env	envelope
Gag	group specific antigen
Gp	glycoprotein
HAART	highly active antiretroviral therapy
HIV	human immunodeficiency virus
HIV-1	HIV type 1
HIV-2	HIV type 2
HLA	human leukocyte antigen
HPD	highest posterior distribution
IDU	injection drug user
IN	integrase
LTR	long terminal repeat
LRT	likelihood ratio test
MHC	major histocompability complex
ML	maximum likelihood
MRCA	most recent common ancestor
mRNA	messenger RNA
MSM	men who have sex with men
Nef	negative factor
NJ	neighbor-joining
PHI	primary HIV infection
PNGS	potential N-linked glycosylation sites
Pol	polymerase
PP	posterior probability
PR	protease
Rev	regulator of virion expression
RNA	ribonucleic acid
RT	reverse transcriptase
SIV	simian immunodeficiency virus
SU	surface unit
Tat	transactivator of transcription
URF	unique recombinant form
V3	third variable region
Vif	virion infectivity factor
Vpr	viral protein R
Vpu	viral protein U

CONTENTS

AIMS

The general aim of this thesis was to investigate the evolution of HIV, both on a population level and within individuals. More specifically, the objectives of this thesis were to:

Paper I: To map the molecular epidemiology of HIV-1 among injecting drug users (IDUs) in Sweden in order to understand the underlying mechanisms of the increase of newly HIV-1 diagnosed IDUs in 2001.

Paper II: To investigate the spread dynamics of the HIV-1 outbreak among IDUs in Stockholm in 2006 by using molecular epidemiology together with observational epidemiological data. In addition, we wanted to investigate if the outbreak in part could be due to the introduction of a new more transmissible HIV-1 variant.

Paper III: To estimate the within-patient evolutionary rate of the HIV-2 envelope gene and to compare it to that of HIV-1.

Paper IV: To investigate the short-term HIV-1 envelope gene evolution in an asymptomatic patient with low viral load and nearly normal CD4+ counts.

1 HIV

1.1 THE HIV PANDEMIC

1.1.1 Origin

The name *human* immunodeficiency virus indicates that there could be viral counterparts in other organisms. Indeed, immunodeficiency viruses in the *Lentivirus* genus can be found in several other species, and interestingly, closely related viruses are found in our closest relatives: the non-human primates [1-6]. These simians are the natural reservoir of many different specific variants of simian immunodeficiency viruses (SIVs). The SIVs are known to naturally infect approximately 40 different species of Old World monkeys and apes, all residing in sub-Saharan Africa [2]. Specific lineages of these SIVs have been introduced to humans through several independent cross-species transmissions as described below, and each successful zoonotic transmission event has resulted in a specific form of HIV (type or group) (Figure 1).

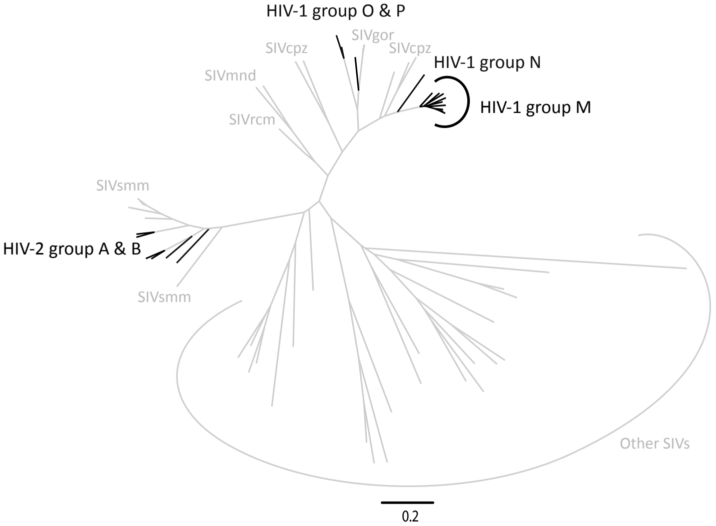

Figure 1. Maximum likelihood tree showing the genetic relationships of a selection of SIV and HIV viruses based on polymerase gene sequences.

West Central African chimpanzees (*Pan troglodytes troglodytes*) are infected with $\text{SIV}_{\text{cpzPtt}}$, which has been introduced to humans, and is now established as HIV type 1 (HIV-1) [7]. More specifically, the closest genetically related $\text{SIV}_{\text{cpzPtt}}$ to HIV-1 group M and N have been found in chimpanzee communities in south-central and southeastern Cameroon [2,8]. Recently, SIV_{gor} was discovered among western lowland gorillas (*Gorilla gorilla gorilla*) and was found to be genetically related to HIV-1 group O and P viruses, however chimpanzees are most likely the original reservoir also of this SIV strain [9,10]. Only one or two cases of group P infections have been discovered so far and therefore group P is not yet formally approved as a fourth group of HIV-1. Among the HIV-1 groups, only group M has showed pandemic spread among humans. HIV type 2 (HIV-2) was introduced to humans from sooty mangabeys (*Cercocebus atys atys*) infected with SIV$_{smm.}$ The introductions are believed to have occurred in West Africa on multiple occasions giving rise to HIV-2 groups A-H. However, only groups A and B have spread effectively in West Africa and beyond, while groups C to H have been identified only in a few individuals [11].

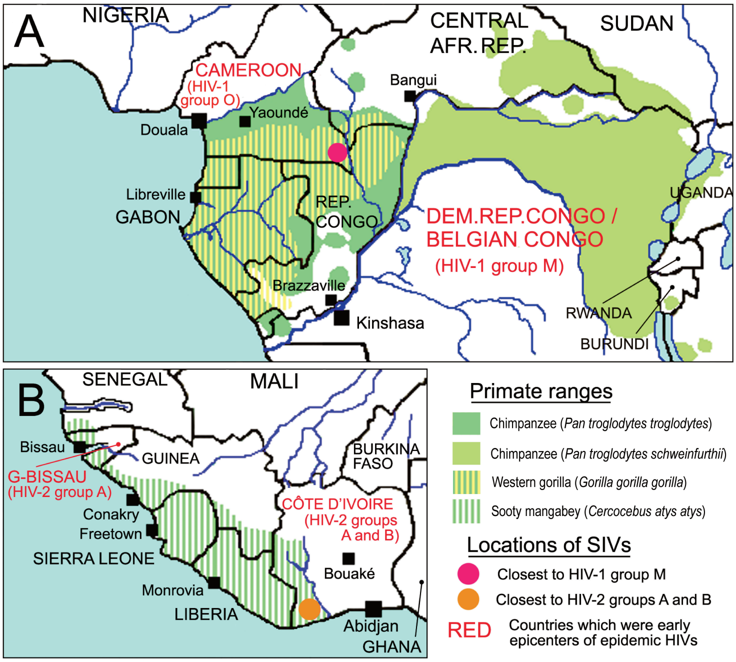

Figure 2. The biogeography of epidemic HIV strains in West Central Africa (A), and West Africa (B). The ranges of the primates that were the source of SIVs that gave rise to HIV strains are indicated. The circles mark the locations where SIVs most closely matching HIV-1 group M and HIV-2 groups A and B were found. Compelling evidence suggests that the countries indicated in red were the most likely epicenters of particular HIV groups. Reprinted from (Sousa JDd, *et al.* (2010) High GUD Incidence in the Early 20th Century Created a Particularly Permissive Time Window for the Origin and Initial Spread of Epidemic HIV Strains. PLoS ONE 5(4): e9936. doi:10.1371/journal.pone.0009936), re-distributed under the terms of the Creative Commons Attribution License: http://creativecommons.org/licenses/by/3.0/.

In the regions in West Central and West Africa where the initial human-to-human spread of HIV-1 and HIV-2, respectively, is believed to have started, genetic diversity of HIV-1 and HIV-2 is the largest in the world (Figure 2) [12]. Using HIV sequences, the first genetic expansions resulting from these zoonotic events have been dated to the late nineteenth through early twentieth century [13-16]. Interestingly, it has recently been shown that SIV has been present in African primates for more than 32 000 years [17]. This indicates that humans may have been exposed to SIV during millennia, which suggests that sporadic cross-species transmissions may have occurred also in the past. The reason why the transmissions that occurred around 100 years ago were maintained in the human population and later showed epidemic spread remains unclear, but social and behavioral changes in form of migration, urbanization [13,16] and the effects of colonization, war and health programs may have been important contributors [18,19].

1.1.2 First discovery

Two years before the identification of HIV in 1983, opportunistic diseases along with immune suppression were observed in young homosexual men in New York City and California [20,21]. The Centers for Disease Control and Prevention (CDC) in Atlanta understood the severity of the situation and published a report in 1981 about the occurrence of pneumocystis carinii pneumonia without identifiable cause [22]. During the following two years, when the cause of the symptoms remained unknown, it became clear that the new disease not only struck the homosexual communities in North America, but also hemophiliacs, injection drug users (IDU) and individuals with Haitian origin [23-26]. Reports from Europe confirmed the presence of the disease among European homosexuals and later individuals of African origin [27,28]. The acronym AIDS (Acquired Immunodeficiency Syndrome) was coined and the first clear evidence that AIDS was caused by an infectious agent came when a child who received blood transfusions died of AIDS-related infections [29]. It was also suggested that the disease could be transmitted heterosexually [30,31]. Finally, in May 1983, a new retrovirus suspected to be the cause of AIDS was isolated at the Pasteur Institute in Paris [32], and a year later the virus and the association with AIDS was confirmed by American colleagues [33]. In 1986, a similar virus (HIV-2) was isolated from AIDS patients in Guinea-Bissau and the Cape Verde Islands [34].

1.1.3 Global spread

Since it was first recognized, HIV/AIDS has become one of the most important infectious diseases with almost 60 million people infected with HIV worldwide and with 25 million deaths due to HIV-related causes. At the end of 2008, an estimated 31.3 million adults and 2.1 million children under 15 years of age were living with HIV. [35]

The HIV prevalence and epidemiological patterns are unevenly distributed across the globe, but also within countries and subregions. For example, two thirds of all people living with HIV can be found in Sub-Saharan Africa and in this region women and girls are affected disproportionately [36]. The epidemic is evolving, and consequently regions are still experiencing epidemiological transitions. In North America and in Western and Central Europe, national epidemics are concentrated around populations at higher risk, such as IDUs, immigrants and men who have sex with men (MSM). Many countries in Western Europe and the North Americas are experiencing a re-emergence of the epidemic among MSM. In Eastern Europe and Central Asia, where the epidemics first involved IDUs, there is now dissemination into the heterosexual population. In Latin America, where MSM has accounted for the largest part of the infections, HIV infections among women and among the indigenous populations are increasing, and heterosexual transmission is becoming a more and more important part of the epidemic. [36] According to UNAIDS, the global spread of HIV appears to have peaked in 1996, however the continuing rise in number of people living with HIV is a reflection of the combined effects of continued high rates of HIV transmission and the beneficial impact of antiretroviral therapy [36].

1.1.4 Genetic variants of HIV

Apart from the heterogeneous epidemiological patterns described above, the regional epidemics around the world are somewhat distinctive in terms of HIV strain composition (Figure 3). This is the result of an uneven spread of HIV strains out of West Central Africa. HIV-2 group A and B are endemic in West Africa and has caused relatively limited spread to other parts of the world, while HIV-1 group M viruses have been more epidemiologically successful and account for almost the entire global epidemic. HIV-1 group M is further divided into nine genetically distinct subtypes A, B, C, D, F, G, H, J and K [37]. In addition, more than 40 circulating recombinant forms (CRFs) have been recognized so far (http://www.hiv.lanl.gov). The CRFs have a mosaic genome composed of regions of different subtypes, are numbered sequentially and named according to their subtype composition, for example subtype A and G: CRF02_AG. When more than three subtypes are present the designation “cpx” (complex) is used, for example, A, G, J and K: CRF06_cpx [37]. Two of the CRFs (CRF01_AE and CRF04_cpx) were initially classified as subtypes (E and I), but complete genome analysis revealed their recombinant nature [38-40]. In addition to the CRFs, there is a multitude of unique recombinant forms (URFs) that have not given rise to a substantial spread. The current HIV-1 epidemic consists of both old and new lineages, where the pure subtypes can be regarded as comparably old lineages with a heritage from the beginning of the group M existence, while the CRFs often are relatively young with contemporary parental sequences [41]. However, the heritable purity of the subtypes has been questioned and they have been proposed to be recombinant viruses themselves [42].

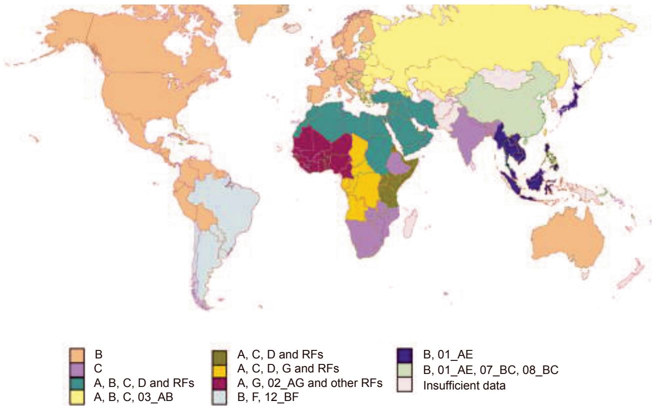

Figure 3. The global distribution of HIV-1 epidemic strains. The distribution shown is a simplification so that 10 major epidemic signatures are identified, but in reality the borders outlines are blurred. In addition HIV-2 is omitted, but would be most significant in the deep purple area, i.e. West Africa. Reprinted with permission: HIV molecular epidemiology: transmission and adaptation to human populations. Woodman, Zenda; Williamson, Carolyn. Current Opinion in HIV & AIDS. 4(4):247-252, July 2009.

The identification of these subtypes and CRFs is important in epidemiological tracking and in the understanding of the ever-changing epidemic. Today, four subtypes and two CRFs dominate the global epidemic: subtypes A-D, CRF01_AE and CRF02_AG [43-47]. Subtype A is concentrated in East Africa and Eastern Europe, while subtype B is widespread globally but dominates the epidemics in the Americas, Western Europe and Australia. Subtype C accounts for approximately 50% of the worldwide infections and strongly dominates the epidemics in Southern and Eastern Africa and India. Subtype D strains are primarily found in East Africa. CRF01_AE and subtype B co-circulate in South-East Asia while CRF02_AG is the most prevalent strain in West Africa. [43,46,47]. As mentioned above, HIV-2 is endemic in certain countries in West Africa, e.g. Guinea-Bissau, Senegal, Guinea and the Gambia. From there HIV-2 has spread to Portugal and regions with past socio-economic ties with Portugal, such as Goa in India, Angola, Mozambique and Brazil [48,49].

Historically many more HIV strains undoubtedly emerged in and from Africa, but chance and possibly lower fitness limited their dispersal. For example, the earliest well documented case of HIV was retrospectively identified in a Norwegian sailor who was infected with a HIV-1 group O virus probably during travels in West Africa during 1961-1962 [50]. However, the Norwegian and most of his family all died in 1976 and caused no or limited spread of the infection. Founder effects, whereby a single chance introduction causes massive spread, can probably account for most of the current geographical distribution of HIV genetic variants, but human genetics and social/behavioral factors are most likely important co-factors to the founding events.

1.2 HIV VIROLOGY

1.2.1 Classification

HIV is a lentivirus, which belongs to the family of *Retroviridae* (retroviruses) and the sub-family of *Orthoretrovirinae*. There are nine species of lentiviruses, where each species infects a certain mammalian (Figure 4).

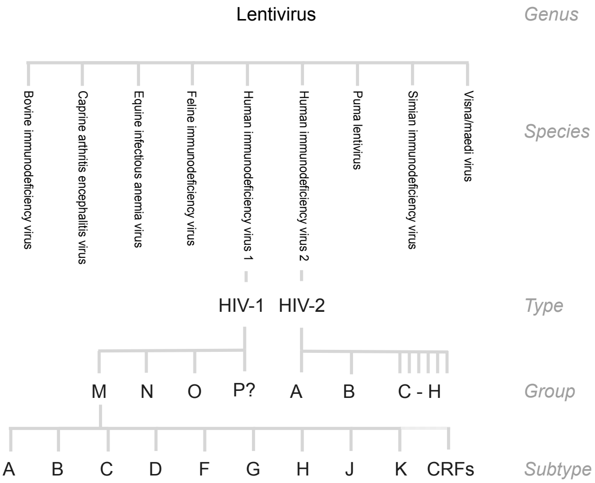
Lentivirus
Genus
Bovine immunodeficiency virus
Caprine arthritis encephalitis virus
Equine infectious anemia virus
Feline immunodeficiency virus
Human immunodeficiency virus 1
Human immunodeficiency virus 2
Puma lentivirus
Simian immunodeficiency virus
Visna/maedi virus
Species
HIV-1
HIV-2
Type
M
N
O
P?
A
B
C - H
Group
A
B
C
D
F
G
H
J
K
CRFs
Subtype

Figure 4. Nomenclature – Lentiviruses; and in more detail the HIV viruses.

The study of retroviruses was initiated in 1908, when cell-free filtrates were able to transmit leukemia among chickens, but the first isolation of a human retrovirus came over 70 years later, in 1979, when the human T-cell lymphotrophic virus type 1 (HTLV-1) was isolated [51]. In common for the retroviruses is the enveloped virion containing the RNA genome, the replication through a DNA intermediate, the integration of the viral genome into its host cell genome and the lifelong infections established. The lentiviruses (lent: slow) are characterized by prolonged sub-clinical infections and are often associated with neurological and immunosuppressive diseases.

1.2.2 Genome and Structure

HIV has a positive sense single stranded genome (+ss) consisting of two RNA copies, each approximately 10,000 nucleotides long. Nine genes are encoded in the compact genome, where all three open reading frames are used (Figure 5). Fourteen proteins are translated, produced through alternative splicing of the mRNAs, bicistronic mRNAs, ribosomal slippage in the translation and cleavage of polyproteins.

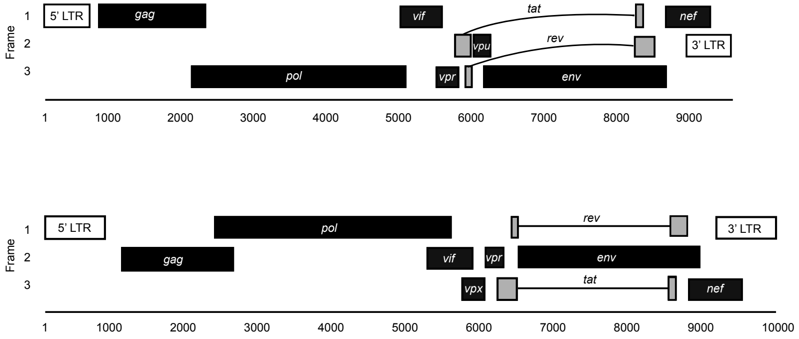

Figure 5. Genomic organization of HIV-1 and HIV-2. The genomes are flanked by the long terminal repeats (LTRs). The scale bar shows approximate nucleotide position.

In common for all retroviruses are the structural or enzymatic proteins, encoded by *gag*, *pol* and *env*. In HIV-1, *gag* encodes the polyprotein precursor p55, which is processed into p24 (capsid), p17 (matrix), p7 (nucleocapsid), and p6 by the viral protease. The Gag proteins are the driving force in HIV virion assembly and release. The viral enzymes protease (PR), reverse transcriptase (RT) and integrase (IN) are encoded by *pol*, but produced as a Gag-Pol precursor polyprotein and processed by the viral PR. These enzymes are essential in the replication of HIV. The envelope gene encodes the viral polyprotein gp160/gp140 that is cleaved to the external glycoprotein (gp) 120 (HIV-1) or gp125 (HIV-2) and the transmembrane gp41 (HIV-1) or gp36 (HIV-2). The envelope glycoproteins are essential for viral attachment to and fusion with its host cell as it contains the binding sites for the CD4 receptor and the chemokine co-receptors for HIV. The surface of gp120/gp125 has five variable loops (V1-V5) and is extensively glycosylated, as described below. The primate lentiviruses encode two regulatory (*tat* and *rev*) and four accessory genes (*nef, vif, vpu, vpr/vpx*). However, the accessory gene *vpu* exists only in HIV-1 and related SIVs, while *vpx* only in HIV-2 and related SIVs. These regulatory and accessory genes are important for the regulation of the viral life cycle (Table 1).

Table 1. Regulatory and accessory gene functions.

