Analysis of HIV-host interaction on different scales

Katarzyna Bożek

PhD thesis

Analysis of HIV-host interaction on different scales

Dissertation
zur Erlangung des akademischen Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.) im Fach Informatik
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes
von

Katarzyna Bożek

Datum der Einreichung:	19.09.2011
Gutachter:	Prof. Dr. Dr. Thomas Lengauer
Gutachter:	Prof. Dr. Michael Lässig
Datum des Kolloquiums:	08.12.2011
Dekan der Fakultät:	Prof. Dr. Holger Hermanns
Vorsitzender des Kolloquiums:	Prof. Dr. Kurt Mehlhorn
Protokollant:	Dr. Nico Pfeifer

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, 09.12.2011

____

Katarzyna Bożek

CONTENTS

ABSTRACT

The human immunodeficiency virus depends on molecular pathways of the host for efficient replication and spread. The intricate network of host-virus interactions shapes the virus' evolution by driving the pathogen to evade immune recognition and constraining it to maintain its capacity to replicate. Study of the HIV-host interactions provides important insights into viral evolution, pathogenicity and potential treatment strategies. This thesis presents an analysis of HIV-host interactions on several scales, ranging from individual protein interactions to whole genomes.

On the scale of individual interaction we analyze structural and physical determinants of the interaction between host TRIM5α and virus capsid – an interaction of potential therapeutic interest due to the capacity of TRIM5α to block retroviral infections. On the scale of viral population we present two studies of a highly variable region of the virus genome involved in the interaction with host cell coreceptors upon virus cell entry. The studies provide insights into the virus evolution and the physicochemical and structural properties related to its interaction with cellular coreceptors. On the scale of the single cell we develop models of HIV cell entry involving virus, host and environmental factors. The models represent a comprehensive picture of the virus phenotype and allow one to view the variability of virus phenotypes on 2D phenotype maps. On the genomic scale we perform a large-scale analysis of all HIV-host interactions. This study reveals insights into general patterns of the host-pathogen evolution and suggests candidate host proteins involved in interactions potentially important for the infection and interesting for further study on other scales.

Interactions and processes crucial for the HIV infection reemerge across the scales pointing to the importance of integrative, multi-scale studies of host-pathogen biology.

ZUSAMMENFASSUNG

Das Humane Immundefizienz-Virus hängt von molekularen Mechanismen des Wirts für seine effiziente Replikation und Ausbreitung ab. Das komplizierte Netzwerk von Wirt-Virus Interaktionen formt die Evolution des Virus, indem es den Erreger dazu bringt, sich der Erkennung durch das Immunsystem zu entziehen und seine Replikationskapazität aufrecht zu erhalten. Das Studium der HIV-Wirt-Interaktionen erlaubt wichtige Einblicke in die viralen Evolution, die Pathogenität des Virus, sowie mögliche Behandlungsstrategien.

Diese Arbeit stellt eine Analyse der HIV-Wirt-Interaktionen in mehreren Größenordnungen vor, von einzelnen Protein-Interaktionen bis hin zur Analyse ganzer Genome. In Hinblick auf einzelne Interaktionen untersuchen wir strukturelle und physikalische Determinanten der Interaktion zwischen dem Wirtfaktor TRIM5α und dem viralen Kapsid - eine Interaktion, die von therapeutischem Interesse ist wegen der Fähigkeit von TRIM5α, retrovirale Infektionen zu blockieren. In Hinblick auf virale Populationen präsentieren wir zwei Studien einer hochvariablen Region des viralen Genoms, die in der Interaktion des Virus mit zellulären Rezeptoren des Wirts beim viralen Zelleintritt involviert sind. Diese Studien geben Einblick in die virale Evolution und die physikalisch-chemischen und strukturellen Eigenschaften des Virus bezüglich dessen Interaktion mit zellulären Ko-Rezeptoren. Auf der Skala der einzelnen Zelle entwickeln wir Modelle des HIV Zelleintritts welche das Virus, den Wirt und Umgebungsfaktoren berücksichtigen. Diese Modelle bieten ein umfassendes Bild des viralen Phänotyps und erlauben es, die Variabilität des Virus auf 2D-Phänotyp-Karten zu visualisieren. Im genomweiten Maßstab führen wir eine groß angelegte Analyse aller HIV-Wirt-Interaktionen durch. Diese Studie erlaubt Einblicke in allgemeine Muster der Wirt-Pathogen-Evolution und identifiziert Kandidaten für Wirtsproteine, deren Interaktionen potenziell wichtig für die virale Infektion sind und deren weitere Untersuchung in anderen Größenordnungen von Interesse ist.

Interaktionen und Prozesse, die von entscheidender Bedeutung für die HIV-Infektion sind zeigen sich wiederholt in allen untersuchten Maßstäben und unterstreichen die Bedeutung einer integrativen und multi-skalaren Untersuchung der Wirt-Pathogen-Biologie.

ACKNOWLEDGEMENTS

In the first place, I would like to thank my advisor Thomas Lengauer for the invaluable support in realizing projects throughout my thesis, the freedom of pursuing my own research ideas and for forstering the scientific exchange between me and other people of different scientific background.

I would like to thank my wet-lab collaboration partners who provided me with data that made this work possible. Firstly I am grateful to Manon Eckhart from the Heidelberg University for her persistence, hard work and long discussions we maintained during the HIV cell entry project. I would like to thank other people involved in this project Hans-Georg Kräusslich and Barbara Müller at the Department of Virology (Heidelberg), Saleta Sierra and Rolf Kaser (Cologne) for their support, involvement and knowledge sharing. I would like additionally thank Saleta and Rolf for providing data and support in other projects of this thesis, for the discussions we had and for creating this friendly collaboration between our groups. I would like to thank my collaborators from University of Osaka in Japan – Tatsuo Shioda, Emi Nakayama, Ken Kono and Ayumu Kuroishi – for sharing their experimental data and view on the HIV research and for their hospitality during my visit in Osaka.

I would like to thank all the members, past and current, of the group who supported this work with invaluable discussions and feedback – Francisco Domingues, Alice McHardy, Christoph Bock, Alexander Thielen, Oliver Sander, Adrian Alexa. Additionally Ruth Schneppen-Christmann, Joachim Büch and Georg Friedrich deserve a big thank you for their other than scientific support of my work.

I would like to thank the proofreaders of this thesis – Sebastian Krämer, Glenn Lawyer, Jarosław Reguła and Marta Bożek – for their patience in reading and helpful comments. Sven-Eric Schelhorn for the German version of the abstract.

I would also thank all the invaluable people I was lucky to meet and spend time with throughout my time in Saarbücken: Brice Bayer, Alexandria Kimsey, Kaleigh Smith, Gabriele Mayr, Simone Schulze.

Work on this PhD thesis was supported by basic funding of the Max Planck Society and by the grant of the German Federal Ministry of Education and Research (BMBF) within the Medical Systems Biology (MedSys) program.

CHAPTER 1 – Introduction

Viruses cause a wide range of diseases. They can give rise to chronic or acute infection, generate no symptoms or result in death, can be persistent or recur in periodic fashion, spread easily in a restricted fashion and infect any type of tissue. Developments in the field of virology incited a revision of the concept of living matter and provoked questions central to understanding all life processes. Virus replication is dependent on regulatory pathways of the host cellular system and is hampered by effective action of the immune system. Study of viruses contributed to our understanding of molecular biology and the origins of infectious diseases (Knipe, Howley et al. 2006).

With the development of technologies for acquisition of molecular data, virology has shifted from a descriptive, observation-based to an integrative and quantitative field of research based on large amounts of measured viral data. Technologies such as DNA microarrays, phenotypic resistance testing, small interfering RNA screens, or deep sequencing underscore the need for computational methods and models capable of handling and integrating the large-scale virus data. Data analysis and modeling advance virology research and support treatment of infectious diseases.

This thesis presents several computational studies of the human immunodeficiency virus (HIV). Basing on a range of methods and data, virus interactions with the host are examined on different scales – from specific proteins to entire genomes, from molecular interactions to the spread of disease. The analyses presented in this thesis contribute to our knowledge about the HIV-host interactions. The results help our understanding of the critical HIV-host interactions involved in blocking the infection, explain the evolution and determinants of interaction phenotypes having an impact on the disease progression and targeted by the anti-HIV treatment, and point to new candidate interactions for further investigation. This chapter presents a brief introduction into the HIV[*] biology and computational approaches in HIV research. A detailed outline of the thesis is provided at the end of this chapter.

1.1 The unique nature of HIV

The isolation of a T lymphotropic retrovirus in 1983 (Barre-Sinoussi, Chermann et al. 1983) and in 1984 (Montagnier, Gruest et al. 1984) from a lymph node of an individual with lymphadenopathy marked the beginning of three decades of intense research on the pathogenesis of the virus causing the Acquired Immune Deficiency Syndrome (AIDS). The disease that was first recognized in patients in the USA in 1981 was indicated by a wide array of deleterious effects that HIV was causing in every component of the host immune system. Further research allowed for recognizing the mode of virus

[*] Unless otherwise stated, HIV is used for HIV type 1 (HIV-1) throughout this thesis.

transmission, its origin in African primate species (Sharp, Bailes et al. 2000), the effect of the virus on the immune system and its interactions with the host (Ptak, Fu et al. 2008). Despite the advances in our understanding of HIV pathogenesis and immunology after the discovery of the virus, a protective vaccine or an effective cure of the disease remain elusive. Here we present several virus properties that contribute to the specific nature of the HIV infection. The description is incomplete, and is only designed to give a background and motivation for the problems addressed further in this thesis.

1.1.1 HIV is a retrovirus

HIV is a member of the genus *Lentivirus* in the *Retroviridae* family. Retroviruses are so called because their genome is encoded in RNA that is transcribed into DNA using the viral enzyme reverse transcriptase (RT) after the virus entry into a host cell. This DNA is then transported to the cell nucleus and integrated into the cellular genome. Figure 1.1 illustrates the lifecycle of HIV.

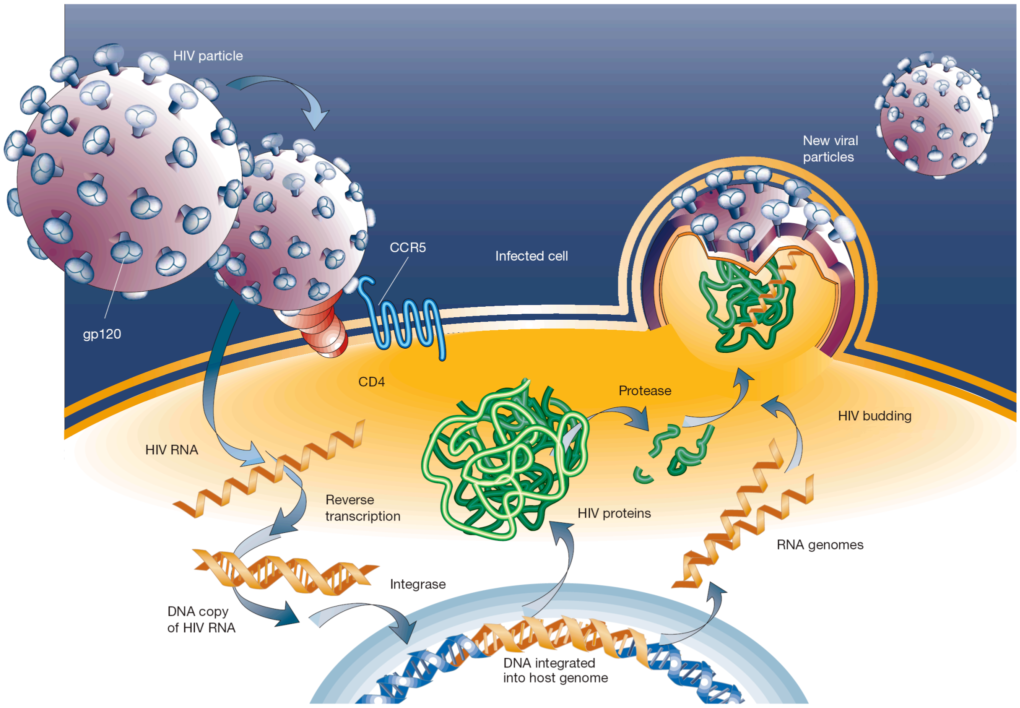


Figure 1.1 HIV life cycle involving cell entry, reverse transcription, replication and budding. Adapted from (Weiss 2001).

A retrovirus, once retro-transcribed into the genome of its host is copied with each host cell division. In effect the HIV infection is life-long. Integrated silent copies of the viral DNA are rapidly established in cellular reservoirs of various tissues and persist throughout the host life. An archive of forms of the virus that evolved thoughout the infection is retained in the host DNA and these forms on occasion can re-emerge (Korber, Theiler et al. 1998; Shankarappa, Margolick et al. 1999; Korber, Muldoon et al. 2000) presumably under changing selection pressures in the host exerted by drugs and the immune system. The integration of the HIV genome into the genome of the host represents an important impediment to achieving complete eradication of the virus in infected individuals, which points to vaccines for preventing infection or to cellular mechanisms blocking viral replication as ways for suppressing the spread of the virus.

1.1.2 HIV replicates in the immune system cells

The hallmark of the HIV infection is its destructive effect on the host immune system. Following transmission, HIV disseminates rapidly in the absence of preexisting immune pressures leading to a burst of viremia and a rapid depletion of CD4$^+$ T cells in the initial phase of the infection (Figure 1.2). After this acute phase host immune response causes the viral load to rapidly decline and reach a steady state termed the *virus set point* (Little, McLean et al. 1999; Fiebig, Wright et al. 2003) which marks the beginning of the chronic phase of the infection. While CD4$^+$ T cell counts rebound to their initial level, the immune system remains chronically activated (Goonetilleke, Liu et al. 2009) due to the virus replication in the lymphoid tissue. By targeting the major constituent of the immune system, HIV gradually destroys and dysregulates its functionality.

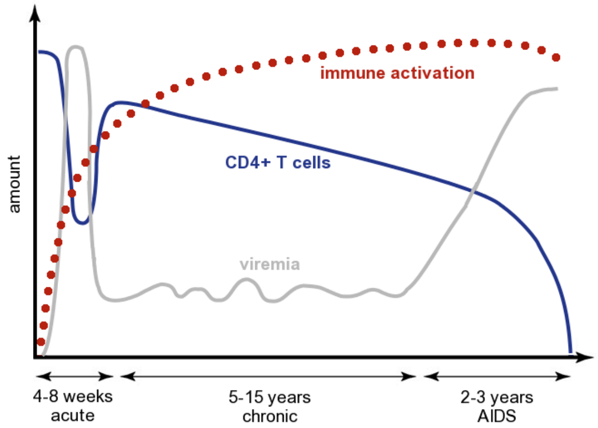


Figure 1.2 Progression of a typical HIV infection. After the initial acute phase characterized by a strong immune response and rapid virus replication comes the chronic phase when the virus can remain undetected for many years. AIDS is characterized by a high immune activation, low number of CD4$^+$ T cells and high virus replication (Forsman and Weiss 2008).

As immune cells are activated they become targets for HIV, so an active immune system paradoxically contributes to a higher viral load. A pool of productively infected cells is constantly replenished and the continuous exposure to viral antigens during the chronic phase of infection leads to T and B cell exhaustion and ultimately to a broad and severe dysfunction of immune responses (Trautmann, Janbazian et al. 2006). Although the underlying causes of HIV-induced cellular hyperactivity remain debatable, the pathogenic consequences of such activity are fairly well understood and include increased cell turnover leading to cellular exhaustion, senescence and low renewal potential. In the absence of antiretroviral treatment (ART) the vast majority of HIV-infected individuals experience progressive loss of CD4$^+$ T cells, increased immunodeficiency leading to AIDS and subsequent development of opportunistic infections and cancers. With the increased availability of ART, there has been a dramatic reduction in the number of people who progress to AIDS. Yet, long-term pathogenic consequences of living with HIV persist in most individuals due to incomplete immunologic recovery and the loss of the immune system's ability to replenish itself. HIV is also known to induce immunologic dysfunction of other immune cells such as CD8$^+$ T cells B cells and natural killer (NK) cells through mechanisms that include increased cell turnover, activation, differentiation and response to infection.

Replication in the immune system cells is not a property unique to HIV. Another human retrovirus, human T-lymphotropic virus type I (HTLV-1) also infects CD4$^+$ lymphocytes, however it does not have the same detrimental effects on the target cells (Jeang 2010). The persistent replication and spread of HIV (Perelson, Essunger et al. 1997) result in the progressive loss of immune system cells. In contrast, HTLV establishes a continuous culture in the target cells not causing their increased turnover. The molecular basis of the chronic activation of the immune system in the pathogenic HIV infection is still not known.

1.1.3 High mutation rate of HIV

HIV is characterized by an exceptionally high genetic diversity. An HIV infection starts out with a homogeneous viral population (Zhang, MacKenzie et al. 1993; Wolinsky, Korber et al. 1996). Over its course, however viruses showing more than 10% mutated DNA bases in the envelope gene arise (Lukashov, Kuiken et al. 1995; Wolinsky, Korber et al. 1996; Shankarappa, Margolick et al. 1999).

The lack of a proof-reading mechanism in the viral RT, and the consequential high error rate (0.2–2 mutations per genome per cycle) (Drake 1993) can explain only part of this diversity. For example, HTLV-1, like HIV, goes though a reverse transcription step, but is far less variable. This has been previously attributed to differences in the dynamics of the two viruses (Wodarz and Bangham 2000). A high replication rate accompanied by rapid viral turnover (Perelson, Essunger et al. 1997), as well as pressure for adaptation and host recognition evasion might be additional forces driving virus genomic change. Beyond the immune escape that can drive selection of specific HIV epitopes (Borrow, Lewicki et al. 1997; Goulder, Sewell et al. 1997; Price, Goulder et al. 1997; Allen, O'Connor et al. 2000) neutral mutations and genetic drift can also contribute to the overall diversity of the virus (Sala and Wain-Hobson 2000; Yang, Nielsen et al. 2000). In addition to base-substitutions, HIV is subject to recombination and relatively large insertions and deletions (indels), which rapidly generate radically divergent forms.

Consequently, instead of a single virus clone, an individual is infected with a virus population termed viral quasispecies. The concept of quasispecies assumes that the prevalence of viral clones in a population is dictated by their relative fitness (Eigen 1996). Under circumstances of selective pressure, such as therapy (Condra 1998) or immune pressure (Price, Goulder et al. 1997; Poignard, Sabbe et al. 1999; Allen, O'Connor et al. 2000), the clone profile in the viral population can shift.

The high variability of HIV represents a major challenge for the immune response and development of an effective vaccine. The challenge is even greater than that presented by influenza, which is also antigenically variable. A comparison of HIV and influenza A evolution reveals very different patterns. HIV evolution is characterized by a radial spread outward from an ancestral node, while influenza is characterized by bottlenecks and global drift from year to year (Grenfell, Pybus et al. 2004). Within a single nine-month flu season, little variation is typically found between geographically distinct influenza isolates after the emergence of the epidemic strain. HIV, on the other hand, shows increasing genetic diversity within a population through time. The diversity of influenza sequences world-wide in any given year appears to be roughly comparable to the diversity of HIV sequences found within a single infected individual at one time point (Korber, Gaschen et al. 2001). Thus, while influenza does have a relatively fast rate of mutation when measured over decades, the vaccine for any given year is targeted towards a relatively homogeneous viral population. Even a small number of changes in the viral amino acid sequence at antigenic sites requires a change in the vaccine strain to induce an immunologically appropriate response for currently circulating strains. This small number of amino acid substitutions that result in a loss of influenza vaccine efficacy indicates the challenge in developing an effective HIV vaccine. The viral diversity which an HIV vaccine must counter is far greater than the diversity countered by the influenza vaccine which underscores the necessity of innovative methods of HIV prevention.

1.1.4 HIV establishes a chronic infection

Most human respiratory and enteric viruses cause short but acute infections. The pathogen multiplies mainly at and around its portal of entry; induces transmission-promoting symptoms such as sneezes, coughs, or diarrhea; kills its target cells; is countered by innate immune responses and cleared by adaptive immune responses that protect against reinfection with the same viral strain (Knipe, Howley et al. 2006).

In contrast, HIV belongs to class of human viral pathogens that are less transmissible but induce persistent rather than acute infections. This class includes all herpes virus family members, the hapatotropic hepatitis B and C viruses (HBV and HCV), the tumorigenic papilloma and polynoma viruses and retroviruses such as HTLV or HIV. Although diverse in nature and in replicative properties, all of these viruses share an ability of escaping adaptive immune responses, affording persistence in the host. How the virus achieves this, however, varies greatly. Herpes simplex virus (HSV) escapes clearance by establishing latency within cells of the central nervous system and undergoes only sporadic reactivation (Efstathiou and Preston 2005). In contrast, HCV does not become dormant but shields from and antagonizes immune responses (Guidotti and Chisari 2006). With HBV age at the time of infection is determinant: infants typically develop chronic infections because their immature immune system is overly tolerant to the virus, whereas infection later in life is spontaneously cleared in 95% of cases (Guidotti and Chisari 2006).

HIV always induces a persistent infection irrespective of the age and immune status of the host. The virus becomes latent in a fraction of infected cells, yet never stops replicating in others. Together, these features allow for viral persistence despite ART capable of keeping viremia undetectable for years (Chun, Nickle et al. 2005). It is this combination of the characteristics of a retrovirus rapidly mutating and replicating in the immune system cells which make HIV a unique pathogen particularly difficult to eradicate and treat. The ineffectiveness of HIV vaccine trials (Barouch 2008) calls for novel approaches in the search for the cure of which computational analyses and mathematical models are an essential part.

1.2 Computational approaches to understanding and treating the HIV infection

Almost three decades of HIV research and the rapid development of data acquisition technologies resulted in an accumulation of large amounts of molecular and clinical data on HIV. Mathematical and bioinformatics tools are of invaluable support not only for the storage and integration of the growing volume of information but also by providing predictive and explanatory models in situations where biological knowledge is inaccurate, missing or cannot be directly inferred. In this section we review several computational studies of high importance for HIV research. These example studies represent a restricted part of the broad field of HIV research. However, they illustrate how each of the characteristics of the virus biology described in the previous section can be analyzed in computational studies on different scales – from individual proteins to interspecies genomic comparisons. Additionally, the work presented in this thesis was partially motivated by several shortcomings of these studies summarized in the last part of this section.

1.2.1 Host restriction factors

Defence against a pathogen begins at the first virus-cell interaction. A mechanism termed *intrinsic resistance* implies immediate cellular responses to the infection acting before the innate and adaptive immune responses are triggered. Some of the intrinsic immunity processes involve so called *restriction factors* – proteins constitutively expressed in some cell types or induced as a part of the innate immune response in others (Bieniasz 2004). The proteins of this front line of antiviral defense exhibit substantial genetic variation among species due to selection pressures of ancient pathogens. HIV is a new pathogen in the human population, therefore human restriction factors fail to prevent the spread of this recently acquired retrovirus. Yet the interspecies variation of host restriction factors and their species-specific responses to lentiviruses suggest that they represent a potential prevention mechanism for the HIV infection in humans.

Several studies analyzed the evolutionary history of the host restriction factors known to block lentiviruses in primates: tripartite motif 5-α (TRIM5α) (Stremlau, Owens et al. 2004; Sawyer, Wu et al. 2005), apolipoprotein B editing complex 3G (APOBEC3G) (Bieniasz 2004; Sawyer, Emerman et al. 2004) and tetherin (BST-1, CD317) (Neil, Zang et al. 2008; McNatt, Zang et al. 2009). Via interactions of still unknown nature with the viral capsid, and the viral proteins Vif and Vpu, respectively, these factors account for resistance to lentiviruses transmitted from other primate host species (Malim and Emerman 2008). The comparative analyses of primate sequences of each of the three restriction factors (Sawyer, Emerman et al. 2004; Sawyer, Wu et al. 2005; McNatt, Zang et al. 2009) pointed to positive selection acting on these proteins and predating the origin of lentiviruses. Statistical models of positive selection (Yang, Wong et al. 2005) identified specific sequence regions showing high levels of positive selection that might form points of physical contacts with the respective viral proteins.

Sequence analyses of individual proteins participating in an early defense can point to specific domains involved in the interactions blocking the infection. Even though the role of these domains in the virus restriction can be tested through mutagenic studies (Sawyer, Wu et al. 2005) such sequence analyses do not provide insights into the actual mechanism of interaction. Recognizing the structural and physicochemical determinants of the interaction of a host protein with a highly mutating virus protein necessitates studies of both interaction partners expanding beyond the sequence analysis.

1.2.2 Prediction of HIV drug resistance

Several drug classes have been developed which target different stages of the HIV replication cycle (Nicol and Kashuba 2010). High mutation rates of HIV allow the virus to evolve resistance to any single drug. Combination therapies known as *highly active antiretroviral therapy* (HAART) were introduced after it became apparent that no single drug could be expected to achieve durable viral suppression and clinical benefits. Combination therapies proved more effective in the long term as multiple mutations required for resistance to three or more drugs are harder for the virus to acquire (Colgrove and Japour 1999).

Virus resistance to antiretroviral drugs can be quantified in experimental phenotypic tests. In such tests the replication of the patient’s virus is compared to the replication of a reference wild type strain in a cell culture in the presence of a varying concentration of a given drug. In the experiment, resistance only to a single drug can be measured, it has to be performed in specialized laboratories, its costs are elevated and the results are only available within weeks. In constrast, genotyping of patient viral samples is a standardized and affordable measurement that is performed routinely in any type of laboratory. Accessibility and utility of genotypic testing motivated introduction into clinical practice of genotype-based methods for prediction of drug resistance.

The increasing number of anti-HIV drugs and their possible combinations contribute to the complexity of regimens and the pathways of the virus evolution to resistance. Computational systems replace currently the manually compiled resistance mutation tables used to derive successful treatment strategies (Lengauer and Sing 2006). These systems base on genotype of an individual virus or a virus population (Altmann, Däumer et al. 2009) and apply statistical learning methods to predict viral resistance to an individual compound measured in phenotypical tests *in vitro* (Beerenwinkel, Däumer et al. 2003) or their combinations observed *in vivo* (Larder, Wang et al. 2007). With the concerted effort to accumulate data on virus resistance and treatment outcome, computational models for predicting virus response to drug regimens become both essential and accurate. Flexibility of the HIV genome results in countless possible mutation pathways and genetic variants among which resistance patterns are hard to discern without the assistance of mathematical tools.

The use of viral resistance models in clinical practice underscores the importance of genotypic resistance analyses. These analyses are based the viral genotype only and do not currently consider additional host and environmental factors which might also play role in the effectiveness of the treatment (Tozzi, Libertone et al. 2008). Although HAART is highly effective, critical questions remain about how patients respond to treatment: genetic and environmental variation among individuals may cause considerable variability in drug pharmacokinetics and pharmacodynamics. In particular, since anti-HIV medicines are used for life, even modest differences in susceptibilities are important. Consequently, identification of host parameters that play role in the HAART could allow for tailoring of the drugs to minimize the long-term toxicities and result in not only virus- but also patient-targeted therapies.

1.2.3 Studies of virus populations

Ever since the first sequencing of the DNA genome of coliphage phiX174 by Fred Sanger in 1977, sampling, sequencing and computer technologies provided the means to identify and sequence entire viral communities without the intervention of costly techniques of isolation and characterization of individual viruses (Wang, Urisman et al. 2003; Edwards and Rohwer 2005; Suttle 2005). Insights into the diversity of virus populations in an environment are essential as studies of isolated viruses in the laboratory do not provide information necessary to accurately assess virus impact on human health.

Recent development of next generation sequencing technologies (Metzker 2010) allowed for study of virus populations on a yet unprecedented level of detail. Next-generation or deep sequencing platforms are now capable of reading millions of base pairs in a more cost-effective and faster way than traditional Sanger sequencing. This technology has already been used to address such topics as resistance mutations in minority populations (Hoffmann, Minkah et al. 2007; Wang, Mitsuya et al. 2007; Johnson, Li et al. 2008; Simen, Simons et al. 2009), evolution of virus population (Poon, Swenson et al. 2009) or coreceptor usage of entire within-patient populations (Archer, Braverman et al. 2009).

The forces shaping virus quasispecies cannot be explained based on virus sequences exclusively. Recognizing the determinants of virus population structure and evolution is not possible without considering host and environmental factors having potential impact on the virus. This calls for integrative studies of not only virus but also host and environment variability and their mutual interactions.

1.2.4 Phylogenetics of lentiviruses

Despite the lack of ancient virus sequences and epidemiologic data on lentivirus spread among human and monkey species in Africa, much insight into the evolution of the virus can be gained through comparative genomic analyses of contemporary lentivirus sequences. Mathematical methods such as maximum likelihood (Felsenstein and Churchill 1996) or Bayesian (Drummond, Rambaut et al. 2005) models were applied to the estimation of virus phylogenies (Van Heuverswyn, Li et al. 2007) and the rate of its evolution (Korber, Muldoon et al. 2000; Wertheim and Worobey 2009). These studies allowed not only to trace the origin of human immunodeficiency viruses (HIV-1 and HIV-2) in the African primates (Sharp, Bailes et al. 2000) but also to estimate the approximate date of virus transmission to humans at the beginning of the century (Korber, Muldoon et al. 2000).

Phylogenetic and molecular clock studies show how computational models can generate hypotheses about the missing knowledge based on the incomplete data. Comparative genetic analyses on the interspecies genomic scale allow for building models of the virus past based on the contemporary virus data. Even though dating of the virus transmission generated hypotheses on the social and economic factors that facilitated the virus spread (Hahn, Shaw et al. 2000), purely phylogenetic analyses give limited information on the biological determinants of the virus interspecies transmission, on the factors and means of a pathogen adaptation to a new host species. An integrative approach of analyzing both host and virus factors determining the infection affords means of explaining the specific transmission pathways of lentiviruses and of providing knowledge potentially important for future prevention of virus transmissions (Garten, Davis et al. 2009).

1.2.5 Aspects addressed in this thesis

The studies described above address different aspects of the host (host restriction factors) and virus (drug resistance, virus population structure, origins of HIV) playing a role in the infection. As critical as the studies are in recognizing, classifying and predicting patterns in the host and virus evolution separately, they provide limited insight into their interdependence and underlying reciprocal host-pathogen mechanisms.

Viruses express only few proteins essential for their replication, spread and escape from host recognition (Figure 1.3). Having no metabolism of their own, viruses are obliged to invade cells and make use of the cellular machinery of the host, subverting it to their own purposes. The replicative cycle of the HIV is largely known. It comprises elaborate interactions between the viral genome, viral proteins and the molecular processes of the host (Knipe, Howley et al. 2006).

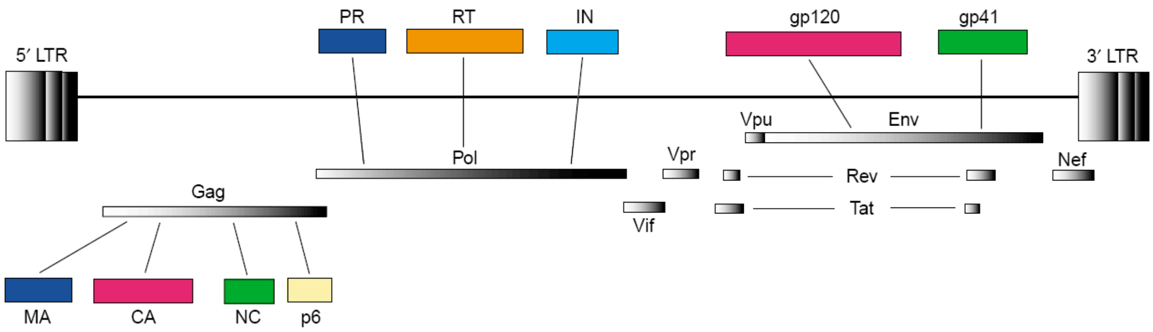


Figure 1.3 Organization of the HIV genome. Gray-shaded proteins Gag, Pol and Env are initially synthesized as polyprotein precursors, cleaved into mature proteins represented by colored boxes. From (Freed 2004).

Drug treatment is only one among many factors shaping HIV evolution. The virus is involved in a constant struggle to escape recognition by the host immune system while simultaneously being subject to functional constraints ensuring its efficient replication and transmission. As an example of immune system influence shaping the evolution of the virus, human leukocyte antigen B (HLA-B) alleles were recognized that are associated with slow disease progression (Heeney, Dalgleish et al. 2006; Fellay, Ge et al. 2009). Such alleles offer protection by binding to highly conserved HIV proteins that are essential to the virus function and mutation of which comes at a cost of the virus fitness. Less protective alleles supposedly bind to pieces of proteins that HIV can change by mutation without affecting its replication capacity. Studies of the host and virus interactions can result in a more complete view of the mechanisms underlying disease progression, virus evolution and spread.

In addition to understanding forces shaping virus evolution, the study of host factors involved in the infection can provide insights into new avenues of treatment and prevention based on host proteins critical for the HIV infection. First drugs targeted at the host proteins – drugs blocking C-C chemokine receptor type 5 (CCR5) one of the coreceptors used by HIV in the cell entry (Chan and Kim 1998) – have been recently introduced into clinical use (Dorr, Westby et al. 2005). Drugs targeted at host factors crucial for the virus replication could potentially be more efficient than those targeted at the virus as they would require the virus to evolve new interaction mechanisms instead of modifying drug attachment sites. This highlights the importance of developing models of HIV-host interactions in addition to the current HIV drug resistance analyses.

Integrated study of host-pathogen interactions allows for the paired recognition of host defense components and viral gene products that counter them. The variety of interactions in which the host and pathogen are involved throughout the course of infection includes virus restriction by the host restriction factors (Malim and Emerman 2008), virus binding to cellular membrane proteins upon cell entry (Chan and Kim 1998), virus down- or up-regulation of host immune system (Kirchhoff 2009) or virus binding to antigens (Barouch 2008). Investigating the combined host and pathogen responses might therefore improve our understanding of the pathogenesis of HIV in humans and provide insights into potential diagnostic and therapeutic targets for viral infections.

1.3 Thesis outline

The high complexity of the HIV interactions with the host can be viewed synoptically by integrating studies at different scales, ranging from individual interactions to interspecies comparisons of many interactions. In this thesis we approach the analysis of HIV-host interactions on four distinctive scales.

In chapter 2, we present an analysis on the scale of a specific interaction – the interaction between host protein TRIM5α and the virus capsid. We perform an in-depth analysis of the structural and physical features of both interaction partners in the search of determinants of the restriction. The study is based on the experimental data collected in the lab of Prof Tatsuo Shioda at the Department of Viral Infections, Research Institute for Microbial Diseases at Osaka University in Japan. The importance of this analysis on the individual interaction scale is underlined by the capacity of TRIM5α to block retroviral infection and the potential impact of this interaction in prevention of the HIV infection.

In chapter 3 we present two studies on the virus population scale of a highly variable region of the virus genome that is involved in the interaction with host cell coreceptors in the process of cell entry. The studies implicitly include the host partner of the interaction by distinguishing between virus variants that interact with a specific coreceptor. Both studies expand beyond sequence-based classification models and provide insights into the virus evolution driven by or resulting from the interaction with the coreceptor as well as the physicochemical and structural properties involved in it.

In chapter 4 we describe a study of the viral cell entry process on the single cell level. The model of HIV cell entry developed in this study includes virus, host and environmental factors playing role in the entry. It is a comprehensive analysis of a specific interaction based on a high-throughput experimental data generated in the lab of Prof. Dr. Hans-Georg Kräusslich and Dr. Barbara Müller at the Department of Virology, Heidelberg University and based on clinical samples provided by the lab of Dr. Rolf Kaiser, Institute of Virology, University of Cologne.

In chapter 5 we perform a large-scale analysis of all HIV-host interactions on the genomic scale. We compare evolutionary patterns of interacting host and pathogen proteins in search of those that might be relevant for the infection and investigate the evolutionary patterns of the interacting host and pathogen proteins. This large-scale analysis presents candidate proteins potentially interesting for further study on other scales and insights into general patters of the host-pathogen evolution.

Each of the studies presented in this thesis relate to different aspects of the HIV biology and its interaction with the host. Every chapter contains therefore an explanatory introduction into the biological aspects and the scale of analysis.

CHAPTER 2 – Individual interaction scale: structural analysis of TRIM5α-CA interaction

In this part of the thesis we present an analysis on the individual interaction scale of the interaction between host TRIM5α and virus capsid. The importance of this interaction manifested in the capacity of TRIM5α to block retroviral infections call for its better understanding. Work presented in this chapter was performed in collaboration with the experimental group of Prof Tatsuo Shioda at the Department of Viral Infections, Research Institute for Microbial Diseases at Osaka University in Japan. We analyzed computationally data from a cellular assay measuring TRIM5α viral restriction elaborated by Dr. Ken Kono, Dr. Kuroishi and Dr. MD Emi Nakayama.

2.1 Background

Persistently infected species evolved a range of cellular defense mechanisms against pathogens. The evolution of these mechanisms has been driven by ancient infections to which the host species were exposed in the past. The infection history varies among species which can have an effect on host response to current infections.

The term *intrinsic immunity* has been assigned to the activity of specific antiviral factors that have evolved as a result of host exposure to pathogens in humans and other mammals. This kind of immunity has an advantage over the innate and acquired immune responses as it does not need to develop the capacity to combat viruses but is already active at the first virus-cell interaction. The host restriction factors being a part this type of immunity are sometimes constitutively expressed and provide an intrinsic pre-mobilized system of defense against retroviral infection (Bieniasz 2004). The importance of restriction factors in the virus response is underscored by their genetic variability accumulated as a result of species exposure to different pathogens and by the complex mechanisms the viruses have evolved in order to evade the antiviral activity of these proteins (Neil and Bieniasz 2009). As a result of the genetic variability of the host restriction factors, their capacity to restrict modern lentiviruses among primates is species-specific. As an example, HIV-1 efficiently enters the cells of Old World monkeys, but encounters a block before reverse transcription caused by an interference with uncoating of the retroviral capsid (CA) (Stremlau, Owens et al. 2004). The understanding of the interaction mechanisms conditioning specific lentivirus restriction by primate restriction factors is still incomplete. Such knowledge however could open ways for novel forms of therapeutic intervention in pathogenic retroviral infections, as well as the development of animal models of human disease, which advocates detailed studies on the scale of individual interaction of the interactions between restriction factors and respective lentivirus proteins.

An example of a retroviral restriction factor showing different restriction patterns against lentiviruses in primates is the tripartite motif (TRIM) 5. TRIM5 was shown to be responsible for various retrovirus restriction phenotypes that can be observed in human and nonhuman primate cells (Besnier, Takeuchi et al. 2002; Stremlau, Owens et al. 2004; Towers 2005). The most studied antiviral TRIM protein, TRIM5α is a retroviral restriction factor that targets the early steps of cellular infection. It recognizes the viral capsid (CA) and promotes premature virus disassembly before its reverse transcription in the host cell (Stremlau, Owens et al. 2004). The TRIM5α-CA interaction is species-specific: human TRIM5α has limited efficacy against HIV-1, whereas some primate TRIM5αs can potently restrict HIV-1.

The TRIM family is a large family of proteins that are characterized by a structure comprising a RING domain, one or two B-box domains and a predicted coiled-coil region (Reymond, Meroni et al. 2001). Analyses of TRIM5α variation among primate species identified three variable regions (V1, V2 and V3) within the B30.2 (SPRY) domain that are the major determinants of its anti-HIV-1 potency (Figure 2.1) (Stremlau, Owens et al. 2004). The genetic composition of these regions was shown to be reflected in the protein restriction specificity of different viruses (Ohkura, Yap et al. 2006). The variable regions in the SPRY domains of primate TRIM5α might have evolved independently to recognize various retroviruses, which resulted in the differing restriction phenotypes against modern immunodeficiency viruses. In the absence of a crystal structure of TRIM5α, and in particular its SPRY domain, the molecular mechanism behind these phenotypic differences remains unclear. Nonetheless, we have a rudimentary understanding of the molecular details by which TRIM5α proteins block retroviral infection. Specifically, a C-terminal domain of the TRIM5α protein directly recognizes the incoming viral CAs (Sebastian and Luban 2005; Stremlau, Perron et al. 2006) and thereby governs antiretroviral specificity (Perez-Caballero, Hatziioannou et al. 2005; Stremlau, Perron et al. 2005; Yap, Nisole et al. 2005; Perron, Stremlau et al. 2006), while a central coiled-coil drives TRIM5α multimerization, which is essential for inhibition.

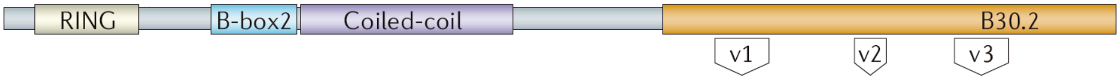


Figure 2.1 TRIM5 protein domains. Three variable regions (V1, V2, V3) in the B30.2 (SPRY) domain are indicated.

The identification of TRIM5α as the protein responsible for the resistance of certain cells to HIV stimulated studies of its evolution and mechanism of action. In this part of the thesis we describe an in-depth analysis of the structural and electrostatic determinants of the TRIM5α-CA interaction. Based on comparative analyses of protein structures and of the electrostatic potential on the surfaces of different variants of both TRIM5α and CA that were experimentally verified for their interaction phenotype, we aimed to identify the sequence and structure determinants of their interaction. The importance of the TRIM5α-

CA interaction for the infection points to in-depth studies on the scale of an individual interaction as a potential means of understanding this restriction mechanism.

2.2 First interaction partner – mutations in the TRIM5α SPRY domain resulting in different protein restriction capacities

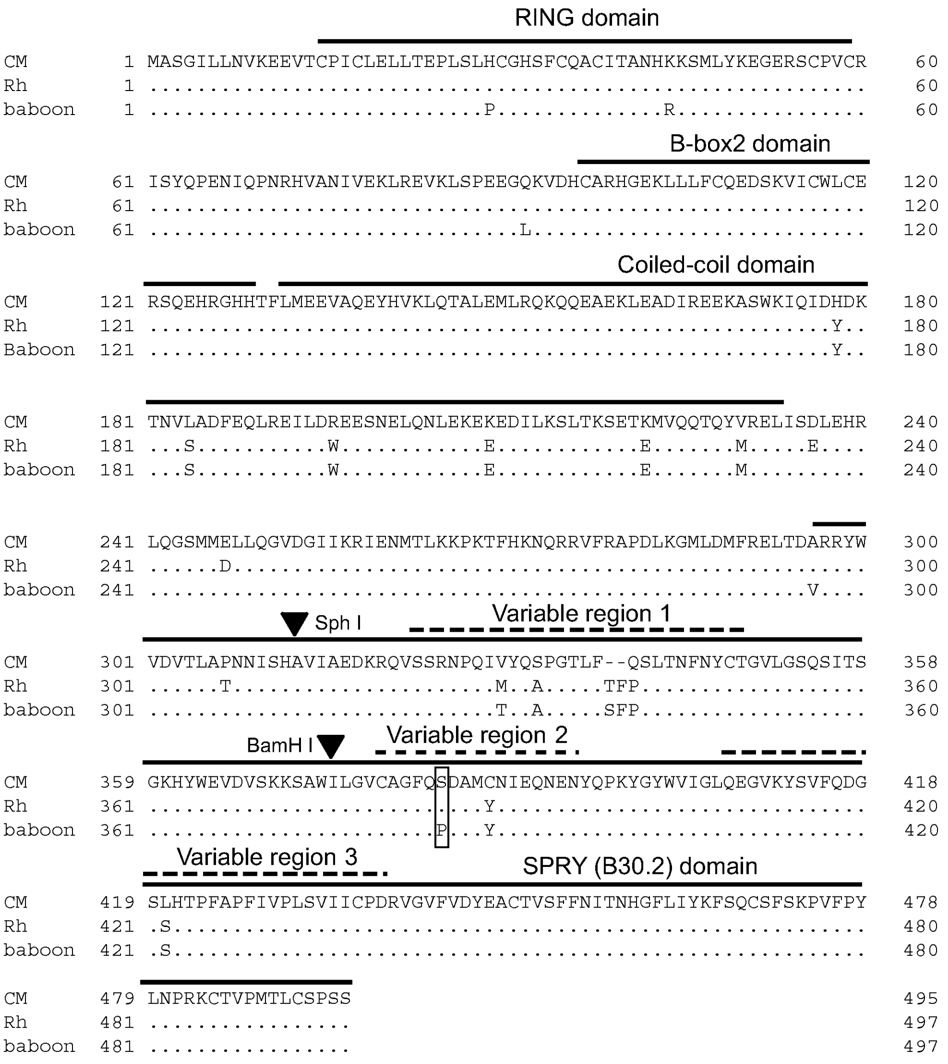
The results of the study described in this section are published in: Kono, K., Bozek, K., Domingues, FS., Shioda, T., Nakayama, EE. Impact of a single amino acid in the variable region 2 of the Old World monkey TRIM5alpha SPRY (B30.2) domain on anti-human immunodeficiency virus type 2 activity. *Virology.* 2009 May 25;388(1):160-8. In this work we performed structural analysis of several chimeric TRIM5αs experimentally tested for their ability to restrict various HIV-1 and HIV-2 CA variants. The analysis was focused on the protein's SPRY domain, given the evidence that this region is the major determinant of the protein restriction specificity (Perez-Caballero, Hatziioannou et al. 2005; Stremlau, Perron et al. 2005). A 17-amino acid region of the African green monkey (AGM) TRIM5α SPRY domain (Nakayama, Miyoshi et al. 2005) as well as position 322 of the human (Stremlau, Perron et al. 2005; Yap, Nisole et al. 2005), 386 and 389 of the orang-utan and gorilla TRIM5α (Ohkura, Yap et al. 2006) were shown to be important for inhibiting several HIV and SIV species.

Previous experiments in the lab of Prof. Shioda testing rhesus monkey TRIM5α (TRIM5α-rh) and cynomolgus monkey TRIM5α (TRIM5α-cm) pointed to a broad spectrum of TRIM5α-rh restriction of HIV-2 as compared to TRIM5α-cm (Kono, Song et al. 2008). A chimeric study of the two proteins showed that the variable region 1 (V1) of the SPRY domain of TRIM5α-rh was determinant for the restriction. Additionally a baboon TRIM5α (TRIM5α-bab) V1 was tested in the TRIM5α-cm background and showed HIV-2 restriction capacity despite the fact that baboons are sensitive to HIV-2 infection (Barnett, Murthy et al. 1994; Locher, Barnett et al. 1998; Locher, Witt et al. 2001). A possible explanation for this discrepancy is that variable region 2 or 3 (V2 or V3) of SPRY domain also contributes to anti-HIV-2 activity of the protein.

2.2.1 Experimental results

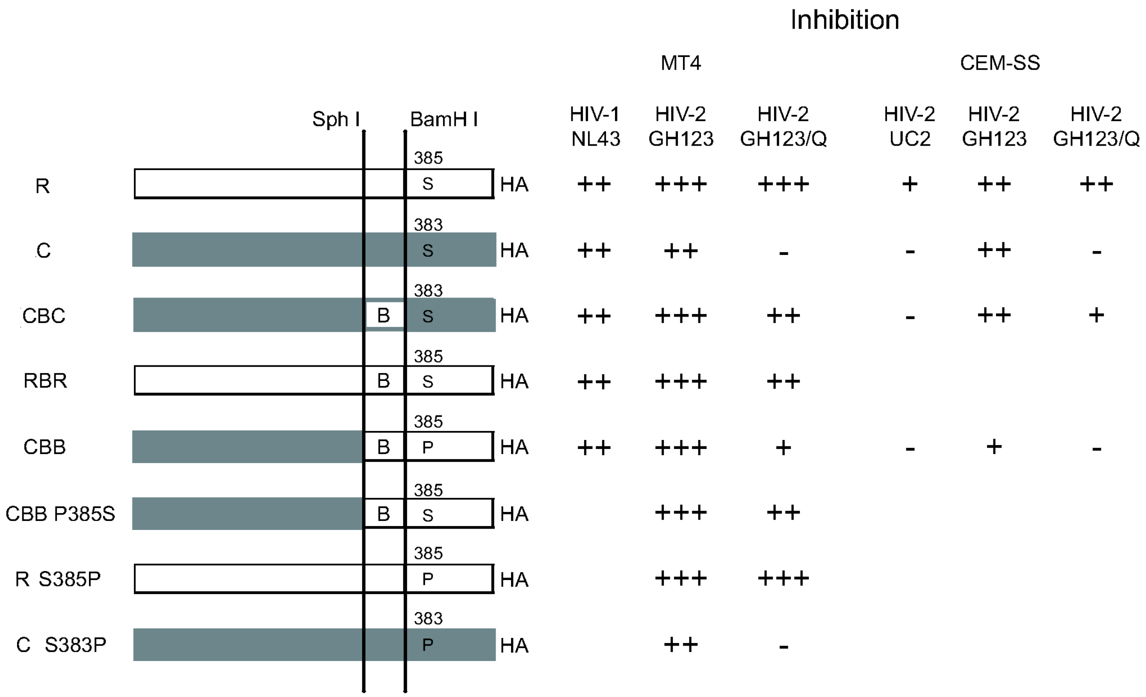
In this study the regions V2 and V3 of TRIM5α SPRY domain were examined for their role in the HIV-2 restriction. Chimeric TRIM5αs composed of parts of the protein of three monkey species were constructed. The TRIM5α chimera was produced using a Sendai virus (SeV) expressing chimeric TRIM5αs composed of protein parts obtained using *Sph* and *BamH* restriction enzymes (Figure 2.2). The restriction enzymes separated the protein into: the N-terminal fragment containing RING, B-box2, and coiled-coil domains, the central fragment containing V1 of the SPRY domain, and the C-terminal fragment containing V2 and V3 of the SPRY domains. To examine the roles of V2 and V3 region of the SPRY domain of TRIM5α in the HIV-2 restriction chimeric proteins were constructed composed of parts of cynomolgus monkey (C), baboon (B) and rhesus macaque (R) TRIM5α (Figure 2.3). Those TRIM5α constructs were tested for their ability to restrict X4-tropic HIV-1 strain NL4-3 and HIV-2 strains GH123 and GH123/Q. Strain

GH123/Q is an HIV-2 GH123 strain with a glutamine on the 120th CA position. This mutation has been previously shown (Kono, Song et al. 2008) to affect the TRIM5α restriction. Viruses were tested in two human T cell lines: MT4 and CEM-SS.



```
CM        1 MASGILLNVKEEVTCPICLELLTEPLSLHCGHSFCQACITANHKKSMLYKEGERSCPVCR   60
Rh        1 ............................................................   60
baboon    1 ............................P..............R................   60

CM       61 ISYQPENIQPNRHVANIVEKLREVKLSPEEGQKVDHCARHGEKLLLFCQEDSKVICWLCE  120
Rh       61 ............................................................  120
baboon   61 ...............................L............................  120

CM      121 RSQEHRGHHTFLMEEVAQEYHVKLQTALEMLRQKQQEAEKLEADIREEKASWKIQIDHDK  180
Rh      121 .........................................................Y..  180
Baboon  121 .........................................................Y..  180

CM      181 TNVLADFEQLREILDREESNELQNLEKEKEDILKSLTKSETKMVQQTQYVRELISDLEHR  240
Rh      181 ...S...........W............E............E.......M.....E....  240
baboon  181 ...S...........W............E............E.......M..........  240

CM      241 LQGSMMELLQGVDGIIKRIENMTLKKPKTFHKNQRRVFRAPDLKGMLDMFRELTDARRYW  300
Rh      241 ......D.....................................................  300
baboon  241 .......................................................V....  300

CM      301 VDVTLAPNNISHAVIAEDKRQVSSRNPQIVYQSPGTLF--QSLTNFNYCTGVLGSQSITS  358
Rh      301 ......T......................M..A....TFP....................  360
baboon  301 .............................T..A....SFP....................  360

CM      359 GKHYWEVDVSKKSAWILGVCAGFQSDAMCNIEQNENYQPKYGYWVIGLQEGVKYSVFQDG  418
Rh      361 ............................Y...............................  420
baboon  361 ........................P...Y...............................  420

CM      419 SLHTPFAPFIVPLSVIICPDRVGVFVDYEACTVSFFNITNHGFLIYKFSQCSFSKPVFPY  478
Rh      421 .S..........................................................  480
baboon  421 .S..........................................................  480

CM      479 LNPRKCTVPMTLCSPSS                                            495
Rh      481 .................                                            497
baboon  481 .................                                            497
```

Figure 2.2 Alignments of amino acid sequences of cynomolgus monkey (CM), rhesus monkey (Rh), and baboon TRIM5αs. The RING, B-box2, coiled-coil and SPRY (B30.2) domains are indicated by labelled bars above the sequences. Variable regions 1, 2, and 3 are indicated by dashed bars above the sequence. Inverted triangles denote Sph and BamH restriction enzyme site, respectively. Dots denote the amino acid residues identical to the cynomolgus monkey sequence, dashes deletions relative to rhesus and baboon TRIM5αs. The box marks the amino acid residue found in this study to affect anti-HIV-2 activity of TRIM5α.



	Inhibition					
	MT4			CEM-SS		
	HIV-1 NL43	HIV-2 GH123	HIV-2 GH123/Q	HIV-2 UC2	HIV-2 GH123	HIV-2 GH123/Q
R	++	+++	+++	+	++	++
C	++	++	-	-	++	-
CBC	++	+++	++	-	++	+
RBR	++	+++	++			
CBB	++	+++	+	-	+	-
CBB P385S		+++	++			
R S385P		+++	+++			
C S383P		++	-			

Figure 2.3 Schematic representation of chimeric and mutant TRIM5αs, and summary of the experimental results. White and grey bars denote rhesus monkey (R) and cynomolgus monkey (C) sequences, respectively. B denotes a baboon sequence. The amino acid at the 385th or 383rd position in V2 of the SPRY domain is indicated in the construct. +++, ++, +, and – denote more than 1000-fold, 100- to 1000-fold, 8- to 100-fold, and less than 8-fold suppression of virus growth, respectively, relative to the negative control lacking SPRY domain. Restriction of the described viruses in two tested cell lines is shown.

In the MT4 cells all chimeric as well as rhesus monkey and cynomolgus monkey TRIM5αs almost completely restricted HIV-1 NL4-3 (Figure 2.3) and HIV-2 GH123. In the case of HIV-2 GH123/Q, rhesus monkey TRIM5α, CBC and RBR chimeric TRIM5αs but not cynomolgus monkey TRIM5α restricted the virus replication. This virus was also only moderately restricted by CBB chimeric TRIM5α (Figure 2.3) which indicates a weaker anti-HIV-2 GH123/Q activity of CBB as compared to CBC and RBR chimeric TRIM5αs. CBB differs from the RBR or CBC chimeric TRIM5αs in the SPRY domain is only at the 385th amino acid residue in V2, with a proline (P) in the baboon and serine (S) in the rhesus monkey and cynomolgous monkey sequences (Figure 2.2, box). These results suggest that the amino acid at this position of TRIM5α affects the protein restriction activity against HIV-2 GH123/Q infection.

To confirm this hypothesis, a mutant CBB TRIM5α was constructed with a serine at the 385th in the place of proline (CBB P385S TRIM5α, Figure 2.3) and compared in its ability to restrict HIV-2 GH123/Q with that of CBB TRIM5α. The inhibition of HIV-2 GH123/Q by the CBB P385S TRIM5α was stronger than that of the CBB chimeric TRIM5α (Figure 2.3) indicating that a single amino acid substitution from S to P at the 385th position of baboon TRIM5α influences its anit-HIV-2 activity.

Subsequently, a cynomolgus monkey and rhesus monkey TRIM5αs were produced with an S-to-P mutation at the corresponding positions of the 385th chimeric protein sequence (383th for the cynomolgus monkey due to a 2-position deletion) (CM S383P TRIM5α and Rh S385P TRIM5α respectively). However, these mutant TRIM5αs did not show significant differences in the anti-HIV-2 activity from the wild types, which indicates that the restriction is an accumulative effect of different parts on the protein.

In order to test the restriction activity of TRIM5αs in different physiological conditions another set of experiments was performed using the CEM-SS cell line. In this T cell line TRIM5α is expressed at lower levels of 0.2 of those in the MT4 cells. All TRIM5αs inhibited HIV-2 GH123 although the extent of inhibition varied among the TRIM5αs. The difference in anti-HIV-2 GH123/Q activity between CBB and CBC chimeric TRIM5α was also observed in CEM-SS cells. This indicates that the chimeric TRIM5αs containing baboon SPRY domain fail to restrict HIV-2 GH123/Q replication.

2.2.2 Structural modeling – sequence alignment

Given the evidence for the relevance of the V1 region and in particular the 385th position effect on the restriction capacity of TRIM5α against HIV-2, we performed structural analysis of the region and the residue of interest.

Based on the sequence search of the protein data bank (PDB) (Berman, Westbrook et al. 2000), we identified four homologous protein structures: PRYSPRY domain of TRIM21 (2IWG) (James, Keeble et al. 2007), mouse TRIM21 (2VOK and 2VOL) (Keeble, Khan et al. 2008) and human PRYSPRY domain (2FBE) (Grutter, Briand et al. 2006). All of the structures showed a highly significant e-value (Table 2.1) and about 35% (2IWG, 2VOK, 2VOL) and 29% (2IWG) of their sequence identity with the human TRIM5α SPRY domain, respectively. All for four structures showed 45%-76% sequence similarity with 2VOK and 2VOL being 99% identical.

template	e-value	identity positions	identity [%]
2IWG	$2.1e^{-25}$	71/295	35
2VOK	$3.1e^{-22}$	62/172	36
2VOL	$8.2e^{-22}$	60/172	35
2FBE	$7.2e^{-12}$	53/182	29

Table 2.1 Similarity of the human SPRY domain to the template structures.

A critical step of structural homology modeling is elaborating of an accurate alignment of the modelled protein to the structural templates. In order to assess the robustness of the alignment of the analyzed primate sequences and the templates we performed two additional alignments. The *basic alignment* apart from the four template sequences included six primate SPRY domain sequences: human, baboon (GenBank AAV91976), cynomolgus monkey (AB210052), rhesus macaque (AY523632), sooty mangabey (AY710303, EF551344) and African green monkey (AB210050). The *enriched alignment* included additional 289 sequences that scored e-value < $1.0e^{-12}$ in the BLAST (Altschul, Gish et al. 1990) search of GenBank (Benson, Karsch-Mizrachi et al. 2009). All

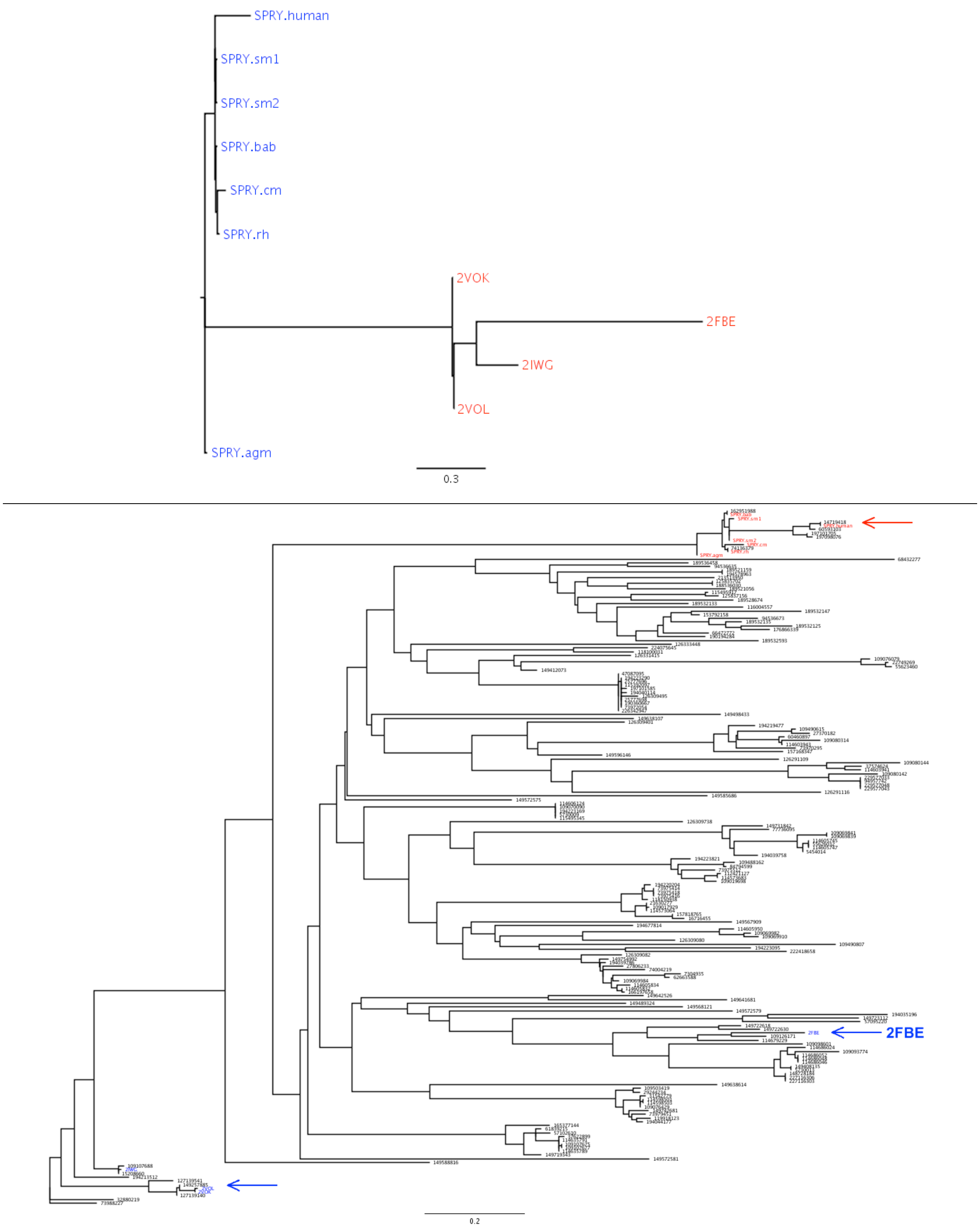


Figure 2.4 Maximum likelihood tree of the basic (top) and enriched (bottom) alignments. Seven primate sequences are labelled in red, template sequences in blue. Respective primate and template sequence clades are indicated with arrows colored accordingly on the enriched alignment tree. Additionally labelled 2FBE template shows the lowest similarity with the primate and other template sequences and is located in separation from other templates on the enriched alignment tree.

sequences in the enriched alignment appeared also among the first 1000 hits of the BLAST query of the used template sequences. Both alignments were computed using the Muscle method (Edgar 2004). In the enriched alignment iteratively sequences of long insertions or deletions (>10bp) relative to the templates were removed and the remaining sequences were realigned, which resulted in 168 homolog sequences used in this alignment.

The homologous sequences of the enriched alignment filled the sequence space between the templates and the primate sequences (Figure 2.4, bottom) especially between the 2FBE template and other used sequences.

We analyzed the difference in the alignment quality, calculated as the percentage of sequence identity, resulting from an increased number of homologous sequences in the enriched alignment (Table 2.2). It appears that the enriched alignment generally showed higher quality of the primate sequence alignment to the templates except for minor quality decrease (<0.20%) of three alignments of 2IWG and two of 2VOL (Table 2.2).

	AGM	baboon	CM	human	RM	SM
2IWG	1.18	-0.14	-0.13	-0.14	0.81	0.04
2VOK	1.25	1.92	1.45	0.42	1.42	1.92
2VOL	0.72	1.31	0.82	-0.19	-0.14	1.49
2FBE	1.35	1.66	2.15	2.77	1.66	2.15

Table 2.2 Alignment difference [%] of the primate SPRY domains and templates between the enriched and basic alignments. Lower alignment quality of the enriched alignment is colored in red.

