From DEPARTMENT OF LABORATORY MEDICINE,
DIVISION OF CLINICAL MICROBIOLOGY,
Karolinska Institutet, Stockholm, Sweden

UNIFYING VIRAL EVOLUTION AND IMMUNOLOGICAL PATTERNS TO INVESTIGATE RISK OF HIV-1 DISEASE PROGRESSION

Melissa M Norström

Karolinska Institutet

Stockholm 2012

The cover picture shows a schematic representation of the research presented in this thesis. It depicts HIV and host immune responses during infection, where viral evolution and immunological patterns are unified to investigate disease progression. The picture was hand drawn by the author.

Published by Karolinska Institutet. Printed by Elanders Sweden AB.

ISBN 978-91-7457-987-1

“None of us knows what might happen even the next minute, yet still we go forward.
Because we trust. Because we have Faith.”
― Paulo Coelho

This work is dedicated to all people affected by HIV

ABSTRACT

After 30 years of research, the exact mechanisms underlying human immunodeficiency virus type 1 (HIV-1) pathogenesis and disease progression remain elusive. In the absence of highly active antiretroviral therapy, most HIV-infected individuals progress to AIDS within 10 years. The clinical course of HIV-1 infection is characterized by considerable variability in the rate of disease progression among patients with different genetic background. It has been shown that the rate of progression can depend on the expression of certain human leukocyte antigen (HLA) class I alleles that present antigen to the host immune system. The HLA-B*5701 allele is most strongly associated with slower progression. Underlying mechanisms are not fully understood but likely involve both immunological and virological dynamics. In this thesis, viral evolution and immunological patterns were studied in the context of HIV-1 risk of disease progression in HLA-B*5701 subjects and non-HLA-B*57 control subjects.

First, HIV-1 *in vivo* evolution and epitope-specific $CD8^+$ T cell responses were investigated in six untreated HLA-B*5701 patients monitored from early infection up to seven years post-infection. The subjects were classified as high-risk progressors (HRPs) or low-risk progressors (LRPs) based on viral load and baseline $CD4^+$ T cell counts. Interestingly, polyfunctional $CD8^+$ T cell responses were more robust in LRPs, who also showed significantly higher interleukin-2 production in early infection compared to HRPs. Additionally, HIV-1 *gag* p24 sequences exhibited more constrained mutational patterns with significantly lower diversity and intra-host evolutionary rates in LRPs than HRPs. Further in-depth analyses revealed that the difference in evolutionary rates was mainly due to significantly lower HIV-1 synonymous substitution [replication] rates in LRPs than HRPs. The viral quasispecies infecting LRPs was also characterized by a slower increase in synonymous divergence over time. This pattern did not correlate to differences in viral fitness, as measured by *in vitro* replication capacity, but a significant inverse correlation between baseline $CD4^+$ T cell counts and mean HIV-1 synonymous rate was found. The results indicate that HLA-linked immune responses in HLA-B*5701 subjects who maintain high $CD4^+$ T cell counts in early infection are more likely to control HIV-1 replication for an extended time.

To further assess these findings and evaluate them in the context of viral population dynamics, a new method was implemented to investigate the temporal structure of phylogenetic trees inferred from HIV-1 intra-host longitudinal samples. The analysis revealed that changes in viral effective population size (*Ne*) over time were more constrained in HLA-B*5701 subjects compared to non-HLA-B*57 controls, possibly due to the different evolutionary dynamics of archival viral strains observed in the two groups of patients.

Explaining the differences in risk of HIV-1 disease progression among HLA-B*5701 subjects, as well as between HLA-B*5701 and non-HLA-B*57 subjects, could have significant translational impact by providing specific correlates of protection that are essential for the successful development of a vaccine. Ultimately, the present work demonstrates that a thorough understanding of HIV-1 pathogenesis and disease progression requires a multidisciplinary approach unifying viral evolution and immunological patterns.

LIST OF PUBLICATIONS

This thesis is based on the following papers, which will be referred to in the text by their Roman numerals (I-IV):

I **Norström MM**, Buggert M, Tauriainen J, Hartogensis W, Prosperi MC, Wallet MA, Hecht FM, Salemi M, Karlsson AC. Combination of Immune and Viral Factors Distinguishes Low-Risk versus High-Risk HIV-1 Disease Progression in HLA-B*5701 Subjects. *J Virol. 2012; 86(18):9802-16.*

II **Norström MM**, Veras NM, Huang W, Prosperi MCF, Cook J, Hartogensis W, Hecht FM, Karlsson AC, Salemi M. Baseline CD4 Counts Determines HIV-1 Synonymous Rates in HLA-B*5701 Subjects with Different Risk of Disease Progression. *Submitted manuscript.*

III **Norström MM***, Prosperi MC*, Gray RR, Karlsson AC, Salemi M. PhyloTempo: A Set of R Scripts for Assessing and Visualizing Temporal Clustering in Genealogies Inferred from Serially Sampled Viral Sequences. *Evol Bioinform. 2012; 8:261-9.* *These authors equally contributed to the work.

IV **Norström MM**, Veras NM, Nolan DJ, Prosperi MCF, Hartogensis W, Hecht FM, Salemi M, Karlsson AC. Different HIV-1 Intra-Host Phylodynamic Patterns between HLA-B*5701 Study and Control Subjects. *Manuscript.*

PUBLICATIONS NOT INCLUDED IN THIS THESIS

Buggert M, **Norström MM**, Czarnecki C, Tupin E, Luo M, Gyllensten K, Sönnerborg A, Lundegaard C, Lund O, Nielsen M, Karlsson AC. Characterization of HIV-specific $CD4^+$ T Cell Responses Against Peptides Selected with Broad Population and Pathogen Coverage. *PLoS One. 2012; 7(7):e39874.*

Norström MM, Karlsson AC, Salemi M. Towards a New Paradigm Linking Virus Molecular Evolution and Pathogenesis: Experimental Design and Phylodynamic Inference. *New Microbiol. 2012; 35(2):101-11.*

Lindkvist A*, Edén A*, **Norström MM**, Gonzalez VD, Nilsson S, Svennerholm B, Karlsson AC, Sandberg JK, Sönnerborg A, Gisslén M. Reduction of the HIV-1 Reservoir in Resting $CD4^+$ T-lymphocytes by High Dosage Intravenous Immunoglobulin Treatment: a Proof-of-Concept Study. *AIDS Res Ther. 2009; 6:15.* *These authors equally contributed to the work.

Pérez CL, Larsen MV, Gustafsson R, **Norström MM**, Atlas A, Nixon DF, Nielsen M, Lund O, Karlsson AC. Broadly Immunogenic HLA Class I Supertype-Restricted Elite CTL Epitopes Recognized in a Diverse Population Infected with Different HIV-1 Subtypes. *J Immunol. 2008; 180(7):5092-100.*

TABLE OF CONTENTS

1 Introduction 1
1.1 HIV virology 1
1.1.1 The Origin and Spread of HIV 1
1.1.2 Structure and Genome 2
1.1.3 Replication Cycle 3
1.2 HIV Immunology and Pathogenesis 5
1.2.1 The Immune System in HIV Infection 5
1.2.2 Antigen Processing and Presentation 5
1.2.3 HIV-1-Specific T Cell Responses 6
1.2.4 Transmission 6
1.2.5 Course of HIV-1 Infection and Disease Progression 7
1.3 HIV Molecular Evolution 9
1.3.1 Genetic Variation 9
1.3.2 Selection Pressure 10
1.3.3 Phylodynamics of HIV-1 Intra-Host Evolution 11
Aims of Thesis 13
2 Materials and Methods 14
2.1 Study Design and Patient Material 14
2.2 Methodologies 15
2.2.1 Single Genome Sequencing 16
2.2.2 Flow Cytometry 16
2.2.3 Gag-Pro Mediated Replication Capacity Assay 16
2.2.4 Phylogenetic Signal and Recombination 17
2.2.5 Phylogeny Inference 17
2.2.6 Selection Analysis 18
2.2.7 Molecular Clock Analysis 19
2.2.8 HIV-1 Intra-host Demographic History 19
2.2.9 Temporal Clustering 20
2.2.10 Statistical Analysis 20
2.2.11 Ethical Considerations 21
3 Results and Discussion 22
3.1 HIV-1 Immune and Viral Factors in HLA-B*5701 Low-Risk and High-Risk Progressors 22
3.2 Developing Tools to Analyze Temporal Structure of Viral Genealogies 26
3.3 HIV-1 Intra-Host Phylodynamic Patterns and In-Depth Temporal Structure Analysis of Viral Genealogies in HLA-B*5701 Subjects and Non-HLA-B*57 Controls 26
4 Conclusions and Future Perspective 29
5 Acknowledgements 31
6 References 31

LIST OF ABBREVIATIONS

AIDS	Acquired immunodeficiency syndrome
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
APC	Antigen presenting cell
CCR5	C-C chemokine receptor type 5
CD4	Cluster of differentiation 4
CTL	Cytotoxic T lymphocyte
CXCR4	C-X-C motif receptor 4
DNA	Deoxyribonucleic acid
Env	Envelope
ER	Endoplasmic reticulum
Gag	Group-specific antigen
Gp	Glycoprotein
HIV-1	Human immunodeficiency virus type 1
HIV-2	Human immunodeficiency virus type 2
HLA	Human leukocyte antigen
IFN	Interferon
IL	Interleukin
IN	Integrase
LTR	Long terminal repeat
MHC	Major histocompatibility complex
mRNA	Messenger RNA
Nef	Negative factor
NKT	Natural killer T cells
PBMC	Peripheral blood mononuclear cell
PCR	Polymerase chain reaction
Pol	Polymerase
PR	Protease
Rev	Regulator of virion expression
RNA	Ribonucleic acid
RT	Reverse transcriptase
SIV	Simian immunodeficiency virus
TAP	Transporter associated with antigen processing

TCR	T cell receptor
Vif	Viral infectivity factor
Vpr	Viral protein R
Vpu	Viral protein U

1 INTRODUCTION

1.1 HIV VIROLOGY

1.1.1 The origin and spread of HIV

The origin of the human immunodeficiency virus (HIV) has been traced to simian immunodeficiency viruses (SIVs), which have been found in African apes and monkeys [1-3]. It is known that SIV naturally infect approximately 40 different species of Old World monkeys in sub-Saharan Africa [4]. Some of these SIVs have, through zoonotic transmission events, resulted in different HIV types and groups. Transmissions from West Central African chimpanzees (*Pan troglodytes troglodytes*) established HIV type 1 (HIV-1), while transmissions from sooty mangabeys (*Cercocebus atys atys*) established HIV type 2 (HIV-2) [1]. By using sequence data with known sampling times, phylogenetic analysis have shown that the time to the most recent common ancestor for HIV-1 dates back to 1910 for HIV-1 and to 1940 for HIV-2 [5-7]. Current estimates suggest that SIV has been present in African primates for more than 32,000 years [8] during which several zoonotic transmissions to humans may have occurred. However, it was a single (or limited) transmission about 100 years ago that eventually gave rise to the current AIDS pandemic, largely due to the worldwide spread of HIV-1 group M subtypes. The evolutionary and ecological forces driving the global dissemination of HIV-1 during the last three decades remain unclear, but may be related to social, historical and behavioural changes including decolonization, migration [9] and urbanization [7, 10], as well as rapid increase in infrastructure and human mobility [11].

Even though HIV-1 was introduced into the human population through several cross-species transmissions, it was only about 30 years ago that recognition and identification of the virus began. In 1981, opportunistic diseases such as *Pneumocystis carinii* pneumonia and Kaposi's sarcoma, combined with immune suppression was reported in young and previously healthy homosexual men in New York City and California [12, 13]. Additional opportunistic complications such as mycobacterial infections, toxoplasmosis, invasive fungal infections and non-Hodgkin's lymphoma were soon described. The disease was given the name acquired immunodeficiency syndrome (AIDS) [14]. However, the cause of the disease remained unknown until end of 1982 when a child who received blood transfusions died of AIDS-related infections, providing the first clear evidence that AIDS was caused by an infectious agent [15]. In 1983, the virus was isolated [16] that later was given the name HIV. Dr. Luc Montagnier and Dr. Francoise Barre-Sinoussi who isolated HIV were awarded the Noble Prize for their finding in 2008.

Today, almost 30 years after the discovery of the virus, there is still no cure or vaccine. HIV is one of the fastest evolving organisms known [17] and its ability to rapidly diversify allows the virus to evade the host immune system [18]. More than 60 million people have been infected with HIV-1 since 1981 and more than 20 million have died

from AIDS-related illnesses. Today, according to UNAIDS, the virus has spread to all continents and about 34.2 million people are infected, with the most affected part of the world being sub-Saharan Africa [19, 20].

1.1.2 Structure and Genome

HIV is a Lentivirus belonging to the *Retroviridae* family. Retroviruses are enveloped viruses containing two identical positive-sense single-stranded (ss) RNA molecules (9.2 kBp) that are non-covalently linked at the 5'-end. The virus is icosahedral with a diameter of approximately 100 nm and the bi-layered lipid envelope, derived from the host cell, contains viral trimeric glycoprotein gp41 covalently linked to the external trimeric gp120 (Figure 1). Inside the envelope, a protective cone-shaped nucleocapsid surrounds the genome and the viral enzymes reverse transcriptase (RT), protease (PR) and integrase (IN). The enzymes are required for specific replication events.

The viral genome is approximately 10,000 nucleotides and has three major structural genes: envelope (*env*), group-specific antigen (*gag*) and polymerase (*pol*). The *env* gene encodes the viral polyprotein gp160/gp140 that is cleaved into the transmembrane gp41 and the external gp120 (Figure 1). The *gag* gene encodes the polymerase precursor p55, which is processed by the viral protease into p24 (capsid), p17 (matrix), p7 (nucleocapsid) and p6. The *pol* gene encodes the viral enzymes RT, PR and IN. The *gag* and *pol* genes are produced as Gag or Gag-Pol precursor polyproteins that are cleaved by the viral PR into functional proteins. HIV-1 also has several regulatory (*tat* and *rev*) and accessory (*vif*, *vpr*, *vpu* and *nef*) genes, which are important for the viral life cycle.

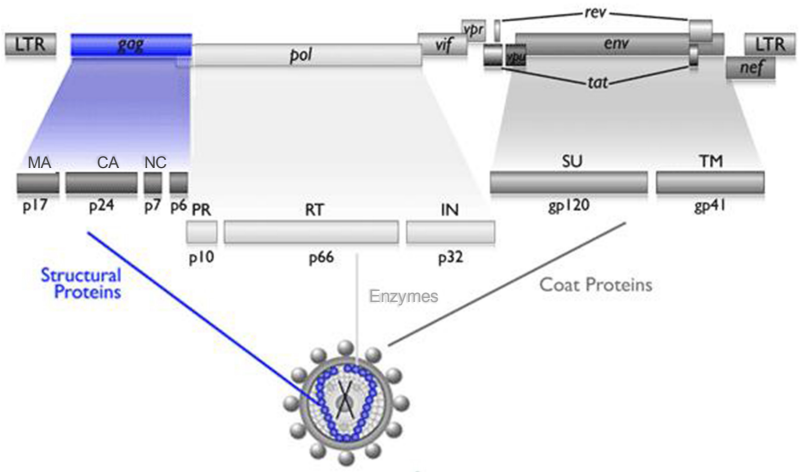

Figure 1. HIV-1 virion and genomic organization. Adapted from MWCHO.

1.1.3 Replication Cycle

The viral life cycle begins when the gp120 protein on the surface of the virus particle binds to the primary receptor, the CD4 molecule, on the target cell (Figure 2). CD4 molecules are found on $CD4^+$ T lymphocytes, macrophages, dendritic cells (DCs) and brain microglia [21, 22]. The virus also requires binding to the co-receptor CCR5 or CXR4 on the host cell for entry. After binding to the CD4 molecule, external gp120 undergoes a conformational change allowing transmembrane gp41 to insert its hydrophobic terminus into the host cell membrane bringing the virion closer to the host cell membrane. This enables fusion of the viral and cellular membranes resulting in the release of the viral nucleocapsid into the cytoplasm.