Gene	Protein	Function
Regulatory		
tat	Tat	Tat is the transactivator of HIV gene expression. It acts by binding to the TAR RNA element to facilitate initiation and elongation of viral transcription.
rev	Rev	Rev binds to the rev response element (RRE) present on unspliced/partially spliced mRNAs to promote their nuclear export. It has also been proposed that Rev mediate inhibition of integration, preventing superinfection [52,53].
Accessory		
nef	Nef	Nef downregulates the CD4 receptor, which prevents superinfection and facilitates the release of new viral particles. MHC class I (HLA-A/B) is also downregulated by Nef, preventing viral induced apoptosis. In the SIVcpz/HIV-1 lineage Nef seems to have lost its ability to downregulate the CD3 receptor, causing enhanced immune activation [54]. However, the role of Nef in downregulating the T-cell receptor, and thus affecting immune activation, is under debate, reviewed in [55].
vif	Vif	The cytoplastic Vif (virion infectivity factor) protein inhibits the antiviral APOBEC protein and thus G-to-A hypermutations.
vpr	Vpr	Causes G2/M arrest, thus preventing cell division. Vpr is also involved in the nuclear import of the pre integration complex (PIC).
vpu (HIV-1)	Vpu	Vpu promotes the degradation of CD4 in ER and enhances virion release from the plasma membrane.
vpx (HIV-2)	Vpx	Vpx is a paralog of Vpr and their functions are thought to be redundant. In addition, Vpx seems to be important in HIV-2/SIV$_{SM}$ infection of macrophages.

HIV is a spherical enveloped virus, with a diameter of approximately 100 nm (Figure 6). The envelope consists of a lipid bilayer derived from the host cell plasma membrane at budding. Therefore, some host cell proteins may be embedded in the lipid bilayer. Spanning the lipid bilayer is the viral glycoprotein gp41/gp36, which is non-covalently bound to the gp120/gp125 facing the outside. The viral glycoprotein heterodimers interact on the surface and are associated as trimers, on average estimated to be 14 spikes per virion [56]. Lining the inside of the envelope is the matrix, which helps stabilize the spherical structure. The cone shaped capsid resides within the matrix and contains the two +ssRNA copies, which are associated with the nucleocapsid proteins. In addition, there are a number of viral proteins contained within the capsid, such as PR, RT and IN, which are needed either for the maturation of the viral particle or at the early phase of the replication.

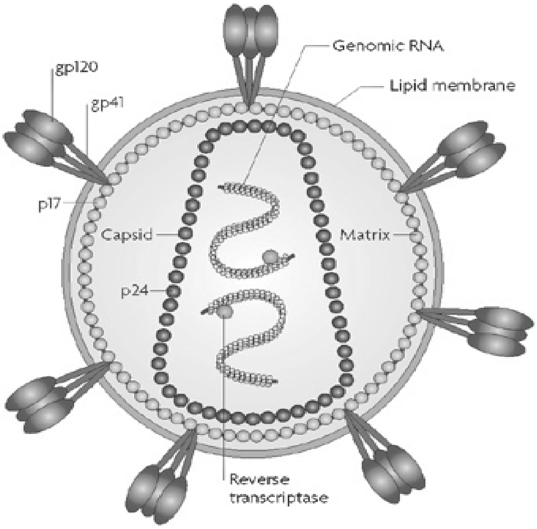

Figure 6. Schematic structure of the HIV particle. Reprinted with permission from Karlsson Hedestam et al., Nature Reviews Microbiology 6 (February 2008).

1.2.3 Viral life cycle

The viral life cycle of HIV starts when the gp120/gp125 trimer binds to the primary receptor CD4, which is present on CD4+ T-lymphocytes, macrophages, monocytes, dentritic cells and brain microglia (Figure 7). The initial interaction occurs mainly through electrostatic interactions, which facilitates CD4 binding. The binding induces an extensive conformational change, whereby the binding site for the co-receptors of HIV is exposed [57]. The chemokine receptors CCR5 and CXCR4 are the most important co-receptors for both HIV-1 and HIV-2. However, both viruses, and especially HIV-2, have been shown to use alternative co-receptors *in vitro* [58,59], but the *in vivo* relevance of these findings are unclear. Co-receptor bindings lead to a second conformational change in gp120/gp125, which triggers insertion of the fusion peptide of gp41/gp36 into the host cell membrane. Consequently, the viral envelope fuses with the host cell and the capsid is released into the cytoplasm. The primary characteristic for retroviruses is the reverse transcription of +ssRNA genome to double stranded (ds) DNA. This is performed by the viral RT enzyme inside the partially opened capsid. Reverse transcription is primed by a human transfer RNA that is bound to the RNA genome inside the virion. Each virion contains two +ssRNA copies of the HIV genome and the RT enzyme jumps between these two templates during the reverse transcription. If the two RNA templates are genetically distinct this will give rise to a new recombinant virus. The RT enzyme contains two domains, a DNA polymerase and a ribonuclease H (RNAse H) domain. Hence, the RNA template is degraded immediately after its transcription to first complementary DNA strand, which then serves as template for the synthesis of the a second DNA strand. The RT enzyme is error-prone and lacks proof-reading ability. This leads to a high mutation rate during reverse transcription. The high mutation rates and the template switching are the foundations of the high diversity seen in HIV, both within an individual and on the global level. During the transcription, long terminal repeats (LTRs) are generated in the 5’ and 3’ ends of the genome. The LTRs are important for the integration of the viral DNA into the host cell genome as well as for its transcription. After reverse transcription the viral DNA genome is contained within the pre-integration complex (PIC), which is translocation into the nucleus. The integration is mediated by the viral protein integrase (IN) preferably into active and thus open regions of the human genome [60], but integration also takes place in resting cells. Once the viral DNA has been integrated into the host cell genome it is referred to as a provirus and can remain transcriptionally silent (i.e. latent) or be immediately transcribed by the cellular machinery.

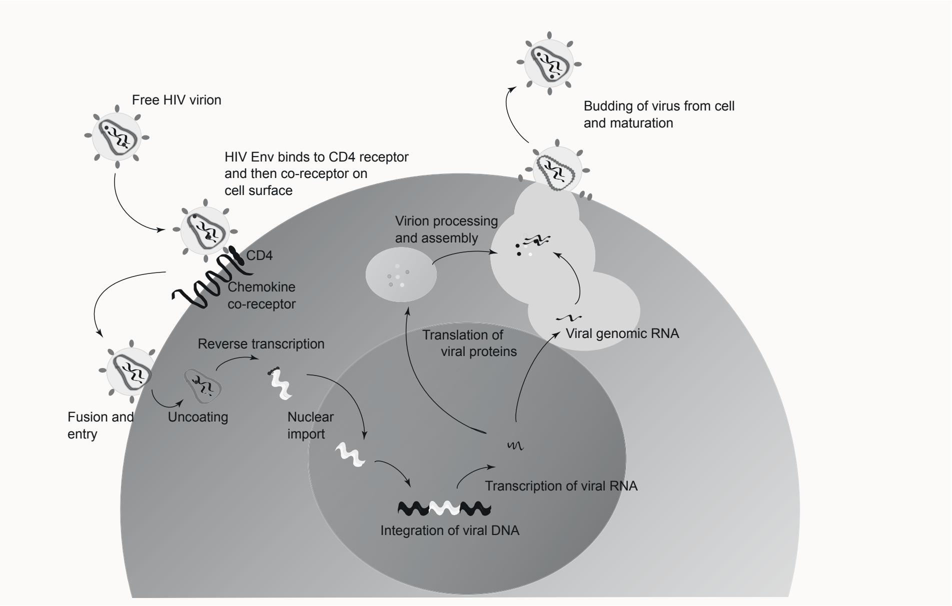
Free HIV virion
Budding of virus from cell and maturation
HIV Env binds to CD4 receptor and then co-receptor on cell surface
Virion processing and assembly
CD4
Chemokine co-receptor
Reverse transcription
Translation of viral proteins
Viral genomic RNA
Fusion and entry
Uncoating
Nuclear import
Transcription of viral RNA
Integration of viral DNA

Figure 7. The life cycle of HIV.

The HIV LTRs act as transcriptional promoter regions and direct the cellular RNA polymerase II to the DNA template, and the first event of the transcription is the synthesis of a full-length RNA copy. The early transcripts are completely spliced into short mRNAs that encode the Nef, Tat, and Rev proteins. Tat binds to the transactivation response region (TAR) downstream of the LTR enhancer regions, and promotes efficient HIV mRNA elongation. The Rev protein binds to the rev responsive element (RRE) in the *env* region of the HIV mRNA and functions as a carrier of the unspliced or partially spliced RNAs between the nucleus and the cytoplasm. Therefore, the accumulation of Rev in the nucleus signals the change from early to late transcription. The late transcription involves the expression of the longer gag, gag-pol, env and the vif, vpr/vpx and vpu mRNAs, which are unspliced or incompletely spliced. All mRNAs are translated in the cytoplasm or near the endoplasmatic reticulum (ER). The Env proteins are heavily glycosylated in the ER and the Golgi apparatus, and trimers are formed. The viral particles are assembled at the plasma membrane where they bud from the cell, consequently acquiring the lipid envelope already containing the gp120/gp125 and gp41/gp36 trimers as well as certain cellular membrane proteins. The maturation is the final step of the viral life cycle and takes place after budding when the PR enzyme cleaves the Gag-Pol polyprotein into it functional proteins.

1.3 HIV INFECTION

1.3.1 Transmission

HIV spreads mainly through sexual contacts, contaminated injection equipment, blood transfusions and mother-to-child transmission during pregnancy, delivery or breastfeeding. Globally, sexual transmission accounts for approximately 80% of the infections, where heterosexual transmission constitutes the major part. HIV infection as a consequence of injecting drug use represents about 10% of the global infections, but this proportion is higher in areas where the IDU population is large as in Eastern Europe and Central Asia. The risk of infection is dependent upon a number of factors, such as the amount of virus in the infecting body fluid, co-infections such as other STDs, behavioral factors, as well as route of infection [61,62]. Commonly, the risk of sexually acquired HIV-1 infection has been estimated to be around 1 in 1000 coital acts, but this represents a lower bound as these numbers often are derived from studies of stable heterosexual couples with low prevalence of high-risk factors [63]. Accordingly, the risk of heterosexual transmission was significantly associated with viral load and stage of infection of infecting partner, and the presence of genital ulcers in studies of sero-discordant couples in Rakai, Uganda [64,65]. Furthermore, contagiousness is higher in the beginning and the end of the infection when virus levels are high. Successful antiretroviral treatment leads to a substantially lowered transmission risk. Male circumcision has been associated with decreased risk of becoming infected, but not in lower transmission risk [66-68]. It is important to stress that behavioral factors are very important determinants for the actual transmission risk. For instance, the transmission risk is close to zero if condoms are correctly used in sexual encounters or sterile injection equipment is used by IDUs.

1.3.2 Pathogenesis

The course of HIV infection can be divided into three stages; the acute phase, the chronic stage and AIDS (Figure 8). The acute phase, i.e. primary HIV infection (PHI), lasts for about four to eight weeks and is characterized by massive HIV replication resulting in high viral levels with 10^7 - 10^8 million RNA copies/ml, and loss of CD4+ T-cells, especially in gut-associated lymphoid tissue [69-72]. Flu-like symptoms, which include fever, body ache and swollen lymph nodes, may appear during this initial phase as a result of immune responses directed against the infection. Eventually the viral load is suppressed to a semi-steady state level (viral setpoint) and the CD4+ T-cell levels are partially restored. The setpoint has been shown to be predictive of disease progression in HIV-1 infection [61], and on average the onset of AIDS takes about 8 years (9-11 years survival infection-death, UNAIDS 2007) without effective antiretroviral treatment. The plasma viral setpoint is lower for HIV-2, often undetectable, and the average rate of disease progression is much lower [73,74]. During the chronic phase the immune system slowly gets exhausted by the constant battle with the infection, i.e., chronic immune activation and depletion of CD4+ T-cells, and finally collapses. Chronic immune activation is believed to be an important factor in pathogenesis, which possibly can explain the differences seen in the rate of disease progression between HIV-1 and HIV-2 since a lower level of immune activation has been observed in HIV-2 as compared to HIV-1 infected patients [54,75-77]. The immunological definition of severe disease progression is a drop of CD4+ T-cell levels below 200 cells/μl plasma [78]. At this stage there is a high risk for development of a number of opportunistic diseases and virus-induced tumors. By definition these diseases mark the onset of AIDS.

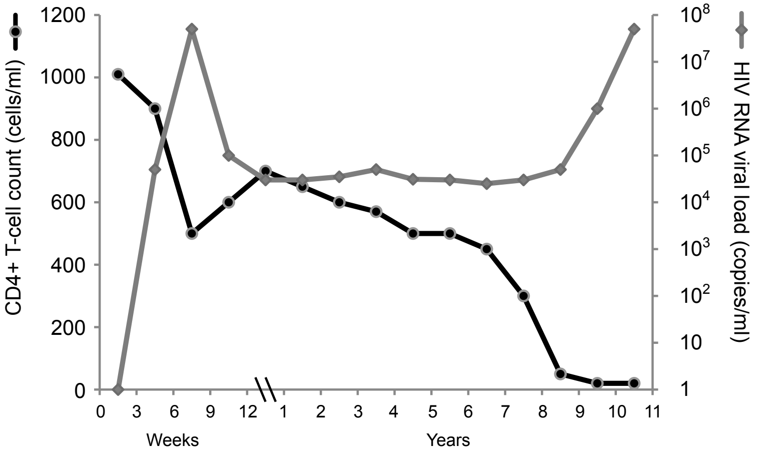

Figure 8. The clinical course of HIV-1 infection.

1.3.3 Treatment

Without the use of effective antiviral treatment, almost all HIV-1-infected patients develop AIDS, which eventually leads to death. The antiviral drugs currently used are aimed to interfere with specific parts of the HIV viral replication cycle. They restrict the production of new viral particles, but will not eliminate infected cells and therefore cannot cure the patient from HIV infection. There are three classes of drugs used today in the first-line treatment; nucleoside analogue reverse transcriptase inhibitors (NRTIs), non-nucleoside analogue reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs). There are two additional classes of licensed antiretroviral drugs, namely entry inhibitors and integrase inhibitors. The entry inhibitors include a fusion inhibitor (enfuvirtide, T20) and a co-receptor antagonist (maraviroc). Only one integrase inhibitor (raltegravir) is currently licensed. Effective HIV therapy requires a combination of at least three drugs from at least two different drug classes (referred to as combination antiretroviral therapy [cART] or highly active antiretroviral therapy [HAART]). HAART was introduced in 1996 and has had an enormous impact on the survival of HIV patients [79]. The treatment is life-long and adherence is very important. If the virus gets a chance to replicate during drug-selective pressure there is a significant risk for selection and *de novo* evolution of resistance mutations. This may lead to treatment failure, which will complicate the HIV medication scheme. The antiretroviral drugs are optimized for HIV-1, and therefore, optimal treatment options for HIV-2 infection are relatively limited.

1.4 THE EVOLUTION OF HIV

Common characteristics of the RNA viruses are the exceptionally high mutation rates, the small genomes and the high number of offspring [80]. This explains why HIV is one of the fastest evolving organisms with an average mutation rate of approximately 0.1-0.3 mutations per genome and replication cycle [81-83]. Less is known about the mutational process of HIV-2. Point mutations are defined as substitutions, insertions or deletions and can be generated during several steps in the replication cycle. The error-prone RT enzyme is considered to account for most of the point mutations. However, the cellular RNA polymerase II, active during the RNA transcription from the proviral DNA, also lacks proof reading capacity, and most likely contribute to viral mutagenesis. In addition, cellular enzymes may induce nucleotide modifications, in particular the ABOBEC3G/F enzymes that specifically introduce G-to-A hypermutations in retroviral RNA [84]. The second major contribution to HIV genetic diversity is the template switching of the RT enzyme between the two RNA genomes during reverse transcription [85]. This will generate a recombinant proviral DNA that has a mosaic genome derived from the two parental genomes. Template switching has been estimated to occur on average two to three times per replication cycle, but both higher and lower rates have been observed [86-88]. However, recombination only contributes to evolution if the two RNA molecules in the virus particle are genetically distinct. As a consequence, when the two parental genomes are genetically distinct, large evolutionary and antigenic leaps may occur in just one round of replication. The effective recombination rate is the product of superinfection of the same cell and number of template switching events, and has been estimated to be similar to the nucleotide substitution rate, ~0.14 recombinations per genome and replication cycle [89]. Combined with the high replication and production rates of HIV, the ability to mutate and recombine is the foundation for the high HIV diversity seen both within an individual and on a global level. However, the reason why certain mutations survive in subsequent generations is a complex process determined by selective forces and chance events that act on HIV to shape its evolution.

1.4.1 Within-patient HIV evolution

The stages of infection are associated with distinct patterns of within-patient HIV evolution (see below). However, evolution on an individual basis will always be distinctive to some extent depending on the host immune system, stochastic events and the infecting HIV type. In a seminal paper by Shankarappa *et al.* (1999) the within-patient diversity (genetic variability at one time point) and divergence (genetic change from a point of reference) of HIV-1 were measured [90]. From nine patients sequences covering the V3 region (C2-C5) of the envelope gene were included and it was shown that the diversity could be as large as 10-15% at one time point with a divergence rate of approximately 1% per year. The diversity and divergence increased linearly during the early phase of chronic infection, followed by stabilization in diversity in parallel with continuing divergence. At later stages of the disease the divergence rate was shown to also level off, which probably is a reflection of the collapsed immune system at this point. As already mentioned, disease progression is slower in HIV-2 infection than in HIV-1 infection, and therefore the viral evolutionary dynamics are likely to be somewhat different. Moreover, the rate of disease progression also varies considerably among individuals infected with HIV-1 and therefore broad generalizations about HIV-1 within-patient evolution will not be valid. Accordingly, the rate of within-patient HIV-1 evolution has been associated with disease progression, both at an absolute rate [91], and when differentiating between synonymous and non-synonymous substitutions [92].

1.4.1.1 Primary infection

The HIV diversity is low during the initial phase of acute infection, i.e., until the setpoint in viral load is reached [93-97]. It is uncertain if this observation is due to transmission of only a single virus particle or outgrowth of a single variant. However, recent studies have employed limiting dilution to accurately sequence many individual viral molecules from acutely infected patients and have shown that many primary infections are likely established by a single virus particle [96,98-101]. Interestingly, these studies suggest that the number of infecting variants is correlated to transmission route, so that higher risk of infection correlates with higher number of infecting particles [102]. Nevertheless, transmissions are associated with severe population bottlenecks. It has been suggested that the transmission bottlenecks is the result of selection acting on the envelope gene, so that only CCR5-using viruses can be transmitted or establish a successful infection [93,94]. This theory is strengthened by the fact that individuals homozygote for a 32-bp deletion in the CCR5-gene, which results in a non-functional protein, seem to be protected against HIV-1 infection [103,104]. Transmission bottlenecks have been seen not only in mucosal transmission, but also in infections through intravenous drug use [105,106].

Already during primary infection, the immune system gets activated, leaving its signature on the evolution of HIV [107,108]. The first line of defense is the innate immune system, followed by the development of the adaptive immune responses. CD8+ T-cell mediated killing of productively infected cells is believed to contribute substantially to the initial decline in plasma viral load [109-111]. Thus, HIV-1 specific T-cell responses develop before seroconversion and just before the peak of viremia is reached. However, usually HIV rapidly escapes these first T-cell responses, which indicates that the targeted epitopes are readily changeable. The T-cell responses change in return, targeting more slowly evolving or invariant epitopes. Thus, if present, these latter T-cell responses may be important in maintaining the already established setpoint [109]. Ongoing viral replication continues from peak of viremia until setpoint, implying a significant role of cellular immunity in control of the virus [95]. Antibodies directed against HIV-1 mediated by the humoral immune response have been seen to arise within 8 days of infection. These first antibodies mostly contribute to the formation of immune complexes and are not likely to have impact on the control of acute phase viremia [112], but early escape from neutralizing antibodies has been reported [113,114].

1.4.1.2 Chronic infection

As the infection progresses, the breadth of T-cell responses generally increases but has been both positively and negatively correlated with viral load [115,116]. Certain genetic traits of the host, especially some HLA types, have been associated with reduced *in vitro* replication capacity of HIV-1 and the rate of disease progression, highlighting the importance of the immune system on viral control [117-120]. During chronic infection the humoral immune system constantly retargets new HIV-1 variants. However, even though they may be neutralizing they lag behind, rarely targeting contemporary viruses [114,121-124]. No association has been seen between natural control of HIV-1 viremia during chronic infection and specific antibody responses. In contrast, autologous neutralization escape has been shown to be rare in HIV-2 infection, but further studies are needed to establish its' importance for the observed low plasma virus levels in HIV-2 infection [125]. Glycosylations of the envelope spikes have been shown to be important for the folding of gp120 upon CD4 binding as well as determinants of the co-receptor usage of HIV-1 [126,127]. Moreover, the host-derived glycans hinder efficient antibody binding. Accordingly, the glycan-shield has been shown to evolve, where glycosylation sites in the HIV-1 envelope gene emerge and disappear during the course of infection [122], but not during HIV-2 infection [125,128]. In HIV-1, the emergence of CXCR4-using viruses is correlated with loss of CD4+ T-cells and faster progression to AIDS [129].