Next, we inspected alignment of specific positions of TRIM5α SPRY domain (Ohkura, Yap et al. 2006) between the basic and enriched alignments (Figure 2.5). The comparison revealed that the major differences between the alignments occur close to the insertions in the primate SPRY domain relative to the template sequences which indicates that the alignments are less reliable close to the regions of these insertions (Figure 2.5). We therefore used both alignments in the further homology modeling to assess the structural variability of the differently aligned sequence regions.

```
                                   _________V1  region_________
human   RRYWVDVTVAPNNI-SCAVISEDKRQVSSPKPQIIYGARGT--RYQTFVNFNYCTGILGSQSITSGKHYWEVDVSKKT
CM      RRYWVDVTLAPNNI-SHAVIAEDKRQVSSRNPQIVYQSPGT--LFQSLTNFNYCTGVLGSQSITSGKHYWEVDVSKKS
RH      RRYWVDVTLATNNI-SHAVIAEDKRQVSSRNPQIMYQAPGTLFTFPSLTNFNYCTGVLGSQSITSGKHYWEVDVSKKS
baboon  RRYWVDVTLAPNNI-SHAVIAEDKRQVSSRNPQITYQAPGTLFSFPSLTNFNYCTGVLGSQSITSGKHYWEVDVSKKS
SM      ----VDVTLAPNNI-SHAVIAEDKRQVSSRNPQIMYQARGTLFSFPSHTNFNYCTGVLGSQSITSGKHYWEVDVSKKS
2IWG    --HMVHITLDPDTANPWLILSEDRRQVRLGDTQQSIPGNEE--RFDS------YPMVLGAQHFHSGKHYWEVDVTGKE
2VOK    HHHMVHITLDRNTANSWLIISKDRRQVRMGDTHQNVSDNKE--RFSN------YPMVLGAQRFSSGKMYWEVDVTQKE
2VOL    --HMVHITLDRNTANSWLIISKDRRQVRMGDTHQNVSDNKE--RFSN------YPMVLGAQRFSSGKMYWEVDVTQKE
2FBE    PEFQVDMTFDVDTANNYLIISEDLRSFRSGDLSQNRKEQAE--RFDT------ALCVLGTPRFTSGRHYWEVDVGTSQ
            ____V2  region___             ___V3  region___
human   AWILGVCAGFQPDAMCNIEKNENY--PKYGYWVIGLEEGVKCSAFQDSSFHTPSVPFIVPLSVIICPDRVGVFLDYEACTV
CM      AWILGVCAGFQSDAMCNIEQNENYQ-PKYGYWVIGLQEGVKYSVFQDGSLHTPFAPFIVPLSVIICPDRVGVFVDYEACTV
RH      AWILGVCAGFQSDAMYNIEQNENYQ-PKYGYWVIGLQEGVKYSVFQDGSSHTPFAPFIVPLSVIICPDRVGVFVDYEACTV
baboon  AWILGVCAGFQPDAMYNIEQNENYQ-PKYGYWVIGLQEGVKYSVFQDGSSHTPFAPFIVPLSVIICPDRVGVFVDYEACTV
SM      AWILGVCAGFQPDAMYNIEQNENYQ-PKYGYWVIGLQEGVKYSVFQDGSSHTPFAPFIAPLSVIICPDRVGVFVDYEACTV
2IWG    AWDLGVCRD-------SVRRKGHFLLSSKSGFWTIWLWNKQKYEA-----GTYP----QTPLHLQVPPCQVGIFLDYEAGM
2VOK    AWDLGVCRD-------SVQRKGQFSLSPENGFWTIWLWQ-DSYEA-----GTSP----QTTLHIQVPPCQIGIFVDYEAGV
2VOL    AWDLGVCRD-------SVQRKGQFSLSPENGFWTIWLWQ-KSYEA-----GTSP----QTTLHIQVPPCQIGIFVDYEAGV
2FBE    VWDVGVCKE-------SVNRQGKIELSSEHGFLTVGCREGKVFAA-----STVP----MTPLWVSPQLHRVGIFLDVGMRS
```

```
                                   _________V1  region_________
human   RRYWVDVTVAPNNI-SCAVISEDKRQVSSPKPQIIYG---ARGT--RYQTFVNFNYCTGILGSQSITSGKHYWEVDVSKKT
CM      RRYWVDVTLAPNNI-SHAVIAEDKRQVSSRNPQIVYQSPGTL----F-QSLTNFNYCTGVLGSQSITSGKHYWEVDVSKKS
RH      RRYWVDVTLATNNI-SHAVIAEDKRQVSSRNPQIMYQ---APGTLFTFPSLTNFNYCTGVLGSQSITSGKHYWEVDVSKKS
baboon  RRYWVDVTLAPNNI-SHAVIAEDKRQVSSRNPQITYQ---APGTLFSFPSLTNFNYCTGVLGSQSITSGKHYWEVDVSKKS
SM      ----VDVTLAPNNI-SHAVIAEDKRQVSSRNPQIMYQ---ARGTLFSFPSHTNFNYCTGVLGSQSITSGKHYWEVDVSKKS
2IWG    ---MVHITLDPDTANPWLILSEDRRQVRLGDTQ---Q---SI----P-GNEERFDSYPMVLGAQHFHSGKHYWEVDVTGKE
2VOK    HHHMVHITLDRNTANSWLIISKDRRQVRMGDTH---Q---NV----S-DNKERFSNYPMVLGAQRFSSGKMYWEVDVTQKE
2VOL    ---MVHITLDRNTANSWLIISKDRRQVRMGDTH---Q---NV----S-DNKERFSNYPMVLGAQRFSSGKMYWEVDVTQKE
2FBE    PEFQVDMTFDVDTANNYLIISEDLRSFRSGDLS---Q---NR----K-EQAERFDTALCVLGTPRFTSGRHYWEVDVGTSQ
            ____V2  region___             ___V3  region___
human   AWILGVCAGFQPDAMCNIEKNENYQPKYGYWVIGLEEGVKCSAFQDSSFHTPSVPFIVPLSVIICPDRVGVFLDYEACTV
CM      AWILGVCAGFQSDAMCNIEQNENYQPKYGYWVIGLQEGVKYSVFQDGSLHTPFAPFIVPLSVIICPDRVGVFVDYEACTV
RH      AWILGVCAGFQSDAMYNIEQNENYQPKYGYWVIGLQEGVKYSVFQDGSSHTPFAPFIVPLSVIICPDRVGVFVDYEACTV
baboon  AWILGVCAGFQPDAMYNIEQNENYQPKYGYWVIGLQEGVKYSVFQDGSSHTPFAPFIVPLSVIICPDRVGVFVDYEACTV
SM      AWILGVCAGFQPDAMYNIEQNENYQPKYGYWVIGLQEGVKYSVFQDGSSHTPFAPFIAPLSVIICPDRVGVFVDYEACTV
2IWG    AWDLGVCR----DSVRRKGH-FLLSSKSGFWTIWLWNKQKYEA--------GTYPQTPLHLQVPPCQVGIFLDYEAGMVS
2VOK    AWDLGVCR----DSVQRKGQ-FSLSPENGFWTIWLW-QDSYEA--------GTSPQTTLHIQVPPCQIGIFVDYEAGVVS
2VOL    AWDLGVCR----DSVQRKGQ-FSLSPENGFWTIWLW-QKSYEA--------GTSPQTTLHIQVPPCQIGIFVDYEAGVVS
2FBE    VWDVGVCK----ESVNRQGK-IELSSEHGFLTVGCREGKVFAA--------STVPMTPLWVSPQLHRVGIFLDVGMRSIA
```

Figure 2.5 Basic (top) and enriched (bottom) alignments of the primate and structure template sequences. For brevity only part of the SPRY domain is shown (residues 295-449) AGM and one of the SM sequences are not shown. Positions highlighted in red are those detected to be important for the TRIM5α lentivirus restriction in previous studies (Ohkura, Yap et al. 2006). Position analyzed in this study is highlighted. Variable regions are indicated by horizontal bars above each alignment.

2.2.3 Structural modeling

Using both alignments we constructed structural models of the primate TRIM5α SPRY domains focusing on the baboon domain due to its phenotypical characteristics found in the mutagenesis experiments. Models were built using Modeller 9v4 (Eswar, Webb et al. 2006) and visualized with PyMol v1.0r2 (DeLano).

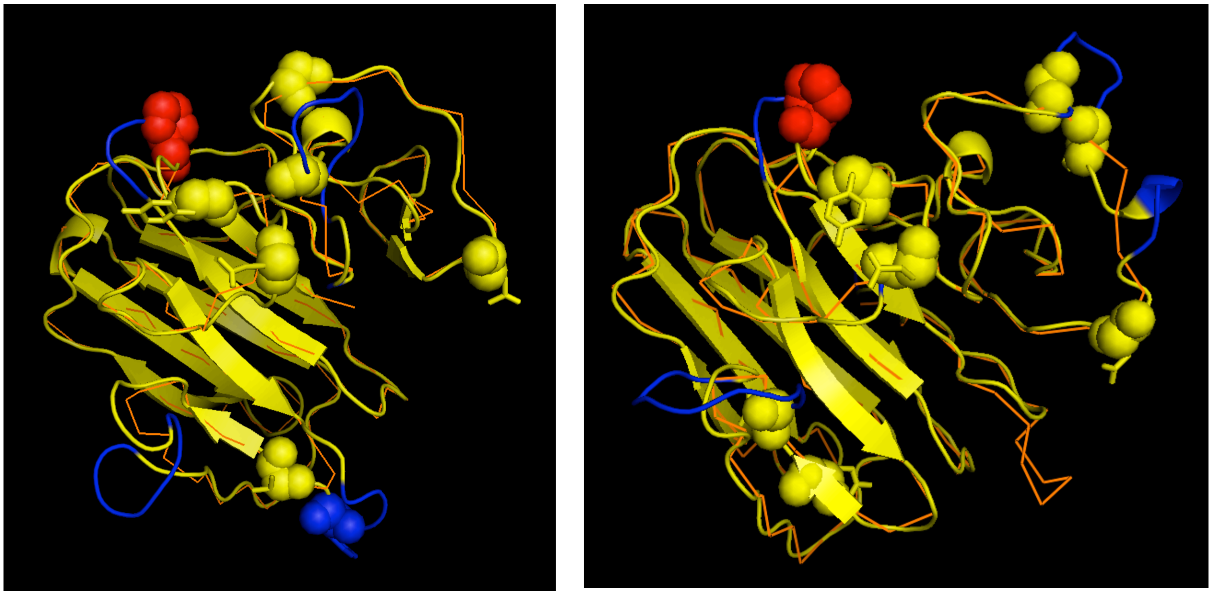


Figure 2.6 Baboon SPRY domain models based on basic (left) and enriched (right) alignments. The model structure is colored in yellow, the insertions relative to the 2IWG template are colored in blue. The backbone of the 2IWG template is traced with an orange line. Atoms of residues important for the restriction (Ohkura, Yap et al. 2006) are represented by spheres, atoms of the 385th residue are additionally colored in red.

Comparison of the models based on different alignments showed visible differences especially in the regions close to the insertions relative to the templates (Figure 2.6). These regions are located on loops that are structurally flexible which hampers precise structure prediction. The insertions are short, hence, based on the models, it can be assumed that the inserted regions are located at the protein surface. In spite of their separation in sequence, regions V1 and V2, appear to be spacially close. The proline residue at the 385th position of baboon TRIM5α is directed towards V1 region.

It was previously speculated that V1 and V2 are included in variable loops which form a binding interface of TRIM and other protein families with the SPRY domain (James, Keeble et al. 2007). A substitution involving proline is expected to affect local backbone conformation, and a P/S substitution in particular results in an additional hydrogen bond donor-acceptor. Overall, this indicates that position 385 is located at a putative binding site, where a P/S substitution might affect the local structure and surface chemistry. This result is consistent with the observed impact of these substitutions on TRIM5α anti-HIV-2 activity. Nevertheless, it is not yet clear how these substitutions affect TRIM5α restriction activity and binding affinity in particular, which calls for further investigation of the protein interactions involving TRIM5α.

2.2.3 Discussion

The study presented here indicates that the anti-HIV-2 activity of chimeric TRIM5α containing the entire baboon SPRY domain is weaker than that of chimeric TRIM5α containing baboon V1 alone. These results are consistent with those of the previous in vivo studies (Barnett, Murthy et al. 1994; Locher, Barnett et al. 1998; Locher, Witt et al.

2001). Additionally a single amino acid in V2 of the SPRY domain was found to affect its anti-HIV-2 activity.

An S-to-P mutation in the CM and baboon SPRY resulted in an enhanced restriction capacity against HIV-2 GH123/Q strain. In contrast to baboon and cynomolgus monkey TRIM5α, there was no evidence of an obvious contribution by the 385th amino residue for anti-HIV-2 activity in the case of rhesus monkey TRIM5α background. The specific combination of amino acids in V1 and V2 might therefore play an important role in determining specificity of restriction (Ohkura, Yap et al. 2006). The combination of amino acids in V1 and V2 of the SPRY domain of baboon and cynomolgus monkey TRIM5α appears to be more stringent in its anti-HIV-2 activity than that of rhesus monkey TRIM5α.

The study adds to previous recombinant studies of TRIM5αs performed by the Osaka group (Nakayama, Miyoshi et al. 2005; Song, Nakayama et al. 2007; Kono, Song et al. 2008) as well as other similar studies (Stremlau, Perron et al. 2005; Yap, Nisole et al. 2005; Ohkura, Yap et al. 2006) that aimed at identifying key differences between homologous TRIM5αs that determine their restriction capacities. With the lack of knowledge on the precise mechanism of TRIM5α-CA binding, mutagenic studies of the SPRY domain supported by structural models provide hypotheses about the potential interaction site. The proposed structural model indicates that the position 385 is located at a putative binding site and the P/S substitution is expected to affect the local structure and protein surface chemistry. Knowledge of the exact binding mechanism would open new potential directions for HIV treatment. Analogous mutagenetic studies of the CA should follow in order to identify respective positions of the virus protein playing crucial role in this interaction.

2.3 Second interaction partner - single substitution in HIV-1 CA affecting its sensitivity to the TRIM5α restriction

The efficient restriction of lentviruses by the TRIM5α protein is dependent on the complementary binding site of both interaction partners. Given the rapid evolution of the virus, mutation of the virus CA might be the main force conditioning the restriction and the virus capacity to infect certain host species. In this study, we analyzed the structural impact of single CA mutations that affected virus restriction by primate TRIM5α. The results of the study described in this section are published in: Kuroishi, A., Bozek, K., Shioda, T., Nakayama, EE. A single amino acid substitution of the human immunodeficiency virus type 1 capsid protein affects viral sensitivity to TRIM5alpha. *Retrovirology.* 2010 Jul 7;7:58.

2.3.1 Experimental results

The HIV-1 CA protein consists of the N-terminal core and C-terminal dimerization domains (Gitti, Lee et al. 1996; Gamble, Yoo et al. 1997). The N-terminal of CA is composed of seven α-helices that are arranged together in a flat triangular shape connected by conformationally flexible loops (Tang, Ndassa et al. 2002). Previous mutagenic experiments reported the importance of the loops between α-helices 4 and 5 (L4/5) and between α-helices 6 and 7 (L6/7) for the virus replication in CM cells (Kuroishi, Saito et al. 2009). The same study showed that introduction of SIVmac L6/7 into an HIV-1 CA resulted in an improved of viral growth in cynomolgus monkey (CM) cells and an attenuated growth in human cells. To gain more insight into the effects of CA loops on viral replication a virus construct was experimentally tested containing SIVmac L4/5, L6/7 and vif in an HIV-1 NL-43 genetic background (NL-4/5S6/7SvifS) (Figure 2.7). The virus was inoculated into CEMss cells and cultivated over 42 days. The progeny virus (NL-4/5S6/7SvifSd42) that was inoculated into fresh CEMss cells showed an increased replicative capability (Figure 2.8).

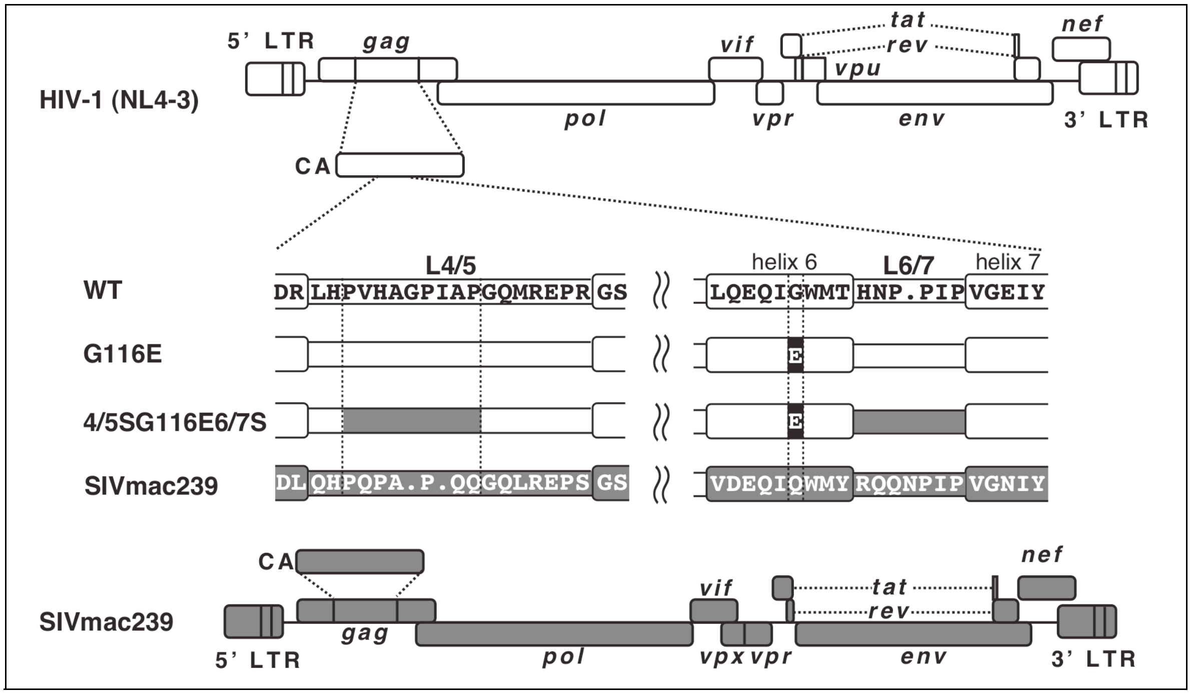


Figure 2.7 Schematic representation of HIV-1 CA constructs. White and grey bars denote HIV-1 (NL4-3) and SIVmac239 sequences respectively. E indicates the amino acid residue at the 116th position of the CA protein.

The progeny virus was next amplified using polymerase chain reaction (PCR) and cloned. Nucleotide sequence analysis of the resultant clones revealed that six out of six independent clones carried a single nucleotide substitution of a glycine (G) to glutamic acid (E) at the 116th position of CA (G116E).

Analysis of 96 HIV-1 strains in Los Alamos HIV sequence database (http://www.hiv.lanl.gov/) including subtype A to subtype K of group M, revealed that there was no HIV-1 strain carrying glutamic acid at the 116th position of CA, although this position was occupied with variable amino acid residues (35 strains carried glycine, 36 alanine, nine threonine, seven arginine, six glutamine, one isoleucine and one aspartic acid).

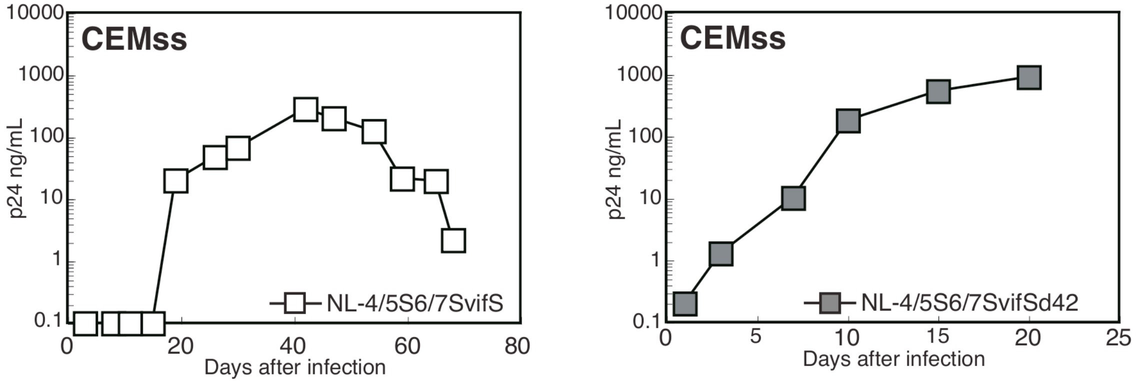


Figure 2.8 Adaptation of HIV-1 derivatives to human cells. NL-4/5S6/7SvifS (left panel), a virus with the SIVmac239 L4/5, L6/7 and vif was inoculated into CEMss cells which were then periodically assayed for the levels of p24 as an indicator of virus replication. Virus in the culture on day 42 after infection (NL-4/5SL6/7SvifSd42) was inoculated into fresh CEMss cells and assayed for the replication based on p24 (right panel).

To validate the effect of the G116E mutation on the virus replication the mutation was introduced into the initial virus NL-4/5S6/7SvifS (NL-4/5SG116E6/7SvifS). In addition, to exclude the effects of SIVmac genomic parts on the TRIM5α mediated restriction, the same mutation was introduced into NL4-3 HIV-1 reference virus (NL-G116E) (Figure 2.7). The two viruses were tested in CEMss cells expressing three types of TRIM5α – human, CM and a negative control TRIM5α with a truncated SPRY domain (CM-SPRY(-)). The NL-4/5SG116E6/7SvifS virus could grow more efficiently in the TRIM5α expressing CEMss cells than the original NL-4/5S6/7SvifS virus. The virus could grow in the cells expressing CM TRIM5α although the levels of virus titer were less than 10% of those in the absence of TRIM5α (Figure 2.9). The growth of NL4-3 virus was not affected by human TRIM5α, while that of NL-G116E was slightly suppressed by human TRIM5α.

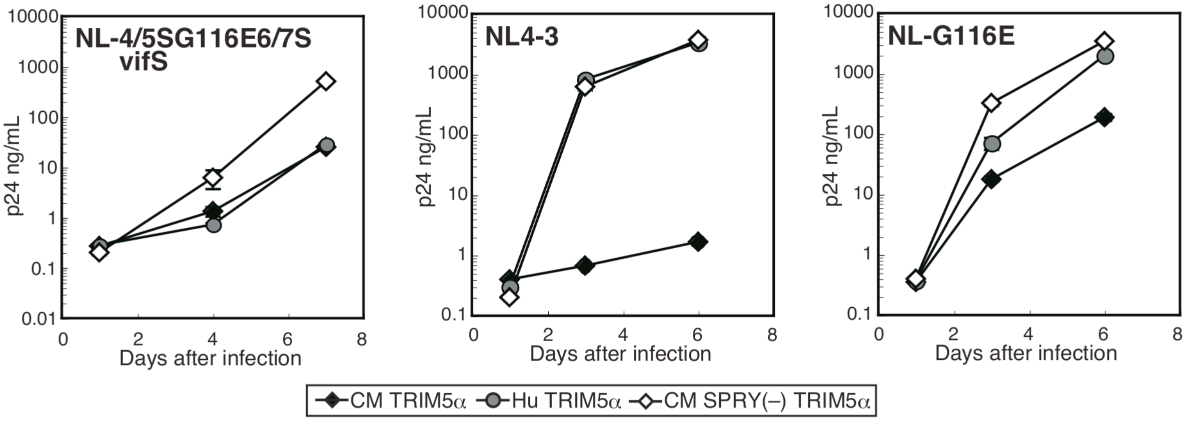


Figure 2.9 Viral growth in the presence of TRIM5α. CEMss cells expressing CM (black diamonds), human (grey circles), or CM-SPRY(-) (white diamonds) TRIM5α were infected with HIV-1 derivatives and assayed for the levels of p24 indicating virus replication.

This suggested the individual effect of the 116th amino acid residue of CA on the TRIM5α mediated restriction. The rescued replicative capability of the virus NL-4/5S6/7SvifSd42 virus in human cells was due, at least partly, to the acquisition of G116E substitution in

CA. The G116E substitution that occurred during long-term virus cultivation in human cells appeared to render the virus more sensitive to TRIM5α restriction.

2.3.2 Structural modeling

In order to investigate the localization of the CA G116E single amino acid substitution and its potential effect on the protein structure, we performed a structural analysis of the capsid protein of NL-G116E. We used the structure of the N-terminal domain of NL4-3 CA protein (1GWP) (Tang, Ndassa et al. 2002) as a template for this study. The goal of this analysis was to inspect the effect of one amino acid change relative to the template, No alignment was therefore needed for this structural modeling. We built the model of NL-G116E using Modeller 9v4 (Eswar, Webb et al. 2006) and visualized with PyMOL v1.0r2 (DeLano).

The position 116 is located in the 6th helix near the L4/5 and L6/7 and is apparently exposed to the protein surface. (Figure 2.10, upper panels). The structural effects of the G-to-E substitution might be due to the long side chain of E in contrast to G lacking side chain. The mutation can result in two possible effects. First, if the residue belongs to the interaction site, it can change the local complementarity between CA and TRIM5α. Second, even if the residue is not directly in the binding site, the side chain and polarity change can influence the configuration of nearby loops and therefore affect the binding site that is located somewhere else on the protein. Notably, the loops being flexible parts of the protein are slightly repositioned in the modeled structure with the G116E substitution (Figure 2.10).

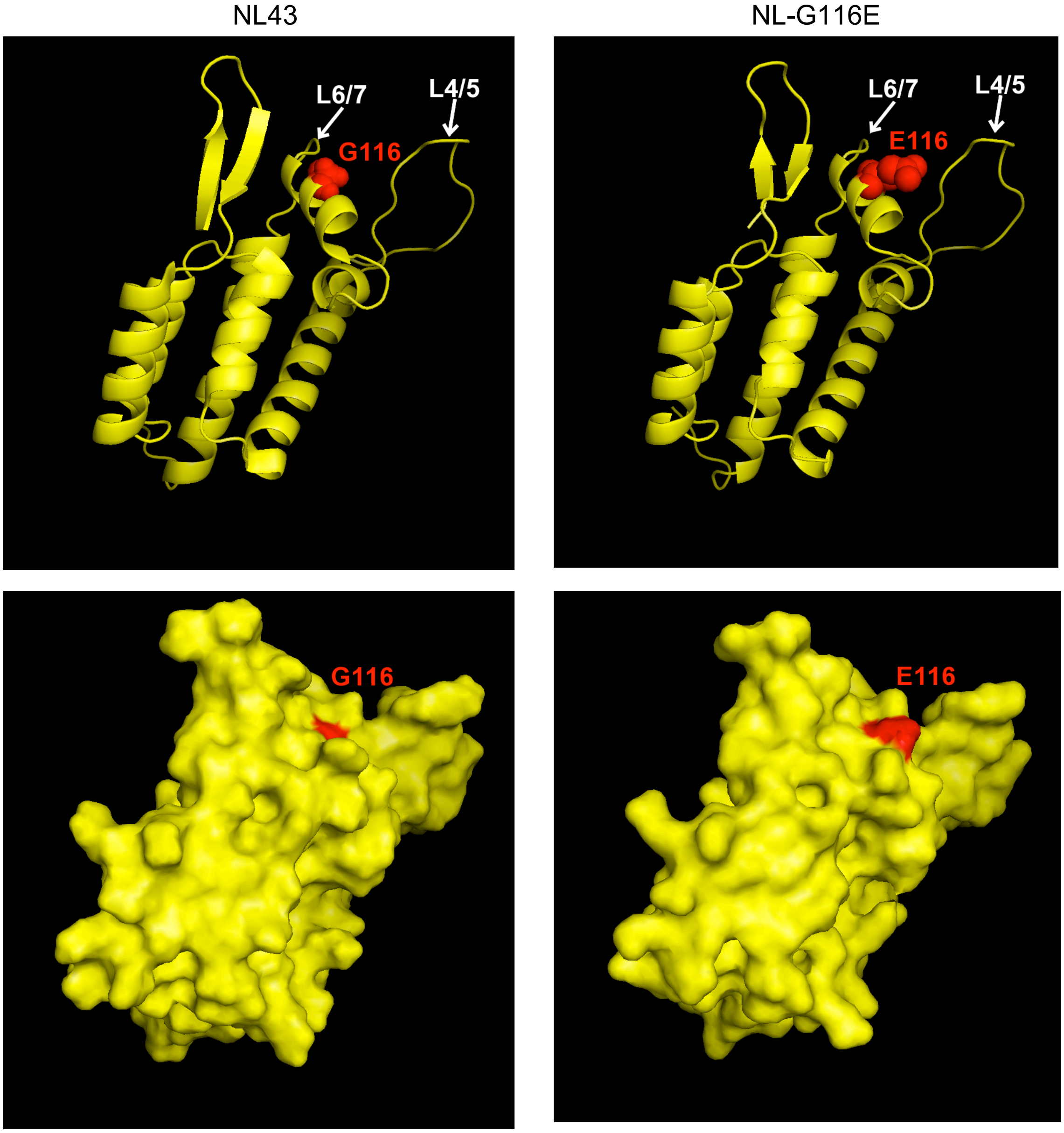


Figure 2.10 Structural model of the N-terminal domain of HIV-1 CA with G116E substitution. Left panels show the template structure of the N-terminal domain of the HIV-1 CA, right panels – of the domain structure with the G116E mutation. Red spheres represent G and E on the 116th position with their side chains. Bottom panels show surface views of the template and model structures with the 116th position indicated in red. The loops between α-helices 4 and 5 (L4/5) and 6 and 7 (L6/7) are labelled.

2.3.3 Discussion

In this study, through a long-term cultivation of human CEMss cells infected with an HIV-1 derivative virus a mutation was found that improved the virus replication. The G-to-E single amino acid substitution at the 116th position of CA appeared to enable the virus to escape the TRIM5α recognition. Although this position is highly variable among natural HIV-1 strains from subtypes A to K, no virus with E at the 116th position was found in the

Los Alamos HIV sequence database (http://www.hiv.lanl.gov/), which suggests its importance for viral growth. It is located at the central position of the CA surface composed of L4/5 and L6/7, a structure considered to be important for TRIM5α binding (Kuroishi, Saito et al. 2009). The amino acid substitution of G to E at the 116th position caused an important change in the structure of the surface through exchanging of a lack of side chain by a long, negatively charged side chain (Figure 2.10). This change in the conformational structure of the L4/5 and L6/7 might affect the interaction between the CA and TRIM5α. Alternatively, this substitution might influence the configuration of surrounding loops by the changes in the side chain and polarity without directly involving the binding site of TRIM5α.

Similar to the mutagenic study of the TRIM5α protein this study pointed to a residue of CA that is potentially involved in the interaction site. Even though the crystal structure of NL4-3 CA is a good template for homology modeling, the uncertainty of the TRIM5α model does not suggest hypotheses about the precise interaction mechanism. However, a comparative study of structural and biochemical properties of a broader spectrum of virus and host protein variants of known interaction phenotype could point to major determinants of the interaction.

2.4 Two interaction partners together – compilation study of the experimentally tested TRIMα and CA sequences

The two studies presented in the previous sections of this chapter analyzed individual genetic changes of each of the interaction partners separately and their effect on the TRIM5α-CA interaction. As the last step of the analysis of this host-virus interaction on the single interaction level we performed a compilation analysis of the experimental results collected by the Osaka group. The group gathered data on the capacity of different CA and TRIM5α pairs to interact by measuring the efficiency of virus carrying a specific CA to infect cells expressing different TRIM5αs. The protein variants were specific to species or obtained through mutagenesis studies. The compiled sequence and phenotype dataset of individually tested proteins allowed for a systematic and statistical search of the sequence and structural determinants of the TRIM5α restriction that might have not been observed in the studies of individual genetic changes.

2.4.1 Sequence analysis

First, we collected sequence data of the TRIM5α and CA variants that were experimentally tested by the Osaka group (Song, Nakayama et al. 2007; Kono, Song et al. 2008; Kuroishi, Saito et al. 2009; Kono, Song et al. 2010; Kuroishi, Bozek et al. 2010). We obtained a list of 35 CA variants of HIV-2 and SIVmac239 that were tested for the restriction by several TRIM5α constructs (*CA dataset*). The *TRIM5α dataset* contained 14 sequences of cynomolgus monkey (CM) and rhesus macaque (RH) TRIM5αs and sequences composed of CM, RH and baboon TRIM5α genetic parts as described in the first part of this chapter. As the experiments were aimed at testing effects of a single protein position on the restriction, the datasets were incomplete and biased in favor of mutations at specific sequence positions of special interest. Additionally, not all protein pairs were tested for their interaction. In order to perform statistical analyses of the sequence in relation to the phenotype we extracted a subset of the data that was fully characterized with the restriction phenotype.

The resulting CA dataset consisted of 35 CA variants that were characterized for being restricted by CM and RH TRIM5α. Restriction was quantified as binary information with 1 indicating more than 8-fold restriction of the virus replication. In the TRIM5α dataset we identified 14 protein constructs that were tested for their restriction of two HIV-2 CA variants: GH123 and GH123_Q. These two CA variants differ only at position 120, with P and Q in GH123 and GH123_Q respectively. Since GH123 was restricted more than 8-fold by all of the 14 TRIM5α constructs, we could only search for TRIM5α positions related to the GH123_Q restriction. Nine of the TRIM5αs showed CA restriction, however, given the limited number of sequences originating from only three primate species and the very specific interaction partner they were tested against, the information obtainable from the TRIM5α dataset was limited.

Both datasets of sequences characterized with their restriction phenotype information, were aligned using Muscle (Edgar 2004) and further analyzed. We performed two tests for relationship of each sequence position with the phenotype. Mutual information (MI) was used as an indicator of the significance of the relation between a change on a given sequence position with the phenotype, without defining which amino acid type corresponds to a specific phenotype. Fisher exact test (FET) was used to test the significance of co-occurrence of a specific amino acid type on each sequence position with the same restriction phenotype and this way to point to specific amino acid types that correspond to restriction.

MI test indicated three CA positions significantly ($p<0.05$) related to the CM TRIM5α restriction and only one related to the RH TRIM5α restriction (Table 2.3). The highest MI was found between the CA 120th position and CM TRIM5α restriction. Other positions mostly in the region close to the first loop (L1) of the CA (residues 5-13) showed elevated MI ($p<0.1$) with the CM TRIM5α restriction. The only position significantly related to the RH TRIM5α restriction in the analyzed dataset is position 97, however other positions in the region of L4/5 showed elevated mutual information ($p<0.1$, Table 2.3) with the RH TRIM5α restriction (residues 82-97). FET pointed to several amino acid types related to the restriction at these positions (Table 2.3).

CA position (amino acid)	MI	p-value	CA position (amino acid)	MI	p-value
5	**0.194**	**0.027**	82 (A)	0.148	0.072
7	0.093	0.078	86	0.137	0.096
11 (V)	**0.185**	**0.026**	90 (L)	0.128	0.061
13	0.114	0.075	**97 (D,E)**	**0.155**	**0.017**
29	0.093	0.076	124	0.090	0.089
75	0.093	0.067	136	0.090	0.074
111	0.104	0.059			
120 (P,Q)	**0.951**	**0**			

Table 2.3 CA positions with a high (p<0.1) MI indicating CM (left) and RH (right) TRIM5α restriction. Positions with a significant p-value (p<0.05) are indicated in bold. Amino acid types significantly (FET, p<0.05) related to the restriction are indicated in brackets.

In the TRIM5α dataset significant (p<0.05) MI related to the GH123_Q restriction was found for positions 330 and 339-342. These positions are located in the V1 region of the protein. Region 339-342 includes an insertion of two amino acids in the RH and baboon proteins relative to CM TRIM5α (Figure 2.11). Notably, an insertion at positions 339 and 340 was found to be significantly related to restriction (FET, p<0.05). Due to the low variability of the TRIM5α dataset, lack of the crystal structure of the SPRY domain and limited information on the phenotype of each protein variant, the information resulting from the analysis of this dataset is limited. It was therefore not analyzed in the further steps of this compilation study.

```
CM TRIM5α        QVSSRNPQIVYQSPGTLF--QSLTNFNYC
RH TRIM5α        QVSSRNPQIMYQAPGTLFTFPSLTNFNYC
BABOON TRIM5α    QVSSRNPQITYQAPGTLFSFPSLTNFNYC
```

Figure 2.11 TRIM5α residues 320-350. Highlighted in red are positions significantly related to GH123_Q restriction (330, 339-342).

2.4.2 Amino acid features

The reason that so few CA positions and amino acid types were found as significantly related to the phenotype might lie in the limited size of the CA dataset. An insufficient variation of amino acid types on important sequence positions does not allow for detecting positions and amino acid types significantly related to the phenotype. A way to compensate the scarceness of sequence variants is to group amino acid types sharing certain properties and test for the relation of properties instead of amino acids to the phenotype. Additionally, properties of amino acids express their physicochemical similarities that are not encoded in the amino acid letter representation.

We subjectively selected seven amino acid indices that might be important for protein-protein binding from the AAIndex database (Kawashima, Ogata et al. 1999). This database catalogues amino acid properties previously tested and published. An index is a numerical quantification of a given physicochemical property of each amino acid type. The database contains about 500 indices many among them being redundant and strongly correlated. The seven indices were selected as an example set of properties potentially relevant for binding and are mainly related to the amino acid structure and hydrophobicity (Table 2.4).

Given a numerical representation of amino acids one can test the correlation of an amino acid property on each sequence position with the quantified phenotype. In this test we expanded the phenotype description to four values denoting the fold of the suppression of virus growth with 3 denoting more than 1000-fold, 2 – 100- to 1000-fold, 1 – 8- to 100-fold, 0 – less than 8-fold. We calculated the correlation of the seven selected amino acid indices on each CA sequence position with the four-value phenotype description and tested for its significance by permutation of amino acid types on a given position in the dataset.

Amino acid index (ID)	Positions - CM TRIM5α restriction	Positions - RH TRIM5α restriction
Heat capacity (HUTJ700101)	7, 13, 27, 29, 75, 109, 111, 120, 124	61, 82, 97
Hydrophobicity factor (GOLD730101)	86, 109, 120, 136	82, 86, 90
Hydropathy index (KYTJ820101)	5, 7, 13, 109, 136	11, 86, 90, 120
Side chain volume (KRIW790103)	11, 13, 27, 29, 109, 111, 124	82, 86, 97
Percentage of buried residues (JANJ780102)	5, 7, 13, 29, 75, 109, 120, 136	11, 86, 97, 120
Transfer free energy (JANJ790102)	5, 7, 13, 111, 120, 136	11, 62, 86, 97, 120, 150
Average interactions per side chain atom (WARP780101)	7, 27, 29, 75, 120, 124	11, 62, 86, 97

Table 2.4 CA positions with a significant (p<0.05) correlation with the restriction by CM and RH TRIM5α. Colors indicate the type of correlation with positive in red and negative in blue.

The amino acid indices analysis pointed to several positions significantly related to the phenotype additional to those inferred from the amino acid type analysis (Table 2.4). Positions important for the CM TRIM5α are located mainly close to the loops joining CA α-helices: L1 – positions 5, 7, 13, loop between helix 5 and 6 (L5/6) – positions 109, 111 and L6/7 – positions 120, 124. Several other important positions seem to occur also in helices (27, 29, 75) however they show small amino acid variation among the CA variants. Positions related to the RH TRIM5α restriction lie mainly close to the L4/5 (82, 86, 90, 97). Even though the specific amino acid properties and their relation to the phenotype are hard to interpret within a limited scope of the CA dataset (Figure 2.12), this analysis indicated the importance of the physicochemical constitution of the CA loops for the TRIM5α interaction. Four CA loops are pointing in the same direction and are in structural vicinity. Changes on the loops exposed to the protein surface (L1, L5/6 and L6/7) might directly affect the binding site, changes on the less exposed loop parts might indirectly affect local configuration of the residues involved in binding.

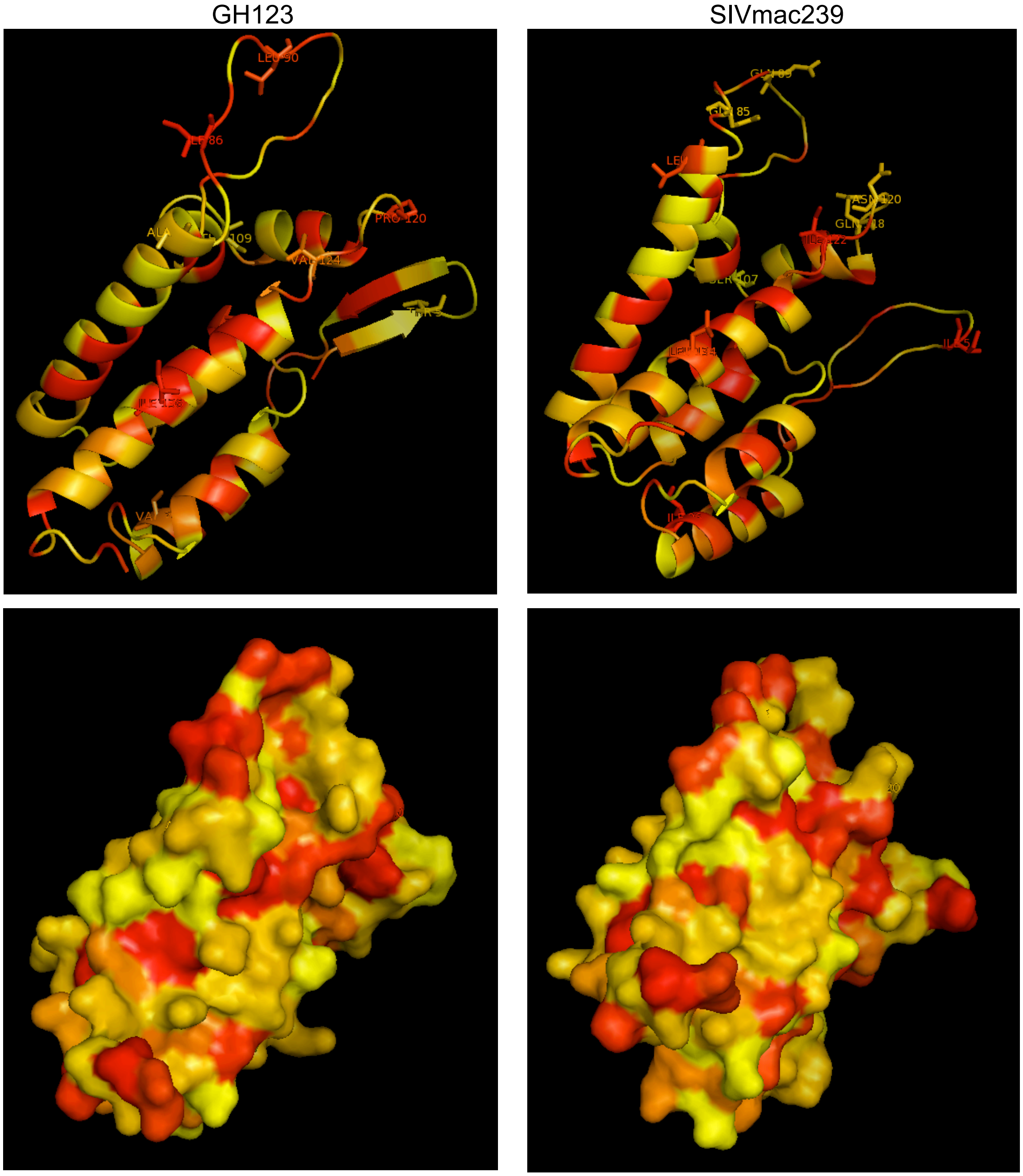


Figure 2.12 Hydrophobicity factor property of the residues of the GH123 (left panels) and SIVmac239 (right panels) CA proteins. Structures are colored according to the amino acid hydrophobicity factor with red indicating high and yellow low values. Positions showing significant correlation to the CM or RH TRIM5α restriction are labeled on the upper panels. Bottom panels show the protein surface colored accordingly. The two CA variants show opposite phenotype with GH123 being restricted by both TRIM5αs and SIVmac239 by none. The hydrophophobicity factor values are generally lower on the loop surface of SIVmac239 which might prevent TRIM5α binding.

2.4.3 Structural analysis

Next, we investigated the structural properties of the variants in the CA dataset. We used the structure of the N-terminal domain of HIV-1 CA protein (1GWP) (Tang, Ndassa et al. 2002) as a template and modeled the structures of each variant in this dataset using Modeller 9v4 (Eswar, Webb et al. 2006). Each variant was aligned to the template individually and the best model was chosen out of ten runs of the modeling procedure.

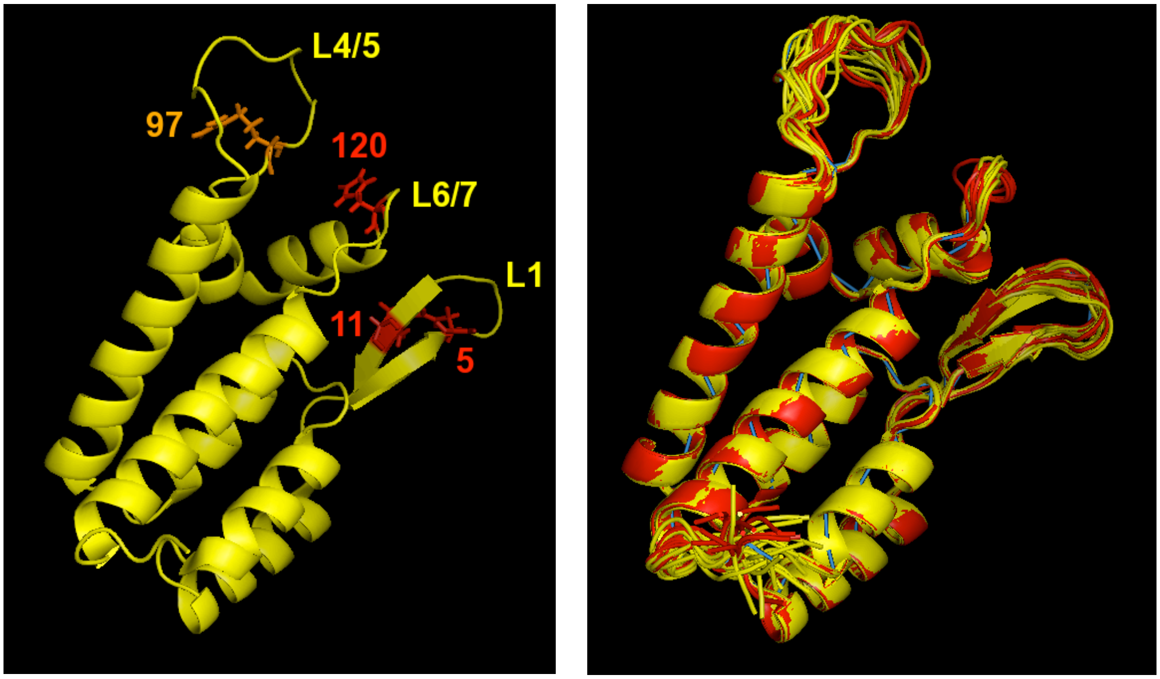


Figure 2.13 Models of the CA variants. Template CA structure used in the modeling is shown on the left panel. Three exposed loops are indicated, side chains of residues significantly related to the phenotype based on MI test are pictured as sticks and labelled. Red side chains belong to residues related to the CM TRIM5α restriction, orange to the RH TRIM5α restriction. Right panel shows all modeled variants superposed. Variants restricted by CM TRIM5α are colored in red, not restricted in yellow. The backbone of the template structure is traced with blue line.

The modeled structures showed certain variability, mainly in the loops (Figure 2.13). Visual comparison of individual variants of opposite phenotypes (Figure 2.12) pointed to specific differences among restricted and non-restricted variants. As an example SIVmac239 CA that is not restricted by any CM or RH TRIM5α appeared to have less extended loops and a more contracted structure than GH123 which is restricted by both TRIM5αs (Figure 2.12). In order to quantify this information for all CA variants, we elaborated a rough measure of *loop extension* as follows. For each of the loops L1, L4/5 and L6/7 we defined start and end residues: 2 and 12 for L1, 83 and 101 for L4/5, 118 and 126 for L6/7. We defined loop extension as the maximum distance of a loop residue to the loop start and end residues. The distances from the loop start and end of each of the three loops in the analyzed CA variants are plotted against each other in Figure 2.14. Variants are labeled and colored according to the CM TRIM5α restriction phenotype. The restricted variants tend to have more extended loops as their loop start and end distances are seemingly larger than those unrestricted on all three loops and on L6/7 in particular (Figure 2.14) suggesting that an extended loop conformation facilitates TRIM5α binding. L6/7 is the central loop between L1 and L4/5 on the CA surface containing previously described (Song, Nakayama et al. 2007) position 120 that has a strong effect on the TRIM5α restriction.

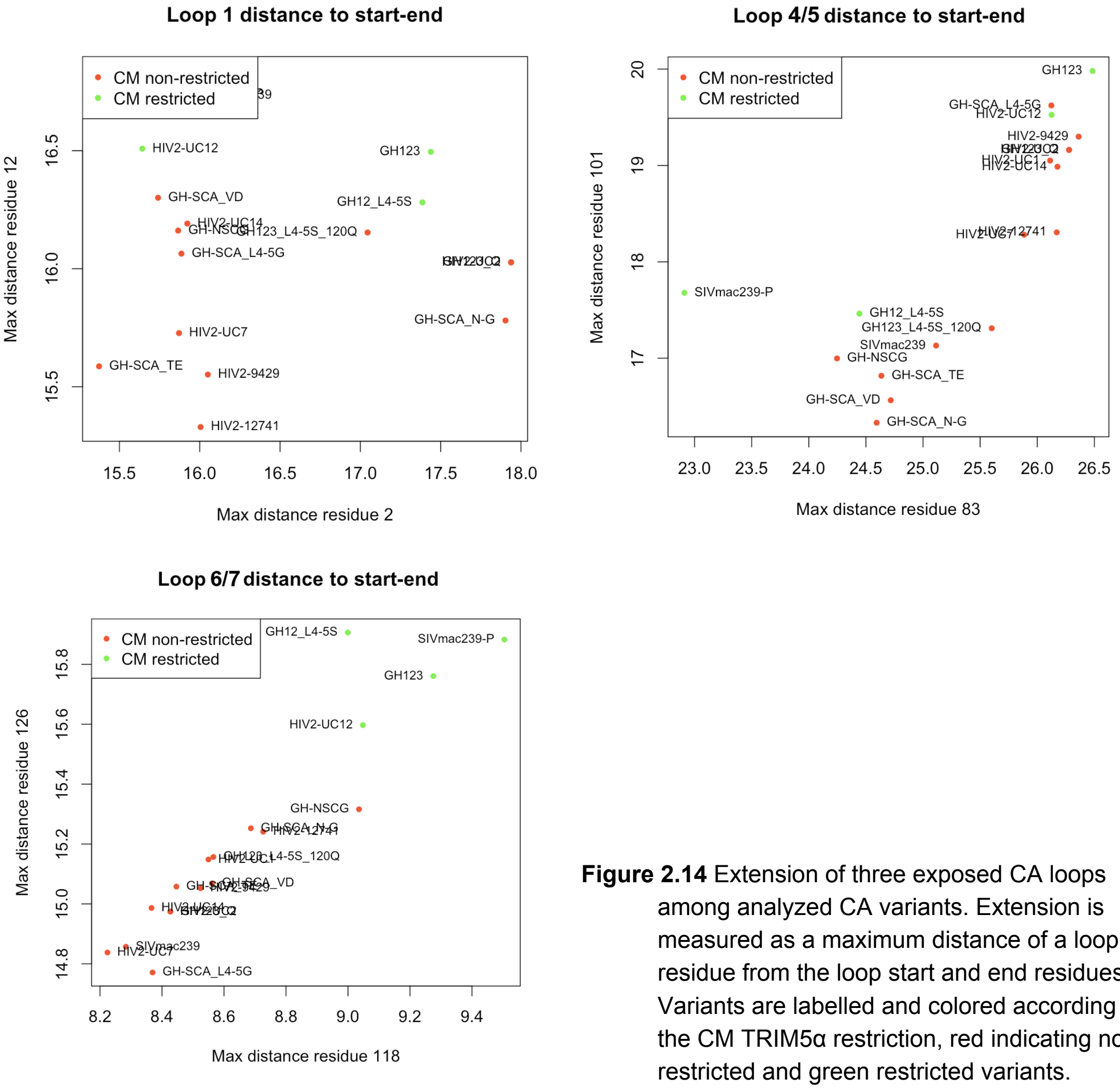


Figure 2.14 Extension of three exposed CA loops among analyzed CA variants. Extension is measured as a maximum distance of a loop residue from the loop start and end residues. Variants are labelled and colored according to the CM TRIM5α restriction, red indicating non-restricted and green restricted variants.

2.3.4 Electrostatic potential

The previous steps of this compilation analysis pointed to specific positions and regions of CA, their physicochemical and structural properties that correspond to the TRIM5α restriction. Apart from the sequence and structural characteristics conditioning protein-protein interaction, an important factor allowing for a molecular interaction is the electrostatic potential at the binding site. The electrostatic potential on protein surface is generated through redistribution of electrons according to local electrical fields. It is defined as the potential energy of a proton at a particular location near a molecule. Negative electrostatic potential results in attraction of the proton by the concentrated electron density. Positive electrostatic potential results in repulsion of the proton by the atomic nuclei in regions where low electron density exists and nuclear charge is incompletely shielded. Electrostatic effects were shown to be a major factor in determining the nature and strength of the interaction between protein surfaces (Dong and Zhou 2002; Kortemme and Baker 2002). A complementary charge on the binding site of both proteins might result in an attraction force allowing for the binding to happen.

In the last part of this study we performed an electrostatic potential analysis of the surface region of the three CA loops. For this analysis we selected two CA variants – GH123 and SIVmac239 as two example strains showing opposite phenotypes of interaction with primate TRIM5α proteins and the most strongly pronounced difference in the shape of the loops surface (Figure 2.15). SIVmac239 CA has a more contracted shape as compared to extended GH123 loop structure. To confirm that this shape difference is not due to modeling noise we remodeled both proteins using each one as a template for the other. The remodeled structures showed similar shape differences, which supports the view that the real structures differ.

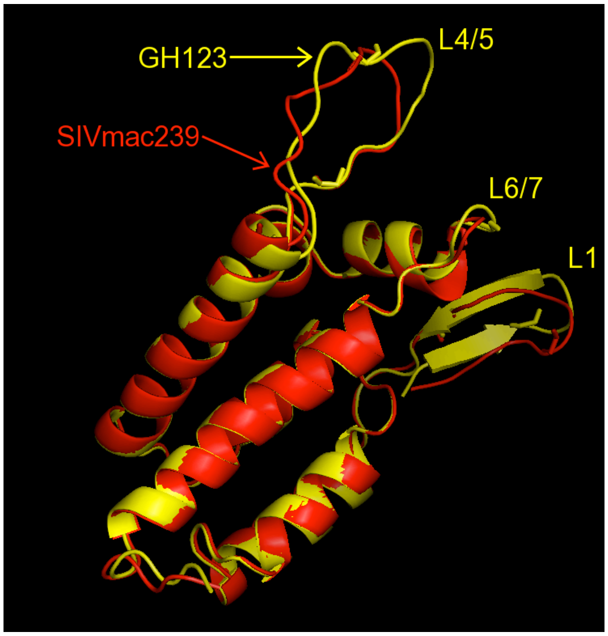


Figure 2.15 Superposition of the GH123 CA (yellow) and SIVmac239 (red) CA modeled structures. The three loops containing sites important for the TRIM5α interaction are labeled.

As the initial step preceding electrostatic potential modeling we added missing hydrogen atoms and estimated the ionization (protonation) of the molecules. We used H++ server (Gordon, Myers et al. 2005) that adds protons to the input structure according to the calculated ionization states at the specified pH of the solvent. The H++ method models molecules as a medium with low dielectricity ε_{in} in a solvent with a high dielectric constant ε_{out}. It additionally allows the user to define the salt concentration of the medium and its pH. We used the biologically relevant parameters in human cells pH=7.2, salinity 1%, molecule dielectric ε_{in}=10 and medium dielectric ε_{in}=80. The dielectric parameters were chosen according to the suggestions of the authors of the H++ method as appropriate for modeling protonation of surface residues. We tested several other parameter combinations and verified visually if they result in radically different electrostatic potential profiles. Other parameter regimes did not produce strongly different electrostatic potentials in the region of interest we therefore chose the initial parameters as the most relevant in biological settings.

Next, we applied two methods of electrostatic potential calculation: the Adaptive Poisson-Boltzmann Solver (APBS) (Baker, Sept et al. 2001) and the nonlocal electrostatics method (Hildebrandt, Blossey et al. 2007). In both methods electrostatic properties are described by Poisson-Boltzmann equation (PBE), a second-order nonlinear partial differential equation. The APBS method solves the equation using finite-element techniques based on parameter discretization and iterative parallel refinement of the equation solution. The nonlocal electrostatics method allows for inclusion of the structure of water molecules in the calculation and describes the system as a continuum. This method captures the effects of the dipole polarization of the water molecules and the effects of surrounding hydrogen bond network and represents therefore a more accurate model of the electrostatic potential estimations close to the molecule-solvent interface.

We visually inspected the results of the electrostatic potential using pymol (DeLano). A strong difference between the two molecules could be observed on the loops surface with the GH123 molecule having predominantly positive and SIVmac239 predominantly negative electrostatic potential on this part of their surface (Figure 2.16).

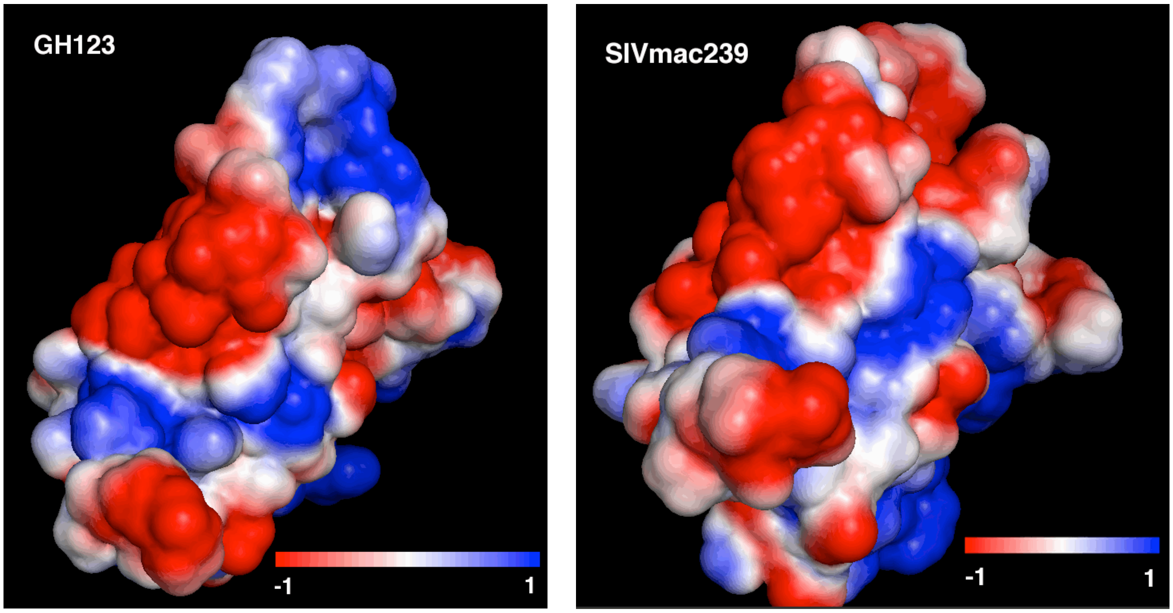


Figure 2.16 Electrostatic potential on the GH123 and SIVmac239 surfaces. Structures are positioned as in Figure 2.15 with the loops directed towards upper right of the image. Electrostatic potential was calculated and visualized using the APBS plugin in pymol software (DeLano).

In order to quantify this observation and to obtain more insight into the specific region where the electrostatic potential differences are strong we extracted the electrostatic potential values on the surface of the two molecules. We used three different surface approximations: solvent-accessible surface (SAS) of two different sizes and solvent-excluded surface (SES) (Figure 2.17). SAS is the surface of a molecule that is accessible to a solvent. It is estimated using a “rolling ball” approach (Shrake and Rupley 1973) in which a sphere of solvent of a particular radius is used to probe the surface of the molecule, the surface is then described by the center of the probing sphere. We used the approximate radius of water molecule of 1.4Å and an additional of 3Å in order to see how the electrostatic potential changes with the distance from the molecule. SES uses the same “rolling ball” approach and describes the surface by the closest point of the probing sphere to the molecule atoms and this way represents the molecular surface not occupied by the solvent. In electrostatic potential calculations this surface defines the interface between molecule and the solvent where the dielectric constant changes from the ε_{in} of the molecule to the ε_{out} of the solvent.

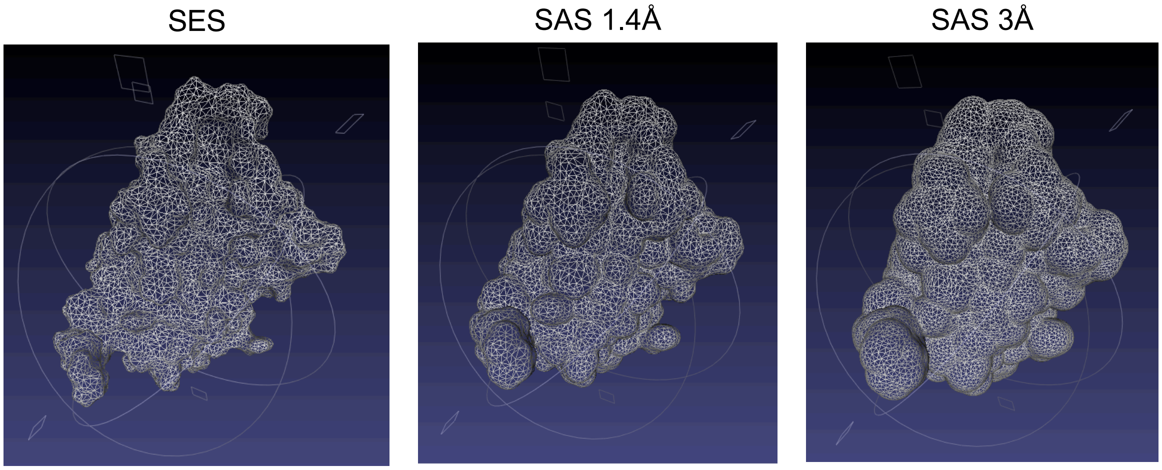


Figure 2.17 Three different protein surface estimations shown of the GH123 CA protein.

From the electrostatic potential values estimated in a grid covering entire space around the molecules we extracted grid points adjacent to the points of the triangulation of each surface type (Figure 2.17). We grouped these electrostatic potential values according to the atoms of loop residue they lie the closest to (Figure 2.18 and 2.19, top-right). The comparison of this way grouped electrostatic potential values of corresponding residues of the two analyzed molecules allowed to confirm quantitatively the difference in the electrostatic potential in the region of interest and to point to specific residues around which the differences are stronger.

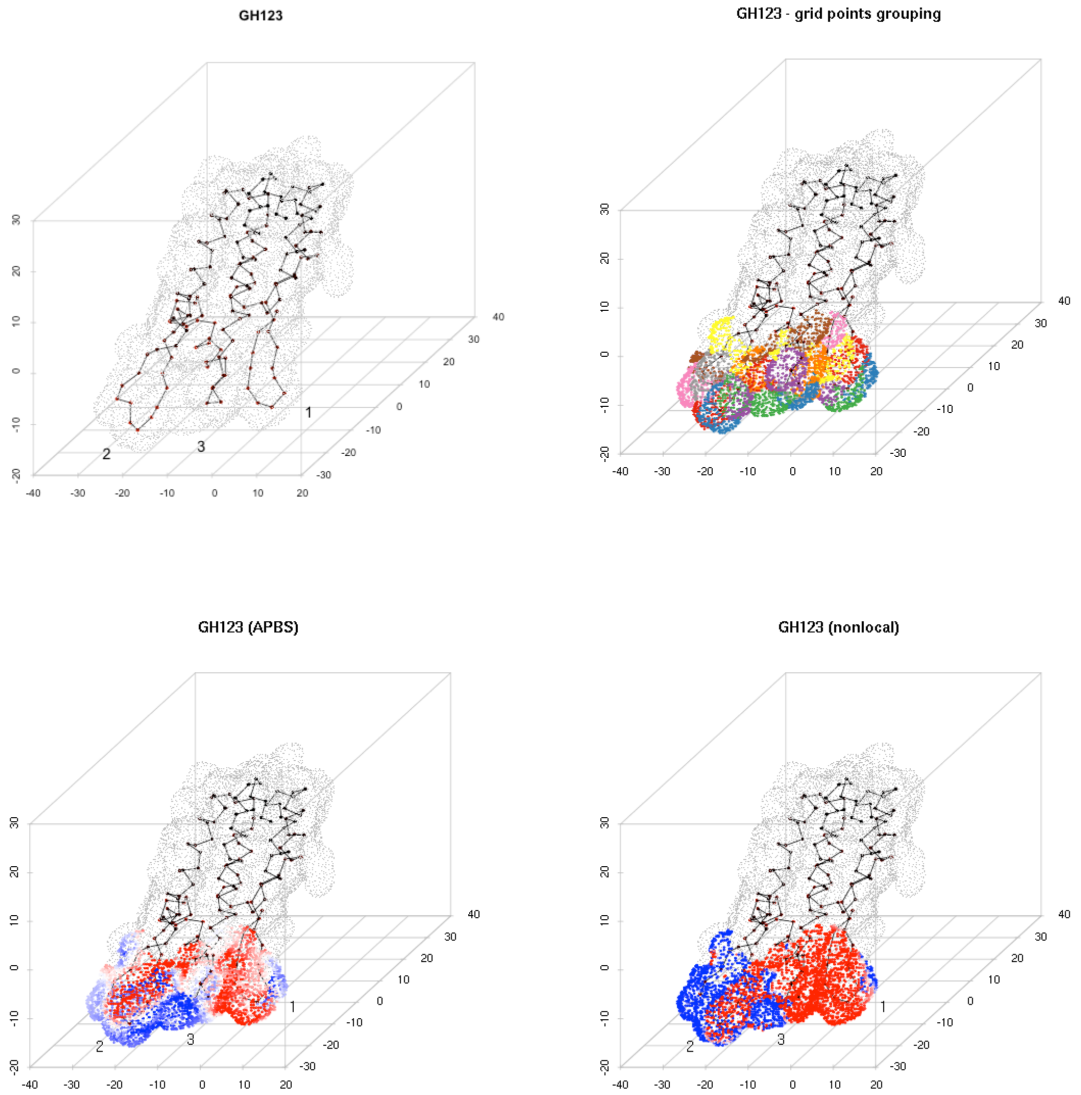
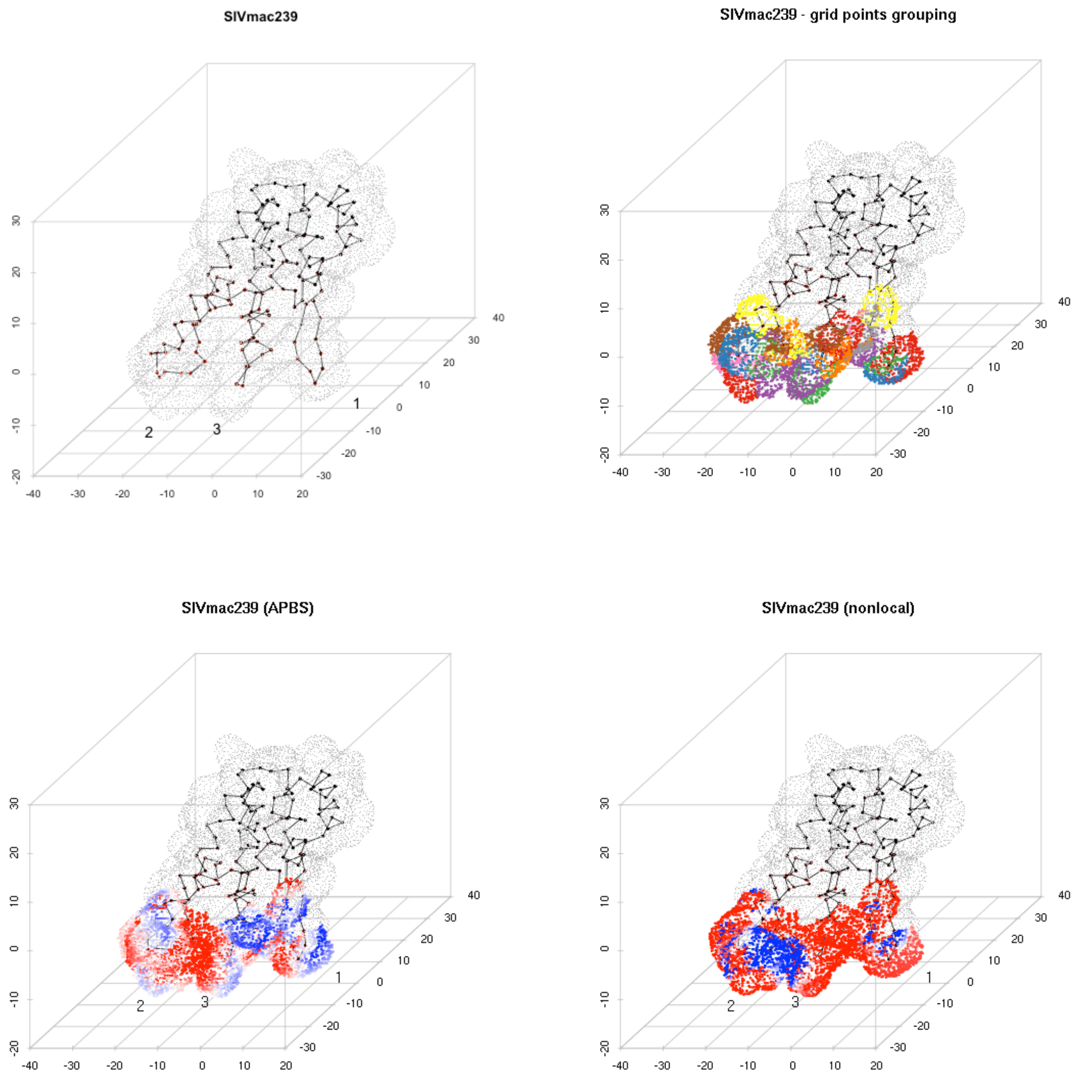


Figure 2.18 GH123 CA backbone and surface. Black lines trace the protein backbone by joining following Cα atoms. Loops are pointing downwards and numbered: 1 – L1, 2 – L4/5, 3 – L6/7. Gray dots mark the SAS of the protein. The coloring of the surface dots corresponds to: the partition of the surface according to the closest protein residue (top-right), electrostatic potential calculated using APBS (bottom-left), electrostatic potential calculated using nonlocal approach (bottom-right). Red color indicates negative and blue positive electrostatic potential.