Once inside the cell, the capsid uncoats and genomic RNA strands, enzymes and additional molecules required for the initiation of transcription are released. The RT enzyme reverse transcribes the ssRNA genome into a complementary strand of DNA [23]. The template RNA is degraded by the ribonuclease H (RNase H) domain of the HIV polymerase and a complementary DNA strand is synthesized, creating a double stranded (ds) DNA of the genome. During the reverse transcription, long terminal repeats (LRTs) are added to both the 5'- and 3'-end of the DNA and are crucial for facilitating the subsequent transcription of the viral genome. Afterwards, IN form a pre-integration complex with the dsDNA and other viral and cellular proteins and enters the nucleus where the HIV-1 genome is inserted into the host genome [23]. After integration, the viral DNA is referred to as a provirus and remains permanently associated with the host genome [24].

The integrated proviral DNA is transcribed by host RNA polymerase II to produce novel viral genomes or viral messenger RNA (mRNA). The regulatory proteins are the first to be translated during this process. These regulatory proteins, amongst other things, facilitate the expression of the late structural viral proteins. The transactivator for transcription (Tat) protein forms complexes with several cellular proteins and enhance transcription of viral RNA by binding the trans-activating response region in the LTRs of the viral genome [25, 26]. The regulator of viral expression (Rev) protein increases the expression of the viral Gag, Env and Pol poly-proteins as it binds Rev-responsible elements. These are present in the viral RNA facilitating the export of unspliced viral mRNA from the nucleus [25, 26]. The negative regulatory factor (Nef) protein accelerates the endocytosis and subsequent degradation of CD4 and major histocompatibility complex (MHC) class I molecules allowing the cell to evade recognition by the immune system [25, 26].

Besides the regulatory proteins, three accessory proteins – viral infectivity factor (Vif), viral protein R (Vpr) and viral protein U (Vpu) – are expressed from the viral genome. Vif counteracts the antiretroviral effect of apolipoprotein B mRNA editing enzyme catalytic polypeptide-like 3G (APOBEC3G), which is a protein that inhibits retroviral infection by hypermutating the negative RNA strand during reverse transcription resulting in deamination of the proviral DNA [27, 28]. The Vpr constitutes a part of the pre-integration complex and the Vpu enhances the release of virions from the surface of the cell [27, 28].

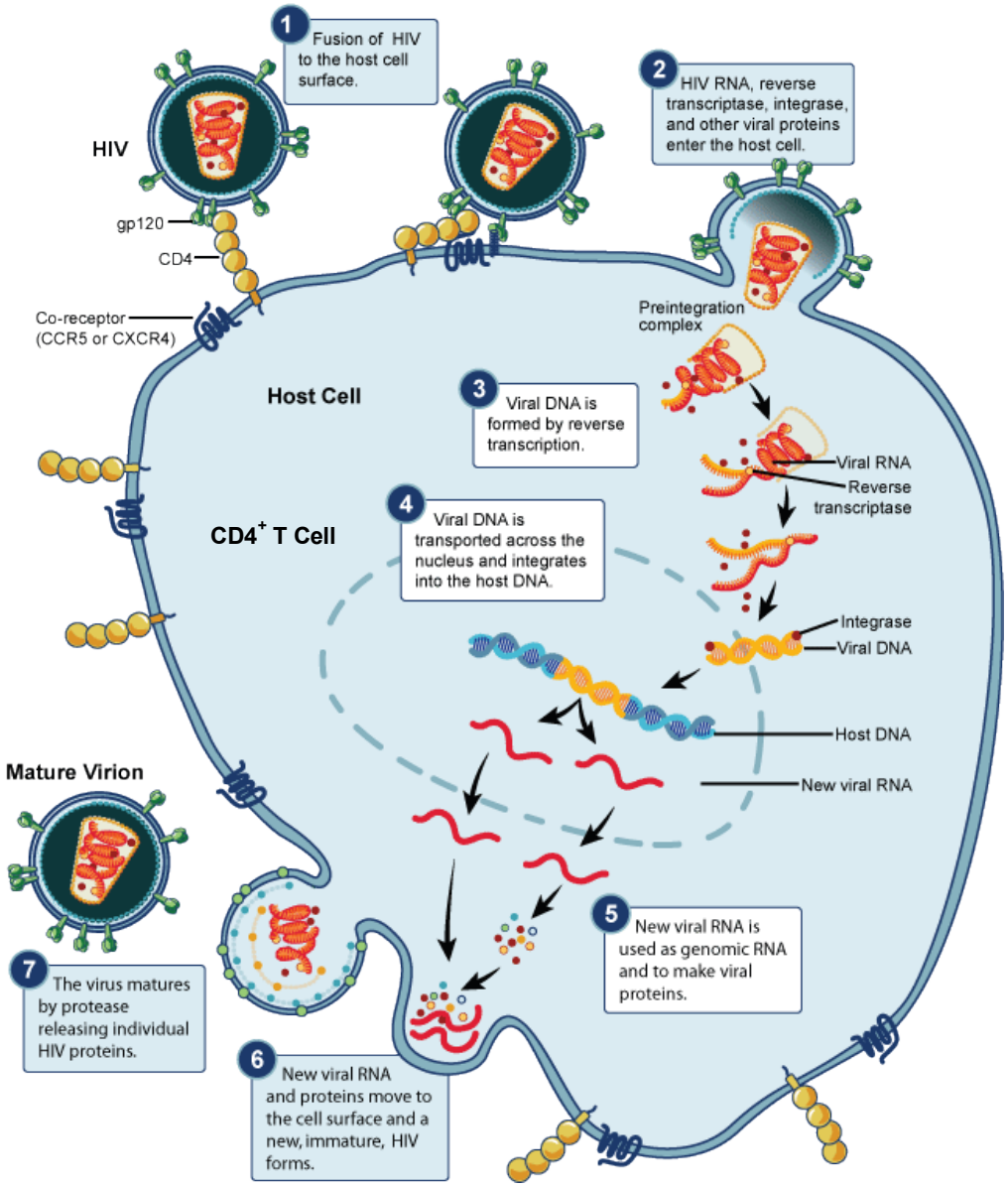

Figure 2. HIV replication cycle: 1) fusion of HIV to the host cell surface, 2) viral RNA, RT, IN and other viral proteins enter the $CD4^+$ T cell, 3) viral DNA is formed by reverse transcription, 4) viral DNA is transported across the nucleus and integrated into the host DNA, 5) new viral RNA is used as genomic RNA and to make viral proteins, 6) new viral RNA and proteins move to surface of the cell and a new virus forms, and 7) the virus matures by protease releasing individual HIV proteins. Adapted from NIAID.

The Env gp160 precursor protein is expressed and glycosylated in the endoplasmatic reticulum (ER) and subsequently cleaved by the cellular protease furin into gp120 and gp41 in the Golgi apparatus. They are transported to the cell surface, where trimers of transmembrane gp41 protein associate to trimers of the extracellular gp120 protein. Concurrently, two copies of the viral genome and Gag (p55) and Gag-Pol (p160) poly-proteins are assembled at the cell membrane. After budding of the immature virions, the viral Protease (PR), which is auto-cleaved from the Pol precursor protein, cleaves the Gag and Gag-Pol poly-proteins into: p17 (matrix protein), p24 (capsid protein), p7, p6 (nucleocapsid protein) and the viral RT, IN and PR enzymes [23]. This last step of the replication completes the HIV life cycle and the mature virion can infect other cells.

1.2 HIV IMMUNOLOGY AND PATHOGENESIS

1.2.1 The Immune System in HIV Infection

The innate immune system is the first line of defence against HIV-1 acting within a few hours after infection. It is often referred to a non-specific response as it recognizes conserved patterns on pathogens and damaged cells. This is followed by the adaptive immune response which needs weeks to fully mature as it is tailored towards the infecting pathogen. There are two arms of the adaptive immune system: the humoral and cellular response. The humoral immune response is built up by antibody-producing B cells, while the cellular immune response is composed of a variety of T cell populations. Both B and T cells originate from the same haematopoietic stem cells in the bone marrow, but B cells develop in the bone marrow while T cells mature in the thymus. Both of these cell types express antigen-specific receptors that, upon encounter with antigen, induce rapid cellular proliferation. This clonal expansion of antigen-specific effector cells is essential to enable control of the HIV-1 infection. T cells can be divided further into several subsets including $CD4^+$ T cells, $CD8^+$ T cells and NKT cells. The focus will, hereafter, be on the adaptive immune system and specifically $CD8^+$ T cells.

1.2.2 Antigen Processing and Presentation

After a $CD4^+$ T cell becomes infected with HIV-1, viral antigens are processed and presented by MHC class I on the surface of the infected cell. MHC class I molecules consist of two polypeptide chains: α and β_2-microglobulin (b_2-m). The heavy α chain consists of the three domains α1, α2 and α3. The two chains are linked non-covalently *via* interaction of b_2-m and the α3 domain. The heavy α chain is highly polymorphic and encoded by a HLA gene, while the light b_2-m subunit is not polymorphic and encoded by the b_2-m gene.

During antigen processing, viral proteins are degraded in the cytoplasm by the proteasome into 8-11 amino acid long epitopes. These antigenic peptides are transported from the cytoplasm into the endoplasmic reticulum (ER) *via* the transporter associated with antigen processing (TAP) proteins, where they are loaded into the MHC class I molecule. The MHC class I-peptide complex is then transported to the cell surface via the Golgi complex and presented to $CD8^+$ T cells. The α3 domain is plasma membrane spanning and interacts with the CD8 co-receptor of T cells, while the α1 and α2 domains fold to make up the peptide (antigen) binding groove.

The HLA class I alleles are the most polymorphic genes in the human genome and can be further divided into A, B and C alleles. The number of different alleles in each individual depends on whether the person is homozygotic or heterozygotic at each locus. Importantly, different HLA alleles have different binding specificities. Individuals with a heterozygote HLA composition may have an advantage in fighting most infections since a broader variety of antigens can be presented compared to homozygotes. The same epitope can be presented by different HLA alleles [29, 30], but

the impact on the pathogen may differ depending on the specific allele presenting the antigen [31].

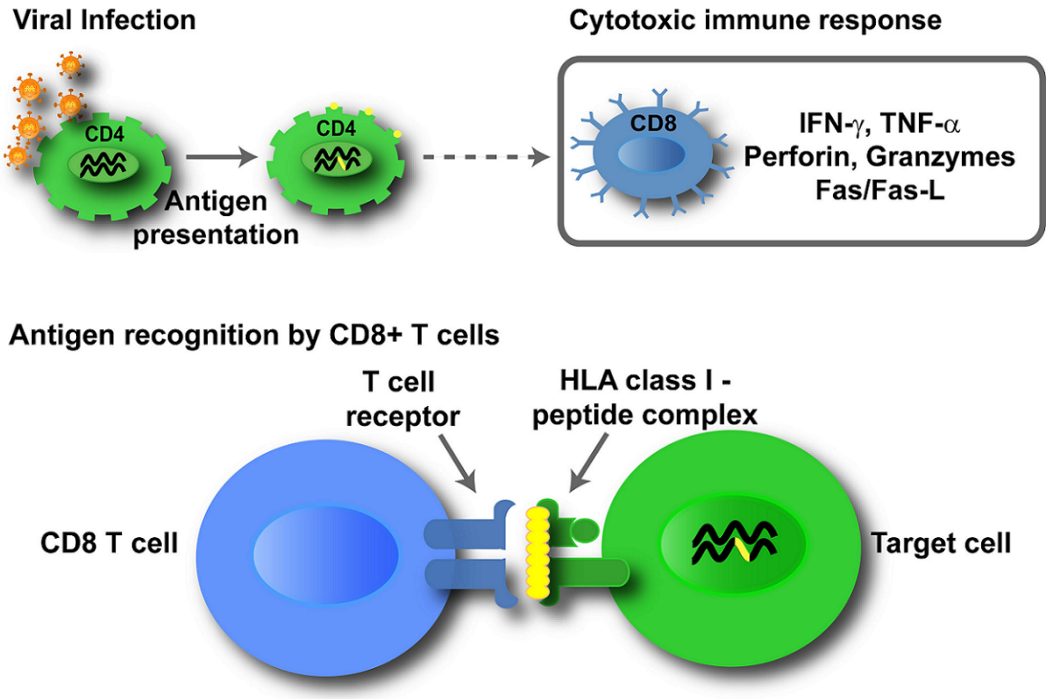

Figure 3. Illustration of antigen presentation in $CD4^+$ T cells, where the viral peptide:MHC class I complex is presented on the surface of the HIV-1-infected $CD4^+$ T cell. $CD8^+$ T cells that bind to the HLA-peptide complex with their T cell receptor can recognize the antigen and become activated. Courtesy of Annika C Karlsson.

1.2.3 HIV-1-Specific T Cell Responses

$CD8^+$ T cells express the CD8 glycoprotein at their surface and recognize their targets by binding to antigen associated with MHC class I. $CD8^+$ T cells are important in the control of the virus during infection [32]. Upon activation, $CD8^+$ T cells (also known as cytotoxic T cells or CTLs) can kill infected cells through the release of cytotoxic granules containing perforin and granzymes [33, 34] and produce antiviral cytokines (*e.g.* IFN-γ and IL-2) and chemokines (*e.g.* MIP-1β). Killing may also occur through the Fas-mediated pathway, where the Fas ligand (Fas-L) on the surface of $CD8^+$ T cells binds to the Fas receptor, a death receptor, on the target cell inducing target cell apoptosis. However, the perforin-mediated killing is believed to be of greater importance in viral infections [35]. $CD8^+$ T cells need three signals to become fully activated. First, the T cell receptor (TCR) will recognize and bind to the proper MHC:epitope complex. Second, co-stimulatory molecules on the T cells interact with molecules on the antigen-presenting cell. Thirdly, cytokines are needed in order to facilitate differentiation and proliferation of the antigen-specific $CD8^+$ T cells.

1.2.4 Transmission

There are three major routes of human-to-human HIV-1 transmission: i) sexual, ii) blood or blood product and iii) mother-to-child. The most common route of HIV

transmission is sexual (both heterosexual and homosexual) intercourse through the mucosa in the genitals, rectal and oral tracts. HIV blood transmission can occur through blood transfusion, organ transplantation, needle exchange among intravenous drug users, and needle accidents in health care and laboratory settings. Finally, mother-to-child transmission can occur during pregnancy, birth or breastfeeding. The risk of transmission depends on the transmission route, the presence of other infections and level of viral load exposure. Genital exposure constitutes a considerably lower transmission risk compared to rectal exposure and there is a higher risk of infection if the viral load during exposure is higher [36]. The viral load is dramatically lowered by antiretroviral treatment, which reduces risk of HIV-1 transmission [37-39].

It has been shown that HIV-1 diversity is low during primary infection [40-44] and most HIV-1 infections are probably established by one or a few virions [41, 45, 46]. It is still unclear if several virions are transmitted during HIV-1 infection. However, only one or a few virions grow out and transmission bottlenecks have been seen in both infection through intravenous drug use and mucosal transmission [47]. In the absence of treatment, the viral diversity increases during the course of the HIV-1 infection [48-50], but decreases later when progressing to AIDS [51].