1.4.1.3 Selective forces

The continuous and changing pressure from the immune response usually leads to selection of new variants throughout the course of infection [130]. Such a selection process is referred to as positive selection. However, in protein coding sequences conservative forces generally dominate, since the functions of the proteins must be preserved in order for the organism to survive. This is termed purifying or negative selection and even in *env*, the most variable gene of HIV-1 and HIV-2, negative selection has been shown to dominate over positive selection [131,132]. However, individual sites may still be under strong positive selection and the variability of HIV further suggests that many mutations are tolerated and thus are not part of natural selection, i.e. they are under neutral evolution. Nucleotide substitutions can either be non-synonymous, which will lead to a change in amino acid, or synonymous, which will not lead to a change in amino acid. Hence, selective forces can be measured on a nucleotide level by comparing homologous sequences, which will be discussed in more detail in section 2.4.1.

It has been suggested that within-patient HIV-1 evolution is dominated by genetic drift [133], which has been supported by *in vitro* culture studies [134]. Genetic drift occurs when the mutational process is stochastic (neutral), thus mutations get fixated in the viral population due to chance. The population size of HIV-1 within a patient is large, with more than 10^8 productively infected cells and more than 10^{10} virions produced daily in untreated patients [135,136]. The principles of population genetics argue that under these circumstances natural (non-neutral) selection will dominate. Hence, the presence of genetic drift seems counter-intuitive. However, it has been proposed that the effective population size of HIV-1 during chronic infection is much lower than the total amount of HIV-1 particles present [133,137-139]. The small effective population size can be viewed as less genetic diversity than expected relative to the total population size, (for a definition of effective population size, see section 2.4.3). However, the methods used to calculate the effective population size often assume neutral evolution and well-mixed populations. In order to address the presence of genetic drift, more appropriate models of population structure and selective forces need to be developed. To date, a few models have tried to unify the estimated small effective population sizes and the strong positive selection believed to act on HIV during chronic infection [133,139]. A model where several different effects are taken into account was proposed by Achaz *et al.* 2004, hence a meta-population model and selective sweeps are proposed to both be factors that act together to reduce the intra-host effective population size of HIV-1 [140].

1.4.2 Evolution of HIV on a population level

Selective forces acting on HIV within the host can also be seen on a population level. One example is the transmission of drug resistant variants of HIV-1, which has been estimated to account for approximately 8% of new infections in Europe [141]. Another example is the suggestion that HIV immune evasion from CD8+ T-cells leaves imprints on the HIV proteome, *i.e.,* that the HIV CTL epitope distribution on a population level is adapting to the human host [142,143]. However, when proteins under neutral selection on a population level are compared to the HIV CTL epitope distribution, these epitope clusterings have been shown not to be significantly different from a random distribution [144]. The question whether HIV is adapting to the human host is still under debate, and it has been proposed that the specific epitope clusterings seen locally are due to shared common ancestry (viral founder effects) instead of HIV adaptation [145]. An important note in this discussion is that many transmissions occur before the virus has experienced pressure of the host immune system, i.e. during early infection when viral levels are high [64]. Another interesting concept is the optimization if viral transmissibility. Transmissibility will depend upon infectiousness and the duration of infection, which are closely linked to setpoint viral load in HIV infection. Thus, mathematical models suggest that HIV-1 maximizes the transmission potential when virus levels in patients are around 30,000 HIV-1 RNA copies per ml plasma, which is the observed mean setpoint viral loads in heterosexual cohorts [146]. This implies that the setpoint viral load may be heritable [147]. However, these findings need to be confirmed in larger studies where several modes of transmission are included.

1.5 DYNAMICS OF HIV SPREAD

Even though selective forces resulting from within-patient evolution might contribute to HIV evolution on a population level, the effect from demographic and spatial history as well as patterns of host behavior are more important in shaping the global and regional HIV epidemics. As already mentioned, the stage of infection is important in determining risk of transmission, with highest risk during early infection. Therefore, a substantial part of infections occur from individuals who themselves are newly infected, and has been estimated to account for 5 - 50% of the transmissions, where the variable proportion is a reflection of the difficulty conducting empirical studies aiming to answer this question as the transmitting individual usually is unaware of the infection [148]. Already in the 1980s simple mathematical modeling was used to understand the epidemic dynamics of HIV-1 [149], and in later years the incorporation of structured social networks is becoming increasingly important in the understanding of the spread of HIV [150]. Thus, the structure of contact networks are essential in the understanding of sexual transmission as well as transmission between IDUs when sharing injection equipment.

1.5.1 Impact of transmission routes

Sexual transmission networks are characterized by heterogeneous partner exchange, where “super-spreaders” who have very high numbers of partners transmit their virus disproportionally in the epidemic [151,152]. Few empirical studies have been made mapping sexual contact networks, but an example is shown in Figure 9 [153,154]. HIV transmission networks can be mapped to some extent through interviews and partner notification so that transmitters of the disease and people who may have been unknowingly infected can be identified, treated, and advised about disease prevention. However, larger transmission networks are harder to study, but one example is the “Swedish transmission chain” [155]. Interestingly, the authors found that HIV sequence data accurately reconstructed the known transmission chain. Thus, the use of sequence data in estimating unknown transmission dynamics may be useful (but see discussion on page 43). For example, sequence data from a large cohort, > 2000, of MSM living in London were compared using molecular phylodynamics and it was found that the transmission dynamics were characterized by episodic sexual transmission in a large number of distinct networks [156].

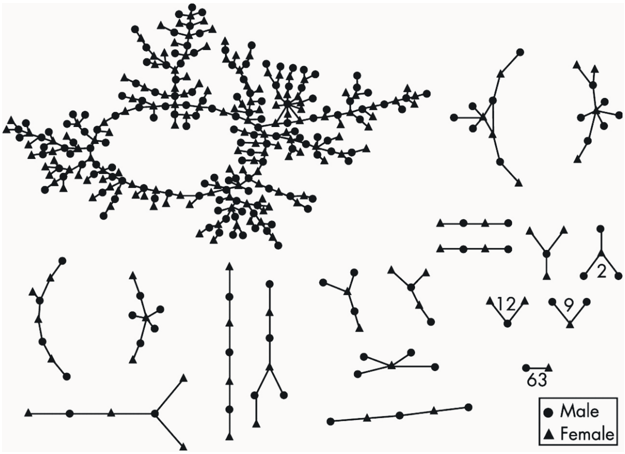

Figure 9. Sexual network of adolescents in the “Jefferson High” study. Reproduced from (Prevention strategies for sexually transmitted infections: importance of sexual network structure and epidemic phase, Ward H, 83, i43, 2007) with permission from BMJ Publishing Group Ltd.

Sero-sorting is a term that is used when the behavior of an individual is influenced by the fictive or true knowledge of the HIV status of a person in his or hers contact network. Sero-sorting is most common among IDUs, but may also be important when choosing sexual partner or whether to use a condom. Thus, a burst of new infections during a short period of time may be due to an introduction of HIV into a standing network with established risk behavior, where the anticipated HIV status of one individual is not longer true. The sharing of injection equipment may be common in many IDU communities, especially where no successful prevention strategies are implemented, and these communities are especially vulnerable for explosive outbreaks of HIV. This has been seen on a number of occasions for example in countries of the Former Soviet Union (FSU), Finland, Thailand and Sweden [157-163] and [II].

1.5.2 Fast and slow spread of HIV

The rate of epidemic spread is likely to have an impact on the rate of evolution of HIV on a population level [164]. An explosive outbreak, where transmission occurs shortly after infection, will lead to transmission of viruses with little experience of the host immune pressure. Consequently, on a population level the viruses will be similar and the evolutionary rate will be slow. In contrast, in a slow spreading epidemic, transmissions will occur more often during chronic infection and with viruses that has adapted to the host immune pressure. Thus, transmitted virus variants will be heterogeneous over time and the estimated evolutionary rate will be faster. [164] This means that sequence data intelligently sampled over time can be used to estimate the rate of spread within a confined epidemic.

1.6 HIV IN SWEDEN

HIV was introduced into Sweden in the late 1970s, and the first AIDS case was diagnosed in 1982. The number of diagnoses peaked in the mid 1980s when HIV testing became available, followed by a rather low incidence rate, which has slowly increased during the last decade. Initially, as in many other western European countries, the epidemic mostly struck the MSM and IDU populations. Still today, most of the domestic transmissions occur within the MSM population, which is the only transmission group that has shown a steady increasing infection trend since 2002. The IDU population experienced an increase of domestic infections in 2006, and rapid response with extensive testing resulted in even more people being diagnosed in 2007. However, in 2008, the epidemic was reversed (Figure 10). Since 1990 the largest proportion of newly infected cases originates from people infected abroad through heterosexual transmissions, mainly immigrants from high-endemic areas of the world. Up until 2002 around 300 new cases were reported annually, while around 400-500 cases have been reported each year since then. Since the beginning of the epidemic until the end of 2009, 8935 HIV infections had been reported and today approximately 5240 persons are living with HIV in Sweden. [165]

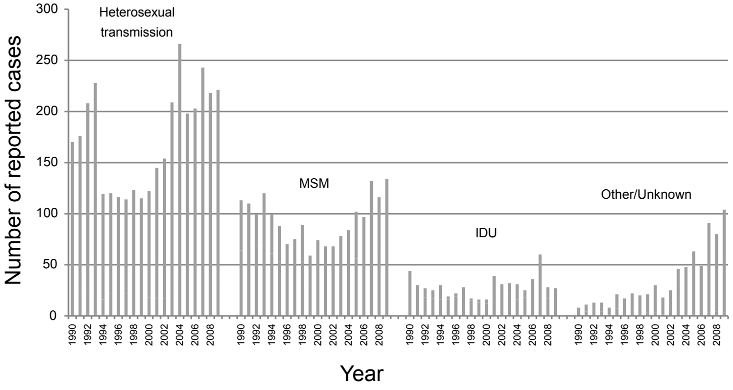

Figure 10. Number of yearly reported cases shown as histograms per transmission route, 1989-2009, compiled by the Swedish Institute for Infectious Disease Control.

2 RECONSTRUCTING HIV EVOLUTION

In this chapter I will give a brief introduction to the field of evolutionary biology, the methods used when studying evolution and the various applications on HIV. The characteristics of HIV replication and spread within and between hosts, as described in Chapter 1, will give rise to both problems and possibilities when inferring a correct evolutionary history of HIV from sequence data. I will mostly focus on the theory behind and the methods used in Papers I-IV.

2.1 THE STUDY OF EVOLUTION

The modern field of evolutionary biology was established in the first part of the 20th century when the theories of Gregor Johann Mendel (1822-1884) and Charles Darwin (1809-1882) were reconciled. The full history is out of the scope of this thesis, but the end result was an evolutionary theory based on a combination of Mendelian inheritance and evolution by natural selection, which is referred to as "the modern evolutionary synthesis" and neo-Darwinism. With the identification of DNA as the carrier of the genetic information and the subsequent publication of its structure in 1953 [166], genetics and molecular biology have gotten a pivotal role in the field of evolutionary biology. Genetic material can be used to infer the evolutionary history of organisms through phylogenetics, where the evolutionary process is regarded as a branching process. Phylogenetic approaches are often preferred to estimate and represent biodiversity and are also essential in descriptive molecular epidemiology, where they can be used to study the evolutionary relatedness of different strains of an organism. Phylogenetics can also be used to answer more detailed questions about the evolutionary process itself. For example, in combination with population genetics (which is the study of the allele frequency distribution in a population), powerful tools arise to qualitatively and quantitatively estimate the history of the relative genetic diversity over time for a whole population from only a small subsample.

In the case of HIV, which is characterized by a high mutation rate, it is possible to follow the evolution in "real-time" and estimate evolutionary parameters directly from sequence data. However, mathematical models are needed to describe the process of nucleotide or amino acid change and they will be presented in more detail in section 2.3.1.

2.1.1 Sequence generation

The genetic material of HIV is retrieved through extraction of viral RNA from virions or proviral DNA from infected cells, often from plasma or cells from patient blood samples. Direct population sequencing, where all genetic variants present in the sample are sequenced simultaneously, will result in a consensus sequence containing polymorphic sites, whereas clonal sequencing can pin-point individual RNA or DNA molecules. Clonal sequencing can either be performed through molecular cloning or through limiting dilution; aka. single genome sequencing (SGS) [167,168]. Chain-termination sequencing is by far the most common sequencing method and is used for both population and clonal sequencing. However, in the last few years, high-throughput sequencing methods have become available where thousands or millions clones can be sequenced at once. These next-generation sequencing methods will likely have enormous impact on the evolutionary biology field. In the case of HIV, the deep sequencing that the new sequencing methods make possible, will for example help understand the intrinsic evolutionary patterns of resistance mutation development within patients [169].

2.2 LET THERE BE HOMOLOGY

The core concept of phylogenetics is the comparison of homologous gene sequences or genetic regions, which share a common ancestor. Furthermore, homologous sequences are orthologous if they were separated by a speciation event, while paralogous sequences were separated by a duplication event. Care must be taken to identify the sites suitable for phylogenetic reconstruction and to put them in their right setting. Genetic shifts, reassortment and recombination will obscure the phylogenetic signal and will lead to false and unpredictable results. Hence, in the case of HIV, identification and removal of putative recombinant sequences before standard phylogenetic tree building, or the use of phylogenetic network models are two ways to account for recombination, as described below.

2.2.1 Alignment

The mutational process of HIV, and other organisms, includes substitutions, insertions, deletions and recombination events. In order to infer these processes, the sequences under study are site-wise compared. This is done by constructing an alignment, where the location of each nucleotide (or amino acid) is positioned correctly in relation to its homologous sites in the other sequences. Gaps are inserted when needed so that homologous sites are placed in successive columns. Figure 11 shows the nucleotide alignment used to infer the phylogenetic tree in Figure 1.

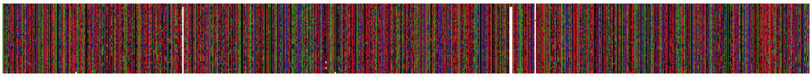

Figure 11. Nucleotide alignment of the pol gene of SIV and HIV viruses. Each colour represents one of the four nucleotides (A, C, G, T). The figure was done using the program Pixel, available at: http://www.hiv.lanl.gov/content/sequence/pixel/pixel.html.

Several different methods are available to automate and optimize the alignment process, and multiple sequence alignment is generally performed using computational algorithms with heuristic optimization, for example HMMER, MUSCLE and MAFFT [170-172]. My experience is that for the highly variable regions of HIV, where deletions and insertions are common especially in the variable loops of the envelope gene, a visual last check is recommended or even mandatory. Furthermore, regions where homology is uncertain or that are difficult to align should usually be removed before continuing with the sequence analyses. In addition, if codon or amino acid models are to be used further on in the analyses it is necessary to make sure that the nucleotides are aligned in frame.

2.2.2 Accounting for recombination

Putative recombinants can be detected through genetic comparison of an *a priori* assigned query sequence with a reference genetic background. The recombinant identification program (RIP) and the bootscan analysis, where the query sequence is compared to an aligned reference dataset consisting of putative parental sequences by a sliding window approach were the firsts of their kind [173,174]. These methods perform perhaps at their best when the genetic distance between the query sequences is substantial, for example when detecting inter-subtype recombinants. Today there are more complex methods to detect between subtype recombination, for example the jpHMM program that is used extensively at the Los Alamos HIV sequence database to correctly classify the recombinants stored [175,176]. Today there are several methods where the *a priori* assumption is not needed and where more closely related sequences may be compared. An example is the program Recco, which given a multiple sequence alignment scores the cost of obtaining one of the sequences from the others by mutation and recombination. The optimal path, defined as the smallest number of evolutionary events, will explain the data best, and putative recombinants along with specific breakpoints will be identified [177]. A widely used program is GARD, which searches for evidence of segment-specific phylogenies by inferring phylogenies for each putative non-recombinant fragment through an iterative approach [178]. Given a maximum number of breakpoints, *B*, the method will search the space for all possible locations for *B* or fewer breakpoints in the alignment. The goodness of fit is done by an information-based criterion, such as the Akaike information criterion (AIC) derived from a maximum likelihood model fit for each segment. In the next round *B*+1 number of breakpoints will be tested. The end result will tell you how many putative breakpoints you have in your alignment and where they are. However, the putative recombinant sequences will not be identified. Instead, you will be forced to infer separate phylogenies for the different non-recombinant fragments. Another approach is to infer phylogenetic networks, such as split and reticulate networks, instead of bifurcating phylogenic trees. In a split network, every edge is associated with a split of the taxa, but there may be a number of parallel edges associated with each split. Thus, a branch in a bifurcating phylogenetic tree is analogous to an edge in a split network. The difference between a split network and a phylogenetic tree becomes apparent when there is conflicting phylogenetic signal in the data, which may arise with recombination events in the evolutionary history of the taxa. An example is given in Figure 12.

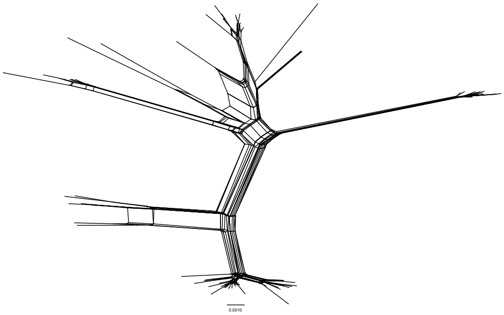

Figure 12. An example of a Neighbor-Net split network. Inferred using Neighbor-Net [179].

In the program Splitstree4 it is possible to infer phylogenetic networks to visualize the evolutionary history of query sequences [180], although the visualization process is difficult and therefore the networks may not be correctly shown in all aspects. Nevertheless, taxa that have been involved in putative recombination events may be identified, as they often are associated with several edges. The presence of recombination signal can then specifically be tested with the pairwise homoplasy index statistic (PHI-stat) [181]. Homoplasy is the result of parallel or convergent evolution but not the result of recombination, however the signals are confounded and are therefore difficult to separate. The PHI-stat is based upon the principle of compatibility and incompatibility. When a pair of sites is incompatible either a recombination event must have taken place or a homoplasy must have occurred in the history of one of the sites. Hudson and Kaplan (1985) came to the conclusion that the degree of genealogical correlation between neighboring sites is negatively correlated with the rate of recombination [182]. The PHI-stat measures the similarity between closely linked sites and the significance of the observed test statistic is obtained by using a permutation test. If there is no recombination in the data the genealogical correlation of adjacent sites is invariant to permutation. But in the presence of finite recombination, the order of the sites is important, and distant sites will tend to have less genealogical correlation than adjacent sites. Thus, putative recombinants may be identified, however the breakpoint sites will remain unknown. A similar approach where recombination only is distance dependent has recently been developed and applied to HIV-1 within patient sequence data [89]. If possible, a combination of the above methods should be used to assure correct handling of recombinant signal in a dataset.

2.3 PHYLOGENETIC INFERENCE

Phylogenetic trees are inferred through several steps. Given the alignment, a model of sequence evolution needs to be chosen, followed by parameter estimations. Parameter estimations may be done in conjunction with tree building, which is the last step. However, work has been done where alignment and the phylogenetic tree are inferred in conjunction, for example [183,184], which probably is the best way to estimate the uncertainty associated with each step.

2.3.1 Models of sequence evolution

One of the simplest models of sequence evolution assumes that the evolutionary process is the same across different regions of the sequences and through different stages of evolution. However, this is very seldom true in nature. An example is the envelope gene, which contains both regions of extremely high variability, but also more conserved regions, for example the CD4 binding site. If this simple model of sequence evolution would be applied to this region the genetic changes would be gravely underestimated. Figure 13, illustrates the discrepancy between the simple uncorrected pairwise p-distance model (Observed distance), which only measure pairwise nucleotide differences, with a more complex (and realistic) model, which puts different weights on different mutations and accounts for multiple substitution at the same site (Correction).

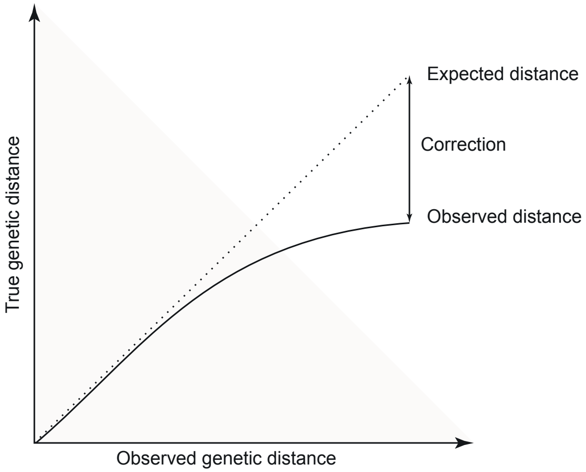

Figure 13. The effect of using over-simplified models on the observed genetic distance.

As discussed above (section 1.4.1.3), selective forces also have an influence on the nucleotide codon positions. For example, under purifying selection, nucleotide changes that will lead to synonymous changes are likely to occur more often than non-synonymous changes. Mutational process, where certain nucleotide changes are more likely than others, will also influence the evolutionary process; in DNA evolution transitions are usually more common than transversion. Thus, transition and transversion biases are often incorporated into a substitution model.