SIVmac239

SIVmac239 - grid points grouping

SIVmac239 (APBS)

SIVmac239 (nonlocal)

Figure 2.19 Corresponding to Figure 2.18 plots of SIVmac239 backbone and surface.

The strongest difference in electrostatic potential between both variants could be observed on the surface of loop 4/5 with GH123 showing positive and SIVmac239 negative electrostatic potential. Six out of nine residues of this loop showed a significant difference in the mean electrostatic potential (Table 2.15) and a clear separation of the electrostatic potential values on the grid neighboring to the loop residues by both local ABPS and nonlocal methods (Figure 2.20).

Residues of loop 6/7 showed similar to loop 4/5 electrostatic potential differences however over a smaller number of residues and showing less agreement between the two electrostatic potential methods. Loop 1 showed an opposite pattern with GH123 having negative and SIVmac239 positive electrostatic potential. The differences were however smaller and only reported by the local APBS method.

Similar electrostatic potential differences although spanning a narrower range of values was observed on the SAS of the 3Å probe radius. This reflects the electrostatic potential decrease with the distance from the molecule surface.

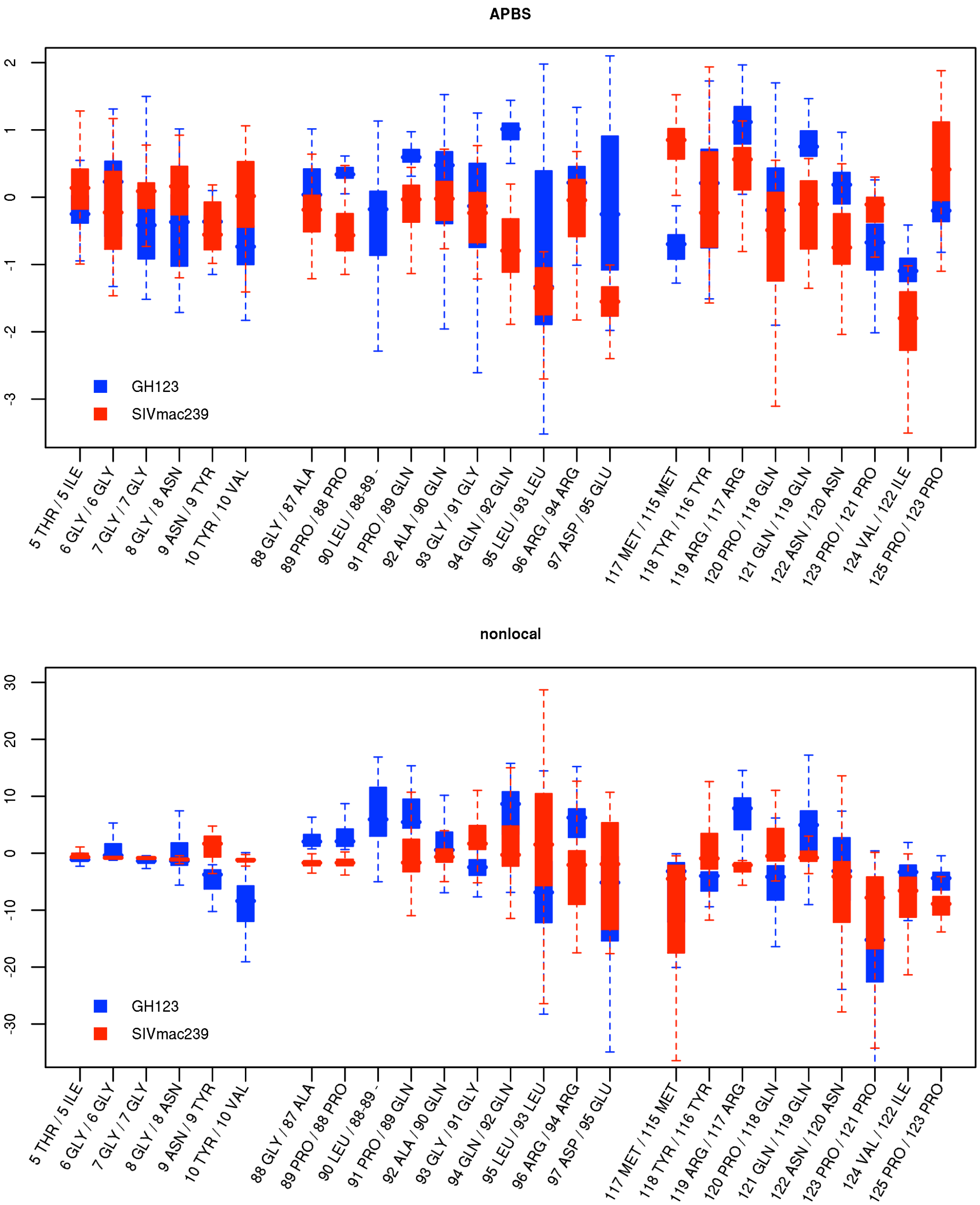


Figure 2.20 Electrostatic potential values on the SAS surrounding following residues on the three CA loops. Blue boxes represent electrostatic potential of GH123, red of SIVmac239. Upper plot shows values calculated using APBS, bottom the nonlocal method.

	Residue (GH123 / SIVmac239)	**APBS**			**nonlocal**		
		GH123	SIVmac329	p-value	GH123	SIVmac329	p-value
Loop 1	5 THR / 5 ILE	-0.206	0.064	<0.001	-1.049	-0.096	<0.001
	6 GLY / 6 GLY	0.025	-0.196	0.006	0.787	-0.805	<0.001
	7 GLY / 7 GLY	-0.315	0.024	<0.001	-1.283	-0.854	<0.001
	8 GLY / 8 ASN	-0.420	0.066	<0.001	0.058	-1.092	0.406
	9 ASN / 9 TYR	-0.463	-0.241	0.741	-5.668	2.697	<0.001
	10 TYR / 10 VAL	-0.782	0.021	<0.001	-8.827	-1.367	<0.001
Loop 4/5	88 GLY / 87 ALA	0.147	-0.248	<0.001	2.906	-1.700	<0.001
	89 PRO / 88 PRO	0.355	-0.522	<0.001	2.879	-0.524	<0.001
	90 LEU / 88-89 -	-0.426	-	-	6.567	-	-
	91 PRO / 89 GLN	0.603	-0.133	<0.001	6.543	-0.673	<0.001
	92 ALA / 90 GLN	0.047	-0.051	<0.001	1.282	-0.418	<0.001
	93 GLY / 91 GLY	-0.230	-0.269	0.076	-2.761	3.070	<0.001
	94 GLN / 92 GLN	0.895	-0.735	<0.001	7.148	0.820	<0.001
	95 LEU / 93 LEU	-0.958	-1.433	0.046	-6.661	2.234	<0.001
	96 ARG / 94 ARG	0.090	-0.227	<0.001	5.805	-3.992	<0.001
	97 ASP / 95 GLU	-0.045	-1.599	<0.001	-8.336	-3.481	0.001
Loop 6/7	117 MET / 115 MET	-0.765	0.799	<0.001	-6.437	-9.665	0.078
	118 TYR / 116 TYR	0.070	-0.069	0.167	-5.037	0.055	<0.001
	119 ARG / 117 ARG	1.022	0.405	<0.001	6.785	-2.802	<0.001
	120 PRO / 118 GLN	-0.094	-0.706	<0.001	-5.178	3.904	<0.001
	121 GLN / 119 GLN	0.802	-0.260	<0.001	4.308	0.340	<0.001
	122 ASN / 120 ASN	0.119	-0.674	<0.001	-4.078	-6.824	0.003
	123 PRO / 121 PRO	-0.782	-0.235	<0.001	-17.281	-11.590	<0.001
	124 VAL / 122 ILE	-1.200	-1.906	<0.001	-6.233	-8.141	0.003
	125 PRO / 123 PRO	-0.250	0.455	<0.001	-4.804	-12.468	<0.001

Table 2.5 Mean electrostatic potential on the surface surrounding residues of the three loops of GH123 and SIVmac239 CA variants calculated using the local APBS and nonlocal methods. Colors indicate significant difference (wilcoxon test) between the electrostatic potential of the two variants with positive electrostatic potential GH123 and negative SIVmac239 marked in red and negative GH123 and positive SIVmac239 marked in blue.

Precise calculations of the interaction electrostatics are very challenging because of the large surfaces involved and the large structural changes that can occur upon binding. Through this quantitative approach and with the use of two methods a trend could be observed of a positive electrostatic potential on the surface of the restricted CA variant and negative of the non-restricted. According to the experimental data SIVmac239 variant was not restricted by neither of the RH or CM TRIM5α proteins as opposed to the GH123 variant that was strongly restricted by both primate TRIM5αs. Presence of electrons on the surface of loop 4/5 and 6/7 as indicated by the electrostatic potential of the GH123 CA might be therefore a prerequisite for the interactions with RH and CM TRIM5αs. These two loops form the most outward pointing part of the CA protein and share neighboring protein surface, while loop 1 is directed slightly apart. Complementary to GH123 surface charge distribution at the binding site of the host protein might be necessary for the binding. Similar studies of TRIM5α surface electrostatics could therefore help to point to the specific site of this interaction.

2.4.5 Discussion

In the compilation study of the experimentally tested variants of the two interaction partners – viral CA and host TRIM5α we identified sequence, structure and electrostatic determinants of the interaction. This systematic analysis of an experimentally phenotyped sequence dataset pointed to specific sequence positions, sequence regions and their physicochemical features, structural shapes and surface electrostatics that appeared to be significantly related to the interaction. The analyzed dataset is highly incomplete with respect to phenotype and sequence variation, however the analysis of the entire datasets allowed for inferring specific features of the CA protein that were not observed in individual experimental studies. Moreover, the initial study of the entire CA dataset allowed to select specific examples for an in-depth electrostatic potential analysis. Due to the lack of TRIM5α crystal structure a similar search of significant structural features of TRIM5α is difficult to perform. The structural physicochemical determinants of the interaction could indicate potential drug strategies based on TRIM5α restriction. Following these results, more experimental testing will be performed in the Osaka lab.

2.5 Conclusions

Even though host and virus interact throughout the entire virus life cycle, the example of TRIM5α-CA interaction shows the importance of detailed studies of individual molecular interactions between the virus and its host. Specific mutations in the two interaction partners show a strong impact on the efficiency of infection.

Mutation studies of the two proteins underscore that the barrier against interspecies transmission is potentially fragile. Studies of TRIM5α-CA can explain the host range of a given lentivirus species which indicates how on the scale of an individual interaction one can observe determinants of virus-host adaptation. Additionally, studies on the scale of an individual interaction can be of therapeutic interest, as knowing the exact interaction mechanism of the CA and TRIM5α could indicate potential ways of for inducing of the HIV restriction in humans.

Among the downfalls of studies on the individual interaction scale is the limitation of the number of analyzed virus variants. Given the variability of the HIV and the cost of analyses of individual virus variants such as presented in this chapter, the results of studies of individual interactions might be difficult to scale to entire virus populations. The structural and biochemical determinants of an interaction of a low number of analyzed virus proteins might not be representative of large HIV populations.

CHAPTER 3 – Virus population scale

HIV-1 coreceptor tropism encoded in the V3 loop sequence and structure

HIV is a highly mutating pathogen (Korber, Gaschen et al. 2001). The lack of a proof-reading mechanism in the reverse transciptase (Drake 1993), the high replication rate, rapid viral turnover (Perelson, Essunger et al. 1997), selection pressure for adaptation and evasion of therapy (Condra 1998) or immune recognition (Price, Goulder et al. 1997; Poignard, Sabbe et al. 1999; Allen, O'Connor et al. 2000) generate virus populations of radically divergent virus forms. Each HIV-infected individual is exposed to a virus population of a diversity equal to that of influenza sequences world-wide in any given year (Korber, Gaschen et al. 2001). Given such a high level of virus variability computational models of entire virus populations are more representative of the host-virus interactions than analyses of individual variants. In this chapter we describe two studies on the scale of virus population. Both studies comprise analyses of a highly variable part of the viral genome involved in an important interaction with the host upon virus cell entry.

3.1 Background

Among virus genes, the highest amount of genetic variability is observed in the envelope gene (Env) (Hahn, Gonda et al. 1985). The env protein is composed of two parts – glycoprotein 41 (gp41) anchored in the virus membrane and the exterior glycoprotein 120 (gp120). The envelope glycoproteins form spikes displayed on the surface of the virion. The surface of the spike is composed of gp120, attached through non-covalent interactions to the gp41 glycoprotein (Kowalski, Potz et al. 1987; Lu, Blacklow et al. 1995). Being exposed on the virus membrane, gp120 can elicit virus-neutralizing antibodies (Profy, Salinas et al. 1990). High genetic variability of this gene is therefore a way of the virus to elude the immune system and escape recognition. On the other hand, the same gene mediates virus entry into host cells. The parts of the gene sequence that involve host receptor binding sites are under selection pressure to preserve the capacity of host cell entry.

HIV entry into human cells is a multi-step process (Figure 3.1). It is initiated through binding of the viral envelope glycoprotein gp120 to the cellular CD4 receptor (Chan and Kim 1998; Pierson and Doms 2003). CD4 binding induces conformational changes in the gp120 glycoprotein (Sattentau and Moore 1995), some of which involve the exposure and/or formation of a binding site for specific chemokine receptors. These chemokine receptors, mainly CCR5 and CXCR4 for HIV-1, serve as obligate second receptors for virus entry (Feng, Broder et al. 1996; Trkola, Dragic et al. 1996; Moore 1997). The interaction of gp120 with the coreceptor induces a series of further rearrangements in the envelope glycoproteins that trigger fusion of virus and cell membranes (Chan and Kim 1998).

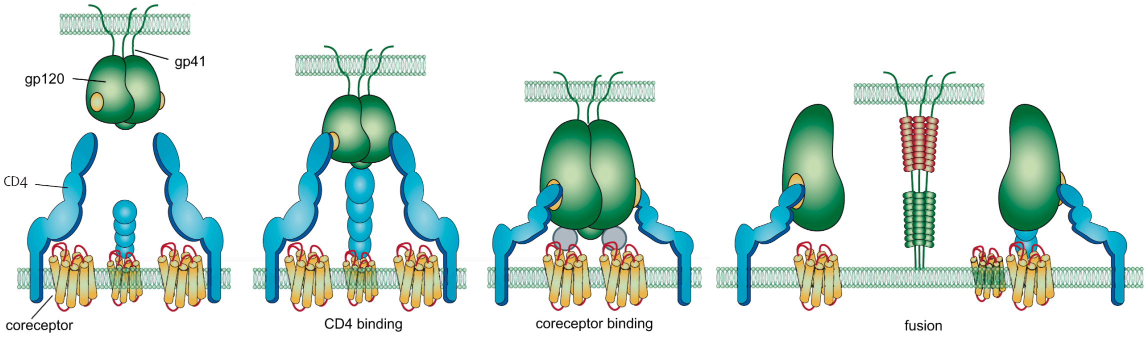


Figure 3.1 Following steps of the HIV cell entry: binding to CD4 receptor, binding to coreceptor and fusion. Adapted from (Este and Telenti 2007).

It has been shown that viruses binding to CCR5 (R5 viruses) are almost exclusively present during the early asymptomatic stage of the infection whereas CXCR4-binding viruses (X4 viruses) may emerge in later phases of the infection and are associated with a CD4$^+$ T-cell decline and progression towards AIDS (Miedema, Meyaard et al. 1994; Berger, Murphy et al. 1999). The specificity of the virus to use one of the coreceptors is termed tropism. Before the coreceptors were identified, two phenotypic variants were recognized according to the virus ability of forming syncytia in MT-2 cells. Already at that time, syncytium-inducing (SI) and non-syncytium-inducing (NSI) viruses were observed to have a different impact on the disease progression in infected people (Schuitemaker, Kootstra et al. 1991). There is a high correlation between CCR5-tropic and NSI viruses, on the one hand, and between CXCR4-tropic and SI viruses, on the other hand. One of the big still unresolved questions of the HIV-tropism research is whether the emergence of X4 and SI virus is a cause of advanced progression towards CD4$^+$ T-cell depletion and the rise of AIDS symptoms or appears as a result of these phenomena (or both). The evolutionary reasons for the development of these two HIV variants remain unknown.

The finding that humans who lack CCR5 expression due to the Δ32 mutation are resistant to HIV-1 infection (Huang, Paxton et al. 1996) stimulated the research of CCR5 inhibitors which culminated with the licensing of Maraviroc (MVC) (Dorr, Westby et al. 2005) for clinical use in 2007. The unknown mechanism driving the coreceptor switch, its role in the disease progression and the new treatment strategies targeting R5 viruses call for efficient and accurate routine monitoring of coreceptor usage and for a better understanding of the determinants of the coreceptor binding. Phenotypic assays exist for experimental testing of viral tropism (Whitcomb, Huang et al. 2007). However phenotyping via such assays involves sending patient samples to specialized laboratories offering satisfying technological and security requirements. Phenotyping results are returned in ten or more days. This procedure of is therefore expensive for patient diagnosis in a daily clinical practice. In contrast to phenotyping, sequencing is cheaper, more accessible and faster and thus has become a standard lab procedure throughout Europe. However sequence data requires the interpretation in terms of the viral tropism which can only be done with the use of computational methods.

The principal determinant of the specificity of HIV to a chemokine receptor is the third variable (V3) loop of gp120 (Fouchier, Groenink et al. 1992; Shioda, Levy et al. 1992; Speck, Wehrly et al. 1997). Binding to CCR5 coreceptor only (R5 viruses), to CCR5 and CXCR4 (dual-tropic viruses) or to CXCR4 only (X4 viruses) is determined predominantly by the sequence and structure of this region (Jensen and van 't Wout 2003). The V3 loop is an extended structure protruding ~30Å from the CD4-bound gp120 core (Huang, Tang et al. 2005). It contains a conserved base, a flexible stem that rigidifies on binding and a tip in a β-hairpin conformation (Figure 3.2).

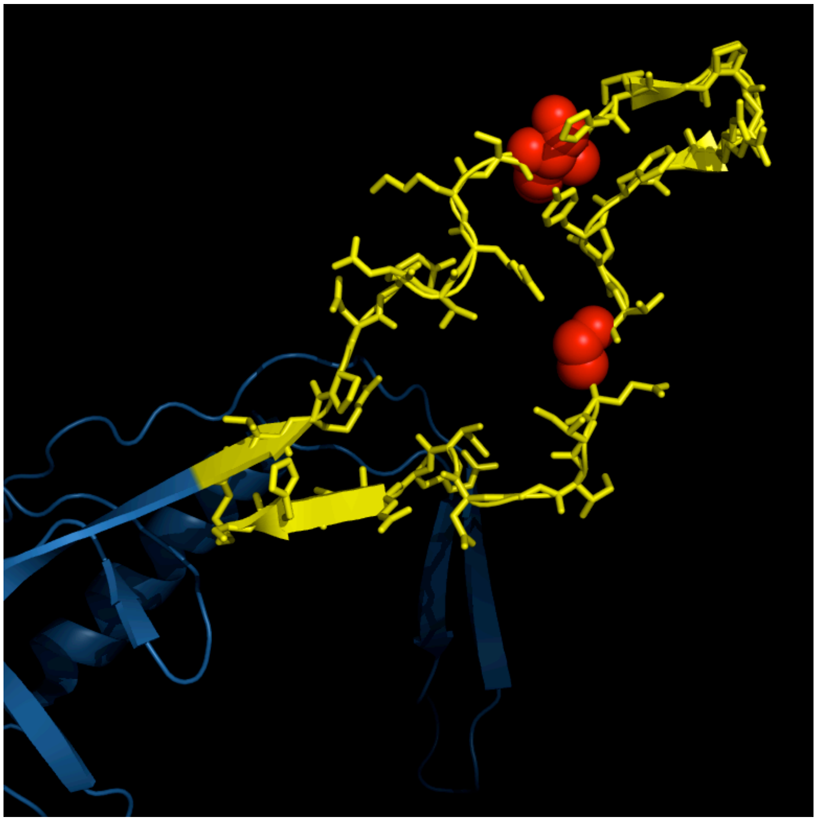


Figure 3.2 V3 loop structure (PDB ID: 2B4C). The loop is colored in yellow, with atoms of residues on position 11 and 25 represented by red spheres. Other parts of the gp120 are traced by a blue ribbon.

As an alternative to costly phenotypic assays for monitoring coreceptor usage, computational methods were developed aiming at predicting the viral tropism based on the V3 sequence. The 11/25 rule was proposed as an initial approach for inferring coreceptor usage. It is based on the observation that a positive charge at either of the 11th or 25th residues in the V3 region is indicative of an X4 virus (Fouchier, Groenink et al. 1992; Shioda, Levy et al. 1992). Due to its simplicity, the 11/25 rule has been commonly used although it has been shown that for many V3 variants changes at positions 11 or 25 are neither necessary nor sufficient for the tropism switch (Jensen and van 't Wout 2003). As more phenotypic data became available more sophisticated methods appeared based on position-specific scoring matrices (PSSM) (Jensen and van 't Wout 2003), decision trees (Pillai, Good et al. 2003), neural networks (Resch, Hoffman et al. 2001) and support vector machines (SVM) (Sing, Beerenwinkel et al. 2004). The sequence-based methods of coreceptor usage prediction rely on binary encoding of the letter representation of amino acids at each position of the V3 loop and use statistical approaches to build predictive models and to infer residues strongly related to the phenotype. One drawback of the sequence-based methods of viral tropism prediction is the limited insight they offer into the sequence determinants and sequence features of the two virus types. Methods encode a classification rule in a form of a multidimensional hyperplane or decision algorithms that are not directly interpretable. The results of classification do not point to specific sequence characteristics of viruses of each class nor to the potential reasons for classification errors. Moreover, the binary encoding of the amino acids in the V3 loop sequence lacks the information on the actual physicochemical similarities among different amino acid types. The sequence-based methods do not account for the physicochemical and structural properties of the loop that might be determinant for coreceptor binding.

Since the structures of gp120 including V3 were determined by x-ray crystallography (Huang, Tang et al. 2005; Huang, Lam et al. 2007) new coreceptor prediction methods were developed (Sander, Sing et al. 2007; Dybowski, Heider et al. 2010) integrating the structural information in the prediction process. Sander at al. (Sander, Sing et al. 2007) proposed a distance-based descriptor of the spatial arrangement of physicochemical properties of the loop. They found that providing the distance information coupled with modeling of the loop side-chains together with the binary sequence encoding outperforms the methods based on sequence alone. Dybowski et al. (Dybowski, Heider et al. 2010) developed a two-level approach that considers two physicochemical properties of the loops (electrostatic potential and hydrophobicity) on the first level and combines the results on the second level of the prediction procedure. This two-level approach resulted in improvement in prediction accuracy as compared to prediction based on sequence alone. Even though including the structural information into the prediction is a step forward to understanding HIV coreceptor usage, both methods have certain limitations. The method by Sander et al. is based on molecular distances that do not offer direct interpretation of the structural determinants of tropism. Dybowski et al. analyze only two physicochemical features in their method while a systematic analysis of a larger feature set of the V3 loop would allow for distinguishing other features important for tropism. Both methods involve costly computational operations as electrostatic potential calculation or side-chain modeling that reduce their efficiency in comparison to sequence-based approaches. With the increasing use of high-throughput sequencing technologies in HIV research and patient diagnosis these methods can be expected not to fulfill the growing efficiency requirement.

All except one (Sing, Beerenwinkel et al. 2004) of the above mentioned methods sequence- and structure-based were developed and assessed on clonally derived data.

This type of data is inferred in lab conditions from amplified and cloned virus strains. In contrast to the clonal data, clinically derived virus data are obtained through bulk sequencing and phenotyping of virus populations obtained directly from patient blood plasma without prior amplification and cloning. Analysis of clinically derived data is burdened with two major problems. First, clinical data contain ambiguities in the sequence and phenotype reflecting the variability of the virus population. When two or more loci display variation, the information on whether and how often these variants occur on the same DNA molecule is lost. Second, small minorities of the virus population might be undetected through bulk sequencing. Sanger capillary sequencing can only detect the mutations of a mixed sample if their frequency exceeds a threshold of ~20%. Clinically derived sequence represents therefore a consensus amino acid composition of the the few dominating virus strains. This shortcoming can result in an inaccurate genotype-phenotype relationship, which cannot be rectified without additional experimental effort, such as individual virus cloning. These problems have been shown to pose additional challenges for *in silico* coreceptor prediction (Sing, Low et al. 2007). Nevertheless, correct prediction of the clonal data is crucial for successful patient surveillance and treatment. Computational methods of coreceptor prediction should therefore be aimed at classifying clinical data.

Since the approval of the CCR5 inhibitor MVC, patients undergo treatment involving this new drug. What level of the X4 viruses in the patient virus population is acceptable for the therapy to succeed and whether the treatment would cause the emergence of the minority population and consecutive therapy failure is still debated. However with the increasing amount of data concerning MVC treatment computational methods should be developed targeted at predicting therapy outcome.

In this chapter we present two studies of the HIV V3 loop on the virus population scale. The studies are aimed at characterizing and classifying virus sequences in relation to the phenotype while addressing the shortcomings of the methods mentioned above – interpretability of the prediction results, efficiency of the structure-based prediction, classification of the clinically-derived data and prediction of the therapy outcome.

3.2 V3 loop sequence space analysis

Here we present the results of a comprehensive analysis of the HIV V3 loop sequence space. This study was published in: Bozek, K., Thielen, A., Sierra, S., Kaiser, R., Lengauer, T. V3 loop sequence space analysis suggests different evolutionary patterns of CCR5- and CXCR4-tropic HIV. *PLoS One.* 2009 Oct 9;4(10):e7387. The aim of this project was to better understand the evolutionary mechanism driving the emergence of X4 viruses. We analyze all available phenotyped V3 loop sequences from the Los Alamos database (http://www.hiv.lanl.gov/). Using different sequence distance measures and visualization methods we describe the arrangement of the sequences in their multidimensional sequence space with regards to the phenotype. The results reveal a relatively high conservation of R5 and NSI strains as compared to more diverse X4 and SI strains evolving in an apparently unconstrained manner. We find that the arrangement of the sequences imparts one of the reasons for the inaccuracy of sequence-based methods for coreceptor prediction. We also use the location of the V3 loop sequence in sequence space to improve the accuracy of the prediction of virus tropism. Finally, we investigate the relation between the location of V3 loop sequences in sequence space and the associated clinical markers such as CD4$^+$ T-cell level. Sequences of patients with a functioning immune system tend to be located close to each other in sequence space and thus are likely to share common features whereas, with decreasing CD4$^+$ T-cell counts the conservation of the V3 loop among patients decreases and the diversity of possible viral genotypes increases. These results support the hypothesis of the immune system initially imposing strong selective pressure on the viral envelope gene. Once the immune system is compromised, this pressure diminishes which enables the virus to undergo less restrained variation.

3.2.1 Data

Using the Los Alamos database (http://www.hiv.lanl.gov/) we defined two sets of labelled V3 loop protein sequences: the labels of the first set (NSI/SI dataset) were attributed according to the annotation of non-syncytium-inducing (NSI) and syncytium-inducing (SI) strains. Those of the second set (R5/X4 dataset) were attributed according to the annotation of the sequences concerning coreceptor usage – CCR5-, CXCR4- and dual-tropic (R5X4) strains. In order to prevent samples from a single patient to dominate any of the two sequence sets and to analyze viral evolution among hosts rather than patient-specific selection pressures, we limited the datasets to contain one randomly chosen sequence from each patient. The two sets contain 1096 and 859 V3 loop amino acid sequences, respectively, with an 85% prevalence of NSI and CCR5-tropic strains, respectively.

3.2.2 Distance distribution

We used four sequence distance measures to compare the V3 sequences: Hamming distance, Blosum62 matrix (Henikoff and Henikoff 1992), difference in amino acid charge and size, and difference in amino-acid composition at positions significant for the phenotype, as reported by Sing et al. (Sing, Low et al. 2007). Different distance measures resulted in the same pattern of sequence separation. Therefore we used the Blosum62 matrix as the only distance measure in other parts of this analysis.

For all considered sequence distance measures the following pattern among the V3 loop sequences was observed. R5 and NSI sequences were positioned close together while, in contrast, X4 as well as SI sequences were much more widely spread out in sequence space. The left panel of Figure 3.3 shows the distribution of Blosum62 distances between pairs of sequences from the NSI/SI dataset of the same and of different phenotypes. The mean distance of pairs of SI sequences (red curve) is almost twice the mean distance of the NSI sequences (blue curve). The distribution of the distances for pairs of SI sequences has also a larger variance than the one of the NSI sequences. The fact that the mean distance between pairs of sequences of opposite phenotypes is smaller than the mean distance of pairs of SI sequences implies that the SI sequences are widely spread out throughout sequence space and show no apparent common pattern of evolution. Distances between sequences in the R5/X4 dataset exhibit a similar pattern (Figure 3.3, right panel) with the dual-tropic sequences spread out less, on average, than the exclusively CXCR4-tropic sequences and more than the R5 sequences. The different distance measures result in the same pattern of sequence separation we therefore chose the Blosum62 matrix as the distance measure in all other parts of this analysis.

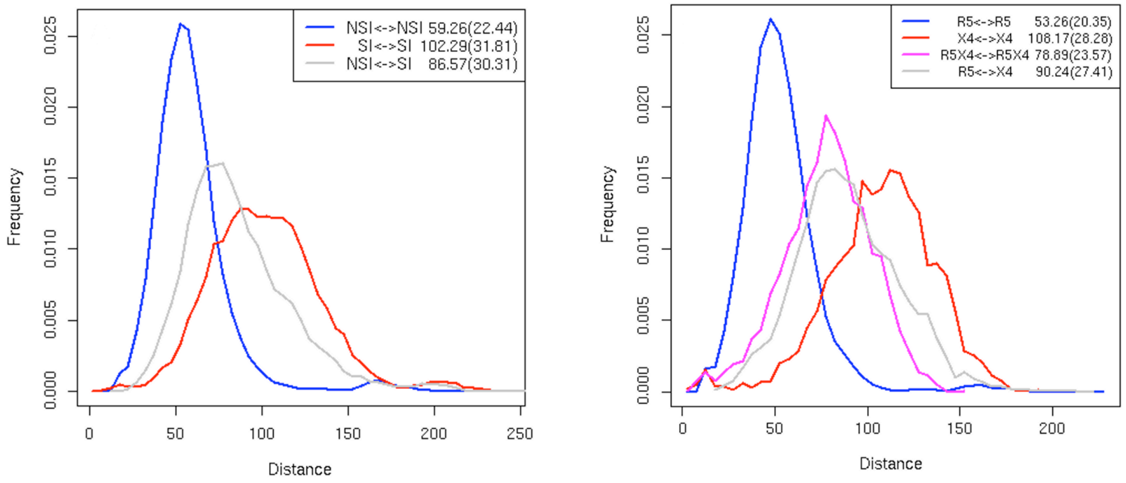


Figure 3.3 V3 loop sequence distance distribution. Shown is the distribution of Blosum62 distances between pairs of sequences of the same (SI vs SI, NSI vs NSI) and different (SI vs NSI) phenotypes (left) and of the same (R5 vs R5, X4 vs X4, R5X4 vs R5X4) and different (R5 vs X4) tropisms (right). The mean value and standard deviation of each of the distributions are indicated in the inserted boxes.

3.2.3 Clustering

To further investigate the observations gained from the sequence distance distribution we performed sequence clustering. Both datasets were clustered hierarchically, using complete linkage clustering. We analyzed the tendency of viral sequences of different phenotypes to form clusters depending on the cluster diameter defined as the distance between the two most distant elements of the cluster. Clustering with a given upper limit for the diameter was achieved in an iterative procedure of merging two closest clusters in each step of the procedure until no two clusters can be merged without generating clusters of a diameter above the predefined limit. In complete linkage clustering, the distance between two clusters is defined as the largest distance between two elements, one in each cluster. Only clusters containing at least 1% of all sequences in the dataset were considered, the sequences belonging to smaller clusters as well as singletons were defined as *unclustered*. Additionally we used a notion of a cluster size to compensate for the imbalance between the amount of R5/NSI and X4/SI sequences. The number of sequences of a given type in a cluster is multiplied by the ratio of the number of sequences of all other types in the full dataset over the number of sequences of the same type. This weighting scheme allows for considering smaller clusters of an underrepresented phenotype as significant.

The measure of silhouette value (Rousseeuw 1987) was used to choose one clustering whose individual cluster structures should be investigated. The silhouette value of a sequence in a cluster is defined as the difference between the average distance of the sequence to sequences in other clusters and to sequences in the same cluster. The silhouette value of a cluster is the average silhouette value of its sequences. The silhouette value of a clustering is the average of the silhouette values of its clusters. This measure can be used as a quantitative indicator of the quality of a clustering - larger values represent partitions containing clearly separated clusters. We calculated the average silhouette values for partitions obtained in successive steps of the hierarchical clustering and selected the clustering showing a maximal silhouette value, among those that cluster more than 50% of the sequences in a dataset and contain more than one cluster. We called the clustering resulting from this procedure the *selected* clustering.

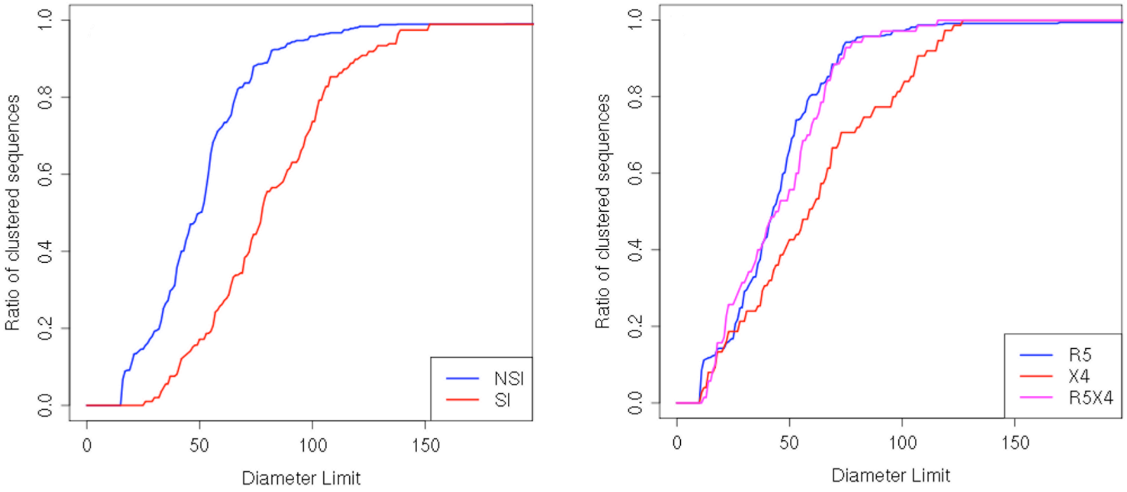


Figure 3.4 Clustering of the V3 loop sequences. The plot illustrates clustering trends of sequences of the NSI and SI phenotype (left) and of R5, X4 and dual-tropic sequences (right). The diameter limit, plotted on the x-axis, is defined as the distance between the two most distant elements of a cluster. The y-axis indicates the fraction of sequences in the dataset falling into any of the clusters below the diameter limit. Minimal cluster size is 1% of all sequences in the dataset, sequences of clusters of a smaller size as well as singletons are considered as unclustered and are not counted.

Clustering of the NSI/SI and R5/X4 datasets displayed a more pronounced grouping trend of the NSI and R5 sequences than of the SI and X4 sequences. In the initial steps of the iterative clustering procedure only tight clusters of highly similar sequences were formed. The clusters contained mainly R5 and NSI sequences. The X4/SI sequences are clustered only in clusters of a relatively large diameter – clusters of diameter large enough to contain 90% of the NSI sequences group only 50% of the SI sequences in the NSI/SI dataset (Figure 3.4, left panel). The clustering of the R5/X4 dataset showed that dual-tropic sequences cluster relatively less than R5 but more than X4 sequences – a clustering containing from 50% to 80% of the CCR5 sequences included on average 10% less of the dual-tropic sequences (Figure 3.4, right panel).

Next, we inspected the selected clustering in detail. In the NSI/SI dataset we obtained one major cluster containing 60% of all sequences, most of them of the NSI phenotype, and two smaller clusters each one containing 15% of all sequences, the first including solely NSI sequences, the second including equal percentage of sequences of both phenotypes. In the R5/X4 dataset the clustering contained two main clusters comprising 35% and 32% of all sequences respectively, containing mainly R5 and dual-tropic sequences, in the range of 19 to 38% of the sequences of each type in the R5/X4 dataset. Details of individual clusters are listed in Table 3.1. We observed that clusters containing mainly X4/SI sequences are rare (only cluster 4 and 5 in the NSI/SI dataset); sequences of this phenotype tend to associate predominantly with clusters of R5 sequences. Additionally, in both datasets X4/SI sequences are highly over-represented among the unclustered sequences (p-value < 0.001, chi-square test).

	NSI/SI			R5/X4			
cluster	all	NSI	SI	all	R5	X4	dual
1	0.59	0.65	0.33	0.35	0.38	0.09	0.30
2	0.15	0.17	0.06	0.32	0.37	0.05	0.19
3	0.14	0.14	0.15	0.06	0.07	0	0.04
4	0.02	0	0.07	0.05	0.06	0	0.01
5	0.01	0.01	0.03	0.05	0.06	0	0

Table 3.1 Five largest clusters in the NSI/SI and R5/X4 dataset clustering. Numbers indicate what fraction of the whole dataset is grouped in a given cluster (column "all") and what is the ratio of the sequences of a given phenotype to all sequences in the respective cluster.

Subsequently, we used data density estimation to examine the structure of individual clusters in the selected clustering. Data density is an indicator of the sequence concentration in a given part of sequence space relative to the rest of the space. We used an unsupervised learning method of data density estimation via classification (Hastie, Tibshirani et al. 2001). This method allows for determining regions in sequence space in which the sequence density is significantly higher than average. For this purpose, we augmented the datasets with 500,000 random reference data points distributed uniformly over the high-dimensional sequence space. The number of the reference data points was chosen as a balance between the computational load and sufficient space population for accurate density estimation. A binary logistic regression model where true data points are assigned the value 1 and the generated reference data points are assigned value 0, was fitted using maximum likelihood (ML) estimation. The value returned by the fitted logistic regression for each of the true data points was treated as the probability of a point to be sampled by a distribution producing the analyzed dataset. The log-odds of this probability for each data point represent the local density of the original data relative to the generated reference data. Data density of a single cluster was calculated as the mean of log-odds of the cluster sequences. The larger the density of a cluster is, the more highly concentrated set of data points it contains.

Additionally, we measured the amount of positive selection among sequences in each cluster. For this purpose each of the protein sequences has been assigned its corresponding DNA sequence from the Los Alamos database. We defined as a cluster center the location of a sequence with minimal distance to all other sequences in the cluster. The amount of positive selection ω_k exerted on a sequence k in a cluster is defined in terms of the mean of the ratio of non-synonymous to synonymous substitution rate ω_{ik} (Yang and Nielsen 2000) of the given sequence and other sequences i having a smaller distance D_i to the cluster center:

$$\omega_k = \frac{\sum_{i=1}^{n} q_{ik}\omega_{ik}}{\sum_{i=1}^{n} q_{ik}},$$

where n is the number of sequences in the cluster, $q_{ij} = 1$ if $D_i < D_j$ and 0 otherwise. Non-synonymous to synonymous substitution rates were calculated using the Yang and Nielsen method (Yang and Nielsen 2000) implemented as a part of PAML software package (Yang 2007).

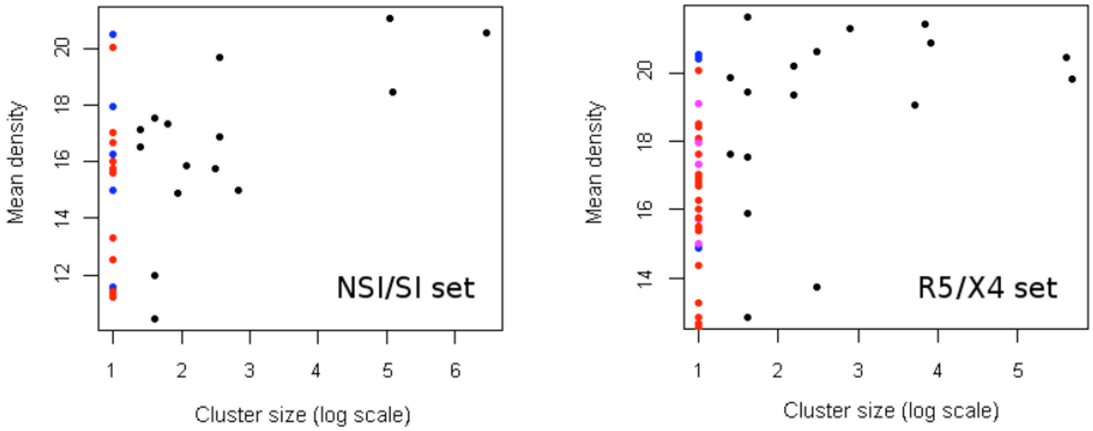


Figure 3.5 Data density in clusters of the selected clustering. Mean data density of sequences in clusters of the selected clustering of the NSI/SI (left) and R5/X4 dataset (right) is plotted against the cluster size. Unclustered sequences are represented by dots at the value 1 on the x-axis with colors corresponding to their phenotype: NSI/R5 sequences in blue, SI/X4 in red, R5X4 in magenta. Clusters are represented by black dots, cluster sizes are displayed in log scale. Large clusters are formed in denser parts of the data space than unclustered sequences. SI/X4 sequences remain predominantly unclustered.

The relation of the cluster size and the density in sequence space is illustrated in Figure 3.5. Unclustered sequences occur in less dense parts of sequence space and are predominantly SI or X4. Detailed inspection of the individual cluster structures allowed for relating data density and the amount of positive selection on a sequence to the position of a given sequence within its cluster. The two largest clusters of the R5/X4

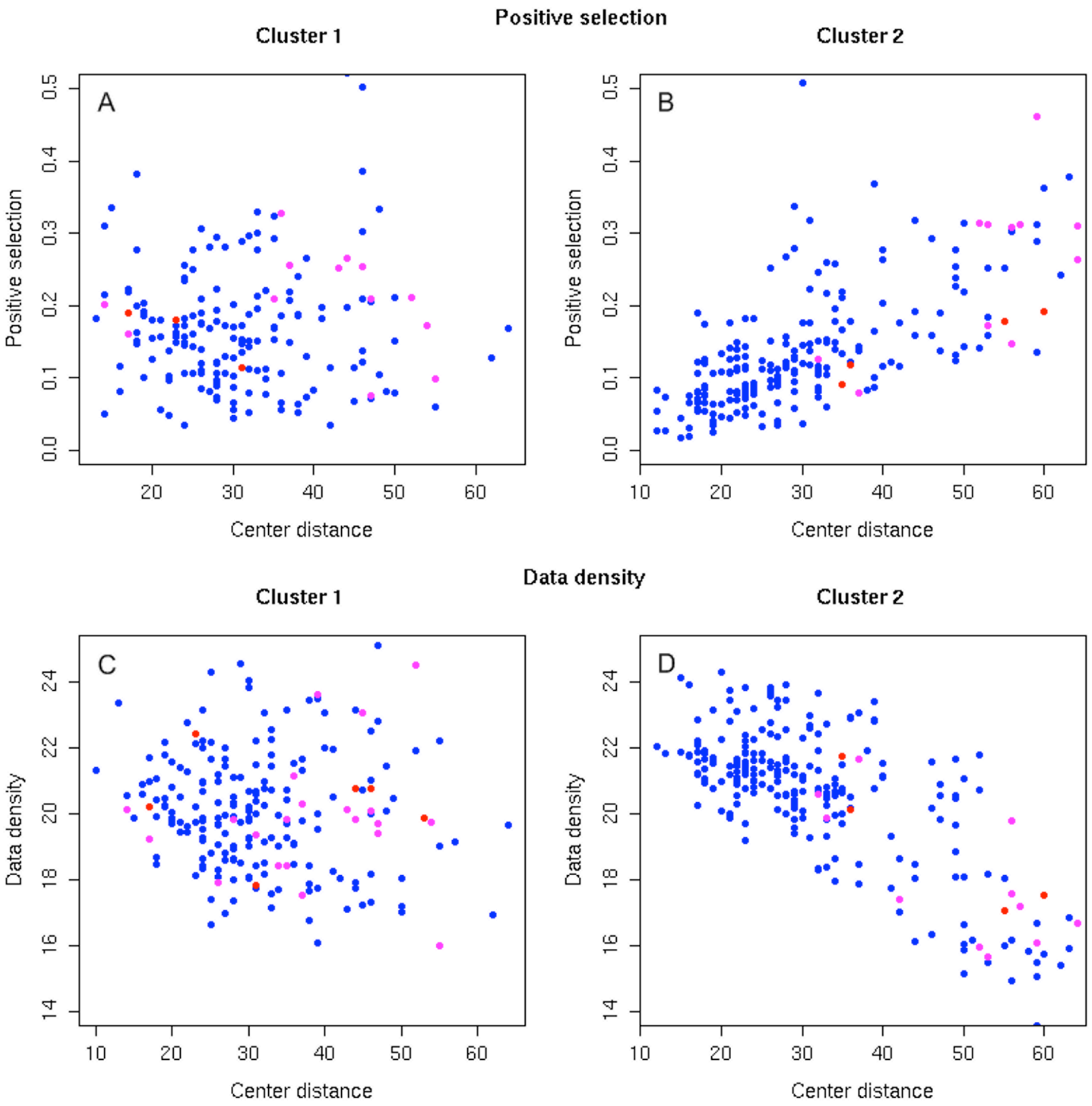


Figure 3.6 Structure of two largest clusters in the R5/X4 dataset. Cluster 2 (right panels) shows a clear correlation of the distance of a sequence to the cluster center with the amount of positive selection on this sequence (B) and with the data density of its sequences (D). Dots on diagrams are colored according to sequence tropism (R5 – blue, X4 – red, R5X4 – magenta). The plots indicate that cluster 2 contains a dense center composed of conserved R5 sequences and sparser brims composed of X4 and R5X4 sequences that are subject to weaker selection pressures. No such pattern can be observed in the case of cluster 1 (A and C) which seems not to contain a coherent centre but comprise a medium density region of sequences of various tropisms.

dataset are depicted in Figure 3.6. In the case of cluster 2 we observed a strong correlation of the sequence distance from the cluster centre with the data density at the location of the sequence as well as with the amount of positive selection exerted on the sequence (correlation coefficient of -0.76 and 0.71, respectively). The results pointed that there was a dense center in cluster 2 grouping most of the cluster sequences and sparse brims where the concentration of sequences was smaller. The selection pressure in the center of such clusters was greater and reduced farther from its center. X4 and dual-tropic sequences preferentially occupy the peripheral regions of the cluster (Figure 3.6). However, no such cluster pattern could be observed in the case of cluster 1. Data density analysis showed that this cluster contained sequences spread over a similar density range independently of their position within the cluster. There was no clear distribution of variation in selection pressures in cluster 1 either.

3.2.4 Phylogenetic analysis

Next, we performed phylogenetic analysis of the V3 loop datasets. In this analysis we used the Splitstree software (Huson 1998). The split decomposition method (Bandelt and Dress 1992) relaxes the usual requirement of representing the data in tree form, in order to elicit where the underlying distance matrix does not reflect a tree structure. A Splitstree network is tree-like, in general, but also represents the divergence of the phylogenetic data from the tree form by sets of parallel edges that expand the tree to a more complex network. The Blosum62 matrix was used as the distance measure in the Splitstree analysis.

The visualization of the V3 loop data via the Splitstree diagrams (Figure 3.7) showed the separation of viral strains of different coreceptor usage. Both datasets were too large to be displayed in one tree. We therefore generated trees of randomly sampled subsets of sequences. Figure 3.7, left panel shows an example Splitstree of a randomly sampled set of 25 sequences of each phenotype in the NSI/SI dataset. Both the lengths of splits (sets of parallel edges in the centre of the graph) and of single branches connecting data nodes to the rest of the tree clearly discriminate between these two types of sequences. SI sequences (represented by red dots) are located on long tree branches that reflect the larger evolutionary distance between them and other sequences in the dataset. NSI sequences (blue dots) are located on shorter branches and grouped in more tree-like clades. A similar tree generated for a sample of 20 sequences of each tropism from the R5/X4 dataset is shown in Figure 3.7, inset panel. The dual-tropic sequences (represented by magenta dots) showed an intermediate character between the R5 (blue dots) and X4-tropic (red dots) sequences. Both their branch lengths and localization on the tree supported the view of the dual-tropic sequences combining characteristics of the two other sequence types or being an intermediate form in their evolution. A test consisting of generating random trees of sequences in both datasets showed that in the R5/X4 dataset the average path joining two X4 sequences on a tree was about 1.35 times longer than the one joining two R5 sequences. In the NSI/SI dataset an average path joining two SI sequences was 1.1 times longer that the one joining two NSI sequences.

The above analysis indicates that both R5 and NSI sequences share common features and form coherent groups in sequence space. In the selected clustering 99% and 98% of the NSI and R5- sequences from each dataset, respectively, are clustered. In the following part of this study, we therefore used these sequences as reference points in sequence space. The mean distance of a sequence from all R5/NSI sequences was considered as a measure of its conservation.

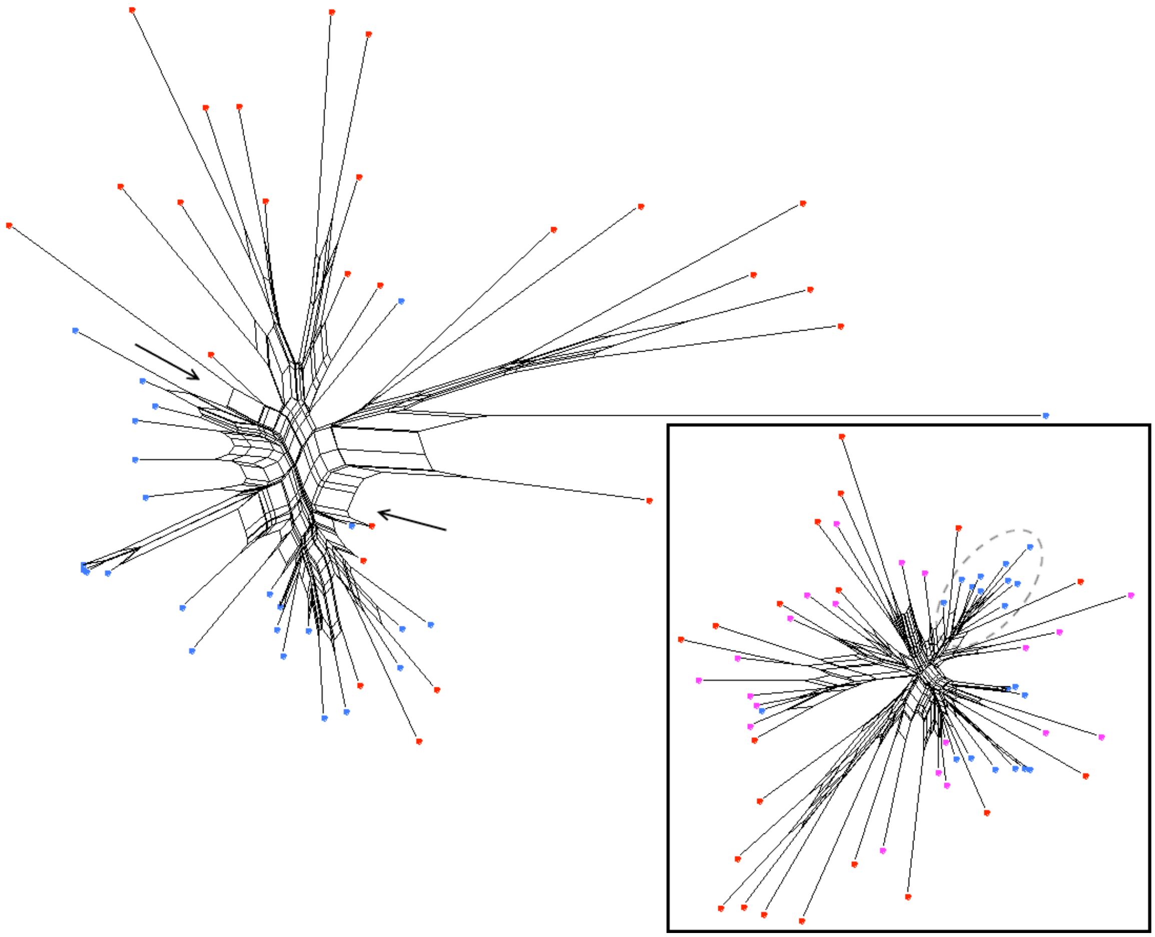


Figure 3.7 Splitstrees of sampled subsets of the NSI/SI and R5/X4 datasets. The splitstree on the left diagram was generated on a randomly sampled set of 25 sequences of each phenotype. NSI sequences are represented by blue dots, SI by red dots. The splitstree on the inset diagram was generated for a randomly sampled set of 20 sequences of each tropism (R5 – blue dots, X4 – red dots, R5X4 – magenta dots). Branch lengths, the number and width of splits (set of parallel edges in the graph) illustrate the evolutionary distance between the parts of the tree they separate. An example split is indicated by two arrows in the left panel. In both cases the SI/X4 sequences are located on branches relatively longer than those of the NSI/R5 sequences and separated by wide splits from the NSI/R5 phenotype (an example is indicated by arrows). NSI/R5 sequences tend to form dense tree-like parts of the Splitstree network containing few short splits (example shown in dashed circle in the inset panel) which suggests their evolutionary proximity.

3.2.5 Accuracy of genotypic coreceptor prediction in sequence space

There is a range of computational methods that aim at distinguishing NSI/CCR5-only from SI/CXCR4-capable sequences based on the V3 loop sequence. We investigated the accuracy of several sequence-based coreceptor prediction methods (11/25 rule, Web PSSM, geno2pheno) (Fouchier, Groenink et al. 1992; Jensen and van 't Wout 2003; Sing, Low et al. 2007) on both V3 loop datasets. Sequences that are incorrectly classified by all considered methods were identified and localized in sequence space using the same sequence distance, clustering and phylogenetic analysis.

Even though the distance distributions as well as the Splitstree diagrams exhibited discernable differences between sequences with different phenotypes, still some exceptions could be observed. We measured the average distance of sequences in the NSI/SI dataset to a reference set composed of NSI sequences from this dataset with a phenotype correctly predicted by all three methods. Figure 3.8 shows the distribution of the average distance of four different groups of sequences to this reference set. The four groups of sequences are: (i) correctly classified NSI sequences (left panel, blue curve), (ii) correctly classified SI sequences (left panel, red curve), (iii) NSI sequences misclassified by all three methods (right panel, blue curve) and SI sequences misclassified by all three methods (iv) (right panel, red curve). Sequences that failed to be correctly classified were located in untypical regions in sequence space – SI sequences classified in discordance with the Los Alamos annotation were closer to correctly predicted NSI sequences than the correctly predicted SI sequences (p-value < 0.05) and, on the other hand, the misclassified NSI sequences were further apart from this reference set in sequence space than the correctly predicted NSI sequences (p-value < 0.05). A similar significant pattern could be observed for the R5/X4 dataset.

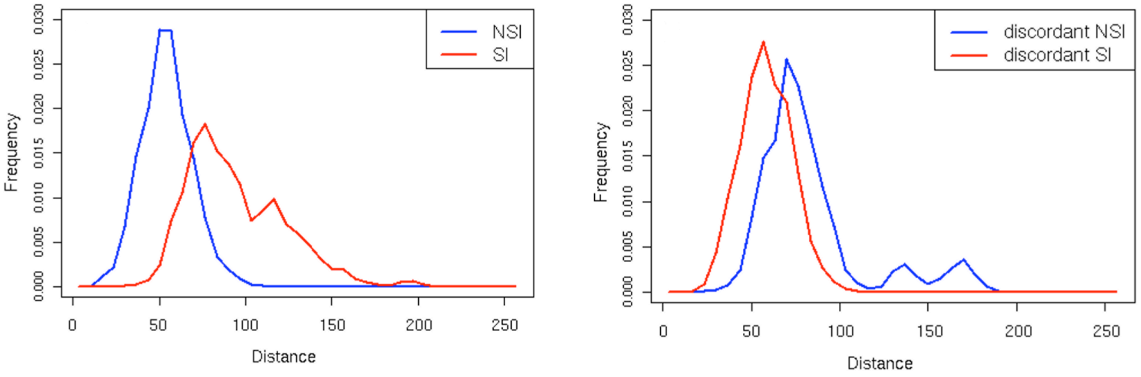


Figure 3.8 Distance distribution of incorrectly predicted V3 loop sequences. The plot contains the distribution of distances of the average distance of correctly (left panel) and incorrectly (right panel) predicted sequences to the set of NSI sequences with a phenotype correctly predicted by all analyzed coreceptor prediction methods (11/25 rule, Web PSSM, geno2pheno). NSI sequences are represented by the blue curve, SI by the red curve.

The above observations were confirmed by the Splitstree analysis (Figure 3.9, left panel). The misclassified sequences showed evolutionary relationships characteristic for the opposite phenotype – NSI sequences occupied longer branches and were located among SI sequences on the tree while the misclassified SI sequences were evolutionarily less distant from NSI clades or lied on boundaries between both phenotypes. The right panel of Figure 3.9 illustrates the clustering patterns of sequences misclassified by geno2pheno as compared to those correctly classified. As observed in the previous analysis the misclassified sequences showed clustering trends uncommon for their phenotype.

To investigate whether the classification error is due to data scarcity, we performed a test of the dependency of the accuracy of the predictions on the amount of data. In this test we used support vector machine (SVM) with linear kernel implemented in the package libsvm and position-specific scoring matrix (PSSM) implemented according to the description in (Jensen and van 't Wout 2003). We examined the performance of two classification methods trained on datasets with various sizes. We sampled subsets of the original NSI/SI dataset, used them as training sets for the SVM and PSSM, and then verified the number of prediction errors of the trained model on the same sequence set. With the increasing size of the training and test dataset we could observe no tendency of a decreasing prediction error. Both methods failed on a similar percentage of sequences independently of the size of the underlying dataset.

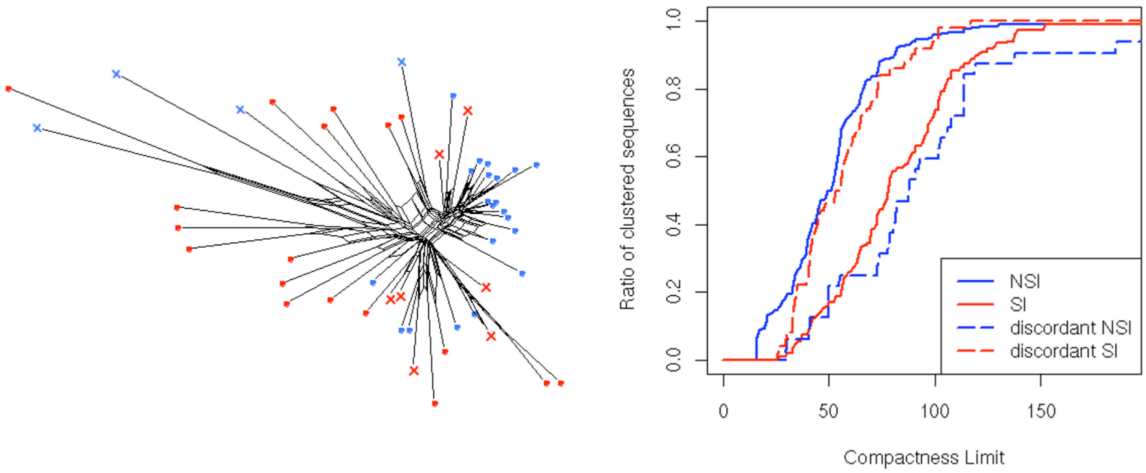


Figure 3.9 Location of the incorrectly predicted sequences in sequence space. The splitstree was generated for a sample of sequences containing misclassified NSI sequences represented by blue crosses and misclassified SI nodes by red crosses. Correctly classified sequences are represented by dots, colors are in accordance to the coloring scheme on previous figures. Right panel shows clustering patterns of sequences misclassified (dashed curves) by geno2pheno plotted against those correctly classified (solid curves).

On the one hand, a possible reason for the errors of the coreceptor prediction tools might be the complexity of factors determining coreceptor usage. For example, other parts of the gp120 protein beside the V3 loop may play a role in coreceptor binding. We repeated a similar sequence space analysis on the sequences of the V2 loop. We analyzed a dataset of phenotyped sequence spanning both V2 and V3 regions. The dataset was retrieved from the Los Alamos database and contained 280 sequences with 212 R5, 34 X4, and 34 dual-tropic. We compared the distribution of distances between sequences of the same and opposite tropisms of the full sequences and their V2 and V3 parts separately.

The analysis showed no clear separation of the two phenotypes in V2 sequence space. Diagrams in Figure 3.10 illustrate the distance distribution among the sequences spanning over both V2 and V3 regions as well as between their V2 and V3 parts separately. V2 sequences did not show the same pattern of distribution with highly divergent X4 and more conserved R5 sequences as the V3 sequences did. The joint V2 and V3 regions had a lower difference in distance distribution between both sequence types than their V3 part separately.

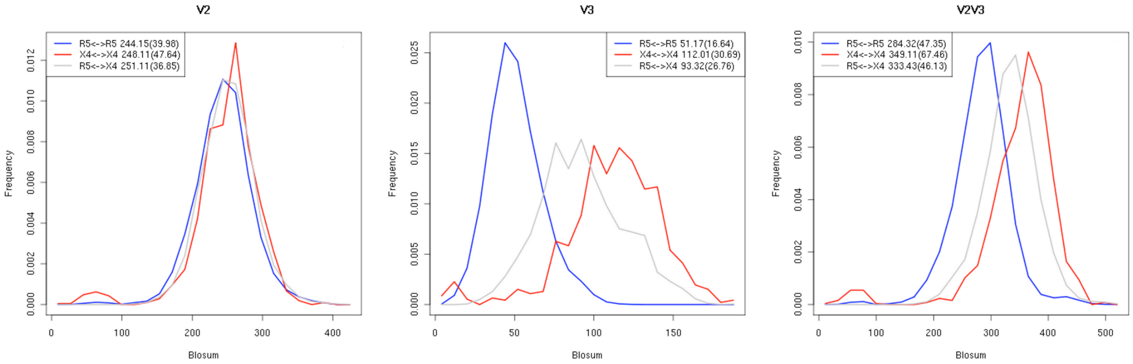


Figure 3.10 Distance distribution of V2-V3 sequences. Shown is the distribution of Blosum62 distances between pairs of V2 (left panel), V3 (middle panel) and V2V3 (right panel) parts of sequences of the same (R5 vs R5, X4 vs X4) and different (R5 vs X4) tropisms. The mean value and standard deviation of each of the distributions are indicated in the inserted boxes.

3.2.6 Coreceptor prediction based on sequence space analysis

The study described above – distance, clustering and phylogenetic analysis showed a clear separation between NSI/R5 and SI/X4 sequences in terms of the distance distribution, clustering steps and locations in splitstrees. NSI/R5 sequences appeared to be more conserved and to form clusters while SI/X4 appeared to diverge in an apparently unconstrained manner occupying distant parts of sequence space. Based on the above observations we tested if the position of a sequence relative to conserved NSI/R5 V3 loop sequences in sequence space conveyed sufficient information for the effective prediction of coreceptor usage. We investigated the predictive power of the three aforementioned methods of characterizing sequence space based on distances, clustering and phylogeny on the NSI/SI and R5/X4 datasets separately. The proposed classification methods aimed at distinguishing NSI/CCR5-only from SI/CXCR4-capable sequences. Dual-tropic sequences were labeled as X4 in this section. The score of each classifier was based on the separation of the sequence from the NSI/R5 sequences in sequence space. Low scores characterized sequences less separated from NSI/R5 sequences and therefore more conserved and probable to be NSI/R5. High scores indicated divergent sequences that, according to the sequence space analysis, are more likely to be SI/X4.

The classifiers were evaluated on both the NSI/SI and R5/X4 datasets using ten times ten-fold (10x10) cross-validation. In the following steps of the cross-validation procedure training and test sets were sampled from the analyzed datasets.

The first classifier was based on the distance measures and predicted the tropism of a sequence depending on its average distance from all NSI/R5 sequences in the training dataset. We constructed classifiers using three distance measures – Blosum62 matrix, Hamming distance and differences on positions significant for the coreceptor tropism according to Sing et al. (Sing, Low et al. 2007). As the coreceptor prediction score we used the mean distance of a sequence to all the NSI/R5 sequences in the training set.

The classification decision of the second classifier was based on the step of the hierarchical clustering algorithm in which the sequence ceases to be a singleton (termed here the *clustering step*). In order for the score to reflect the divergence of a sequence from the NSI/R5 sequences, only these sequences from the training set were used in the prediction procedure.

The third classifier used the Splitstree method for estimating the phylogenetic distances between pairs of sequences. Due to the high computational cost of a large splitstree construction, the phylogenetic distance between two sequences of a large dataset was calculated as an average distance between those two sequences in trees of randomly sampled sequence subsets composed in half of NSI/R5 and in half of SI/X4 sequences. The training procedure consisted of 100 iterations of sequence sampling, tree construction and tree distances extraction. First, subsets of 100 sequences of the training dataset (50 NSI/R5 and 50 SI/X4) were sampled. Then a Splitstree was constructed for each of the sampled sets and for each sequence pair in the trees the information on phylogenetic distance between the two sequences was extracted. We tested the predictive power of two different measures of phylogenetic distance: the sum of the lengths of the splits separating two sequences and the number of splits between them. After the iterative sampling and tree construction procedure, additional trees were constructed containing the sequences in the training dataset that did not appear in any of the sampled trees. The distance between two sequences was estimated as the mean of distances between those two sequences in the trees in which both sequences appeared. This way a phylogenetic distance matrix of the training set was assembled. In the prediction step we selected a subset of 90 sequences from the distance matrix: 45 NSI/R5 that were the most conserved (showed the least number of splits or shortest splits separating them from other sequences) and 45 SI/X4 that were the most diverse (showed the largest number of splits or longest splits separating them from other sequences). These sequences were used in the prediction procedure of the test set. In the prediction procedure subsets of 10 sequences from the test set were added to the selected 90 sequences of the training set and a splitstree was constructed for the merged set. The proportion of the test to train set sequences on this tree was chosen as optimal after testing several other proportions for the accuracy of predictions. The mean number of splits and the mean sum of lengths of splits between a sequence and the NSI/R5 sequences on the tree were used as the score predictive of the coreceptor usage.