1.2.5 Course of HIV-1 Infection and Disease Progression

The natural course of HIV-1 infection in untreated subjects includes three main stages: the acute infection, clinical latency and progression to AIDS (Figure 4). During the acute stage of infection, the viral load reaches a peak level and there is a massive destruction of $CD4^+$ T cells [52, 53]. Around 80% of the $CD4^+$ T cells are depleted in the gut-associated lymphoid tissue (GALT) [54, 55]. During the acute phase many individuals experience "flu-like" symptoms [56]. After a few weeks, when the adaptive immune response has matured, the peak viremia drops to a steady state level [57] resulting in partial recovery of the $CD4^+$ T cells. The subsequent chronic phase (clinical

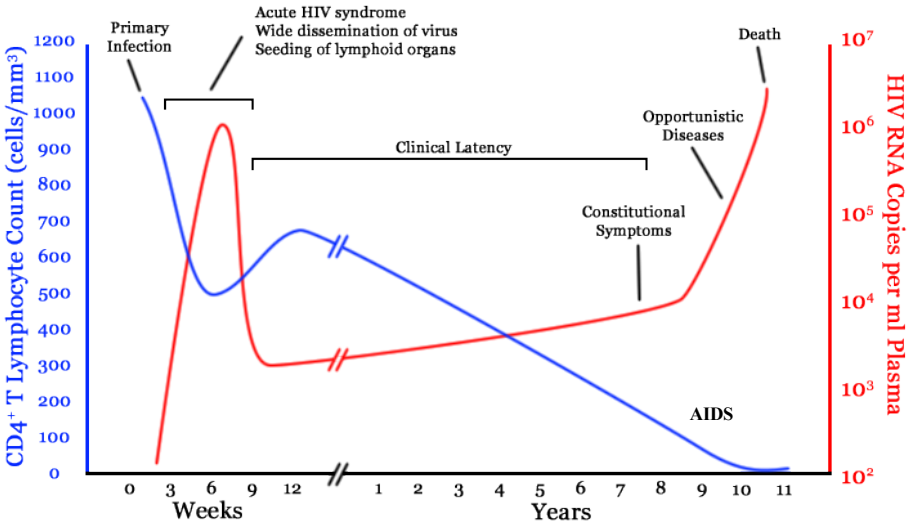

Figure 4. Clinical course of HIV-1 infection showing $CD4^+$ T cell count (blue) and viral load levels (red) throughout disease progression. Adapted from [58].

latency), characterized by slow but constant depletion of $CD4^+$ T cells, can last for years. During the chronic phase, the immune system is continuously activated due to on-going viral replication eventually leading to exhaustion and a general defect in immune responsiveness [59] and the potential onset of several opportunistic infections. According to current guidelines, the AIDS phase begins when the $CD4^+$ T cell count drops below 200 cells/mm^3.

A small percentage (≤ 1 %) of HIV-infected individuals spontaneously control viral replication in the absence of antiretroviral therapy. While there is no universally accepted definition for this rare group of HIV-infected individuals, they generally are called elite controllers (EC), elite suppressors (ES) or elite non-progressors (ENP). These terms cover individuals with viremia below the detection limit of standard viral load assays (50 or 75 copies/mL) for one year or more [60]. Some patients are also referred to as HIV controllers and are able to maintain viremia at ≤ 400 copies/ml for five years or more after infection [61]. The definition of these patients is virological and should not be confounded with another group of patients whose condition progresses slowly to AIDS. They are called long-term survivors (LTS), long-term asymptomatics (LTA) and long-term non-progressors (LTNP) and the definition is based on a $CD4^+$ T cell count greater than 500/mm^3 for several years without antiretroviral treatment. These patients were first described in the 1990s, but gradually the majority experienced a decrease in $CD4^+$ T cell counts with a significant fraction progressing to AIDS. In contrast, HIV controllers appear to have a considerably lower risk of progressing to AIDS [62].

There are several known host genetic factors associated with HIV-1 control. The HLA allele B*57, and to a lesser extent B*27, are associated with slower rate of HIV-1 disease progression [63]. In contrast, there are HLA alleles associated with an increased rate of progression, such as B*35, where patients progress to AIDS within 2-3 years [64]. Another host factor is the delta-32 deletion on the CCR5 gene (CCR5Δ32). Studies have shown that individuals infected with HIV-1, who have specific genetic mutations in one of their two copies of the CCR5 gene, progress to AIDS slower than individuals with two normal copies. There are also rare individuals with two mutant copies of the CCR5 gene who (in most cases) appear to be protected from HIV infection [65, 66]. Gene mutations in other HIV co-receptors, such as CXCR4, may also influence the rate of disease progression. Additionally, it has also been shown that plasma viral load and $CD4^+$ T cell counts at baseline can be prognostic markers of HIV-1 infection [67].

1.3 HIV MOLECULAR EVOLUTION

1.3.1 Genetic Variation

HIV-1 is characterized by high genetic variability as well as rapid evolution and diversification [68]. The rapid evolution of the viral genome is the result of several factors, including elevated error rate of the reverse transcriptase, recombination and rapid turnover of HIV-1 in infected individuals [69]. The HIV-1 evolutionary rate is estimated to be approximately one million times faster than the rate of cellular genes in higher organisms (Figure 5) [70]. Introduction of point mutations into the viral genome is mediated by the RT enzyme during the reverse transcription phase of the viral replication cycle, where the error frequency of RT has been estimated to be as high as 3.4×10^{-5} substitutions per site per replication cycle [71]. Considering that the size of the HIV-1 genome is approximately 10,000 bases, such an error rate results in the introduction of one nucleotide substitution every second to third newly synthesized viral genome. Moreover, highly variable regions, e.g. the hypervariable domains in the surface glycoprotein gp120, can display significant length polymorphism due to the frequent occurrence of RT-mediated insertions and deletions (indels). HIV-1 genetic diversity is further escalated by recombination, which is the result of strand switching during reverse transcription in superinfected cells. It has been estimated that recombination events may occur two to three times per replication cycle [72] and can significantly impact the immune systems' ability to control the infection, the emergence of drug-resistant viral variants and allow viruses to survive genomic damage [73].

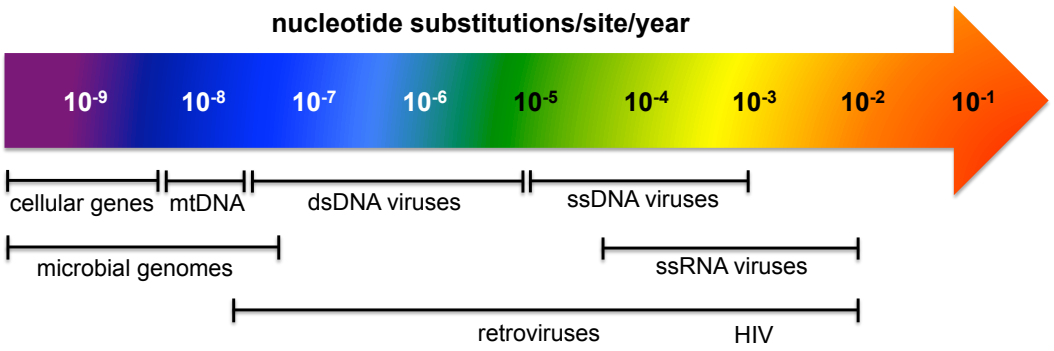

Figure 5. Evolutionary rates of different organisms. Adapted from [74].

HIV-1 fast replication rate and high number of viral particles produced per unit time are two other important factors contributing to the rapid evolution of the virus. The time from the release of a virion until it infects a new cell and eventually releases a progeny of its own (generation time) has been estimated to be approximately 2-3 days [75, 76] and every day ~10^9 new viral particles are produced [75, 77]. As a consequence of all the factors mentioned above, HIV-1 positive patients are usually infected with a highly heterogeneous viral population consisting of a pool of genetically distinct yet related viruses called quasispecies [78, 79]. The genetic variability of the HIV-1 quasispecies, that can reach up to 10% nucleotide diversity within an infected subject, provides the viral population with the ability to adapt rapidly to changes in its environment, and is the main challenge for the development of a vaccine or a treatment able to eradicate the infection.

1.3.2 Selection Pressure

Changes in the environment impose selective pressure on evolving populations resulting in the fixation of genetic variants with genes best adapted to the new milieu they were selected for. Adaptation at the molecular level can be studied by analyzing synonymous to nonsynonymous substitutions. Synonymous substitutions, also called silent substitutions, do not alter the encoded amino acid and occur mostly at the third position of a codon. Nonsynonymous substitutions, on the other hand, result in a change of the encoded amino acid. Synonymous substitutions are usually neutral, or nearly-neutral, and are fixed in by random genetic drift. According to the neutral theory of molecular evolution, in the absence of positive selection, the evolutionary (nucleotide substitution or fixation) rate is equal to the mutation rate. Therefore, the synonymous substitution rate of any retrovirus is expected to be equal to the RT error rate, which is in turn proportional to the viral replication rate [80]. On the other hand, the rate of nonsynonymous substitutions may depend on selective pressure that can increase (positive selection) or decrease (negative selection) the fixation probability of specific amino acid changes, in which case nonsynonymous substitution rates can be used as a measure of the adaptation rate.

In practice, the ratio of nonsynonymous and synonymous substitutions (dN/dS) is often employed to investigate natural selection and random genetic drift at the molecular level. A ratio around one indicates that neutral (dS) and adaptive (dN) mutations occur at the same rate, which is only possible in the absence of selection. On the other hand, $dN/dS<1$ indicates that the synonymous substitution rate is greater than the nonsynonymous one, which is expected in the presence of negative (purifying) selection removing genetic variants with new amino acid changes. The majority of amino acid replacements are usually under negative selection because of protein structural constraints necessary to maintain the biological function. A $dN/dS>1$ is evidence of positive (diversifying) selection occurring when amino acid changes increase fitness.

HIV-1 intra-host evolution is driven by the dynamic interplay between viral evolution and host immune system, which can result in the selection of viral variants with reduced sensitivity to CTLs and neutralizing antibodies. It has been shown that emergence of viral escape mutants, in case of CTLs, is associated with disease progression [81, 82]. CTL escape mutants can emerge shortly after acute infection and become the dominant viral strains in the infected individual [83]. Rapid emergence of CTL escape variants indicates their pre-existence in the viral population and points to the dominant role of CTLs as a selective force in the infection [81]. It is also important to mention that the emergence of CTL escape mutations is also correlated with loss of viral replication fitness [84]. Antiretroviral treatment also exerts strong selective pressure on the infecting quasispecies that can lead to the emergence and fixation of low-frequency viral variants carrying drug resistance mutations.

1.3.3 Phylodynamics of HIV-1 Intra-Host Evolution

Phylodynamics encompass both phylogeny inference and the interaction between evolutionary (i.e. mutation, genetic drift, selection) and ecological (population dynamics and environmental stochasticity) processes, which shape the spatiotemporal and phylogenetic patterns of infectious disease dynamics, both at the intra- and inter-host level [85, 86]. Phylodynamics allow the study of viral evolution by investigating the topology and branch lengths of genealogies to infer the viral evolutionary and population dynamic patterns. Several studies have shown that staircase topology is typical of HIV-1 intra-host evolution [51, 87], which suggests viral quasispecies undergo continual immune-driven selection through sequential population bottlenecks [88] (Figure 6). Usually, HIV-1 intra-host genealogies of longitudinally sampled sequences display strong temporal structure, where sequences from the same sampling time tend to cluster together and are the direct ancestors of sequences from the following time point [89]. However, the degree of temporal structure can vary among genealogies inferred from different data sets. Unfortunately, no previous study has investigated the temporal structure of HIV-1 intra-host genealogies in-depth, because topological differences among genealogies are difficult to quantify. The *Temporal Clustering* (TC) statistic has recently been developed to provide a quantitative measure of the degree of topological 'temporal clustering' in a serially sampled genealogy [89].

HIV-1 divergence and diversity have been shown to follow distinct patterns during the infection [90]. Divergence of the virus describes its evolution from a founder strain, while diversity is a measure of genetic variation within the virus population at a specific time point. When an individual becomes infected with HIV-1, a relatively homogenous population of the virus is harbored because transmission is usually associated with a significant population bottleneck [47]. The viral diversity in *gag* and *env* genes has been estimated to be less than 1% during transmission of HIV-1 [43, 44, 91]. During the early period of the chronic phase, viral diversity and divergence have been shown to linearly increase with similar rates. The intermediate period is characterized by stabilization of viral diversity, while the divergence from the founder strain continually increases at the same pace.

The chronic phase of HIV-1 evolution is characterized by a progressive increase in both viral divergence and diversity. It has been suggested that this phase is dominated by continuous pressure from the host immune system resulting in rapid turnover of the infecting quasispecies. The effect of intense immune-mediated positive selection on HIV-1 within a patient is reflected by the temporal structure of the viral genealogies inferred from longitudinal samples, where a main lineage usually propagates successfully along the trunk (backbone) of the genealogy while other lineages become extinct (Figure 6). In other words, the phylogenetic trees display a topological signature consistent with the occurrence of sequential population bottlenecks that would be the result of an evolutionary process driven mainly by continual immune selection rather than random genetic drift.

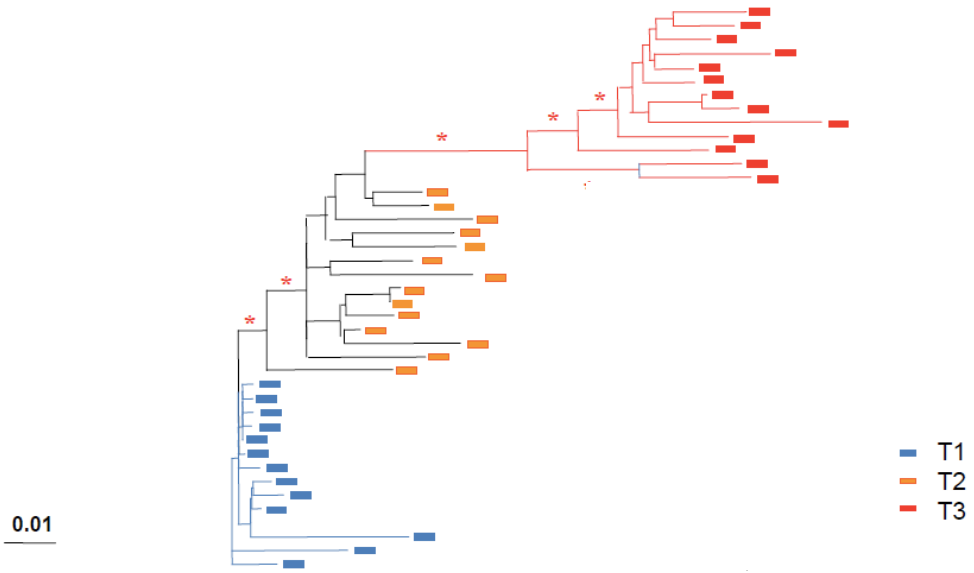

Figure 6. Evolution of *env* quasispecies in plasma. This is an example of a phylogenetic tree with staircase evolution and perfect temporal structure. The tree was inferred by ML using longitudinal plasma samples of an HIV positive pediatric patient infected by mother-to-child transmission. Different colors of tip labels represent sampling times of the viral strains according to the legend in the figure, where T1=0.3, T2=1.2 and T3=7.0 years post-infection. Asterisks along the backbone indicate major population bottlenecks supported by high bootstrap values (>85%). Courtesy of Marco Salemi.

The last phase of the infection, when the immune system collapses and progression to AIDS begins, involves stabilization of viral divergence and decline of viral diversity. This observation has been explained as a consequence of $CD4^+$ T cells depletion, which results in less effective selection pressure on the virus, as well as significant decrease in target cells capable of sustaining viral replication. In support of such a hypothesis, it has been shown that late infection is, in fact, characterized, by a decrease in HIV-1 evolutionary rate [90].

AIMS OF THESIS

The main objective of the present work was to unify viral evolution and immunological patterns to investigate risk of HIV-1 disease progression. Four specific aims were developed:

Paper I To investigate HIV-1 *in vivo* evolution and functional profiles of epitope-specific $CD8^+$ T cell responses in untreated HLA-B*5701 subjects with different risk of progression.

Paper II To investigate how risk of HIV-1 disease progression in HLA-B*5701 subjects correlates to HIV-1 intra-host synonymous and nonsynonymous substitution rates.

Paper III To implement a new method to investigate differences in HIV-1 population dynamics through the analysis of the temporal structure of viral genealogies.

Paper IV To characterize HIV-1 population dynamics in subjects with different genetic background, by investigating temporal structure and intra-host phylodynamic patterns in HLA-B*5701 subjects and non-HLA-B*57 controls.