Markov models are often used when modeling sequence evolution. These are memoryless stochastic processes, which in the case of evolution is a reasonable assumption as evolution in general is memoryless [185]. The different Markovian models differ essentially in the parameterization of the rate matrix and in the modeling of rate variations across sites. The rate matrix **M** can be described by 8 parameters μ_{xy} corresponding to the 5 change of relative nucleotide base change, and the 3 parameters π_X corresponding to the relative base composition frequencies, as described below. However, depending on the data at hand, 8 parameters may represent an over-parameterization and more economical parameterizations are often desirable. The first model proposed was the simple Jukes-Cantor (JC), which assumes a constant rate for every possible change μ_{xy} [186]. To account for the transition vs. transversion biases Kimura used two parameters (the K2 model) [187]:

$$\mu_{xy} = \begin{cases} \alpha & \text{for transitions} \\ \beta & \text{for transversions} \end{cases} \tag{2.1}$$

To account for unbalanced base composition the HKY model includes three more parameters, the stationary frequencies: π_A, π_G, π_C and $\pi_T=1-(\pi_A+\pi_G+\pi_C)$. The most generalized model is REV or GTR (general time-reversible) model, where $\mu_{xy} = s_{xy}\,\pi_y$ and $s_{xy} = s_{yx}$. The twelve nondiagonal entries of **M** can therefore be described by 8 independent parameters under the assumption of reversibility, with 5 exchangeability terms s_{xy} and 3 stationary frequencies π_x. Note that as one of the s_{xy} terms is set to 1, such that the other terms become relative to 1, and in the case of π_X the sum of frequencies add to 1, each of the exchangeability terms and stationary frequencies are given as balance, reducing the number of free parameters to be estimated from data. In addition to these parameters it is often realistic to add rate heterogeneity across sites-parameters (for instance to account for the above described heterogeneity in *env*). This can be done in different ways, a popular and parameter saving option is to describe it using a Gamma distribution and include an invariant class (+G +I) [188].

Protein coding genes can also be analyzed at an amino acid or codon level. A codon is a triplet of nucleotide bases encoding for a specific amino acid. There are 61 codons that are classified into 20 groups, where each group encodes the same amino acid. Codon changes within a group are called synonymous and changes between groups are called non-synonymous. Goldman and Yang described the first codon model [189], which has 63 parameters, namely 60 stationary frequency parameters π_{xyz}, one transition rate α and one transversion rate β. In addition the non-synonymous/synonymous substitution ratio ω is estimated.

One fundamental assumption of the Markov models is stationarity where the base composition is assumed to be at equilibrium throughout the tree. This results in the likelihood independence from the location of the root. Model choice is important and can for example be done though comparisons of nested models. The DNA Markov models described above are special cases of REV/GTR and can be compared using likelihood ratio tests (LRTs). The AIC is a related likelihood-based measure appropriate for both nested and non-nested models.

2.3.2 Tree building

As already mentioned, a phylogenetic tree is a representation of the genealogical relationship among sequences. The tree consists of edges that connect the nodes. The branching-pattern of the tree is called the topology and the length of the branches may either be genetic divergence or the time covered by a branch. Sometimes only the topology is shown, such a tree is called a cladogram (Figure 14).

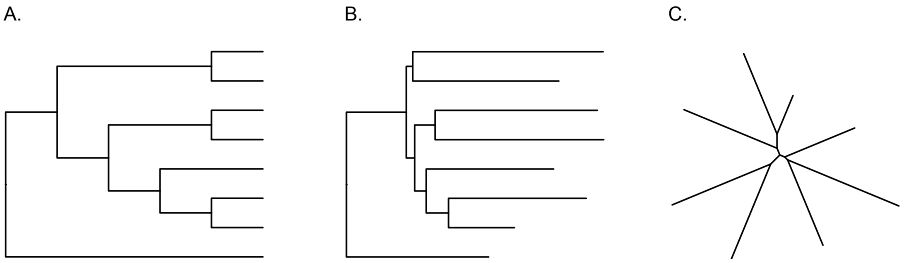

Figure 14. The same tree is shown in different ways: Cladogram (A), Phylogram (B), Unrooted tree (C).

There are various ways to infer a phylogenetic tree, which include but is not limited to distance methods, likelihood methods and Bayesian methods. In short, distance based methods calculates pairwise genetic distances between all sequences into a distance matrix, and a clustering algorithm is subsequently used to infer the tree. The most widely used are UPGMA and neighbor-joining [190,191]. The other methods mentioned above are character based methods and are often more time-consuming.

2.3.2.1 The maximum likelihood method

The likelihood is defined as the probability of observing the data when the parameters are given and is thus a function of the parameter values. The data consists of *s* aligned homologous sequences, each *n* nucleotides long, and can be represented as a *s* x *n* matrix $X = \{x_{jh}\}$, where x_{jh} is the *h*th nucleotide in the *j*th sequence. Let $\mathbf{x}_h$ denote the *h*th column in *X*. To define the likelihood one has to specify a model by which the data are generated, for example the K2 model specified above. In addition we have to assume that each site and each branch of the tree evolves independently. The length of the branch leading to node *i* is denoted t_i, defined as the expected number of nucleotide substitutions per site. Hence, the parameters in the model include the branch lengths and the substitution parameters and are collectively denoted θ. As we made the assumption of independent evolution among sites, the probability of the whole data is the product of the probabilities of data at individual sites and the log likelihood is a sum over sites in the sequence:

$$\mathrm{l} = \log(L) = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h|\theta)\} \tag{2.2}$$

The maximum likelihood (ML) method estimates θ by maximizing the log likelihood l, often using numerical optimization algorithms [192] Thus, the ML method uses an optimality criterion to assess a tree's fit to the data. However, an exhaustive tree search, where every possible tree is assessed is too computationally intensive. Instead heuristic searches with algorithms that explore parts of the tree space are used. These searches are done in various ways, but examples are branch-swapping by subtree pruning and regrafting (SPR), nearest neighbor interchange (NNI) and tree bisection and reconnection (TBR).

2.3.2.2 Bayesian inference

In the Bayesian approach a likelihood function $p(D \mid \theta)$ describes the probability of the data *D* given the parameters of θ. The prior distribution p(θ) expresses the uncertainty in the parameters prior to the observation of the data. Bayes' theorem provides the form of the posterior distribution $p(\theta \mid D)$, which describes the uncertainty in the parameters after observing the data:

$$p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{p(D)} \tag{2.3}$$

The denominator *p*(*D*) is the marginal probability of the data, averaged over all possible parameter values weighted by their prior distribution and is a normalizing constant. However, it is not achievable to compute $p(D)$ directly. In the mid-1990s Markov chain Monte Carlo (MCMC) methods were developed to calculate posterior probabilities of phylogenies on the basis of aligned DNA sequence data. This made it feasible to apply Bayesian inference to phylogenetic reconstruction. The MCMC is a general computational technique for evaluating sums and integrals. The *Monte Carlo* implies that the method is using random sampling and the *Markov chain* indicates a dependent sampling scheme. The primary ideas behind MCMC were created by physicist Nicholas Metropolis and colleagues over fifty years ago at the Los Alamos National Laboratory as part of a solution to a problem in physics [193]. [194]

2.3.2.3 Assessing the robustness of the tree

A result of the MCMC Bayesian inference of phylogeny is the generation of a posterior probability distribution of trees, weighted according to their posterior probability. Hence, you will not get one most likely tree, but a set of trees that explains the data best. Accordingly, you will get an estimation of the uncertainty in the inferred phylogeny (Figure 15).

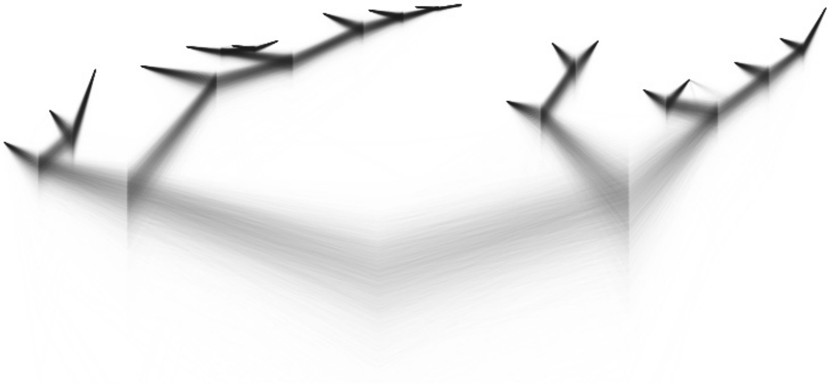

Figure 15. More than 10000 Bayesian trees shown in one picture, darker areas indicate high posterior probability, whereas diffuse parts illustrate the wider interval of branch lengths or topology miss-matches. The figure was produced using the DensiTree software [195].

In contrast, the robustness of the inferred phylogeny by distance and likelihood-based methods is often assessed separately from the reconstruction of the most likely tree. A widely used technique is the bootstrap, where columns of the sequence alignment are randomly sampled with replacement to form a new alignment [196,197]. The bootstrap is repeated multiple times (100-10000) and a new phylogeny is inferred for each bootstrap replicate. Thus, the percentage of times a specific branch exists in the bootstrap trees is the estimated bootstrap value for that branch. The resulting statistics is not clearly understood until today, thus the interpretation of bootstrap values are difficult, but are generally considered to be conservative [198]. LRTs have been used to assess if the length of a branch of interest is equal to zero (zero-branch length tests) and recently a fast approximate LRT have been introduced [199] and incorporated in the ML tree search algorithm [200].

2.4 THE USE OF PHYLOGENETICS

2.4.1 Detection of selective forces

Selection can be estimated by comparing homologous sequences using codon models of sequence evolution. By using the synonymous rate as the background rate, it is possible to estimate if the fixation of non-synonymous substitutions is influenced by natural selection, *i.e.* either negative or positive selection. Thus, one can estimate dS (rate of synonymous substitutions) and dN (rate of non-synonymous substitutions) at single codon sites. The ratio ω (dN/dS) measures the selective pressure at the amino-acid level. A $\omega = 1$ indicates that there is no influence of natural selection, while $\omega < 1$ indicates that non-synonymous mutations are deleterious and are removed from the population under a purifying or negative selection pressure. If the $\omega > 1$ the non-synonymous mutations are favored and is an indication of adaptive protein evolution and positive selection. Selection pressure can be calculated globally for the whole sequence alignment, but also at specific branches of a phylogenetic tree and at specific sites in the alignment.

2.4.2 Evolutionary rate estimation

Divergence describes the genetic distance from a reference point and is measured in substitutions per site. In order to get an estimate of evolutionary rate given in units of time it is necessary to incorporate a time model in your analysis. For the rapid evolving HIV virus it is possible to calculate the expected number of substitutions directly between sequences sampled at different points in time, and then calculate the evolutionary rate given in substitutions per site and time unit. This can be done through inferring a tree without a time model incorporated in the phylogenetic inference, and then using the dates of the tips to estimate the evolutionary rate measured in time.

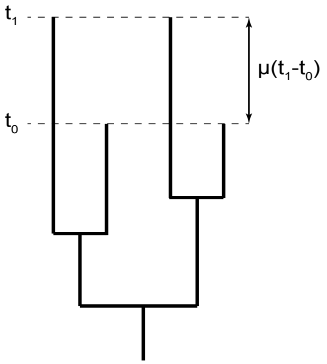

Figure 16. Schematic representation of a four taxa example, with taxa sampled at two different time points: t_0 and t_1. μ (t_1-t_0) is the genetic distance measured in substitution per site between the time separated samples.

The program TreeRate [201] is an example of this, which uses a rooting algorithm to find the best root of the tree based on time stamps of the tips followed by a calculation of genetic divergence between these tips. Moreover, there are inference methods where a time model is incorporated in the phylogenetic tree building. Thus, a clock-constraint is imposed and will influence the phylogenetic inference so that tips sampled at the same time will be equidistant from the root. The constraint would be a strict or constant clock where there is only a single rate of evolution [202], and where the substitutions would follow a Poisson distribution according to the Markov models of substitution [203]. However, the rate of which substitutions accumulate over time may not be constant. Thus, relaxed molecular clocks have been developed that allow the substitution rate to vary over the tree [204-208]. In conclusion, when sequences are sampled at different times, an independent measure of time derives from the intervals between the times of sampling (Figure 16).

2.4.3 Coalescent based methods

Coalescent theory is a part of theoretical population genetics, which studies the forces that produce and maintain genetic variation within a population. The coalescent describes the genetic ancestry of a sample under a specific model and makes predictions about patterns of genetic variation. The standard coalescent is based on the Wright-Fisher model that assumes that the generations are non-overlapping and that the population size is constant over time and finite [209-211]. As the reproduction is assumed to be random, genetic lineages will disappear by chance, which also is called random genetic drift. Coalescent theory makes it possible to infer population-level processes from a small random sample of sequences. Figure 17 gives an example of a coalescent model with constant population size and its relation to an inferred time-resolved phylogenetic tree.

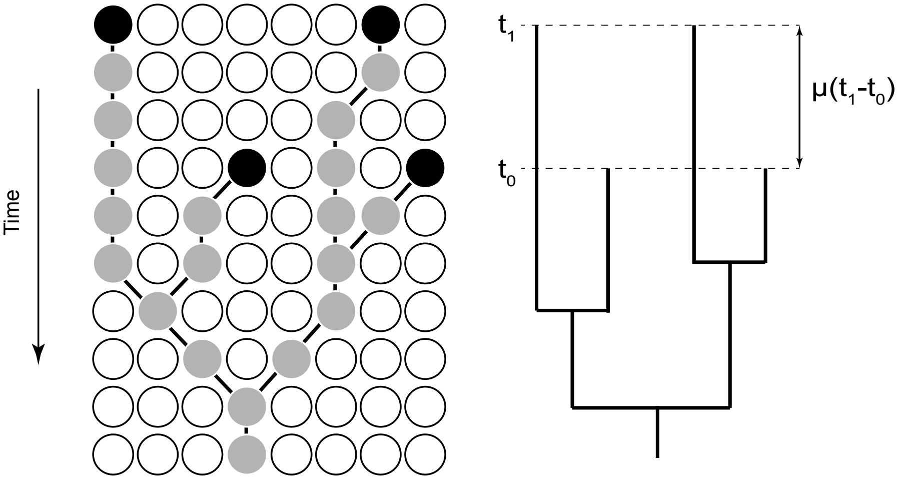

Figure 17. The coalescence process for the four taxa example in Figure 16. The effective population size over time is in this example constant. In each generation, a parent is chosen at random. Thus, coalescent events occur when more than one progeny randomly chooses the same parent.

It is possible to estimate phylogeny and coalescent parameters in conjunction [212]. Phylodynamics is a term coined by Grenfell *et al.* in 2004, where the link between molecular evolution of a pathogen and the population dynamics of the disease is aimed to be formalized [213]. The coalescent theory is central in phylodynamics as it represents a direct link between the gene sequences of the pathogen and the population dynamics of the pathogen, Figure 18, or host.

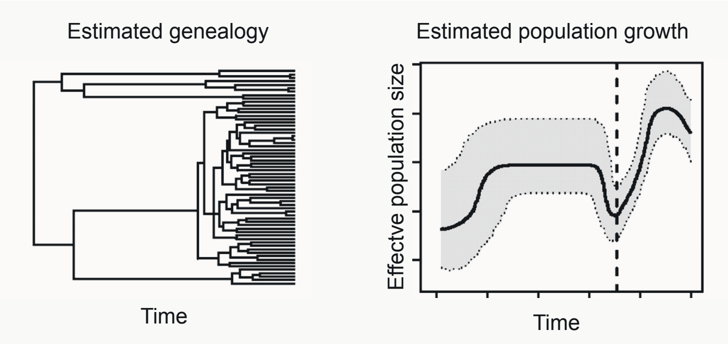

Figure 18. An example of an estimated genealogy (Left panel) and its inferred population growth over time using coalescent model (Right panel). Thus, the right panel shows the estimated effective viral population size (relative genetic diversity) over time. Sequences are derived from the HCV epidemic in Egypt, and the dashed line indicates the time when parenteral antischistosomal therapy was started to be administered intravenously in Egypt in 1920. Figure reproduced with permission from Minin V.N., et al. Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics Mol Biol Evol (2008) 25 (7): 1459-1471.

A reoccurring term in population genetics is the effective population size (N_e), which is usually smaller than the absolute population size N. The effective size of a population was originally defined to be the size of a Wright-Fisher population that would produce the same rate of genetic drift as the population of interest. Thus, it is the number of reproductive genetic variants in an ideal population that would show the same amount of dispersion of allele frequencies (i.e. diversity) under random genetic drift as the population under study.

3 RESULTS AND DISCUSSION

3.1 THE HIV-1 EPIDEMIC AMONG IDUs IN SWEDEN [I, II]

Two of the studies in this thesis focus on the HIV-1 spread among IDUs in Sweden and Stockholm. The IDU group has since the beginning of the Swedish HIV epidemic been affected disproportionally in comparison to the general population. Thus, active drug use is regarded to be associated with higher risk of acquisition of HIV infection, especially when injection equipment is shared [214]. This is not only true for Sweden but for many countries around the world. There have been a number of examples where the nature of the epidemics among IDUs has been explosive, involving fast spread between individuals [157,158,160,162,215,216]. On two occasions since the turn of the twentieth century, there has been a concern about increasing numbers of newly HIV-1 diagnosed individuals in Sweden with intravenous drug use as reported transmission route. In addition to pure epidemiological studies where the number of infected individuals and social- as well as other risk-factors associated with infection can be studied, HIV sequence data has been shown to be a great aid in estimating the history and dynamics of the events leading up to these increases of newly diagnosed HIV-1 cases. The first study where sequence data was used to study the HIV-1 epidemic among IDUs in Sweden was initiated after the increase of newly diagnosed cases in 2001 [I]. This study was followed up after a second increase among IDUs in Stockholm in 2006 [II]. Paper I included 47 IDUs diagnosed in 2001 and 2002 along with 50 local control sequences sampled between 1987 and 2004. Paper II included HIV sequence data from 70 individuals living in Stockholm diagnosed in 2004-2007 along with demographic and clinical information as well as the earlier data set.

The spread of subtype B

Phylogenetic analysis showed that the majority of the sequences in Paper I were of subtype B, and that the majority of the variants spreading in 2001-2002 were clustered in three regionally and genetically distinct Swedish transmission chains, Stockholm clusters I, II and the Sundsvall cluster. Furthermore, when looking over the whole sampling period, i.e. 1987-2004, there had been many imports from abroad with over 30 potential introductions. It is difficult to know whether these introductions, often seen as single Swedish sequences in the phylogenetic tree, were dead-end introductions or a reflection of limited sampling. However, focusing only on the study period of 2001-2002, where 66% of all diagnosed cases in Sweden were included, there had been 12 independent introductions of subtype B. Forty patients were diagnosed with subtype B infection over this time, and sequences from 27 of these (68%) clustered in the three large transmission chains, and sequences from an additional seven patients clustered in smaller Swedish transmission chains. Thus, during this study period, there was mostly domestic spread of subtype B where most transmissions were local. This tells us that the increase of HIV-1 among IDUs was mostly the consequence of spread within Sweden of already established local HIV-1 variants rather than extensive import from the outbreaks flourishing in nearby countries in Eastern Europe, which was a concern that led to the initiation of the study.

The dates of the most recent common ancestors (MRCAs) of the three larger Swedish transmission chains were estimated using time stamped sequence data. The Stockholm clusters were the oldest with MRCAs dated to the mid 1990s, while the Sundsvall cluster was younger with a MRCA existing around 1999. The date represents the first divergence event within that cluster. Thus, the estimated time is dependent on the sampling and could be moved back if there exist additional (unsampled) individuals whose viral sequences are basally positioned in the clusters. Such unsampled individuals are not unrealistic since there has been significant mortality among HIV-1 infected IDUs in Stockholm, especially before the advent of cART in 1996. The Sundsvall cluster had been followed up carefully and epidemiological information available including last HIV negative tests agreed well with the estimated onset of local HIV spread in Sundsvall. In addition, since we estimated that the evolutionary rate in the Swedish transmission chains was relatively high we concluded that the rate of spread in the Swedish subtype B epidemic was likely to have been slow [164].

Import of CRF01_AE from Helsinki

Already in Paper I, we documented that three Swedish IDUs diagnosed in 2002 were infected with a CRF01_AE variant that also was present among IDUs in Helsinki, Finland, which had experienced an HIV-1 outbreak among IDUs a few years earlier. Four years later, in the summer of 2006, an outbreak of CRF01_AE was discovered in Stockholm. Extensive HIV testing in the IDU communities in Stockholm was started and by 2007 more than 70 new cases of HIV-1 had been reported.

Paper II covered the study period of 2004-2007 and we found that 46 of the 70 study subjects (66%) were infected with the imported Helsinki CRF01_AE variant. Through phylogenetic inference we found that the Swedish CRF01_AE sequences formed a monophyletic cluster within Finnish CRF01_AE sequences sampled between 1998-2007, whereas the three Swedish CRF01_AE sequences from 2002 clustered within the Finnish sequences. Therefore, the three Swedish IDUs, who were infected with CRF01_AE already in 2002, were likely not carriers of the founding CRF01_AE virus. Instead, a separate introduction was the founder of the outbreak among IDUs in Stockholm in 2006 and a more complex exchange between Helsinki and Stockholm explains the multiple introductions into Sweden (Figure 19).