All three classifiers were tested on both the NSI/SI and R5/X4 datasets using ten times ten-fold (10x10) cross-validation. We compared performance of sequence space classifiers to three existing methods – SVM, PSSM and 11/25 rule. SVMs were trained using the package libsvm with linear kernel. PSSMs were implemented according to the description in (Jensen and van 't Wout 2003). All methods were evaluated using receiver operating characteristic (ROC) curves focusing on the trade-off between false positive (FPR) and true positive rates (TPR). This trade-off can be controlled by choosing a prediction cutoff and turning the continuous scores into actual class predictions. The area under ROC curve (AUC) was taken as a cutoff-independent class separation criterion. Averaged ROC curves were estimated from the 10x10 individual cross-validation curves using vertical averaging. In the analysis we used the ROCR package (Sing, Sander et al. 2005).

All distance classifiers based on different distance measures showed comparable performance. At the FPR of the 11/25 rule (0.05 in the NSI/SI and 0.04 in the R5/X4 dataset) the distance methods showed the TPR between 0.51 and 0.56 in both the NSI/SI and the R5/X4 dataset. The areas under the ROC curve (AUC) were between 0.85 and 0.88 in both datasets. This performance was slightly worse than that of the SVM and PSSM methods that show TPR between 0.71 and 0.76 at the FPR of 11/25 rule and the AUC of about 0.90 and 0.92 (SVM and PSSM respectively) in both datasets (Figure 3.11, left panel).

The cluster-based classifier showed a similar performance to the distance based ones with the TPR of 0.54 and 0.53 and AUC 0.86 and 0.87 in the NSI/SI and R5/X4 datasets respectively.

The classifier based on the number of splits separating a sequence from the R5 sequences performed markedly worse than the distance-based methods. It yielded an AUC of 0.65 and 0.62 in the NSI/SI and R5/X4 dataset, respectively, and a TPR of 0.19 and 0.07 respectively at the FPR of the 11/25 rule. This suggested that similar ranges of split numbers can separate the sequences of both classes from the conserved R5 sequences. A better prediction performance was achieved with the use of the sum of lengths of splits that separate a sequence from these R5 sequences – TPR of 0.47 and 0.45 in NSI/SI and R5/X4 datasets respectively with the AUC of 0.74 and 0.76 respectively. However this result was lower then other sequence space-based methods.

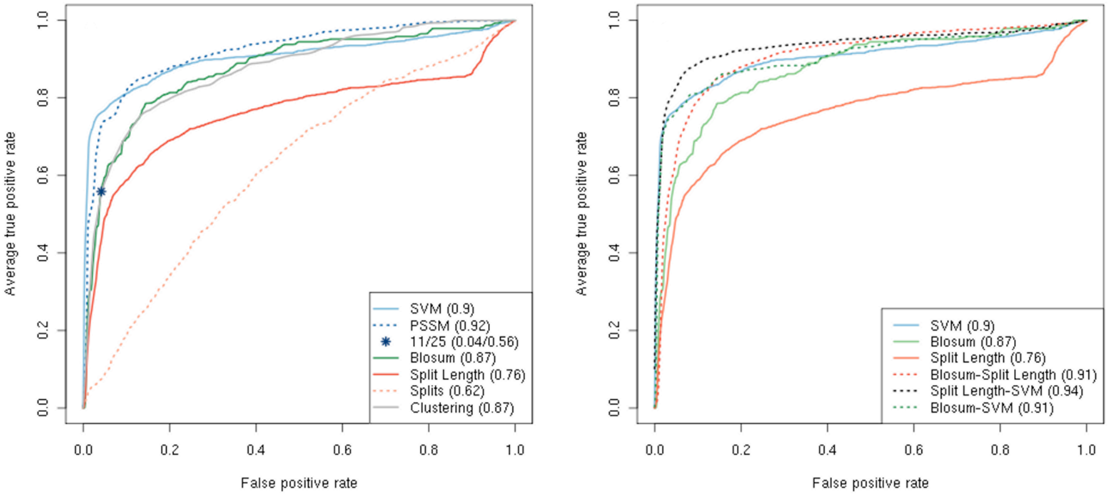


Figure 3.11 Performance of sequence space-based coreceptor prediction methods. Performance of the

individual coreceptor prediction methods (left panel) and their selected combinations (right panel) on the R5/X4 dataset is illustrated with ROC curves. The AUC of each method is indicated in the inserted box.

No sequence space-based methods achieved the performance level of the commonly used methods such as SVM and PSSM (see Figure 3.11, left panel). We therefore tested if adding the sequence space information to the SVM or combining several prediction methods into one can result in improved predictions. In the first approach we added to the binary feature vector coding the given sequence for the SVM the description of its location in sequence space. As the description of the location in sequence space we tested both the output score of the proposed classifiers, quantifying the separation of a sequence from the NSI/R5 sequences, as well as the sequence and phylogenetic distance to each of the NSI/R5 sequences. In the second approach we combined scores of several predictors, trained on the same training set, into one score. The scores of each predictor were normalized to the 0-1 interval with the higher scores representing X4 sequences. We tested several methods of combining prediction scores, such as min, max, mean and Euclidian distance from the origin of the score space. We restricted the classifier combination methods to the simple, non-trainable combiners, bearing in mind their generally good performance (Altmann, Rosen-Zvi et al. 2008).

The description of the position of a sequence in sequence space, either in terms of the score of each of the sequence space prediction methods or in terms of a vector of distances to each NSI/R5 sequence, coupled with the binary feature vector coding the given sequence for the SVM did not result in an improved performance over predictors based on the sequence only.

In the test of combined scores of several prediction methods we first compared the scores returned by different classification methods within the same cross-validation run. All the methods showed a high correlation in the scoring (Pearson correlation coefficient > 0.8) with the exception of phylogeny-based and SVM classifiers (correlation of about 0.25 and 0.7, respectively). Despite the low prediction power of the phylogeny-based method, we observed several sequences (about 15 in both datasets in each cross-validation run) for which the score returned by this classifier was 50% more accurate than the score of the SVM. However, since the phylogeny-based classifier contained a stochastic step, the scores returned by this method showed a high variation between different cross-validation runs and this result was not reproducible in each run. Next, we tested several methods of combining prediction scores and found the Euclidian distance from the origin of the score space to be the method achieving the best discrimination between the two sequence classes. Finally, we examined the performance of all possible combinations of predictors. Combining classifiers improved their performance in general. Three distance-based classifiers coupled together achieved better results than each one individually (Table 3.2). Joining the phylogeny-based method with another classifier resulted in the highest improvement in the predictive power – up to 0.03 increase in the AUC in the case of SVM (Figure 3.11, right panel). The largest predictive power was achieved by merging the distance and phylogeny scores with the SVM methods (Table 3.2).

predictor	R5/X4		NSI/SI	
	sensitivity	AUC	sensitivity	AUC
Blosum	0.5544	0.8747	0.5606	0.8647
Hamming	0.5241	0.8842	0.5144	0.8644
significant positions	0.5524	0.8765	0.5499	0.8532
split number	0.0712	0.6245	0.1932	0.6486
split length	0.4262	0.7591	0.4722	0.7449
clustering	0.5510	0.8725	0.5399	0.8606
SVM	0.7607	0.9038	0.7407	0.8887
PSSM	0.7276	0.9199	0.7140	0.9131
Blosum-Hamming-significant positions	0.6014	0.8990	0.5808	0.8745
Blosum-split lengths	0.5986	0.9063	0.6076	0.8768
Blosum-SVM	0.7517	0.9062	0.7313	0.8944
Split Length-SVM	0.8062	0.9355	0.7929	0.9156
Blosum-split length-SVM	0.8076	0.9420	0.7778	0.9224
all methods	0.7607	0.9404	0.7369	0.9178

Table 3.2 Performance of coreceptor prediction methods and their combinations on the NSI/SI and R5/X4 datasets. The performance is assessed as the sensitivity at specificity of 11/25 rule in the main dataset (0.05 in the NSI/SI and 0.04 in the R5/X4 dataset) and as the area under ROC curve. The table lists performances of several classification methods and their combinations. The predictions of the combined methods are calculated as the Euclidian distance from the origin of the prediction score space of each of the individual methods.

3.2.7 CD4$^+$ T cell counts in sequence space

In the last part of the study we related the location of a patient V3 loop sequence in sequence space to the patient CD4$^+$ T cell count. From the Los Alamos database we selected a set of 7003 V3 loop sequences with a reported CD4$^+$ T cell count. In addition, we selected sequence samples and the corresponding CD4$^+$ T-cell counts of 88 patients (225 sequences) from the University of Cologne. Since both Los Alamos and Cologne sequence sets exhibited similar patterns in the sequence space arrangement we merged the sets into a single dataset which we called the *full dataset*. For the purpose of the longitudinal study we allowed this dataset to contain more than one sequence of the same patient. From the full dataset we selected therapy-naïve patient samples (2213 sequences). We termed this subset of the full dataset the *therapy-naïve dataset*. In both the full and the therapy-naïve datasets we distinguished longitudinal (time-series) data comprising sample sequences of the same patient spanning several years (72 patients in the full dataset, 16 therapy-naïve patients with an average of 3.1 and 3.9 sequences per patient respectively). We analyzed the longitudinal patient data in both datasets to investigate how sequences of viral variants of an individual patient trace paths in sequence space in association with the progression of the disease.

As in the analysis of misclassified sequences in the NSI/SI and R5/X4 datasets, we used the mean distance to the sequences annotated as NSI or R5 in the full dataset (653 sequences) as a measure of sequence conservation. For each sequence in the full dataset, we calculated this distance and plotted it against the CD4$^+$ T cell count (Figure

3.12, top panels). Among sequences collected from highly immunosuppressed patients (T cell count below 200cells/mm^3) we observed a large range of sequence conservation spanning conserved, NSI/R5-like, as well as highly divergent sequences. Among the patients with higher T cell counts this range was narrower and included only conserved sequences.

Next, we analyzed the longitudinal patient data in the full datasets to see how divergence of viral variants inside an individual patient can change in association with the disease progression. In the full dataset we counted 76 patients (both therapy-naïve and therapy-experienced) with data occupying several time points. We examined the conservation of sequential measurements of each of the selected patients against the corresponding CD4$^+$ T cell count. For 21 patients we observed an increase in sequence divergence with decreasing CD4$^+$ T cell count to the immunodeficiency level – below 200cells/mm^3 (example in Figure 3.12, top-left). Five patients showed an opposite trend with a decrease in mutations co-occurring with a decrease of the number of CD4$^+$ T cells (example in Figure 3.12, top-right). Remaining patients showed no or various conservation change with varying CD4$^+$ T cell count level.

In order to examine how drug therapies can influence the relationship between sequence conservation and CD4$^+$ T cell levels, we performed the same study on the data of untreated patients included in the therapy-naïve dataset. This dataset contained 2213 sequences, 112 of which are annotated as NSI or R5. Other sequences were annotated as SI/X4 or had no annotation. The NSI/R5 sequences were used as a reference set in the sequence space. In the therapy-naïve dataset the lack of conservation of the sequences of patients with an impaired immune system was more pronounced. Among severely immunosuppressed patients (CD4$^+$ T cell count below 200cells/mm^3), almost no highly conserved sequences can be observed (Figure 3.12, bottom panels). Again, highly divergent viral strains were found uniquely in patients showing low T cell counts.

In the longitudinal dataset, among 16 patients with more than one time point measurement, four showed an increase in the sequence divergence with the transition to the immunodeficiency state (T cell count below 200cells/mm^3) (example in Figure 3.12, bottom-left). None of the patients showed the opposite trend.

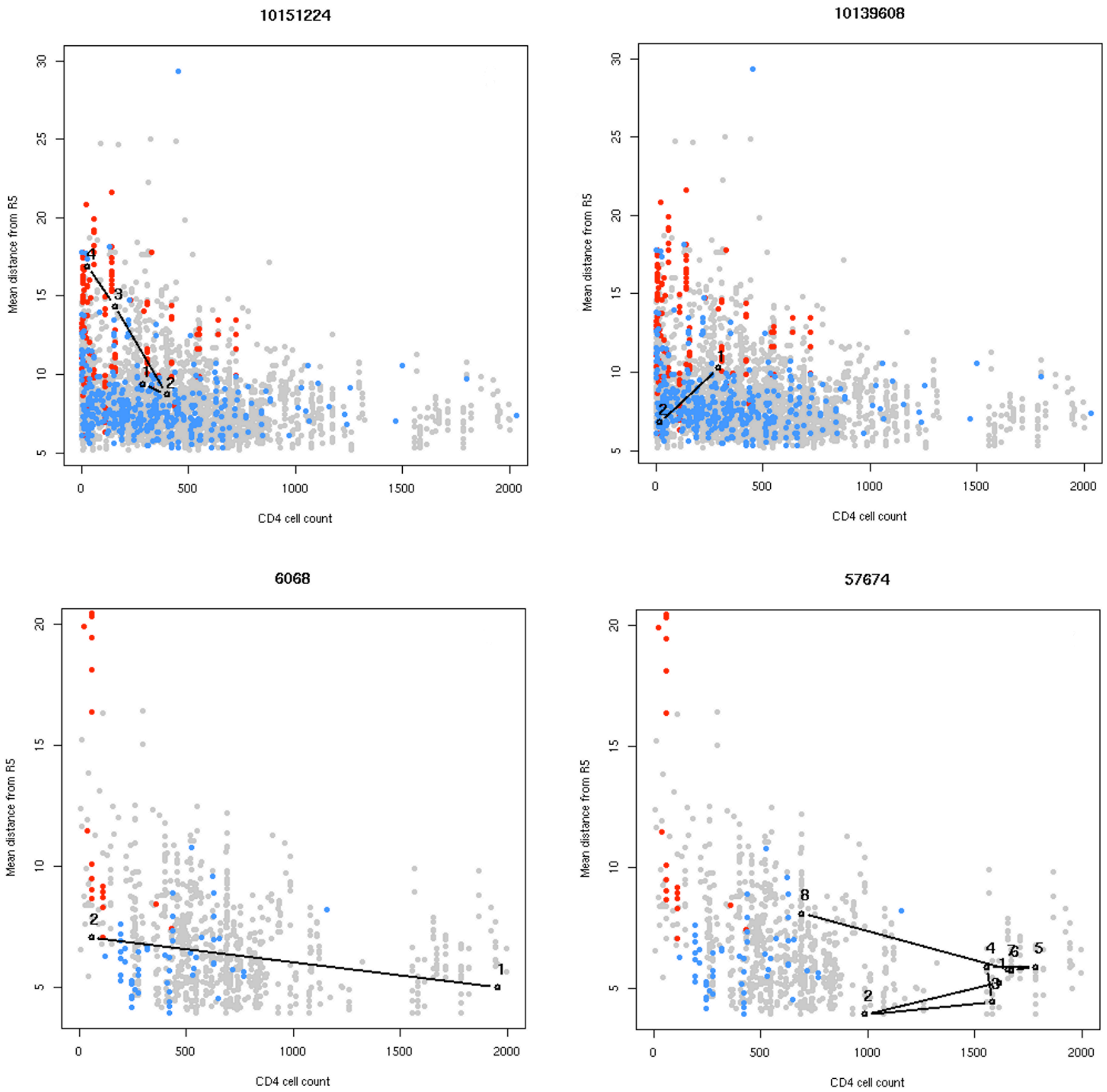


Figure 3.12 Patient sequence evolution in sequence space. CD4$^+$ T cell count is plotted against mean distance to NSI/R5 sequence set in the full (top panels) and the therapy-naïve (bottom panels) datasets. Sequences with annotated phenotype or tropism are marked with colors (blue for NSI/R5 sequences, red for SI/X4). Sequential data points of the same patients are marked in black and connected with a solid line: top-left panel – therapy-experienced patient showing an increase in sequence variability with decreasing CD4$^+$ T cell count, top-right panel – therapy-experienced patient with an opposite trend, bottom-left panel – therapy-naïve patient showing an increase in sequence variability with decreasing CD4$^+$ T cell count, bottom-right panel – therapy-naïve patient with a mixed trend. Patient identifiers are indicated above the plots.

As in the previous part of this study (sequence clustering) we selected a clustering of the full and therapy-naïve sequence sets using the silhouette value and associated a cluster sequence composition to the CD4$^+$ T cell counts. Both sequence sets showed one major cluster containing 86% and 52% of all sequences, respectively, and three smaller clusters of about 3% and 5% in each set respectively. We observed an important overrepresentation of samples collected from ill patients (T cell count below

500cells/mm^3) among the unclustered sequences ($p<0.001$, chi-square test) which reflected their high evolutionary divergence.

3.2.8 Discussion

In this study, by means of different distance measures, clustering and phylogenetic methods, we illustrated and interpreted patterns in V3 loop sequence space of the HIV-1 envelope gene. The analysis confirmed a relatively high conservation of R5 and NSI viral sequences as compared to highly divergent X4 and SI sequences. According to this analysis, the NSI/R5 sequences appear to share common features and the SI/X4 sequences to be highly divergent and not showing a unique mutation pattern. Previous studies reported at least several possible V3 loop mutation pathways (Milich, Margolin et al. 1993) and indicated the twofold larger heterogeneity of the X4 over the R5 viruses. The lack of common features among the X4 sequences, as well as their high divergence, render these sequences impossible to group in coherent clusters. A statistical model of X4 sequences is therefore difficult to obtain. This divergence pattern has already been reported in previous studies (Chesebro, Wehrly et al. 1992; Nelson, Fiscus et al. 1997; Resch, Hoffman et al. 2001), in contrast to these analyses done on small data samples, here we provided support for these hypotheses based on a large-scale analysis of V3 loop sequence data.

The analysis of cluster structure pointed to dense regions of sequence space occupied by R5 sequences sparsely surrounded and interspersed by X4 and dual-tropic sequences. The sparse outer boundaries of these regions are less highly conserved and undergo positive selection. The finding of such regions suggests that the coreceptor usage switch which is correlated with high sequence divergence might be driven by the lack of selective pressure on the viral V3 loop. This finding supports other studies that reported presence of selective pressure maintaining the relative homogeneity of the R5 viruses as well as the correlation of the emergence of the X4 strains with the accelerated V3 loop evolution (Chesebro, Wehrly et al. 1992; Shankarappa, Margolick et al. 1999).

Localization of the misclassified sequences in sequence space indicated a possible reason for errors of V3 loop sequence-based coreceptor prediction tools. These sequences are located in parts of the sequence space untypical for their phenotype and show an inverse pattern in their distances distribution as compared to the distances among correctly predicted sequences (Figures 3.8 and 3.9). Only sequences misclassified by the 11/25 rule were not characterized by such a distance inversion which suggests that in certain cases mutations at positions 11 and 25 are insufficient for the change of phenotype and that the accumulation of other mutations on the V3 loop might drive the coreceptor switch. Previous studies (Pillai, Good et al. 2003) reported dual-tropic V3 loop sequences as being predominant among the sequences misclassified by different prediction methods. The sequences observed in this study to be located in untypical for their tropism regions of sequence space might therefore represent an intermediate form between the two mono-tropic types. Other studies (Low, Marchant et al. 2008) pointed to dependence of the predictive value of positions 11 and

25 on CD4$^+$ T cell level, suggesting that individual patient parameters might influence the viral coreceptor usage. Another possible reason for errors in predicting coreceptor usage on the basis of V3 sequence might be the occurrence of complementary mutations in other parts of the gp120 protein. However, a similar inspection of the sequence space of the V2-V3 sequences showed no separation between the R5 and X4 phenotypes in the V2 part (Figure 3.10). A similar observation was previously reported (Hoffman, Seillier-Moiseiwitsch et al. 2002) in a study of the V1-C3 region of the gp120 protein sequences showing a relative conservation of the V2 region with only a few positions in the V1/V2 stem being significant for the coreceptor usage. This strongly points to the V3 loop being the region of the gp120 protein crucial for the viral tropism that is under selection pressures driven by the interaction with the host.

Sequence location in sequence space can be used for the coreceptor usage prediction. The prediction methods we presented in this study were based on sequence space location determined by the means of sequence distance measures, phylogenetic distance and clustering. The predictive power of the methods is below that of SVM and PSSM which is not surprising as the prediction score based on sequence space is obtained by averaging over many sequence distances – an operation in which information on a single position in a sequence is lost. Adding the sequence space location descriptor to the sequence-based SVM did not improve the accuracy of the prediction which might be due to the fact that the sequence space location is drawn from the genetic information already provided to the SVM. However combining prediction methods, in particular the phylogeny-based method with other classifiers, resulted in a performance increase. Nevertheless the stochastic step involved in the phylogeny-based method renders its predictions less reliable which is reflected by the weak predictive power of the method by itself.

Relating patient clinical markers to the respective sequence position in sequence space showed higher sequence variability among patients with an impaired immune system. Other studies have reported the emergence of highly mutated viruses in the later stage of infection (McNearney, Hornickova et al. 1992; Shankarappa, Margolick et al. 1999). The presented analysis shows the association of this emergence with the drop in patient CD4$^+$ T cell count. This association might be due to a selection pressure exerted on the viral V3 loop that disappears with the gradual erosion of the immune system. With the attenuation of this selection force the virus is apparently undergoing an unrestricted evolution on the V3 loop which traverses distant parts of sequence space. The existence of a similar selection mechanism has been suggested in other studies (Zhang, Diaz et al. 1997).

The nature of the selection pressure limiting the viral mutation in the early stage of infection is not clear. The observation that the development of a highly mutated X4 virus is associated with low CD4$^+$ T cell numbers and therefore with the impairment of the immune system suggests an immunological component of this mechanism. There is some evidence for selective pressure against the emergence of X4 strains having an immunological basis (van Rij, Hazenberg et al. 2003) and decrease of positive selection accompanying the drop in CD4$^+$ T cell count. This supports the hypothesis that the emergence of mutated strains late in infection is related to the limitation in the suppressive capacity of the immune system. However, other studies (Shepherd, Jacobson et al. 2008) showed examples of patients infected with an X4 virus at relatively high CD4$^+$ T cell counts. Such cases could be explained by a successful antiretroviral treatment, as viral tropism appears not to impact the response to the treatment in terms of CD4$^+$ T cell count (Waters, Mandalia et al. 2008). In the datasets used for this study we find patients (mostly therapy-experienced) who exhibit an increase in sequence conservation with a reduction of the CD4$^+$ T cell counts (example in Figure 3.12, top-right). Other parameters may therefore influence the evolution of the viral V3 loop. Notably, high sequence conservation has been observed among patients under long-term successful therapy (Chun, Nickle et al. 2005).

This large-scale analysis of sequence space of the V3 loop provides a comprehensive description of R5 and X4 viral phenotypes. By characterizing X4 viruses as highly variable and dispersed in sequence space we provide further evidence for the fact that not only is this phenotypic change predictive of disease progression but also that it comes as a result of an extensive evolution of the V3 loop sequence and a decrease of the selective pressure on the viral envelope genome.

3.3 Structural descriptor of V3 loop

In the second study on the virus population scale, also described in Bozek, K., Lengauer, T., Sierra, S., Kaiser, R., Domingues, FS. Analysis of physicochemical and structural properties determining HIV-1 coreceptor usage. *submitted*, we constructed a structure-based predictor of the HIV coreceptor tropism. The predictor was designed such as to address the drawbacks of the existent structure-based methods (Sander, Sing et al. 2007; Dybowski, Heider et al. 2010), namely the interpretability of results and efficiency of prediction procedure. In this work, we performed a systematic approach to incorporating physicochemical and structure properties into the coreceptor usage prediction. We mapped 56 amino acid indices representing their physicochemical properties onto the V3 loop structure and used machine learning methods to extract the most informative for the coreceptor usage. The extracted set of amino acid and loop location features was strongly reduced as compared to the initial one and reached higher accuracy with decreased computational load. This structural descriptor offers a direct interpretation by pointing to two regions in the loop stem and their physicochemical properties that play crucial role in determining of the coreceptor usage. The method was developed on clonal data, however we also applied it to clinically derived data and tested its usability for prediction of the MVC therapy outcome. Finally we implemented the method as a server application available for public use.

3.3.1 Main dataset

To construct the dataset on which the method was developed we searched the Los Alamos database (http://www.hiv.lanl.gov/) for all phenotyped V3 loop sequences. In order to avoid bias due to overrepresentation of a single patient data we filtered the dataset leaving one randomly chosen sequence per patient. This way obtained dataset contained 1186 sequences annotated with the coreceptor tropism, with 215 annotated as X4 viruses. This dataset was termed the *main dataset*. We aligned the dataset and the sequence of the crystallized loop using clustalw (Thompson, Higgins et al. 1994) obtaining an alignment of length 50. In this study the positions in the gp120 sequence are numbered relative to the reference as previously described (Korber, Foley et al. 1998).

3.3.2 Structural descriptor

To represent the amino acids with their physichochemical properties rather than their one-letter code, we used the amino acid indices collected in the AAindex database (Kawashima, Ogata et al. 1999). In this database various physicochemical and biochemical properties of amino acids are stored in the form of numerical indices. Due to the high redundancy of over 500 indices in the database we used a representative and interpretable subset of 54 indices, selected using multivariate statistical analysis (Atchley, Zhao et al. 2005). Two of the selected indices – “Normalized frequency of beta-turn” and “Free energy in beta-strand region” had double entries in the AAindex database (AAindex entries: CHOP780101/CHOP780203 and MUNV940104/MUNV940105 respectively) showing minor differences. In order to avoid an arbitrary decision between the double entries we used both of the ambiguous indices which resulted in a set of 56 indices selected for this study.

Each amino acid of the V3 loop sequence was represented as a vector of the 56 preselected amino acid indices. The structural information was integrated into the V3 loop structural descriptor based on the published structure of the V3 loop with PDB (Berman, Westbrook et al. 2000) code 2B4C (Huang, Tang et al. 2005). To construct the structural descriptor for each V3 loop sequence we used spheres defining structural proximities on the loop within which the physicochemical properties of residues are averaged. The spheres are positioned along the reference loop backbone and centered on its residues. Positions of the residues were defined as the position of the representative atom of each residue of the structure – Cα atom for Glycine and Cβ for other amino acid types. Positions of insertions relative to the reference structure were inferred based on the positions of representative atoms of the residues at both ends of the insertions (flanking atoms). First, a line connecting the flanking atoms was calculated. Then the following insertion positions were placed along the line at equidistant spacing. This way calculated coordinates of the following positions of the V3 loop sequence (residues and insertions) were used as centers of the spheres on the loop. In addition to the set of spheres corresponding to alignment positions additional spheres were positioned at the midpoints of lines connecting centers of each pair of consecutive alignment spheres. This resulted in a set of 99 spheres – 50 corresponding to alignment positions and 49 positioned in-between consecutive alignment positions. Example spheres are illustrated in Figure 3.13.

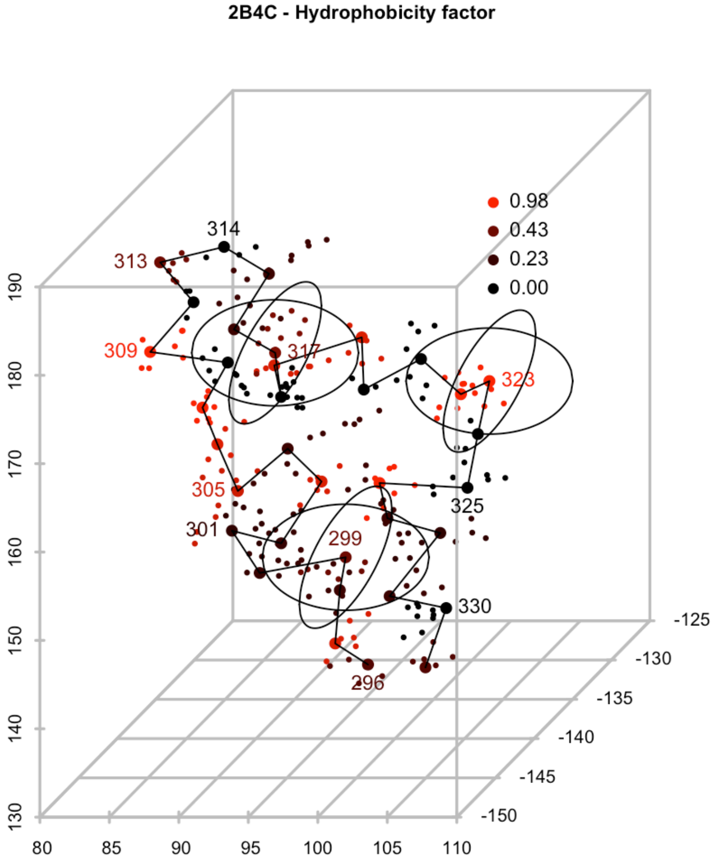


Figure 3.13 Schematic illustration of the structural proximities of the structural descriptor. Atoms of the 2B4C V3 loop structure are represented by dots, representative atoms by larger dots. Black line connects representative atoms of the following loop residues. Atoms of each residue are colored according to the "Hydrophobicity factor" amino acid index as indicated in the legend. Sphere-shaped proximities were centered along the V3 loop backbone, three examples centered on representative atoms of residues 299, 317 and 322 are shown. The physicochemical features of residues within each sphere were summed and used as a part of the structural descriptor.

The radius of the spheres was chosen such as to provide a balance between too small proximities that do not incorporate structural information and too large proximities that reduce spatial resolution. We investigated the number of residues included in each sphere as a function of the sphere radius and selected the radii of 8Å and 10Å for initial testing. We additionally tested other radii (3Å, 7Å, 15Å) to assess the performance of the chosen parameter values in a cross validation setting.

Each V3 sequence position was mapped to a sphere if the corresponding representative atom was located within the given sphere. Within each sphere the vectors of amino acid indices of the mapped residues were normalized using Gaussian smoothing. The vector of amino acid indices representing a mapped residue was multiplied by the value of a normalized Gaussian function applied to the distance between the representative atom of the residue and the sphere center. We inspected the cumulated contribution of each V3 residue to the descriptor as a function of the variance of the Gaussian function. We chose the variance equal to the sphere radius as resulting in the most uniform contribution of each atom and therefore not giving undue priority to any of the loop residues. We tested an additional set of variances (ratio 0.5 and 0.75 of the sphere radius) to assess the impact of this parameter on prediction performance.

Figure 3.14 illustrates the rationale of the parameter choice. Black histograms represent the distribution of the number of residues included in proximities of a radius indicated on the corresponding plot on the left. Spheres of radius 6 include five residues on average with a 4.56 variance among spheres, which might be too small a number to represent a structural binding site. On the other hand, spheres with radius 12 contain 16 residues on average, which corresponds to almost half of the loop. Sphere radius 8 was chosen for further testing. Spheres of this radius contain 8.4 residues on average with a relatively small variability among spheres. Red histograms illustrate the sum of Gaussian normalizing factors per residue. This sum should be similar for each residue such that no individual residue is weighted markedly higher than others in the structural descriptor. The narrowest distribution was obtained if the variance equals the radius (R = 8, var = 8). The chosen set of parameter values of the sphere radius and variance of the smoothing function was used in model testing, the final model parameters were selected as those showing the highest prediction accuracy on our dataset.

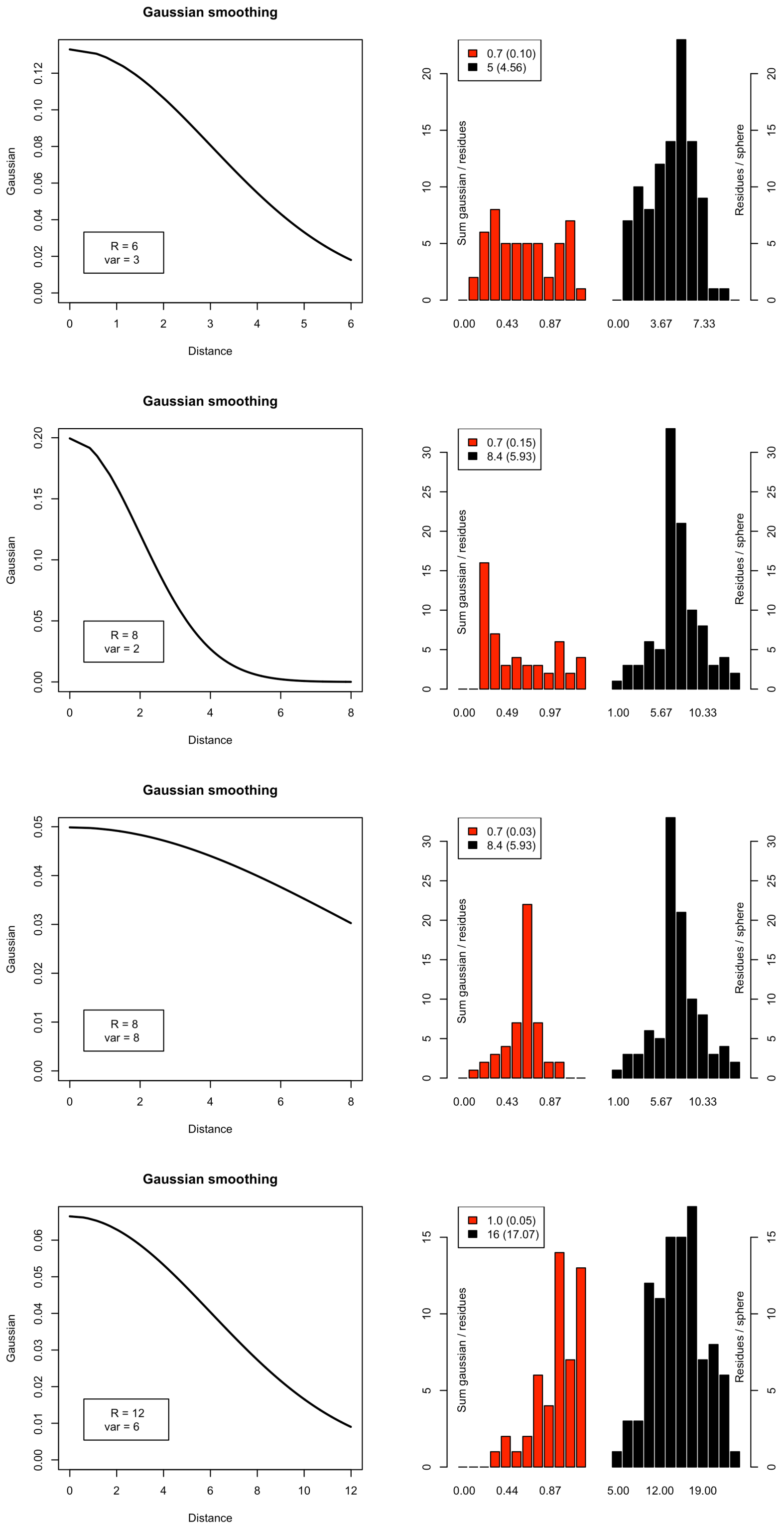


Figure 3.14 Choice of sphere radius and Gaussian smoothing parameters. Left diagrams show shapes of Gaussian function for the parameters indicated in the legend. Right diagrams contain histograms showing sum of normalizing factor per each residue (red) and number of residues in each sphere (black). Mean with variance in brackets of each distribution are indicated in legends.

After Gaussian smoothing the amino acid index vectors of the mapped residues were summed within each sphere. The sphere vectors were then concatenated into a single V3 loop vector and used as the V3 loop structural descriptor in the statistical model for coreceptor usage prediction.

For comparison we implemented two sequence-based descriptors of the V3 loop. The indicator model represents each amino acid with a binary vector of size 20 in which the position of a one indicates the amino acid it encodes. The aaindex model encodes each amino acid as a 56-long vector of amino acid indices used in the structural descriptor.

The structural descriptor-based model distinguishing between R5 and X4 viruses was constructed as an SVM (Boser, Guyon et al. 1992) implemented in the R package e1071 (Dimitriadou, Hornik et al. 2005). In the model evaluation we used ROC curve illustrating the trade-off between specificity and sensitivity. The area under ROC curve was taken as a cutoff-independent class separation criterion. The specificity at the sensitivity of the 11/25 rule was used as an additional measure of the model performance. We used the the R package ROCR (Sing, Sander et al. 2005) for visualization and evaluated the models with ten times ten-fold (10x10) cross-validation. Each descriptor feature was normalized to 0-1 within the training dataset.

First, we tested the performance of the preselected parameter set on the main dataset. Apart from radii of 8Å and 10Å that were preselected based on the average number of residues included in each sphere, we also tested predictive performance of several other values (3Å, 7Å, 15Å) in a cross validation setting on the main dataset. Among these sphere radii the radius of 8Å showed the highest performance with AUC of 0.847 and sensitivity at the specificity of 11/25 rule of 0.587 (Figure 3.15 and Table 3.3). Much smaller and larger radii significantly reduced prediction performance ($p < 0.001$ for R=3Å and R=15Å, paired wilcoxon test).

Radius (Å)	AUC	sensitivity
3	0.833	0.548
7	0.847	0.572
15	0.789	0.470
8	**0.847**	**0.587**
10	0.814	0.483

Table 3.3 Performance of models based on different radii tested on the full set of features. The radius selected for further analysis is marked in bold. The sensitivity is at the specificity of the 11/25.

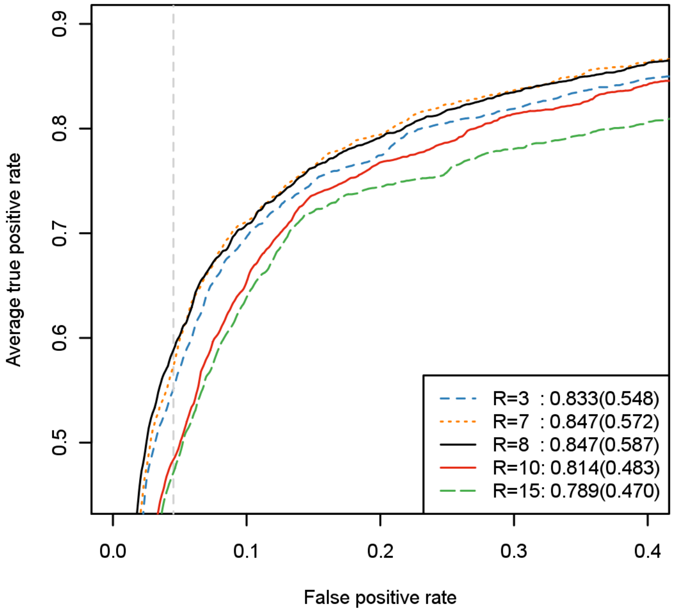


Figure 3.15 ROCRs of models based on structural proximities of different radii. The selected radius of 8Å is traced with a black solid line. AUC and sensitivity at the specificity of 11/25 rule in brackets are indicated in the legend.

3.3.3 Feature selection

In order to reduce the highly redundant feature vector of the structural descriptor and to investigate which features are informative for coreceptor usage we applied several feature selection procedures. We used two classification methods performing feature ranking: Random Forests (RF) (Breiman 2001) with the mean decrease in Gini index and linear SVM with its feature weights (Guyon, Weston et al. 2002). We also used Lasso regression (Tibshirani 1996) that performs feature selection by assigning zero to the less important feature parameters. In case of the methods producing feature ranking (RF and linear SVM) we tested two cut-offs for the selected features: top 1% and top 5% of a gamma distribution fitted to parameters of all the features using maximum likelihood. We used all features selected by the Lasso regression method. The feature ranking of the SVM and Lasso regression methods was obtained as an average of a 10x10 cross-validation. The RF construction method performs internal randomization, its feature ranking was therefore inferred from a single run of the method. We tested the performance of models based on subsets of features selected by different methods and cut-offs separately and combined. Models based on subsets of selected features were named after the feature selection method with the percent cut-off indicated in parentheses (e.g. SVM(1)). Names of models based on combinations of feature sets selected using several feature selection methods were composed of the corresponding feature selection methods separated by a "_" (e.g. SVM(1)_Lasso). The best performing model constructed on the main dataset and based on the selected feature set was termed the *clonal model* and used as the final structural descriptor model. As the analysis of the features selected for the clonal model was a goal of our study, feature selection was performed on the entire clonal dataset. To assess how the choice of the set of sequences on which the features are selected impacts the model's prediction accuracy, we performed two different types of tests. In the first test, features of the model were reselected on the training set in each cross validation run on the clonal set. In the second test we applied the features of the clonal model on other sequence sets.

The comparison of the performance of the structural descriptor based on the full set of features (*full model*) to the descriptor based on separate and combined subsets of features selected through three feature selection methods showed that overall, reducing the feature set results in an improvement in the prediction accuracy over the full set of features and over the indicator model (Table 3.4). The SVM(1) model based on the top 1% ranking features performed better than the SVM(5) model based on a larger feature set of the top 5% ranked features. The Lasso model based on the most highly reduced feature set (102 featues) selected via Lasso regression resulted in the highest performance with AUC 0.893 and sensitivity 0.674 at the specificity of the 11/25 rule. Models based on features selected via RF ranking showed the poorest predictive performance of all models tested (Table 3.4). We performed the same feature selection procedures on models based on the structural descriptor with a sphere radius of 10Å. The results of models based on this radius showed analogous patterns of performance although with worse results (data not shown). In the rest of this study we therefore used models based on the 8Å radius.

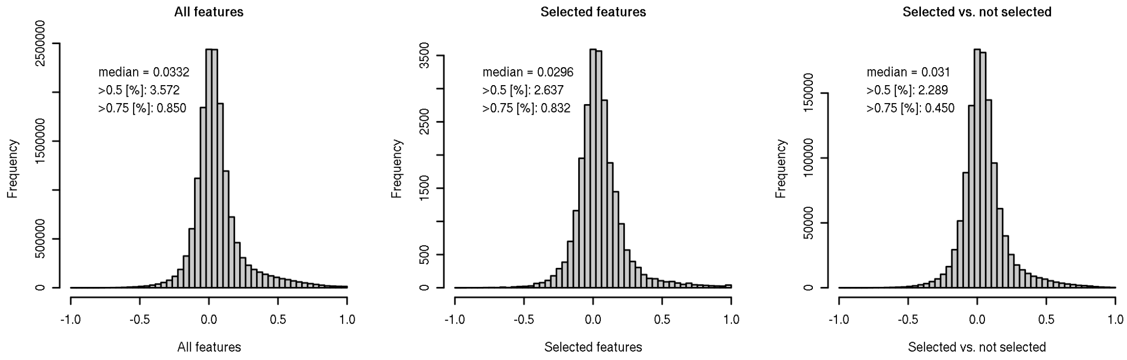


Figure 3.16 Correlation of features in the initial feature set. Histograms show distribution of the Pearson correlation of all features of the structural descriptor (left panel), of the features of the clonal model (middle panel) and of the features of the clonal model with the remaining features of the structural descriptor (right panel). Median, percentage of feature pair with correlation >0.5 and >0.75 are indicated in the legend.

Features selected using the three methods showed a limited overlap ranging from ~1% between SVM and RF, 3.5% between SVM and Lasso and ~9% between RF and Lasso. This limited overlap might be due to correlation of features in the initial feature set, which was previously reported to additionally reduce the predictive accuracy of the underlying model (Tolosi and Lengauer 2011). Features of the initial structural descriptor showed low correlation (median 0.03, see Figure 3.16) and models constructed on clustered features (Tolosi and Lengauer 2011) showed no improvement in accuracy compared to the models based on unclustered features. Therefore, we chose a more parsimonious approach of not involving feature clustering.

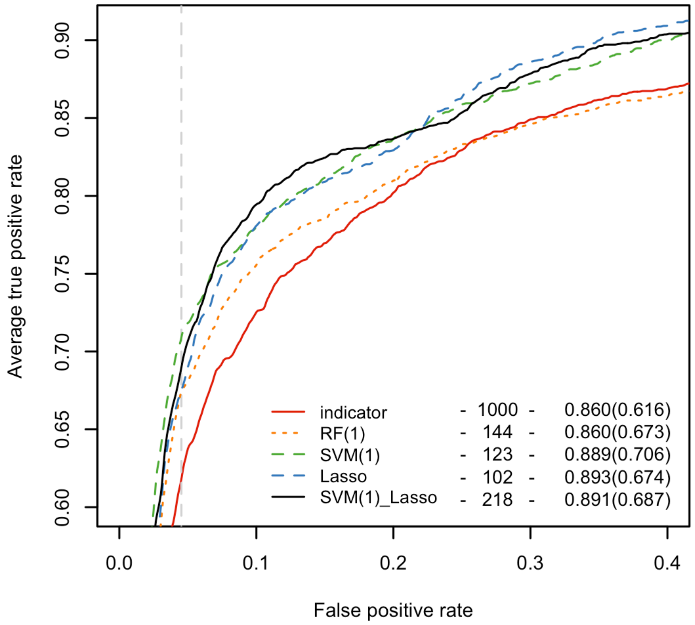


Figure 3.17 ROCR of models based on features selected using RF, SVM and Lasso. Number of features, AUC and sensitivity at the specificity of 11/25 rule in brackets are indicated in the legend. Clonal model is represented by a black solid line, indicator by the red solid line.

Next, we inspected the predictive performance of combined sets of features selected using different methods (Table 3.4). Based on the performance of the tested models we selected the SVM(1)_Lasso model combining the set of top 1% of the SVM-ranked features and Lasso-selected features as the result model termed the *clonal model* (Figure 3.16). The AUC and sensitivity at the specificity of 11/25 rule of this model showed a significant increase over the sequence-based indicator and aaindex models (p < 0.01, paired Wilcoxon test). Even though the performance of the clonal model is not significantly higher than that of the Lasso model that is based on a smaller set of features, we chose the SVM(1)_Lasso feature set offering higher sensitivity at the specificity of 11/25 rule. The AUC and sensitivity at the specificity of 11/25 rule of this model were significantly higher than those of the indicator and aaindex models (p < 0.01, paired Wilcoxon test). A test of reselecting the SVM(1)_Lasso features on the training set within each cross validation run resulted in a ~1.6% decrease of the AUC. We consider this difference as a potential uncertainty of the accuracy estimation inherent to the feature selection procedure. The accuracy within the indicated uncertainty is however significantly higher (p~0.003) than the accuracy of the sequence-based prediction suggesting that the selected structural and physicochemical features are more informative of tropism than the sequence alone (Sing, Low et al. 2007).

Model	# features	AUC	sensitivity
indicator	1000	0.860	0.616
aaindex	2800	0.829	0.565
full	5544	0.847	0.587
RF(1)	144	0.860	0.673
RF(5)	241	0.863	0.634
SVM(1)	**123**	**0.889**	**0.706**
SVM(5)	362	0.879	0.674
Lasso	**102**	**0.893**	**0.674**
RF(1)_SVM(1)	264	0.878	0.683
RF(5)_SVM(5)	588	0.875	0.668
RF(1)_Lasso	226	0.883	0.652
RF(5)_Lasso	315	0.874	0.639
SVM(1)_Lasso*	**218**	**0.892**	**0.686**
SVM(5)_Lasso	448	0.881	0.674
RF(1)_SVM(1)_Lasso	340	0.868	0.644
RF(5)_SVM(5)_Lasso	653	0.883	0.685

Table 3.4 Performance of models based on feature sets and combination of feature sets selected using different feature selection methods. Models are named after the feature selection method with the number in parentheses indicating the percentage cutoff of the ranked features. Sensitivity shown is at specificity of 11/25 rule in the main dataset. The clonal model is indicated with an asterisk.

3.3.4 Combining structure and sequence

In order to test how much performance improvement can be gained from incorporation of another V3 loop structure, we repeated the same model construction and feature selection procedure using a more recently reported V3 loop structure (PDB ID 2QAD) (Huang, Lam et al. 2007). The descriptor vectors based on the two structures had a mean correlation of 0.71 ($p < 0.01$) for the V3s in the main dataset which reflected the similarity of the two descriptors. Moreover, models based on the 2QAD structure showed a similar performance to those based on the initially used 2B4C structure. The 2QAD-based SVM(1)_Lasso model showed an AUC of 0.872 and sensitivity of 0.639 at the specificity of 11/25 rule. The selected features of models based on the two structures had an overlap of 5-8% in case of features selected through SVM and Lasso methods and 20-23% for the RF-selected features.

We also tested the performance of combined structure and sequence descriptors. A combination of descriptors was obtained from concatenation of the vectors of respective descriptors structure- and/or sequence-based.

As already indicated by the correlation of structural descriptors, the combination of two structures did not result in improvement of the prediction. The combination of the clonal models (SVM(1)_Lasso) based on two different structures obtained an AUC of 0.857 and sensitivity of 0.625 at the specificity of 11/25 rule. Combining the indicator descriptor with the 2B4C-based structural descriptor or with the combined structural descriptors reached in both cases an AUC of 0.875 and sensitivity of 0.630 at the specificity of 11/25 rule.

3.3.5 Other datasets

In order to compare the performance of the proposed structural descriptor with other methods of coreceptor usage prediction, we tested the method on the datasets of the study of Sander et al. (Sander, Sing et al. 2007) (*Sander dataset*) and the study of Dybowski et al. (Dybowski, Heider et al. 2010) (*Dybowski dataset*). Summary statistics of all the tested datasets is presented in Table 3.5.

dataset	# sequences	X4 sequences	R5 sequences
main	1188	215	973
Sander	1357	205	1152
Dybowski	515	151	364
HOMER	954	167	787
HOMER-filter	412	39	373
MVC	53	3	25

Table 3.5 Summary statistic of the used datasets. Only a subset of the MVC dataset was phenotyped which is reflected by the number of X4 and R5 viruses that is lower than the total number of sequences in the dataset.

In order to avoid overtraining and to test the performance of the model independent of the training dataset, we did not repeat the feature selection procedure on the new datasets but used the features of the clonal model. Structural descriptors of the Sander and Dybowski datasets were built and limited to the clonal model features and tested in a 10x10 cross-validation setting.

The clonal model showed a slightly poorer performance than the original method on the Sander dataset with the AUC of 0.901 (0.923 reported by Sander et al.) and sensitivity of 0.782 at the specificity of 11/25 rule on this dataset (0.774 reported by Sander et al.), see Table 3.6. The result of Sander et al. was obtained on a dataset with no insertions or deletions relative to the reference structure and involved costly side chain modeling steps. In contrast, clonal model was constructed based on a different dataset, prediction procedure did not involve any additional to V3 encoding and model training computational steps. On the Dybowski dataset the clonal model reached better performance as compared to the original method with AUC of 0.948 (0.937 reported by Dybowski et al.) and specificity of 0.838 at the sensitivity of 11/25 rule (0.810 reported by Dybowski et al.), see Table 3.6. Again, this result was obtained efficiently, without additional modeling steps employed by the original method.

Both the Sander and Dybowski datasets were composed of clonal sequences from the Los Alamos database. Next, we tested the method on clinically derived patient data from the HOMER cohort (Brumme, Dong et al. 2004) (*HOMER dataset*). The dataset was filtered to contain one sequence per patient which resulted in a set of 954 sequences out of which 167 of X4 viruses. Each sequence in the clinical dataset represents a population of variants genotyped and phenotyped in bulk, an approach used in standard clinical practice. Such sequences contain ambiguous positions with alternative amino acids corresponding to different variants in the population. Ambiguous positions were represented by an average of vectors of indices of all alternative amino acids on a given position. Due to these differences between the clinically and clonally derived data, we repeated the feature selection on this dataset and chose the best performing model (Lasso) as the *clinical model*. The clinical model showed AUC of 0.774 and sensitivity of 0.463 at the specificity of 11/25 rule on this dataset, a result significantly higher ($p < 0.01$, paired Wilcoxon test) than the one of the indicator method (Sing, Low et al. 2007) with AUC 0.743 and sensitivity of 0.451 (Table 3.6). Similar to the clonal model, the test of reselecting the Lasso features within the cross validation runs resulted in a ~1.8% decrease of performance scoring nevertheless significantly higher (p~0.002) above the indicator model.

As previously shown (Sing, Low et al. 2007), genotypic methods of tropism prediction perform worse on clinically derived data. This performance decrease might be due to noise introduced by ambiguous positions and to the presence of minorities undetectable by bulk genotyping of viral populations. In order to test whether the proposed strategy of handling sequence ambiguities is a source of performance decrease we tested two sequence sets derived from the original dataset and not containing ambiguities. The HOMER-filter set was obtained from the HOMER set after removing all sequences containing ambiguous positions. It contains 412 sequences with 39 X4 virus sequences (Table 3.5). The HOMER-gap set was derived from the HOMER set by replacing all ambiguous positions with gaps. The same steps involving construction of structural descriptors, feature selection using Lasso and evaluation using cross validation were performed on the two datasets not containing ambiguous positions. Lower prediction performance was observed on both datasets (Table 3.6), suggesting that ambiguous positions carry information important for the tropism prediction and that the presence of undetectable minorities, rather than ambiguous positions, might be the reason for the low predictive performance of models applied to clinically derived data.

As shown by Sing et al. (Sing, Low et al. 2007), the performance of coreceptor prediction methods on clinical data improves after enhancing the sequence information with clinical correlates, such as VL or $CD4^+$ T cell counts. Accordingly, adding these clinical data as new features of the clinical model significantly improved the predictive performance ($p < 0.001$, paired Wilcoxon test) over the clinical model to AUC 0.803 and sensitivity 0.474 at specificity of the 11/25 rule (Table 3.6). This result was also significantly higher than the result of indicator method containing clinical correlates both in AUC and sensitivity at the 11/25 rule specificity ($p < 0.001$, paired Wilcoxon test), see Figure 3.18. This demonstrates the higher prediction accuracy of the proposed method based on preselected structural and physicochemical features of the V3 loop over the commonly used sequence-based methods (Sing, Low et al. 2007).

dataset	model	features used	AUC	sensitivity	original method / geno2pheno	
					AUC	sensitivity
Sander	clonal	218	0.901	0.782	0.923	0.774
Dybowski	clonal	218	0.948	0.838	0.937	0.810
HOMER	clinical	59	0.774	0.463	0.743	0.451
HOMER-filter		95	0.657	0.313		
HOMER-gap		106	0.774	0.303		
HOMER	clinical (CD4, VL)	61	0.803	0.474	0.781	0.442

Table 3.6 Performance of the clonal and clinical models on different datasets. Clonal model performance on the datasets of clonal origins (Sander and Dybowski) is compared to the performance of the original methods developed on these datasets. Clinical model constructed on the HOMER dataset and with or without gaps is compared to the geno2pheno performance. Clinical (CD4, VL) is the clinical model of the HOMER dataset coupled with clinical correlates ($CD4^+$ T cell counts and VL).

As the last test set we used a clinical dataset of a German cohort undergoing MVC therapy (*MVC dataset*) and tested the capacity of the method to predict therapy outcome based on sequences obtained at the therapy start. In this dataset we identified 53 cases of patients under MVC whose therapy outcome could be assessed based on the VLs. Therapy success was defined as a 2 log decrease in VL with respect to the level at the therapy start or a VL drop below 50 copies/ml measured three months after the therapy start. We classified the virus sequences at the therapy start for their coreceptor tropism in order to investigate the capacity of the structural descriptor to predict the therapy outcome. Since the MVC dataset was derived through clinical bulk sequencing we used the clinical model to predict the phenotype of the sequences in this dataset. We used prediction score at the specificity of 11/25 rule which corresponds to a false positive rate (FPR) of 6.28%, in the HOMER dataset as a classification cut-off between the R5 and X4 viruses. The FPR is an estimation of the expected proportion of sequences incorrectly classified as X4 viruses with a given cutoff and is calculated as the ratio of R5 viruses scored above the threshold to all viruses scores above the threshold in a 10x10 cross-validation. As a measure of general prediction quality on the MVC dataset we used Matthews correlation coefficient (MCC) that expresses the correlation of the observed and predicted binary classification and is suited for a dataset with an unbalanced proportion of elements from two classes.

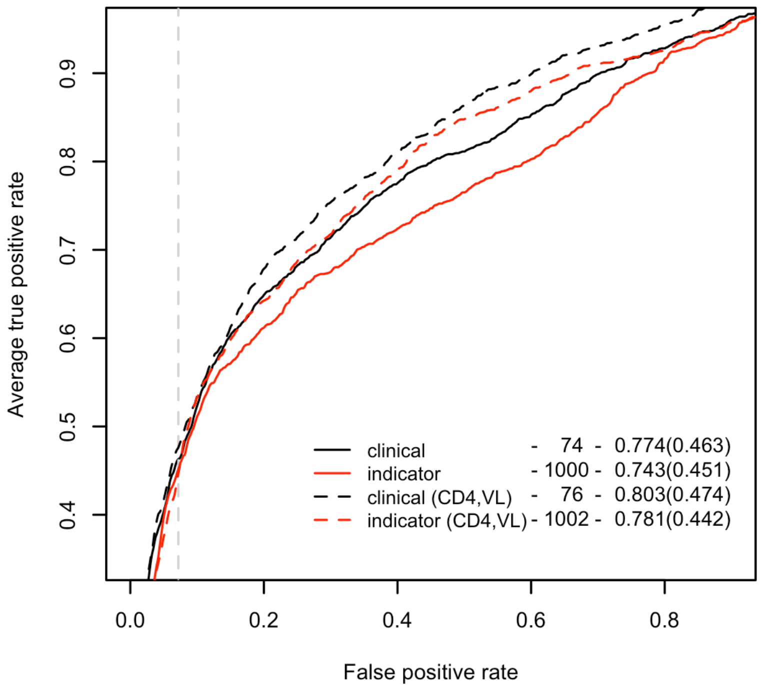


Figure 3.18 Performance of clinical (black) and indicator (red) models on the HOMER dataset without (solid line) and with patient clinical correlates (dashed lines). Number of features, AUC and sensitivity at the specificity of 11/25 rule (gray dashed line) in brackets are indicated in the legend.

Out of 53 patient cases in the MVC dataset 5 experienced a therapy failure (Table 3.7). With the decision cutoff based on the 11/25 rule specificity (specificity 0.928, score 0.097) two of the therapy failure sequences were predicted as X4 viruses. The other three cases of therapy failure were classified as R5 viruses with a high FPR (~80%). Notably both cases (32 and 34) were also phenotyped as R5 virus which suggest the presence of undetectable minorities as the reason for the classification error.

The remaining 48 patients experienced therapy success. 41 of the cases were classified as R5 viruses by the clinical model with a median FPR ~55%. Seven cases were classified as X4 viruses with a median FPR of 3%. Three of these cases were phenotyped as R5 viruses, two as dual viruses. The therapy outcome prediction showed an overall accuracy of MCC=0.34. We also predicted tropism of the sequences in this dataset using the indicator method. We observed higher sensitivity of the proposed structure-based method in finding therapy failure cases with also a higher rate of falsely predicted therapy success cases compared to the indicator method trained on the same clinical data. The indicator method reported correctly only two therapy failure cases however showed a lower number of incorrectly predicted therapy success cases (four). Given the imbalance between the numbers of therapy success and failure cases indicator method showed a generally lower accuracy compared to clinical model with MCC=0.29.

Both clonal model and indicator method trained on clonal data showed a generally lower capacity of detecting therapy outcome. Both models predicted correctly only one therapy failure case. Partial results of the MVC dataset analysis are provided in Table 3.7.

case	therapy outcome	phenotype	score	FPR (%)	predicted phenotype
33	-		0.731	0.4	X4
46	-		0.187	2.9	X4
58	-	R5X4	0.176	3.0	X4
32	-	R5	-0.013	79.6	R5
34	-	R5	-0.015	81.1	R5
52	+		0.672	0.6	X4
17	+	R5	0.553	0.9	
44	+	R5X4	0.228	2.2	
24	+		0.195	2.8	
15	+	R5X4	0.175	3.0	
2	+	R5	0.139	4.4	
36	+	R5	0.100	7.1	
1	+		0.083	9.3	R5
29	+	R5	0.075	10.6	
30	+		0.075	10.6	
45	+		0.0.73	11.1	
…	+	…	…	…	

Table 3.7 Patient cases in the MVC dataset. First five rows of the table show patient cases that experienced therapy failure (-) three of which were correctly predicted by the clinical model. First eleven cases of the remaining 48 cases that experienced therapy success are presented in a decreasing order of the prediction score. Seven cases show score above the selected cutoff and are hence predicted as X4 viruses. Two of the seven cases (44, 15) are classified as R5X4 viruses despite the therapy success. Remaining therapy success cases are correctly predicted with a median FPR ~55%.

3.3.6 Feature analysis

In order to facilitate the interpretation of the large number of selected features we clustered the 56 amino acid indices into similarity groups. As a similarity score among the properties we used the absolute value of their correlation. Thus indices that express the same distances among amino acids are considered similar. We performed hierarchical clustering of the 56 amino acid indices and computed silhouette values (Rousseeuw 1987) in order to select the best set of clusters. The highest silhouette value was obtained for a partitioning of indices into 12 clusters. The highest silhouette value assigned to partitioning into less than 12 clusters was obtained for four clusters. We selected this number of clusters for further analysis as it contained a small number of interpretable attribute groups that are relatively well separated as inferred from the silhouette value. The selected clustering is presented in Figure 3.19.

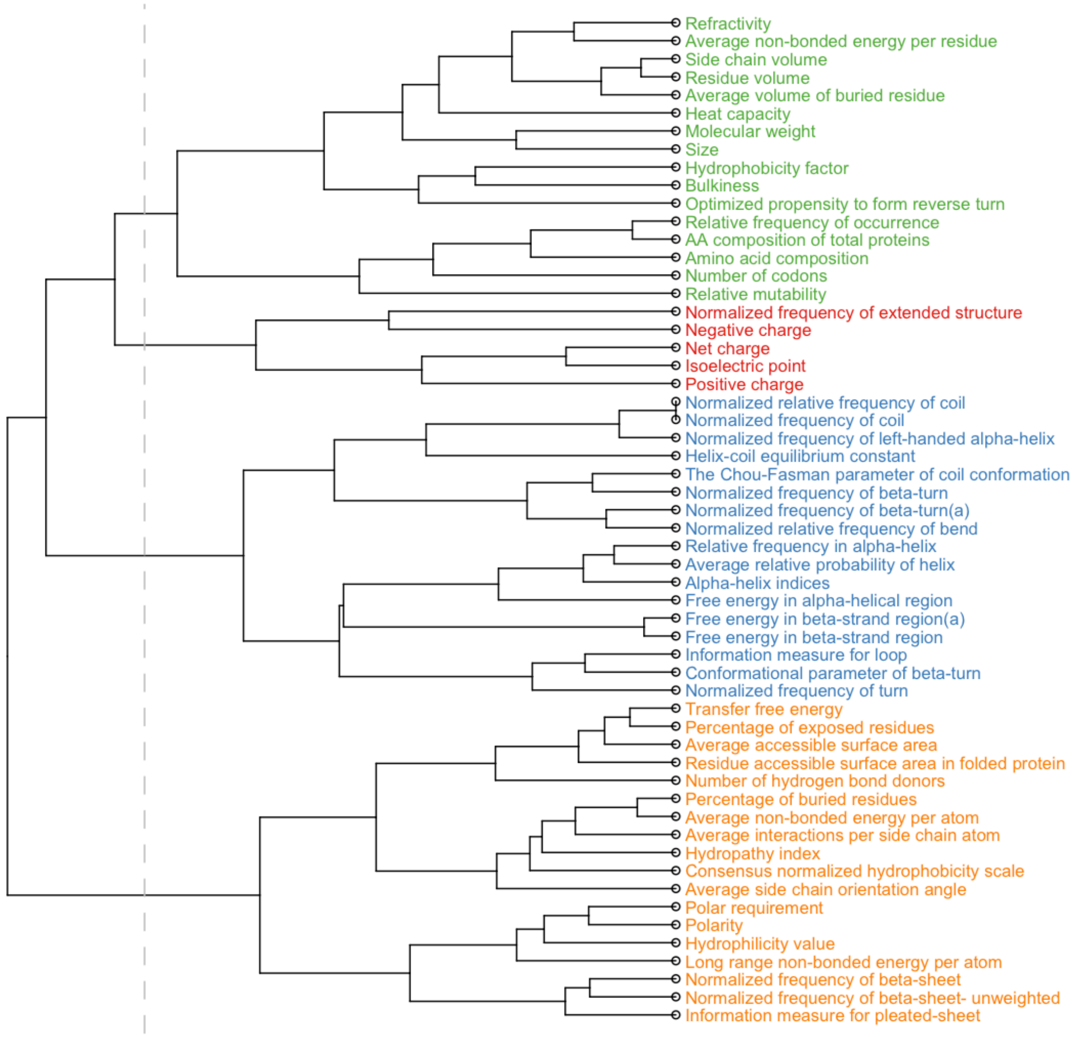


Figure 3.19 Hierarchical clustering of the amino acid indices. Vertical line indicates the separation of the tree into clusters. Labels of the tree are colored according to the clusters – cluster 1 in green, 2 in red, 3 in blue, 4 in orange.

Cluster 1 is composed of two types of indices – related to residue size and volume and to residue occurrence in proteins. Cluster 2 contains the smallest number of indices and is composed of indices related to residue charge. Indices of cluster 3 are related to secondary and tertiary structure of proteins. Cluster 4 contains indices related to different structural properties e.g. residue occurrence in β-sheet, solvent accessibility, amino acid polarity or hydrophobicity.

By combining amino acid indices with specific positions in the V3 loop the proposed features provide description of physicochemical properties along the structure of the loop. Since features selected for the clonal model are the most informative for the prediction of coreceptor usage, their analysis can provide insights into the physicochemical and structural determinants of virus tropism. Features of the clonal model were selected based on two different feature selection methods – Lasso and SVM methods. Among 218 features in this model seven were selected by both methods. Three of the features describe electrical charge along positions 319-322. Two of the features are structure-related (“Free energy in β-strand region”, “Normalized frequency of turn”) and describe positions 304 and 305. The remaining two features selected by both methods are based on amino acid indices “Number of codons” at position 297 and “Relative mutability” at position 307.

Both feature selection methods allow for feature ranking based on the feature coefficients in the respective linear models. We inspected the top-scoring features in both rankings (Figure 3.20). SVM scoring follows a gamma distribution with a shape parameter of 2.2. The selected features follow a close to uniform distribution in the range of values above the chosen cutoff. Feature scores based on Lasso selection are distributed over a wider range of values and contain several high-scoring outliers.