2 MATERIALS AND METHODS

2.1 STUDY DESIGN AND PATIENT MATERIAL

The interplay between HIV-1 evolution and the host immune repertoire determines the course of disease [92]. Therefore, it may be insufficient to focus separately on immunological or evolutionary patterns for correlates of protection. A rigorous study design was developed to take full advantage of the phylodynamic framework (outlined in Figure 7) by following specific guidelines described in Norstrom *et al.* (2012) [74]. Plasma samples were needed to obtain data for the virological studies on the HIV-1 population in circulation at a given time point. Additionally, peripheral blood mononuclear cell (PBMC) samples were necessary to obtain immunological data. Plasma samples (stored at -80°C) and cryopreserved PBMC samples were collected longitudinally from subjects in the San Francisco-based HIV-1-infected cohort OPTIONS at the Positive Health Program, University of California [93]. The OPTIONS cohort contains samples from more than 600 subjects followed from primary HIV-1 infection. For each subject in the cohort, the time since infection was estimated as the midpoint between reported negative and positive tests based on data from serologic tests, HIV-1 RNA testing, and prior antibody testing history. In addition, $CD4^+$ T cell counts (cells/mm^3) and viral load (copies/mL) measurements were performed regularly during the course of infection for the patients included in the cohort.

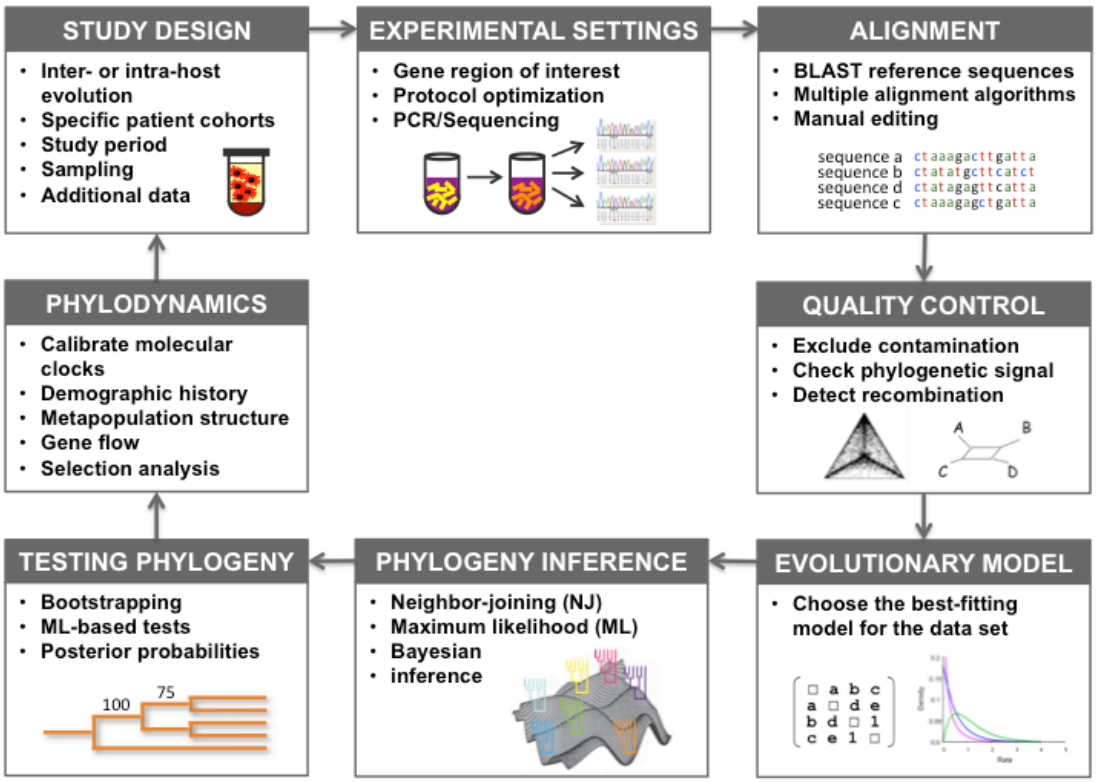

Figure 7. Flow-chart representing the major steps in phylodynamic inference linking experimental design and data analysis. Adapted from [74].

On-going viral intra-host evolution is difficult to characterize in samples with low viral copy numbers and limited genetic heterogeneity [94]. All selected subjects had to have detectable viral load during the study period in order to increase the chance to obtain sufficient phylogenetic signal for a robust evolutionary analysis. There were also specific requirements for an in-depth phylodynamic analysis. First, longitudinal samples were necessary for the calibration of the molecular clock and, in general, to infer viral evolutionary and population dynamic patterns. Second, time intervals between samples had to be optimal to make sure that the quasispecies was a measurably evolving population (MEP), i.e. that a statistically significant number of mutations could be detected between sequences obtained at different time points [95]. If the samples were too distant, important evolutionary events may be missed. It has been shown that a complete viral population turnover within an HIV-1 infected patient requires a time interval of 6-22 months [96]. Therefore, samples from at least three different time points were selected for each subject, time intervals were optimized, and special care was taken to match the data between HLA-B*5701 subjects and non-HLA-B*5701 controls. Unfortunately, some limitations were unavoidable due to the actual samples available in the OPTIONS cohort. A brief description of the patient material included in the different studies is given below. Full details can be found in the Materials and Methods of **Paper I-IV**.

In **Paper I**, six HIV-1 infected individuals carrying the B*5701 allele were selected. All subjects were treatment naïve excluding one who received antiretroviral treatment for 14 months; samples from that period were excluded in the study. Longitudinal plasma and PBMC samples were obtained from early infection up to seven years. For virological analysis, plasma samples were selected from three to six time points for each patient. For immunological analysis, PBMC samples were selected from three time points during the course of the infection for each subject.

In **Paper II**, high-resolution phylogenetic analyses were performed on sequences obtained from the plasma samples collected from the HLA-B*5701 subjects described in **Paper I**. Replication capacity experiments were also performed on additional plasma samples collected from one to four time points for each subject.

In **Paper III**, a method was developed to investigate temporal structure and tested on the HLA-B*5701 data sets (**Paper I** and **Paper II**), as well as SIV data sets downloaded from the Genbank database.

In **Paper IV**, six additional HIV-1 infected individual were selected as a control group. All subjects were treatment naïve and did not carry the B*57 allele. Longitudinal plasma samples were selected from early infection up to six years.

2.2 METHODOLOGIES

A wide range of different virological, immunological and mathematical methods was applied. The sections below provide a brief overview of the main methods in **Paper I-IV**, as well as the advantages of using them in the different studies. More detailed

information about the specific methods can be found in Materials and Methods in the respective papers (Appendix I-IV).

2.2.1 Single Genome Sequencing

Plasma samples from the HLA-B*5701 subjects had, in general, low viral load and only 1 mL plasma was available for most time points. In order to obtain sufficient template recovery, substantial effort was invested to develop sensitive and robust RNA extraction, cDNA synthesis and PCR amplification methods [97]. Characterization of the genetic heterogeneity of HIV-1 intra-host quasispecies is usually achieved by obtaining multiple sequences through PCR/cloning or single genome sequencing (SGS) [98]. SGS permits individual cDNA molecules, derived from defined regions of the genome, to be amplified by PCR and sequenced. This significantly reduces the probability of re-sampling (particularly likely in samples with low viral load), as well as the occurrence of PCR-mediated recombination [97, 99, 100]. SGS protocols were developed to obtain the HIV-1 *gag* p24 sequences required for the studies in **Paper I-IV**. Briefly, sequences were amplified from viral cDNA by limiting-dilution digital nested PCR. To obtain PCR products derived from single cDNA molecules, the cDNA was diluted until approximately 30% of the PCR reactions yielded DNA product [99]. The Gag p24-region was focused on since several HLA-B*5701-restricted epitopes (ISW9, KF11, TW10 and QW9) are located in this region of the HIV-1 genome.

2.2.2 Flow Cytometry

Flow cytometry is a powerful technique for the analysis of multiple parameters of individual cells within heterogeneous populations. It can be used to measure the production of several effector molecules simultaneously in T cell populations stimulated with autologous peptides. In **Paper I**, functional profiles of HIV-1-specific $CD8^+$ T cell responses were identified in six HLA-B*5701 subjects using a flow cytometry assay. Longitudinal PBMC samples were analyzed on a standardized 8-color CantoII (BD Biosciences) to identify epitope-specific $CD8^+$ T cell responses towards the Gag p24-region. Using flow cytometry we were able to distinguish different T lymphocyte populations ($CD3^+$, $CD4^+$ and $CD8^+$ T cells) at a single cell level and to simultaneously identify production of several effector molecules (IFN-γ, MIP-1β, IL-2 and perforin). All analyses were conducted after removal of non-viable cells, stained with Vivid. For each HLA-B*5701 subject, PBMCs from three different time points were analyzed. For more detailed information see the Materials and Methods in **Paper I**.

2.2.3 Gag-Pro Mediated Replication Capacity Assay

A novel assay, developed by Monogram Biosciences, was used to measure the viral replication capacity (RC) in the HIV-1 Gag-Pro region [101]. In **Paper III**, plasma samples were selected from all HLA-B*5701 subjects. Briefly, *gag* and protease sequences were amplified by RT-PCR and transferred into a resistance test vector

(RTV) containing a luciferase reporter gene. Transfections of HEK293 cells with patient-derived *gag-pro* RTVs and an amphotropic murine leukemia virus envelope expression vector were performed to generate pseudovirus stocks for infection of HEK293 cells. Gag-Pro mediated RC was determined by measuring the viral infectivity (luciferase activity) of patient-derived pseudoviruses relative to NL3-4, the reference control.

2.2.4 Phylogenetic Signal and Recombination

HIV-1 *gag* p24 is usually one of the most conserved regions of the viral genome. Hence, the first step in the analysis was to investigate whether aligned data sets from the study subjects displayed sufficient phylogenetic signal to allow reliable inferences. In particular, it was necessary to assess that viral sequences had sufficient genetic variability, within and between longitudinal samples, to be considered a MEP. Several methods have been developed to measure the phylogenetic signal in nucleotide and amino acid sequence data sets. Likelihood mapping [102], transition/transversions versus divergence plots [103], and the Xia test for saturation [102, 104] are often employed to assess the reliability of an alignment for phylogeny inference [105, 106]. In **Paper I** and **Paper IV**, the phylogenetic signal in each data set was investigated by likelihood mapping method implemented in the program TREEPUZZLE [107]. Detailed information can be found in Materials and Methods of **Paper I**.

It is also important to keep in mind that recombination violates the basic assumption of phylogeny inference (ancestry from a common ancestor). Using algorithms that do not explicitly model recombination (*e.g.* BEAST) can bias molecular clock and coalescent estimates [108-110]. Recombinant strains should, therefore, be excluded and analyzed separately or with more complex coalescent models [111, 112]. In **Paper I** and **Paper IV**, the presence of potential recombinant sequences was investigated with the PHI test based algorithm [113] and calculations were performed with the SplitsTree package version 4.8 [114]. It has been shown that the PHI test is the most robust method for detection of recombinants within intra-host sequences that are closely related and display lower diversity [115].

2.2.5 Phylogeny Inference

HIV-1 intra-host genealogies were inferred with several tree-building algorithms. Neighbor-Joining (NJ) trees were constructed to exclude contamination and confirm that all strains were subtype B (**Paper I** and **Paper II**). NJ is a fast, bottom-up clustering algorithm that can be used to quickly analyze large data sets. It is based on the computation of pair-wise distances, with an explicit evolutionary (nucleotide substitution) model, which are in turn employed to infer the phylogenetic tree [116]. It has been shown that NJ trees are accurate as long as the input distance matrix is correct and "nearly additive", i.e. if each entry in the distance matrix differs from the true distance by less than half of the shortest branch length in the tree [117]. Since these properties are seldom satisfied in real data sets, phylogenies were also obtained by more sophisticated character-based methods using the maximum likelihood (ML)

optimality criterion [117, 118]. ML is statistically sound, makes use of all sequence information, but is slower than NJ and must rely on heuristic algorithms (which do not guarantee to find the true ML tree) for data sets including more than 8-12 sequences. In order to assess the robustness and statistical significance of the inferred tree topologies, ML and NJ trees obtained from each alignment were compared for topological consistency. Bootstrapping (500 replicates), as well as the approximate likelihood ratio test (aLRT) were used to assess the reliability of specific monophyletic clades. In particular, the Shimodaira-Hasegawa-like aLRT compares the likelihoods of the best and the second best alternative arrangements around the branch of interest. For both NJ and ML trees, the best-fitting evolutionary model was chosen with the hierarchical likelihood ratio test described by Swofford and Sullivan [117]. ML and NJ calculations were performed with PAUP* 4b10, written by David L. Swofford, and MEGA 5.0 [119], respectively. In **Paper I, Paper II** and **Paper IV**, HIV-1 genealogies were also inferred by Bayesian inference, which generates a posterior distribution for a specific parameter (such as a phylogenetic tree and a model of evolution), based on the prior distribution for that parameter and the likelihood of the data (i.e. the multiple sequence alignment). Bayesian genealogies were inferred with the program BEAST [120, 121], which implements a Markov chain Monte Carlo (MCMC) algorithm [122] to sample trees from a coalescent-based prior. In brief, after the posterior tree distribution is obtained with the MCMC, a maximum clade credibility tree (also called MAP or MCC tree), i.e. the tree with the greatest posterior probability averaged over all branch lengths and substitution parameter values, is chosen from the distribution. The posterior probability of a specific clade in the tree can also be calculated easily from the posterior distribution. MAP trees were selected with the TreeAnnotator program distributed within the BEAST package. Bayesian inference has several advantages over classic NJ and ML methods. First of all, it explicitly models for uncertainty in the phylogenetic reconstruction, since a posterior distribution of possible trees, rather than a single tree (as in clustering on ML-based algorithms), is obtained. In addition, the BEAST software allows for the calibration of clock-like trees (see below) [123, 124] and can estimate absolute evolutionary rates when sequences collected longitudinally with known sampling times are available. Full details on the settings used in each analysis can be found in the Materials and Methods of the respective papers.

2.2.6 Selection Analysis

ML-based methods were employed to investigate selection at the molecular level. In particular, for each patient-specific data set, it was interesting to identify potential HIV-1 *gag* p24 sites under positive selection, as well as to compare selective pressure among groups of patients with different risk of disease progression. In **Paper I**, the presence of sites under positive selection was investigated using different ML codon substitution models implemented in the codeml program of the PAML 4.2 software package [125]. Initial codon-based branch lengths and nonsynonymous/synonymous (dN/dS) ratio were estimated for each ML tree with the M0 (one ratio) model, which assumes the same dN/dS along each branch of the tree [126, 127]. Positive selection analysis was then performed using the tree with codon-based estimated branch lengths. Three *sites* models were compared [128, 129]: M7, assuming a beta distribution of substitution

rates across sites; M8, assuming a beta distribution and dN/dS > 1 (positive selection) across sites; M8a, assuming a beta distribution and dN/dS = 1 (neutrality). The M7 (null hypothesis) and M8 model were compared with a chi-square test with 2 degrees of freedom (M7 rejected when LR > 5.99). The M8a (null hypothesis) and M8 model were compared using a 50:50 mixture of point mass 0 and a chi-square test with a critical value of 2.71 at the 5% level (see PAML documentation for more details). Specific amino acid changes along the internal branches of the tree were inferred by maximum likelihood reconstruction of ancestral sequences using PAML.

2.2.7 Molecular Clock Analysis

In **Paper** I, **Paper II** and **Paper IV**, for each of HIV-1 *gag* p24 patient-specific data set molecular clock analysis was performed using the MCMC approach implemented in BEAST version 1.7 [121]. The analyses were performed with the same nucleotide substitution model selected for the phylogeny inference. Viral evolutionary rates were estimated by enforcing either a strict or a relaxed molecular clock (assuming, across the phylogeny, constant evolutionary rate or branch-specific evolutionary rates drawn from a lognormal prior distribution, respectively) and different population size coalescent priors (constant, exponential and non-parametric Bayesian skyline plot with four bin categories) [130]. For each analysis two independent MCMC were run, which were combined with the LogCombiner program in the BEAST package. The effective sample size (ESS) value for each parameter was > 500 indicating sufficient mixing of the Markov chain. The molecular clock hypothesis was then tested by comparing the marginal likelihood of the strict and the relaxed clock model. Estimated marginal likelihoods were used to compute the Bayes Factor (BF) where evidence against the null hypothesis (strict clock) is assessed in the following way: $2 < BF < 6$ indicates positive evidence against the null hypothesis, $6 < BF < 10$ indicates strong evidence against the null hypothesis, $BF > 10$ indicates very strong evidence against the null hypothesis. In **Paper I**, marginal likelihoods were estimated by bootstrapping *via* importance sampling and in **Paper IV** they were estimated with the newly developed stepping stone model [131], which is more reliable. Additionally, in **Paper I**, ML estimates of the coefficient of variation (CoV) under the relaxed clock model were obtained to assess the overall degree of rate heterogeneity across the genealogies [123]. In **Paper II** and **Paper IV**, for each patient data set, absolute rates of synonymous and nonsynonymous substitutions were estimated using 200 randomly chosen trees from the posterior distribution obtained with BEAST and implementing the method described by Lemey et al. (2007) [132].