The local spread of the CRF01_AE variant causing the outbreak in Stockholm was dated to have started at or before February of 2003 (95% highest posterior distribution (HPD): July-2001, July-2004). The time-resolved phylogeography of the introductions of the Finnish CRF01_AE variant into Sweden can be seen in Figure 20.

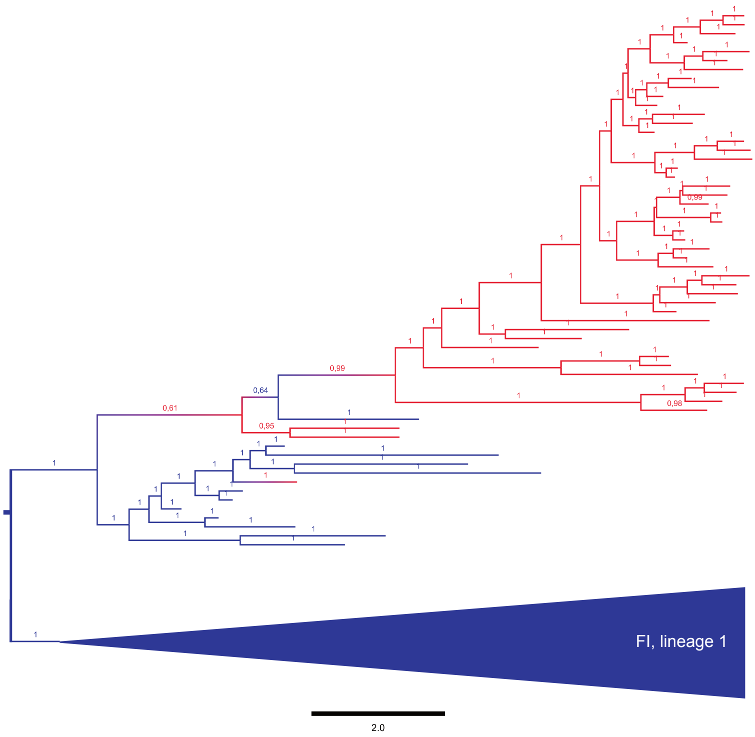

Figure 19. Bayesian maximum clade credibility tree of CRF01_AE V3 sequences from Stockholm and Helsinki. Red branches represent Swedish sequences, while Finnish sequences are colored blue. Posterior probability values of state changes are shown above each branch. Lineage 1 that only contains Finnish sequences is collapsed for readability.

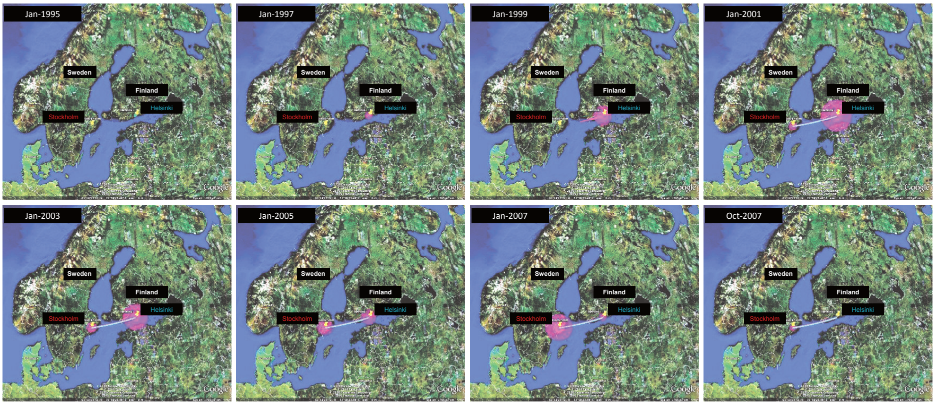

Figure 20. Phylogeographic analysis of the Finnish-Swedish CRF01_AE cluster inferred with the program BEAST [212,217]. Two geographical states are shown, Helsinki and Stockholm. The maps are based on satellite pictures made available in Google Earth (http://earth.google.com).

Parallel epidemics of HIV-1

Parallel to the CRF01_AE outbreak in 2006, a continued transmission of subtype B was observed, i.e. 23 of the 70 study subjects were infected with subtype B. All but one of the subtype B sequences clustered with previously identified (i.e. Paper I) local transmission chains in Stockholm. Hence, the earlier established Stockholm cluster I and II had continued to spread among Stockholm IDUs. This subtype B spread was in part characterized by fast spread, but the majority of the infections appeared to have been involved in slower spread.

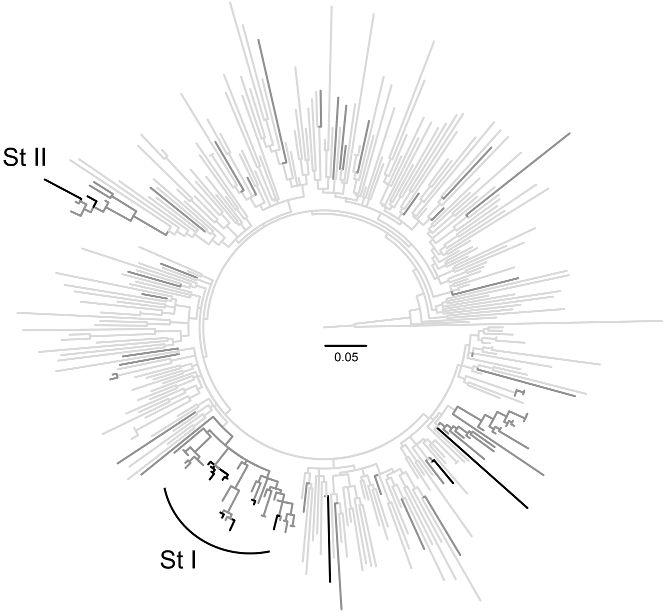

Figure 21. Maximum likelihood tree of subtype B V3 sequences from Stockholm and closely related database sequences. Black branches represent sequences from study [II], while dark grey branches represent sequences from study [I], other sequences are colored light grey. St I and II refer to Stockholm cluster I and II, respectively. The tree was rooted using subtype D sequence ELI (A07108).

The demographic characteristics of patients infected with subtype B or CRF01_AE were compared to investigate if the two variants had spread in different IDU subgroups in Stockholm. We found that a majority of the heroin users had CRF01_AE infections (83%), while amphetamine users had similar proportions of CRF01_AE (52%) and subtype B infections. When the parameters were mapped over the phylogenetic trees, there was no significant clustering of any demographic parameter in either CRF01_AE or in a subtype B sub-cluster of Stockholm cluster I. Thus, evident subgroups could not be established either between or within subtypes, possibly indicating interaction between drug users using different types of main drugs (*i.e.* amphetamine and heroin), and with different ages, housing status and gender. However, it is likely that there was some subgrouping even though we did not detect it with our baseline epidemiological information. It is unrealistic to envision panmixia, i.e. that all infected IDUs had similar degree of sharing of injection equipment with all other IDUs.

IDUs in the Finnish CRF01_AE outbreak had been reported to have higher plasma viral levels than Dutch IDUs infected with subtype B [218]. However, in our study there was no significant difference in plasma HIV-1 RNA levels between Swedish IDUs infected with subtype B or CRF01_AE nor CD4 counts. Moreover, we did not observe any significant clustering according to plasma HIV-1 RNA levels or CD4 counts in the phylogenetic trees.

The rapid spread of CRF01_AE was mirrored in a phylogenetic cluster with short internal and external branches. This indicates that the phylogenetic resolution of eventual subgroupings among those infected may have been poor. Nevertheless, if a structured standing network existed, the rate of spread between those communities must have been relatively large, and the connectivity high.

3.2 PHYLODYNAMIC TOOLS USED TO DESCRIBE AN HIV EPIDEMIC [I, II]

With fast evolving pathogens, such as HIV, influenza, dengue and hepatitis C, the viral genetic information can be used to study their spread in the host population [164,219-222]. In our studies we have developed our own phylogenetic method [I] and extended the use and interpretation of phylogenetic trees [II].

Estimating number of introductions [I]

Usually, estimations of independent introductions of HIV are done by comparing query sequences with sequences derived from earlier publications that are available in public databases such as Genbank. By choosing the sequences that are genetically related to the query sequences, it is possible to infer a phylogenetic tree that will put them in perspective to the global epidemic. However, this method is heavily dependent on the quality and quantity of earlier published sequences and how they relate to the query sequences. In Paper I we developed a tree-independent method, which is not as reliant on specific global reference sequences. Instead, a universal reference dataset was generated, that was matched to subtype and sampling year of the query dataset. Next, the data set was filtered to include only one sequence per patient. Thus, the sequences in this universal reference dataset can be considered to be epidemiologically unlinked, where branching events probably predate independent introductions into specific populations. From this dataset a node height distribution was calculated, which describes what is commonly considered epidemiologically unlinked branching events. Consequentially, the upper 95% percentile of this distribution was regarded as a cut-off for epidemiologically unlinked cases. By comparing the node height distributions of the universal reference dataset with that of the query dataset, it was possible to determine if branching events could be considered epidemiologically unlinked, or if a branching event had taken place closer in time and more likely in a local transmission chain. The method was validated using two previously published datasets consisting of Russian and Estonian sequences and applied on the subtype B sequences sampled in 2001-2002 from Swedish IDUs. The estimated number of introductions agreed well between the original phylogenetic analyses and our new method. We estimated that there were 12 independent subtype B introductions into the Swedish IDU population among the active lineages sampled in 2001-2002.

The method is still dependent on the quality of the universal reference dataset, and perhaps more important, it is still heavily dependent on the sequences sampled in the query dataset. Moreover, it is likely that the subtype under study must have had a star-like global spread. For example, it was not possible to apply the method on the CRF01_AE query dataset. CRF01_AE spread from Africa almost exclusively to South-East Asia (i.e. Thailand) where it caused a larger epidemic than in any other region on the globe. Thus, a careful evaluation of the universal reference dataset is crucial for a correct estimation of the number of independent introductions in a query dataset.

Detailed interpretation of phylogenetic trees [II]

In Paper II we had the possibility to compare two separate outbreaks of the same CRF01_AE variant spreading through the same transmission mode but that were at different phases of their epidemics. Thus, the peak of the Finnish outbreak took place around 1998 and was halted quickly in comparison to other outbreaks in nearby countries [216,218]. The Swedish outbreak was discovered in 2006 and continued until 2007 [II]. The dynamics of the two local epidemics were investigated using the genetic distances of a ML tree (Figure 22A).

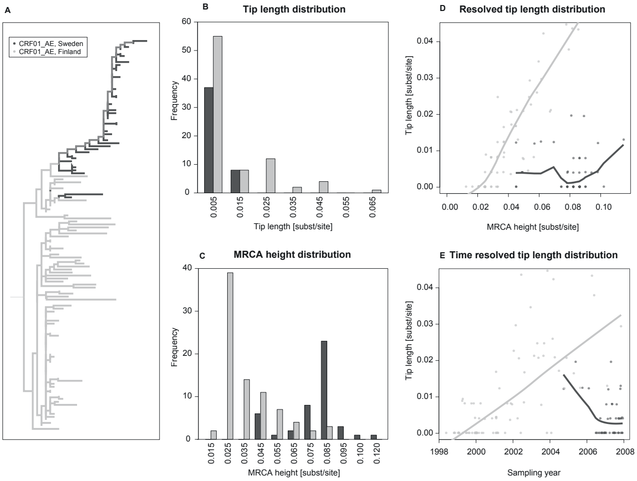

Figure 22. Detailed analysis of the CRF01_AE maximum likelihood tree (A). The tip length distribution for the Swedish and Finnish sequences is shown on intervals of 0.01 subs. site^{-1} (B). The number of tips at a certain height from MRCA$_{\text{Scand}}$, measured to the internal node of the tip (C). Furthermore, the tip length distribution was resolved according to the height from MRCA$_{\text{Scand}}$ to the internal node of each tip (the MRCA of each tip edge). A loess

regression is shown for each geographical region (D). The tip length distribution resolved according to sampling date of the taxa with loess regressions (E).

First of all, we could investigate how time from infection to diagnosis differed between the Helsinki and the Stockholm epidemic (Figure 22B). We found that long tip lengths were more common in the Finnish epidemic, which was not surprising as this epidemic is older. Furthermore, the coverage was not as high in the Helsinki epidemic. Hence, these long tips may be a reflection of missing taxa in the tree. However, the majority of tips were short in both epidemics, i.e. <0.005 subst. site^{-1} year^{-1}, indicating short time from infection to diagnosis [164,223].

Next, we investigated how the number of tips was distributed over the tree on a height basis (MRCA height), (Figure 22C). The MRCA height is a relative genetic measure of the genetic distance from the beginning of the Helsinki epidemic to each branching event resulting in a tip. In accordance with the number of reported cases in the two outbreaks, the Helsinki epidemic had culminated before the outbreak in Stockholm started (p<0.001, Wilcoxon rank sum test). Interestingly, when the tip lengths were resolved according to MRCA height (Figure 22D) and time of sampling (Figure 22E) it was evident that the two local outbreaks had somewhat different dynamics regarding time from infection to diagnosis. In the beginning of the Helsinki outbreak, tip lengths were short, however, as the epidemic progressed time to diagnosis increased. As mentioned above, missing taxa can give the same signal, however, independent data supports the conclusion that time to diagnosis has increased over the course of the Helsinki epidemic [224]. Both in Helsinki and Stockholm, rapid diagnosis of newly infected individuals took place during the peak of the outbreak. Notably, in Stockholm, time to diagnosis during pre- and post-outbreak was characterized by a prolonged period from date of infection to time of diagnosis. It is therefore possible that we will observe a similar increase in tip lengths in Stockholm as we have documented in Helsinki.

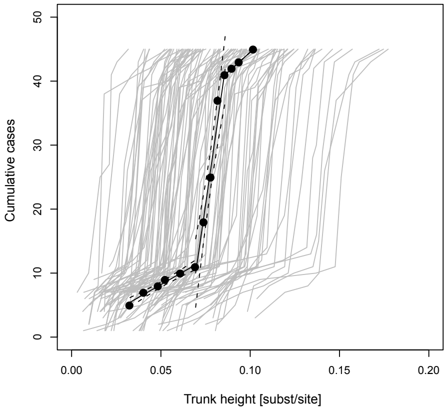

Figure 23. The cumulative history of the Swedish CRF01_AE cases. Red points show the height of the trunk from $MRCA_{Scand}$ in Panel A. The red lines are ordinary least squares estimates of the incidence rates inferred from the ML tree in Fig 22A, with the 95% regression confidence in dashed lines. The slope of phase 1 (pre-outbreak) was 160 cases/height (R^2=0.97, p<0.01, F-test), phase 2 (the outbreak) had a slope of 1930 cases/height (R^2=0.97, p<0.01, F-test) and phase 3 (post-outbreak) was at 251 cases/height (R2=1, p<0.01, F-test). The grey lines represent each of 100 bootstrap replicate trees. The three inferred bootstrap slopes were also well separated (p<0.001, Wilcoxon rank sum test). The initial spread rate of CRF01_AE in the ML tree (pre-outbreak) was used to normalize the slopes in the figure to allow for relative rate estimation. The outbreak accumulated 12 times more infections then during the pre-outbreak phase.

Focusing on the Stockholm epidemic, we measured the number of cumulative cases throughout the new CRF01_AE cluster, i.e. during the study period of 2004-2007. By normalizing the slope to 1 during the initial phase we could show that there was an increase in incidence rate with a factor of 12 as the second phase started, which shows that the relative force of the outbreak was strong. In agreement with the number of reported cases during 2008 and 2009 the incident rate decreased with a factor of more than 7 during the post-outbreak phase, almost returning to the pre-outbreak case incidence rate. Accordingly, the initial and final phase of the CRF01_AE incidence rate was similar to the case incidence rate measured in the subtype B transmission clusters I and II, within a factor of 1-2 (data not shown). This indicates that the CRF01_AE variant could continue to spread among IDUs in Stockholm at a similar rate as the subtype B variants that have been present for over a decade among Stockholm IDUs.

We used a ML tree accompanied by bootstrap trees but the use of Bayesian inference can be an alternative method to account for phylogenetic uncertainty, which probably would have been less conservative. By extending the interpretation of the inferred ML tree, we were able to measure the dynamics and the relative "force" of an HIV-1 outbreak. This can help to quickly get a measure of the urgency and potential an outbreak has. The interpretation scheme we applied here can readily be made for other types of phylogenetic methods and other rapidly evolving infectious agents as well.

3.3 WITHIN-PATIENT EVOLUTION OF HIV-2 [III]

As mentioned in the introduction, HIV-2 is the second causative agent of AIDS. The two human lentiviruses share similar genome organization and structure and the genetic identity in the homologous domains ranges from 30-40% in the more variable genes to 60% in the more conserved genes (*gag* and *pol*) [225]. The natural history of HIV-2 infection differs from that of HIV-1 in that the disease progression in general is slower, but the reasons for this important difference are still largely unknown [226,227], reviewed in [228,229]. In Paper III, we focused on the within-patient evolutionary rate of HIV-1 and HIV-2, in order to obtain a deeper understanding about whether evolutionary processes differ between the two viruses. To be able to compare the two viruses, we matched each HIV-2 patient to HIV-1 patients according to genetic region analyzed, RNA viral load, CD4 count, antiretroviral treatment and time of sampling. Two matching sequence datasets were included, one covering the gp125/gp120 (surface unit: SU), while the other covered the V3 region in *env*. An individual tree was inferred for each patient, while the relaxed molecular clock distribution was shared for each HIV type and each genetic region. Consequently, within-patient evolutionary rates were estimated for the SU region for HIV-2 and HIV-1, respectively, and compared to each other, and the same was done for the V3 dataset. The evolutionary rate comparisons were done by randomly sampling from each posterior distribution with replacement, thus we computed the posterior probability (PP) that one rate exceeded the other.

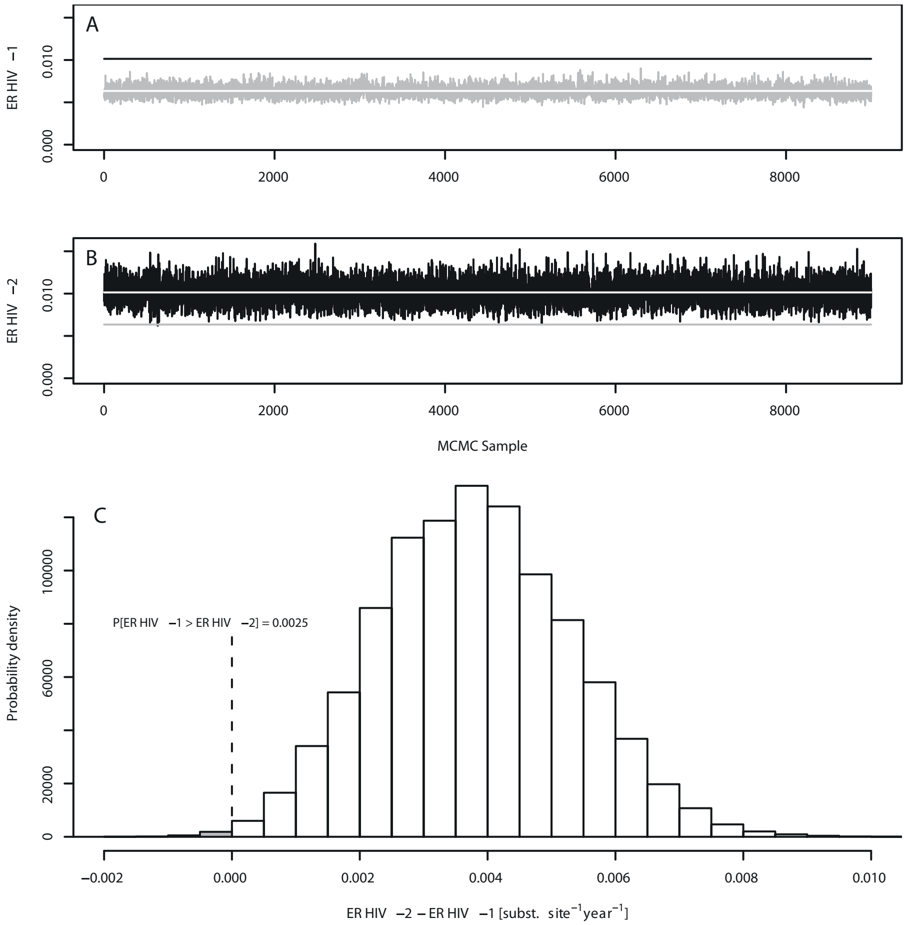

Figure 24. Comparing Bayesian estimates of HIV-2 and HIV-1 SU evolutionary rates. A) MCMC results after burn-in of the mean evolutionary rate of the hyper-parameter (ER) of HIV-1 in gray (μ_{HIV-1}) in each of 8999 sampled trees (ESS=3790). B) MCMC results after burn-in of the mean HIV-2 ER in black (μ_{HIV-1}) in each of 8999 sampled trees (ESS=7947). The white line shows the overall mean rate across all MCMC samples: $\overline{ER}_{HIV-2} =$ 0.01015 and $\overline{ER}_{HIV-1} =$ 0.00636 substitutions site^{-1} year $^{-1}$. The additional lines indicate the mean ER of HIV-2 (black) and HIV-1 (gray) in panels A and B, respectively. C) The significance level of $\overline{ER}_{HIV-2} > \overline{ER}_{HIV-1}$ was assessed by calculating μ_{HIV-2} - μ_{HIV-1} by sampling independently with replacement 1,000,000 times from $\pi(\mu_{HIV-2} | X_{HIV-2})$ and $\pi(\mu_{HIV-1} | X_{HIV-1})$. The fraction of values that are negative (in gray) thus estimates the probability of H_0 ($\overline{ER}_{HIV-1} > \overline{ER}_{HIV-2}$), which was p<0.01.