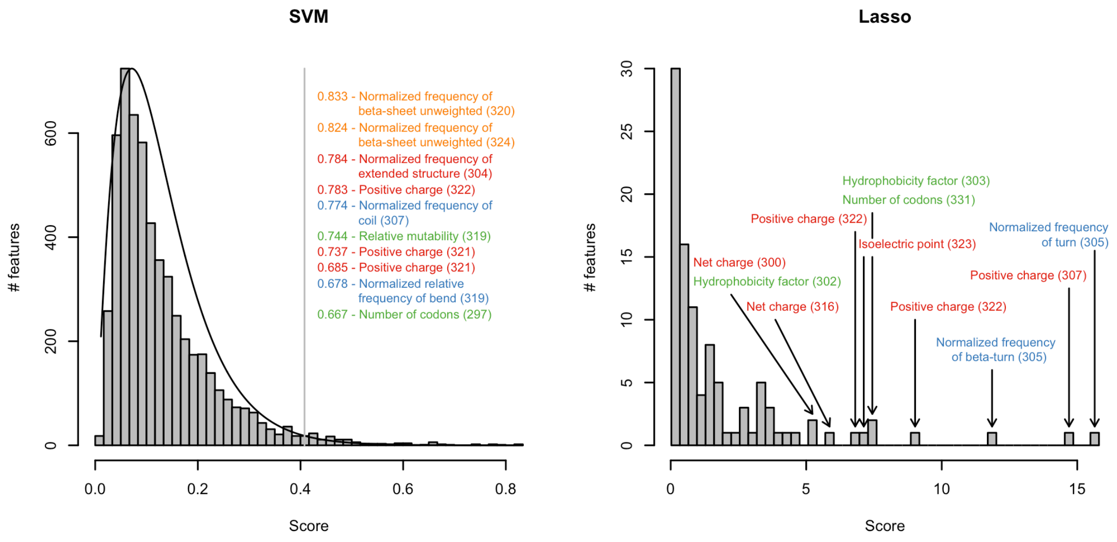


Figure 3.20 Distribution of scores of features selected using SVM (left panel) and Lasso (right panel) methods. On the left panel the vertical line indicates the cutoff for features selected for the clonal model. The scores of the top-scoring features are listed. On the right panel top-scoring features in the distribution are indicated. Positions of the features mapped on the V3 loop structure are indicated in brackets, labels are colored according to their clusters as in Figure 3.19.

Among the top-scoring features selected by both methods we found “Positive charge” at the stem position 322 corresponding to the position 25 in the consensus sequence. Highly ranked features in the SVM scoring include also “Positive charge” at the position 321. Additionally SVM scoring pointed to secondary structure propensities and mutability at the loop stem (“Normalized frequency of coil”, “Normalized frequency of β-sheet unweighted”, “Normalized relative frequency of bend” at positions 307, 319-320 and 324).

Among the high-ranking features in the Lasso scoring we found mostly charge indices at the loop stem (“Positive charge”, “Isoelectric point” and “Net charge” at positions 307, 316, 322-323) and at the loop base (“Net charge” at position 300). Additionally we found “Hydrophobicity factor” at the loop base positions 302-303. Two structure-related features based on “Normalized frequency of turn” and “Normalized frequency of β-turn” amino acid indices at the base position 305 were also scored high by the Lasso method.

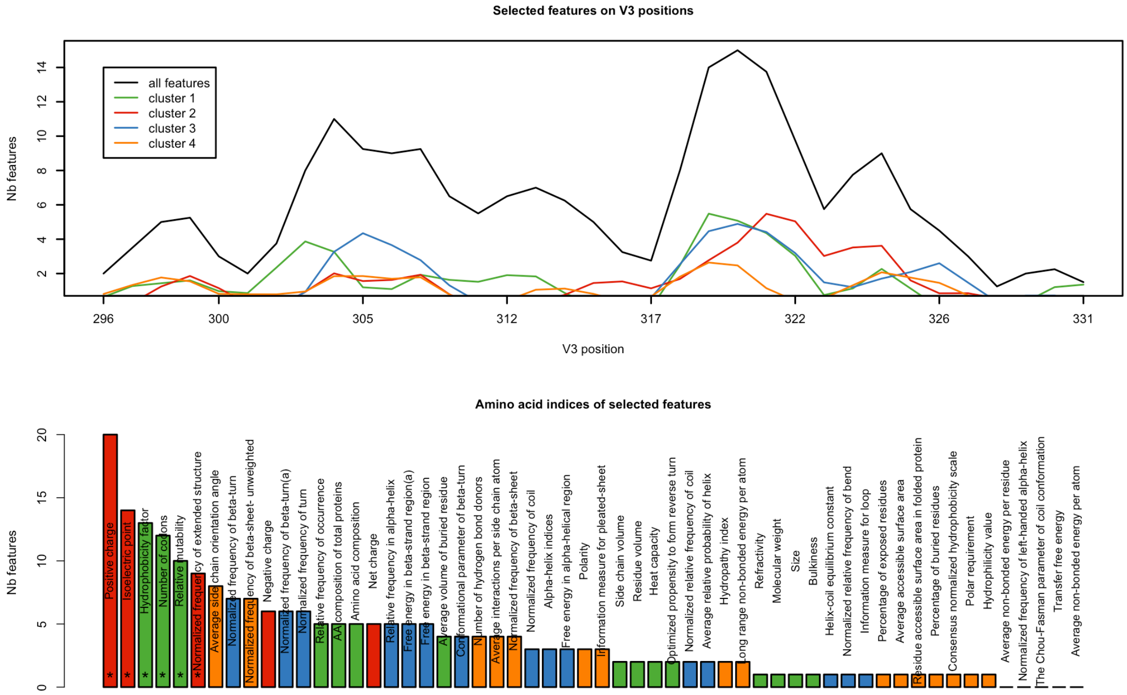


Figure 3.21 Overrepresentation of features of different amino acid index clusters among features selected for the clonal model. On the upper plot positions of the proximities of features of the clonal model were mapped on the positions on the reference sequence. Numbers of selected features on the following sequence positions were summed and averaged over a window of three neighboring sequence positions. This way obtained distribution of all features is represented by the black line, distributions of features of four clusters are represented by lines in colors as in Figure 3.19. Bottom histogram shows the number of features related to amino acid indices among features selected for the clonal model. Bars are colored according to the amino acid index cluster. Significantly overrepresented amino acid indices are indicated with an asterisk.

Next, we inspected the distribution of selected features of the clonal model along the V3 loop and the amino acid indices most often appearing among the selected features (Figure 3.21). The selected features appear to group around two regions in the sequence: between positions 303-312 and even more strongly between positions 318-324. These regions correspond to two strands of the V3 stem. In the first region (303-312) the selected features are based mostly on indices from clusters 1 and 3. In the second region (318-324) there is also a high number of features based on indices from cluster 2 which are predominantly associated with residue charge. Among individual amino acid indices we found six that are significantly enriched among selected features:

"Positive charge", "Isoelectric point", "Hydrophobicity factor", "Number of codons", "Relative mutability" and "Normalized frequency of extended structure" (Figure 3.21, bottom panel). Notably all of the significantly overrepresented indices belong to clusters 1 or 2.

The location of the features in the V3 structure is illustrated in Figure 3.22. Most selected features describe positions in two regions on both sides of the loop stem located in proximity of positions 304, 307 and 319-321, respectively (Figure 3.22A). We labelled these regions core sites (CSs) 1 and 2. In the bound conformation of the loop (PDB code 2QAD) CS1 and CS2 are located closer to each other than in the open conformation (Figure 3.22B). We investigated the interactions of the residues of the two sites (Word, Lovell et al. 1999) and found that in the bound conformation residues of CS1 and CS2 form interacting pairs between two sides of the central loop stem. In particular residues 304 and 307, which are located in one side of the loop stem, form van der Waals interactions with residues 319 and 320, which are located on the other strand of the stem. In the open conformation CS1 and CS2 are more widely separated and the interactions between two sides of the loop are not observed. Position 324 is also associated with a high number of selected features and is located on the loop stem however does not interact with CSs in either of the conformations.

The analysis of the location in the loop structure of the features based on significantly overrepresented indices confirmed the importance of CS1 and CS2 in determining coreceptor usage. These overrepresented indices are related to residue charge and hydrophobicity (e.g. "Positive charge", "Isoelectric point", "Hydrophobicity factor"), secondary structure (e.g. "Normalized frequency of β-turn", "Normalized frequency of coil") or mutability (e.g. "Number of codons", "Relative mutability"), and are all found in CS1 or CS2 as well as residue 324 (Figure 3.22).

We also investigated which clusters of indices were significantly overrepresented among selected features. The only cluster significantly overrepresented among the selected features was cluster 2 ($p < 0.05$). Three out of five features of this cluster are also overrepresented individually in the full set of selected features – "Positive charge" (22 features), "Isoelectric point" (14 features) and "Normalized frequency of extended structure" (9 features). Most of the selected features from this cluster describe residues of CS1 and CS2, in particular positions 319-320 and position 324 (Figure 3.22).

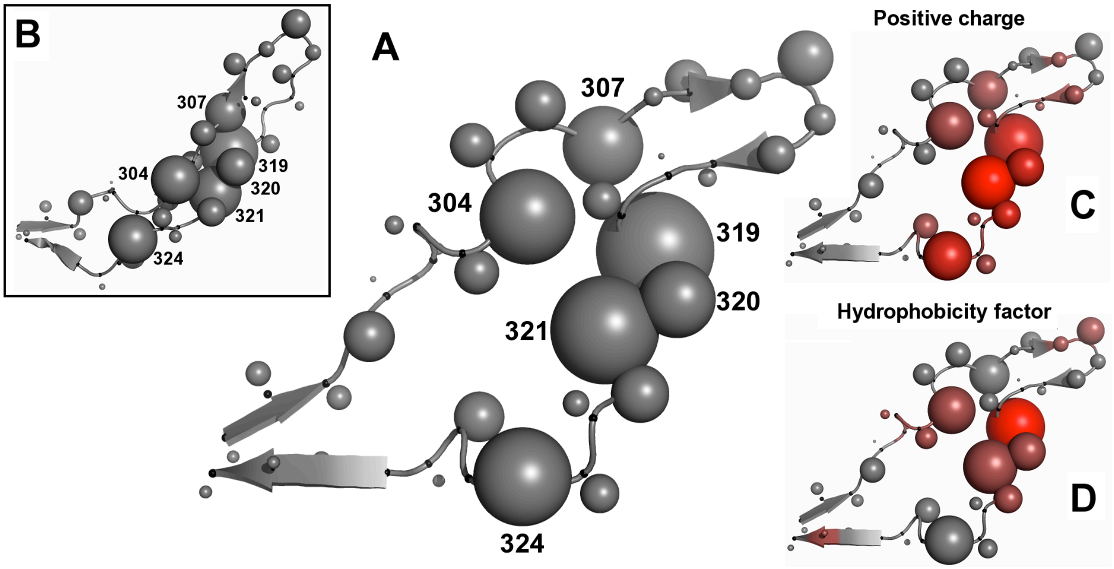


Figure 3.22 Important V3 positions and amino acid indices in the V3 structure. (A) 2B4C V3 structure used in this study. Cα atoms are marked with small black spheres in the loop backbone in gray. Representative atoms are represented by gray spheres with the size proportional to the number of selected features mapped on the respective V3 position. Positions showing largest number of selected features are numbered. (B) V3 structure in a bound conformation (2QAD) (Huang, Lam et al. 2007) with the same sphere representation as the structure in panel (A). Positions important for the tropism come close to each other in this conformation. (C, D) Structure representation of V3 as in panel (A) with positions of the loop colored according to the ratio of selected features related to "Positive charge" (C) and "Hydrophobicity factor" (D) to the overall number of the selected features present on the respective V3 position with red indicating high proportion of given features and gray none.

3.3.7 Discussion

Physicochemical and structural properties of proteins determine their binding affinities. Prediction methods of HIV-1 coreceptor usage based uniquely on V3 sequence do not account for this type of properties nor provide the information about loop features that are crucial for the interaction. Here we present a predictive model of HIV coreceptor usage based on V3 sequence and structure (Huang, Tang et al. 2005). The method is constructed upon a set of features that was selected from a large initial feature set. This way reduced model shows better performance than the one based on the initial feature set, both in terms of prediction accuracy and computational efficiency. Moreover, the model affords an interpretable set of physicochemical properties pertaining to two regions of the loop structure that play crucial role in determining viral tropism. This study reaches beyond the classical prediction methods by providing additional insights into the determinants of viral tropism.

The initial set of features is large and highly redundant which is reflected by the small overlap of features selected by different feature selection methods. The proposed structural descriptor appears to encode both structure and sequence information as adding the sequence binary encoding to the descriptor does not improve its performance. In contrast, the distance-based descriptor of Sander et al. (Sander, Sing et al. 2007) is complementary to the sequence which is demonstrated by the improved performance of the descriptor combined with the sequence.

The prediction performance of the proposed method is significantly higher than that based on sequence only (Sing, Low et al. 2007) both on the clonal and clinical datasets with and without patient clinical markers. It shows also a higher or similar prediction performance to that of other structure-based methods (Sander, Sing et al. 2007; Dybowski, Heider et al. 2010) without modeling steps that increase the computational cost of the prediction. Therefore both the accuracy and efficacy of our approach allow for its usage as a server application.

We assessed the capacity of our method to predict therapy outcome of a patient cohort treated with MVC. This analysis is limited due to the small number of cases in this dataset and the lack of a comparable study. With increasing use of entry inhibitors, therapy outcome data are expected to become more abundant and the capacity to train models predicting therapy outcome will improve. The higher performance of the clinical over the clonal model in predicting therapy outcome suggests that targeted datasets can produce more reliable models.

The analysis of features important for the coreceptor tropism points to two critical sites in the loop stem, consisting of residues 304, 307 and 319-322, respectively and to position 324 located closer to the base of the stem. Residues on both sides of the stem form interactions in the bound conformation of the loop probably contributing to the rigid form of the loop upon binding. Predominantly charge- and hydrophobicity-related indices in the CS1 and CS2 around positions 307 and 319-322 respectively, and secondary structure of 304 in CS1 and of residue 324 appear to be important for coreceptor usage

The results of other studies of structural features related to the tropism are in general accordance to our results. A recently published method (Masso and Vaisman 2010) predicts coreceptor usage based on a perturbation vector reflecting relative change in compatibility of a given V3 the sequence and structure with the reference structure (Vranken, Budesinsky et al. 1995). The ten most important positions for the coreceptor usage, according to this study are positions 302-304, 306-307, 309, 312, 322, 324-325. No additional interpretation of the characteristics of these positions is however provided. Sander et al. (Sander, Sing et al. 2007) rank high the distances from residues 298, 302, 306, 308, 315, 317, 319, 321, 322 and 328. The regions found in my analysis are close or in-between the positions listed by Sander et al. However the ranking of Sander et al. is based on the importance of distances among functional atom types of the V3 residues which is not equivalent to the importance of the residue itself. Dybowski et al. (Dybowski, Heider et al. 2010), point to electrostatics hulls that are of highest importance for the classification: around positions 306, 321 and 322, and between position 301 and 326. Additionally, hydrophobicity of residues 303 and 307 appears to be important for the virus tropism based on Dybowski et al. analysis. Notably, the feature selection methods used in this study applied on Dybowski dataset, showed a strong overrepresentation of

“Hydrophobicity factor” amino acid index (over 2-fold higher than any other) among selected features in this dataset.

Given the considerable structural flexibility and sequence variability of the V3 loop, individual features of this region distinguishing between two virus phenotypes are hard to define. We performed a comprehensive analysis of a large number of physicochemical residue characteristics in various locations on the loop and pointed to those that are the most informative of the tropism. The resulting method offers better performance over sequence-based method with a comparable efficiency and a direct interpretation of structural and physicochemical determinants of tropism. The method was implemented as a server application.

3.4 Conclusions

Variability of HIV underscores the importance of population scale studies comprising comprehensive sets of viral variants. V3 loop is an example of a particularly highly variable part of the virus genome. Its evolution is accelerated by evasion of the host immune recognition, constrained by the need of host receptor binding and subject to drift due to the lack of other functional restrictions. The multiplicity of forces influencing V3 composition is reflected in the complex evolutionary trajectories of sequences in sequence space where a clear pattern of X4 virus evolution is difficult to discern. V3 loop appears to be highly flexible, both in sequence and structure, which contributes to the limited accuracy of the coreceptor usage prediction methods.

The two studies presented in this chapter were performed on sequence data on the virus population level. The analyses extend beyond the classical coreceptor tropism classification methods based on sequence and/or structural information. Consequently they offer biological and evolutionary insights into the virus coreceptor tropism. Even such in-depth analyses fail however to explain all virus variants. One potential reason for this might be the incompleteness of the sequence data. Another reason might lie in other host and environmental factors impacting HIV tropism in addition to the viral V3 loop and not being accounted for in the classification models.

The first reason is well demonstrated by the inaccuracy of coreceptor prediction on the clinically derived data as compared with the clonal data. Clonal data represent individually cloned and tested virus strains, the sequences are therefore unambiguous and their phenotypes reliably tested. Clinically derived data represent entire virus populations sampled from patient plasma, the sequences and phenotypes represent therefore a consensus picture of the dominant circulating virus. Such simplified and often incomplete view of a virus population derived through bulk virus sequencing and phenotyping results in an important decrease in the accuracy of coreceptor prediction methods. With the development of the next generation sequencing (Metzker 2010) virus studies on the population scale enter currently a new stage. Next-generation or deep sequencing platforms can read millions of base pairs in a more cost-effective and faster way than traditional Sanger sequencing. This technology has already been used to address such questions as resistance mutations in minority populations (Hoffmann, Minkah et al. 2007; Wang, Mitsuya et al. 2007; Johnson, Li et al. 2008; Simen, Simons et al. 2009), evolution of virus populations (Poon, Swenson et al. 2009) or coreceptor usage characterization of entire within patient populations (Archer, Braverman et al. 2009). The increased use of the next generation sequencing technologies will allow to resolve virus populations more accurately and overcome the drawbacks of using the consensus sequence representation.

The second potential reason for the incapacity of population-based models to correctly classify all virus variants is the lack of host and environmental factors in the analysis. Even though studies of massive amounts of virus population data generated by next generation sequencing can provide a picture of the evolution of virus populations to an unprecedented level of detail, without accounting for other environmental and host factors affecting virus tropism, these analyses cannot explain the biological mechanisms of the coreceptor switch. Factors that trigger the virus tropism switch in the later stages of infection are still unclear. However studies of virus population sequences independent of their level of detail do not afford causal information of the biological reasons for this switch. Virus evolves in response to specific conditions in its environment such as host immune system, availability of specific host cells or the presence of drugs. Expanding the virus population studies to include host and environmental parameters would provide a more comprehensive picture of the virus evolution with a potential of explaining the basis of the coreceptor switch.

An additional drawback of virus population-based methods aimed at predicting the virus coreceptor tropism is their limited capacity to predict therapy outcome. Certain patients analyzed using structural descriptor described in this chapter experienced therapy outcome unexpected from the virus phenotype. This indicates that therapy outcome might be affected by other than V3 loop environmental and patient factors. General virus fitness or availability of receptors on the surfaces of host cells are other examples of factors that might influence response to treatment. With the increasing use of MVC therapy, it is important to develop more comprehensive models of virus phenotype comprising not only V3 sequence but also virus fitness, drug response and coreceptor dependencies. Data required for such models should involve experimental testing beyond phenotypical coreceptor usage however they will allow for predicting the therapy outcome with a higher accuracy as compared to current coreceptor usage prediction methods.

CHAPTER 4 – Single cell scale

host and drug factors affecting viral cell entry in addition to the virus genome

The V3 loop of the gp120 protein is recognized as the major determinant of the HIV coreceptor tropism. Population scale studies of viral V3 sequences such as described in the previous chapter provide classification methods that support the clinicians in the patient treatment with therapies based on entry inhibitors. The accessibility of bioinformatical prediction methods and the low cost of genotyping are among the major advantages of using sequence-based computational approaches for predicting coreceptor usage. While being an invaluable support in the clinics, the genotype-based methods are subject to two major limitations. Firstly, binary classification of virus tropism provides a simplified picture of the virus phenotype and limited insights into the biological determinants of the coreceptor tropism. Secondly, sequence-based statistical models are based on observed or functionally tested cases are therefore prone to bias inherent to the data and, more importantly, are unable to provide reliable predictions of cases unobserved in the past and unenclosed in the train dataset. The latter limitation becomes important in the view of changing therapies and fast evolving virus.

The limitations of sequence-based prediction methods call for development of improved models providing biological description of the phenotype, capable of characterizing cases unobserved in the data and this way of discovering new undescribed cases and producing testable hypotheses about them. Development of such models requires expanding of the binary sequence and phenotype data with information regarding the mechanism of the interaction and a broader spectrum of phenotypes. The enhanced models would result in a comprehensive picture of the biology of virus tropism allowing for characterizing of viruses and hence for deriving personalized treatment strategies.

In this chapter we present an analysis of the HIV-host interaction at the scale of single cell. In this project several virus V3 loop variants were examined with respect to the efficiency of their entry into host cells in the presence of entry inhibitors of different types and in varying concentrations. The data from a high-throughput single cell assay was collected, processed and used for building models of virus cell entry efficiency. The models reflect virus phenotype involving host and drug parameters and, in this way, extend the classical coreceptor tropism classification based on genotype.

This project was performed in collaboration with the lab of Prof. Dr. Hans-Georg Kräusslich and Dr. Barbara Müller at the Department of Virology, Heidelberg University and the virology group of Dr. Rolf Kaiser, Institute of Virology, University of Cologne. The wet-lab experiments were performed by Dr. Manon Eckhart in the Heidelberg lab, phenotyped patient viral samples were provided by the Cologne lab. The manuscript describing our results is currently in preparation: Bozek, K., Eckhardt, M., Sierra, S., Kaiser, R., Müller, B., Kräusslich, HG., Lengauer, T. A comprehensive model of HIV cell entry phenotype based on multi-parameter single cell data.

4.1 Background

Several factors affect HIV cell entry. In addition to the composition of V3 region of the virus gp120 protein, analyzed in the previous chapter, factors such as coreceptors on the host cell and entry inhibitors play a role in virus cell entry.

In addition to the primary cellular receptor used by HIV, CD4 (Maddon, Dalgleish et al. 1986), two major coreceptors, the cytokines CCR5 and CXCR4, condition successful virus cell entry. Although CCR5 and CXCR4 share little sequence homology (Doms and Moore 1997), they can be used interchangeably by certain HIV strains. Both coreceptors are expressed on the surface of primary T cells and macrophages (Douek, Picker et al. 2003), CXCR4 can also be found on other types of human cells. Only a small percentage of CD4$^+$ T cells express CCR5 while CXCR4 is present on 80-90% of circulating naïve CD4$^+$ T cells (Nicholson, Browning et al. 2001). The Δ32 mutation in the CCR5 coreceptor is a naturally occurring genetic variation in the human population (Samson, Libert et al. 1996) that emerged as a result of ancient selection pressures. 1% of the European Caucasian population is homozygous and 15% heterozygous for the CCR5Δ32 mutation, which however appears not to have an effect on the life expectancy or health of the affected individuals (Liu, Paxton et al. 1996). This mutation prevents functional expression of the CCR5 chemokine receptor, resulting in resistance to HIV infection (Liu, Paxton et al. 1996) which underscores the relevance of cell surface receptor and coreceptor expression for the infection.

Blocking HIV cell entry represents a potential therapeutic strategy for HIV treatment. All steps of virus entry have been regarded as potential targets for anti-HIV intervention (Este 2003) with cell membrane fusion and coreceptor binding among them. Enfuvirtide is a fusion inhibitor (also known as Fuzeon, T-20, or ENF) that was approved as the first HIV entry inhibitor for use in patients (Fletcher 2003). It inhibits fusion of the virus with the host cell by binding to gp41 and preventing the creation of an entry pore for the capsid of the virus and this way keeping it out of the cell. Another class of entry inhibitors target cellular coreceptors. AMD-3100, a drug blocking CXCR4, is the first coreceptor antagonist described (De Clercq, Yamamoto et al. 1992). However, due to the severe side effects it was never approved for clinical use (Dai, Dou et al. 2005).

Blocking the CCR5 coreceptor is a more promising strategy for HIV entry inhibition based on the observation that host defects in CCR5 expression severely reduce acute HIV infection without impacting health or life span of an individual (Liu, Paxton et al. 1996). Several CCR5 antagonists have entered clinical trials (Este and Telenti 2007), however only MVC has been approved for patient treatment so far (Westby and van der Ryst 2005). This drug was shown to be active against a broad spectrum of R5 strains in vitro and in vivo with no observable adverse effects in patients (Fatkenheuer, Nelson et al. 2008). The majority of patients experiencing unsuccessful MVC treatment show a switch of viral coreceptor usage towards CXCR4 which is not affected by the drug. These X4 viruses might emerge from reservoirs in the patient present at undetectable levels before the start of the treatment (Westby, Lewis et al. 2006). However mutations conferring resistance to MVC were also reported in the V3 loop (Trkola, Kuhmann et al. 2002; Westby, Smith-Burchnell et al. 2007) enabling the virus to use the drug-bound coreceptor and thus obviating the need of tropism switch. This underscores that monitoring of coreceptor usage might be insufficient for accurate prognosis of therapy outcome. With the increasing number of patients under MVC treatment a better understanding of viral drug resistance as well as the development of algorithms predicting the efficacy of entry inhibitors are both timely and important.

In this project we developed models of the dependence of HIV entry efficiency on the following host and drug parameters: host cell receptor and coreceptor expression level, presence of entry inhibitors of different types in varying concentrations, and viral V3 loop variant. Models were based on single-cell level data generated in high-throughput experimental assays. The models provide a comprehensive description of virus phenotype that expands beyond the classical classification of coreceptor usage based on virus genotype only by including other host and environmental factors that affect HIV cell entry. The resulting picture of virus phenotype contributes to our understanding of the determinants of virus tropism, offers an improved capacity of identifying X4 viruses and allows to infer the phenotype from V3 sequence – a feature that is essential for the practical use of this method in patient treatment.

4.2 Clone selection

As the first step of this study we selected viral samples for experimental testing using the sequence space and coreceptor prediction methods described in the previous chapter. 94 sequences of phenotyped virus isolates derived from therapy-experienced patients were provided by the Cologne lab. We analyzed the sequences with respect to their position in sequence space (Figure 4.1) and reported and predicted phenotype. In order to locate the provided sequences in sequence space the R5/X4 we used the dataset from the sequence space analysis in chapter 3. Each clone was described by the mean distance of its sequence to all R5 and X4 sequences in the R5/X4 dataset and by the step of the clustering procedure in which its sequence joins a cluster as described in chapter 3. Small distances and cluster steps are typical for R5, large for X4 viruses. Three different prediction methods (geno2pheno (Sing, Low et al. 2007), WebPSSM (Jensen and van 't Wout 2003) and 11/25 rule (Fouchier, Groenink et al. 1992; Shioda, Levy et al. 1992)) were additionally used to predict the tropism of each clone.

For the initial testing eight samples were selected: two clearly R5, three clearly X4 and three of questionable tropism where the predicted and tested tropism as well as the position in sequence space were contradictory or close to the decision boundary. The following rounds of experiments were aimed at populating the sequence and phenotype space with X4 viruses and viruses of questionable tropism. In these experiments nine clones were selected of clearly X4 (252, 381, 391, 420, 719, 309, 315, 376, 631) and five of dubious (239, 334, 446, 468, 541) tropism. Out of the initially selected clones seven in the first round of experiments (220, 286, 651, 685, 822, 838, 924) and nine in the following experiments (252, 381, 391, 309, 315, 376, 631, 468, 541) (Table 4.1) showed entry efficiency on the level of 20-30%, a sufficient level for data analysis and modeling. Remaining clones showed entry efficiency below 5% and were therefore excluded from this study. Additionally, five lab-adapted clones of known characteristics (JRFL, SF162, HxB2, BaL, YU-2) and two reference clones (NL4-3 and NL4-3 R5) were added to the clone pool resulting in a set of 23 clones tested and characterized in our experimental and computational pipeline.

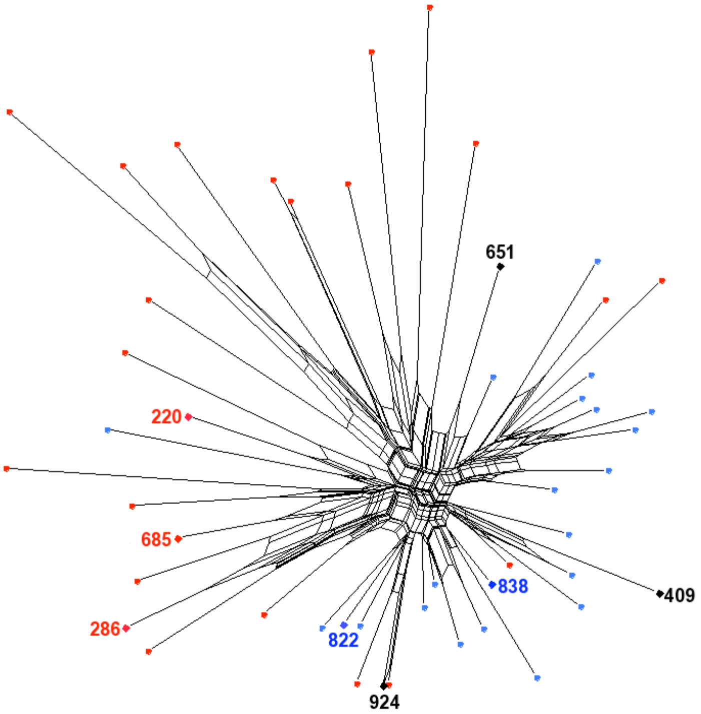


Figure 4.1 Sequence space analysis of the clones provided for experimental testing. The clones selected for the first set of experiments are located on a splitstree (Huson 1998) together with 25 X4 (red dots) and 25 R5 viruses (blue dots) from the dataset used in the sequence space analysis presented in chapter 3. Three of the selected clones (220, 286, 685) were phenotyped as X4 viruses and were found in locations typical for X4 viruses in sequence space which is characterized by long branches separate from R5 virus clades in the tree. Two selected clones (822, 838) were phenotyped as R5 viruses and clustered with R5 viruses in sequence space. Three remaining clones were phenotyped as either X4 (409, 651) or R5 (924) viruses but appeared in unexpected parts of sequence space – close to R5 viruses or on the boundaries between the phenotypes. These viruses were selected for experimental testing as viruses of dubious tropism.

Figure 4.2 shows another visualization of the proximities of the tested clones in V3 loop sequence space. The sequences of the Bonn cohort patient samples and the used lab-adapted clones were incorporated into the R5/X4 dataset used in the sequence space analysis presented in chapter 3. The dataset was hierarchically clustered and an optimal clustering of the sequences into 70 clusters was chosen based on the silhouette value (Rousseeuw 1987). 11 of the clusters contained clones tested in this study. Three sequences from these 11 cluster were sampled randomly. The sampled sequences together with the sequences of the selected clones are plotted in Figure 4.2. Clones selected as R5 (822, 838) as well as clone 924 and R5-tropic lab-adapted clones (JRFL, YU-2, BaL, SF162) are located in the part of the tree containing predominantly R5 sequences showing close similarity as reflected by short branch lengths. X4 clones (NL4-3, 286, HxB2) are located in clades occupied by predominantly X4 sequences and showing longer branch legths. Notably, HxB2 and NL4-3 show high similarity. Among the remaining clones, clones 252, 381, 315 and 376 form a closely related group seemingly dual-tropic. Other clones are located in parts of the tree occupied by both R5 and X4 sequences and showing intermediate branch lengths.

clone	phenotype	R5 distance	clustering step	geno2pheno score (prediction)	WebPSSM score (prediction)	11/25 positions (prediction)
220	X4	0.354	0.153	0.700 (X4)	-2.87 (X4)	GK (X4)
286	X4	0.400	0.280	0.973 (X4)	2.63 (X4)	RG (X4)
651	X4	0.346	0.153	0.170 (R5)	-2.80 (X4)	SD (R5)
685	X4	0.344	0.142	0.983 (X4)	0.74 (X4)	RK (X4)
822	R5	0.261	0.056	0.055 (R5)	-10.64 (R5)	SQ (R5)
838	R5	0.227	0.027	0.082 (R5)	-12.71 (R5)	SE (R5)
924	R5	0.275	0.074	0.120 (R5)	-8.25 (R5)	GE (R5)
252	X4	0.276	0.102	0.970 (X4)	-2.58 (X4)	RD (X4)
381	X4	0.276	0.102	0.950 (X4)	-2.66 (X4)	RQ (X4)
391	X4	0.372	0.280	0.990 (X4)	-1.01 (X4)	RR (X4)
468	X4	0.368	0.280	0.658 (X4)	-6.82 (R5)	SQ (R5)
541	X4	0.301	0.010	0.856 (X4)	-6.92 (R5)	GR (X4)
308	X4	0.328	0.108	0.658 (X4)	0.50 (X4)	SK (X4)
315	X4	0.302	0.102	0.983 (X4)	-3.93 (X4)	RQ (X4)
376	X4	0.277	0.102	0.910 (X4)	-5.32 (X4)	RQ (X4)
631	X4	0.317	0.142	0.960 (X4)	-1.49 (X4)	RT (X4)
NL4-3	X4	0.359	0.153	0.928 (X4)	0.50 (X4)	SK (X4)
NL4-3 R5	R5	0.230	0.052	0.090 (R5)	-12.17 (R5)	GE (R5)

Table 4.1 Selected clones for the initial (upper part of the table, above the double line) and the following rounds (bottom part, below the double line) of testing and the two reference clones (NL4-3, NL4-3 R5). Viruses were phenotyped by the Cologne lab. In order to locate the provided sequences in sequence space the R5/X4 dataset from the sequence space analysis in chapter 3 was used. Column “R5 distance” indicates the mean distances of a clone sequence to all R5 and X4 sequences in the R5/X4 dataset, “clustering step” indicates the step of the clustering procedure in which a sequence joins a cluster (see chapter 3). Three remaining columns indicate the score and phenotype predicted by three different prediction methods (geno2pheno (Sing, Low et al. 2007), WebPSSM (Jensen and van 't Wout 2003) and 11/25 rule (Fouchier, Groenink et al. 1992; Shioda, Levy et al. 1992)). The color indicates the type of clone: blue – clearly R5, red – clearly X4, magenta – dubious.

The selected V3 loop sequences were inserted into the genetic backbone of an X4 reference virus NL4-3 that is a standard laboratory virus (Adachi, Gendelman et al. 1986). This X4 reference virus, as well as R5 reference virus NL4-3 R5, were used as control X4 and R5 viruses respectively.

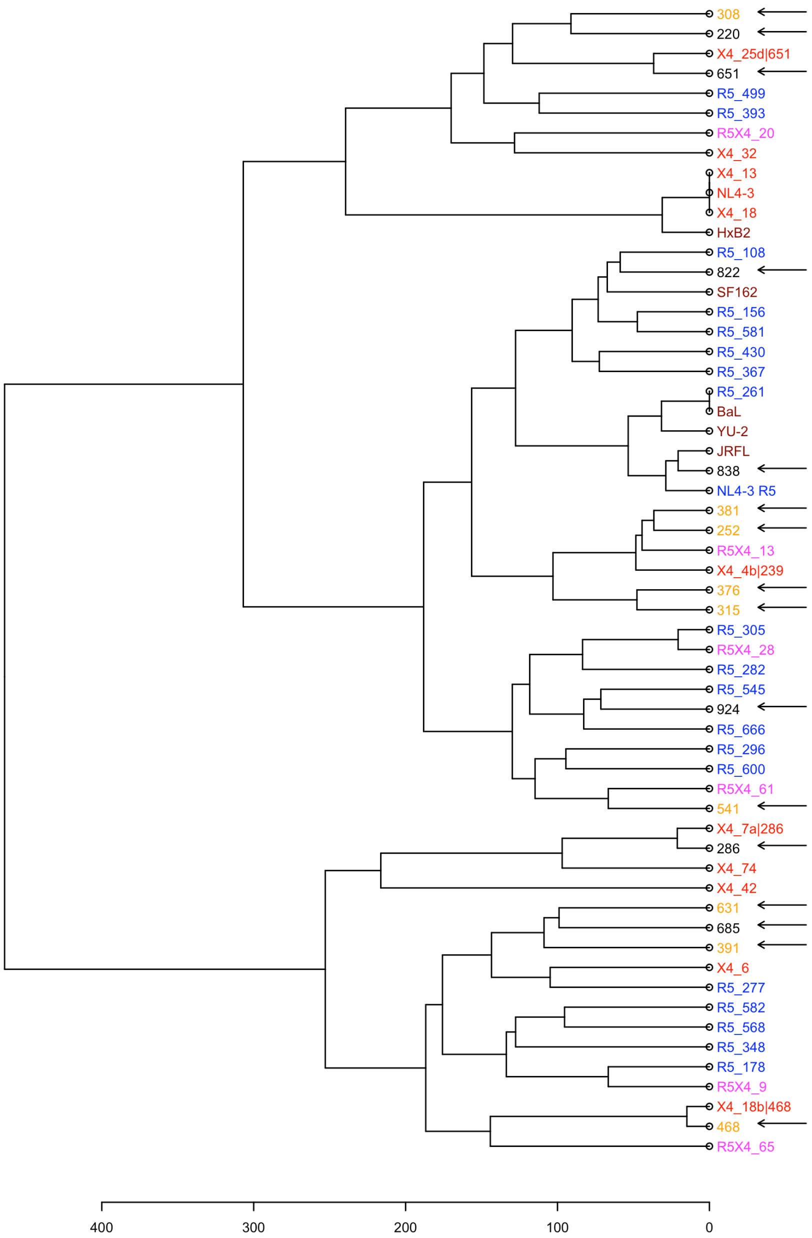


Figure 4.2 Dendrogram of the tested clones and several V3 sequences hierarchically clustered as described in the text. R5 sequences are colored in blue, X4 in red, dual-tropic in magenta. Clones are indicated with arrows. Those tested in the first set of experiments are colored in black, those tested in the second set of experiments in orange, lab-adapted clones in dark red.

4.3 Experimental assays

In the Heidelberg lab two experimental assays for high-throughput data measurement at the single cell level were established – based on *fluorescence microscopy* and on *quantitative flow cytometry*. In each of the assays a different cell line and different cellular markers of virus cell entry were used. Due to the higher biological relevance of the cell line used in quantitative flow cytometry assay and to its higher measurement capacity, the second round of experiments and the modeling were based on this assay only. Fluorescence microscopy was used in the initial set of experiments for the purpose of additional validation and for characterization of the tested clones.

4.3.1 Fluorescence microscopy

In the fluorescence microscopy assay a lab-derived cell line was used called Affinofile cells (Johnston, Lobritz et al. 2009). This cell line offers the possibility of artificially inducing CD4 and CCR5 expression to different levels and independently of each other (Figure 4.3) and thus for obtaining broader ranges of CD4 and CCR5 expression levels for further modeling.

Surface expression of CD4 and CCR5 in the Affinofile cell line is induced by Tetracycline (Tet) and Ponasterone A (Pon), respectively. Amounts of Pon in concentrations from 0 to 1µM, induce the expression of CCR5 in Affinofile cells in a range from ~600 to ~7000 molecules/cell. Higher concentration of Pon leads to saturation of the CCR5 expression. Concentrations of Tet between 0 and 10ng/ml show a linear dependence with the CD4 expression ranging from ~1200 to ~37000 molecules/cell (Table 4.2). CXCR4 is expressed at low levels in this cell lines and cannot be artificially induced.

Expression of CCR5 and CD4 in this cell line was measured using immunostaining. Mouse antibodies raised against CCR5 and CD4 were bound to molecules emitting light in different spectra: Alexa 568 for CCR5 and Alexa 647 for CD4. Emission of light by the coupled antibodies reflects the surface expression of the corresponding protein. Another stain Hoechst was used to mark cell nuclei and derive cell positions on microscope images.

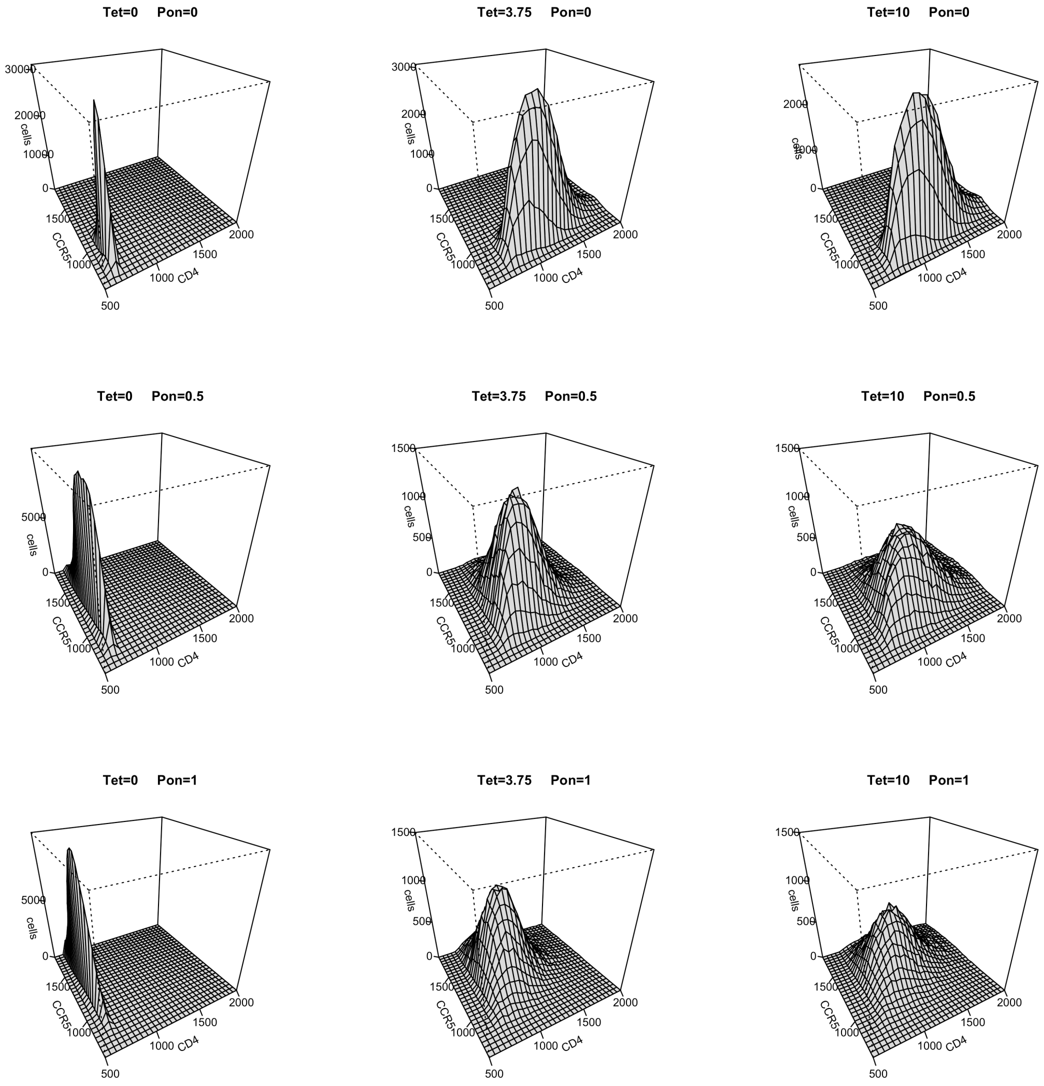


Figure 4.3 Induction of expression of CD4 and CCR5 in the Affinofile cells. 2D density plots of cell surface expression of CD4 and CCR5 are shown for three different concentrations of Pon and Tet and their combinations. Amounts of Pon are given in µM, of Tet in ng.

Green fluorescent protein (GFP) was used as a marker for measuring the presence of virus in a cell in this cell line. Into each viral particle a self inactivating (SIN) vector encoding GFP (Zufferey, Dull et al. 1998) was incorporated and viral entry could be detected by the green fluorescence of target cells after the virus transduction.

Cells stained for CD4 and CCR5 expression were seeded on a plate composed of 96 wells. Each well was filled with the required concentrations of Pon and Tet and then transduced with GFP-labeled viruses with varying V3 sequences. Next, cell data within each well was measured using automated fluorescence microscopy. Data acquisition and preprocessing was done in the lab of Fred Hamprecht, Heidelberg Collaboratory for Image Processing (HCI), Interdisciplinary Center for Scientific Computing (IWR), University of Heidelberg. The high-throughput image was obtained using an epifluorescence Scan^R screening microscope equipped with the Scan^R acquisition software (Olympus Biosystems GmbH, Münster, Germany). The surface of each well was covered by 16 images in four different channels filtering for stained nuclei, infected cells, CCR5 and CD4 immunostains. Microscope images were processed in two steps. First, cells were segmented based on the cell nuclei and antibody stains (Borner, Hermle et al. 2010). Then, within each segment the strength of CD4 and CCR5 stains and GFP was quantified (Figure 4.4).

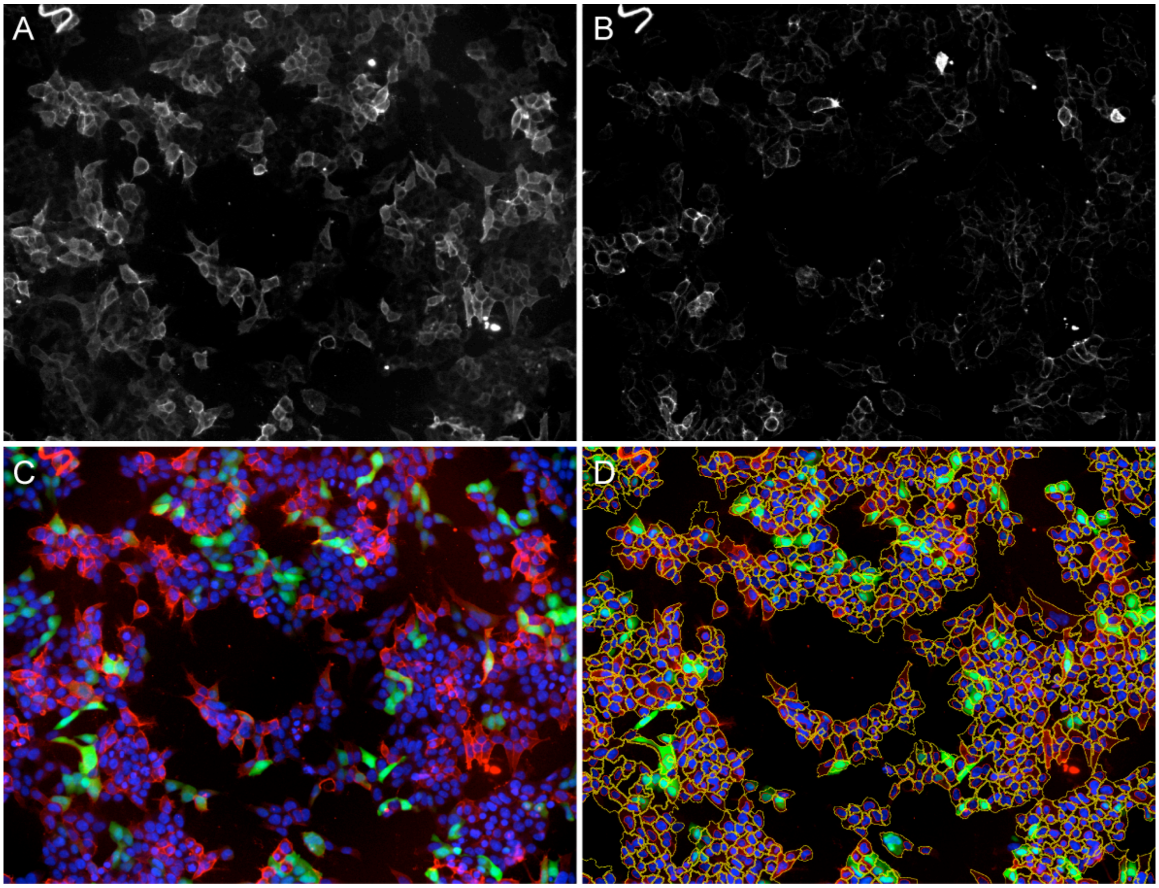


Figure 4.4 Intermediate steps of the microscope image analysis. A and B show the images of CCR5 and CD4 expression respectively. C shows an overlay of pictures taken in the channels for the nuclei stain (blue), GFP (green) and the maximum of CD4 and CCR5 stains (red). D shows the segmentation of the image into cells (yellow lines) based on the respective staining.

On average ~800 cells were measured on each image. As a result of the image analysis, the levels of CD4 and CCR5 expression and the virus presence in a cell were quantified for each cell based on the signal strength in the respective channel.

4.3.2 Quantitative flow cytometry

In the flow cytometry assay human T cells were used – C8166 and C8166-R5 for development of data processing methods and SupT1-R5 cells for the quantitative measurements. Both cell types originate from the human immune system, express receptors and coreceptors used by HIV in sufficient amounts (Table 4.2) and are therefore appropriate for HIV cell entry testing.

cell line	CD4	CCR5	CXCR4
Affinofile (no induction)	1 241.7	547.1	950.8
Affinofile (1µM PonA)	1 151.3	7 298.3	1 128.8
Affinofile 10ng/ml Tet	36 722.6	577.2	1 056.6
SupT1-R5	38 212.7	2 323.1	22 411.6
C1866	195 278.6	318.6	10 833.5
C1866-R5	23 112.5	1 761.9	22 883.6

Table 4.2 Surface expression of measured receptor and coreceptors on the cell lines tested. Values are displayed in molecules per cell.

As a marker of virus presence in a cell in this data acquisition system a virion fusion assay termed β-Lactamase (BlaM) assay was used (Cavrois, De Noronha et al. 2002). This assay is well established in the HIV research and is based on the bacterial β-Lactamase enzyme which is incorporated into the virus and delivered to the target cell after virus and cellular membranes fusion. Target cells are loaded with a cleavable fluorescent substrate (CCF2) that emits green light in its uncleaved state and blue light once it has been cleaved with β-Lactamase. Emission of blue light in a cell is a marker of a cell being infected.

Cells exposed to BlaM-carrying virus were additionally stained with labeled antibodies attaching to CD4 receptor, CCR5 and CXCR4 coreceptors. This way treated cells were next scanned using the fluorescence activated cell sorting (FACS) machine. In the machine, cells are suspended in a liquid that separates them from each other and are pumped through a flow chamber at a speed of ~500 cells per second (Figure 4.5). Laser beams hit the cells as they pass through the flow chamber and their reflection angle provides the information on the physical characteristics of each cell. Light reflected at small angles is termed *forward scatter* (FSC), light reflected in other directions is termed *side scatter* (SSC). Forward scatter provides a measure of cell size, side scatter of its granularity. Different cell types in a cell population can be characterized in terms of separate clusters of unique combinations of these two parameters. In addition to the physical characteristics of the cells, their biological and chemical properties can be measured by color detectors detecting light emitted by biological markers attached to or incorporated into each cell. In our assay, β-Lactamase, each receptor and coreceptor-specific antibodies emit light of a different color. Therefore, the intensity of the light signal at the respective frequency quantifies the cell surface expression of CD4, CCR5 and CXCR4 as well as the virus presence in a cell.

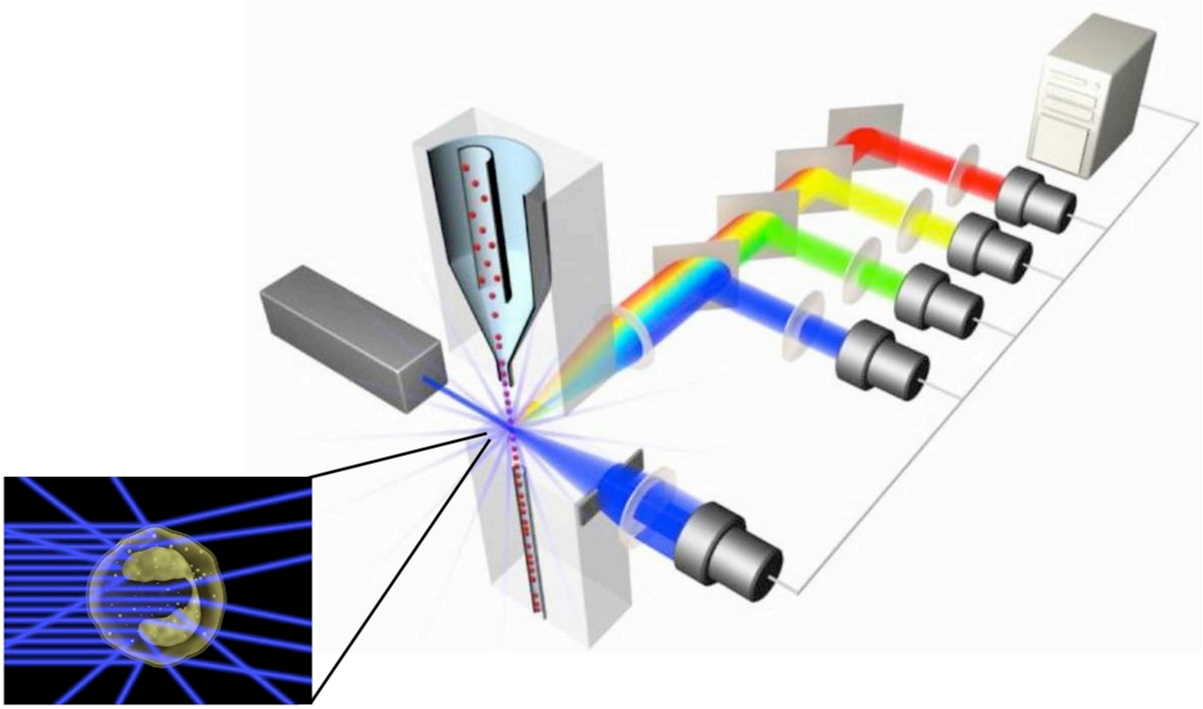


Figure 4.5 Schematic depiction of the FACS machine. Beam scatter provides information on the physical properties of each cell (inserted image, bottom-left). Light detectors allow to measure light intensity of light in different spectra emitted by the cells. (From http://en.wikipedia.org/wiki/Fluorescence-activated_cell_sorting).

FACS output files contain information on the forward and side scatter, green and blue light emission of the BlaM assay, CD4, CCR5 and CXCR4 staining signal for each of the cells scanned in a high-throughput manner. Numbers of cells captured in individual FACS machine measurements are listed in Table 4.4 in section 4.5.2.

4.3.3 Comparison of experimental assays

Overall, both fluorescence microscopy and flow cytometry offer efficient methods of acquiring high-throughput data at the single cell level. Technical limitations of the fluorescence microscopy include the lack of an additional light channel to measure the expression of CXCR4 coreceptor and the low expression level of this coreceptor in Affinofile cells. The possibility of inducing higher levels of CD4 and CCR5 expression presents an efficient way to extend the range of measured parameter values that allows for construction of more general models. However artificially high levels of surface CD4 and CCR5 expression do not occur in vivo, moreover, the lack of CXCR4 expression on the Affinofile cells limits the relevance of this assay for the true environmental settings of an HIV infection.

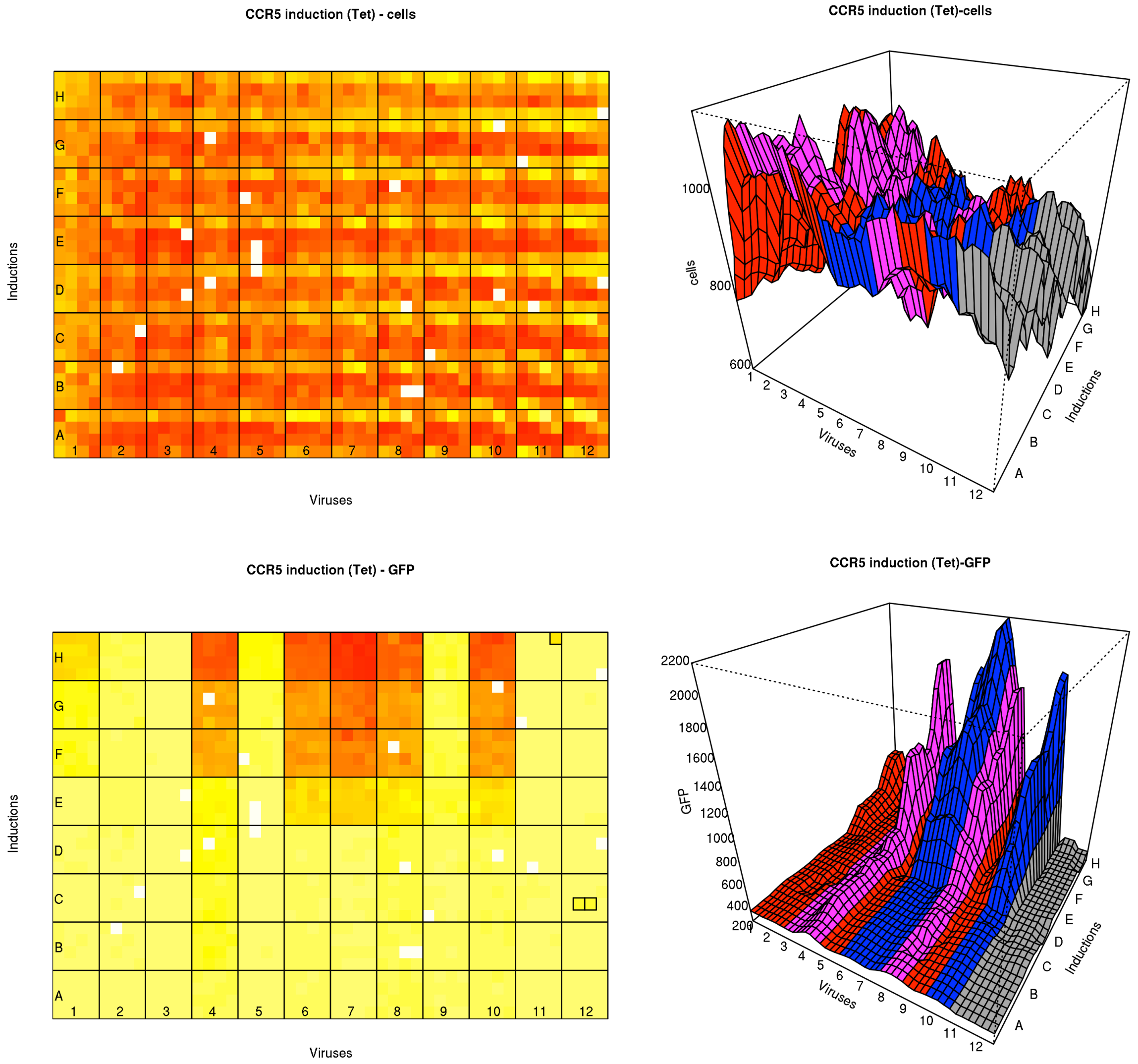
Cells of the T cell family used in the fluorescence microscopy assay are not adapted for capturing in the microscope and do not offer the possibility of enhancing receptor and coreceptor expression. However natural variability of the receptor and coreceptors expression, sufficient levels of CXCR4 expression and the immune system origin of these cells present more physiologically relevant conditions for testing of the HIV cell entry. In addition, FACS machine offers a sufficient number of measurable light spectra to capture all parameters that are relevant for this study. It is not charged with image acquisition and processing errors and offers higher throughput measurements at a lower cost.

4.4 Data preprocessing

In order to analyze data at the single cell level we developed methods that allow for automated processing of data measured in both high-throughput single cell data acquisition systems. Instead of a classical approach of averaging measurements of entire cell populations (Johnston, Lobritz et al. 2009), we developed methods that allow for accessing the data at the single cell level and using their natural variability for generation of models of HIV cell entry. Models based on single-cell level data offer higher accuracy by accounting for the variation of individual cells and this way allow for deriving more detailed information based on individual experiments as compared to the averaging approach.

4.4.1 Fluorescence microscopy

As a first step of the analysis of fluorescence microscopy data we inspected the distribution of cells, of the GFP and other light signals on the plates. Figure 4.6 shows the distribution of cells (upper panels) and the GFP signal (bottom panels) on a microscope plate. The wells of the two rightmost columns on the plate contain control measurements: in column 11 a virus with no envelope gene (Env(-)) is titrated, column 12 contains cells without any virus (no virus). GFP signal in those two columns is expected to be at a constant background level. Microscope pictures that show a higher, above the background, level of GFP might therefore contain errors that change the average level of the signal of the controls. Since the control measurements are used to establish a cutoff for classification of cells into infected and not infected, control images with unexpectedly high signal levels were removed before further classification as they might introduce an overestimation of the background entry efficiency levels.


CCR5 induction (Tet) - cells

CCR5 induction (Tet)-cells

CCR5 induction (Tet) - GFP

CCR5 induction (Tet)-GFP

Figure 4.6 Example distributions of number of cells (upper panels) and the GFP signal (bottom panels) on a microscope plate. Colors of the heat maps indicate the number of cells (upper panels) and the strength of the GFP signal (bottom panels) with red representing higher and yellow lower values. White squares indicate missing images. Plots on the right hand side represent 3D visualizations of the corresponding heat maps on the left hand side and are colored according to the type of virus titrated in each column of the plate with red representing X4 viruses, blue R5 viruses, magenta dubious viruses and gray control measurements – Env(-) and no virus. In each column of the plate a different clone was tested, the following rows of this plate were titrated with an increasing amount of Tet which resulted in increased entry efficiency of R5 viruses.

In addition to the noise in the control measurements certain plates showed an uneven distribution of cells among wells. On the plate shown in Figure 4.7, row G contains up to twice as many cells as row C. On the same plate a negative correlation between the number of cells in a well and the strength of the staining signal could be observed (Figure 4.6, bottom panels). Rows containing lower numbers of cells showed an increased signal of CCR5 expression at a constant Tet induction and the same amount of the antibody titration on the entire plate. This negative correlation might be due to the higher amounts of antibody per cell in the wells where the number of cells is lower. However this hypothesis was not tested experimentally.

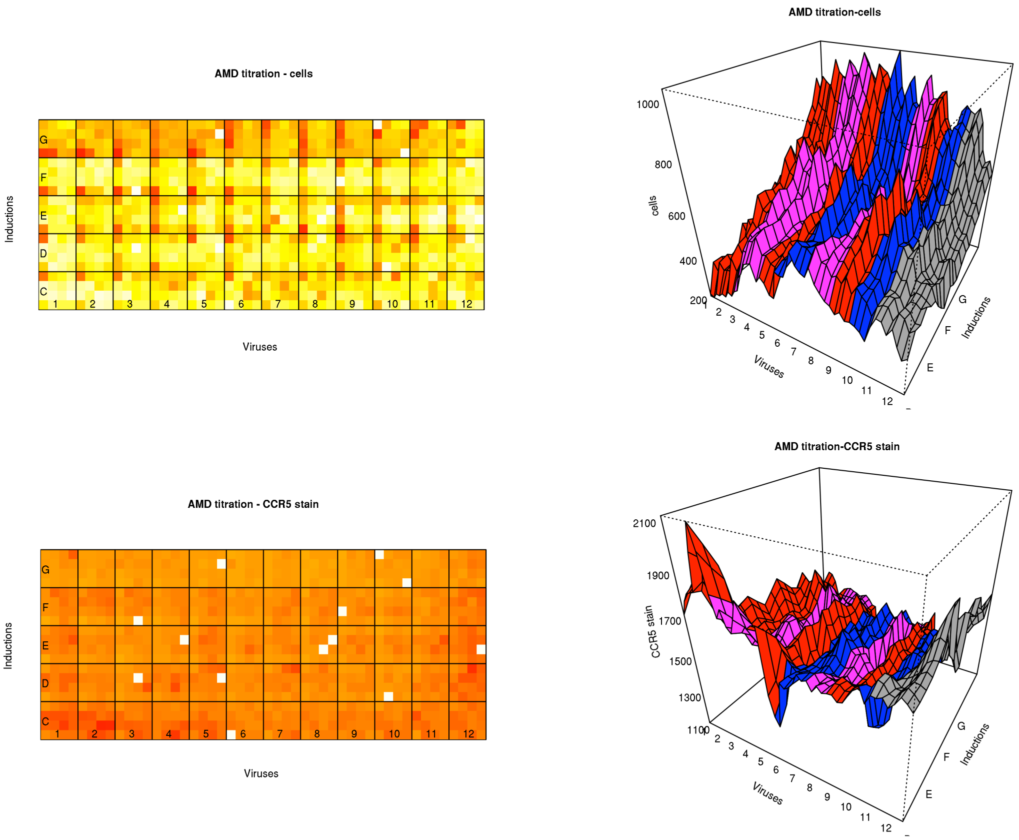


Figure 4.7 Example distribution of cells (upper panels) and the corresponding CCR5 stain signal (bottom panels). The coloring scheme is in accordance with Figure 4.5. Lower amounts of cells in the bottom rows of this plate correspond to higher signal of the CCR5 expression.

With the lack of experimental measurements allowing for correction of the biased stain signal values, we introduced a computational correction procedure. In this procedure we fitted a linear function expressing the dependence of the average stain value on the number of cells in each well of a plate. Next, individual values of a stain intensity y of the cells were corrected based on the fitted function and the number of cells x_{well} in the corresponding well of a given cell according to the formula:

$$\overline{y} = y - \left(a \cdot x_{well} + b - y_{mid}\right)$$

where a is the parameter of the fitted linear function and b its offset and y_{mid} is the middle value of the stain value of all the wells on a given plate. Intuitively, the stain values in wells are pivoted around a central stain value by an angle dictated by the strength of correlation. This operation produces uncorrelated values of stain and cell numbers (Figure 4.8).