2.2.8 HIV-1 Intra-host Demographic History

The demographic history of a population can be inferred from the genealogical relationships of sampled individuals by applying coalescent theory [133]. A genealogy reconstructed from randomly sampled HIV sequences, for example, contains information about population-level processes such as change in population size and growth rate [134]. In **Paper IV**, three different demographic models or each data set were investigated for the HIV-1 intra-host quasispecies by enforcing the best fitting

molecular clock model: constant population size, exponential population growth, and Bayesian skyline plot (BSP). Both parametric (constant or exponential model) and non-parametric BSP estimates of demographic history were performed and compared by BFs as described in the previous section.

2.2.9 Temporal Clustering

Evolutionary trees estimated from serially sampled sequence data are shaped by a complex interaction of demographic factors and selective pressure. Those obtained from genes under strong and continual positive selection are reported to exhibit a 'ladder-like' shape, characterized by (i) phylogenetic asymmetry and (ii) a tendency for sequences sampled at similar times to cluster together [88]. The temporal clustering (TC) statistic, based on parsimony estimates, was recently developed to quantify 'ladder-likeness' of a phylogenetic tree topology by taking into account the sampling time of the tips [89]. However, an improved TC statistics based on Maximum Likelihood (ML) estimates of ancestral characters was implemented in the R language by developing a set of R scripts, called PhyloTempo. Several additional topological measures were also integrated in a user-friendly graphical interface. The TC statistic assesses the temporal structure of a phylogenetic tree by assigning a discrete character corresponding to the sampling time (T_1, T_2, … T_n) to each tip and using ML ancestral trait mapping [135] to calculate the number of time transitions (from T_i to T_j with $i<j$) in the genealogy. If there are n different states at the tips of a phylogeny, the minimum number of ancestral state transitions observed across the phylogeny would be n-1. A greater number of ancestral state transitions indicate a deviation from a perfect temporal structure. For detailed information about the temporal clustering statistics see Materials and Methods in **Paper III**. In **Paper IV**, the TC calculations for each HIV-1 genealogy were performed with the PhyloTempo software running under the R package [135]. Additionally, by recording the number of ancestral state transitions (NAST) in the genealogy for each pair of discrete characters (i.e., T_i and T_j with $i<j$), it is possible to assess whether the observed NAST is significantly greater ($p<0.05$) than the one expected in the null distribution of 1000 trees with randomly shuffled tip characters. NAST significantly exceeding the null expectation indicates the emergence (or re-emergence) of archival viral strains.

2.2.10 Statistical Analysis

In **Paper I**, experimental variables between two groups of individuals were analyzed using unpaired t-test, Student's t-test, Mann-Whitney U-test and Wilcoxon matched-pairs rank test. One-way ANOVA with *post hoc* Dunn's multiple comparison tests (non-parametric) was used to analyze three or more groups. Correlations were assessed using non-parametric Spearman rank tests. All pie charts were analyzed by permutation tests using the data analysis program SPICE version 5.2 (11). In **Paper II**, experimental variables between groups of individuals were compared using the Mann-Whitney U-test. Mean nonsynonymous and synonymous rates for different set of branches in the phylogenetic trees (all, internal, backbone and external) were compared to the corresponding clinical parameters of each patient (baseline CD4 count, baseline VL,

CD4 slope, VL slope and baseline T cell activation) using Pearson's linear correlation to calculate the associated t-values and assess significance. All *p*-values obtained from applying any test statistic multiple times were adjusted with the Bonferroni correction. In **Paper I** and **Paper IV**, slopes of $CD4^+$ T cell counts and viral load (VL) were obtained from least squares regression of log-transformed CD4 counts and VL over time (years). Model coefficients were back transformed and converted from proportions to percentage effect by subtracting one and multiplying by 100 to obtain individual estimates of percent change over time. Full details on the statistical test used in each analysis can be found in the Materials and Methods of the respective papers.

2.2.11 Ethical Considerations

The University of California, San Francisco (UCSF), Committee on Human Research and the Regional Ethical Council in Stockholm, Sweden (2008/1099-31), approved the studies in **Paper I-IV**. For the studies included in **Paper IV**, ethical approvals were obtained by the Committee on Human Research at University of Florida (258-2012).

3 RESULTS AND DISCUSSION

The clinical course of HIV-1 infection is characterized by considerable variability in the rate of disease progression among patients with different genetic background, where HLA-B*5701 is the allele most strongly associated with slower progression. The exact mechanisms underlying HIV-1 pathogenesis are not fully understood, but they likely involve the interplay between host immune system and viral evolution. Most of the previous studies have focused either on intra-host viral evolution or on functional aspects of the immune system, which may be insufficient for identifying correlates of protection. The present work focuses on HIV-1 disease progression in HLA-B*5701 and non-HLA-B*57 subjects, which was investigated through a multidisciplinary approach unifying viral evolution and immunological patterns.

The first studies aimed at characterizing HIV-1-specific $CD8^+$ T cell responses and performing advanced evolutionary analysis in six HLA-B*5701 subjects with different risk of disease progression (**Paper I** and **Paper II**). Next, a new method to investigate differences in HIV-1 population dynamics through the analysis of the temporal structure of viral genealogy was developed (**Paper III**). Finally, this method was applied to investigate intra-host phylodynamic patterns and the emergence and population dynamics of archival viral strains in HLA-B*5701 and non-HLA-B*57 subjects (**Paper IV**). The results will be presented and discussed in the following main sections: i) HIV-1 immune and viral factors in HLA-B*5701 low-risk and high-risk progressors; ii) Developing tools to analyze temporal structure of viral genealogies; iii) HIV-1 intra-host phylodynamic patterns and in-depth temporal structure analysis of viral genealogies in HLA-B*5701 subjects and non-HLA-B*57 controls.

3.1 HIV-1 IMMUNE AND VIRAL FACTORS IN HLA-B*5701 LOW-RISK AND HIGH-RISK PROGRESSORS

In **Paper I**, HIV-1 *in vivo* evolution and epitope-specific $CD8^+$ T cell responses were investigated in six untreated HLA-B*5701 patients monitored from early infection for up to seven years. Three subjects were classified as high-risk progressors (HRPs) and three as low-risk progressors (LRPs) based on viral load and baseline $CD4^+$ T cell counts (10-11 weeks post-infection, wpi) [67]. The focus was on the Gag p24-region since several HLA-B*5701-restricted epitopes (ISW9, KF11, TW10 and QW9) are located in this region of the viral genome. Immune responses were identified against wild-type and autologous variants of the epitopes. Interestingly, polyfunctional $CD8^+$ T cell responses toward wild-type HLA-B*5701-restricted epitopes were found to be more robust in LRPs than HRPs. The fraction of $CD8^+$ T cells coproducing IFN-γ, MIP-1β and IL-2 in response towards the epitopes TW10 and QW9 were significantly higher in the LRP group (Figure 8). The LRPs also showed a significantly higher fraction of IL-2 producing $CD8^+$ T cells (in response to wild-type epitopes) in early infection compared to HRPs. Importantly, a significant positive correlation between the fraction of IL-2-producing $CD8^+$ T cells and the $CD4^+$ T cell count was found when all subjects were included in the analysis.

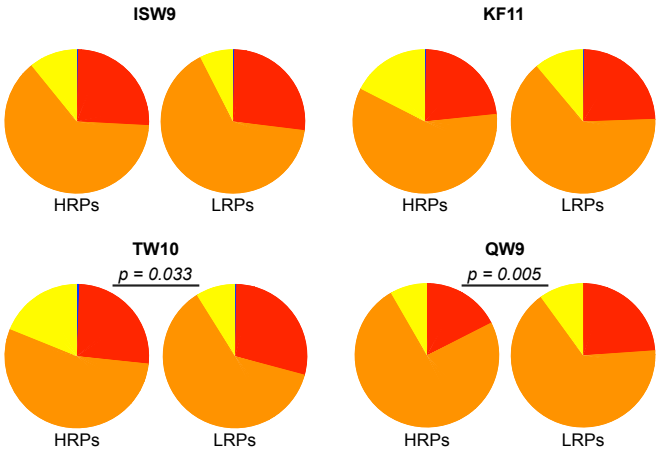

Figure 8. Functional discrepancies of epitope-specific $CD8^+$ T cell responses between HRPs and LRPs. Pie charts representing the fraction of $CD8^+$ T cell responses towards each HLA-B*5701-restricted wild-type epitope in Gag p24. One to four functions are illustrated by the colors yellow, orange, red and blue, respectively. Adapted from [136].

Parallel to the immunological findings, interesting evolutionary patterns were uncovered. In all study subjects, HIV-1 *gag* p24 sequences exhibited more substitutions in flanking regions than in HLA-B*5701-restricted epitopes, but sequence diversity and mean evolutionary rates were significantly lower in LRPs compared to HRPs (**Paper I**). Sequences also exhibited a higher degree of homoplasy in LRPs. More in-depth analyses revealed that this difference was mainly determined by significantly lower synonymous substitution rates (**Paper II**). The difference in synonymous substitution rates between the two groups of patients was especially significant for backbone branches, which represent the surviving viral population successfully emerging over time from sequential population bottlenecks (Figure 9, left). The viral quasispecies infecting LRPs were also characterized by a slower increase in synonymous divergence over time (Figure 9, right).

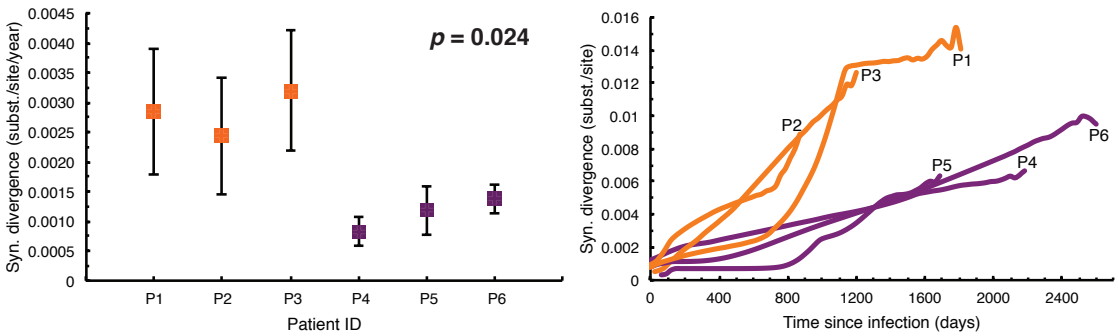

Figure 9. Mean synonymous substitution rates (left) and synonymous divergence over the course of HIV-1 infection (right) for backbone branches in HLA-B*5701 subjects. HRPs (P1-P3) and LRPs (P4-P6) are shown in orange and purple, respectively. Adapted from [136].

Additionally, p24 sequences exhibited more constrained mutational patterns along the major bottlenecks, as defined by the topological structure of the viral genealogy, of the quasispecies infecting LRPs (**Paper I**). Mutational events leading to changes in TW10 and ISW9 (including upstream position 146) and the CypA-binding loop appeared to occur in a specific order (Figure 10, right). On the other hand, in HRPs, viral evolutionary patterns along population bottlenecks appeared to be less constrained and were characterized by the more frequent emergence of variants with amino acid replacements in the CypA-binding loop and/or other flanking regions (Figure 10, left). It is also important to mention that sequences in HRPs displayed the eventual

emergence of QW9 variants, while this epitope region was conserved in LRPs. The results indicate that HIV-1 evolution in HLA-B*5701 patients with a high baseline $CD4^+$ T cell count was constrained by polyfunctional $CD8^+$ T cell responses with a maintained ability to coproduce IL-2. This is strengthened by the significant correlation found between the fraction of IL-2-producing $CD8^+$ T cells and $CD4^+$ T cell count, which in turn could be associated with a lower synonymous substitution rate and risk of disease progression. There are also several studies that have underlined the important role of IL-2-producing $CD8^+$ T cell functional subsets to mediate protective immunity at different stages of infection [137-139].

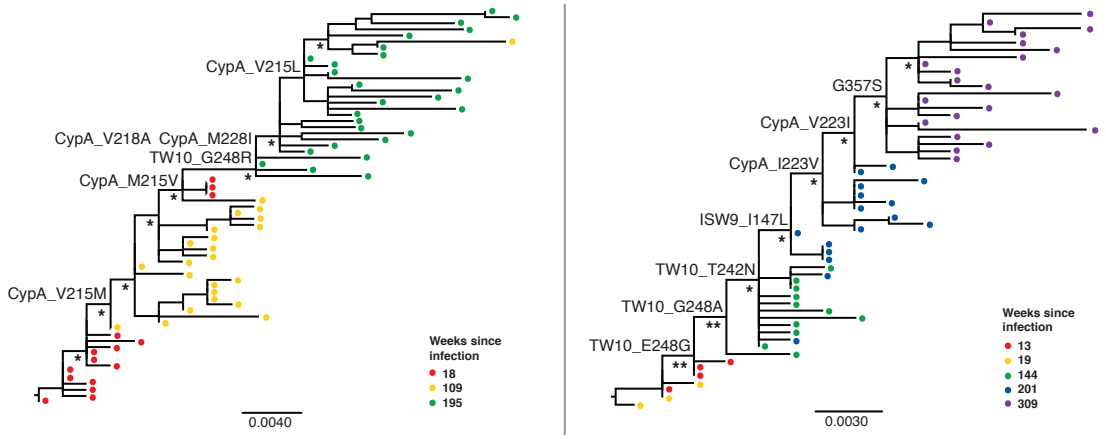

Figure 10. Maximum likelihood *gag* p24 genealogies (using only unique sequences) for HRP (left) and a LRP (right). Statistical support for specific branches are indicated by * (bootstrap > 65% or aLRT p > 0.75) or ** (bootstrap > 65% and aLRT p > 0.75). Amino acid replacements along supported internal branches are indicated; the numbering refers to Gag amino acid position in HXB2, used as reference. Adapted from [136].

Several underlying mechanisms could potentially account for the observed differences in synonymous substitution rates and divergence over time between the two groups of patients. Since synonymous substitutions are neutral or nearly neutral [140], the HIV-1 synonymous substitution rate is expected to be proportional to the viral replication rate [141]. The higher synonymous rates in HRPs could be the consequence of an infection with fitter viral variants characterized by faster replication, but *in vitro* replication capacity (RC) did not display any differences between the two groups of patients (**Paper II**). The pattern in synonymous substitution rates between HRPs and LRPs could not be explained by differences among subjects in $CD4^+$ and $CD8^+$ T cell activation or selection pressure either. However, there are some limitations in these studies that should be mentioned. First, the lack of significance for some analysis might be due to the small sample size and lack of statistical power (**Paper I** and **Paper II**). Second, no samples from the acute phase of HIV-1 infection (i.e. earlier than 11 wpi) were available to assess early events in terms of viral evolution and immune responses (**Paper I**) or to compare whether RC or T cell activation differed between the two groups of patients during primary infection (**Paper II**). Third, the RC assay only tested one part of the viral genome that may not fully capture the total viral replication capacity (**Paper II**). Nevertheless, the data suggest that neither T cell activation nor an initial infection with fitter viral variants would explain the difference in synonymous rates between HLA-B*5701 HRPs or LRPs.