Our analyses showed that the HIV-2 virus evolved at a significantly higher rate than HIV-1, in both genetic regions analyzed. The SU region evolved at a rate of 10.2 x 10^{-3} as compared to 6.4 10^{-3} substitutions site^{-1} year^{-1} for HIV-2 and HIV-1, respectively (PP > 99 %), (Figure 24). The V3 region evolved two to three times faster, i.e. 29.4 x 10^{-3} substitutions site^{-1} year^{-1} for HIV-2 compared to HIV-1 at 12.3 x 10^{-3} substitutions site^{-1} year^{-1} (PP > 99).

Selection acting on the SU protein was estimated using the rates of synonymous and non-synonymous substitutions. We found that the faster evolutionary rate of HIV-2 as compared to HIV-1 was more prominent in synonymous sites than in non-synonymous sites, and that the non-synonymous rate was lower than the synonymous rate within the HIV-2 patients. These results are in agreement with earlier studies, which have found negative selection acting on the envelope gene of HIV-2 [230,231]. HIV-1 showed a more neutral selective signal. The stronger purifying selection on HIV-2 may be a reflection of more specific functions that need to be maintained in the envelope protein of HIV-2 and is supported by the rare escape from autologous neutralizing antibodies seen in HIV-2 infection [232-234].

Again, the faster evolutionary rate of HIV-2 was more pronounced at synonymous sites. Thus, the differences are likely to involve aspects influencing rate of mutations, such as replication error frequency, production rates and generation times [235]. As reported by Koblavi-Deme *et al.* 2004 the amount of HIV-2 RNA levels in plasma may depend on the immune status of the patient and they saw that for HIV-2 patients that had quantifiable levels of RNA, the amount of immune activation markers were similar to that seen in HIV-1 infected patients [75]. This may indicate that the possibility for HIV-2 to replicate increases as disease progresses. In addition, according to Sankale *et al*. 1995, HIV-2 infected persons with AIDS-related symptoms displays higher sequence heterogeneity than asymptomatic patients [236]. The most evident difference between HIV-1 and HIV-2 infection are the levels of viral load, where HIV-2 viral loads often are so low that they cannot be quantified. However, our matched datasets accounted for this difference, especially for the SU dataset, and all HIV-2 patients had quantifiable viral loads. Thus, our study involved somewhat unusual HIV-2 patients with progressive disease. Hence, differences in degree of immune activation and immunosuppression could possibly explain the differences between our results and those obtained by MacNeil *et al.* 2007a and Lemey *et al.* 2007, since their studies included few HIV-2 infected patients with AIDS related symptoms [92,128].

The limited number of studies of HIV-2 evolution in the literature is probably partly a reflection of the difficulties to extract and sequence viral HIV-2 directly from plasma. In addition, there are much fewer HIV-2, than HIV-1, infected persons in the world and furthermore, HIV-2 infections primarily occur in resource-poor settings, which complicates research on HIV-2. Ideally, single molecule sequencing of viral particles directly from plasma should be done to get a picture of the viral diversity and divergence as close to an *in vivo* setting as possible. In Paper III, we had to use primary isolates to be able to extract enough viral material and therefore chose to use population based sequencing, as there probably had been selection already in the isolation step. By including HIV-1 sequence data handled in the same way, we were able to compare our estimated HIV-1 evolutionary rate to earlier published results where sequencing of HIV-1 viral particles directly from plasma had been done. Our estimates of the *env* evolutionary rate of HIV-1 agreed well with these earlier studies both in the SU and the V3 region of HIV-1 [90,91,100,237]. Thus, the fact that we obtained our sequences from virus isolates, rather than directly from patient samples, probably had minor impact on the estimated evolutionary rate for both HIV-1 and HIV-2.

The difference in evolutionary rate between HIV-2 and HIV-1 that we have observed is probably not the reason for the differences seen in disease progression between the two viruses. More likely, the difference is a consequence. Furthermore, we only included patient at advanced disease stages. Thus, it would be interesting to investigate if these differences also can be seen during chronic infection, and to obtain a better understanding about why the potential difference in evolutionary rate exists, since this may be directly linked to the lower virulence of HIV-2.

3.4 WITHIN-PATIENT EVOLUTION OF HIV-1 [IV]

Many studies have focused on the within-patient evolutionary dynamics of HIV-1 over time frames of months to years. However, little is known about short-term evolution. In Paper IV, we generated sequence data from a chronically infected patient who was sampled with a time resolution of days, which is unique to my knowledge. Over the study period, which spanned 3 years, sampling was performed on days: -622, 1, 2, 3, 11, 18, 25, 32 and 522, where day 1 was defined as the starting day of the study. Thus, the sample day -622 represented a stored sample sequenced in retrospect. Approximately 7-11 sequences covering the entire envelope gene were generated for each day, except for day -622 where only three sequences could be amplified. We found evidence for subpopulation structure in the data. Six clades, named A through F, were identified, and if day -622 was excluded, four out of five subpopulations were present over the whole study period, and genetic distances within the subpopulations were relatively short (Figure 25).

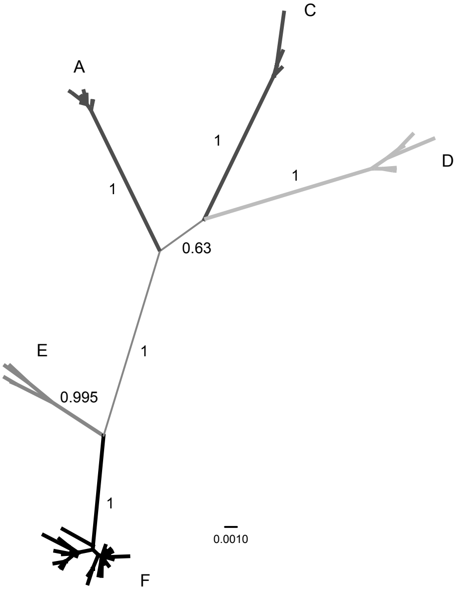

Figure 25. Maximum likelihood tree of the *env* sequences in Paper IV, excluding the 8 putative recombinants. ML bootstrap replicates (1000) were performed and the ratios are showed next to branches connecting subpopulations.

The evolutionary rate analyses were interesting, as it was possible to infer divergence estimates for sequences sampled only days apart. However, the overall signal for temporal evolution was missing in our data, in part due to the subpopulation structure. However, by focusing on individual subpopulations we could infer a within-subpopulation evolutionary rate in a similar way as in Paper III, i.e. using a hyperparameter to infer the evolutionary rate using BEAST. The evolutionary rate was estimated to be 2.3 x 10^{-3} (95 % HPD: 0.94, 3.7) x 10^{-3} substitutions site^{-1} year^{-1}. The estimated rate is a little lower, i.e. one forth to one third, compared to what others have reported [90], [III]. Then again, this is not surprising, as we did not calculate the combined substitution rate for all the sampled HIV-1 sequences, but instead concentrated on individual subpopulations. Interestingly, when focusing on the largest subpopulation (F) the temporal structure measured in divergence (substitution per site) in the envelope gene of HIV-1 could be resolved at about one month. A molecular clock did not fit the whole sequence dataset, consequently an overall temporal evolution was missing in our data. Instead, the viral population seemed to evolve using subpopulation frequency fluctuations. We investigated if these fluctuations were significant and we found that during day 1 through day 32, the fluctuations observed were consistent with constant within-patient frequencies. Conversely, between days 1 to 32 and day 522 the fluctuations became significant.

Under a neutral model of evolution, the size of the effective population N_e will influence the level of subpopulation fluctuations. Thus, large N_e will result in small fluctuations, while small N_e will result in large fluctuations, and given that no new genetic variants arise, one subpopulation will eventually take over and eliminate all other diversity. Assuming a neutral model, there were a number of perhaps unlikely events observed at day 522 and we used population genetics simulation to calculate the likelihood of each such event over a range of plausible values of N_e, i.e. 1 - 10^5. Interestingly, for a specific range of N_e values, the data observed was not unlikely under a neutral model of evolution. This range spanned the values ~600 - 1100, which interestingly coincided with independent estimates of N_e using sequence data assuming neutral evolution (i.e. 512 with 2σ range ± 162). Hence, the significant subpopulation frequency fluctuations observed at day 522 were considered consistent with a model of neutral evolution.

Sequence based estimates of selection were performed, which confirmed that within subpopulations the site-specific evolution was consistent with neutral evolution. However, evidence for site-specific selection was observed if the deep branches that connected the subpopulations were included. Furthermore, potential N-glycosylation sites (PNGS) were over-represented among positively selected sites, even though positive selection also was mapped to other amino acid changes over the tree. Additionally, the signal for positive selection was stronger on branches between subpopulations than on branches within the subpopulations. This indicates that positive selection on both non-PNGS and PNGS sites may have been important in the formation of the subpopulations. In summary, we could conclude that the presence of multiple subpopulations that fluctuated significantly over a time frame of 1.5 years was consistent with a neutral model of evolution, but that natural evolution was likely also to have been involved over longer time periods, i.e. as the subpopulations were formed.

We chose to use a limiting dilution approach (also known as single genome sequencing [SGS]) to study the HIV-1 population dynamics within a patient, partly because we wanted to sequence the whole envelope gene and partly because we wanted to minimize sequencing errors [167,168]. An alternative technique would have been ultra-deep sequencing but that would have provided us with much shorter sequences and higher frequency of sequencing errors. As a result thereof, the power of the analysis would increase with more sequences, i.e., more precise subpopulation frequencies, but decrease with shorter fragments, i.e., lowering the phylogenetic signal. In addition, the higher error rate of this approach would further decrease the phylogenetic signal and also make calculations of evolutionary rates very uncertain. As comparison, based on our original alignment excluding putative recombinants, we cut out short (200 and 500 nt) fragments, in the V3 and the more constant region in the beginning of *env*. As expected, the inferred ML trees based on shorter alignments did not have the same resolution and robustness as the full *env* alignment.

The patient was treatment naïve and had been HIV-1 infected for approximately 7 years at the start of the study (i.e. day 1). Hence, the within-patient evolutionary dynamics described here occurred under chronic and asymptomatic infection. In addition, it is important to point out that the study describes only one case of HIV-1 infection, thus, even though the high-frequent sampling has provided us with new insights in the short-term evolutionary dynamics of HIV-1 in chronic infection, the findings need to be confirmed in additional patients to allow more generalized conclusions. Nevertheless, the subpopulation structure and frequency shift dynamics described in this patient may have implications not only when estimating within-patient evolutionary rates but also on the inference of transmission histories between individuals using viral sequences. The effect of subpopulation frequencies where lineages are unavailable for sampling at one time-point but present at another may obscure transmission histories, in the sense that lineage-sorting effects may occur. Studies using a known transmission chain are currently underway in order to investigate these effects in detail (Maljkovic Berry et al. unpublished data).

As detailed in section 1.4.1, the within-patient evolution of HIV-1 has been suggested to involve both natural and neutral evolution during chronic infection. The seminal paper by Shankarappa presented a fairly distinctive model of the HIV-1 dynamics during the course of infection [90]. These sequences have been used over and over again to look at within-patient evolution of HIV-1, which is not surprising as it is a remarkable dataset. The proposed stabilization of divergence made by Shankarappa has been challenged by others that differentiated between synonymous and non-synonymous sites and using the same dataset, found that the stabilization of divergence at advanced disease stage only was true for non-synonymous sites [238]. Again using the same dataset but including follow-up samples, Lemey *et al.* separated synonymous and non-synonymous rates, and found a strong correlation between rates of synonymous substitutions and disease progression [92]. Furthermore, by looking at separate data, they saw higher non-synonymous rates in viruses with rapid phenotypic escape from autologous nAbs as compared to those with slow escape, but no correlation to disease progression was seen. In all, positive selection as a result of immune pressure from neutralizing antibodies on the envelope protein occurs, and has been shown in other study groups as well [121,124]. Thus, over longer time periods (i.e. months to years) natural selection occurs, in agreement with what we observed in Paper (IV). However, a neutral model of evolution has been proposed to be the general mode to accumulate mutations, again using the Shankarappa material [133]. Hence, 85% of the samples under investigation evolved in a manner consistent with neutral evolution. Nevertheless, evidence for re-occuring selective sweeps with fixation events occurring over the course of the infection were observed in seven out of nine individuals. These results agree with our estimates of short-term evolution in Paper IV, with neutral evolution occurring over short time and within subpopulations, whereas evidence for positive selection was found on the long branches that represent evolutionary history further back in time and over longer time periods. Paper III included estimates of within-patient *env* (SU) evolution of HIV-1, however, these patients were at a relatively advanced disease stage and were studied over a time period of years. According to the above reasoning, the evolution should be experiencing a slow down at the non-synonymous sites, while the synonymous sites would continue to diverge at similar rates. In agreement, the synonymous sites evolved slightly faster than the non-synonymous sites in gp120 (PP = 85%). As comparison I calculated the difference of the estimated synonymous and the non-synonymous rates of the C2-C5 region of HIV-1 after onset of disease progression presented by Lemey *et al.* [92], and found that they were not significantly different (p=0.3, Wilcoxon signed-rank test). Thus, the similar rates of evolution at synonymous and non-synonymous sites estimated in Paper III appear to agree with a model where the divergence rate at non-synonymous sites slows down as disease progresses.

4 CONCLUSIONS AND FUTURE PERSPECTIVE

The HIV virus displays a huge amount of genetic diversity and the combination of its high production rate, short generation time and high mutation rate gives the virus the opportunity to rapidly adapt to changes in its environment. A reoccurring theme throughout this thesis has been to describe HIV evolution, both on a population level and within single individuals. More specifically, the aim has been to apply carefully selected computational methods on sequence data in order to draw detailed conclusions and gain otherwise unobtainable insights into the underlying dynamics of HIV evolution and its relation to epidemic spread, HIV type and disease progression as well as to evolutionary forces acting during chronic HIV-1 infection. In order to achieve high quality analyzes, both existing and newly developed computational methods were used.

In Paper I we found that the HIV-1 epidemic among IDUs in Sweden has been characterized by spread of subtype B up until 2006. Since the beginning of the epidemic there have been many introductions of subtype B into Sweden, but only a few introductions have been established and caused further spread. Our molecular analyses indicated that during the study period of 2001-2002 at least 85% of those diagnosed had been infected through domestic transmission, which agrees with reported epidemiological data (88%) compiled by the Swedish Institute of Infectious Disease Control. Through phylogenetic inference we could also estimate how the spread had taken place within Sweden. Most domestic spread was local, where three larger transmission chains that had existed at least since the mid or end of the 1990s dominated the local transmissions. In addition, the rate of spread between individuals in these transmission chains appeared to have been relatively low. Several imported non-subtype B infections were discovered in Paper I, but none of these caused the massive outbreak among IDUs in Stockholm in 2006. Instead a new introduction or an unsampled case in Paper I was the founder of the CRF01_AE outbreak. Through phylogenetic inference in Paper II, we found that the founder virus originated from the IDU epidemic in Helsinki and had been present in Stockholm for some time before the onset of the outbreak. However, more than one introduction from Helsinki had taken place before the outbreak but had caused no or limited spread within Sweden. When the outbreak was discovered in Stockholm, time from infection to diagnosis was short, indicating an efficient outreach and discovery of those infected during this time. However, we saw an indication that time from infection to diagnosis increased shortly after the outbreak, leaving individuals HIV positive but unaware of their infection for longer time periods before diagnosis. IDUs must rely on their own clean injection equipment or to sero-sorting when sharing equipment, perhaps especially in the Stockholm area where no official needle-exchange programs have yet been established. Hence, the presence of individuals unaware of their change from HIV negative to positive HIV status may lead to further spread in these communities. When the CRF01_AE outbreak in Stockholm begun the incidence rate increased by a factor of 12. Thus, prevention measures will be most effective if they are in place already before outbreaks are starting, and early responses when indications of an outbreak arise are important in order to halt the course of an outbreak. In summary, our results from Paper I and II suggest that both subtype B and CRF01_AE will continue to spread at similar rates in the IDU communities in Stockholm. In the last two years (2008-2009) 16-17 cases per year have been reported to be newly HIV-1 infected through intravenous drug use in Sweden, which is in the same range as before the outbreak started in Stockholm in 2006 but not as low as it was the years prior (1998-2000) to the increase in 2001. In line with this, preliminary data from surveillance of transmitted drug resistance in 42 patients who have been newly diagnosed during 2008-2010 indicate that 12 (29%) were infected with subtype B and 25 (60%) with CRF01_AE. Encouragingly, it has recently been decided to implement needle-exchange programs in the Stockholm area, and that may hopefully lead to a structured harm reduction initiative, which may reduce the risk of outbreaks of HIV-1 among IDUs in the future. The combination of epidemiological data with phylogenetic inference has deepened our understanding of the many faceted HIV-1 epidemic among IDUs in Sweden. Importantly, Paper I and Paper II, illustrates the added value of using molecular epidemiology when conducting surveillance and prevention. Both studies were initiated due to an increase of newly diagnosed cases, but the reasons for these increases differed. Thus, the increase in 2001-2002 was due to continued spread of local variants in several small networks, while the second increase in 2006-2007 among IDUs in Stockholm was due to a larger outbreak of a newly introduced variant. The first increase was due to a catch-up of infections that had taken place during a longer period (reflected as long tips in the tree) and was not really a new outbreak, whereas the newly diagnosed individuals during the increase in 2006 were newly infected (with short tips in the tree), and a real outbreak. This shows that molecular epidemiology can give hard-to-reach information, which can be put into practice as means in infectious disease surveillance and prevention.

In order to enrich and develop the field of phylodynamics, it is essential to have good comparative data. Preferably in the form of high coverage in sequence sampling (dense sampling) and high quality epidemiological information about the individuals affected by and perhaps even those susceptible to the epidemic in order to incorporate realistic network models in the future. In addition, by bridging the still existing gap between the fields of epidemiology and phylogenetics and commence a joint effort to take advantage of the strengths in both fields a higher quality prevention strategy could be achieved. Sequence data is only one piece of the puzzle when describing and understanding an infectious disease epidemic, but still important. By using phylodynamics, not only as means to describe an epidemic in retrospect, but rather by doing so in real-time, a powerful tool arise which can be used to monitor potential outbreak scenarios and help understand where the potential risks may lay. In this thesis the epidemic among IDUs has been investigated. The step to applying the same methods to other transmission groups is not far. In Paper II, we chose to only include individuals with intravenous drug use as most likely route of infection. However, it would be interesting to study individuals infected through other routes in Stockholm in order to achieve an even larger picture of the spread dynamics during this time.

HIV-2 has been known to be less pathogenic than HIV-1 for more than 20 years [34,239,240]. However, the reasons for this difference remain largely unclear. The strongest correlation between disease progression and a parameter linked to the immune system is immune activation [54,76,77,241], where HIV-2 infection displays lower amount of immune activation than HIV-1 and thus has a lower risk to cause exhaustion of the immune system. A consequence has been proposed to involve a slow amount of synonymous changes in HIV-2 as a result of longer generation times due to a greater proportion of the viral population being in latently infected cells [92]. In Paper III, we showed that HIV-2 patients during accelerated disease progression had higher evolutionary rate than HIV-1 patients matched to viral load, CD4 count and antiretroviral treatment, especially at synonymous sites. However, the HIV-2 patients included were all atypical in the sense that they had quantifiable viral loads and usually CD4 counts below 500. The surprising finding and discrepancy with earlier studies may be due to different levels of immune activation. Unfortunately we did not measure markers for immune activation in Paper III, which would have been an interesting inclusion and perhaps a future line of research. An important conclusion drawn from Paper III is that HIV-2 within-patient evolution can be rapid especially at synonymous sites, which implies that the *in vivo* replication capacity of HIV-2 at least under certain circumstances may be large. Further studies with larger cohorts, follow-up over the whole course of infection and with patients with different rates of disease progression would be preferable to better understand the relationship between disease progression and evolutionary rate of both HIV-2 and HIV-1.