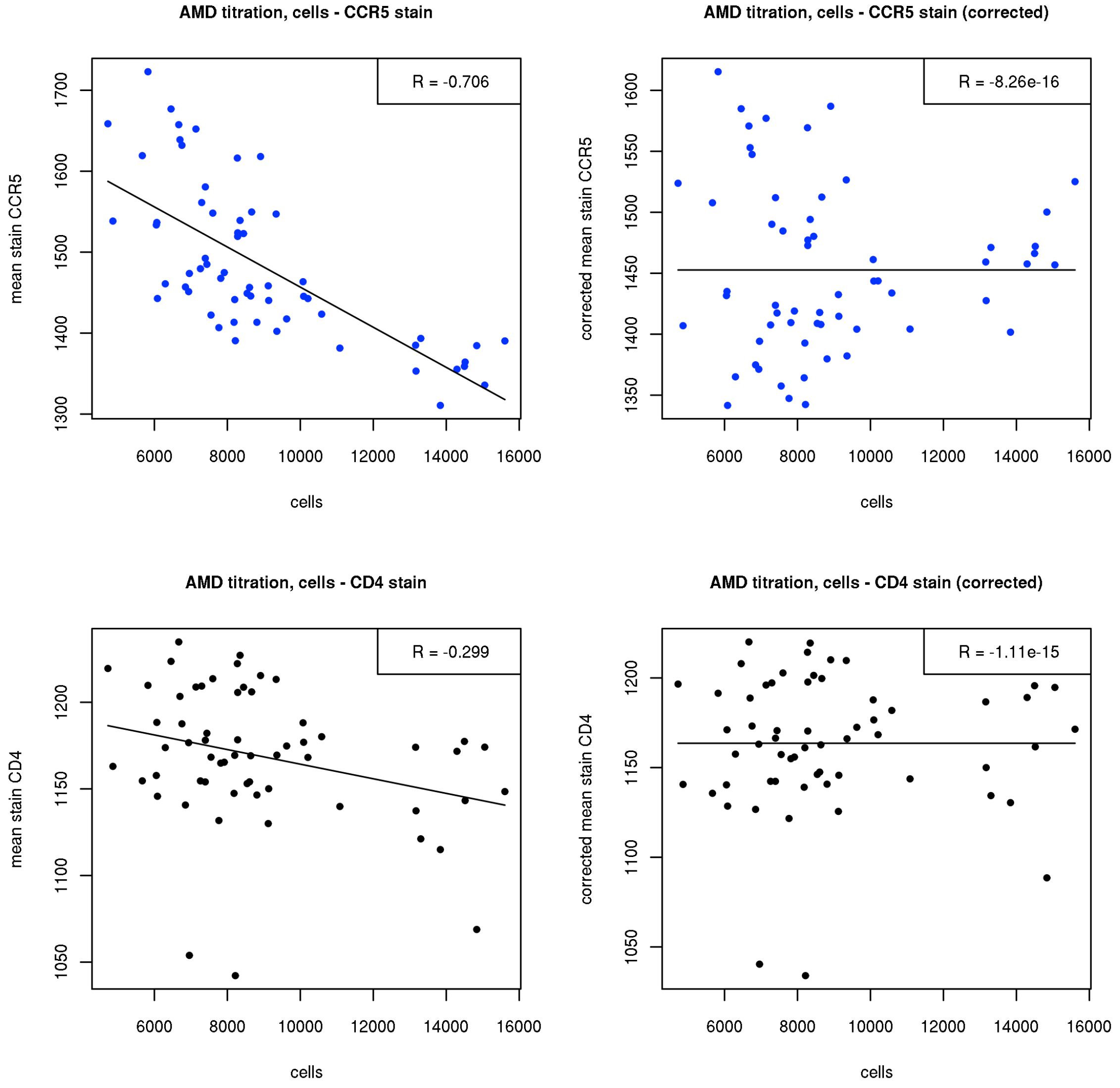


Figure 4.8 Negative correlation of the number of cells in a well with the CCR5 (upper panels) and CD4 stain (bottom panels) on the plate shown in Figure 4.6. The correlation is indicated in the upper right corner of each diagram. The black line represents a linear function derived based on the least-squares fit. The right-hand panels show the corresponding data after the correction.

After the two correction steps – removal of flawed control images and correction of stain signal, cells were classified into infected and not infected based on the GFP signal intensity. Classification cutoff was established for each plate based on the distribution of the GFP signal in the control measurements of Env(-) and no virus. First, gamma distribution was fitted to the GFP values of all cells in the control measurement. Gamma distribution offers the largest variation of shapes and is therefore appropriate for fitting to data with unknown distribution. Next, a cutoff was chosen at the value which is exceeded by a fraction of 0.0001 of the distribution (Figure 4.9). The cutoff was established based on control measurements (Env(-) and no virus) on each plate, cells in other parts of the same plate were classified accordingly.

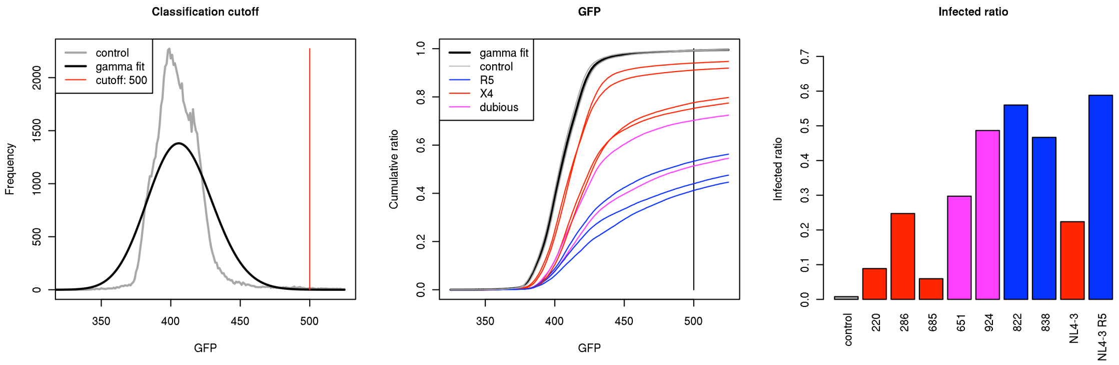


Figure 4.9 Classification of cells into infected and not infected based on the GFP signal. Gamma distribution is fitted to the control measurements, the cutoff is defined at the 0.0001 upper fraction of the fitted distribution (left panel). Middle panel shows the cumulative distribution of the GFP signals in cells titrated with each of the tested clones. The curves are colored according to the clone phenotype as indicated in the legend. Proportion of cells infected with each of the clones according to the established cutoff is shown on the right panel.

Classification of cells allowed for assessment of the experimental setup and preliminary inspection of virus entry efficiencies (Figure 4.10). Specifically, Pon and Tet induction showed a pronounced effect on the entry efficiency of the R5 viruses and less of the X4 viruses (Figure 4.10, upper panels). X4 viruses showed a generally lower entry efficiency in this cell type possibly due to the low levels of CXCR4 expression.

The analysis of the fluorescence microscopy data at the single cell level offered a detailed picture of the dependencies of virus entry efficiency (Figure 4.10, bottom panels). The discussion of the results follows in section 4.5.

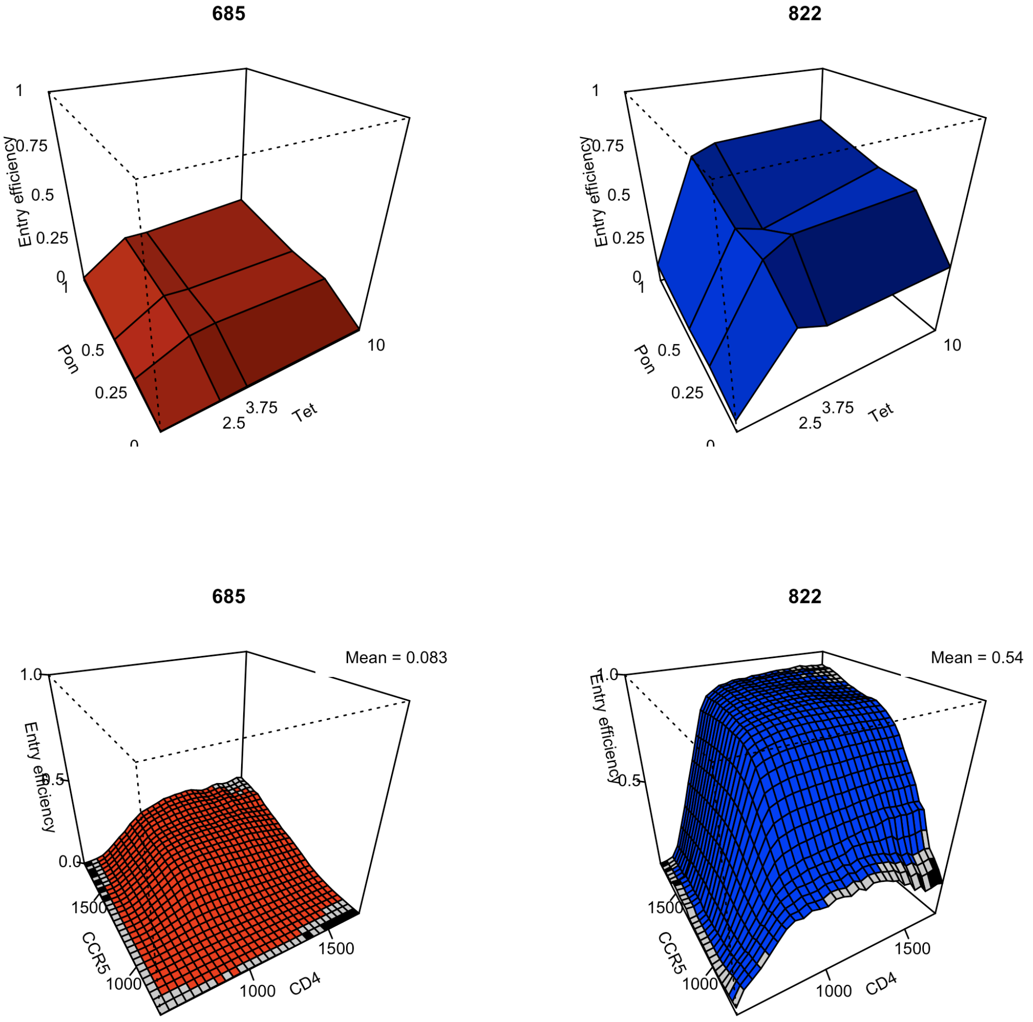


Figure 4.10 Entry efficiency of two example clones 685 and 822 tested in the fluorescence microscopy assay. Upper panels show an increase in the entry efficiency of the clones with the induction of CCR5 and CD4 expression using Pon and Tet, respectively. Entry efficiency here is represented by an averaged ratio of infected cells for each Pon and Tet induction level. Bottom panels show the same experimental data with the clone entry efficiency plotted as a function of the CCR5 and CD4 stain signal. These plots were generated based on the single cell data, the ratio of infected cells was averaged within 30 ranges of stain values spanning over the full range of stain values obtained in this experiment. The color scheme of the plots are detailed in section 4.4.3.

4.4.2 Flow cytometry

Similar to the fluorescence microscopy, single cell data measured with the flow cytometry assay was filtered and classified before further analysis. As the initial step of data preprocessing, cell populations were filtered according to the FSC and SSC parameters, an operation termed *gating*. The distribution of the combination of these two parameter values was analyzed in order to identify the major cell population and filter out cells not belonging to the major population as they might be defunct or be of a different cell type. A classical way to perform gating consists of manually defining the *gate*, i.e. the contour surrounding the major population based on subjective visual inspection (Figure 4.11). Although this approach is predominant in the FACS community, it is not an objective or systematic procedure. Therefore the results of manual gating can be questionable and hard to reproduce.

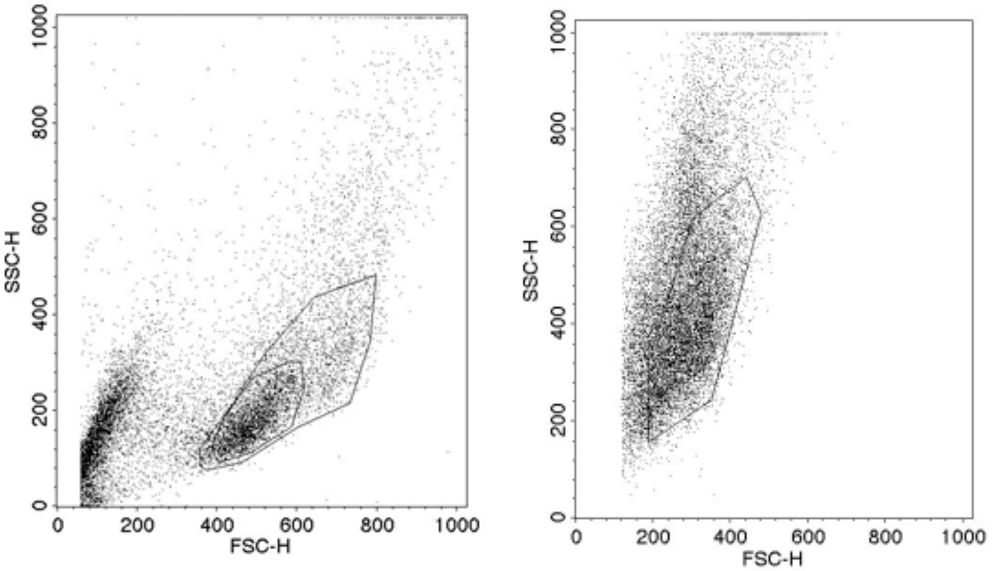


Figure 4.11 Examples of manually gated cell populations.

In order to address the shortcomings of manual gating, we developed a computational method for gating of cell populations as follows. Cell populations are commonly plotted on 2D scatter plots with FSC on the x-axis and SSC on the y-axis. In the proposed automated gating procedure cells are first filtered through a user-defined square window defined on the FSC and SSC values. In the case of SupT1-R5 cells we used a filter of 300 < FCS < 1200 and 0 < SSC < 900. Next, a grid of FSC and SSC values is defined and the numbers of cells in each bin of this two-dimensional grid are calculated, which can be presented in form of a heat map (Figure 4.12). Each bin of the grid is of width and height 5, a value that was chosen among several others based on its optimal gating agreement with the manual method. Next, for following values x_i on the x-axis (FSC) a grid bin is found with position that contains the maximum number of cells among all bins at the given x_i. For each such bin position (x_i, y_i) minimum distance d_i is found such that no cells are found in the bins at equidistant positions $(x_i, y_i - d_i)$ or $(x_i, y_i + d_i)$. The positions $(x_i, y_{i,1}) = (x_i, y_i - d_i)$ and $(x_i, y_{i,2}) = (x_i, y_i + d_i)$ are termed surrounding points. The y-coordinates of the surrounding points are next averaged: $y_{i,k}$ is replaced by a mean value of itself and two neighboring surrounding points $\frac{1}{3}\sum_{j=i-1}^{j=i+1} y_{j,k}$, where $k \in \{1,2\}$.

This averaging results in a smooth gating line. The same search for surrounding points and averaging procedure is repeated along the y-axis. A gate is defined as the minimal contiguous area on the FSC-SSC grid encircled by the line connecting surrounding points. Example results of the automated gating procedure are shown in Figure 4.12. The no virus measurement was used as the control for establishing the gate. All measurements in the same experiment were gated accordingly. For reading the FACS files R package *prada* was used, a part of Bioconductor (Gentleman, Carey et al. 2004).

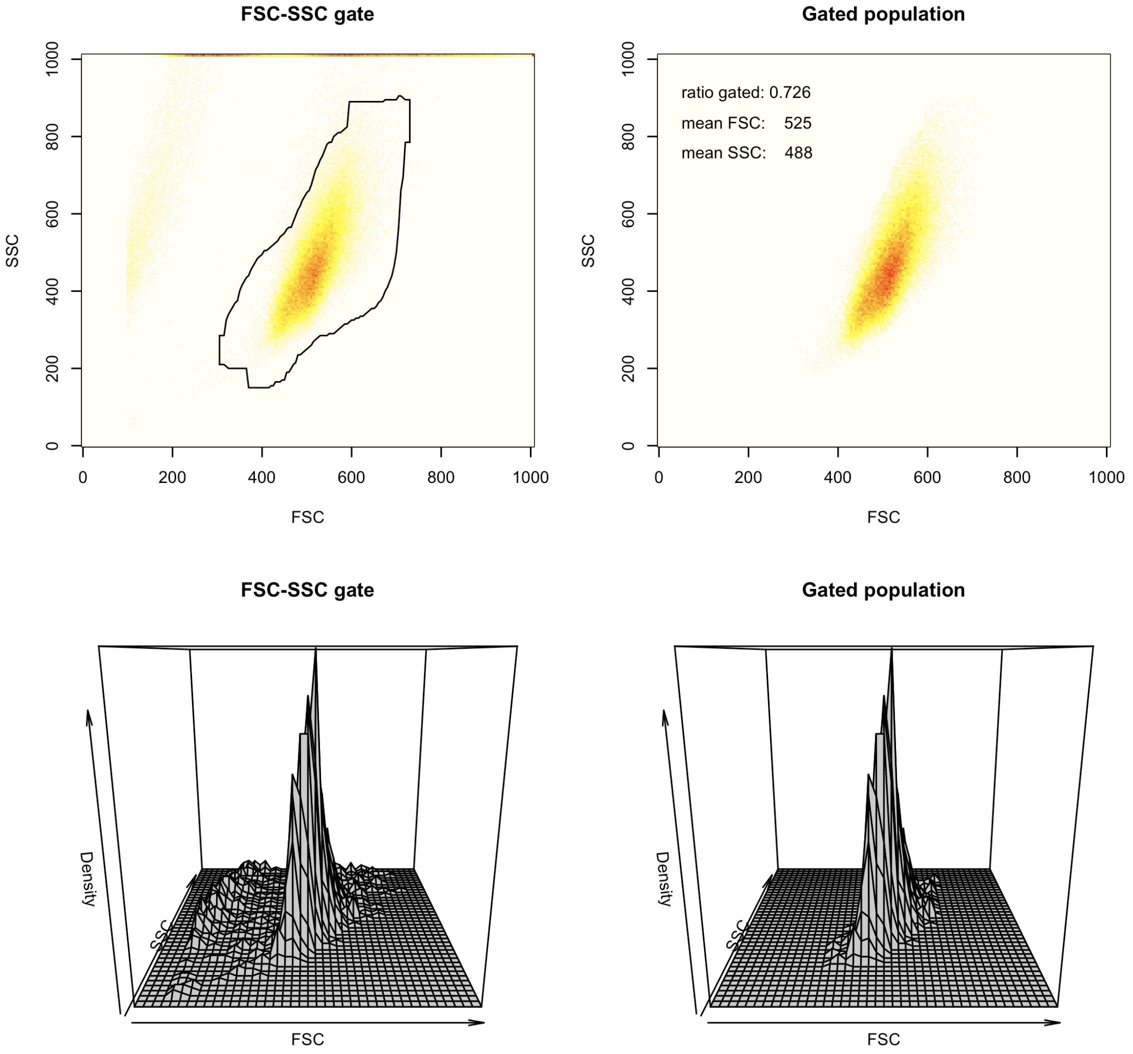


Figure 4.12 Automated gating of FACS data. Top-left panel shows a heat map of the cell density showing specific FCS and SSC values before gating. Calculated gate is traced with a black line. Heat map of the gated cell population is shown on the top-right panel. Bottom panels show 3D views corresponding to the heat maps above. Cell population with a small FSC and high SSC values as compared to the major population is removed in this example.

After gating of the cell populations, cells were classified into infected and not infected based on the BlaM marker. Each cell in the FACS assay carries a green dye that turns into blue after cleavage with the β-Lactamase enzyme incorporated into the virus. A shift from green to blue light is therefore indicative of a cell being infected. Similar to gating, a common approach to classification of the BlaM-stained cells entails a manually established decision boundary based on the control measurement not containing the virus. The decision boundary is defined on the plot of blue against green light signal of the cells in the control measurement. It delineates the region of uninfected as a minimal region cells such that $y > x$ and a proportion of ~0.01% of the control cells are located outside of this region. Measurements of cells titrated with viruses are classified according to such a control-based decision boundary.

In order to efficiently classify a large number of FACS measurements and to obtain objective and reproducible results we established an automated method of BlaM assay classification. The steps of the classification procedure are illustrated in Figure 4.13. Cells showing specific values of the green and blue light are depicted as points in a scatter plot with blue light intensities on the x-axis and green on the y-axis (Figure 4.13, panel 6). A corresponding heatmap representation (Figure 4.13, panel 1) illustrates the density of thousands of cells that appear as overlapping dots on the scatter plot. First, a linear function $y = ax + b$ is fitted to the points representing the cells of two control measurements – not containing the virus and unstained (Figure 4.13, panel 1). For each point on the fitted line (x_i, y_i), where $y_i = ax_i + b$ and $x_i = 1..1200$, the data points (x_{p_i}, y_{p_i}) on the scatter plot are found lying on the line perpendicular to the fitted line and intersecting it at (x_i, y_i): $y = -\frac{1}{a}x + b_i$, where $b_i = y_i + \frac{1}{a}x_i$, such that are the most distant from the point (x_i, y_i) and $y_{p_i} < x_{p_i}$ (Figure 4.13, panel 2). These points represent cells showing the highest shift in the blue light relative to the green light in the control. The distances of these points from the fitted line (Figure 4.13, panel 3) are next smoothed in a sliding window approach by averaging the values within each window and adding of one standard deviation of the values within that window (Figure 4.13, panel 4). The added standard deviation produces a margin beyond the control cells that ensures a required low proportion of false positives in the control measurement (~0.01%). Window size of 30 was selected as the size resulting in the best performance. The smoothed points projected back onto the plot of green and blue light signals define the cutoff decision line (Figure 4.13, panel 5) – cells represented by data points located in the part of the plot below the decision line are classified as infected, those located above the line are classified as not infected. The method design and the choice of parameters were dictated by the comparison with manual classification and the best performance on a large number of measurements. The classification based on this procedure assigns a binary value to each cell, 0 representing not infected and 1 infected cells and was termed *binary classification*.

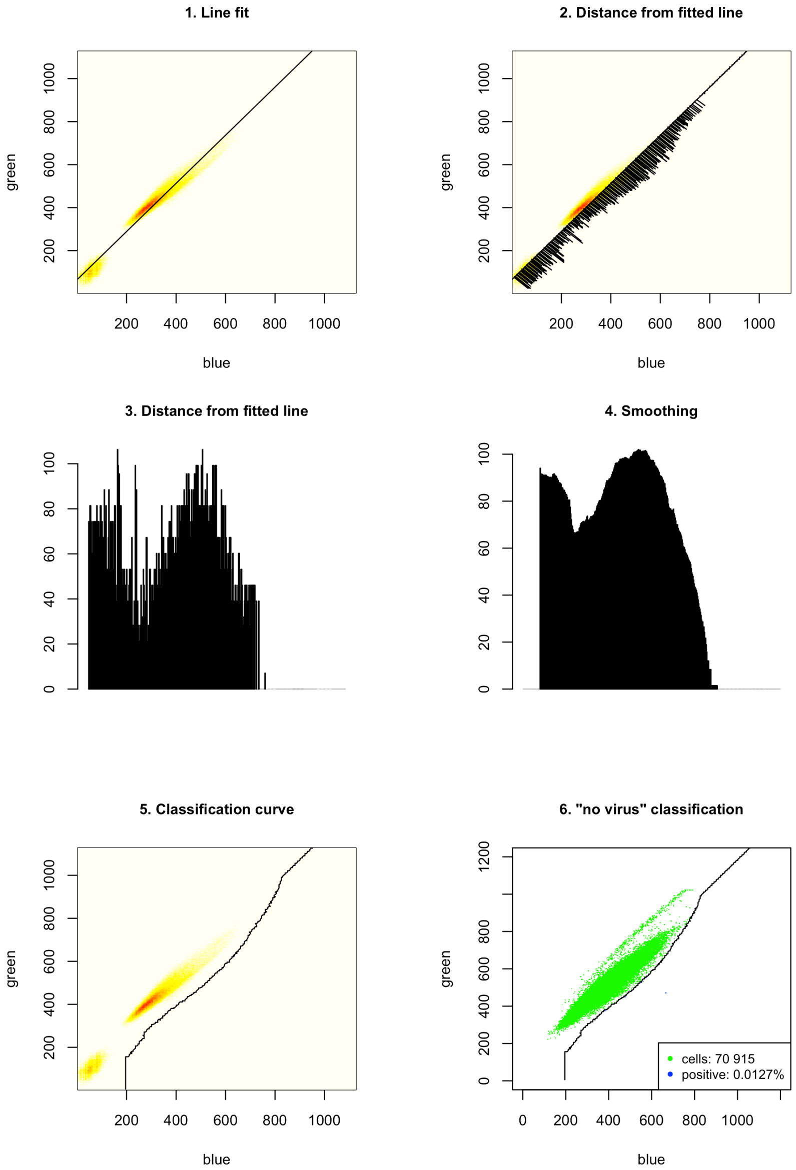


Figure 4.13 Following steps of the classification method for the BlaM assay. (1) Heat map of the density of cells of the control measurement showing given combinations of blue and green light signals. The line represents the linear function fitted to the values of the blue and green light signal. (2) Most distant points from the fitted line are found showing the highest blue-to-green light ratio. (3) Distances of these points are measured and smoothed (4) in a sliding window approach. (5) Smoothed points mapped back onto the blue-green light heatmap define the classification curve. (6) Plot of blue and green light values of the cells of the control measurement with no virus. The classification curve estimated in the previous steps is shown in black, the number of cells and the percentage of control cells outside the classification curve are indicated in the legend.

Two examples of classification of virus measurements using the classification curve depicted in Figure 4.13 are shown in Figure 4.14.

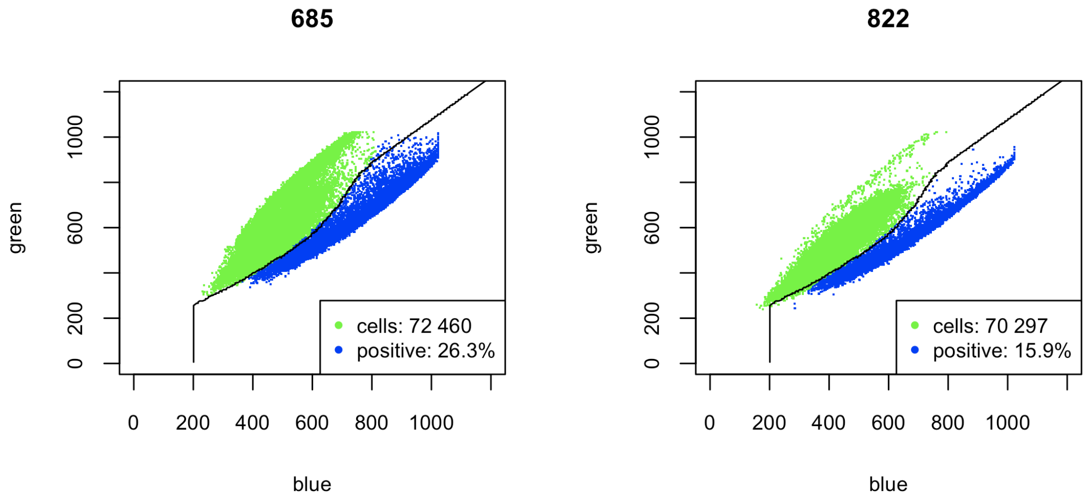


Figure 4.14 Classification of two virus measurements based on the classification curve depicted in Figure 4.13. Cells classified as infected are represented by blue dots, uninfected cells by green dots. Percentage of the infected cells (positive) is indicated in the legend.

Comparison of the automated method with manual classification indicated a lower agreement between the manual and the automatic method for samples with high cell concentration close to the decision line. In order to account for potential errors of the automated method for these cells, we developed a second classification method that instead of a cutoff curve calculates a region delimited by two parallel margins surrounding the decision curve of the binary method (Figure 4.15). Cells located within the region delineated by these two curves are assigned a value between 0 (not infected) and 1 (infected) that estimates the probability that a given cell is infected. The probability estimate is calculated as proportional to the distance of a cell from the upper boundary of the region (lower blue-to-green ratio) with probability 0 at this boundary and 1 at the bottom boundary (higher blue-to-green ratio). We selected 5 as the optimal width on the region that compensates the difference in the classification results between the binary automated method and manual classification. This approach was termed *margin classification*.

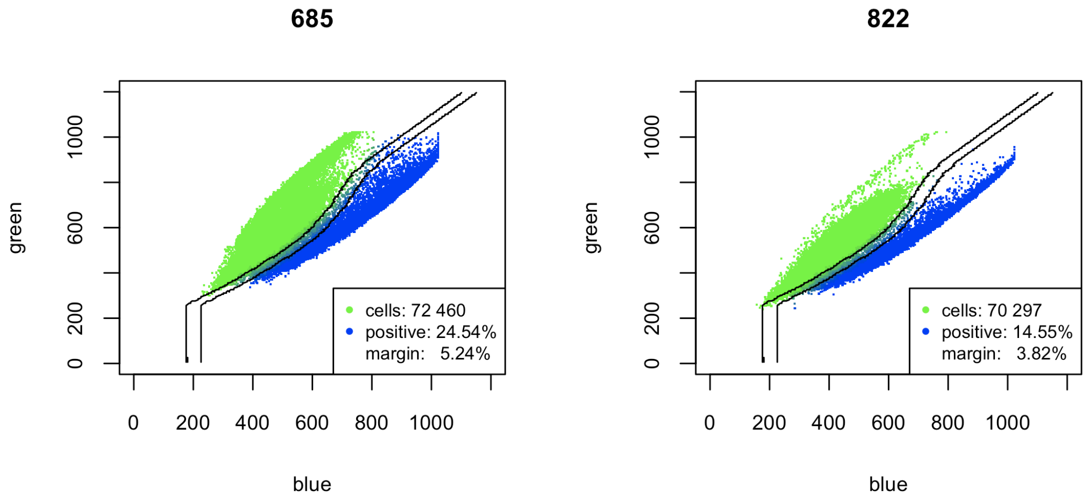


Figure 4.15 Classification of two example viruses shown also in Figure 4.14, using the margin method. Two black lines trace the margin between not infected (green) and infected (blue) cells. Cells located within the margin are colored according to the assigned probability with a green shade representing values closer to 0 and blue closer to 1.

4.4.3 Visualization methods

As the next step of the study we developed visualization methods that allowed for preliminary inspection of the tested clones with respect to their overall entry efficiency in dependence on the measured parameters. This initial inspection was informative for further model development. Throughout the next section we visualize the experimental results using 3D plots of virus entry efficiency in relation to two chosen parameters. Colors of the plots are dictated by virus phenotype – red for X4 viruses, blue for R5, magenta for dubious. Individual cells of a microscope or FACS measurement are localized in a 30x30 grid of values spanning the range of values of the two parameters. Virus entry efficiency is calculated as the mean classification value (assigned using GFP-based, binary or margin classification) of cells within each bin of the grid. Prior to plotting, the grid is smoothed by averaging values from neighboring bins of the grid. In order to account for the differing numbers of cells that show a given combination of parameter values, parts of the grid that contain less than a selected minimum number of cells are colored in gray. The selected minimum number of cells is 10% of the expected number of cells if the cells were evenly distributed over the grid.

4.5 Data analysis

4.5.1 Fluorescence microscopy

In the fluorescence microscopy assay four experiments were performed (Table 4.3). As an experiment we define a set of measurements in which certain parameter or combination of parameters is varied. Table 4.3 shows the average number of cells measured per each parameter value combination in the performed experiments. This number is both a measure of how high-throughput the assay is and an indicator of how much information is processed in the analysis at the single cell level. In the fluorescence microscopy assay only the first set of 8 selected clones was tested. The results were used for the clone characterization, however due to the limitations of this assay (described in subsection 4.3.3), the modeling and following experiments of this study were done using only the flow cytometry assay.

experiment	plates	parameters varied	parameter values	cells per measurement
Pon-Tet	2	Pon, Tet	16	5 379
C46	1	C46	2	30 223
AMD	1	AMD	2	47 835
MVC	2	MVC	5	41 607
AMD-MVC	8	AMD, MVC	16	16 543

Table 4.3 List of measurements performed using fluorescence microscopy. Parameters varied in each experiment are indicated in the "parameters varied" column. The column "cells per measurement" shows the number of cells captured for each clone and individual level of drug or Pon/Tet induction, which represents the efficiency of this experimental assay.

Example figures of the Pon-Tet experiment were shown in the previous section (Figure 4.10). Induced expression of CD4 and CCR5 increased the entry efficiency of the R5 viruses substantially to high levels close to 100%, levels that do not occur in vivo and do not provide sufficient output variability for modeling. Since the induction did not increase the entry efficiency of X4 viruses and R5 viruses showed sufficient efficiency at the natural expression levels of Afinofile cells, the induction was reduced to a low constant level of 2.5 ng/ml of Tet and 250 nM of Pon in the following experiments.

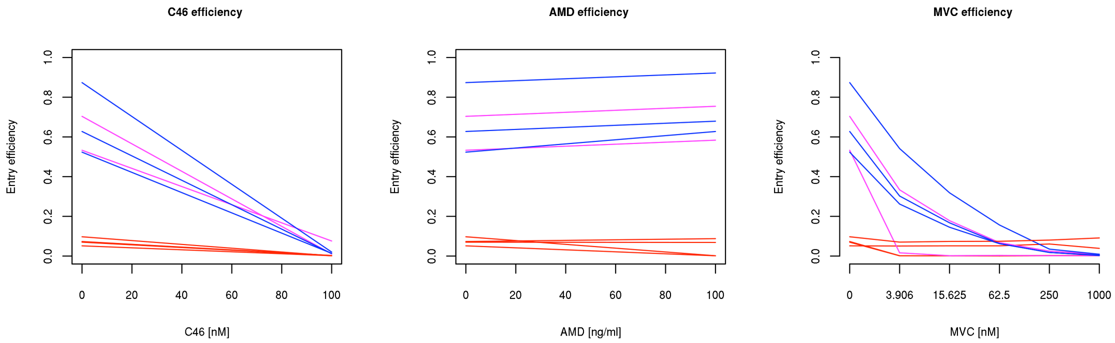


Figure 4.16 Effect of three types of entry inhibitors on the entry efficiency of the tested clones. Line colors correspond to the virus type: blue - R5 viruses, red - X4 and magenta - dubious.

Next, three experiments were performed testing three different entry inhibitors: fusion inhibitor C46 (Fletcher 2003), CXCR4 blocker AMD (De Clercq, Yamamoto et al. 1992) and CCR5 blocker MVC (Westby and van der Ryst 2005). Only a single high concentration of C46 and AMD was tested, MVC was tested at several levels of concentration (Figure 4.16). As expected, all viruses were blocked by C46. AMD restricted only two X4 viruses that showed low overall entry efficiency in Affinofile cells: the reference virus NL4-3 and clone 286. The remaining R5 and dubious clones were efficiently restricted by MVC (Figure 4.16, right panel) with R5 clones showing generally higher entry efficiency and a weaker response to the drug at low concentrations as compared to dubious clones.

Inspection of entry efficiency dependencies of different clones at the single cell level showed a strong relationship between the level of CCR5 expression and entry efficiency of all viruses blocked by MVC (Figure 4.17).

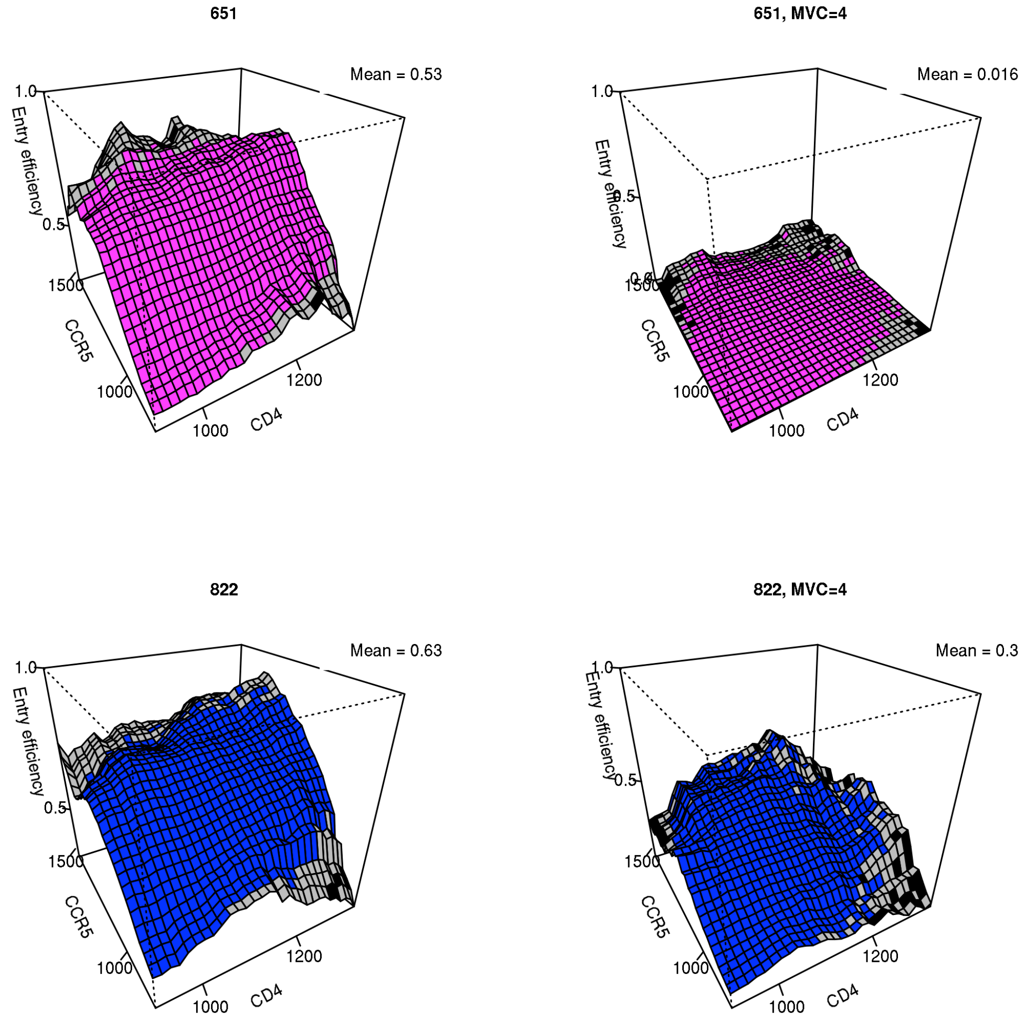


Figure 4.17 Entry efficiency of clones 651 (upper panels) and 822 (bottom panels) in relation to CD4 and CCR5 expression without drugs (left panels) and in the presence of MVC at a concentration of 4nM (right panels). Both clones show a positive relation of CCR5 expression to entry efficiency and a strong drug response.

The last experiment in this assay tested virus responses to combinations of MVC and AMD on different levels (Figure 4.18). As expected, based on the results of previous experiments, all except two viruses (NL4-3 and 286) showed a strong response uniquely to MVC.

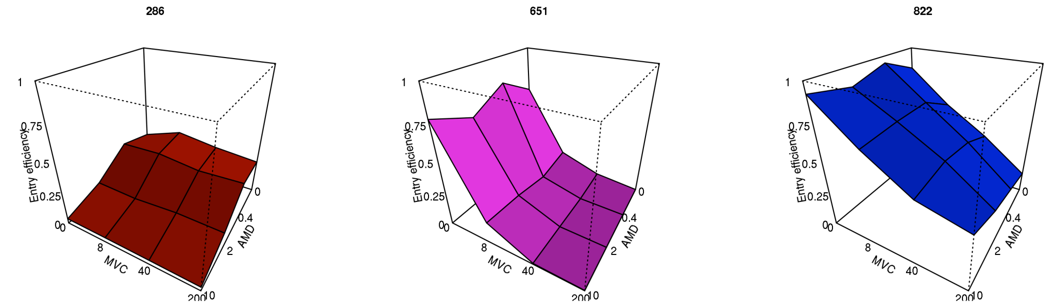


Figure 4.18 Response of clones 286 (left), 651 (middle) and 822 (right) to combination of different concentrations of MVC and AMD. AMD concentrations are in ng/ml, MVC in nM. Plots are colored according to the clone phenotype (see section 4.4.3).

Overall, these results indicate that only two of the tested viruses are purely X4. Clones selected as X4 viruses show overall low entry efficiency suggesting their dependence on the CXCR4 that is expressed only at low levels in Affinofile cells. Two of the X4 clones showed a decrease of the entry efficiency in the presence of CCR5 blocker suggesting their misclassification. R5 viruses showed high entry efficiency and a strong response to MVC. Dubious clones showed medium-level entry efficiency and a strong response to MVC already at low concentrations.

These results indicate that two of the viruses selected as X4 (220 and 685) do not exclusively use CXCR4 but also show the capacity to bind to the CCR5 coreceptor. This binding might be less efficient in case of these viruses which is demonstrated by their overall low entry efficiency. Viruses selected as dubious show binding affinity to CCR5 that is weaker than binding of R5 viruses and stronger than the binding of X4 viruses.

Differences in the overall entry efficiency of different types of viruses can be explained by the insufficient CXCR4 expression in this cell line. The overall entry efficiency appeared to be a discriminatory feature between the R5 and X4 clones. Since the models were aimed to capture the dependence of virus entry efficiency on other parameters there should be no variance on the overall entry efficiency among the clones. In order to avoid this problem in the further experiments, virus input levels were adjusted for each clone to reach similar general entry efficiency in the flow cytometry assay.

4.5.2 Flow cytometry

The data measured using the flow cytometry assay was used to generate models of entry efficiency. In this assay two rounds of experiments were performed. In the first round the first set of 8 selected clones was tested (Table 4.1) on which initial models were developed. In the second round of experiments a new set of clones was tested that was used for model validation and assessment of their capacity to distinguish virus phenotypes.

Table 4.4 lists example experiments performed in this assay – from the initial and following rounds of experiments. In the first measurement of all clones (e.g. experiment 100720) different levels of input virus are tested to establish the input levels of each clone that was maintained constant in the following experiments. Next, repeated measurements of virus response to different levels of individual drugs and drugs in combinations were performed. The SupT1-R5 cell line used in this assay does not offer the possibility to induce receptor and coreceptor expression. Models were therefore based on the natural variation of protein expression measured by antibody staining.

experiment	parameters varied	parameter values	cells per measurement
100720	input virus	8	35 719
100817	MVC	8	53 111
100824	AMD	8	77 443
100827	AMD	7	74 084
100831	AMD	7	74 805
100831	MVC	8	75 495
100903	MVC	8	74 407
100910	MVC, AMD	15	21 266
101013	MVC, AMD	11	47 639
101020	MVC, AMD	14	38 424
101021	MVC, AMD	15	39 313
101027	MVC, AMD	15	42 944

Table 4.4 Example measurements performed using flow cytometry. “Experiment” and “cells per measurement” are defined in the previous section (Table 4.3). The column “parameter values” indicates the number of values or combinations of values of the parameters varied in the respective experiment. The upper part of the table (above the double line) contains the first set of experiments used for model development; the bottom part contains example experiments used for validation and evaluation of the modeling approach.

In the 100720 experiment, different levels of input virus were tested (Figure 4.19). 30% was chosen as the optimal entry efficiency of a virus in the absence of drugs. This proportion of infected cells offers sufficient variation in the response variable for modeling while minimizing the probability of saturation by several virus particles infecting the same cell. The horizontal line in Figure 4.19 indicates the input virus levels chosen for each of the tested clones that were fixed in the following experiments.

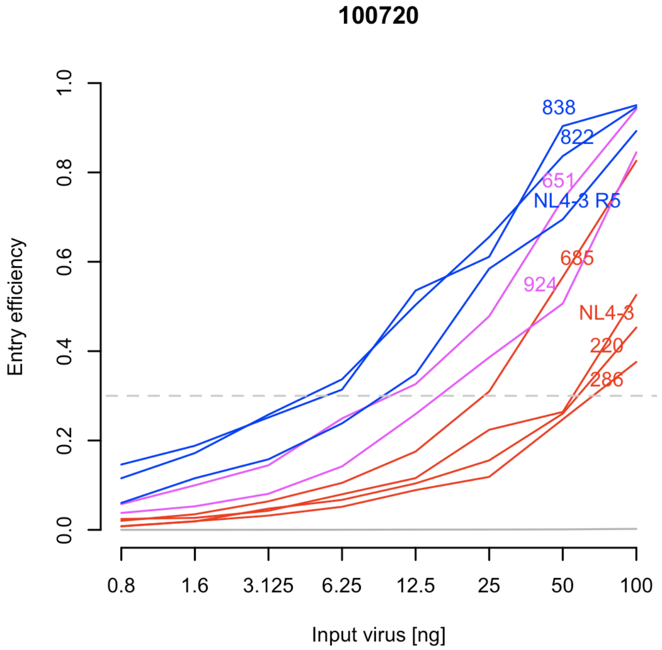


Figure 4.19 Entry efficiency of tested clones depending on the virus input levels. Optimal entry efficiency was established at the level of 30%. Points of intersection of the horizontal line with the curves representing each clone define virus input levels selected for each clone for the further experiments.

We analyzed the dependence of the entry efficiency of each clone on the measured parameters. First, we inspected the variation induced by different expression levels of the two coreceptors – CCR5 and CXCR4 in the absence of drugs. The relevant dependency plots (Figure 4.20) point to differences among the clones of R5 and X4 viruses. Clone 286 selected as an X4 virus, showed a dependence uniquely on the expression level of CXCR4 (Figure 4.20, left panel). Clone 838 being an R5 virus showed a notable dependence on the expression level of CCR5 (Figure 4.20, middle panel), 685 selected as an X4 virus showed a strong dependence on the expression level of both coreceptors (Figure 4.20, right panel).

In three experiments virus response to different levels of AMD was tested. Similar to the fluorescence microscopy results, only clone 286 and the NL4-3 reference virus showed response to this drug (Figure 4.21, left panel). 286 virus showed a weaker response to the drug than the reference virus with about ~5% infected cells at the highest drug concentration (1000 ng/ml). Clone 220 showed an unexpected increase in entry efficiency with higher drug concentration that was attributed to an unknown error in the assay.

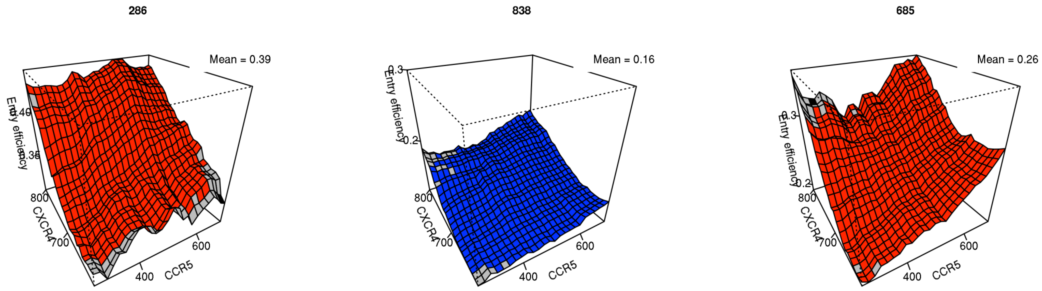


Figure 4.20 Dependence of the entry efficiency of the clones 286, 838 and 685 on the expression levels of two coreceptors – CCR5 and CXCR4. Plots are based on the data illustrated in Figure 4.21 not containing drugs. Clone 286 (left panel) responding to AMD showed dependence on the expression level of CXCR4 for its cell entry efficiency, clone 838 (middle panel) blocked by MVC showed dependence on the expression level of CCR5, clone 685 (right panel) responding only to the combination of drugs showed dependence on the expression of both coreceptors.

Three more experiments tested virus response to different concentrations of MVC. Here in addition to X4 viruses 286 and NL4-3, clone 685 showed no response to the drug (Figure 4.21, right panel). Noteworthy, this virus was not responsive to the AMD drug either, suggesting a dual-tropic nature of this virus – its capacity to alternate the use of two coreceptors. Other R5 and dubious viruses showed responses to MVC of similar strength.

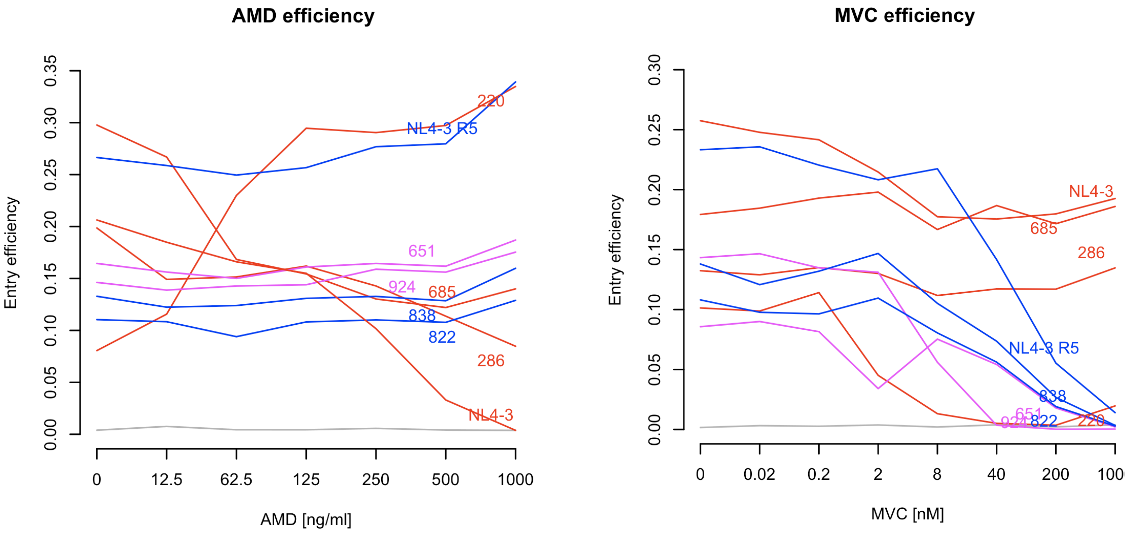


Figure 4.21 Effect of AMD (left panel) and MVC (right panel) on the average entry efficiency of the tested clones in the FACS assay. Curves show averaged entry efficiency of each clone based on three experiments testing AMD response and three experiments testing MVC.

Next, several combinations of AMD and MVC drug concentrations were tested. In contrast to the microscopy assay where rows and columns of a plate are measured simultaneously, FACS measurements are done sequentially. While the microscope plate provides a natural layout for testing combinations of two parameter values, FACS is a more time-consuming method. The number of combinations of drug concentration was therefore reduced from a full grid (30 measurements) to 15 measurements with a full range of MVC concentrations at a low (100 ng/ml) and a high (800 ng/ml) AMD concentration and four measurement of medium levels of concentrations of both drugs (Figure 4.22).

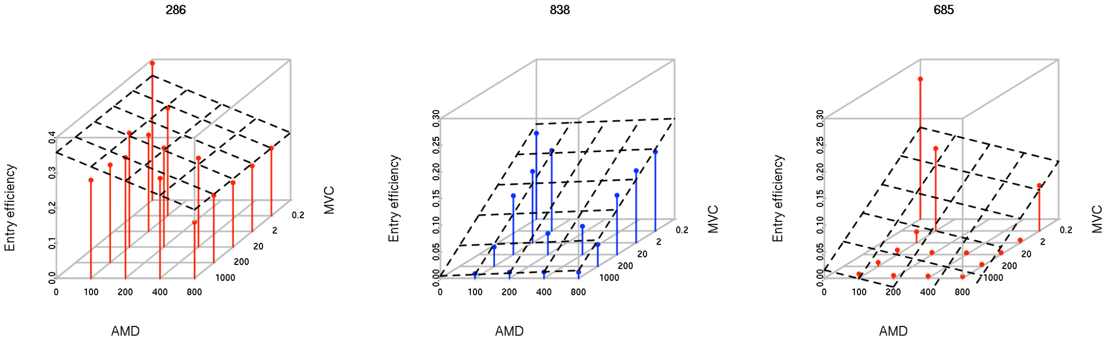


Figure 4.22 Entry efficiency of three example clones in the presence of two drugs at different concentrations. Planes traced with dashed lines are fitted using linear regression and indicate the direction of the decrease of the entry efficiency with the increase of each drug concentration. Virus 286 (left panel) shows response uniquely to AMD, 838 (middle panel) uniquely to MVC, 685 (right panel) to both drugs.

This combination test showed that the blocking is independent of the concentration of a second drug for the clones responding to a specific drug individually. Viruses that responded to AMD (286, NL4-3) showed the same relationship between AMD concentration and entry efficiency also in the presence of MVC (Figure 4.22, left panel). All viruses blocked by MVC showed response uniquely to this drug in the presence of AMD (Figure 4.22, right panel). Clone 685 that did not respond to AMD or MVC in the previous tests, was efficiently blocked by drug combinations even at low concentrations (Figure 4.22, right panel).

4.5.3 Observations

Observations collected in the tests of the initial set of viruses directed the development of models of virus entry efficiency. Experiments showed the predominance of R5 viruses in this clone set. Only one of the selected clones (286) showed a purely X4 phenotype with strong CXCR4 dependency and response to AMD. Clones selected as R5 viruses showed CCR5 dependency and response to MVC. All clones selected as dubious (651, 924) and one selected as X4 virus (220) showed a phenotype of an R5 virus however with a generally lower entry efficiency and a stronger response to MVC even at low drug concentrations. This points to a weaker binding affinity of these viruses to the CCR5 coreceptor as compared to R5 viruses. The fact that these viruses were not clearly classified as R5 viruses suggests a potential error of phenotypic assays on viruses with low CCR5 binding affinity.

One of the tested clones (685) showed no response to an individual coreceptor blocker and dependence of the entry efficiency on both of the coreceptors. However, combination of coreceptor blockers even at low concentrations was efficiently blocking the virus. This indicates the dual-tropic nature of this virus demonstrated by its capacity to shift its coreceptor use in the presence of coreceptor blockers.

Overall, the picture of virus phenotype emerging from the described experiments does not convey a clear distinction between X4 and R5 viruses. Viruses of the same tropism showed differences in general entry efficiency, strength of drug response and dependencies on coreceptor expression. This variety of profiles of virus entry efficiency calls for more elaborate models of virus phenotype extending beyond the mere tropism classification.

4.6 Models

Based on the above observations, we developed models of the dependence of the virus entry efficiency on the measured parameters. The models were aimed at distinguishing among the variety of virus phenotypes that are only partially captured by the tropism classification. They were based on the data generated by the flow cytometry assay, the initial set of viruses was used for training, remaining viruses for the validation and testing the models in their capacity to describe virus phenotypes.

Models were developed based on multivariate regression and on several quality measures for selecting the best model. The measured parameters were used as input variables, the entry efficiency as the response variable. Models were first trained and tested on the NL4-3 (*X4 model*) and NL4-3 R5 (*R5 model*) reference virus data. Based on these data the best model was determined according to predefined model selection criteria. Next, the selected model was trained on the cell entry data of each of the clones separately. The vector of input variable coefficients of the trained model of each virus was used as a multidimensional descriptor of the virus phenotype. This descriptor expresses virus entry dependency on each of the measured parameters in the form of the magnitude and sign of the respective coefficient. Coefficient vectors were next compared with the coefficient vectors of the reference viruses (X4 and R5 models) and plotted on 2D *phenotype maps* that illustrate similarity of virus phenotypes to the reference R5 and X4 phenotypes. Finally, based on the limited number of tested clones we developed a procedure for predicting phenotype vectors based on sequence. Such model for predicting virus phenotype without experimental testing is the eventual goal of this study. Details of the successive steps of the modeling procedure are described in the following sections.

4.6.1 Data preparation

Classification methods described in the section 4.4.2 assign to each cell either a binary value (binary classification) or a value in the range [0,1] (margin classification) that estimates the probability of a cell being infected. To reduce the number of highly redundant data points and to compensate for the potential noise in the measurement at the single cell level, the data was aggregated into a multidimensional grid defined on aggregated receptor and coreceptor parameter values – CD4, CCR5 and CXC4 and on all measured drug concentration levels. CD4, CCR5 and CXCR4 expression levels were first scaled to standard normal distribution, top and bottom 2.5% outliers were removed. Next a five-dimensional grid was defined – spanning over all tested concentration levels of AMD and MVC and over a predefined number of values of the CD4, CCR5 and CXCR4 expression levels dividing their range of expression into bins of equal size. We tested four numbers of bins – 5, 10, 20, 50 – for the quality of the resulting models. Classification values assigned to individual cells were averaged in each grid bin.

4.6.2 Model selection criteria

The best model was selected according to two criteria: accuracy of model fit to the data and separation of the X4 and R5 models.

The first criterion – accuracy of model fit to the data – was based on R^2 measure estimated as:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}},$$

where

$$SS_{err} = \sum_i \left(y_i - f_i\right)^2$$

is the sum of squared residuals with y_i being observed and f_i estimated output, and

$$SS_{tot} = \sum_i \left(y_i - \bar{y}\right)^2$$

is the total sum of squares proportional to the sample variance with $\bar{y} = \frac{1}{n}\sum_i^n y_i$ being the sample mean. R^2 was used as a measure of agreement between the observed and modeled values with higher values representing a better agreement. Models of cell entry efficiency were used as a description of a virus phenotype rather than for predicting the virus entry efficiency for a specific input. For this reason, model agreement with training data rather than its predictive power in a cross validation setting was used to evaluate the model.

The second criterion – separation of the R5 and X4 models – was used in order to obtain models capable of distinguishing between the two contrasting phenotypes. To measure the separation of the R5 and X4 models we used two measures of similarity of coefficient vectors: Pearson correlation and Euclidian vector distance. The two measures are related, R5 and X4 models are chosen such as to show the lowest possible correlation and the highest distance from each other. Models were tested on the merged data of all experiments performed on the first set of viruses (Table 4.4).

Additionally, the significance of model variable coefficients was inspected. Models showing insignificant coefficients were reduced by removing the respective input variables and recalculated. We used the analysis of variance (ANOVA) method on nested models excluding each of the input variables separately. F-test was used to assess the significance of a coefficient in the not reduced model as compared to the respective nested model.

4.6.3 Tested models

There is a large number of possible models that can be defined on the basis of five parameters. Here, we limited the tested models to those varying in six different respects. The corresponding model properties and their possible variations are listed in Table 4.5, numbered from 1 to 6 and discussed below.

	model property	values tested
1	aggregation level	5 / 10 / 20 / 50
2	response variable	binary / margin
3	underrepresented grid cells	0 / 10 / 20
4	drug concentration scale	linear / logarithmic
5	drug-coreceptor combination	yes / no
6	drug-drug combination	yes / no

Table 4.5 Varying elements of the tested models.

The aggregation level (1) corresponds to the number of bins into which the coreceptor and receptor expression levels are separated to produce data grid, as explained in the data preparation section 4.6.1. The tested bin numbers are 5, 10, 20 and 50. The response variable (2) is based on the classification method – binary or margin. The “underrepresented grid cells” property (3) defines the filter cutoff for the parts of the data grid that contain numbers of single cell measurements below a given value. The basis for introducing this filter is the observation that the distribution of residuals of the models showed a tail of high residual values supposedly due to the high variance of entry efficiencies estimated from a low number of single cell measurements. Elimination of parts of the data grid with a low number of cells resulted in a normal distribution of residuals. “Drug concentration scale” (4) involves two scales of the drug concentration parameter – linear and logarithmic. As more measurements were done for low drug concentrations the logarithmic scale produced a more balanced coverage of the range of the MVC and AMD variables’ values, which might result in a better model fit. The properties “drug-coreceptor combination” (5) and “drug-drug combination” (6) describe models in which in addition to the individual effects of the coreceptor expression levels and drug concentrations on the response, their combined effect is also tested. These models include the effect of all combinations of values of a coreceptor expression level and the corresponding drug concentration i.e. combinations of MVC concentration with CCR5 expression and AMD concentration with CXCR4 expression, as well as of the combinations of values of the two drugs. The coefficient values of the combined input variables are more difficult to interpret, however, they might contribute to a better model fit and R5-X4 models separation. The combination of the varied model properties resulted in 192 models tested, 48 for each aggregation level.

Models including insignificant ($p > 0.05$) variable coefficients were reduced accordingly. Strong data aggregation resulted in a higher number of reduced models – among models based on data aggregated to 5 bins 64.6% were reduced, among those based on data aggregated to 50 bins 25.0% were reduced (Table 4.6). Notably only the coefficients of combined variables were insignificant, coefficients of coreceptor and receptor expression levels as well as drug concentrations were significant in all tested models. Table 4.6 lists the quality of models based on different levels of aggregation. It appears that aggregation 50 results in models of a markedly lower fit to the data (R^2 ~ 0.10 on average) as compared to stronger data aggregations 5, 10 and 20 (R^2 ~ 0.47 on average). In the following analysis only models based on the aggregations 5, 10 and 20 were considered.

aggre-gation	reduced models [%]	max R^2 (mean)		min R5-X4 correlation (mean)	max R5-X4 distance (mean)
		R5 model	X4 model		
5	64.6	0.726 (0.516)	0.622 (0.521)	0.161 (0.224)	1.715 (1.585)
10	58.3	0.711 (0.467)	0.632 (0.474)	0.135 (0.222)	1.781 (1.594)
20	33.3	0.676 (0.406)	0.615 (0.426)	0.045 (0.202)	1.943 (1.623)
50	25.0	0.131 (0.102)	0.126 (0.108)	0.128 (0.200)	1.805 (1.636)

Table 4.6 Quality of models based on different levels of data aggregation. “Reduced models” indicates the percentage of models that were reduced by removing of the the variables showing insignificant coefficients. Model quality is based on the model fit to the data (R^2) and separation of the R5 and X4 models (R5-X4 correlation and distance).

Table 4.7 lists the quality of models grouped according to the five remaining model properties. Among them, filtering of the underrepresented data grid cells (3) showed the strongest pattern of increase in the model fit and R5-X4 model separation with the stringency of the filter. Filtering of the grid bins that contain below 20 data points resulted in mean R^2 ~ 0.57 and mean R5-X4 model correlation ~0.18 as compared to the mean R^2 ~ 0.30 and the mean R5-X4 model correlation ~0.26 of the models based on data where only grid bins containing below 5 data points were filtered. In the data aggregation to 5, 10 and 20 bins the filters removed <1%, 2-6.5% and 25-50% single cell measurements respectively. Also the logarithmic drug concentration scale (4) resulted in models of a higher fit to the data (mean R^2 ~ 0.54) as compared to the linear scale (mean R^2 ~ 0.39). Among the remaining model properties (2, 5, 6) a clear pattern of model fit and the R5-X4 model separation for models of various features was not observed.

			max R^2 (mean)		min R5-X4 correlation (mean)	max R5-X4 distance (mean)
			R5 model	X4 model		
2	response	binary	0.716 (0.456)	0.652 (0.464)	0.045 (0.211)	1.943 (1.614)
		margin	0.726 (0.469)	0.662 (0.484)	0.064 (0.222)	1.904 (1.587)
3	under-represented grid cells	5	0.468 (0.281)	0.491 (0.320)	0.211 (0.259)	1.655 (1.519)
		10	0.690 (0.530)	0.643 (0.527)	0.088 (0.205)	1.655 (1.519)
		20	0.726 (0.577)	0.662 (0.574)	0.045 (0.185)	1.943 (1.659)
4	drug concent-ration scale	linear	0.480 (0.369)	0.563 (0.425)	0.045 (0.201)	1.943 (1.660)
		logarithmic	0.726 (0.557)	0.662 (0.523)	0.131 (0.231)	1.737 (1.541)
5	drug-coreceptor combination	no	0.722 (0.462)	0.661 (0.473)	0.045 (0.210)	1.943 (1.621)
		yes	0.726 (0.463)	0.662 (0.475)	0.135 (0.223)	1.781 (1.580)
6	drug-drug combination	no	0.723 (0.462)	0.662 (0.473)	0.078 (0.229)	1.885 (1.579)
		yes	0.726 (0.463)	0.662 (0.475)	0.045 (0.204)	1.943 (1.622)

Table 4.7 Quality of models varying in the five model properties (2-6). Models are based on the data aggregation to 5, 10 and 20 segments and those not satisfying the parameter coefficient criterion are reduced accordingly. Maximum and mean model fit to the data ("R^2") and the best and mean separation of the R5 and X4 models (R5-X4 correlation and distance) are shown for models grouped according to each of the properties.

Due to the lack of clear best model properties, we inspected the distribution of the model features among models of different fit to the data and R5-X4 model separation (Figure 4.23). We selected two models for further inspection – one offering the best R5-X4 model separation and another one showing a lower separation and a better model fit. The models are indicated with arrows in Figure 4.23 and differ in model property "drug concentration scale" (4) with the logarithmic scale showing a better model fit (R^2 ~ 0.62 vs 0.49) and linear scale a better model separation (R5-X4 model correlation of ~ 0.05 vs 0.13). We termed these models *logarithmic model* and *linear model* respectively. Remaining properties are shared by both models and involve:

1. aggregation level: 20
2. response variable: binary method
3. underrepresented data grid cells: 20
5. drug-coreceptor combination: yes
6. drug-coreceptor combination: yes

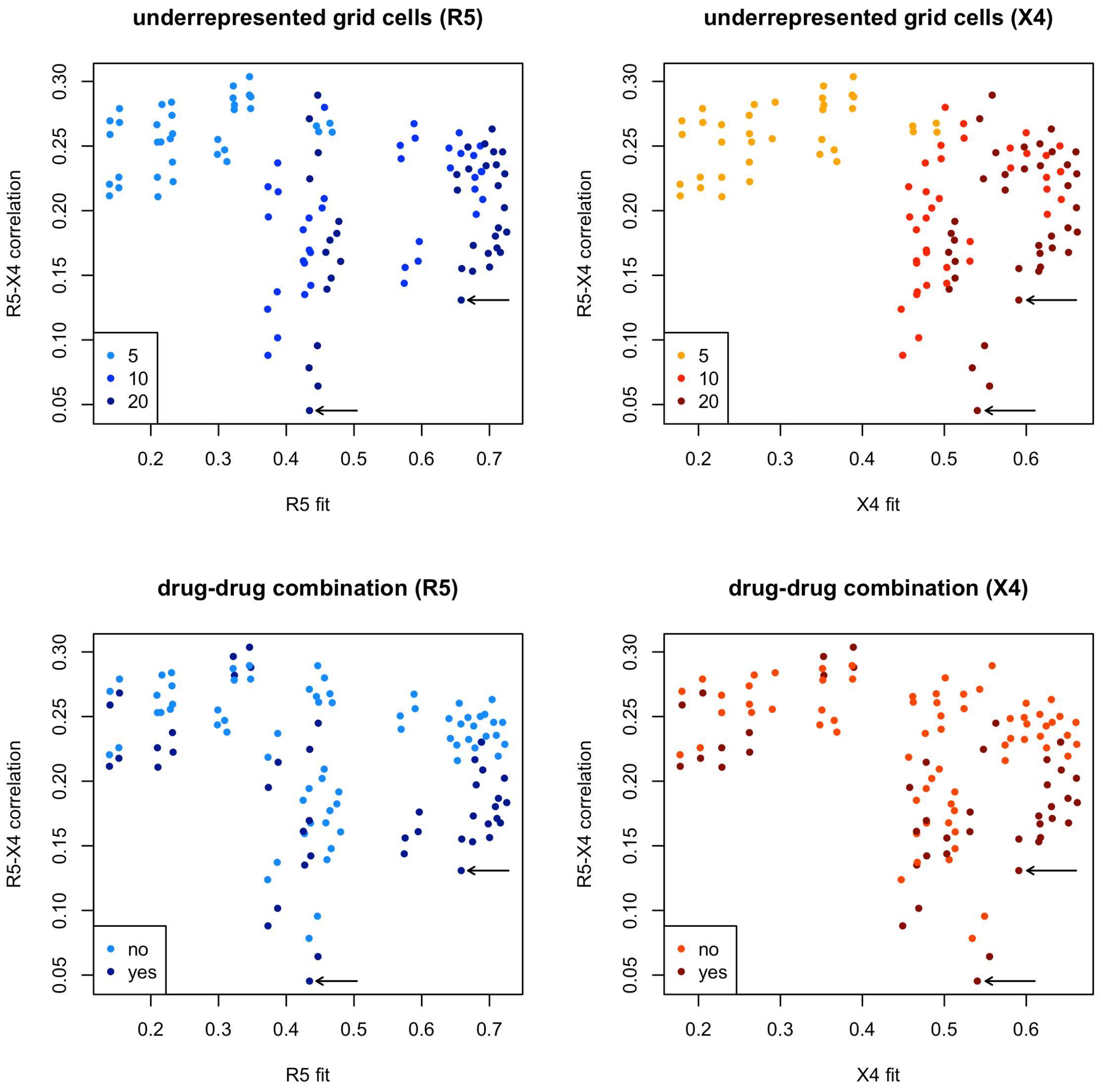


Figure 4.23 Fit of the R5 and X4 models vs their separation. Dots represent the tested models. Left hand side panels show R5 and right hand side X4 models. Color shades correspond to values of two properties of the models: underrepresented grid cells (3, upper panels) and drug-drug combination (6, bottom panels). Optimal model is the one offering low correlation and high fit to the data – models located in the bottom right corners of the plots. The property “underrepresented grid cells” (upper panels) shows a clear pattern of a better model fit with a higher filter value while the property “drug-drug combination” (bottom panel) is more evenly distributed among models of different fit and separation. Selected models – logarithmic (better fit) and linear (better separation) – are indicated with arrows.