In order to identify other potential mechanisms behind the observed differences in viral evolutionary rate, the correlation between mean nonsynonymous or synonymous rates for different branch sets (all, internal, backbone and external) and clinical parameters for each patient were analyzed (**Paper II**). It was found that higher baseline $CD4^+$ T cell counts (10-11 wpi) were strongly correlated with lower HIV-1 mean synonymous substitution rates ($R^2 = 0.9$, $p = 0.002$) along backbone branches (Figure 11). These results indicate lower replication rates and/or longer viral generation times in the LRPs compared to HRPs. They also suggest that HLA-B*5701 subjects with higher baseline $CD4^+$ T cell counts (> 750 cells/mm^3 at 10-11 wpi) are capable of keeping HIV-1 replication under better control during the course of the infection. Such an observation is in agreement with results presented in **Paper I** showing more constrained viral evolution in LRPs, probably linked to a more robust HLA-B*5701-specific $CD8^+$ T cell response. There is, indeed, evidence that the emergence of escape mutations in p24, as a consequence of $CD8^+$ T cell responses, can negatively affect viral fitness [142], and thereby be indirectly responsible for control of viral replication, longer generation times, and lower risk to progress to AIDS.

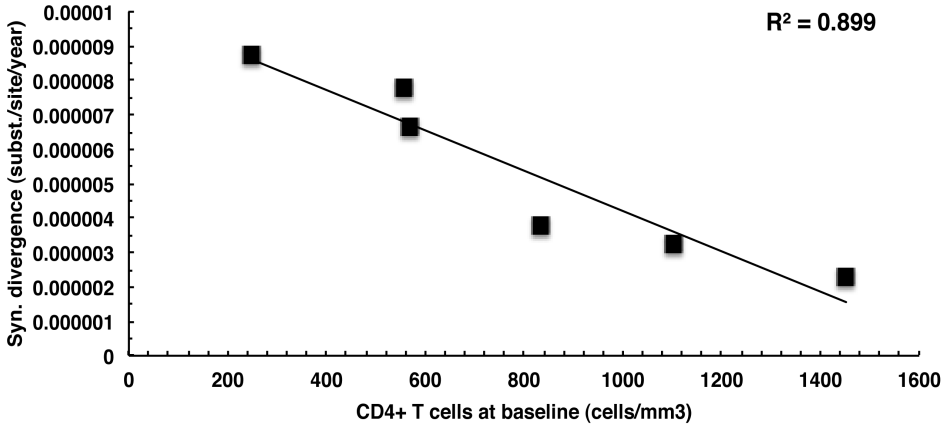

Figure 11. Mean synonymous rates for backbone branches vs. $CD4^+$ T cell count at baseline for each subject. Strong inverse correlation between mean synonymous rates and baseline $CD4^+$ T cell counts (10-11 wpi) by pooling together estimates for all HLA-B*5701 subjects. The correlation is highly significant ($p = 0.002$).

The overall findings may indicate that polyfunctional $CD8^+$ T cell responses, directed towards HLA-B*5701-restricted epitopes, with specific functional profiles are the underlying mechanism of the observed difference in viral evolutionary rates between HLA-B*5701 subjects with different baseline $CD4^+$ T cell counts. However, constrained HIV-1 evolution could also be the direct or indirect result of other host factors, which subsequently affect $CD8^+$ T cell polyfunctionality. It is also important to consider that the focus and study design of the present work cannot exclude the role of individual responses restricted by alleles other than HLA-B*5701.

In summary, the first two papers showed that subjects with high baseline $CD4^+$ T cell counts (LRPs) displayed viral quasispecies evolving along more constrained mutational patterns, characterized by lower synonymous substitution rates, less changes in flanking regions, and no emergence of QW9 variants. The findings in **Paper I** and **Paper II** also revealed that these subjects had specific HLA-linked immune responses, lower replication rate and lower risk of HIV-1 disease progression. The integration of

experimental data with coalescent-based estimates allowed the development of, for the first time, a possible explanation for the correlation between HIV-1 *in vivo* replication rate and different risk of disease progression in HLA-B*5701 subjects.

3.2 DEVELOPING TOOLS TO ANALYZE TEMPORAL STRUCTURE OF VIRAL GENEALOGIES

To further assess the findings of **Paper I** and **Paper II** and evaluate them in the context of HIV-1 intra-host population dynamics, a novel methodology was adopted to investigate the temporal structure of phylogenetic trees inferred from longitudinal samples.

In **Paper III**, a new topological measure, called temporal clustering (TC), which characterizes the temporal structure of a serial samples phylogeny, was implemented in the R language. A set of R scripts, called PhyloTempo, integrated TC calculations and several other tree topological measures in a user-friendly interface. The comparison of the TC statistic with other measures provides multifaceted insights on the evolutionary and population dynamic processes shaping serially sampled genealogies of pathogenic viruses.

The features and applicability of PhyloTempo were tested on longitudinal intra-host human and simian immunodeficiency virus population data sets. The intra-host HIV-1 data set consisted of *gag* p24 sequences from the HLA-B*5701 subjects described in **Paper I** and **Paper II**. PhyloTempo estimates the TC statistic and implements a graphical output interface displaying phylogenetic trees with maximum likelihood reconstructed ancestral states (**Paper III**). TC analysis can reveal important information of archival viral strains and the impact of continuous immune selection on viral population dynamics.

By running the HLA-B*5701 data sets, several biological insights emerged. First, TC did not correlate with any previously described topological tree measure, implying that the new TC statistics evaluates aspects of the evolutionary process not captured by existing methods. Second, despite several studies having interpreted temporally structured phylogenies as evidence of sequential viral population bottlenecks driven by continuous selection pressure [73, 88, 90, 143], TC did not correlate with estimate dN/dS ratios in the different data sets analysed. Additionally, archival viral strains expressed in cellular reservoirs would decrease the temporal structure of serially sampled genealogies, because sequences from later time points may share a most recent common ancestor with sequences collected much earlier in infection [144].

3.3 HIV-1 INTRA-HOST PHYLODYNAMIC PATTERNS AND IN-DEPTH TEMPORAL STRUCTURE ANALYSIS OF VIRAL GENEALOGIES IN HLA-B*5701 SUBJECTS AND NON-HLA-B*57 CONTROLS

To obtain insights into how host genetic factors may impact HIV-1 population dynamics and disease progression, a control group of non-HLA-B*57 subjects was also

included in the study cohort. In **Paper IV**, temporal structure of HIV-1 intra-host genealogies, as well as viral phylodynamic patterns, were investigated in the six HLA-B*5701 subjects and six additional non-HLA-B*57 controls to compare HIV-1 population dynamics in subjects with different genetic backgrounds. The HIV-1 intra-host demographic history was inferred for each subject. A peak in viral effective population size (*Ne*), representing the number of effectively infectious viral genomes (*i.e.* the ones effectively contributing to the next generation), was observed within the first year of infection in control subjects. This peak in *Ne* was also often followed by a second increase later in infection. In the HLA-B*5701 subjects, on the other hand, either constant *Ne* or *Ne* increase two to four years post-infection was observed (Figure 12).

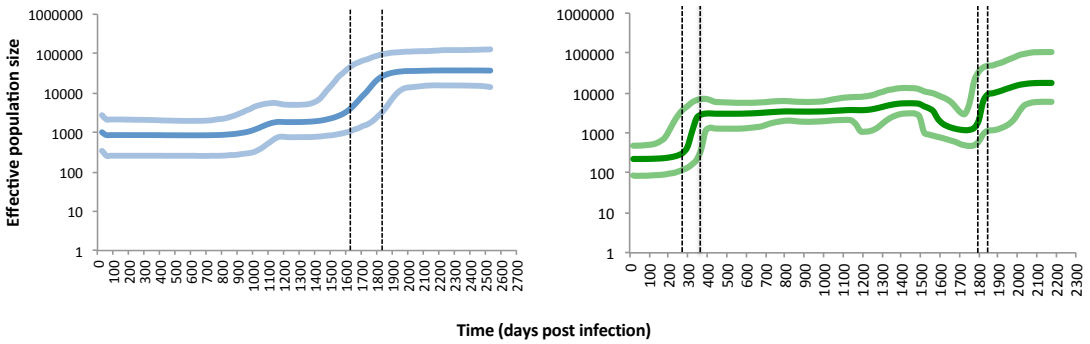

Figure 12. HIV-1 demographic history in a HLA-B*5701 subject (left) and a non-HLA-B*57 control (right). Bayesian skyline plots (BSPs) representing non-parametric estimates of the viral effective population time (*Ne*) over time. The thick line represents median *Ne* estimates, while the thin lines indicate the upper and lower 95% high posterior density (HPD) for the estimate of *Ne*. The estimate of *Ne* is on the y-axis and time is on the *x*-axis.

In-depth TC analysis revealed earlier emergence of archival viral strains in non-HLA-B*57 (1.8 – 2.5 years post infection) than in HLA-B*5701 subjects (2.7 – 5.5 years post-infection (Figure 13). The TC analysis also showed that re-emergence of archival viral strains appeared over a significantly shorter time (p = 0.015) in non-HLA-B*57 controls (1 year) compared to HLA-B*5701 subjects (2 years) (Figure 14).

The earlier *Ne* peak observed in non-HLA-B*57 control subjects was likely influenced by multiple factors, including earlier and regular detection of archival viral strains in the plasma. The emergence or re-emergence of archival strains seemed to contribute less often to the HIV-1 population diversity infecting subjects that carry the HLA-B*5701 protective allele. It is possible that in non-HLA-B*57 control subjects, potential reservoirs may be re-activated earlier and/or more often than in subjects carrying the protective allele, which would result in the replication of the archival viruses detected in the longitudinal plasma samples.

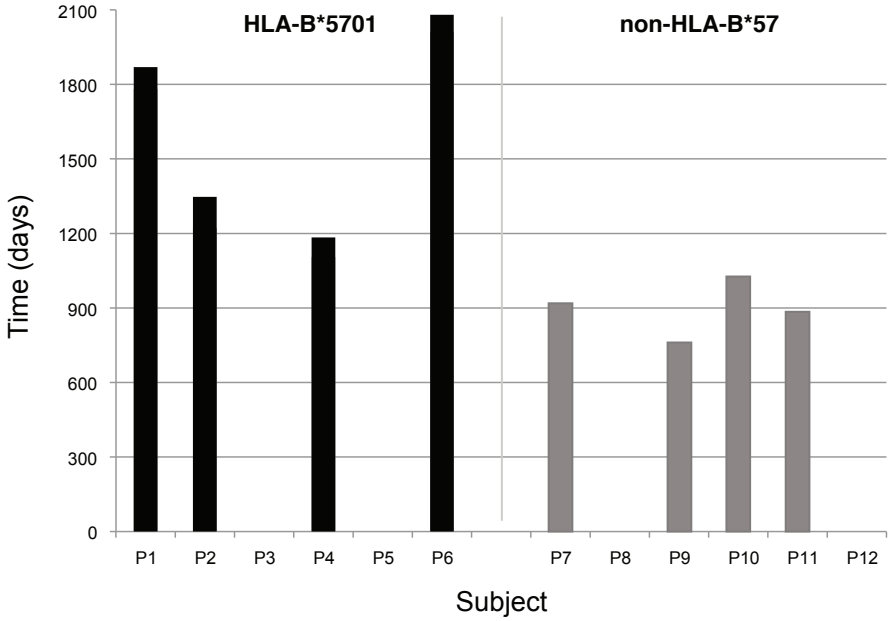

Figure 13. Temporal Clustering analysis in twelve patients: six HLA-B*5701 subjects (P1-P6) and six non-HLA-B*57 control subjects (P7-P12). Each bar indicates the earliest sampling time (Ti) of an inferred time transition (from Ti to Tj with i<j) that was significantly more frequent ($p<0.05$) than expected by chance across the HIV-1 genealogy from each subject (x-axis). This is the first sampling time at which an archival viral sequence could be detected in the genealogy. Time (y-axis) is given in days post-infection.

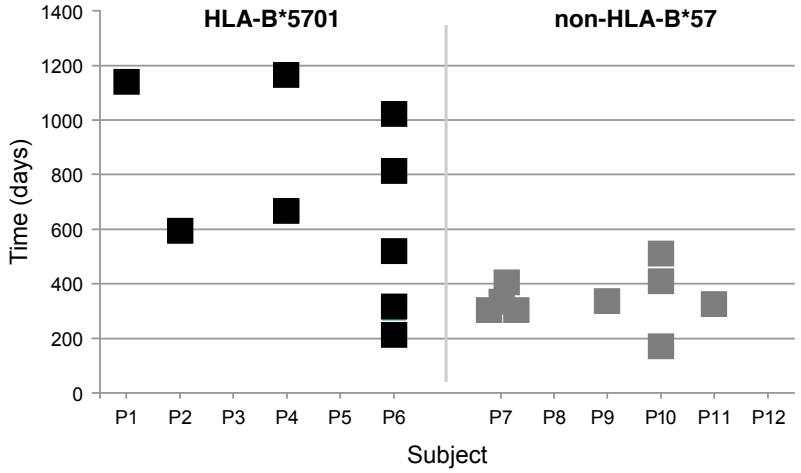

Figure 14. Length of inferred ancestral time transitions across genealogy samples. Squares indicate inferred ancestral time transitions (from *Ti* to *Tj* with *i<j*) that were significantly more frequent ($p<0.05$) than expected by chance across the HIV-1 genealogy from each subject (x-axis). The y-axis indicates the time interval ($\Delta T = Tj-Ti$ with *i<j*) for each transition and represents the waiting time for the re-emergence of archival strains.

The different dynamics in HLA-B*5701 and non-HLA-B*57 subjects is possibly the result of the different genetic backgrounds and the interplay between immune system and viral quasispecies. Therefore, the study in **Paper IV** provided novel insights into how host genetic factors may impact viral population dynamics and HIV-1 disease progression.

4 CONCLUSIONS AND FUTURE PERSPECTIVE

The exact mechanisms underlying HIV-1 pathogenesis and varying disease outcome still remain elusive after 30 years of research. Correlates of protection remain undefined, but investigating HIV-1 disease progression may reveal essential knowledge for successful vaccine design, and potentially lead to the development of therapies able to eradicate the infection. However, it may require a multidisciplinary approach combining both fields of HIV virology and immunology to obtain a better picture of the complex interplay between HIV-1 intra-host evolution and immune system. This thesis summarizes several studies where viral evolution and immunological patterns were unified to investigate risk of HIV-1 disease progression in HLA-B*5701 subjects. A method was also implemented to investigate the temporal structure and phylodynamic patterns of HIV-1 intra-host genealogies from subjects with different genetic backgrounds. The thesis' work resulted in several main findings:

- HLA-B*5701 subjects with higher baseline $CD4^+$ T cell counts had significantly lower HIV-1 intra-host *gag* p24 sequence diversity, lower evolutionary rate and more constrained mutational pattern compared to subjects with lower baseline $CD4^+$ T cell counts (**Paper I**).
- The fraction of polyfunctional and IL-2-producing $CD8^+$ T cell responses, particularly to TW10 and QW9 epitopes, were more robust in HLA-B*5701 subjects with higher baseline $CD4^+$ T cell counts (**Paper I**).
- A significant correlation was found between the fraction of IL-2-producing $CD8^+$ T cells and $CD4^+$ T cell count for the HLA-B*5701 subjects (**Paper I**).
- HLA-B*5701 subjects with higher baseline $CD4^+$ T cell counts had significantly lower HIV-1 synonymous substitution rates compared to subjects with lower baseline $CD4^+$ T cell counts (**Paper II**).
- A significant inverse correlation between baseline $CD4^+$ T cell counts and mean HIV-1 synonymous substitution rates was found for the HLA-B*5701 subjects (**Paper II**).
- PhyloTempo, a new program implemented to analyze temporal structure of viral genealogies, revealed that the TC statistic did not correlate with any previously described topological tree measure and may evaluate aspects of the evolutionary process not captured by existing methods (**Paper III**).
- Changes in viral effective population size (*Ne*) over time were more constrained in HLA-B*5701 subjects compared to non-HLA-B*57 controls (**Paper IV**).
- HLA-B*5701 subjects displayed later and less regular re-emergence of archival sequences compared to non-HLA-B*57 controls (**Paper IV**).