The dynamics of chronic HIV-1 infection is often described as a race between the virus ability to evolve and the immune systems ability to react to new viral variants [90]. Thus, the result would be a consecutive replacement of viral variants that would be reflected as a ladder-like phylogenetic tree. Ladder-like trees can be found in other viral spread situations as well. For example the yearly global epidemic of the influenza A virus that jumps between susceptible persons though the population, but does not establish chronic infection and that leaves recovered individuals with an immune protection against it [213]. Sequence data from HIV epidemics on the other hand mostly result in non-ladder-like phylogenetic trees, where many lineages survive and propagate over time, an example being the slow spread of subtype B in Sweden, presented in Paper I. However, exceptions exist, for example the ladder-like structure of the epidemic spread of CRF01_AE in Stockholm described in Paper II, resulting from a fast spread in a standing network. Thus, generalizations are difficult when talking about the tree structure either on a population basis or on an individual basis and the structure of a tree has important information. Already the non-ladder-like structure seen in the phylogenetic tree inferred from sequences in a chronic HIV-1 patient in Paper IV, gave a clue that these viral variants was experiencing other processes than strong positive selection. This was also shown through genetic population simulations, and a neutral model of evolution was consistent with the subpopulation structure and fluctuations seen in this patient. Others have already shown that selective sweeps followed by prolonged periods of neutral evolution may describe the evolutionary process in the chronic stage of HIV-1 infection [133]. Our results agree with this finding as we could detect variable selection over sites, only when including the long branches connecting the subpopulations. In all, over longer periods of time, i.e. months to years, natural selection is likely to have an influence on the evolution of viral variants in the body, whereas over shorter periods of time neutral evolution and random genetic drift is likely to be dominant.

Finally, HIV sequence data can be used in order to estimate processes of evolution on several different levels. However, in order to understand differential dynamics, sequence data often needs to be linked to other parameters, such as time of sampling, virological or clinical parameters and epidemic and demographic data. I hope that the work in this thesis has contributed to the understanding of HIV infection and spread, and that the research field will continue to deepened and broaden their knowledge about HIV/AIDS. I also hope that the results may be translated into more effective prevention against the further spread of the virus in the future.

5 ACKNOWLEDGEMENTS

The theoretical framework that needs to be in place in order to study evolutionary processes is as wide and complex as the dynamics of HIV itself. Thus, there are many people that I would like to acknowledge and that have contributed to the work in this thesis or influenced my learning curve in a positive way:

Jan Albert: You have been a rock of knowledge and expertise throughout the years, and you have always been amazingly available for questions and thoughts about Science, HIV and evolution. In addition, you have encouraged me to go abroad to Los Alamos, courses, workshops and conferences, which have been so important for me. I am immensely grateful for all the support you have shown me during my years as a PhD student.

Thomas Leitner: Thank you for welcoming me to the HIV database and for being such an inspiring teacher and discussion evoker when I have been there. Your metaphors are legendary (at least according to me) and your passion for research and straightforward ways are motivating. I hope at least some of your eye for details has gotten through to me.

I would like to say thank you to all the **patients** that have contributed to make these studies possible, without your participation none of this work could have been done.

The **co-authors** for fruitful collaborations, especially Tim Wallstrom and Ryan Gutenkunst for guiding me in Bayesian statistics and population genetics, and *Maria Axelsson* and *Mattias Mild* for excellent research and for always being there to discuss everything about anything and for reminding me that I'm actually pretty good at what I'm doing at least from time to time, and especially Mattias who was the proud reader of this thesis at almost final stage, and encouraging me at time when I needed it :-).

My co-workers at SMI over the years, *Eleonor Brandin* for introducing me to the lab and the work-place during my Master's project a long, long time ago. *Kajsa Apéria* for being such an essential part of the lab and *Afsaneh Heidarian* for being so supportive and compassionate. I also wish to say thank you to *Marianne Jansson* and *Annika Karlsson* for contributing to the group in many ways.

Dace Balode, thank you for always being there and for standing all my crazy talk. *Carina Peréz*! always on the fly and always leaning over when listening. *Salma Nowroozalizadeh,* your integrity and wisdom are a joy for a lost sailor in disguise. *Charlotte Hedskog*, thank you for bringing structure to the group. To all the other students in the large group: *Wendy Murillo*, *Leda Parham*, *Melissa Norström*, *Johanna Jernberg*, *Lina Josefsson*, *Viktor Dahl* and *Marcus Buggert*. It has been a pleasure sharing the same room with you all and I will miss it.

To the T-6 group at Los Alamos National Laboratory, thank you for taking such good care of me when I've been there, especially *Ina*, you have been a great fellow colleague and I hope we can collaborate in the future, *Werner*, who endlessly has helped me to install new software every time I show up. *Ming*, *Gayathri*, *Carla*, *John*, *Brian* and *Brian* and everyone else in the group for making it such a great working environment. Many of you have invited me to 3 o'clock tea, almost everyday, even though I've been boring, staying in my cubicle, and thank you for throwing me goodbye potlucks I think more than once…

Thank you to all the friends that I've gotten to know in the fairy land of New Mexico. I have been very, very lucky in meeting you all, and Santa Fe/Los Alamos feels like my second home by now.

Till mina vänner här hemma! *Emma*! Du har varit helt ovärderlig speciellt under denna fina höst. TACK. Till Organismen. Ni är bäst, och jag hoppas verkligen att vi fortsätter vara ett miss-masch av idéer och känslor för alltid. *Fredrik*, tack för att du är du. Mitt band som är så fint, hoppas vi fortsätter att äta en massa middagar ihop!

Min större familj: *Jürgen*, *Lena*, *Sara* och *Laura*. Tack för att ni är så roliga och snälla.

Max, tack för all kärlek!

Pappa, tack för att du finns. Jag älskar dig. **Mamma** och **Peter**, ni finns alltid där för mig och jag är oerhört tacksam för det! Jag älskar er också!

6 REFERENCES

1. 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.

2. Takehisa J, Kraus MH, Ayouba A, Bailes E, Van Heuverswyn F, et al. (2009) Origin and biology of simian immunodeficiency virus in wild-living western gorillas. J Virol 83: 1635-1648.

3. Sharp PM, Shaw GM, Hahn BH (2005) Simian immunodeficiency virus infection of chimpanzees. J Virol 79: 3891-3902.

4. Huet T, Cheynier R, Meyerhans A, Roelants G, Wain-Hobson S (1990) Genetic organization of a chimpanzee lentivirus related to HIV-1. Nature 345: 356-359.

5. Hirsch VM, Olmsted RA, Murphey-Corb M, Purcell RH, Johnson PR (1989) An African primate lentivirus (SIVsm) closely related to HIV-2. Nature 339: 389-392.

6. Gao F, Yue L, White AT, Pappas PG, Barchue J, et al. (1992) Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. Nature 358: 495-499.

7. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, et al. (1999) Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 397: 436-441.

8. Van Heuverswyn F, Li Y, Bailes E, Neel C, Lafay B, et al. (2007) Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. Virology 368: 155-171.

9. Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, et al. (2006) Human immunodeficiency viruses: SIV infection in wild gorillas. Nature 444: 164.

10. Neel C, Etienne L, Li Y, Takehisa J, Rudicell RS, et al. (2010) Molecular epidemiology of simian immunodeficiency virus infection in wild-living gorillas. J Virol 84: 1464-1476.

11. Santiago ML, Range F, Keele BF, Li Y, Bailes E, et al. (2005) Simian immunodeficiency virus infection in free-ranging sooty mangabeys (Cercocebus atys atys) from the Tai Forest, Cote d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. J Virol 79: 12515-12527.

12. Hahn BH, Shaw GM, De Cock KM, Sharp PM (2000) AIDS as a zoonosis: scientific and public health implications. Science 287: 607-614.

13. 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.

14. Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, et al. (2003) Tracing the origin and history of the HIV-2 epidemic. Proc Natl Acad Sci U S A 100: 6588-6592.

15. Wertheim JO, Worobey M (2009) Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. PLoS Comput Biol 5: e1000377.

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

17. Worobey M, Telfer P, Souquiere S, Hunter M, Coleman CA, et al. (2010) Island biogeography reveals the deep history of SIV. Science 329: 1487.

18. de Sousa JD, Muller V, Lemey P, Vandamme AM (2010) 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 5: e9936.

19. Marx PA, Alcabes PG, Drucker E (2001) 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 356: 911-920.

20. CDC (1981) Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men--New York City and California. MMWR Morb Mortal Wkly Rep 30: 305-308.

21. Hymes KB, Cheung T, Greene JB, Prose NS, Marcus A, et al. (1981) Kaposi's sarcoma in homosexual men-a report of eight cases. Lancet 2: 598-600.

22. CDC (1981) Pneumocystis pneumonia--Los Angeles. MMWR Morb Mortal Wkly Rep 30: 250-252.

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

24. Pitchenik AE, Fischl MA, Dickinson GM, Becker DM, Fournier AM, et al. (1983) Opportunistic infections and Kaposi's sarcoma among Haitians: evidence of a new acquired immunodeficiency state. Ann Intern Med 98: 277-284.

25. CDC (1982) Pneumocystis carinii pneumonia among persons with hemophilia A. MMWR Morb Mortal Wkly Rep 31: 365-367.

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

27. Gerstoft J, Malchow-Moller A, Bygbjerg I, Dickmeiss E, Enk C, et al. (1982) Severe acquired immunodeficiency in European homosexual men. Br Med J (Clin Res Ed) 285: 17-19.

28. Clumeck N, Sonnet J, Taelman H, Mascart-Lemone F, De Bruyere M, et al. (1984) Acquired immunodeficiency syndrome in African patients. N Engl J Med 310: 492-497.

29. CDC (1982) Possible transfusion-associated acquired immune deficiency syndrome (AIDS) - California. MMWR Morb Mortal Wkly Rep 31: 652-654.

30. Harris C, Small CB, Klein RS, Friedland GH, Moll B, et al. (1983) Immunodeficiency in female sexual partners of men with the acquired immunodeficiency syndrome. N Engl J Med 308: 1181-1184.

31. CDC (1983) Immunodeficiency among female sexual partners of males with acquired immune deficiency syndrome (AIDS) - New York. MMWR Morb Mortal Wkly Rep 31: 697-698.

32. Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, et al. (1983) Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220: 868-871.

33. Gallo RC, Salahuddin SZ, Popovic M, Shearer GM, Kaplan M, et al. (1984) Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. Science 224: 500-503.

34. Clavel F, Guetard D, Brun-Vezinet F, Chamaret S, Rey MA, et al. (1986) Isolation of a new human retrovirus from West African patients with AIDS. Science 233: 343-346.

35. UNAIDS (2008) Report on the global HIV/AIDS epidemic 2008. Geneva.

36. UNAIDS (2009) AIDS Epidemic Update 2009. Geneva.

37. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, et al. (2000) HIV-1 nomenclature proposal. Science 288: 55-56.

38. Carr JK, Salminen MO, Koch C, Gotte D, Artenstein AW, et al. (1996) Full-length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. J Virol 70: 5935-5943.

39. Gao F, Robertson DL, Morrison SG, Hui H, Craig S, et al. (1996) The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J Virol 70: 7013-7029.

40. Kostrikis LG, Bagdades E, Cao Y, Zhang L, Dimitriou D, et al. (1995) Genetic analysis of human immunodeficiency virus type 1 strains from patients in Cyprus: identification of a new subtype designated subtype I. J Virol 69: 6122-6130.

41. Zhang M, Foley B, Schultz AK, Macke JP, Bulla I, et al. (2010) The role of recombination in the emergence of a complex and dynamic HIV epidemic. Retrovirology 7: 25.

42. Carr JK, Wolfe ND, Torimiro JN, Tamoufe U, Mpoudi-Ngole E, et al. (2010) HIV-1 recombinants with multiple parental strains in low-prevalence, remote regions of Cameroon: evolutionary relics? Retrovirology 7: 39.

43. McCutchan FE (2006) Global epidemiology of HIV. J Med Virol 78 Suppl 1: S7-S12.

44. Thomson MM, Perez-Alvarez L, Najera R (2002) Molecular epidemiology of HIV-1 genetic forms and its significance for vaccine development and therapy. Lancet Infect Dis 2: 461-471.

45. Osmanov S, Pattou C, Walker N, Schwardlander B, Esparza J (2002) Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000. J Acquir Immune Defic Syndr 29: 184-190.

46. Geretti AM (2006) HIV-1 subtypes: epidemiology and significance for HIV management. Curr Opin Infect Dis 19: 1-7.

47. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM (2008) The challenge of HIV-1 subtype diversity. N Engl J Med 358: 1590-1602.

48. Schim van der Loeff MF, Aaby P (1999) Towards a better understanding of the epidemiology of HIV-2. AIDS 13 Suppl A: S69-84.

49. De Cock KM, Brun-Vezinet F (1989) Epidemiology of HIV-2 infection. AIDS 3 Suppl 1: S89-95.

50. Jonassen TO, Stene-Johansen K, Berg ES, Hungnes O, Lindboe CF, et al. (1997) Sequence analysis of HIV-1 group O from Norwegian patients infected in the 1960s. Virology 231: 43-47.

51. Poiesz BJ, Ruscetti FW, Gazdar AF, Bunn PA, Minna JD, et al. (1980) Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc Natl Acad Sci U S A 77: 7415-7419.

52. Grewe B, Uberla K (2010) The human immunodeficiency virus type 1 Rev protein: menage a trois during the early phase of the lentiviral replication cycle. J Gen Virol 91: 1893-1897.

53. Levin A, Hayouka Z, Friedler A, Brack-Werner R, Volsky DJ, et al. (2010) A novel role for the viral Rev protein in promoting resistance to superinfection by human immunodeficiency virus type 1. J Gen Virol 91: 1503-1513.

54. Schindler M, M¸nch J, Kutsch O, Li H, Santiago ML, et al. (2006) Nef-Mediated Suppression of T Cell Activation Was Lost in a Lentiviral Lineage that Gave Rise to HIV-1. Cell 125: 1055-1067.

55. Arhel NJ, Kirchhoff F (2009) Implications of Nef: host cell interactions in viral persistence and progression to AIDS. Curr Top Microbiol Immunol 339: 147-175.

56. Klein JS, Bjorkman PJ (2010) Few and far between: how HIV may be evading antibody avidity. PLoS Pathog 6: e1000908.

57. Eckert DM, Kim PS (2001) Mechanisms of viral membrane fusion and its inhibition. Annu Rev Biochem 70: 777-810.

58. Morner A, Bjorndal A, Albert J, Kewalramani VN, Littman DR, et al. (1999) Primary human immunodeficiency virus type 2 (HIV-2) isolates, like HIV-1 isolates, frequently use CCR5 but show promiscuity in coreceptor usage. J Virol 73: 2343-2349.

59. Blaak H, Boers PH, Gruters RA, Schuitemaker H, van der Ende ME, et al. (2005) CCR5, GPR15, and CXCR6 are major coreceptors of human immunodeficiency virus type 2 variants isolated from individuals with and without plasma viremia. J Virol 79: 1686-1700.

60. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, et al. (2002) HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110: 521-529.

61. Mellors JW, Rinaldo CR, Jr., Gupta P, White RM, Todd JA, et al. (1996) Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. Science 272: 1167-1170.

62. Gray RH, Wawer MJ, Brookmeyer R, Sewankambo NK, Serwadda D, et al. (2001) Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda. Lancet 357: 1149-1153.

63. Powers KA, Poole C, Pettifor AE, Cohen MS (2008) Rethinking the heterosexual infectivity of HIV-1: a systematic review and meta-analysis. Lancet Infect Dis 8: 553-563.

64. Wawer MJ, Gray RH, Sewankambo NK, Serwadda D, Li X, et al. (2005) Rates of HIV-1 transmission per coital act, by stage of HIV-1 infection, in Rakai, Uganda. J Infect Dis 191: 1403-1409.

65. Serwadda D, Gray RH, Sewankambo NK, Wabwire-Mangen F, Chen MZ, et al. (2003) Human immunodeficiency virus acquisition associated with genital ulcer disease and herpes simplex virus type 2 infection: a nested case-control study in Rakai, Uganda. J Infect Dis 188: 1492-1497.

66. Bailey RC, Moses S, Parker CB, Agot K, Maclean I, et al. (2007) Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. Lancet 369: 643-656.

67. Gray RH, Kigozi G, Serwadda D, Makumbi F, Watya S, et al. (2007) Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. Lancet 369: 657-666.

68. Wawer MJ, Makumbi F, Kigozi G, Serwadda D, Watya S, et al. (2009) Circumcision in HIV-infected men and its effect on HIV transmission to female partners in Rakai, Uganda: a randomised controlled trial. Lancet 374: 229-237.

69. Daar ES, Moudgil T, Meyer RD, Ho DD (1991) Transient high levels of viremia in patients with primary human immunodeficiency virus type 1 infection. N Engl J Med 324: 961-964.

70. Little SJ, McLean AR, Spina CA, Richman DD, Havlir DV (1999) Viral dynamics of acute HIV-1 infection. J Exp Med 190: 841-850.

71. Mehandru S, Poles MA, Tenner-Racz K, Horowitz A, Hurley A, et al. (2004) Primary HIV-1 infection is associated with preferential depletion of CD4+ T lymphocytes from effector sites in the gastrointestinal tract. J Exp Med 200: 761-770.

72. Guadalupe M, Reay E, Sankaran S, Prindiville T, Flamm J, et al. (2003) Severe CD4+ T-cell depletion in gut lymphoid tissue during primary human immunodeficiency virus type 1 infection and substantial delay in restoration following highly active antiretroviral therapy. J Virol 77: 11708-11717.

73. Andersson S, Norrgren H, da Silva Z, Biague A, Bamba S, et al. (2000) Plasma Viral Load in HIV-1 and HIV-2 Singly and Dually Infected Individuals in Guinea-Bissau, West Africa: Significantly Lower Plasma Virus Set Point in HIV-2 Infection Than in HIV-1 Infection. Arch Intern Med 160: 3286-3293.

74. Marlink R, Kanki P, Thior I, Travers K, Eisen G, et al. (1994) Reduced rate of disease development after HIV-2 infection as compared to HIV-1. Science 265: 1587-1590.

75. Koblavi-Deme S, Kestens L, Hanson D, Otten RA, Borget MY, et al. (2004) Differences in HIV-2 plasma viral load and immune activation in HIV-1 and HIV-2 dually infected persons and those infected with HIV-2 only in Abidjan, Cote D'Ivoire. AIDS 18: 413-419.

76. Michel P, Balde AT, Roussilhon C, Aribot G, Sarthou JL, et al. (2000) Reduced immune activation and T cell apoptosis in human immunodeficiency virus type 2 compared with type 1: correlation of T cell apoptosis with beta2 microglobulin concentration and disease evolution. J Infect Dis 181: 64-75.

77. Sousa AE, Carneiro J, Meier-Schellersheim M, Grossman Z, Victorino RM (2002) CD4 T cell depletion is linked directly to immune activation in the pathogenesis of HIV-1 and HIV-2 but only indirectly to the viral load. J Immunol 169: 3400-3406.

78. WHO (2007) WHO case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children. Geneva.

79. Palella FJ, Jr., Delaney KM, Moorman AC, Loveless MO, Fuhrer J, et al. (1998) Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. N Engl J Med 338: 853-860.

80. Duffy S, Shackelton LA, Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9: 267-276.

81. Mansky LM, Temin HM (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol 69: 5087-5094.

82. Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH (2010) The Nature, Position and Frequency of Mutations Made in a Single-Cycle of HIV-1 Replication. J Virol.

83. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148: 1667-1686.

84. Harris RS, Liddament MT (2004) Retroviral restriction by APOBEC proteins. Nat Rev Immunol 4: 868-877.

85. Chen J, Powell D, Hu WS (2006) High frequency of genetic recombination is a common feature of primate lentivirus replication. J Virol 80: 9651-9658.

86. Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, et al. (2002) Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J Virol 76: 11273-11282.

87. Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, et al. (2000) High rate of recombination throughout the human immunodeficiency virus type 1 genome. J Virol 74: 1234-1240.

88. Levy DN, Aldrovandi GM, Kutsch O, Shaw GM (2004) Dynamics of HIV-1 recombination in its natural target cells. Proc Natl Acad Sci U S A 101: 4204-4209.

89. Neher RA, Leitner T (2010) Recombination rate and selection strength in HIV intra-patient evolution. PLoS Comput Biol 6: e1000660.

90. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, et al. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73: 10489-10502.

91. Lee HY, Perelson AS, Park SC, Leitner T (2008) Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. PLoS Comput Biol 4: e1000240.

92. Lemey P, Kosakovsky Pond SL, Drummond AJ, Pybus OG, Shapiro B, et al. (2007) Synonymous Substitution Rates Predict HIV Disease Progression as a Result of Underlying Replication Dynamics. PLoS Comput Biol 3: e29.

93. Zhu T, Mo H, Wang N, Nam DS, Cao Y, et al. (1993) Genotypic and phenotypic characterization of HIV-1 patients with primary infection. Science 261: 1179-1181.

94. Zhang LQ, MacKenzie P, Cleland A, Holmes EC, Brown AJ, et al. (1993) Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J Virol 67: 3345-3356.

95. Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, et al. (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. PLoS ONE 5.

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

97. Kearney M, Maldarelli F, Shao W, Margolick JB, Daar ES, et al. (2009) Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. J Virol 83: 2715-2727.

98. Abrahams MR, Anderson JA, Giorgi EE, Seoighe C, Mlisana K, et al. (2009) Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. J Virol 83: 3556-3567.