Figure 4.24 shows the parameter coefficient estimation of the reference R5 and X4 logarithmic (upper panels) and linear (bottom panels) models. Logarithmic and linear drug concentration scales result in different magnitudes of coefficients however the relative values of the coefficients are similar between the scales. R5 reference clone is characterized by inhibition by MVC and dependence on CCR5, which is reflected by negative MVC and positive CCR5 coefficients. Conversely, X4 reference clone is characterized by inhibition by AMD and dependence on CXCR4, which is reflected by negative AMD and positive CXCR4 coefficients. This way, the coefficients represent a multivariate phenotype description of the V3 loop variants.

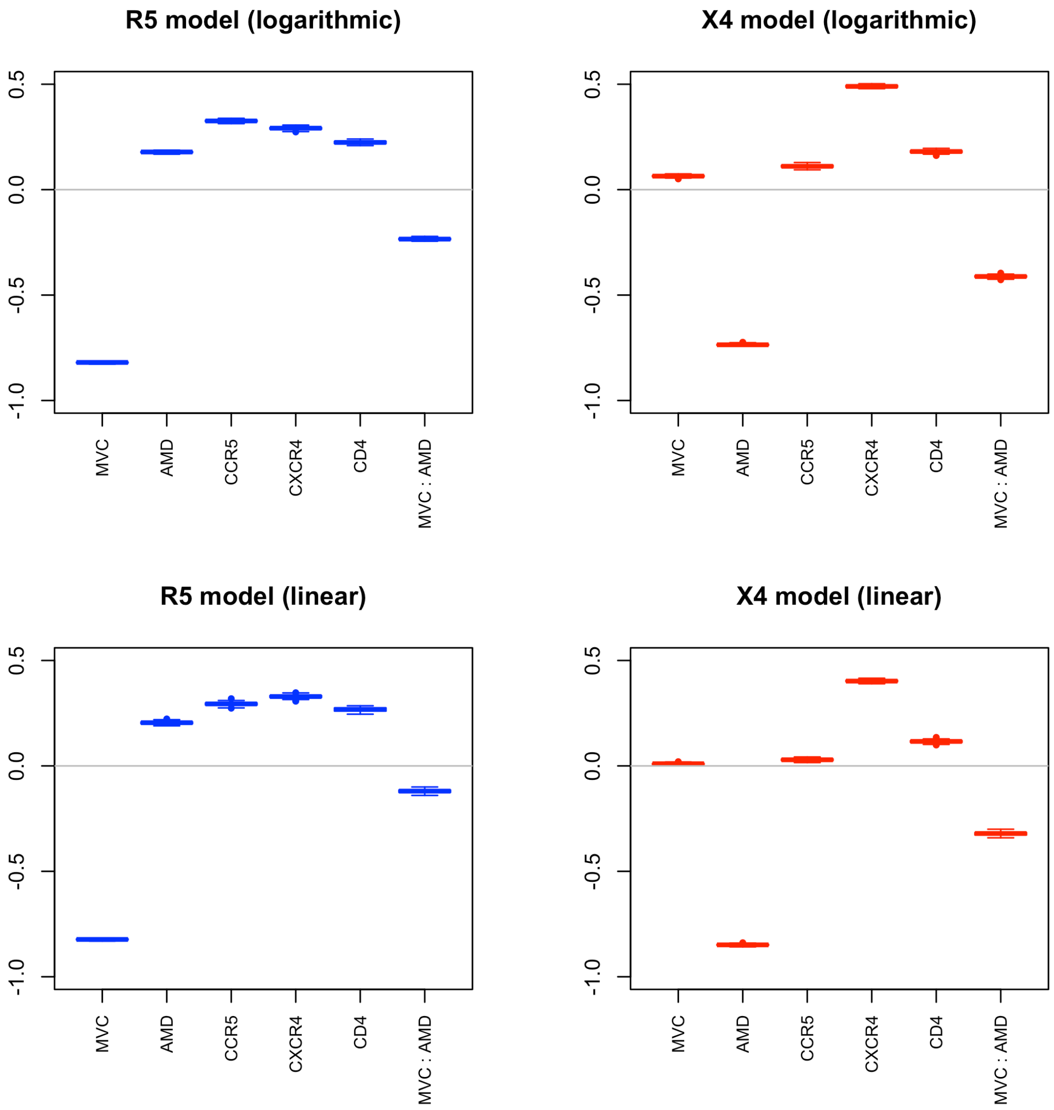


Figure 4.24 Parameter coefficient estimation of R5 and X4 selected models with the logarithmic (upper panels) and linear (bottom panels) scales of drug concentration. Boxplots show parameter coefficient variation in a bootstrapping test of 100 times sampling of 75% data points and model estimation.

4.6.4 Phenotype map

Models selected in the previous step of this analysis were next trained on the entry efficiency measurements of each of the clones separately. This resulted in a parameter coefficient vector representation of each clone. The coefficient vector reflects the dependence of the entry efficiency of a virus on the measured parameters and is termed here *phenotype vector*. The phenotype vectors of the reference viruses – NL4-3 R5 and NL4-3 – represent reference R5 and X4 phenotypes respectively. Correlation and distance between the phenotype vectors of tested clones and the two reference phenotypes were used as measures of clone similarity to both reference phenotypes. Each clone was next represented by two numbers expressing its similarity to the R5 and X4 phenotypes. This representation allows for visualizing clone phenotype on convenient 2D plots termed here *phenotype maps*. The linear model produced a stronger separation of the reference models, for convenience in the following part of this section phenotype maps based on the linear model are presented.

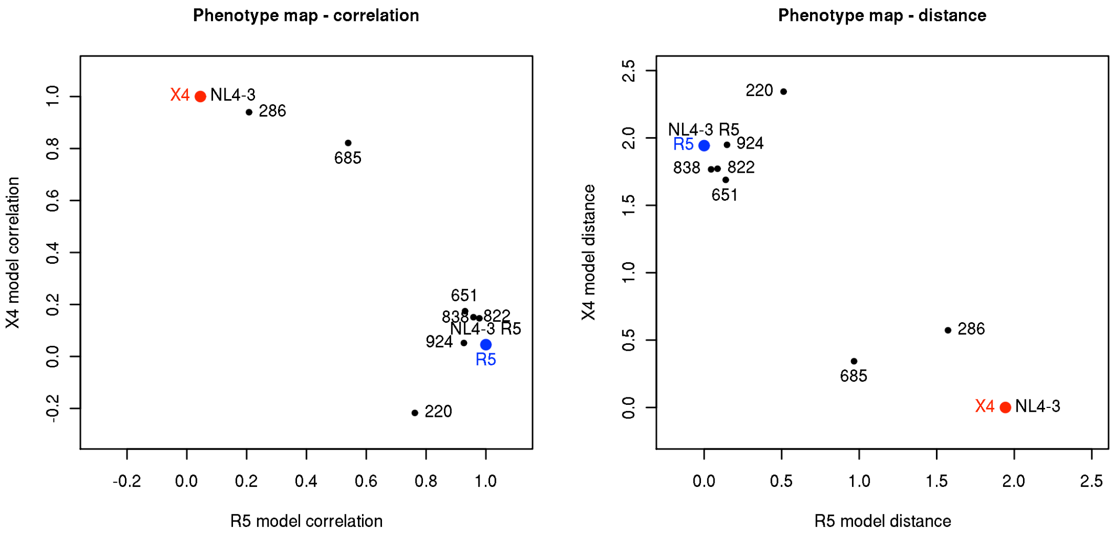


Figure 4.25 Phenotype map of the first set of data based on correlation (left panel) and distance (right panel) with R5 and X4 reference models. Reference phenotypes are marked with colored dots, red for X4 and blue for R5 phenotype. Clones are represented by black dots and labeled.

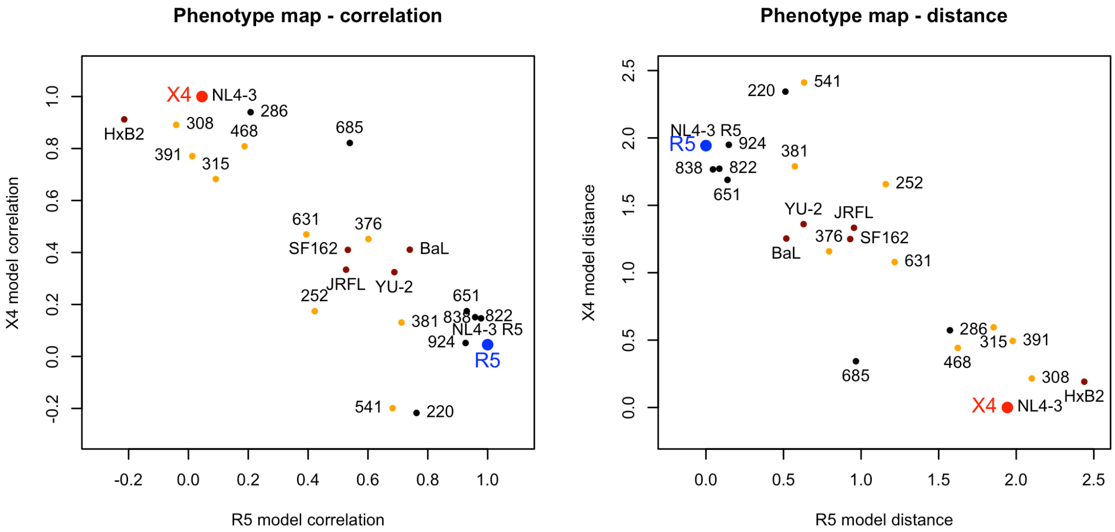
Figure 4.25 shows phenotype maps of the initial set of clones. Most of the clones showed a phenotype typical for R5 viruses which is reflected by their high correlation with and a small distance to the R5 reference virus. Only one clone in this dataset, clone 286, showed X4 phenotype which is reflected by its position close to the X4 reference phenotype on the map. Clone 685 showed the phenotype of a dual-tropic virus, responding only to the coreceptor blockers in combination. This clone appears in between two phenotypes on the map closer however to the X4 phenotype.

4.7 Prediction of phenotype vectors

The analysis of the first set of clones showed the capacity of our experimental and computational pipeline of distinguishing between R5, X4 and dual-tropic viruses. Virus position on a phenotype map indicates its entry dependence on coreceptors and susceptibility to inhibition by each of the drugs. Phenotype maps can be therefore used for visual display of the boundary between clones sensitive and non-sensitive to MVC treatment. Since experimental testing of an individual clone is a costly process, circumventing the testing by prediction of the phenotype vector based on the V3 loop sequence is essential for the practical use of our method for patient treatment. In this section we present a prediction model of virus phenotype based on sequence. Only few clones were phenotypically tested in this study so far, we therefore first address the problem of high dimensionality of the data with a low number of observations. Next, we estimate the number of clones necessary for obtaining satisfactory prediction accuracy.

4.7.1 Populating of the phenotype map

In order to gain insight into a potential separation of the phenotypes on the map, additional phenotype vectors of the clones tested in the following rounds of experiments were used to populate the phenotype map. The experimental measurements of the clones were subject to data processing pipeline described above. The selected models were next trained on the data of each clone and the resulting phenotype vectors were plotted on the phenotype map of the clones tested in the first round of experiments (Figure 4.26).


Phenotype map - correlation

Phenotype map - distance

Figure 4.26 Phenotype map of the first set of data (Figure 4.25) enriched with clones (orange) and lab-adapted viruses (dark red) tested in the following rounds of experiments. Reference phenotypes are marked with larger red and blue dots.

All the clones in the second dataset, except the lab-adapted clones, were selected as X4 viruses. Based on visual inspection of the clone tested in our assay, four of the clones showed X4 (308, 315, 391 and 468), three R5 (376, 381 and 541) and two dual-tropic phenotype (252 and 631). The dual-tropic clones were not inhibited or only partially inhibited by individual drugs however showed a strong response to drugs in combination.

Figure 4.27 illustrates the strength of the inhibition by MVC and AMD of the clones on the phenotype map. The dots on both panels are colored in a range of colors between red and blue. On the left plot blue dots represent clones susceptible to MVC, red – non-susceptible. On the right panel red dots represent clones susceptible to AMD, blue – non-susceptible. Two groups of clones can be distinguished on the map, responsive to MVC and responsive to AMD (colored in different shades of blue and red respectively). Coloring of clones 685, 631 and 252 suggests either an attenuated response to both drugs (685 and 631) or no response to any of the drugs (252). These clones are also positioned between the two groups of R5 and X4 viruses on the map, clones 252 and 631 closer to R5 clones, clone 685 closer to X4 clones. This illustrates the way clone position on the map provides information on its susceptibility to drugs and where a potential boundary between phenotypes can be situated.

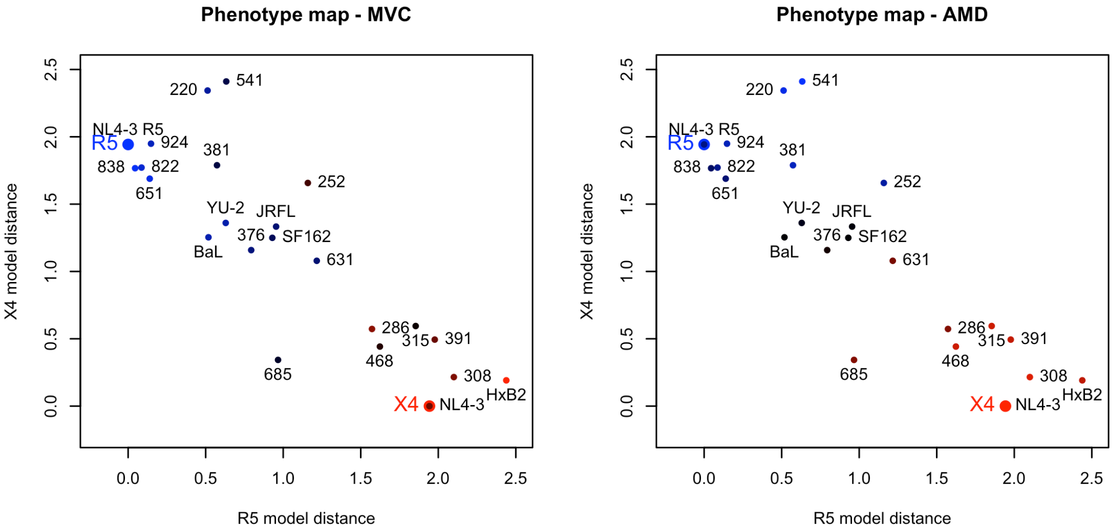


Figure 4.27 Phenotype map based on distances to the reference models colored according to the MVC (left panel) and AMD (right panel) coefficients. On the left panel gradient of color between blue and red indicates MVC coefficient with low coefficient indicating susceptibility to MVC colored in blue. On the right panel gradient of color between red and blue indicates AMD coefficient with low coefficient indicating susceptibility to AMD colored in red.

4.7.2 Phenotype prediction

Given the essential information for patient treatment that can be derived from the phenotype map and the cost of phenotypic testing of an individual virus variant, a computational method for phenotype prediction based on V3 loop sequence is important for the practical use of our approach. Since the variability of V3 loop, and in particular of X4 strains, is high, clones tested within this part of the study might be insufficient for accurate prediction of the phenotype. To address this problem the predictive model developed here is based on shrinkage methods that allow for reducing the high number of dimensions of the sequence data compared to the low number of measured clones. Throughout this section the phenotype vectors estimated using linear model are used.

The binary sequence encoding of the V3 loops of the 23 tested clones includes 88 positions that show variation among the clones. Since the number of clones (23) is lower than the length of the sequence vector, the problem of phenotype prediction is underdetermined and cannot be solved by least square approach. Instead of classical regression we used two methods that involve regularization procedures: Ridge regression (Tychonoff 1943) and Lasso (Tibshirani 1996). These methods were trained on the binary sequence encoding of the clones with the respective phenotype vectors as the output variables. For each position of the phenotype vector the penalty parameter λ

resulting in minimal prediction error in leave-one-out cross validation (LOOCV) was chosen from a sequence of 100 values. The chosen λs were used in further phenotype prediction.

In addition to Ridge and Lasso, we tested the performance of linear regression based on a reduced number of input variables. The input variables were reduced to those showing significant ($p < 0.01$) pearson correlation with any of the output variables. Significance was calculated in 1000 permutation tests. The reduction procedure resulted in 26 and 17 input variables in the logarithmic and linear models of phenotype vectors respectively.

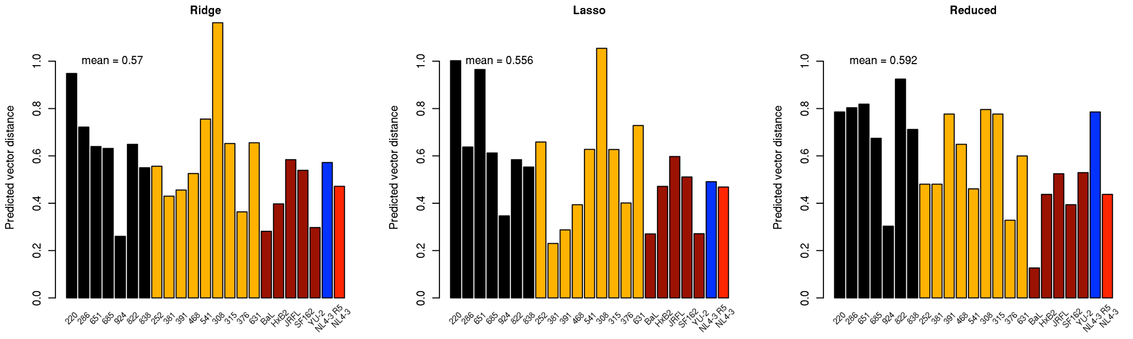


Figure 4.28 Error of the predicted phenotype vectors of three models – Ridge (left panel), Lasso (middle panel) and reduced linear model (right panel). Predictions are calculated in LOOCV setting, error is estimated as Euclidian distance between the predicted and observed phenotype vectors. Bar colors represent clone groups: black bars depict clones tested in the first, orange in the following rounds of experiments, dark red bars depict lab-adapted clones, blue and red bars depict R5 and red X4 reference clones respectively.

Figure 4.28 shows prediction errors of the three models on the tested clones. The reduced linear model showed the lowest accuracy, presumably due to the number of input variables exceeding the number of observations in this model. Among three prediction models similar relative prediction errors among clones were observed. Low prediction error was obtained for clones phylogenetically close to any other clones in this test set or being situated in uniform R5 clusters as pictured in Figure 4.2. Three clones showing a markedly elevated prediction error are clones 220, 651 and 308. Notably these clones were phenotyped and classified as X4 viruses however two of them, 220 and 651, showed R5 phenotype in our assay. In sequence space these clones are located in a sparse cluster occupied by both R5 and X4 viruses (Figure 4.2) suggesting the sequences of these viruses represent boundary cases whose minor changes might result in tropism switch. These cases should be analyzed in detail for further characterization of the changes determining their tropism. A training set of clones covering the boundary parts of V3 sequence space is therefore critical for obtaining an accurate prediction method. Model based on Lasso shrinkage showed the highest accuracy in the linear and logarithmic models and was used in the following steps of this analysis.

In this study phenotype map is used for visualizing and approximating virus phenotype. Next, we inspected how a clone position on the phenotype map differs between the observed and predicted phenotype vectors. Figure 4.29 shows phenotype maps with three examples of predicted phenotypes indicated with arrows. Predicted phenotypes of the clones 838 and 685 show a relatively small change of position on the phenotype map, which does not result in the misclassification of virus tropism. The prediction error of clone 286 is higher – the location of the predicted phenotype on the map suggests it is dual-tropic contrary to the observed virus X4 phenotype.

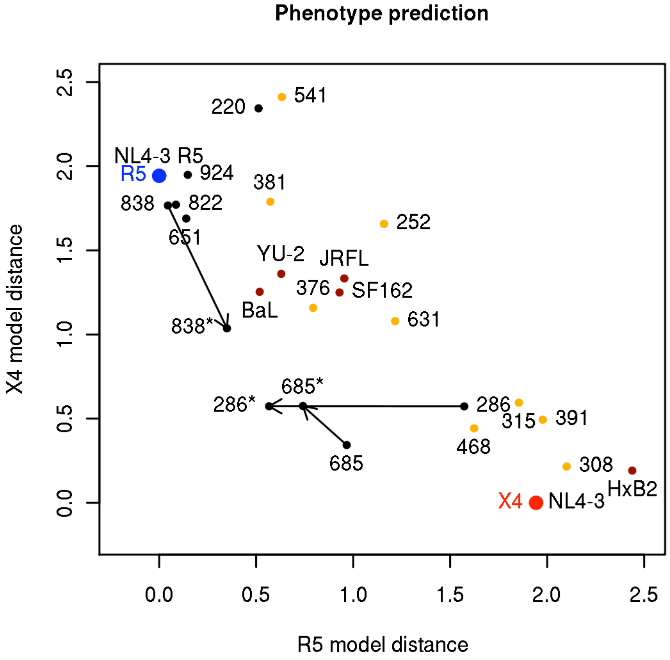


Figure 4.29 Position of predicted phenotypes of three clones 838, 685 and 286 on phenotype map. The arrows connect the positions of the observed with predicted phenotypes labeled with respective clone name and an asterisk.

4.7.3 Improving phenotype prediction

Finally, in order to test the feasibility of an accurate sequence-based phenotype prediction method, we approached the question of the relationship of the size of the training with the accuracy of prediction. In order to define an accurate prediction, we introduced the notion of *borderline phenotype*. A borderline phenotype is derived from the observed clone phenotype by reducing the coefficients characterizing given phenotype to zero. For an R5 virus these coefficients are CCR5 and MVC, for an X4 virus CXCR4 and AMD, for a dual-tropic virus MVC and AMD. This way derived borderline phenotypes are located in the region between R5 and X4 viruses on the phenotype map (Figure 4.30) and represent the minimal change in a virus phenotype vector that leads to virus tropism misclassification.

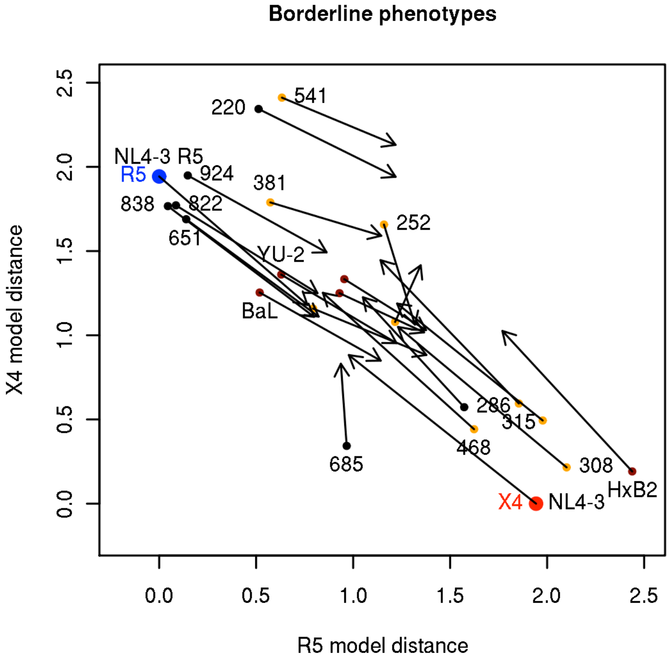


Figure 4.30 Positions of borderline phenotypes on the phenotype map. Arrows connect the observed clones' phenotypes with the respective borderline phenotypes.

Next, we investigated the relationship between the size of the training set and the prediction accuracy. We devised a procedure of predicting each clone's phenotype based on a sampled training set of an increasing size between 2 and 22. Training sets of each size were sampled multiple times; the prediction error of each clone was averaged for each size of the training set. A polynomial function $f(x) = ax^b$ was next fitted to the relationship between the size of the training set and the prediction error. In Figure 4.31 three examples of fitted functions are shown showing different gradients of the error decrease with the increasing train set size.

Three of the tested clones – 220, 651 and 308 – showed an increasing error function with $b > 0$ suggesting that an accurate prediction of their phenotype is not possible based on the current train set. Notably these three clones are located on a clade of the tree presented in Figure 4.2 occupied by both X4 and R5 viruses and showing relatively long branches. Virus 308 shows X4 phenotype, viruses 220 and 651 an R5 phenotype despite being classified as X4 viruses based on sequence. This indicates that there is a lack of clones of sufficient sequence similarity to the clones 220, 651 and 308 to obtain their accurate phenotype predictions. Shrinkage methods failed to select features predictive of the phenotype of these clones as they represent outliers in the sequence-phenotype pairing. These cases were excluded from further error estimation it should be however considered that the averaged error and train set size might be underestimated due to the exclusion of these cases.

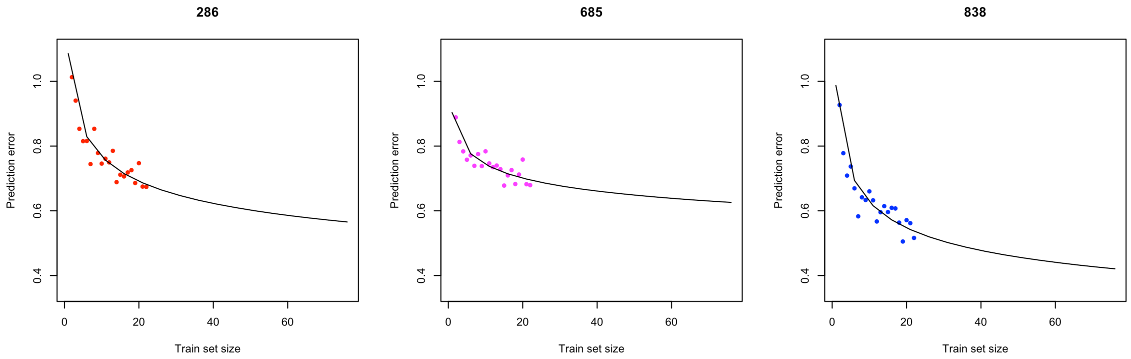


Figure 4.31 Fitted function of the prediction error against the size of training set for three example clones 286 (left panel), 685 (middle panel) and 838 (right panel). Dots represent averaged prediction error obtained for sampled training sets of respective sizes. Black lines represent the fitted function plotted in the range (0, 75] of the train set sizes.

The b parameter of the functions of the remaining clones was located in the range (-0.454, -0.059) with a mean of -0.244. The mean was similar for the R5 and X4 clones (-0.256 and -0.261, respectively) even though R5 clones showed a lower prediction error on average in the LOOCV setting, which suggests that with a representative train set covering large parts of sequence space an accurate prediction of the X4 phenotype is possible.

The fitted functions were next used to approximate the prediction error of models constructed on training sets of sizes exceeding 23. Figure 4.32 shows averaged fitted functions of all (thick black line) of R5 (dashed blue line) and of X4 (dashed red line) clones plotted in the range (0, 1250]. As previously described, borderline phenotypes represent the minimal change in a virus phenotype vector that results in virus tropism misclassification. Vector distance between the observed and respective borderline phenotypes was therefore used to define a cutoff for prediction accuracy. We defined two cutoffs: the minimal error of clones with an X4 or R5 phenotype (*R5/X4 cutoff*) and the minimal error of dual-tropic clones (*dual-tropic cutoff*). Dual-tropic clones are situated between the R5 and X4 phenotypes on the map, the difference between their observed and borderline phenotypes is therefore small and represents a conservative criterion for prediction accuracy. R5 and X4 clones show greater differences with borderline phenotypes and the error margin allowing for their correct tropism prediction is larger. R5/X4 cutoff is therefore less stringent than the dual-tropic cutoff however might result in misprediction of the dual-tropic clones. R5/X4 and dual-tropic cutoffs are represented by dashed horizontal lines colored in black and magenta respectively in Figure 4.32. The averaged function reaches the R5/X4 cutoff at the train set size of 50 and the dual-tropic cutoff at the train set size of 360. The upper quantile of the fitted functions (black thin line) intersects the R5/X4 cutoff at the train set size 290. These results were obtained for linear models of the clone entry efficiency. Logarithmic models showed train set sizes of 110 and 530 for R5/X4 and dual-tropic cutoffs respectively. These results indicate that in order to obtain satisfactory prediction accuracy on average a training set of ~100 clones is required. However a reliable prediction of the X4 and dual-tropic phenotype can only be achieved by augmenting the training set to the size of ~400-500 clones.

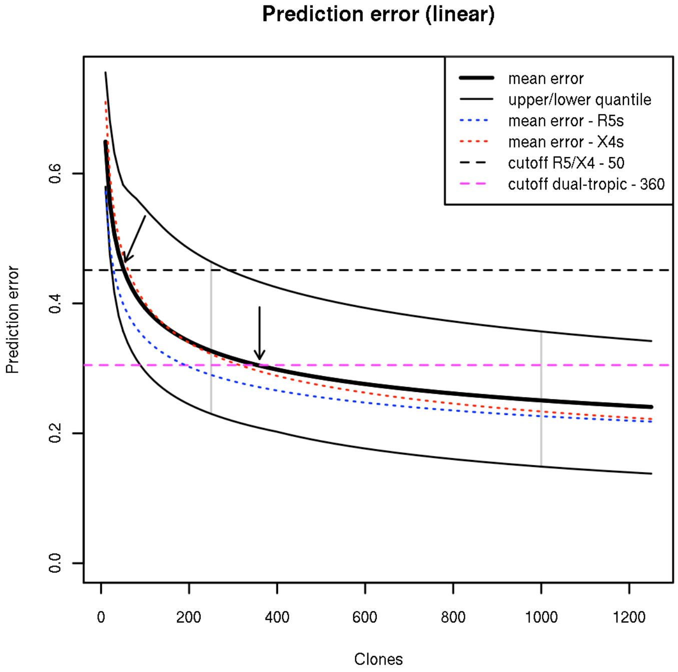


Figure 4.32 Estimation of the prediction error with an increasing number of clones. Thick black line represents the average of fitted functions of all tested clones, thin black lines their quantiles. Two vertical gray lines show the distance between the quantiles. Dashed red and blue lines represent the average of fitted functions to the X4 and R5 clones respectively. Dashed horizontal lines represent R5/X4 (black) and dual-tropic (magenta) cutoffs. Train set size of the averaged function at two cutoffs are pointed by arrows and indicated in the legend.

As the last step of this study, we elaborated a quantitative description of the clones whose phenotype is difficult to predict based on current set of tested clones. The goal of this analysis was to characterize clones such as 220, 651 and 308 that show a lack of relation between the accuracy of their phenotype prediction and the size of the underlying train set which suggests that the sequences in the current train set provide insufficient information for the accurate phenotype prediction of these three clones. We based the analysis on two factors that might affect the accuracy of a clone phenotype prediction: presence of a genetically similar clone in the train set, and the position of the clone in V3 sequence space. Genetically similar clones in the train set might provide information on the sequence features that are determinant for the phenotype of a given clone. Location of a clone sequence in V3 sequence space reflects either its close affinity to sequences of a particular tropism or its location on the tropism boundaries in which case distinguishing sequence features proper to a specific phenotype is more difficult.

clone	closest clone	distance to	cluster			
	(distance)	cluster center	nb	score	R5s	X4s
838	JRFL (0.19)	0.44	1	0.95	96	1
JRFL	838 (0.19)	0.32				
SF162	JRFL (0.25)	0.40				
BaL	JRFL (0.14)	0.56				
YU-2	BaL (0.17)	0.84				
NL4-3 R5	JRFL (0.13)	0.48				
822	JRFL (0.19)	0.63	2	0.93	139	1
924	NL4-3 R5 (0.22)	0.69	3	0.92	44	0
631	685 (0.48)	0.80	4	0.63	11	1
541	NL4-3 R5 (0.31)	0.14	5	-0.13	9	12
220	376 (0.48)	0.53	6	-0.10	7	9
651	SF162 (0.54)	0.67				
252	381 (0.18)	0.26	7	-0.36	1	5
315	376 (0.22)	0.91				
376	315 (0.22)	0.74				
381	252 (0.18)	0				
391	685 (0.42)	0.51	8	-0.17	4	6
468	631 (0.59)	0.81				
685	391 (0.42)	0				
308	BaL (0.47)	1.00	9	0.79	12	1
HxB2	NL4-3 (0.14)	0	10	-1.00	0	2
NL4-3	HxB2 (0.14)	1.00				
286	685 (0.47)	0	11	-0.75	0	3

Table 4.8 Characterization of clones in V3 loop sequence space. Clones are grouped according to the cluster they belong to as described in section 4.2 of this chapter. Closest clone among the tested clones is indicated in column "closest clone" with the distance normalized to [0,1] in the V3 sequence dataset. Distance to the cluster center and cluster score are measured as described in the text above. Cluster number, amounts of R5 and X4 sequences are indicated in respective columns.

Table 4.8 lists characteristics of the sequences of the tested clones. Clones are grouped according to the cluster they belong to as described in the clustering procedure in section 4.2 of this chapter. Column "closest clone" indicates the genetically closest clone among the tested clones and its distance normalized to [0,1] interval among all distances in V3 sequence space. Following columns depict the clusters that the clones belong to and the clone position in the clusters. Column "distance to cluster center" lists the distances of clone sequences to the respective cluster center relative to the cluster radius. *Cluster center* is defined here as the sequence of a minimal mean distance to other sequences within the cluster. *Cluster radius* is defined as the maximal distance of a cluster sequence to the cluster center. Distance 0 in the column "distance to cluster center" indicates cluster center, distance 1 indicates the most distant sequence of a cluster, lying on the cluster boundary. To depict the composition of a cluster we introduced cluster score as:

$$c_i = {}^{(r5_i - x4_i)}\Big/_{(r5_i + x4_i + r5x4_i)},$$

where $r5_i, x4_i, r5x4_i$ are the numbers of R5, X4 and dual-tropic sequences respectively in cluster i. Cluster score close to 0 is characteristic of mixed and dual-tropic clusters, close to -1 of predominantly X4 clusters, close to 1 of predominantly R5 clusters.

Clusters 1-4, 7, 9-11 represent relatively uniform X4 or R5 clusters as reflected by the value of the cluster score clearly distinct from 0. It can be assumed that sequence space location of the clones that pertain to these clusters provides information for their phenotype prediction. Clone 308 belonging to cluster 9 represents the most distant sequence from the cluster center lying on the cluster boundaries (distance = 1). This cluster is predominantly R5, however the closest neighbor of 308 in this cluster is the only X4 sequence in this cluster suggesting the boundary location of this clone (Figure 4.33, middle panel). In cluster 10, clone HxB2 is also located on the cluster boundaries, however, unlike clone 308, this clone is closely related to another clone in this cluster (NL4-3) that provides information for its phenotype prediction.

Among the remaining clusters, cluster 6 shows the score closest to 0 (score = -0.1) among the analyzed cluster, suggesting its mixed composition. This cluster is also sparse as reflected by long branches of the respective tree (Figure 4.33, left panel). Two clones contained in this cluster – 220 and 651 – are in addition relatively distant from other clones (distance >= 0.48) which underlies the difficulty of their phenotype prediction. Clone 541 belongs to another mixed cluster 5. However this clone is closely located to clone NL4-3 R5 in sequence space which provides information for its phenotype prediction. Similarly, clones contained in cluster 8 that is also relatively mixed (score = -0.17) are each other's closest neighbors among the tested clones which might contribute to the accurate prediction of their phenotypes.

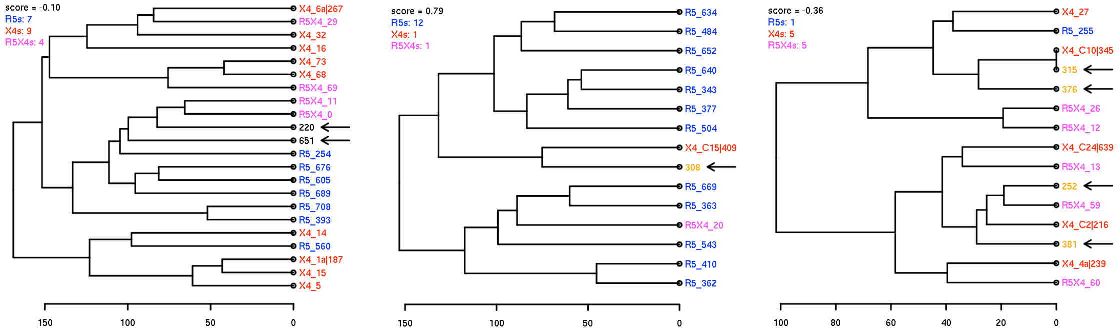


Figure 4.33 Dendrograms of three example clusters from the clustering described in section 4.2 of this chapter. Cluster 6 (left panel) is a mixed and sparse cluster as reflected by the equal proportion of R5 and X4 sequences and long tree branches. Cluster 9 (middle panel) is a uniform R5 cluster with clone 308 located on its boundaries. Prediction error of the clones 220, 651 and 308 does not improve with the increasing size of the train set in this study. Cluster 7 (right panel) is predominantly X4 and contains four closely related clones from the train set. Phenotypes of the clones in this cluster are predicted more accurately with the increasing size of the train set.

Overall, this analysis shows that sequence location on a tropism boundary (clone 308) or a location in a mixed cluster with no close neighbors among the tested clones (clones

220, 651) contribute to the difficulties in the phenotype prediction. With more sequences tested that would populate a larger number of clusters, a systematic approach to establishing a confidence score for the phenotype prediction will be possible. Based on the current clone set two rules for recognizing clones of high prediction error can be derived:

- sequences located on the boundaries of clusters not containing other clones from the train set, or

- sequences belonging to mixed clusters (absolute cluter score < 0.15) and far from other tested clones (distance > 0.45).

These two rules characterize clones 308, 220 and 651 in our dataset and 68 out of 880 sequences in the V3 sequence dataset used for this analysis. Testing of these sequences is of high interest for further development of the proposed approach.

4.8 Discussion

The difficulty of predicting MVC therapy outcome (Trkola, Kuhmann et al. 2002; Westby, Lewis et al. 2006; Westby, Smith-Burchnell et al. 2007), the flexibility of the virus to use other receptors (Turville, Cameron et al. 2002) and high variability of V3 loop suggest that the virus cell entry phenotype is more complex than the CCR5/CXCR4 coreceptor tropism. The study presented in this chapter is an attempt to address the shortcomings of sequence-based HIV tropism classification, the need of more accurate recognition of X4 viruses and better virus characterization for the effective use of coreceptor blockers. The study is based on high-throughput data on the single cell level. The analysis on the single cell level affords higher accuracy as compared to the approach based on averaged measurements of entire cell populations and provides additional cellular information used for the construction of the expanded picture of virus phenotype. Novel methods were developed for data analysis, visualization and modeling of virus cell entry phenotypes. The major result is the virus phenotype representation in the form of a multivariate phenotype vector expressing the determinants of virus cell entry efficiency. The multivariate phenotype vector and its position on the phenotype map provide a comprehensive picture of virus phenotype reaching beyond tropism classification and distinguishing among a spectrum of virus phenotypes. Tests of predicting phenotype vector based on sequence show that with an enhanced dataset of experimentally tested clones this approach affords practical application in predicting virus response to coreceptor blockers. A method of predicting virus phenotype vector from V3 sequence is of high interest for effective patient treatment as viruses of similar locations on the phenotype map represent also potentially similar treatment targets.

Among the tested clones we found scarcity of strictly X4 viruses. Clones selected as dubious showed an R5 phenotype, several clones selected as X4 viruses showed R5 or dual-tropic characteristics. A potential explanation for this observation is the fitness cost coming with maintaining of the X4 phenotype and its high sequence diversity. It is therefore likely that in the evolutionary process in which the X4 viruses emerge from R5 viruses, the capacity of using CCR5 is often not entirely lost. The exclusive use of CCR5 is more common among the tested clones, majority of the R5 clones show also relatively high sequence conservation. This suggests the existence of a conserved part of the phenotype map occupied by R5 viruses that are uniform in sequence and show fitness advantage over highly variable X4 viruses occupying other parts of the map.

The presented approach of phenotype inference based on sequence suggests that with more data points on the phenotype map a systematic method of clone phenotype prediction based on sequence can be developed. Approach presented in this chapter is based on shrinkage methods that allow for reducing the data dimensionality with a limited number of observations. With a larger training set, more accurate and customized models can be developed, for example targeted at characterization of dual-topic viruses representing the boundary between X4 and R5 phenotypes. Detailed description of boundary phenotypes would allow for delineation of viruses susceptible to either CCR5 or CXCR4 blockers and would be therefore of high interest for effective patient treatment with entry inhibitors.

4.9 Conclusions

The study presented in this chapter analyzes virus V3 loop interaction with a host coreceptor on a single cell level. The experimental and computational pipeline developed for this analysis is laborious however results in an insightful view of the virus phenotype involving several determinants of virus cell entry and expanding beyond the binary coreceptor tropism classification.

As opposed to high-throughput data based on recently developed deep sequencing technology, here we examine data obtained from a long-time established technology but the study relies on novel methods of information classification, visualization and analysis. The advantage of the single cell data over sequence data is the insight it offers into the host molecular and environmental parameters conditioning virus cell entry in addition to the virus sequence variability. This kind of data allows for construction of a comprehensive picture of the virus phenotype describing virus cell entry dependence on the expression of cell surface proteins and on the presence of blocking them molecules.

The cost of testing of a single virus variant presents a limitation to the developed approach. Given the variability of V3 loop a complete phenotype map of virus variants is hard to obtain, nevertheless, with an increased number of tested clones and the knowledge on the possible phenotypes and their relation to the sequence, a more targeted selection of viruses for testing can be employed that might result in a representative phenotype map without the need of analyzing a large set of clones. Such selection should aim at covering different clades of the V3 phylogenetic tree (Figure 4.2), obtaining a large variability of X4 clones and obtaining clones from sequence cluster boundaries as described in the last part of this study.

Similar to the sequence space analysis presented in the previous chapter, this study of cell entry efficiency points to the predominant proportion of R5 virus phenotypes –

several viruses predicted as X4 showed R5 or dual-tropic characteristics in our assay, while the opposite situation was never observed. This suggests a fitness advantage of the R5 phenotype over other types of viruses which results in higher preservation of this phenotype. When a yet unknown mechanism triggers change of a virus coreceptor tropism from the CCR5 usage it evolves into highly variable variants in the process involving certain fitness loss. The fact that only few of the tested clones showed a purely X4 phenotype suggests that due to the fitness advantage of using the CCR5 coreceptor the capacity of using this coreceptor is rarely completely lost in the evolutionary process of the coreceptor switch. This way the analysis of virus phenotypes on the single cell level complements the sequence study on the level of virus population. Phenotypic testing of additional virus variants characterized by sequence space location should further expand our knowledge on the potential relationship between sequence and phenotype.

This study represents a step towards building virtual phenotype – a computational model of virus biology incorporating comprehensive biological information. Virtual phenotype of the virus cell entry described in this chapter contributes to our understanding of virus tropism and to predicting of a variety of virus phenotypes reaching beyond binary tropism classification. Given the limitations of sequence-based methods of virus tropism classification, expanding models with other than sequence molecular information represents a potentially promising avenue for improving our understanding of the phenotype and its prediction.

The studies presented in this chapter and chapter 3 analyze a specific host-virus interaction on a population and single-cell levels. In spite of the high relevance of the interaction between the V3 loop and host coreceptors for the disease development and of its importance in the MVC patient treatment, this interaction represents a limited fraction of the highly complex host-virus interaction network. The short genome of retroviruses encodes only a small number of essential proteins, the virus relies strongly on the host molecular machinery for its successful replication and spread. Throughout its lifecycle HIV is involved in a multitude of molecular interactions with the host, certain of which are a part of the virus replication process, others involve host immune recognition that might counteract the infection. The molecular host-virus interactions present therefore evolutionary constrains and pressures on the virus that shape its genetic constitution, lifecycle and infection strategies. Knowledge of other host-virus interactions beyond the ones presented in the previous chapters is therefore crucial for better understanding of the virus biology, its mechanisms of infection and pathogenicity.

CHAPTER 5 – Genome scale

all known host-virus interactions among multiple host and virus species

In the previous chapter we presented a study of a specific host-virus interaction that comprised host, virus and environmental parameters. This thorough analysis of a specific interaction produced a detailed picture of virus phenotype that is of high interest for the application of therapies based on entry inhibitors. The current success of MVC-based treatment creates a good prognosis for a new generation of HIV treatment that instead of virus proteins would be targeted at host factors involved in crucial interactions for the HIV life cycle. Such type of treatment that efficiently blocks crucial host factors might prove more efficient and long-term effective since it would require the virus to evolve a new biological function instead of simply mutating the drug binding site.

HIV is involved in a multitude of molecular interactions with different host proteins throughout its life cycle. Starting from binding to the host receptors through replication to budding, the virus subjugates and exploits host cellular machinery in order to propagate. Knowledge of the mechanisms of all HIV-host interactions is of great not only scientific but also therapeutical interest. In this chapter we present a study of host-virus interactions on the broadest, genomic scale. We perform a comparative genomics analysis of all reported HIV-host interactions and point to individual ones that are potentially interesting for further study. We additionally correlate the amounts of genetic change between the interacting host and virus genes and point to the parts of host-pathogen biology that undergo accelerated evolution as opposed to those that are relatively conserved. The results of this study are published in: Bozek, K., Lengauer, T. Positive selection of HIV host factors and the evolution of lentivirus genes. *BMC Evol Biol.* 2010 Jun 18;10:186.

5.1 Background

Phylogenetic studies have shown that HIV emerged in humans through at least eleven cross-species transmission events of simian immunodeficiency virus (SIV) from non-human African primates. Three transmissions of chimpanzee SIV (SIVcpz) from the central African chimpanzee subspecies (*Pan troglodytes troglodytes)* gave rise to the HIV-1 groups M, N and O (Gao, Bailes et al. 1999), with the group M causing the AIDS pandemic. Other transmissions of sooty mangabey (*Cercocebus torquatus atys*) SIV (SIVsmm) gave rise to HIV-2 groups A-H (Hirsch, Olmsted et al. 1989 ; Gao, Yue et al. 1992) (Figure 5.1).

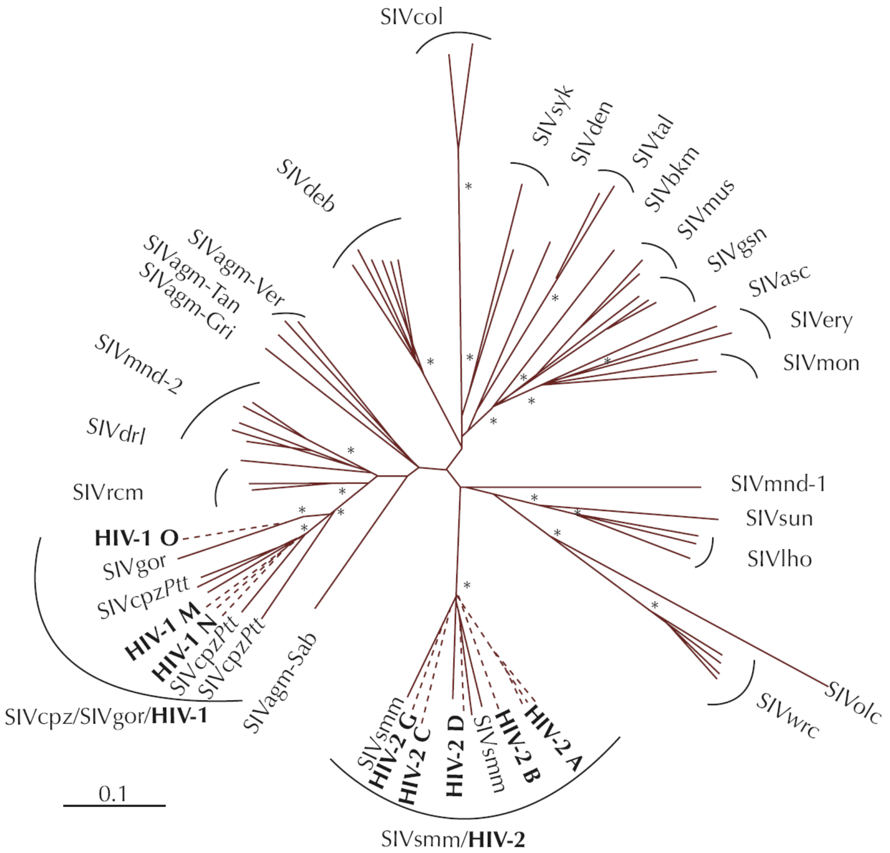


Figure 5.1 Phylogenetic tree of different SIV and HIV lineages based on neighbor joining analysis of partial pol sequences. The scale bar indicates 0.1 substitutions per site, asterisks indicate bootstrap support of >85%. The positions of different HIV-1 and HIV-2 lineages (marked in bold) interspersed with the SIVcpz/SIVgor and SIVsmm lineages respectively, indicate the origins of the human viruses. From (Van Heuverswyn and Peeters 2007).

SIV infection of African non-human primate host species (including sooty mangabeys, African green monkeys, mandrills, and several others) is non-pathogenic despite high levels of viremia (Silvestri 2005; Silvestri, Paiardini et al. 2007; Pandrea, Sodora et al. 2008). Different levels of pathogenicity of immunodeficiency viruses in their host species (Pandrea, Sodora et al. 2008; Keele, Jones et al. 2009) as well as the lack of adaptation to their non-natural species (Bogerd, Doehle et al. 2004; Jia, Serra-Moreno et al. 2009) show how interspecies differences can impact viral infectivity and drive adaptation. Host genetic differences between individuals also affect the dynamics of disease progression (Heeney, Dalgleish et al. 2006; Fellay, Ge et al. 2009). There is a growing list of genes and alleles for which there is evidence of a positive or negative effect on infection and disease progression. Among them, several host factors block or restrict retroviral infections in primates (see chapter 2): TRIM5α (Stremlau, Owens et al. 2004); APOBEC3G (Sheehy, Gaddis et al. 2002) and tetherin (BST-2, CD317) (Neil, Zang et al. 2008). These restriction factors constitute defense mechanisms of the host acting in a species-specific manner (Bogerd, Doehle et al. 2004; Jia, Serra-Moreno et al. 2009) blocking the viruses from replication in their non-natural host species and thus being potential agents of anti-HIV defense.

A feature of pathogenic HIV-1 infection that distinguishes it from non-pathogenic SIV infections is the high level of chronic immune activation associated with accelerated T cell turnover rates and apoptosis (Silvestri 2005). The basis for this difference in pathogenicity is not understood, however deciphering which viral and host factors are responsible for the nonpathogenic course of natural SIV infections could prove useful in developing more effective treatments and prevention strategies for AIDS.

Positive selection, demonstrated in part by the rapidly evolving immune system genes (Endo, Ikeo et al. 1996), reflects the evolution of the host defense against various infections. Several HIV restriction factors have been shown to be under positive selection throughout primate evolution (Sawyer, Emerman et al. 2004; Zhang and Webb 2004; Sawyer, Wu et al. 2005; McNatt, Zang et al. 2009). Due to the relatively long generation times of primate species with slow rate of genetic evolution in contrast to the short generation times of viruses with high rates of genetic evolution and the potentially recent introduction of SIV into primates (Wertheim and Worobey 2009), the impact of the selection pressures solely from SIV on the host species is likely to be negligible. However, genetic polymorphisms in genes interacting with the virus can influence traits relevant for the susceptibility to lentiviral infection and point to a potential role of a gene in infection and its contribution to disease. Comparative genomics can offer insights into disease mechanisms by correlating molecular differences that arose during primate evolution with the variation in disease susceptibility.

There is ample scientific knowledge on HIV-1 human protein interactions. The *HIV-1 Human Protein Interaction Database* is a catalogue of over 1400 human proteins that participate in approximately 3000 unique HIV-1-to-human protein interactions reported in peer reviewed scientific literature (Fu, Sanders-Beer et al. 2009). The size and scope of this database allows for large-scale analyses of HIV-host molecular interactions. Together with several fully sequenced primate genomes it allows for a systematic search for host factors under positive selection that might be relevant for infection and merit further investigation. Previous studies of positive selection in the HIV host factors focused on individual examples (Sawyer, Emerman et al. 2004; Zhang and Webb 2004; Sawyer, Wu et al. 2005; McNatt, Zang et al. 2009), as well as on a set of 140 proteins compiled from the literature (Ortiz, Guex et al. 2009). Here we analyze all 1439 genes available in the *HIV-1 Human Protein Interaction Database*.

In this study we explore genetic differences of HIV-interacting genes among primates. We perform a comparative genomics analysis of the HIV-interacting proteins in search of positively selected genes in four different primate species. The positively selected genes are next characterized by their biological function, role in the protein-protein interaction networks and interactions with the virus. We also analyze the relationship between the strength of positive selection in the host proteins with the evolutionary rates of the interacting proteins of five immunodeficiency viruses in search of patterns in the evolution of host-pathogen interactions.

5.2 Primate sequence analysis

In the first part of this study we analyzed HIV-interacting host genes. We extracted and aligned sequences of all genes reported to interact with the virus from four primate species genomes. Positive selection in the sequences of the four primates and human genes separately is next estimated. Genes under positive selection were next analyzed in terms of their interactions with the virus, functionality and role in the protein-protein interaction network.

5.2.1 Sequences

From the University of California, Santa Cruz (UCSC) *Genome Browser Database* (GBD) (Karolchik, Kuhn et al. 2008) we extracted the gene sequences of human proteins reported in the *HIV-1 Human Interaction Database* (Fu, Sanders-Beer et al. 2009) to interact with HIV-1 as well as of the respective homologs in three non-human primate species: chimpanzee (*Pan troglodytes*), orangutan (*Pongo pygmaeus abelii)* and rhesus macaque *(Macaca mulatta)*. We used all available primate species for which genome sequences were publicly available and at least partially annotated. Homologous sequences were aligned using *Threaded Blockset Aligner* (TBA) (Blanchette, Kent et al. 2004) with the human sequence as the reference and trimmed to their coding parts based on the human gene annotation. We excluded sequences of genes not identified in the human genome or in more than one non-human primate species. We also excluded sequences composed of > 50% gaps as compared to the human sequence. This filter helped to ensure that the results of the analysis are not influenced by the lack of proper homolog identification. The information on the alignment quality is provided in Table 5.1. Of the 1439 human proteins and 3643 unique interactions in the interaction database, 1182 proteins involved in 2596 unique interactions fulfilled these alignment criteria.

	all	chimp	orangutan	macaque
removed gene sequences [%]	18.1	7.8	14.8	14.2
mean gap ratio	0.0183	0.0234	0.0241	0.0284
sliding window score - gap ratio correlation	0.0101(0.293)	0.0021(0.473)	-0.0196(0.829)	0.0292(0.07)
site-based score - gap ratio correlation	-0.0977(0.962)	-0.1123(0.984)	-0.1402(0.999)	0.0107(0.406)

Table 5.1 Alignment quality and independence of the positive selection score from the number of gaps in a sequence. Genes with no sequence found in the human genome were excluded from the analysis as were genes with more than one unidentified homolog among the three non-human primate species. Alignment of homologous sequences containing more than 50% gaps as compared to the human gene sequence were also excluded from the tests for positive selection. The first row of the table lists the percentage of the sequences excluded according to these criteria. The second row shows the quality of the alignment of the remaining sequences in terms of the mean ratio of the number of gaps over the length of a gene sequence. The last two rows show the correlation between the quality of alignment and the result of the positive selection tests. It appears that high scores of positive selection are not correlated with poor sequence alignment. P-values are given in parentheses.

5.2.2 Positive selection in primate species

Next, we performed the search of the host factors under positive selection among the 1439 host factors reported in the *HIV-1-Human Protein Interaction Database*. We used *Bayes Empirical Bayes* (BEB) approach for inference of amino acids under positive selection (Yang, Wong et al. 2005) implemented in the PAML package (Yang 2007). This approach uses a statistical distribution to describe the variation of the dN/dS ratio (non-synonymous to synonymous nucleotide substitution ratio) among sites and a likelihood ratio test (LRT) to compare two distributions: a distribution that allows a subset of sites to have dN/dS > 1 and a null model that does not. If the result of the LRT is statistically significant then one can infer a gene to be under positive selection. An empirical Bayes test is then used to calculate the probability that a site belongs to the class of sites whose dN/dS ratio is larger than 1. We applied two LRTs implemented in the PAML package: M1a (NearlyNeutral) to M2a (PositiveSelection) comparison and M7 (beta) to M8 (beta&ω) comparison retaining the results of the higher LRT. The first LRT compares the NearlyNeutral null model which assumes two site classes one with 0 < dN/dS < 1 and one with dN/dS = 1 to the alternative model PositiveSelection which adds a site class of dN/dS > 1. The second LRT compares the beta null model which assumes a beta distribution for dN/dS in the interval (0,1) to the alternative model beta&ω which adds a site class with dN/dS > 1. In order to assess the amount of positive selection in the HIV-interacting genes and to rank the genes accordingly, we established a *site-based score*. This score is based on the weighted sum of sites with dN/dS > 1 normalized by the sequence length. The sites are weighted according to the calculated probability P of a site being under positive selection: by a factor of 3 for the sites with P > 0.99, a factor of 2 for the sites with P > 0.95, other sites by a factor of 1.

In order to assess the robustness of, and provide additional support for, the site-based score we compared it with a *sliding window score*. The sliding window score was based on the dN/dS ratio averaged over each sliding window across the protein gene sequence. We used the method of estimating the dN/dS substitution rates of entire sequences by Yang and Nielsen (Yang and Nielsen 2000) also from the PAML package (Yang 2007). The calculation of the dN/dS ratio was done in windows of 150 base-pair length (50 amino acids) slid along the sequence by a step of 30 base-pairs (10 amino acids). This test facilitates the localization of regions with a high dN/dS ratio in gene sequences rather than specific sites returned by the BEB approach.

While there are potential false positives associated with the LRT some genes might be missed by the sliding window approach due to a small number of positively selected sites. We considered the LRT approach to be more relevant for long and highly conserved primate sequences and therefore used the more stringent LRT significance as the indicator of positive selection, the sliding window measure was used as a supporting score. We used both tests to search for positive selection in all four primate species and subsequently in the human and chimpanzee gene sequences separately.

The site-based search for genes under positive selection returned 152 genes having sites under positive selection among all four primate species and in 97 genes having sites under positive selection in the human-chimp comparison with an overlap of 49 genes. The genes were next ranked according to their site-based and sliding window scores. The ranks of genes obtained with the sliding window score correlate with the site-based score ranks with a correlation coefficient of 0.65 in all species and 0.52 in the human-chimp comparison ($p < 0.01$). This positive correlation of ranks obtained with two different methods, together with the high scores assigned to proteins reported to be under positive selection in other studies (APOBEC3G (Sawyer, Emerman et al. 2004; Zhang and Webb 2004), TRIM5α (Sawyer, Wu et al. 2005) – Figure 5.2, left panel) suggests that the ranks used in further analyses are robust with respect to the scoring method.

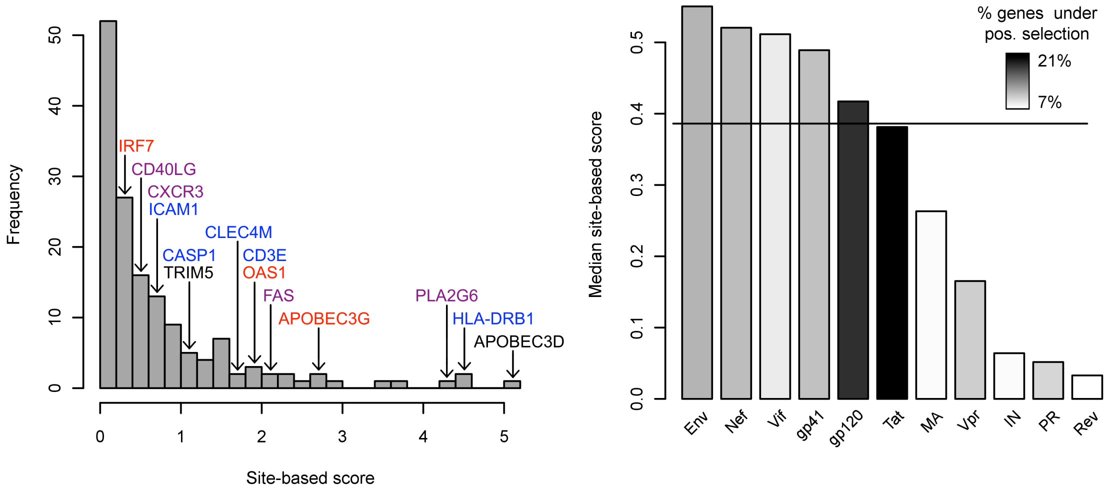


Figure 5.2 Distribution of the site-based score in the genes of four primate species. Left panel shows the distribution of all gene scores. Positions of important host factors are indicated. Colors indicate the functional group to which the factors belong: blue - membrane-related proteins, red – innate immune response proteins, magenta – both, black – none of the groups. Right panel shows median values of site-based scores in the interaction grouping. Shown are groups of the size > 2% of all interactions. Bar colors indicate the percentage of the genes under positive selection in each group as inferred by LRT (listed in Table 5.2). CA protein (not shown on the figure) has less that 1% interactions reported in the dataset showing a high median of 1.3.

The recently identified host restriction factor tetherin (BST-2, CD317) (Neil, Zang et al. 2008) was not included in the *HIV-1 Human Interaction Database*. We separately extracted, aligned and performed the positive selection tests on the tetherin sequences of the four primate species. Even though positive selection in the primate tetherins has been reported before (McNatt, Zang et al. 2009), the site-based approach did not result in significant LRT. However the sliding window test showed this protein to be under positive selection with a rank of 62 among the full list of 1182 genes (APOBEC3G rank 6, TRIM5α - 38).

In order to inspect the distribution of positive selection scores in subsets of the HIV-interacting genes we established an *interaction grouping* based on with which viral protein the host proteins interact. The interaction grouping showed variation in the distributions of site-based scores as well as in the ratio of positively selected host proteins among groups (Figure 5.2, right panel). Permutation tests revealed a significantly lower mean of the ranks based on site-based scores of host genes interacting with gag protein and a comparatively higher mean for integrase (IN), protease (PR), vpr and rev proteins (Table 5.2). The ranks based on sliding window scores were additionally significantly lower for the Env-, gp120-, gp41- and CA-interacting genes and higher for Vif- matrix (MA)- and nucleocapsid (NC)-interacting genes. The site-based ranking was limited to the 152 genes under positive selection. The discrepancies of the significance of mean ranks of gene groups between the two scorings were due to the differing numbers of genes included in both rankings. The mean group ranks in both scorings are positively correlated with r = 0.52 (p ≈ 0.07). If only the 152 genes under positive selection are considered the correlation increases substantially to r = 0.92 (p < 0.05).

virus gene	rank	interacting genes	% genes under pos. sel.	mean site-based rank	mean sliding window rank
gp120	1	434	12.9	70.11	489.1
Vpu	2	22	18.2	45.75	529.2
Env	3	145	15.9	69.52	505.7
p6	4	12	16.7	40.50	692.8
Rev	5	59	6.8	113.75	770.1
gp41	6	123	16.3	66.40	496.3
Nef	7	168	12.5	67.14	534.4
Gag	8	45	13.3	17.00	625.0
Vif	9	55	16.4	54.56	670.1
Tat	10	636	11.0	80.81	616.0
MA	11	69	7.2	84.00	754.0
NC	12	19	15.8	64.67	772.0
Vpr	13	152	11.2	100.06	624.4
RT	14	33	9.1	93.67	631.6
Pol	15	1	100.0	30.00	96.0
CA	16	21	28.6	37.00	379.2
PR	17	71	21.1	109.47	639.8
IN	18	66	7.6	112.60	738.9

Table 5.2 Ranking of viral proteins and characterization of interacting host genes. Positively selected host genes are inferred from the significance of the LRT. Colors indicate significantly high (red) and low (blue) mean ranks or ratios of interacting genes under positive selection.

5.2.3 Positive selection in humans

In order to additionally estimate positive selection that acted along the more recent timescale after the primate interspecies split we searched for single nucleotide polymorphisms (SNPs) in the human gene sequences. We used the haplotype map HapMap Phase II (Sabeti, Varilly et al. 2007) to search for the SNPs located in the gene regions (both introns and exons) of the analyzed HIV-interacting human proteins (Fu, Sanders-Beer et al. 2009). We then used the integrated haplotype score (iHS) introduced by Voight et al. (Voight, Kudaravalli et al. 2006) to identify those SNPs that might have emerged as a result of positive selection. The iHS score is based on the idea that a favored allele increases its frequency in a population fast and therefore tends to be located in unusually long haplotypes of low diversity in contrast to the more variable haplotypes of the unselected background. Therefore the decay of identity of a haplotype as a function of distance from a given allele reflects the strength of positive selection acting on the allele. High levels of haplotype homozygocity extending much farther than expected under neutral model are an evidence for selection acting at or near the SNPs in the allele. The cutoff of |iHS| ≥ 2 was used to choose SNPs under positive selection.

Out of 1335 SNPs in the analyzed genes 88 showed |iHS| ≥ 2, slightly more (6.6%) than expected from the standard normal distribution of the iHS. For each gene we calculated the number of positively selected SNPs, and used their presence as evidence of positive selection acting on a gene in the human lineage. We found 61 genes containing positively selected SNPs with a maximum number of 5 SNPs in a gene (BRCA1) and the majority of genes having only one positively selected SNP (43 out of 61). Among genes found to be under positive selection using site-based method in the four primates and in the human-chimp comparison, 10 and 4 genes, respectively, were found to contain positively selected SNPs. We found no significant correlation between the presence of SNPs and positive selection in primates.