Most studies so far have focused either on intra-host viral evolution or on functional aspects of the immune system. Few attempts have been made to apply the full phylodynamic framework specifically developed to investigate the interplay between pathogen evolution and host immune response [88]. Immunological (REFs) and evolutionary studies [90, 132, 145] have provided important insights in the HIV field, but correlates of protection still remain elusive because, as demonstrated in the present work, they require study design and experimental settings able to integrate both immunological and viral evolutionary findings.

Paper I provided insights into the roles of polyfunctional $CD8^+$ T cells and IL-2-producing T cells in constraining the tempo and mode of HIV-1 evolution in HLA-B*5701 subjects. This suggests that these biomarkers may serve as additional measures of risk for disease progression, including both viral evolutionary dynamics and host immune responses [92]. Additionally, **Paper II** provided a possible mechanism for different risk of HIV-1 disease progression in HLA-B*5701 subjects. This may have significant translational impact on clinical practice since synonymous rates, which are proportional to *in vivo* replication rates, could be used as an evolutionary marker of disease progression in HLA-B*5701 subjects. Finally, **Paper IV** provided insights into how host genetic factors impact viral population dynamics and HIV-1 disease progression.

Correlates of protection, in other words, likely involve both evolutionary and immunological biomarkers. Through a study design including extensive clinical and immunological observations (measurement of polyfunctional $CD8^+$ T cell responses, T cell activation, HLA-typing, $CD4^+$ T cell counts, viral load), as well as high-resolution phylodynamic analysis (genealogies from longitudinal samples, Bayesian estimates of evolutionary rates and effective population size, and temporal structure analysis of viral genealogies) it is possible to unify viral and immunological patterns within a coherent model. By using this approach, new models could be developed to explain HIV-1 pathogenesis and risk of disease progression, as well as shed new light on how host genetic factors impact HIV-1 intra-host population dynamics.

The studies described in this thesis have provided valuable insights in the HIV field, but also call for further investigation, as several key questions remain unanswered. First, in order to confirm the findings in **Papers I-IV** and increase statistical power, the same type of studies would need to be performed on a larger number of patients. Second, samples from acute infection should be included, if possible, to characterize the earliest immunological and evolutionary events. Third, it is important to further investigate the viral dynamics in the non-HLA-B*57 control subjects by obtaining non-synonymous and synonymous divergence, replication capacity and T cell activation data. Then, the same methods and analysis described in **Paper II** for the HLA-B*5701 subjects could be applied to the control data to obtain these results. It is also important to stress that all the studies included in this thesis are based on the *gag* p24-region. Therefore, knowledge of immunological and viral factors in other regions of the HIV-1 genome would be of value and the data could be obtained by applying the methodologies described in **Paper I**. It is already known, for example, that several HLA-B*5701-restricted epitopes are located in *nef* [146] and it would be interesting to investigate if any differences in this genomic region could be found in the HLA-B*5701 subjects with different risk of disease progression.

Finally, this thesis shows the power of using a multidisciplinary approach that may have potential translational impact not only in the HIV field, but also in the study of other fast-evolving viruses by providing a general framework unifying viral evolution and immunological patterns.

5 ACKNOWLEDGEMENTS

Annika Karlsson, thank you for supporting me during these years and for accepting me as a PhD student in your group. You truly gave me the freedom to explore research, to find my own way, and this has shaped me into the scientist I am today. I am deeply grateful for your willingness to always let me pursue the research projects and collaborations I wanted. Thank you for everything!

I would also like to thank all my co-supervisors: **Klas Kärre, Adnane Achour and Jakob Michaëlsson**. You all believed in me at the very beginning of my PhD studies and supported me in the KID application. Thank you Adnane for introducing me to Annika! Thank you Klas for always being at my yearly seminars.

Marco Salemi, aka OFL – while you have not officially been my supervisor, there are no words to express how grateful I am for all the support and inspiration you have given me. Many of the studies in this thesis would not have been possible if it was not for you. I love to work with you and I am so happy you welcomed me to join your research group at University of Florida. Thank you for being such a fantastic mentor!

Rick Hecht, a big thank you for the patient samples, this work would not have been possible without you, and always giving me valuable input and perspective on the studies I performed during these years. Thank you for your enthusiasm and always being so involved!

I would also like to thank **all co-authors** for their help, especially the people at Monogram. Thank you **Wei Huang** for all the RC experiments and such a great communication, and **Wendy Hartogensis** for all the statistical analysis. Thank you **Mark Wallet** for your valuable immunological expertise!

Sven Britton, a big thank you for taking me onboard and showing me Ethiopia and Ghana. Both travels have been life changing for me and it was a joy to listen to you share your knowledge and experience. Thank you so much!

Poker, thank you for the great lunches and coffee breaks! It is always fun to see and talk to you. I am sure we will stay in touch in the future ☺

I would like to thank all other current and former members in Annika's group: **Marcus**, thank you for sharing these years with me by always listening and helping me! I would also like to thank **Leda**, and hope we will meet soon when you come to Sweden. **Johanna T**, thank you for being such an excellent student helping me in my projects. I wish you all good luck with your research projects!

A big thank you to all people I got to know at **MTC**, **SMI** and **LABMED** during my years as a PhD student. Thank you **Charlotte** and **Salma** for being amazing friends! We had such great times living and travelling together . Thank you **Viktor**, for being such a fantastic friend sharing the desk next to me during my first year as a PhD student. I would also like to thank **Lina, Johanna, Wendy**, **Helena**, **Mattias, Malin**, **Karin, Tara** and everyone else I got to know during these years. I cannot mention you all, but you for sure contributed to the atmosphere! Even though it is not possible to mention all of the people that have contributed to the work in this thesis and that have

supported me over the years… You all know who you are and I am grateful for meeting you on this journey!

The Florida crew, especially **David**! You are sooo freaking great! I am so happy I got to know you and I really value you as both a co-worker and a friend. My time in Gainesville would not have been the same without you. Dr. Ama misses you like crazy!!! Thank you **Mattia**, for all the great science we have done together and of course the funny discussions! I hope you are enjoying Manchester! I would also like to thank the hot redhead **Audrey**, **Sam**, **Nazle**, **Taj**… and of course **Sofia** – my Swedish friend ☺

Agnes, thank you for being so enthusiastic and the highlight of my summer. I really enjoyed it and hope to see you soon! We will always be partners in crime.

David, I made it! Do you remember the very first summer we worked together with all the craziness? I had so much fun and thank you for being such an engaged student that made me grow and learn new things.

I would like to thank all people I have gotten to know in all boards and committees. Thank you to all people I worked with in **MF**, **GSA**, **SSCO**, **SFS-DK** and all boards at **KI**. My activity in these boards was a true learning experience for me and I am grateful for the diversity of people I met as well as the huge range of experiences.

I am extremely grateful for all the scholarships I have been granted during my years as a PhD student. Thank you **Erik and Edith Fernstrom**, **Sven Gard**, **KI funds**, **Johanna Björkman**! Thanks to you I have been able to travel and meet international collaborators. I was also able to stay at University of Florida for more than a year!

Thank you all crazy people in the **TNT-family**. I love you all and **TNT** completely changed my life during my time as a PhD student. Serving has taken on a whole new meaning for me. What would I have done during my free time without you guys? ☺ I would especially like to thank my lovely sisters: **Anastasia**, **Police woman**, **Maidenhead**, **Amazon**, **Dominatrix**, **Linda L**, **Olive**. I would like to thank my amazing brothers: **Pan**, **Popeye**, **Dexter, Abe** and of course… my buddy **C Tiger** - you have helped me see myself and it has been a joy to get to know you and share the things we have done the past few months. I really love you all!

Thank you to the Belgians **Guido**, **Anita**, **Koen**, **Katrien** and most importantly **Falco-biche** for the super delicious dinners, fun trips and a lot of joy and laughter ☺

I would also like to thank my family: **Robin** and **Miranda**, you are my crazy siblings that are with me no matter what I do or wherever I decide to travel in the world. I am so happy to have you in my life and I really love you from the bottom of my heart. Thank you **Mom**, for your willingness to work things out and I am grateful for our contact over the past few months. I have missed you!

Bruno, even though this journey was far from easy I have always felt your support. Thank you so much for being at my side and showing me what unconditional love is! I never thought I would be able to experience that with someone. I truly love you so much! ♥

6 REFERENCES

1. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, Michael SF*, et al.* Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* 1999,**397**:436-441.
2. Gao F, Yue L, White AT, Pappas PG, Barchue J, Hanson AP*, et al.* Human infection by genetically diverse SIVSM-related HIV-2 in west Africa. *Nature* 1992,**358**:495-499.
3. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML*, et al.* Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science* 2006,**313**:523-526.
4. Takehisa J, Kraus MH, Ayouba A, Bailes E, Van Heuverswyn F, Decker JM*, et al.* Origin and biology of simian immunodeficiency virus in wild-living western gorillas. *J Virol* 2009,**83**:1635-1648.
5. Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, Mboup S*, et al.* Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 2001,**15**:276-278.
6. Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, Vandamme AM. Tracing the origin and history of the HIV-2 epidemic. *Proc Natl Acad Sci U S A* 2003,**100**:6588-6592.
7. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, Bunce M*, et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 2008,**455**:661-664.
8. Worobey M, Telfer P, Souquiere S, Hunter M, Coleman CA, Metzger MJ*, et al.* Island biogeography reveals the deep history of SIV. *Science* 2010,**329**:1487.
9. Gray RR, Tatem AJ, Lamers S, Hou W, Laeyendecker O, Serwadda D*, et al.* Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. *AIDS* 2009,**23**:F9-F17.
10. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A*, et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* 2000,**288**:1789-1796.
11. Tatem AJ, Hemelaar J, Gray RR, Salemi M. Spatial accessibility and the spread of HIV-1 subtypes and recombinants in sub-Saharan Africa. *AIDS* 2012.
12. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men--New York City and California. *MMWR Morb Mortal Wkly Rep* 1981,**30**:305-308.
13. Hymes KB, Cheung T, Greene JB, Prose NS, Marcus A, Ballard H*, et al.* Kaposi's sarcoma in homosexual men-a report of eight cases. *Lancet* 1981,**2**:598-600.
14. Update on acquired immune deficiency syndrome (AIDS)--United States. *MMWR Morb Mortal Wkly Rep* 1982,**31**:507-508, 513-504.
15. Possible transfusion-associated acquired immune deficiency syndrome (AIDS) - California. *MMWR Morb Mortal Wkly Rep* 1982,**31**:652-654.
16. Barre-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J*, et al.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* 1983,**220**:868-871.
17. Holmes EC. What does virus evolution tell us about virus origins? *J Virol* 2011,**85**:5247-5251.

18. Nowak MA, Anderson RM, McLean AR, Wolfs TF, Goudsmit J, May RM. Antigenic diversity thresholds and the development of AIDS. *Science* 1991,**254**:963-969.
19. UNAIDS. World AIDS day report 2011. *Geneva* 2011.
20. *Global report : UNAIDS report on the global AIDS epidemic*: Geneva : UNAIDS, 2010-.
21. Dalgleish AG, Beverley PC, Clapham PR, Crawford DH, Greaves MF, Weiss RA. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature* 1984,**312**:763-767.
22. Freed EO. HIV-1 and the host cell: an intimate association. *Trends Microbiol* 2004,**12**:170-177.
23. Tang H, Kuhen KL, Wong-Staal F. Lentivirus replication and regulation. *Annu Rev Genet* 1999,**33**:133-170.
24. Sierra S, Kupfer B, Kaiser R. Basics of the virology of HIV-1 and its replication. *J Clin Virol* 2005,**34**:233-244.
25. Emerman M, Malim MH. HIV-1 regulatory/accessory genes: keys to unraveling viral and host cell biology. *Science* 1998,**280**:1880-1884.
26. Seelamgari A, Maddukuri A, Berro R, de la Fuente C, Kehn K, Deng L*, et al.* Role of viral regulatory and accessory proteins in HIV-1 replication. *Front Biosci* 2004,**9**:2388-2413.
27. Kamp W, Berk MB, Visser CJ, Nottet HS. Mechanisms of HIV-1 to escape from the host immune surveillance. *Eur J Clin Invest* 2000,**30**:740-746.
28. Steffens CM, Hope TJ. Recent advances in the understanding of HIV accessory protein function. *AIDS* 2001,**15 Suppl 5**:S21-26.
29. Perez CL, Larsen MV, Gustafsson R, Norstrom MM, Atlas A, Nixon DF*, et al.* Broadly immunogenic HLA class I supertype-restricted elite CTL epitopes recognized in a diverse population infected with different HIV-1 subtypes. *Journal of immunology* 2008,**180**:5092-5100.
30. Hoof I, Perez CL, Buggert M, Gustafsson RK, Nielsen M, Lund O*, et al.* Interdisciplinary analysis of HIV-specific CD8+ T cell responses against variant epitopes reveals restricted TCR promiscuity. *Journal of immunology* 2010,**184**:5383-5391.
31. Leslie A, Price DA, Mkhize P, Bishop K, Rathod A, Day C*, et al.* Differential selection pressure exerted on HIV by CTL targeting identical epitopes but restricted by distinct HLA alleles from the same HLA supertype. *J Immunol* 2006,**177**:4699-4708.
32. Koup RA, Safrit JT, Cao Y, Andrews CA, McLeod G, Borkowsky W*, et al.* Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J Virol* 1994,**68**:4650-4655.
33. Sandberg JK, Fast NM, Nixon DF. Functional heterogeneity of cytokines and cytolytic effector molecules in human CD8+ T lymphocytes. *J Immunol* 2001,**167**:181-187.
34. Voskoboinik I, Smyth MJ, Trapani JA. Perforin-mediated target-cell death and immune homeostasis. *Nat Rev Immunol* 2006,**6**:940-952.
35. Harty JT, Tvinnereim AR, White DW. CD8+ T cell effector mechanisms in resistance to infection. *Annual review of immunology* 2000,**18**:275-308.
36. Mellors JW, Rinaldo CR, Jr., Gupta P, White RM, Todd JA, Kingsley LA. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* 1996,**272**:1167-1170.

37. Cohen MS, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, Kumarasamy N*, et al.* Prevention of HIV-1 infection with early antiretroviral therapy. *The New England journal of medicine* 2011,**365**:493-505.
38. Fang CT, Hsu HM, Twu SJ, Chen MY, Chang YY, Hwang JS*, et al.* Decreased HIV transmission after a policy of providing free access to highly active antiretroviral therapy in Taiwan. *The Journal of infectious diseases* 2004,**190**:879-885.
39. Bunnell R, Ekwaru JP, Solberg P, Wamai N, Bikaako-Kajura W, Were W*, et al.* Changes in sexual behavior and risk of HIV transmission after antiretroviral therapy and prevention interventions in rural Uganda. *AIDS* 2006,**20**:85-92.
40. Kearney M, Maldarelli F, Shao W, Margolick JB, Daar ES, Mellors JW*, et al.* Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *Journal of virology* 2009,**83**:2715-2727.
41. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG*, et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences of the United States of America* 2008,**105**:7552-7557.
42. Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T*, et al.* Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PloS one* 2010,**5**:e12303.
43. Zhang LQ, MacKenzie P, Cleland A, Holmes EC, Brown AJ, Simmonds P. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. *Journal of virology* 1993,**67**:3345-3356.
44. Zhu T, Mo H, Wang N, Nam DS, Cao Y, Koup RA*, et al.* Genotypic and phenotypic characterization of HIV-1 patients with primary infection. *Science* 1993,**261**:1179-1181.
45. Herbeck JT, Rolland M, Liu Y, McLaughlin S, McNevin J, Zhao H*, et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *Journal of virology* 2011,**85**:7523-7534.
46. Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, Li H*, et al.* Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *The Journal of experimental medicine* 2009,**206**:1273-1289.
47. Masharsky AE, Dukhovlinova EN, Verevochkin SV, Toussova OV, Skochilov RV, Anderson JA*, et al.* A substantial transmission bottleneck among newly and recently HIV-1-infected injection drug users in St Petersburg, Russia. *The Journal of infectious diseases* 2010,**201**:1697-1702.
48. Kouyos RD, von Wyl V, Yerly S, Boni J, Rieder P, Joos B*, et al.* Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2011,**52**:532-539.
49. Nowak P, Karlsson AC, Naver L, Bohlin AB, Piasek A, Sonnerborg A. The selection and evolution of viral quasispecies in HIV-1 infected children. *HIV medicine* 2002,**3**:1-11.
50. Schuitemaker H, van 't Wout AB, Lusso P. Clinical significance of HIV-1 coreceptor usage. *Journal of translational medicine* 2011,**9 Suppl 1**:S5.
51. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H*, et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology* 1999,**73**:10489-10502.