99. Haaland RE, Hawkins PA, Salazar-Gonzalez J, Johnson A, Tichacek A, et al. (2009) Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. PLoS Pathog 5: e1000274.

100. Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, et al. (2008) Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J Virol 82: 3952-3970.

101. Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, et al. (2009) Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. J Exp Med 206: 1273-1289.

102. Li H, Bar KJ, Wang S, Decker JM, Chen Y, et al. (2010) High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men. PLoS Pathog 6: e1000890.

103. Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, et al. (1996) Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell 86: 367-377.

104. Huang Y, Paxton WA, Wolinsky SM, Neumann AU, Zhang L, et al. (1996) The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. Nat Med 2: 1240-1243.

105. Sagar M, Kirkegaard E, Long EM, Celum C, Buchbinder S, et al. (2004) Human immunodeficiency virus type 1 (HIV-1) diversity at time of infection is not restricted to certain risk groups or specific HIV-1 subtypes. J Virol 78: 7279-7283.

106. Masharsky AE, Dukhovlinova EN, Verevochkin SV, Toussova OV, Skochilov RV, et al. (2010) A substantial transmission bottleneck among newly and recently HIV-1-infected injection drug users in St Petersburg, Russia. J Infect Dis 201: 1697-1702.

107. Koup RA, Safrit JT, Cao Y, Andrews CA, McLeod G, et al. (1994) Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. J Virol 68: 4650-4655.

108. Borrow P, Lewicki H, Wei X, Horwitz MS, Peffer N, et al. (1997) Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. Nat Med 3: 205-211.

109. Goonetilleke N, Liu MK, Salazar-Gonzalez JF, Ferrari G, Giorgi E, et al. (2009) The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. J Exp Med 206: 1253-1272.

110. Streeck H, Jolin JS, Qi Y, Yassine-Diab B, Johnson RC, et al. (2009) Human immunodeficiency virus type 1-specific CD8+ T-cell responses during primary infection are major determinants of the viral set point and loss of CD4+ T cells. J Virol 83: 7641-7648.

111. Borrow P, Lewicki H, Hahn BH, Shaw GM, Oldstone MB (1994) Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. J Virol 68: 6103-6110.

112. Tomaras GD, Yates NL, Liu P, Qin L, Fouda GG, et al. (2008) Initial B-cell responses to transmitted human immunodeficiency virus type 1: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. J Virol 82: 12449-12463.

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

114. Richman DD, Wrin T, Little SJ, Petropoulos CJ (2003) Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc Natl Acad Sci U S A 100: 4144-4149.

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

116. Karlsson AC, Iversen AK, Chapman JM, de Oliviera T, Spotts G, et al. (2007) Sequential broadening of CTL responses in early HIV-1 infection is associated with viral escape. PLoS ONE 2: e225.

117. Brockman MA, Brumme ZL, Brumme CJ, Miura T, Sela J, et al. (2010) Early selection in Gag by protective HLA alleles contributes to reduced HIV-1 replication capacity that may be largely compensated in chronic infection. J Virol.

118. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317: 944-947.

119. Carrington M, O'Brien SJ (2003) The influence of HLA genotype on AIDS. Annu Rev Med 54: 535-551.

120. Migueles SA, Sabbaghian MS, Shupert WL, Bettinotti MP, Marincola FM, et al. (2000) HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. Proc Natl Acad Sci U S A 97: 2709-2714.

121. van Gils MJ, Bunnik EM, Burger JA, Jacob Y, Schweighardt B, et al. (2010) Rapid escape from preserved cross-reactive neutralizing humoral immunity without loss of viral fitness in HIV-1-infected progressors and long-term nonprogressors. J Virol 84: 3576-3585.

122. Wei X, Decker JM, Wang S, Hui H, Kappes JC, et al. (2003) Antibody neutralization and escape by HIV-1. Nature 422: 307-312.

123. Dhillon AK, Donners H, Pantophlet R, Johnson WE, Decker JM, et al. (2007) Dissecting the neutralizing antibody specificities of broadly neutralizing sera from human immunodeficiency virus type 1-infected donors. J Virol 81: 6548-6562.

124. Frost SD, Wrin T, Smith DM, Kosakovsky Pond SL, Liu Y, et al. (2005) Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection. Proc Natl Acad Sci U S A 102: 18514-18519.

125. Shi Y, Brandin E, Vincic E, Jansson M, Blaxhult A, et al. (2005) Evolution of human immunodeficiency virus type 2 coreceptor usage, autologous neutralization, envelope sequence and glycosylation. J Gen Virol 86: 3385-3396.

126. Li Y, Luo L, Rasool N, Kang CY (1993) Glycosylation is necessary for the correct folding of human immunodeficiency virus gp120 in CD4 binding. J Virol 67: 584-588.

127. Pollakis G, Kang S, Kliphuis A, Chalaby MI, Goudsmit J, et al. (2001) N-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization. J Biol Chem 276: 13433-13441.

128. MacNeil A, Sankale JL, Meloni ST, Sarr AD, Mboup S, et al. (2007) Long-term intrapatient viral evolution during HIV-2 infection. J Infect Dis 195: 726-733. Epub 2007 Jan 2018.

129. Scarlatti G, Tresoldi E, Bjorndal A, Fredriksson R, Colognesi C, et al. (1997) In vivo evolution of HIV-1 co-receptor usage and sensitivity to chemokine-mediated suppression. Nat Med 3: 1259-1265.

130. Williamson S (2003) Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. Mol Biol Evol 20: 1318-1325.

131. Edwards CT, Holmes EC, Pybus OG, Wilson DJ, Viscidi RP, et al. (2006) Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. Genetics 174: 1441-1453.

132. Barroso H, Taveira N (2005) Evidence for negative selective pressure in HIV-2 evolution in vivo. Infect Genet Evol 5: 239-246.

133. Shriner D, Shankarappa R, Jensen MA, Nickle DC, Mittler JE, et al. (2004) Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. Genetics 166: 1155-1164.

134. Voronin Y, Holte S, Overbaugh J, Emerman M (2009) Genetic drift of HIV populations in culture. PLoS Genet 5: e1000431.

135. Chun TW, Carruth L, Finzi D, Shen X, DiGiuseppe JA, et al. (1997) Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature 387: 183-188.

136. Haase AT, Henry K, Zupancic M, Sedgewick G, Faust RA, et al. (1996) Quantitative image analysis of HIV-1 infection in lymphoid tissue. Science 274: 985-989.

137. Kouyos RD, Althaus CL, Bonhoeffer S (2006) Stochastic or deterministic: what is the effective population size of HIV-1? Trends Microbiol 14: 507-511.

138. Nijhuis M, Boucher CA, Schipper P, Leitner T, Schuurman R, et al. (1998) Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. Proc Natl Acad Sci U S A 95: 14441-14446.

139. Shriner D, Liu Y, Nickle DC, Mullins JI (2006) Evolution of intrahost HIV-1 genetic diversity during chronic infection. Evolution 60: 1165-1176.

140. Achaz G, Palmer S, Kearney M, Maldarelli F, Mellors JW, et al. (2004) A robust measure of HIV-1 population turnover within chronically infected individuals. Mol Biol Evol 21: 1902-1912.

141. Vercauteren J, Wensing AM, van de Vijver DA, Albert J, Balotta C, et al. (2009) Transmission of drug-resistant HIV-1 is stabilizing in Europe. J Infect Dis 200: 1503-1508.

142. Yusim K, Kesmir C, Gaschen B, Addo MM, Altfeld M, et al. (2002) Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation. J Virol 76: 8757-8768.

143. Kawashima Y, Pfafferott K, Frater J, Matthews P, Payne R, et al. (2009) Adaptation of HIV-1 to human leukocyte antigen class I. Nature 458: 641-645.

144. Schmid BV, Kesmir C, de Boer RJ (2009) The distribution of CTL epitopes in HIV-1 appears to be random, and similar to that of other proteomes. BMC Evol Biol 9: 184.

145. Bhattacharya T, Daniels M, Heckerman D, Foley B, Frahm N, et al. (2007) Founder effects in the assessment of HIV polymorphisms and HLA allele associations. Science 315: 1583-1586.

146. Fraser C, Hollingsworth TD, Chapman R, de Wolf F, Hanage WP (2007) Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. PNAS 104: 17441-17446.

147. Hollingsworth TD, Laeyendecker O, Shirreff G, Donnelly CA, Serwadda D, et al. (2010) HIV-1 transmitting couples have similar viral load set-points in Rakai, Uganda. PLoS Pathog 6: e1000876.

148. Sagar M (2010) HIV-1 Transmission Biology: Selection and Characteristics of Infecting Viruses. The Journal of Infectious Diseases 202: S289-S296.

149. May RM, Anderson RM (1987) Transmission dynamics of HIV infection. Nature 326: 137-142.

150. Doherty IA, Padian NS, Marlow C, Aral SO (2005) Determinants and consequences of sexual networks as they affect the spread of sexually transmitted infections. J Infect Dis 191 Suppl 1: S42-54.

151. Liljeros F, Edling CR, Amaral LA, Stanley HE, Aberg Y (2001) The web of human sexual contacts. Nature 411: 907-908.

152. Keeling MJ, Eames KT (2005) Networks and epidemic models. J R Soc Interface 2: 295-307.

153. Ward H (2007) Prevention strategies for sexually transmitted infections: importance of sexual network structure and epidemic phase. Sex Transm Infect 83 Suppl 1: i43-49.

154. Bearman PS, Moody J, Stovel K (2004) Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks. American Journal of Sociology 110: 44-91.

155. Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J (1996) Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci U S A 93: 10864-10869.

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

157. Bobkov A, Cheingsong-Popov R, Selimova L, Ladnaya W, Kazennova E, et al. (1997) An HIV type 1 epidemic among injecting drug users in the former Soviet Union caused by a homogenous subtype A strain. AIDS Res Hum Retroviruses 13: 1195-1201.

158. Balode D, Ferdats A, Dievberna I, Viksna L, Rozentale B, et al. (2004) Rapid Epidemic Spread of HIV Type 1 Subtype A1 among Intravenous Drug Users in Latvia and Slower Spread of Subtype B among Other Risk Groups. AIDS Res Hum Retroviruses 20: 245-249.

159. Bobkov A, Kazennova E, Selimova L, Bobkova M, Khanina T, et al. (1998) A sudden epidemic of HIV type 1 among injecting drug users in the former Soviet Union: identification of subtype A, subtype B, and novel gagA/envB recombinants. AIDS Res Hum Retroviruses 14: 669-676.

160. Kalish ML, Baldwin A, Raktham S, Wasi C, Luo CC, et al. (1995) The evolving molecular epidemiology of HIV-1 envelope subtypes in injecting drug users in Bangkok, Thailand: implications for HIV vaccine trials. AIDS 9: 851-857.

161. Liitsola K, Ristola M, Holmstrom P, Salminen M, Brummer-Korvenkontio H, et al. (2000) An outbreak of the circulating recombinant form AECM240 HIV-1 in the Finnish injection drug user population. AIDS 14: 2613-2615.

162. Liitsola K, Tashkinova I, Laukkanen T, Korovina G, Smolskaja T, et al. (1998) HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad. AIDS 12: 1907-1919.

163. Lukashov V, Karamov E, Eremin V, Titov L, Goudsmit J (1998) Extreme founder effect in an HIV type 1 subtype A epidemic among drug users in Svetlogorsk, Belarus. AIDS Research and Human Retroviruses 14: 1299-1303.

164. Maljkovic Berry I, Ribeiro R, Kothari M, Athreya G, Daniels M, et al. (2007) Unequal Evolutionary Rates in the Human Immunodeficiency Virus Type 1 (HIV-1) Pandemic: the Evolutionary Rate of HIV-1 Slows Down When the Epidemic Rate Increases. J Virol 81: 10625-10635.

165. Socialstyreslen (2010) Ungass - Country Progress Report 2010 Sweden Stockholm.

166. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171: 737-738.

167. Simmonds P, Balfe P, Peutherer JF, Ludlam CA, Bishop JO, et al. (1990) Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. J Virol 64: 864-872.

168. Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, et al. (2005) Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. J Clin Microbiol 43: 406-413.

169. Hedskog C, Mild M, Jernberg J, Sherwood E, Bratt G, et al. (2010) Dynamics of HIV-1 quasispecies during antiviral treatment dissected using ultra-deep pyrosequencing. PLoS ONE 5: e11345.

170. HMMER (2010) http://hmmer.janelia.org/.

171. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

172. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33: 511-518.

173. RIP http://www.hiv.lanl.gov/content/sequence/RIP/RIP.html.

174. Salminen MO, Carr JK, Burke DS, McCutchan FE (1995) Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses 11: 1423-1425.

175. Schultz AK, Zhang M, Leitner T, Kuiken C, Korber B, et al. (2006) A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. BMC Bioinformatics 7: 265.

176. Schultz AK, Zhang M, Bulla I, Leitner T, Korber B, et al. (2009) jpHMM: improving the reliability of recombination prediction in HIV-1. Nucleic Acids Res 37: W647-651.

177. Maydt J, Lengauer T (2006) Recco: recombination analysis using cost optimization. Bioinformatics 22: 1064-1071.

178. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics 22: 3096-3098.

179. Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21: 255-265.

180. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254-267.

181. Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. Genetics 172: 2665-2681.

182. Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111: 147-164.

183. Suchard MA, Redelings BD (2006) BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics 22: 2047-2048.

184. Lunter G, Miklos I, Drummond A, Jensen JL, Hein J (2005) Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics 6: 83.

185. Lio P, Goldman N (1998) Models of molecular evolution and phylogeny. Genome Res 8: 1233-1244.

186. Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN, editor. In *Mammalian protein metabolism*. New York: Academic Press. pp. 21-123.

187. Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111-120.

188. Leitner T, Kumar S, Albert J (1997) Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J Virol 71: 4761-4770.

189. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725-736.

190. Sokal RR, Sneath PHA (1963) Numerical Taxonomy; Co. WHFa, editor. San Fransisco, CA.

191. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

192. Yang Z (2006) Computational Molecular Evolution. New York: Oxford University Press.

193. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics 21: 1087-1092.

194. Larget B (2005) Introduction to Markov Chain Monte Carlo Methods in Molecular Evolution. In: Nielsen R, editor. Statistics for Biology and Health. New York: Springer Science+Business Media, Inc.

195. Bouckaert RR (2010) DensiTree: making sense of sets of phylogenetic trees. Bioinformatics 26: 1372-1373.

196. Efron B. The jacknife, the bootstrap, and other resampling techniques; 1982; Philadelphia. Society of Industrial and Applied Mathematics.

197. Felsenstein J (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39: 783-791.

198. Hillis DM, Bull JJ (1993) An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. Systematic Biology 42: 182-192.

199. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol 55: 539-552.

200. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704.

201. Maljkovic Berry I, Athreya G, Kothari M, Daniels M, Bruno WJ, et al. (2009) The evolutionary rate dynamically tracks changes in HIV-1 epidemics: Application of a simple method for optimizing the evolutionary rate in phylogenetic trees with longitudinal data. Epidemics In Press, Corrected Proof.

202. Rambaut A (2000) Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16: 395-399.

203. Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Vogel VBaHJ, editor; 1965 New York. Academic Press. pp. 97-166.

204. Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. Molecular Biology and Evolution 14: 1218-1231.

205. Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol. pp. e88.

206. Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol 19: 101-109.

207. Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol 15: 1647-1657.

208. Drummond AJ, Suchard MA (2010) Bayesian random local clocks, or one rate to rule them all. BMC Biol 8: 114.

209. Ewens WJ (1979) Mathematical Population Genetics. Berlin: Springer-Verlag.

210. Kingman JFC (1982) The coalescent. Stochastic Processes and their Applications 13: 235-248.

211. Kingman JFC (1982) On the Genealogy of Large Populations. Journal of Applied Probability 19: 27-43.

212. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.

213. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, et al. (2004) Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303: 327-332.

214. Norden L, Lidman C (2005) Differentiated risk behaviour for HIV and hepatitis among injecting drug users (IDUs). Scand J Infect Dis 37: 493-496.

215. Tran TT, Maljkovic I, Swartling S, Phung DC, Chiodi F, et al. (2004) HIV-1 CRF01_AE in intravenous drug users in Hanoi, Vietnam. AIDS Res Hum Retroviruses 20: 341-345.

216. Kivela P, Krol A, Simola S, Vaattovaara M, Tuomola P, et al. (2007) HIV outbreak among injecting drug users in the Helsinki region: social and geographical pockets. Eur J Public Health 17: 381-386.

217. Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009) Bayesian phylogeography finds its roots. PLoS Comput Biol 5: e1000520.

218. Kivela PS, Krol A, Salminen MO, Geskus RB, Suni JI, et al. (2005) High plasma HIV load in the CRF01-AE outbreak among injecting drug users in Finland. Scand J Infect Dis 37: 276-283.

219. Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, et al. (2009) Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog 5: e1000590.

220. Bennett SN, Drummond AJ, Kapan DD, Suchard MA, Munoz-Jordan JL, et al. (2010) Epidemic dynamics revealed in dengue evolution. Mol Biol Evol 27: 811-818.

221. Pybus OG, Drummond AJ, Nakano T, Robertson BH, Rambaut A (2003) The epidemiology and iatrogenic transmission of hepatitis C virus in Egypt: a Bayesian coalescent approach. Mol Biol Evol 20: 381-387.

222. Rambaut A, Holmes E (2009) The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. PLoS Curr Influenza: RRN1003.

223. Leitner T, Albert J (1999) The molecular clock of HIV-1 unveiled through analysis of a known transmission history. PNAS 96: 10752-10757.

224. Kivela PS, Krol A, Salminen MO, Ristola MA (2009) Determinants of late HIV diagnosis among different transmission groups in Finland from 1985 to 2005. HIV Med.

225. Guyader M, Emerman M, Sonigo P, Clavel F, Montagnier L, et al. (1987) Genome organization and transactivation of the human immunodeficiency virus type 2. Nature 326: 662-669.

226. Marlink R, Kanki P, Thior I, Travers K, Eisen G, et al. (1994) Reduced rate of disease development after HIV-2 infection as compared to HIV-1. Science 265: 1587-1590.

227. Jaffar S, Wilkins A, Ngom PT, Sabally S, Corrah T, et al. (1997) Rate of decline of percentage CD4+ cells is faster in HIV-1 than in HIV-2 infection. J Acquir Immune Defic Syndr Hum Retrovirol 16: 327-332.

228. Reeves JD, Doms RW (2002) Human immunodeficiency virus type 2. J Gen Virol 83: 1253-1265.

229. de Silva TI, Cotten M, Rowland-Jones SL (2008) HIV-2: the forgotten AIDS virus. Trends Microbiol 16: 588-595.

230. Barroso H, Taveira N (2005) Evidence for negative selective pressure in HIV-2 evolution in vivo. Infect Genet Evol 5: 239-346.

231. Choisy M, Woelk CH, Guegan JF, Robertson DL (2004) Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J Virol 78: 1962-1970.

232. Cavaleiro R, Brunn GJ, Albuquerque AS, Victorino RM, Platt JL, et al. (2007) Monocyte-mediated T cell suppression by HIV-2 envelope proteins. Eur J Immunol 37: 3435-3444.

233. Cavaleiro R, Sousa AE, Loureiro A, Victorino RM (2000) Marked immunosuppressive effects of the HIV-2 envelope protein in spite of the lower HIV-2 pathogenicity. AIDS 14: 2679-2686.

234. Shi Y, Brandin E, Vincic E, Jansson M, Blaxhult A, et al. (2005) Evolution of human immunodeficiency virus type 2 coreceptor usage, autologous neutralization, envelope sequence and glycosylation. J Gen Virol 86: 3385-3396.

235. Elena SF, Wilke CO, Ofria C, Lenski RE (2007) Effects of population size and mutation rate on the evolution of mutational robustness. Evolution 61: 666-674.

236. Sankale JL, de la Tour RS, Renjifo B, Siby T, Mboup S, et al. (1995) Intrapatient variability of the human immunodeficiency virus type 2 envelope V3 loop. AIDS Res Hum Retroviruses 11: 617-623.

237. Ince WL, Zhang L, Jiang Q, Arrildt K, Su L, et al. (2009) Evolution of the HIV-1 env Gene in the Rag2-/-{gamma}C-/- Humanized Mouse Model. J Virol.

238. Williamson S, Perry SM, Bustamante CD, Orive ME, Stearns MN, et al. (2005) A statistical characterization of consistent patterns of human immunodeficiency virus evolution within infected patients. Mol Biol Evol 22: 456-468.

239. Barin F, M'Boup S, Denis F, Kanki P, Allan JS, et al. (1985) Serological evidence for virus related to simian T-lymphotropic retrovirus III in residents of west Africa. Lancet 2: 1387-1389.

240. Kong LI, Lee SW, Kappes JC, Parkin JS, Decker D, et al. (1988) West African HIV-2-related human retrovirus with attenuated cytopathicity. Science 240: 1525-1529.

241. Koblavi-Deme S, Kestens L, Hanson D, Otten RA, Borget MY, et al. (2004) Differences in HIV-2 plasma viral load and immune activation in HIV-1 and HIV-2 dually infected persons and those infected with HIV-2 only in Abidjan, Cote D'Ivoire. AIDS 18: 413-419.