52. Guadalupe M, Reay E, Sankaran S, Prindiville T, Flamm J, McNeil A*, et al.* 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. *Journal of virology* 2003,**77**:11708-11717.
53. Schacker TW, Hughes JP, Shea T, Coombs RW, Corey L. Biological and virologic characteristics of primary HIV infection. *Annals of internal medicine* 1998,**128**:613-620.
54. Alimonti JB, Ball TB, Fowke KR. Mechanisms of CD4+ T lymphocyte cell death in human immunodeficiency virus infection and AIDS. *The Journal of general virology* 2003,**84**:1649-1661.
55. Brenchley JM, Schacker TW, Ruff LE, Price DA, Taylor JH, Beilman GJ*, et al.* CD4+ T cell depletion during all stages of HIV disease occurs predominantly in the gastrointestinal tract. *The Journal of experimental medicine* 2004,**200**:749-759.
56. Kahn JO, Walker BD. Acute human immunodeficiency virus type 1 infection. *The New England journal of medicine* 1998,**339**:33-39.
57. Lyles RH, Munoz A, Yamashita TE, Bazmi H, Detels R, Rinaldo CR*, et al.* Natural history of human immunodeficiency virus type 1 viremia after seroconversion and proximal to AIDS in a large cohort of homosexual men. Multicenter AIDS Cohort Study. *The Journal of infectious diseases* 2000,**181**:872-880.
58. Fauci AS, Pantaleo G, Stanley S, Weissman D. Immunopathogenic mechanisms of HIV infection. *Annals of internal medicine* 1996,**124**:654-663.
59. Douek DC. Disrupting T-cell homeostasis: how HIV-1 infection causes disease. *AIDS reviews* 2003,**5**:172-177.
60. Deeks SG, Walker BD. Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy. *Immunity* 2007,**27**:406-416.
61. Lambotte O, Boufassa F, Madec Y, Nguyen A, Goujard C, Meyer L*, et al.* HIV controllers: a homogeneous group of HIV-1-infected patients with spontaneous control of viral replication. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2005,**41**:1053-1056.
62. Okulicz JF, Marconi VC, Landrum ML, Wegner S, Weintrob A, Ganesan A*, et al.* Clinical outcomes of elite controllers, viremic controllers, and long-term nonprogressors in the US Department of Defense HIV natural history study. *The Journal of infectious diseases* 2009,**200**:1714-1723.
63. Kiepiela P, Leslie AJ, Honeyborne I, Ramduth D, Thobakgale C, Chetty S*, et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* 2004,**432**:769-775.
64. O'Brien SJ, Nelson GW. Human genes that limit AIDS. *Nature genetics* 2004,**36**:565-574.
65. Hutter G, Nowak D, Mossner M, Ganepola S, Mussig A, Allers K*, et al.* Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *The New England journal of medicine* 2009,**360**:692-698.
66. Piacentini L, Biasin M, Fenizia C, Clerici M. Genetic correlates of protection against HIV infection: the ally within. *Journal of internal medicine* 2009,**265**:110-124.
67. Mellors JW, Munoz A, Giorgi JV, Margolick JB, Tassoni CJ, Gupta P*, et al.* Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Annals of internal medicine* 1997,**126**:946-954.

68. Seillier-Moiseiwitsch F, Margolin BH, Swanstrom R. Genetic variability of the human immunodeficiency virus: statistical and biological issues. *Annual review of genetics* 1994,**28**:559-596.
69. Peeters M, Sharp PM. Genetic diversity of HIV-1: the moving target. *AIDS* 2000,**14 Suppl 3**:S129-140.
70. Li WH, Tanimura M, Sharp PM. Rates and dates of divergence between AIDS virus nucleotide sequences. *Mol Biol Evol* 1988,**5**:313-330.
71. Mansky LM, Temin HM. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 1995,**69**:5087-5094.
72. Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, Dougherty JP. High rate of recombination throughout the human immunodeficiency virus type 1 genome. *Journal of virology* 2000,**74**:1234-1240.
73. Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. *Nature reviews. Genetics* 2004,**5**:52-61.
74. Norstrom MM, Karlsson AC, Salemi M. Towards a new paradigm linking virus molecular evolution and pathogenesis: experimental design and phylodynamic inference. *The new microbiologica* 2012,**35**:101-111.
75. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* 1996,**271**:1582-1586.
76. Markowitz M, Louie M, Hurley A, Sun E, Di Mascio M, Perelson AS*, et al.* A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. *J Virol* 2003,**77**:5037-5038.
77. Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P*, et al.* Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* 1995,**373**:117-122.
78. Holland JJ, De La Torre JC, Steinhauer DA. RNA virus populations as quasispecies. *Curr Top Microbiol Immunol* 1992,**176**:1-20.
79. Goodenow M, Huet T, Saurin W, Kwok S, Sninsky J, Wain-Hobson S. HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. *J Acquir Immune Defic Syndr* 1989,**2**:344-352.
80. Salemi M, Lewis M, Egan JF, Hall WW, Desmyter J, Vandamme AM. Different population dynamics of human T cell lymphotropic virus type II in intravenous drug users compared with endemically infected tribes. *Proceedings of the National Academy of Sciences of the United States of America* 1999,**96**:13253-13258.
81. Borrow P, Lewicki H, Wei X, Horwitz MS, Peffer N, Meyers H*, et al.* Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus. *Nat Med* 1997,**3**:205-211.
82. Goulder PJ, Phillips RE, Colbert RA, McAdam S, Ogg G, Nowak MA*, et al.* Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nat Med* 1997,**3**:212-217.
83. Karlsson AC, Iversen AK, Chapman JM, de Oliviera T, Spotts G, McMichael AJ*, et al.* Sequential broadening of CTL responses in early HIV-1 infection is associated with viral escape. *PloS one* 2007,**2**:e225.
84. Liu Y, McNevin J, Zhao H, Tebit DM, Troyer RM, McSweyn M*, et al.* Evolution of human immunodeficiency virus type 1 cytotoxic T-lymphocyte epitopes: fitness-balanced escape. *J Virol* 2007,**81**:12179-12188.

85. Raffel TR, Martin LB, Rohr JR. Parasites as predators: unifying natural enemy ecology. *Trends in ecology & evolution* 2008,**23**:610-618.
86. Kuhnert D, Wu CH, Drummond AJ. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 2011,**11**:1825-1841.
87. Williamson S. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Mol Biol Evol* 2003,**20**:1318-1325.
88. Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA*, et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 2004,**303**:327-332.
89. Gray RR, Pybus OG, Salemi M. Measuring the Temporal Structure in Serially-Sampled Phylogenies. *Methods Ecol Evol* 2011,**2**:437-445.
90. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H*, et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol* 1999,**73**:10489-10502.
91. Edwards CT, Holmes EC, Wilson DJ, Viscidi RP, Abrams EJ, Phillips RE*, et al.* Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. *BMC evolutionary biology* 2006,**6**:28.
92. Makedonas G, Betts MR. Living in a house of cards: re-evaluating CD8+ T-cell immune correlates against HIV. *Immunol Rev* 2011,**239**:109-124.
93. Hecht FM, Busch MP, Rawal B, Webb M, Rosenberg E, Swanson M*, et al.* Use of laboratory tests and clinical symptoms for identification of primary HIV infection. *AIDS* 2002,**16**:1119-1129.
94. Munoz A, Kirby AJ, He YD, Margolick JB, Visscher BR, Rinaldo CR*, et al.* Long-term survivors with HIV-1 infection: incubation period and longitudinal patterns of CD4+ lymphocytes. *Journal of acquired immune deficiency syndromes and human retrovirology : official publication of the International Retrovirology Association* 1995,**8**:496-505.
95. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. *Trends in ecology & evolution* 2003,**18**:481-488.
96. Achaz G, Palmer S, Kearney M, Maldarelli F, Mellors JW, Coffin JM*, et al.* A robust measure of HIV-1 population turnover within chronically infected individuals. *Molecular biology and evolution* 2004,**21**:1902-1912.
97. Lindkvist A, Eden A, Norstrom MM, Gonzalez VD, Nilsson S, Svennerholm B*, et al.* Reduction of the HIV-1 reservoir in resting CD4+ T-lymphocytes by high dosage intravenous immunoglobulin treatment: a proof-of-concept study. *AIDS research and therapy* 2009,**6**:15.
98. Jordan MR, Kearney M, Palmer S, Shao W, Maldarelli F, Coakley EP*, et al.* Comparison of standard PCR/cloning to single genome sequencing for analysis of HIV-1 populations. *J Virol Methods* 2010,**168**:114-120.
99. Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, Bazmi H*, et al.* Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* 2005,**43**:406-413.
100. Shriner D, Rodrigo AG, Nickle DC, Mullins JI. Pervasive genomic recombination of HIV-1 in vivo. *Genetics* 2004,**167**:1573-1583.
101. Choe S, Fransen S, Toma J, Petropoulos CJ, Huang W. Assessing replication capacity and susceptibility to maturation and protease inhibitors using a Phenosense HIV assay that captures contiguous gag-pro sequences. XVIIth

Conference on Retroviruses and Opportunistic Infections, San Francisco, CA, USA. 2010.
102. Strimmer K, von Haeseler A. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A* 1997,**94**:6815-6819.
103. Salemi M, Lamers SL, Huysentruyt LC, Galligan D, Gray RR, Morris A*, et al.* Distinct patterns of HIV-1 evolution within metastatic tissues in patients with non-Hodgkins lymphoma. *PLoS One* 2009,**4**:e8153.
104. Xia X, Xie Z, Salemi M, Chen L, Wang Y. An index of substitution saturation and its application. *Mol Phylogenet Evol* 2003,**26**:1-7.
105. Salemi M, Vandamme AM. Hepatitis C virus evolutionary patterns studied through analysis of full-genome sequences. *Journal of molecular evolution* 2002,**54**:62-70.
106. Gray RR, Veras NM, Santos LA, Salemi M. Evolutionary characterization of the West Nile Virus complete genome. *Mol Phylogenet Evol* 2010,**56**:195-200.
107. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 2002,**18**:502-504.
108. Posada D. Unveiling the molecular clock in the presence of recombination. *Mol Biol Evol* 2001,**18**:1976-1978.
109. Posada D, Crandall KA, Holmes EC. Recombination in evolutionary genomics. *Annu Rev Genet* 2002,**36**:75-97.
110. Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. *J Mol Evol* 2002,**54**:396-402.
111. Carvajal-Rodriguez A, Crandall KA, Posada D. Recombination estimation under complex evolutionary models with the coalescent composite-likelihood method. *Mol Biol Evol* 2006,**23**:817-827.
112. Kuhner MK. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 2006,**22**:768-770.
113. Salemi M, Gray RR, Goodenow MM. An exploratory algorithm to identify intra-host recombinant viral sequences. *Mol Phylogenet Evol* 2008,**49**:618-628.
114. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 2006,**23**:254-267.
115. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. *Genetics* 2006,**172**:2665-2681.
116. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987,**4**:406-425.
117. Salemi M, Vandamme A-M, Lemey P. *The phylogenetic handbook : a practical approach to phylogenetic analysis and hypothesis testing*. 2nd ed. Cambridge, UK ; New York: Cambridge University Press; 2009.
118. Felsenstein J. *Inferring phylogenies*. Sunderland, Mass.: Sinauer Associates ; [Basingstoke : Palgrave, distributor]; 2004.
119. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 2011,**28**:2731-2739.
120. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 2007,**7**:214.
121. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol* 2012,**29**:1969-1973.

122. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* 2002,**161**:1307-1320.
123. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol* 2006,**4**:e88.
124. Drummond AJ, Suchard MA. Bayesian random local clocks, or one rate to rule them all. *BMC Biol* 2010,**8**:114.
125. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007,**24**:1586-1591.
126. Yang Z. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *J Mol Evol* 2000,**51**:423-432.
127. Nielsen R, Yang Z. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 1998,**148**:929-936.
128. Swanson WJ, Nielsen R, Yang Q. Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol* 2003,**20**:18-20.
129. Wong WS, Yang Z, Goldman N, Nielsen R. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 2004,**168**:1041-1051.
130. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 2005,**22**:1185-1192.
131. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol Biol Evol* 2012,**29**:2157-2167.
132. Lemey P, Kosakovsky Pond SL, Drummond AJ, Pybus OG, Shapiro B, Barroso H*, et al.* Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS Comput Biol* 2007,**3**:e29.
133. Griffiths RC, Tavare S. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci* 1994,**344**:403-410.
134. Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 2000,**155**:1429-1437.
135. Norstrom MM, Prosperi MC, Gray RR, Karlsson AC, Salemi M. PhyloTempo: A Set of R Scripts for Assessing and Visualizing Temporal Clustering in Genealogies Inferred from Serially Sampled Viral Sequences. *Evol Bioinform Online* 2012,**8**:261-269.
136. Norstrom MM, Buggert M, Tauriainen J, Hartogensis W, Prosperi MC, Wallet MA*, et al.* Combination of immune and viral factors distinguishes low-risk versus high-risk HIV-1 disease progression in HLA-B*5701 subjects. *Journal of virology* 2012,**86**:9802-9816.
137. Cellerai C, Perreau M, Rozot V, Enders FB, Pantaleo G, Harari A. Proliferation capacity and cytotoxic activity are mediated by functionally and phenotypically distinct virus-specific CD8 T cells defined by interleukin-7R{alpha} (CD127) and perforin expression. *Journal of virology* 2010,**84**:3868-3878.
138. Makedonas G, Banerjee PP, Pandey R, Hersperger AR, Sanborn KB, Hardy GA*, et al.* Rapid up-regulation and granule-independent transport of perforin to the immunological synapse define a novel mechanism of antigen-specific CD8+ T cell cytotoxic activity. *Journal of immunology* 2009,**182**:5560-5569.
139. Makedonas G, Hutnick N, Haney D, Amick AC, Gardner J, Cosma G*, et al.* Perforin and IL-2 upregulation define qualitative differences among highly

functional virus-specific human CD8 T cells. *PLoS pathogens* 2010,**6**:e1000798.
140. Sanjuan R, Moya A, Elena SF. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America* 2004,**101**:8396-8401.
141. Lemey P, Kosakovsky Pond SL, Drummond AJ, Pybus OG, Shapiro B, Barroso H*, et al.* Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics. *PLoS computational biology* 2007,**3**:e29.
142. Martinez-Picado J, Prado JG, Fry EE, Pfafferott K, Leslie A, Chetty S*, et al.* Fitness cost of escape mutations in p24 Gag in association with control of human immunodeficiency virus type 1. *Journal of virology* 2006,**80**:3617-3623.
143. Salemi M, Burkhardt BR, Gray RR, Ghaffari G, Sleasman JW, Goodenow MM. Phylodynamics of HIV-1 in lymphoid and non-lymphoid tissues reveals a central role for the thymus in emergence of CXCR4-using quasispecies. *PLoS One* 2007,**2**:e950.
144. Blankson J, Persaud D, Siliciano RF. Latent reservoirs for HIV-1. *Current opinion in infectious diseases* 1999,**12**:5-11.
145. Lee HY, Perelson AS, Park SC, Leitner T. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. *PLoS Comput Biol* 2008,**4**:e1000240.
146. Salgado M, Brennan TP, O'Connell KA, Bailey JR, Ray SC, Siliciano RF*, et al.* Evolution of the HIV-1 nef gene in HLA-B*57 positive elite suppressors. *Retrovirology* 2010,**7**:94.