5.2.4 Gene Ontology

Subsequently, we searched for Gene Ontology (GO) (Ashburner, Ball et al. 2000) terms enriched among the positively selected HIV-interacting genes. GO separates biological roles performed by genes of different organisms into three separate ontologies: *biological process*, *molecular function* and *cellular component*, each organized in a hierarchical manner with more general terms preceding more specific terms in the GO graph. We used the R package topGO (Alexa, Rahnenführer et al. 2006) to score GO terms by their overrepresentation in groups of genes. The method makes use of the hierarchical structure of the GO by first grouping genes related to each of the terms in the GO graph and then processing the nodes bottom-up. Iteratively genes annotated to significant GO terms are removed from more general parent terms to test how enriched a node is if the genes from its children nodes are not considered. We applied two statistical tests, Fisher’s exact test (FET) and Kolmogorov-Smirnov test (KS), to test for the overrepresentation of terms in the three groups of positively selected genes: in all four primates, in human-chimp comparison and human only as inferred from the SNPs analysis resulting in six enrichment tests in total. The FET and KS test for two different aspects of enrichment – enrichment of genes under positive selection and enrichment among high-ranking genes based on a positive selection score. The GO annotation of the genes from the *HIV-1-Human Protein Interaction Database* has been previously reported (Ptak, Fu et al. 2008), the goal of this analysis was to find which terms were specific for the three groups of genes being under positive selection as compared to the full set of HIV-interacting genes. We therefore assessed term enrichment among genes under positive selection using the full set of HIV-interacting genes as a control. This helps to ensure that the significance of a term is not due to the general abundance of genes assigned to this term in the full set of HIV-interacting genes. Here we discuss selected terms that were significantly enriched ($p < 0.05$) in at least three out of the six tests.

In the *biological process* ontology several terms related to immune response were found to be enriched among the positively selected genes (e.g. “antigen processing and presentation”, “immune response” – Figure 5.3, left panel) and several immune response terms were enriched among the high-ranking genes in all three groups of genes (e.g. “innate immune response” and “defense response to virus”).

Among the terms of the *cellular component* ontology we found evidence of genes related to the cellular membrane being under positive selection. Eight out of 14 terms of this ontology that were overrepresented in at least four out of six enrichment tests were associated with the cellular membrane (Figure 5.3, right panel). The MHC protein complex group of terms in this gene ontology also appeared to be enriched among positively selected HIV-interacting genes.

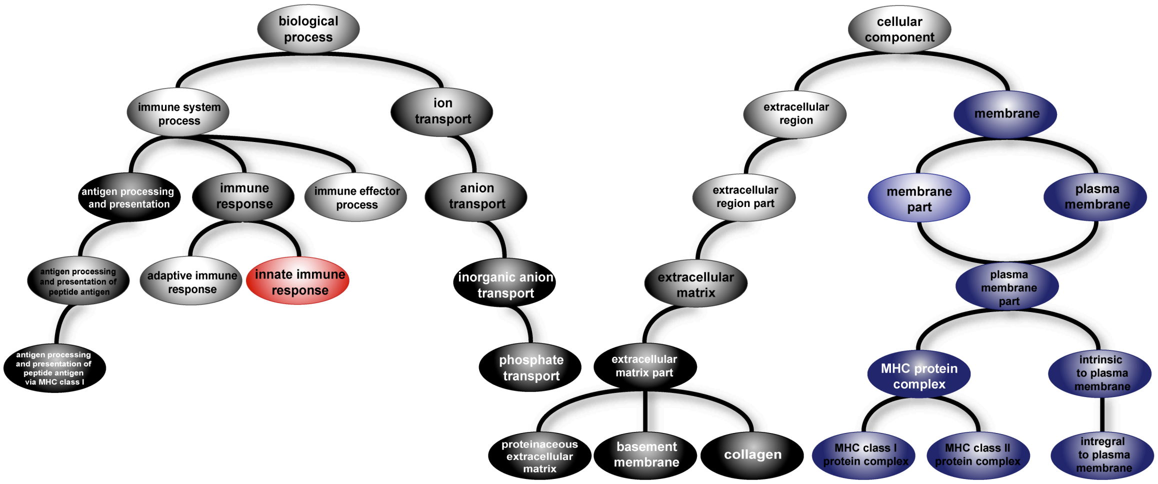

biological process
immune system process
ion transport
antigen processing and presentation
immune response
immune effector process
anion transport
antigen processing and presentation of peptide antigen
adaptive immune response
innate immune response
inorganic anion transport
antigen processing and presentation of peptide antigen via MHC class I
phosphate transport
cellular component
extracellular region
membrane
extracellular region part
membrane part
plasma membrane
extracellular matrix
plasma membrane part
extracellular matrix part
MHC protein complex
intrinsic to plasma membrane
proteinaceous extracellular matrix
basement membrane
collagen
MHC class I protein complex
MHC class II protein complex
intregral to plasma membrane

Figure 5.3 GO graphs of terms of the biological process (left) and cellular component (right) gene ontologies enriched among the positively selected HIV-interacting genes. Brightness of the nodes indicates the number of enrichment tests in which a term was found to be overrepresented ($p < 0.05$), ranging from three (light nodes) to five (darkest nodes) significant results out of six enrichment tests. The red node represents the innate immune response term, blue nodes represent membrane-related terms that are discussed in detail in the text.

In the *molecular function* ontology the terms enriched among the positively selected genes are predominantly related to receptor or signaling activity, e.g. “receptor activity”, “MHC class I and II activity”, “transmembrane receptor activity”, “signal transducer activity”.

Given the high enrichment of the terms “membrane” and “innate immune response” we further analyzed genes in those two functional groups. To identify the first group we extracted all genes annotated with the GO term GO:0016020 (“membrane” from the *cellular component* ontology) and its subterms. To identify the innate immune response genes we performed a two-step procedure. First, we extracted all genes annotated with the GO term GO:0045087 (“innate immune response” from the *biological process* ontology) and its subterms and then we searched the dataset for all proteins known to participate in the innate immune response (Katze, Fornek et al. 2008; Sadler and Williams 2008).

We found the membrane-related proteins to be overrepresented among the positively selected genes in all four primates (chi-square test $p < 0.05$) and innate immune response to be overrepresented in all three groups of positively selected genes (chi-square test $p <0.001$).

5.2.5 Protein-protein interaction networks

To characterize the importance of the HIV-interacting proteins for the host and those among them being under positive selection, we analyzed their position in the network of human protein-protein interactions (PPIs) and human-pathogen PPIs, based on the analysis of human-pathogen interactions for 190 pathogens, both bacterial and viral, merged into 54 pathogen groups based on taxonomic similarity (Dyer, Murali et al. 2008). Dyer et al. reported that both viral and bacterial proteins preferentially interact with human proteins that are either *hubs* – i.e., those involved in many interactions, or *bottlenecks* – i.e., those central to many human PPI network pathways suggesting that pathogens may have evolved to interact with proteins controlling critical processes in the host cell as a mechanism of disrupting the key elements of the host cellular machinery. We inspected the distribution of the local connectivity, centrality, number of interacting pathogens, and number of pathogen groups for the host proteins in the interaction grouping. Local connectivity of a protein is defined as the number of human PPIs in which it participates; centrality is the fraction of shortest paths in human PPI network between all protein pairs that pass through the given protein. High centrality is characteristic of a bottleneck in an interaction network, high connectivity – of a hub.

	group	proteins	local connectivity	centrality	pathogens	pathogen groups
Pol-encoded	PR	27	<0.01	<0.01	-	-
	RT	20	0.01	<0.01	-	-
	IN	49	<0.01	-	-	<0.01
	Nef	97	-	0.03	-	-
	Rev	42	<0.01	0.01	<0.01	<0.01
	Tat	455	-	-	-	0.01
	Vif	45	<0.01	-	-	<0.01
	Vpr	98	<0.01	0.03	0.03	0.01
	Vpu	20	-	<0.01	0.02	0.01
Gag-encoded	MA	54	-	-	-	<0.01
	CA	18	0.03	-	-	-
	NC	16	-	-	<0.01	-
	p6	8	0.01	<0.01	<0.01	<0.01
	Gag	26	0.01	-	0.01	<0.01
Env-encoded	gp120	175	-	-	-	-
	gp41	61	<0.01	-	0.01	<0.01
	Env	57	0.04	-	-	-

Table 5.3 Interaction specificities of HIV-related host factors in the interaction grouping. Significant p-values for local connectivity, centrality, number of interacting pathogens and pathogen groups are shown in red for high values and in blue for low values.

We found that IN, p6, PR, Rev, reverse transcriptase (RT), Vif, Vpr interact with host proteins of a significantly high local connectivity. p6 and Rev interact with proteins that are also highly central (Table 5.3). In contrast, gp41, Env, CA and Gag seem to interact with host proteins of a significantly lower local connectivity.

Given the low local connectivity of envelope-interacting host proteins, we investigated the distributions of the local connectivity, centrality, number of interacting pathogens and pathogen groups of the HIV-interacting proteins annotated with membrane-related GO terms. These proteins were of 0.73-fold lower degree ($p < 0.01$), interact with 0.69-fold lower number of pathogens ($p < 0.01$) and 0.93-fold lower number pathogen groups ($p < 0.03$).

All of the positively selected genes in the four primates tend to be less connected (0.76-fold difference, $p \approx 0.05$) and less centrally located (0.74-fold difference, not significant, $p \approx 0.16$) in the human PPI than the genes that show no positive selection. No clear patterns were observed in the number of interacting pathogens and pathogen groups or in the positively selected genes in the human-chimp comparison.

5.3 Virus sequence analysis

In order to further characterize the host proteins interacting with the virus, we analyzed the evolution of corresponding virus genes of several virus species. We aligned 175 genomes of five species of HIVs and SIVs and the genomes of the HIV-1 and SIVcpz only. Viral gene sequences extracted from the alignments were ranked according to their *relative evolutionary rates*, a measure of genetic variability in the genes based on their phylogenetic distances corresponding to the positive selection measure in the primate genes.

5.3.1 Viral sequences

We searched the Los Alamos HIV sequence database (http://www.hiv.lanl.gov/) for complete genomes of the following viral species: HIV-1 and HIV-2, SIVcpz, rhesus macaque SIV (SIVmac) and SIVsmm. Since our primary interest was in estimating general patterns in large data sets rather than in the analysis of individual genomes, we excluded SIV species for which less than five genomes were available. In order to minimize the bias associated with the overrepresentation in the public databases of the HIVs and to obtain the highest variability of the HIV-1 genomes with approximately equal distribution among groups, we additionally filtered the HIV-1 group M genomes by retaining only therapy-naïve patient sequences, only one sequence annotated with both the same country and year and only one sequence per patient. In case of several sequences of the same year-country or patient category the longest genome sequence was selected. This filter was applied only to the HIV-1 group M sequences. All available HIV-1 group N and O complete genome sequences were kept as well as the reference HIV-1 sequence HXB2-LAI-HXB2R (accession number NC_001802). We additionally included one available *Colobus guereza* SIV (SIVcol) genome in the dataset (GenBank AF301156) as an outgroup genome for further analyses. From the complete genomes we extracted individual protein sequences according to the original genome annotation. We aligned these gene sequences using MUSCLE (Edgar 2004) and removed those that contained more than 50% insertions or deletions as compared to the reference virus HXB2-LAI-HXB2R gene sequence. We refer to the alignment of all viruses the *full alignment*. We computed an additional alignment of the HIV-1 and SIVcpz viruses only (*HIV-1/SIVcpz alignment*). The compiled dataset contained 76 HIV-1 genomes out of which 54 were M, 16 of type O and 6 of type N, 30 HIV-2, 19 SIVcpz, 36 SIVmac and 14 SIVsmm genomes. The filter applied to the HIV-1 group M genomes resulted in a representative diversity of the subtypes, with 35% subtype B sequences, 17% C, 8% A, one sequence of the D, E, F subtypes and 33% circulating recombinant forms of subtypes A to G. After gene extraction, alignment and filtering there were, on average, 160 sequences of each protein in the full alignment (90 in the HIV-1/SIVcpz alignment) approximately 45% of which are HIV-1 sequences (80% in the HIV-1/SIVcpz alignment).

5.3.2 Evolutionary rates

Unlike primate gene sequences, lentiviral genomes are highly variable, relatively short and contain overlapping reading frames, e.g Tat protein has overlapping coding regions in different frames. Calculation of its dN/dS substitution rates can be therefore strongly biased.

Different methods of estimating divergence in HIV genes have been previously applied. Several studies used the dN/dS ratio estimates within different HIV genes in order to quantify the positive selection related to the virus immune evasion (Simmonds, Balfe et al. 1990). Interpretation of the averaged values of dN/dS over gene regions is complicated as positive selection can be limited to narrow immunogenic domains within proteins, and average values of these ratios for full genes can be misleading. Many forces concurrently influence the evolutionary pattern of the virus in vivo, such as counterbalancing influences of retention of protein structure and function. Hence, other approaches estimating these parameters for each site were also used (Zanotto, Kallas et al. 1999; Yamaguchi-Kabata and Gojobori 2000; Yang and Nielsen 2000).

Given the difficulty of estimating precise selection pressures acting on highly variable lentivirus genomes exposed to complex evolutionary pressures, instead of using classical methods of estimating positive selection we developed a surrogate measure of *relative evolutionary rate*. The aim of this measure was to quantify the relative accumulation of genetic change of each virus gene without indicating specific sites or evolutionary forces that might have acted on each of the gene.

Across all host species, each protein of the immunodeficiency virus has the same specific biological role in the viral life cycle and host species adaptation – a role that necessitates interaction with a set of host proteins. These roles and the resulting interactions present specific constraints and selection pressures on viral genes that contribute to the accumulation of genetic change with a rate characteristic to each viral protein. Even though the date of the introduction of the viral species into its respective host species and the duration of the infections in each host individual are unknown, the common functionality of a viral protein in different host species determines its ability to evolve at a specific rate. Therefore, we assumed that the genetic change in the viral gene relative to the genetic change in a reference gene in the same viral genome is similar among different viruses (Muller-Trutwin, Corbet et al. 1996). Due to its role each protein shows itself specific patterns and rates of evolution. Using this assumption we introduced the relative evolutionary rate measure and used it to assess the rate of accumulation of genetic change in different viral protein sequences independently of the host species and the time of infection. This measure affords ranking viral genes according to the amount of genetic change accumulated among viral species.

We based the measure of relative evolutionary rate on maximum likelihood (ML) trees of nucleotide sequences estimated using the dnaml program (Felsenstein and Churchill 1996), part of the PHYLIP package (Felsenstein 2005). Due to the potential errors in dating of the SIV sequences and the difficulty of estimating evolutionary parameters with incomplete data from highly variable viral populations (Muller-Trutwin, Corbet et al. 1996; Rey-Cuille, Berthier et al. 1998), more advanced phylogenetic methods (Drummond and Rambaut 2007) were not applied. The trees constructed using the dnaml program (Felsenstein and Churchill 1996) are based on a Hidden Markov Model (HMM) inferring different rates of evolution at each alignment site. We used the gamma distribution for approximating the distribution of evolutionary change at different sites with shape parameter 1 and nine rate categories. These settings were selected because they resulted in the trees with the highest likelihood in several tests over a range of parameters on different viral genes.

We inferred a phylogenetic tree for each viral gene separately. Corresponding genes of the SIVcol genome were used as an outgroup in each of the trees. The distance between two gene sequences was defined as the sum of branch lengths between the nodes representing those sequences in the ML tree. Branches with low significance ($p > 0.05$) were excluded; pairs of sequences having such a branch between them were excluded from the calculation. Since the dataset of viral sequences was based on complete viral genomes, each of the gene trees contained corresponding sequences for each virus. Next, we calculated the relative evolutionary rate of each single sequence pair by dividing the distance between the two sequences of a viral gene by the distance between the reference gene sequences for the same pair of viruses. IN was chosen as the reference gene because its phylogenetic tree had the shortest branches on average resulting in mean relative evolutionary rates > 1 for all other genes.

To account for the differing numbers of sequences of each viral species in the dataset we introduced a weighting scheme for the sequence distances to reduce the bias in the mean evolutionary rate due to the overrepresentation of the HIV-1 sequences. The weighting resulted in only minor changes in the viral gene ranking that did not influence the overall results. Thus, we chose the more parsimonious approach of not weighting.

We first inspected the capacity of the relative evolutionary rate measure to quantify the amounts of genetic change within among different virus genes. We defined *diversity* and *divergence* as the within and between species variation respectively. Diversity and divergence of a gene is based the distances between all pairs of sequences of the same and different viral species respectively. Examples of diversity and divergence values distributions are shown in Figure 5.4.

Divergence distributions of three example proteins gp120, Nef and PR picture the differences among the divergence of a fast, medium and slow evolving proteins (Figure 5.4, upper panel). Gp120 shows the highest average divergence among virus species, about four times higher than the reference IN gene. Protease (PR) gene shows lower and less widely distributed divergence values. Notably, PR is an essential protein for the HIV lifecycle – it cleaves newly synthesized polyproteins to create the mature protein components of an infectious HIV virion (Kräusslich, Ingraham et al. 1989). Nef is an accessory protein, known to modulate virus pathogenicity (Kirchhoff 2009) and shows intermediate divergence between PR and gp120.

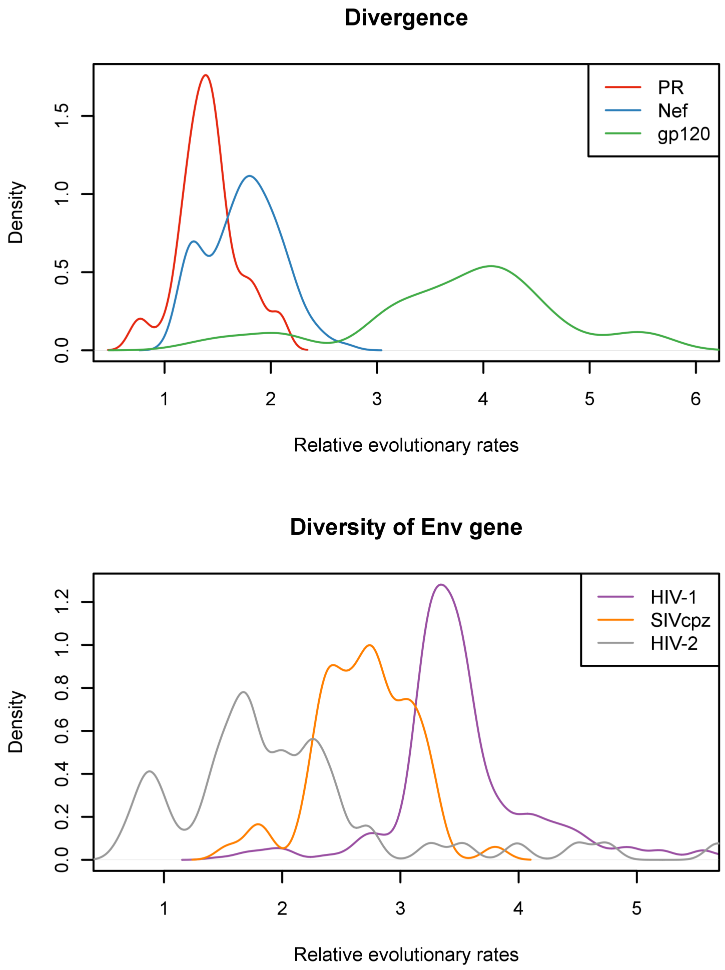


Figure 5.4 Relative evolutionary rates of virus genes. Top panel shows the density estimation of the divergence values of three example proteins in the full alignment. Bottom panel shows the density estimation of diversity values of the envelope gene in three example virus species. Densities are estimated using Gaussian kernel.

Diversity values distribution affords means of distinguishing different levels of gene variability among virus species (Figure 5.4, bottom panel). These differences are however harder to interpret and might be biased by a limited number of sequences of certain virus species.

The final viral gene ranking is based on the average of the relative evolutionary rates over all pairs of sequences (Table 5.2). In order to determine how this ranking depends on the distance measure, we recalculated the relative evolutionary rates based on Hamming distances. The ranks based on distances inferred from phylogenetic trees correlated with the rates based on Hamming distance ($r = 0.93$, $p < 0.001$ in the alignment of all viruses and $r = 0.94$, $p < 0.001$ in the HIV-1/SIVcpz alignment). The main goal of the ranking was to compare the relative variability of the viral genes in order to assess the correlation with the positive selection of the interacting host genes. Regardless of the distance measure used in both alignments, Env and its two subproteins together with Vpu and Rev were among the top 6 in the protein ranking. Pol and Pol-encoded proteins (RT, PR, IN) together with CA were among the 5 lowest ranked proteins. The only discordance among the distance measures and alignments were observed among the middle-ranked genes (Nef, Gag, Vif, Tat, MA, NC, Vpr, RT).

5.4 Gene rank correlation

To investigate the relative rates of evolution among the interacting host and virus genes we next calculated the correlation between the strength of positive selection acting on the host factors and the rate of evolution of the interacting viral genes. We calculated the correlation of the host factor ranks assigned according to the site-based and the sliding window scores and the viral gene ranks assigned according to the relative evolutionary rate measure. In the site-based ranking we restricted the host factors to the ones under positive selection as inferred from the LRT. Both the viral and host gene rankings were normalized to the [0,1] range. The same procedure was repeated on the positively selected genes inferred in the human-chimp comparison and correlated to the viral gene ranking based on the HIV-1/SIVcpz alignment.

The ranks of interacting host and viral proteins showed only minor but significant ($p < 0.01$) positive correlation ($r = 0.18$ in the site-based and $r = 0.17$ in the sliding window scoring). In the human-chimp comparison the sliding window ranking produced the only significant correlation ($r = 0.14$, $p < 0.01$).

5.4.1 Interaction binning test

Given the small but significant correlation of the ranks of interacting host and virus proteins we devised an *interaction binning* test to investigate the proximate relationships between interacting gene ranks. The interactions were grouped into bins defined by viral gene ranking. The ranks of host genes within bins were averaged. Bins of different sizes were slid along the viral gene ranking scale, advancing from one gene to adjacently ranked gene in each step. For each bin size we calculated the correlation between viral gene ranks and the averaged host gene ranks. Because of the small number of viral genes (18) a symmetrical approach of averaging over viral gene ranks was not performed. We tested all 18 possible bin sizes and used the permutation procedures described below to test for the significance of the correlation obtained for the averaged binned ranks of each bin size.

The test of all 18 bin sizes showed a markedly increased correlation of approximate ranks of interacting viral and host genes (Figure 5.5, left panel). For example, binning interactions over a bin of size 0.083, that averages host gene ranks interacting with 3 viral genes neighboring in the viral gene ranking increases substantially the correlation to r = 0.78 in the site-based and r = 0.92 in the sliding window scoring (Figure 5.5, right panel). Permutation tests showed the significant correlations (p ≤ 0.05) in 14/18 bin sizes in the site-based ranking and in 13/18 in the sliding window ranking. In the human-chimp comparison only the correlation based on the sliding window score was significant for 16/18 bin sizes, with the correlation of 0.93 on average and of 0.88 for the 0.083 bin size. The high correlations obtained after binning suggest that the viral genes of different evolutionary rates tend to interact with host factors under commensurate levels of positive selection.

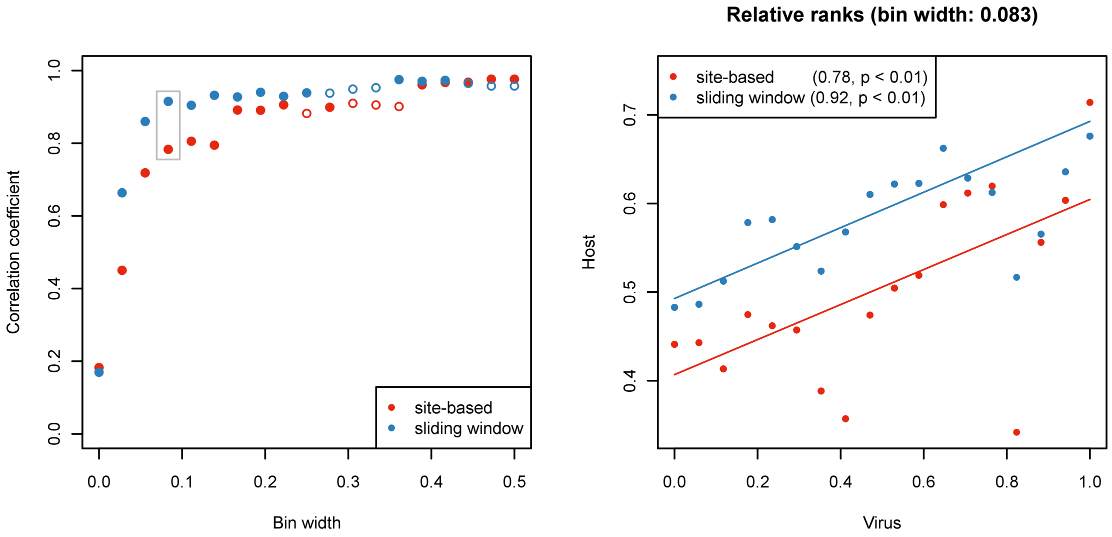


Figure 5.5 Correlation of the normalized ranks of the interacting host and viral proteins obtained from site-based and sliding window scoring of the host genes and relative evolutionary rates of the viral genes in the comparison of the all four primate species and the alignment of all viruses. Left panel shows correlation coefficients of the ranks of interacting proteins averaged by bin size. Significant correlation coefficients (p ≤ 0.05) are indicated by filled dots, the gray box indicates correlation coefficients of ranks plotted in the right panel. Right panel shows a plot of viral and host binned ranks. The correlation coefficients obtained for each host gene scoring is indicated in the brackets, lines are fitted according to least sum of squares.

5.4.2 Permutation procedures

In order to assess the statistical significance of results of this study we developed permutation tests of the HIV-human interactions. The HIV-human interaction data can be represented by a bipartite graph with nodes representing host and viral proteins and edges connecting interacting host and viral proteins. We designed two procedures of permuting the host-virus interaction network. The *host-oriented* test consists of retaining the degree of each of the host gene nodes in the network and randomly sampling a corresponding number of interacting viral genes from the set of all viral genes. The *virus-oriented* test consists of retaining the degree of each viral gene node and randomly sampling a corresponding number of interacting host genes. Performing two different permutation tests allowed for testing if certain results were due to the differing numbers of interactions reported for different host and viral proteins. We developed additional permutation tests allowing for random node degrees in the network and found the permutation tests conserving aspects of the network topology to be more stringent in assessing the statistical significance of the observations in this study. We therefore used the host- and virus-oriented tests to assess statistical significance of the results of the presented analyses.

5.5 Discussion

Understanding the genome-wide selection pressure on HIV-interacting proteins can provide insights into the evolutionary dynamics of host factors, the genetic basis of differences between nonpathogenic and pathogenic lentivirus infection and the roles of individual genes in host-pathogen interaction and immunopathogenesis. Only two of the host species analyzed in this study are naturally infected with HIV/SIV (human and chimpanzee) moreover because these viruses have been recently introduced into their host species (Keele, Jones et al. 2009) the selection pressures observed in this study are not driven by modern lentiviruses. Even though the interaction data used is mainly human and HIV-1 specific, the majority of interactions are shared with SIVs and with other viruses. Many of the host factors found to be under positive selection interact with multiple pathogens so the selection pressures on the host factors are likely to be driven by those pathogens. However the comprehensive HIV interaction data offers opportunity for a broad study of the evolution of host-pathogen interactions. The interactions in the *HIV-1 Human Protein Interaction Database* are catalogued manually based on a literature screen and cannot be considered as fully validated. Nevertheless this screen for positive selection points to a narrow set of potentially interesting interactions that can be examined and validated individually.

In this analysis we identify ~10% of the 1439 HIV-interacting genes as being under positive selection based on LRT of sites under positive selection in four primate species. Ortiz et al. (Ortiz, Guex et al. 2009) reported a similar fraction of genes as being under positive selection based on the analysis of 140 genes. Among the 62 proteins analyzed in both this study and by Ortiz et al. 13 are identified as being under positive selection in the Ortiz et al. study. Among these, 11 are confirmed in my study either by showing significant LRT or being within the upper quintile of the sliding window scores. The reasons for discrepancies between the two studies might lie in different criteria of which genes to analyze and an expanded number of species for the analysis of individual genes in the Ortiz et al. study.

Three screens using small interfering RNA (siRNA) have been reported (Brass, Dykxhoorn et al. 2008; König, Zhou et al. 2008; Zhou, Xu et al. 2008) that search human genes having effect on HIV infection – so called HIV dependency factors (HDFs). The overlap among the HDFs identified in the three studies and the proteins in the HIV interaction database is known to be small (Bushman, Malani et al. 2009). We scanned the results of one of the siRNA screening studies (Brass, Dykxhoorn et al. 2008) for the presence of genes under positive selection detected in my analysis. Among 32 genes common between this and Brass et al. studies only one protein (SP110) appeared to be under positive selection.

Several HIV restriction factors have been shown to be under positive selection throughout primate evolution (Sawyer, Emerman et al. 2004; Zhang and Webb 2004; Sawyer, Wu et al. 2005; McNatt, Zang et al. 2009). TRIM5α, discussed in detail in chapter 2, restricts post-entry activities of the retroviral CAs in a species-specific manner (Stremlau, Owens et al. 2004; Towers 2005), and has apparently undergone multiple episodes of positive selection that predate the estimated origin of primate lentiviruses (Sawyer, Wu et al. 2005). The genetic changes that emerged as an effect of selection pressures by previous viral epidemics resulted however in a species-specific restriction of modern lentiviruses. Another type of restriction factors, cytidine deaminase enzymes APOBEC3G and APOBEC3F block the virus at a later than TRIM5α stage of reverse transcription and are packaged into nascent virions. The APOBEC family in primates consists of nine cytosine deaminases (cystosine and uracil) and two others that possess in vivo DNA editing functions (Sheehy, Gaddis et al. 2002; Turelli and Trono 2005). In the absence of the lentivirus accessory gene virion infectivity factor (Vif), APOBEC3G becomes incorporated into nascent virions and inhibits HIV activity by causing hypermutations that are incompatible with further replication. Thus, APOBEC3G and Vif are under selection to decrease and enhance, respectively, their interaction with one another, a genetic conflict resulting in the rapid fixation of mutations commonly seen in host-pathogen interactions. As with the primate TRIM5α family, APOBEC3G activity shows species-specific adaptations (Sawyer, Emerman et al. 2004) emphasizing that coevolution of lentiviruses was a prerequisite for adaptation to a new host after cross-species transmission (Bogerd, Doehle et al. 2004). Thus, similar to TRIM5α, APOBEC3G clearly possessed an ancient role in defence against RNA viruses, a function that predates estimates of the emergence of today's primate lentiviruses, that resulted in species-specific restriction of current RNA viral infections (Sawyer, Emerman et al. 2004; Zhang and Webb 2004). Both TRIM5α and APOBEC genes are reported in the study presented here to be under positive selection (Figure 5.2), which suggests that this large scale approach can point to factors relevant for the immunodeficiency virus infection.

Several host factors not identified as being under positive selection in the LRT showed an elevated dN/dS ratio in the sliding window test. For example, tetherin, previously reported to be under positive selection (McNatt, Zang et al. 2009), did not show significant LRT but ranked among the top 5% of all HIV-interacting proteins in the sliding window test. However, in order to be able to identify genes with small number of polymorphic sites in relatively long and conserved sequences we relied on the criterion of the significance of the LRT to assess positive selection.

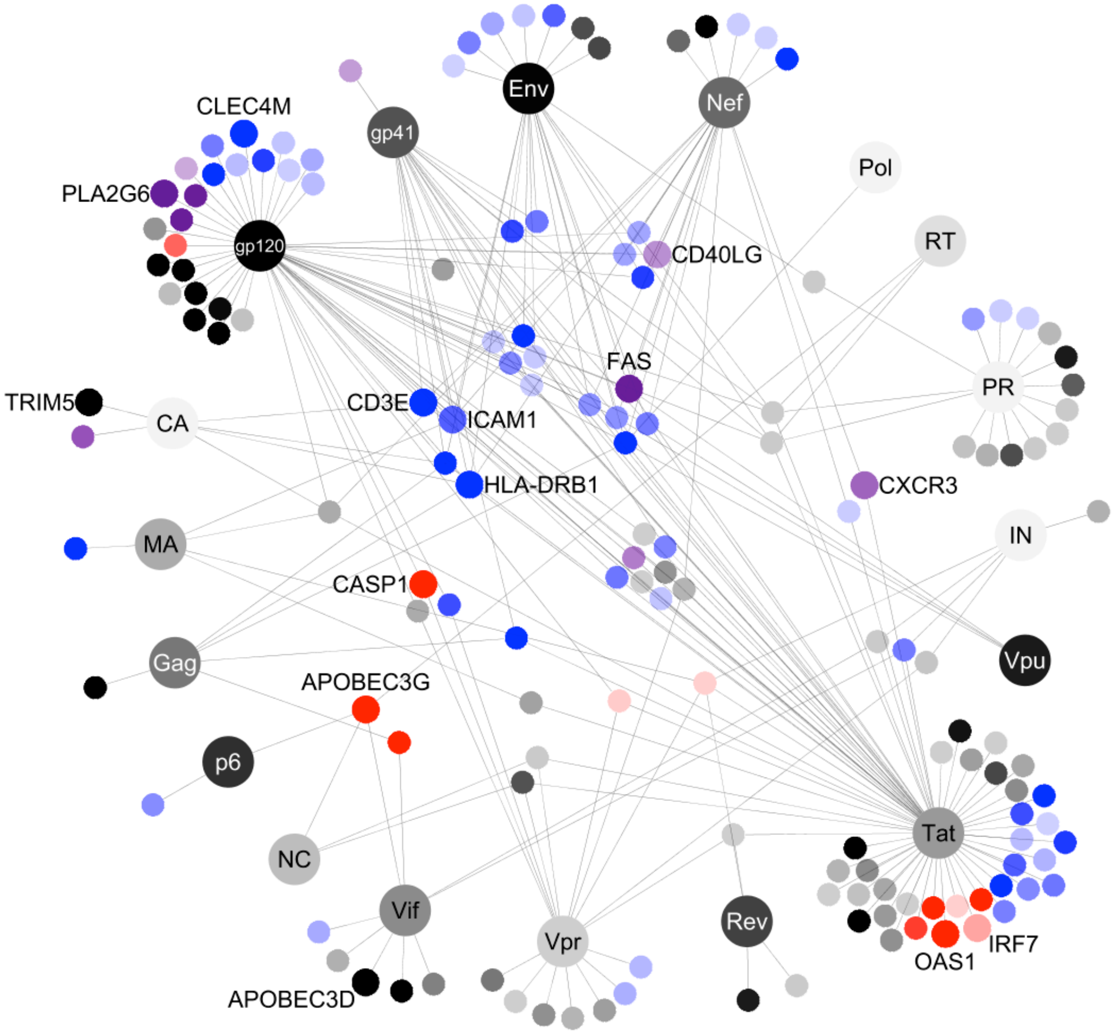


Figure 5.6 HIV-human protein interaction network. Viral proteins are represented by the largest dots. The shade of each dot indicates the rank of relative evolutionary rates of the viral proteins with dark colors representing a high rank and light colors representing a low rank. Host proteins are represented by smaller dots positioned among the interacting viral proteins. Nodes of the host factors discussed in the text are enlarged. Colors indicate the functional group to which the factors belong: blue - membrane-related proteins, red – innate immune response proteins, magenta – both. The intensity of the colors indicates the positive selection score in the site-based scoring with stronger colors representing a high rank and less intensive colors representing a low rank. Only host proteins with significant LRT in the comparison of the four primates are shown. The network was rendered using Cytoscape (Shannon, Markiel et al. 2003).

The functional analysis of the HIV-interacting host factors under positive selection pointed to two functional groups showing evidence of positive selection: membrane-related proteins and innate immune response proteins. These functional groups were overrepresented among the subsets of genes positively selected in all four primates, in human-chimp comparison and human only as inferred from the SNPs analysis (Figure 5.3). The interaction network of the virus proteins and the positively selected host proteins discussed below is visualized in Figure 5.6. Individual host factors pointed out in the text are labeled in this Figure.

Among the membrane-related HIV-interacting proteins under positive selection we observed several that are known to interact with other pathogens. C-type lectin (CLEC4M) a transmembrane receptor expressed on the surface of dendritic cells and macrophages, known to bind gp120 (Gattegno, Ramdani et al. 1992), is a part of signaling pathways induced by other pathogens such as *Mycobacterium tuberculosis* (Srivastava, Manchanda et al. 2009), *hepatitis C virus* (Pöhlmann, Zhang et al. 2003) and *ebola* virus (Alvarez, Lasala et al. 2002). Intercellular adhesion molecule 1 (ICAM1), a cell surface glycoprotein expressed in endothelial cells and cells of the immune system involved in a range of interactions with the HIV (Chirmule, Oyaizu et al. 1994; Fais, Capobianchi et al. 1995; Lafrenie, Wahl et al. 1996; Tardif and Tremblay 2003) is a receptor used by the *rhinovirus* (Greve, Davis et al. 1989). Chemokine (C-X-C motif) receptor 3 (CXCR3), a G-protein-coupled receptor expressed in activated T cells, NK cells, and dendritic cells, suggested to interact with Nef (van Marle, Henry et al. 2004) and Tat (Poggi, Carosio et al. 2004), participates in the signaling cascade of the T cell activation in *genital herpes simplex virus type 2* (Thapa and Carr 2009) and *hepatitis C virus* (Perney, Turriere et al. 2009) infections. Binding to the surface receptors of the host cell and internalization into that cell are the first steps of the viral infection; the genetic variability of the proteins expressed on the cell surface therefore represents a potential host defense mechanism to infection, preventing viral recognition and efficient cell entry. The genetic variation in membrane genes has been previously reported (Murphy 1993). This variation might have resulted from previous infections however can have impact on modern lentivirus restriction and primate species susceptibility to infection.

We detected a substantial amount of genetic variation in the membrane gene CD3E (rank 17 in the site-based, 3 in the sliding window score), coding for a part of the T cell receptor CD3 complex (CD3-TCR). The highest concentration of positively selected residues was found in the protein extracellular and transmembrane domains (Figure 5.7). The CD3-TCR complex is known to participate in the mechanism responsible for different levels of immune activation and T cell apoptosis in pathogenic HIV-human and nonpathogenic SIV-monkey infections (Schindler, Münch et al. 2006; Schindler, Schmökel et al. 2008). This and other examples of membrane proteins being under positive selection suggests that other proteins in this functional group might be relevant for differences in HIV pathogenesis.

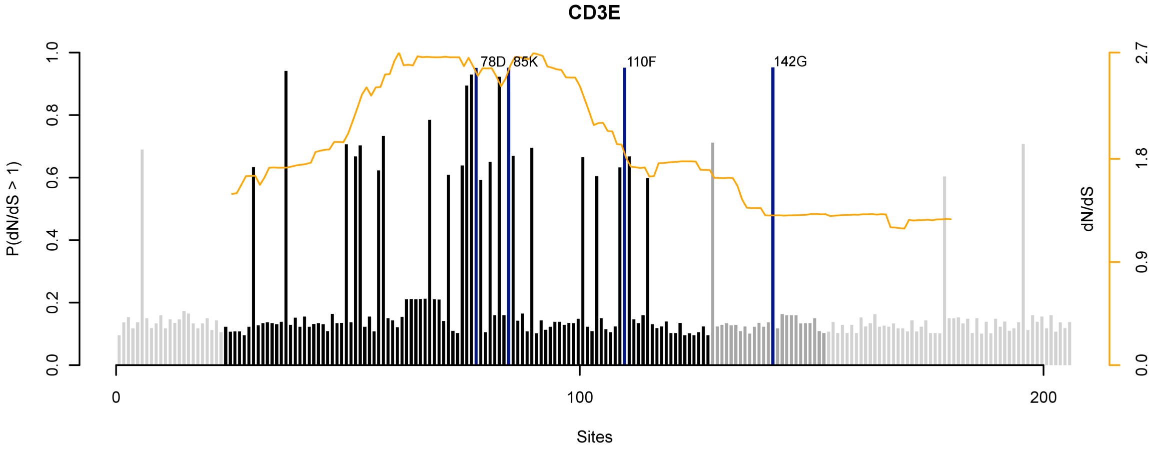


Figure 5.7 Positive selection of the CD3E gene. Bars indicate the dN/dS ratio of the following sites in the protein. Different shades of the bars indicate following domains of the protein: signal peptide (light gray), extracellular (black), transmembrane (dark gray) and intracellular (light gray). Sites with a high probability of being under positive selection ($p < 0.05$) are indicated with blue bars and labeled with their position and amino acid in the human protein. Orange line traces the dN/dS ratio calculated in the sliding window test. Highest accumulation of positive selection can be observed in the extracellular domain of CD3E.

Similar to the membrane-related genes, we found the innate immune response gene group to be significantly enriched among the positively selected genes. The innate immune response is the first line of defense against viral infection detecting virus particles, modulating signaling pathways leading to an increased interferon (IFN) production and inhibition of virus spread to limit pathogenesis. There is growing evidence that the variability in the innate immune response proteins among primate species might contribute to different levels of immune activation in immunodeficiency virus infections (Mandl, Barry et al. 2008; Ortiz, Guex et al. 2009).

We found interferon regulatory factor 7 (IRF7) to be under positive selection both in the test of four primate species and in the human-chimp comparison. IRF7 is a transcriptional factor involved in signaling triggered by the Toll-like receptor (TLR) TLR7 and TLR9 in response to the SIV infection (Mandl, Barry et al. 2008). Previous studies have shown that the genetic change in the primate IRF7 is correlated with the immune activation in SIV infection of primates (Mandl, Barry et al. 2008). Mutations in IRF7 might therefore be responsible for an altered TLR7 and TLR9 signaling leading to different levels of immune activation between pathogenic and nonpathogenic infections.

Caspase 1 (CASP1), a member of cysteine-aspartic acid protease family is an example of innate immune response gene under positive selection in primates (rank 31 in site-based and 26 in sliding window scoring) that is also located on a positively selected allele in humans. CASP1 is known to play a role in cellular apoptosis induced by HIV-1 Vpr (Stewart, Poon et al. 2000) and gp120 (Ohnimus, Heinkelein et al. 1997). Caspase activation has also been shown to participate in the cellular defense against *cowpox* (Komiyama, Ray et al. 1994), *baculovirus* (Cartier, Hershberger et al. 1994), and *dengue* *virus* (Nasirudeen and Liu 2009). HIV-1 interaction with the caspase-regulated apoptosis might play a role in limiting the host immune response to the virus and facilitating viral persistence.

Among the high-scoring innate immune response genes under positive selection we also found Fas (APO-1, CD95) (rank 13 in the site-based and 46 in the sliding window scoring), a member of the tumor necrosis factor (TNF) receptor family. The Fas-mediated pathway plays an important role in HIV-1 immunopathogenesis. Fas has been suggested to contribute to the loss of CD4$^+$ T cells in progression to AIDS as a part of the TCR-CD3 signaling pathway (Badley, Pilon et al. 2000; Petit, Corbeil et al. 2001). High levels of positive selection in Fas among primates and its regulatory relationship with TCR-CD3 point to this molecular pathway being possibly differently regulated among primates and partially responsible for immunopathogenesis.

Hyperactivation of the immune system in response to a pathogen recently introduced from another species can result in severe pathogenesis (e.g. hantavirus (Nichol, Spiropoulou et al. 1993)). Disentangling the interaction of the HIV with the innate immune response and the host-specific differences in primate species might thus be of therapeutic interest. The examples of innate immune response genes under positive selection relevant for diverse immune activation in HIV infection suggest that other genes reported by our analysis might merit further study.

To gain further insight into the positively selected genes we grouped them according to their interacting viral genes and characterized each group with the use of the positive selection score, gene roles in the PPI network and the evolutionary rate of the corresponding viral gene. Among the gene groups, genes interacting with the viral envelope and its subunits show elevated scores of positive selection. In contrast, we found limited evidence of positive selection in the genes interacting with viral Pol-encoded genes, with PR and IN in particular.

In the context of their role in the PPI network and interactions with other pathogens the envelope and gp41-interacting genes revealed a significantly low local connectivity, centrality and number of interacting pathogens and pathogen groups. The same pattern was observed for the group of membrane-related genes. This less crucial role in the PPI and high specificity for the interacting pathogens might be key factors allowing for the accelerated evolution of the membrane and envelope-interacting genes. The less critical role of these proteins in the host biology makes them potential drug targets, for example, the chemokine receptor CCR5 is targeted by MVC (Westby and van der Ryst 2005).

In contrast to the host proteins interacting with the viral envelope and its subunits we found significantly higher local connectivity and centrality in human PPIs of the host genes interacting with viral Pol-encoded genes. Together with their higher conservation this suggests that the crucial role of these highly connected genes in host cellular biology could be an evolutionary constraint. Pathogens might therefore interact preferentially with highly connected human proteins not only as a strategy to control critical host molecular processes (Dyer, Murali et al. 2008) but also because the lack of variation makes them a static target for virus-host interaction.

The high variability of the viral envelope gene and its subunits gp120 and gp41 (Hahn, Gonda et al. 1985; Holmes, Zhang et al. 1992) is in accordance with the high positive selection score of envelope-interacting and host membrane genes in general. Envelope proteins that participate in the crucial steps of binding to the cell receptors and coreceptors and membrane fusion (Deng, Liu et al. 1996; Berger, Murphy et al. 1999), are located on the surface of the virion and contain recognition sites for various adaptive immune responses. Thus changes in optimal host cell receptor affinity and evasion of host immune responses create selection pressure on the Env gene. Together with the positive selection acting on the host membrane genes it points to the viral recognition and host cell receptor affinity as processes in which both the viral and host genes undergo accelerated evolution driven by viral evasion and host suppression.

Conversely, Pol-encoded proteins show the least genetic variation among viral proteins. These proteins perform essential enzymatic functions common to all retroviruses, such as RNA retrotranscription, DNA integration and protein maturation and are unsurprisingly among the most conserved HIV/SIV proteins. Host factors interacting with the Pol-encoded proteins also show less genetic variation among the HIV-interacting genes, suggesting that these interactions tend to be the conserved parts of the host-pathogen interface.

The significant correlation of the viral and host gene ranks based on the rates of gene evolution further supports the hypothesis of reciprocal evolutionary effects between interacting host and pathogen proteins (Woolhouse, Webster et al. 2002). Highly conserved processes tend to be those crucial for viral replication; those less conserved might not be essential for the virus survival but involve accessory proteins and might contribute to pathogenesis. The highest rates of genetic change tend to be in the processes acting on the viral envelope and host cell membrane as they involve viral evasion and host pathogen immune recognition.

In this study we searched and characterized host defense factors under positive selection potentially involved in different responses to immunodeficiency virus infections in primates. Identifying genetic differences in the interacting proteins can open the way to biological testing of hypotheses regarding their role in various SIV/HIV infection phenotypes with different levels of immune activation. In addition to providing new insights into viral pathogenesis and host immunity, the approach presented here provides the potential for discovering new targets for antiviral therapies based on the knowledge of crucial elements of the host-pathogen interface and the pace of their evolution.

5.6 Conclusions

HIV has entered human population recently which likely contributes to its pathogenic nature (Hahn, Shaw et al. 2000). The fact that closely related to humans African primate species are infected with viruses ancestral to HIV and do not develop the disease offers an invaluable setting for comparative studies in the search of differences crucial for the development of immunodeficiency. As illustrated by host restriction factors (Sheehy, Gaddis et al. 2002; Stremlau, Owens et al. 2004; Neil, Zang et al. 2008) the interspecies genetic differences in the HIV-interacting genes might account for the HIV pathogenicity in humans.

The study presented in this chapter, comparing ~2500 HIV interactions among several host and virus species, was possible to accomplish only with the recent accumulation of the genomic (Karolchik, Kuhn et al. 2008) and interaction data (Fu, Sanders-Beer et al. 2009). Even though the amount and type of data allows for large-scale genomic analyses of the host-pathogen interactions, one should bear in mind potential errors inherent to this kind of data. Interactions collected in the HIV-1 human interaction database were reported by different labs, according to varying experimental protocols and criteria. Primate genomes are not yet fully annotated and the inference of homologous genes based on alignment with human genome might fail if the genome sequence is incomplete. Not fully validated interactions can therefore result in false positives in a large-scale interaction analysis; missing gene sequences hinder detection of potentially important interactions. The results of a large-scale analysis on not entirely confirmed data should be subjected to careful interpretation and additional validation. Nevertheless, as underscored by the presented examples confirmed by other experimental studies, the approach shown here affords indicating interactions relevant for the differences in the virus pathogenicity among primates. Given the accumulation of sequence data, analyses on the scale of genome represent a powerful method to filter out information potentially interesting for the disease.

Progressive growth and refinement of the genomic and interaction data will contribute to more reliable comparative studies of HIV-host interactions. Better validated and characterized interactions and a larger set of genomes of primate species naturally infected with SIV can increase the accuracy of this kind of analysis. A more targeted genomic study of SIV-infected monkey species is of high medical interest with a potential of providing information for the disease treatment.

The study presented in this chapter pictures HIV-host interactions on the broadest, genomic scale. In contrast to the previous chapters, here all known HIV-host interactions are analyzed among all species whose genomic data was available. This large-scale analysis affords means to generate novel hypotheses on the reasons of virus pathogenicity, filter out host factors potentially important for the disease and direct further detailed research.

CHAPTER 6 – Summary, conclusions and outlook view across the scales

One hundred years ago pathogens such as polio or smallpox presented a substantial threat to human life worldwide. Among the most remarkable achievements of modern medicine is the complete eradication of natural smallpox (Behbehani 1983) owing to the worldwide vaccination program. The smallpox vaccine is based on live viruses that induce antibodies protecting an individual from future infections. This way induced immunity is effective and long-term because the *Variola* virus which causes smallpox does not evolve to escape prior immunity (Ryan and Ray 2004).

In contrast, HIV appears to have an endless capacity to evolve to highly variable forms. Inducing long-term immunity against HIV is currently not possible due to the genetic flexibility of the virus and its capacity to rapidly escape host immune responses. Its high mutation and replication rates allow HIV to change continually in response to the selection pressure of host immune recognition and under the functional restriction of preserving its replicative fitness. Having a genome that codes only for few essential proteins, the virus is involved in a complex network of interactions with the host. Given the high variability of HIV and its dependence on the host for replication and spread only combined studies of the host and pathogen can provide causal information on the virus evolution. Understanding evolutionary forces shaping evolution of HIV might offer important insights for the search for effective and long-term anti-HIV treatments.

Evolutionary constraints and selection pressures on the virus operate on different scales from individual proteins through viral populations to species. On the scale of the individual protein, an interacting molecule is under constraints to retain its biological function while evading host restriction and immune recognition. On the population scale, selection promotes survival of the fittest virus able to replicate, infect new cells and evade host recognition in the most efficient way. On the scale of species, those virus variants are selected that are capable of passing the bottleneck of interspecies transmission and adapting to a new host species. Only through comprehensive analyses of host-pathogen interactions on different scales are we able to approach the complexity of HIV evolution and pathogenicity.

The work presented in this thesis represents an approach to analyze HIV-host interactions on different scales. Here we present a brief summary of the results, discuss their possible developments and propose their synthesis across the scales.

6.1 The scales individually

On the scale of individual interaction, the structural study of host TRIM5α and viral CA protein variants of different phenotypes pointed to their potential interaction site and structural and physical determinants of their binding. The analysis of this interaction should improve with the determination of crystal structure of the SPRY domain of TRIM5α or its close homolog and an increased number of protein variants tested for their interaction. The crystal structure would improve the accuracy of the homology-based models, larger sets of experimentally tested protein variants would allow for a systematic search for sequence and physicochemical determinants of the interaction. In the absence of knowledge of the exact mechanisms of an interaction, structural models and their detailed analyses provide insights into the potential binding sites and the factors conditioning the interaction. However given the high variability of the virus and the large number of potentially interesting host-pathogen interactions, this approach is limited to their unrepresentative subset.

On the scale of virus population, two studies presented in this thesis analyzed the highly variable V3 region of the HIV genome that is crucial for virus binding to the host coreceptors. The analysis of sequence space of the V3 loop pointed to the large variability of CXCR4- as compared to CCR5-binding viruses. Structural analysis of the V3 loop on the virus population scale resulted in a structure-based classification method of virus tropism and identification of physicochemical properties on the loop stem critical for the coreceptor usage. Both analyses reach beyond the classical methods for prediction of coreceptor usage and offer biological and evolutionary insights into coreceptor tropism. A potentially interesting direction of expanding the sequence space analysis is analyzing in the same manner virus population data obtained through deep-sequencing of patients of known clinical condition. Tracing of evolutionary trajectories of entire virus populations in sequence space in relation to the patient clinical attributes would validate and expand the observations gained from the presented analysis. A highly interesting way of developing the structure-based coreceptor prediction method is to target it at predicting the MVC therapy outcome instead of the virus phenotype. With the increasing use of this entry inhibitor in patient treatment and the growing data documenting their clinical history, such models can gain both high interest and improved accuracy.

The scale of virus population involves studies of seemingly unlimited number of virus variants. Such studies provide integrative models of virus quasispecies, enabling the analysis of their variation, dynamics and evolution. As relevant as these models are for the highly variable and large HIV populations in human hosts, models of viral populations in separation from the host and environmental factors that dictate their composition cannot explain the underlying biological mechanisms. The studies of virus populations presented in chapter 3 implicitly involve the virus interaction with the host chemokine coreceptors through virus tropism classification. The search for sequence and structure patterns within the virus classes allowed for drawing hypotheses about evolutionary patterns driving the emergence of each of the virus types and the structural and physicochemical determinants of the virus-coreceptor interaction.

On the scale of single cell, the study presented in this thesis integrated virus, host and environmental factors playing role in the virus cell entry. This study resulted in a comprehensive picture of the virus cell entry phenotypes. It indicated a preservation of the CCR5-tropic viruses as compared to more rarely observed CXCR4- and dual-tropic viruses. The models developed on the on scale of single cell represent a comprehensive description of the virus phenotype compared to the tropism classification, however, at a higher cost of describing an individual virus. With the increased number of viruses tested and characterized in the presented experimental and computational pipeline this approach will allow for prediction of virus phenotype based on V3 loop genotype. Such a predictive model represents a practical application of our approach in patient treatment with coreceptor blockers.

On the broadest scale of genome all known HIV-human interactions were analyzed in the context of the evolutionary change in the interacting host and virus proteins. This large-scale comparative genomics study of host-pathogen interactions identified and characterized host factors harboring higher amounts of genetic change and thus being potentially important for the infection. Additionally, by linking evolutionary rates of interacting host and pathogen proteins, patterns in the evolution of the host-pathogen biology were found, involving conservation of processes crucial for the virus replication and higher genetic variability of proteins involved in interactions related to virus cell entry and immune recognition. The accuracy of this genomic analysis could be improved with the use of fully validated and characterized interactions and a larger set of genomes of primate species naturally infected with SIV. Given the role of retroviruses in primate evolution, the identification of genomic signatures in the proteins relevant for the infection can provide additional layers of data for analysis of contemporary susceptibility to HIV. Analysis on this scale offers a broad view of the HIV interaction with the host and a filter of the large amount of reported interactions to the potentially interesting ones.

6.2 Multi-scale analysis

Analyses of HIV-host interactions limited to an individual scale provide incomplete information on different aspects of specific interactions. A better understanding of the HIV infection can be obtained through a systems biology approach of integrated studies covering different scales. In the work presented in this thesis, several patterns emerge across the scales.

The TRIM5α restriction factor, analyzed on the scale of individual interaction appeared also among the results of the study on the genomic scale. The importance of this restriction mechanism was both reflected by the variety of restriction phenotypes of chimeric TRIM5α proteins and by the high amounts of positive selection of this protein among the primate species. This indicates how an individual mutation changing the conformation or electrostatics at the protein interaction site can have implications on virus restriction and this way affect the spread of the pathogen in a population. A complete understanding of this defense mechanism can be therefore obtained through a multi-scale study – from protein to population.

The virus cell entry process is another key interface between the host and virus pictured on different scales in this thesis. High genetic variability of host membrane genes and virus envelope protein shown on the interspecies genomic and virus population scales reflects their complex evolutionary trajectories shaped by the dynamics of host recognition and pathogen evasion. On the scale of single cell the analysis of integrated effects of host, virus and environmental factors on the virus cell entry pointed to a comparable variability of the phenotypes of virus cell entry mechanism. This indicates how the complexity of the virus cell entry mechanism can be visible across the scales – from sequence to phenotype, from single cell to species.

The examples of patterns visible across scales demonstrate that although there is much to learn on each scale, the deepest understanding of the HIV-host interactions can be gained from integrating across the scales. In the next section we describe an outlook on a non-reductionist, systems biology approach to study HIV-host interactions through integrating heterogeneous biological information.

6.3 Outlook – systems biology approach to HIV research

Possible scales and types of analyses of the HIV-host interactions on each of them are countless. An emerging approach to study HIV-host interactions is based on genome-wide high-throughput information gained from recent and fast developing technologies of genome, transcriptome, proteome and siRNA screening. Large data sets generated using these new technologies can be analyzed in isolation however an integrative approach affords a broad, systems biology picture of HIV-host interaction measured genome-wide (Bushman, Malani et al. 2009; Telenti 2009).

Genomic analyses advanced to the point where comprehensive analyses of the role of host genetic variation in the HIV infection can be performed. Coupling new developments in the sequencing technology with large amounts of patient data collected over the last decades of HIV surveillance allows for systematic search of human variations determining different clinical outcomes of the infection. Several studies searched for human genetic factors that modulate the disease based on genome-wide association studies (GWAS) of infected cohorts (O'Brien and Nelson 2004; Marmor, Hertzmark et al. 2006; Fellay, Ge et al. 2009). Fellay et al. (Fellay, Ge et al. 2009) examined over 2500 infected individuals of Caucasian origins for common genetic traits that would explain the variability in the viral load at set point and the disease progression. The study confirmed the essential role in the HIV infection played by the MHC region (Martin and Carrington 2005) especially variants located near HLA-B and HLA-C genes (de Bakker, McVean et al. 2006; Colombo, Rauch et al. 2008), and the chemokine receptor cluster on chromosome 3. The interpretation of the polymorphisms is however not clear (Stranger, Forrest et al. 2005). Moreover, the reported genetic variation accounted for only ~13% of the differences in the clinical outcomes.

Beyond the information generated in host genetic studies, microarray technologies allow currently for genome-wide analyses of gene expression in infected cells in vitro and in vivo in individuals with HIV (Hyrcza, Kovacs et al. 2007; Giri, Nebozyhn et al. 2009; Rotger, Dang et al. 2010). Several gene expression studies (Hyrcza, Kovacs et al. 2007; Giri, Nebozyhn et al. 2009; Rotger, Dang et al. 2010) pointed to a broad modulation of antiviral defense system, such as interferon response and intrinsic cellular defense, to modulation of genes involved in the cell cycle and degradation pathways and to the absence of expression patterns typical for effective control of viral replication. Similar comparative studies of pathogenic and non-pathogenic SIV monkey infections showed evidence of downregulation of interferon response in the non-pathogenic infection (Mandl, Barry et al. 2008). Notably, the IRF7 gene involved in this response was characterized as a host factor harboring high amounts of genetic change in the study on the genomic scale described in chapter 5. This indicates how gene expression and comparative genomic studies can provide complementary information in the search for differences in the lentivirus primate infection.

Proteome analyses are still limited by the number of measurable proteins. Several studies (Chan, Qian et al. 2007; Ringrose, Jeeninga et al. 2008) screened expression levels of 2000-3200 proteins of HIV-infected individuals. The screens detected 15-21% of the measured proteins to be differentially expressed upon infection. Additionally changes were found in the abundance of proteins with known interactions with HIV. No integrated approaches have been used so far to analyze proteomic data in the context of other genome-wide studies of HIV infection.

SiRNA transfection screen is a recent high-throughput technology enabling the identification of proteins potentially necessary for virus replication. Three siRNA studies (Brass, Dykxhoorn et al. 2008; König, Zhou et al. 2008; Zhou, Xu et al. 2008) targeted at coding RNA of >20,000 human proteins identified ~1000 proteins as potentially related to optimal viral replication. Only a small overlap across studies was found (Bushman, Malani et al. 2009) supposedly due to the novelty of the technology and differences in study design. However, among those genes that were shared by one or more studies, a functional pattern was observed (Fellay, Shianna et al. 2010). Notably, none of the screens identified genes that would restrict viral replication, i.e. whose silencing would result in greater viral production.

The technologies described above allow for acquisition of different type of data on a genome-wide scale. None of the studies based on a particular technology fully explains the variation in the response to infection, which points to integrative approaches to this type of heterogeneous biological data as a potential way to obtain complementary information about the HIV infection. Individual examples of integrative studies of genome-wide data exist, e. g. combining transcriptome analysis with genetic variation in respective regulators (Rotger, Dang et al. 2010), however finding a limited overlap among different types of data and therefore offereing a limited interpretation of the results.

The patterns of HIV interaction with the host are intricate. Their identification requires combining several layers of information – from genetic variation, through expression to function. Eventually, integration of genome-wide data should provide comprehensive results that will complement or guide further experiments under the systems biology paradigm to study HIV infection.

PUBLICATIONS

Journal articles related to this thesis:

Bozek, K., Thielen, A., Sierra, S., Kaiser, R., Lengauer, T. V3 loop sequence space analysis suggests different evolutionary patterns of CCR5- and CXCR4-tropic HIV. *PLoS One.* 2009 Oct 9;4(10):e7387

Bozek, K., Lengauer, T. Positive selection of HIV host factors and the evolution of lentivirus genes. *BMC Evol Biol.* 2010 Jun 18;10:186.

Bozek, K., Lengauer, T., Sierra, S., Kaiser, R., Domingues, FS. Analysis of physicochemical and structural properties determining HIV-1 coreceptor usage. *submitted*

Bozek, K., Eckhardt, M., Sierra, S., Kaiser, R., Müller, B., Kräusslich, HG., Lengauer, T. A comprehensive model of HIV cell entry phenotype based on multi-parameter single cell data. *in preparation*

Kono, K., Bozek, K., Domingues, FS., Shioda, T., Nakayama, EE. Impact of a single amino acid in the variable region 2 of the Old World monkey TRIM5alpha SPRY (B30.2) domain on anti-human immunodeficiency virus type 2 activity. *Virology.* 2009 May 25;388(1):160-8

Kuroishi, A., Bozek, K., Shioda, T., Nakayama, EE. A single amino acid substitution of the human immunodeficiency virus type 1 capsid protein affects viral sensitivity to TRIM5alpha. *Retrovirology.* 2010 Jul 7;7:58

Other journal articles published during my work on this thesis:

Bozek, K., Kielbasa, SM., Kramer, A., Herzel, H. Promoter Analysis of Mammalian Clock Controlled genes. *Genome Inform.* 2007;18:65-74

Bozek, K., Relogio, A., Kielbasa, SM., Heine, M., Dame, C., Kramer, A., Herzel, H. Regulation of clock-controlled genes in mammals. *PLoS One.* 2009;4(3):e4882

Bozek, K., Rosahl, AL., Gaub, S., Lorenzen, S., Herzel, H. Circadian Transcription in Liver. *Biosystems.* 2010 Jul 21

REFERENCES

"Los Alamos National Laboratory HIV Sequence Database." from http://www.hiv.lanl.gov/.

Adachi, A., H. E. Gendelman, et al. (1986). "Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone." J Virol **59**(2): 284-291.

Alexa, A., J. Rahnenführer, et al. (2006). "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure." Bioinformatics **Jul 1;22**(13): 1600-1607.

Allen, T. M., D. H. O'Connor, et al. (2000). "Tat-specific cytotoxic T lymphocytes select for SIV escape variants during resolution of primary viraemia." Nature **407**(6802): 386-390.

Altmann, A., M. Däumer, et al. (2009). "Predicting the response to combination antiretroviral therapy: retrospective validation of geno2pheno-THEO on a large clinical database." J Infect Dis **199**(7): 999-1006.

Altmann, A., M. Rosen-Zvi, et al. (2008). "Comparison of classifier fusion methods for predicting response to anti HIV-1 therapy." PLoS One **3**(10): e3470.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Alvarez, C. P., F. Lasala, et al. (2002). "C-type lectins DC-SIGN and L-SIGN mediate cellular entry by Ebola virus in cis and in trans." J Virol **76**(13): 6841-6844.

Archer, J., M. S. Braverman, et al. (2009). "Detection of low-frequency pretherapy chemokine (CXC motif) receptor 4 (CXCR4)-using HIV-1 with ultra-deep pyrosequencing." AIDS **23**(10): 1209-1218.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet. **May;25**(1): 25-29.

Atchley, W. R., J. Zhao, et al. (2005). "Solving the protein sequence metric problem." Proc Natl Acad Sci U S A **102**(18): 6395-6400.

Badley, A. D., A. A. Pilon, et al. (2000). "Mechanisms of HIV-associated lymphocyte apoptosis." Blood **Nov 1;96**(9): 2951-2964.

Baker, N. A., D. Sept, et al. (2001). "Electrostatics of nanosystems: application to microtubules and the ribosome." Proc Natl Acad Sci U S A **98**(18): 10037-10041.

Bandelt, H. J. and A. W. Dress (1992). "Split decomposition: a new and useful approach to phylogenetic analysis of distance data." Mol Phylogenet Evol **1**(3): 242-252.

Barnett, S. W., K. K. Murthy, et al. (1994). "An AIDS-like condition induced in baboons by HIV-2." Science **266**(5185): 642-646.

Barouch, D. H. (2008). "Challenges in the development of an HIV-1 vaccine." Nature **455**(7213): 613-619.

Barre-Sinoussi, F., J. C. Chermann, et al. (1983). "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)." Science **220**(4599): 868-871.

Beerenwinkel, N., M. Däumer, et al. (2003). "Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes." Nucleic Acids Res **31**(13): 3850-3855.

Behbehani, A. M. (1983). "The smallpox story: life and death of an old disease." Microbiol Rev **47**(4): 455-509.

Benson, D. A., I. Karsch-Mizrachi, et al. (2009). "GenBank." Nucleic Acids Res **Jan;37**(Database issue): D26-31.

Berger, E. A., P. M. Murphy, et al. (1999). "Chemokine receptors as HIV-1 coreceptors: Roles in viral entry, tropism, and disease." Annu Rev Immunol **17**: 657-700.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Res **28**(1): 235-242.

Besnier, C., Y. Takeuchi, et al. (2002). "Restriction of lentivirus in monkeys." Proc Natl Acad Sci U S A **99**(18): 11920-11925.

Bieniasz, P. D. (2004). "Intrinsic immunity: a front-line defense against viral attack." Nat Immunol **5**(11): 1109-1115.

Blanchette, M., W. J. Kent, et al. (2004). "Aligning multiple genomic sequences with the threaded blockset aligner." Genome Res **Apr;14**(4): 708-715.

Bogerd, H. P., B. P. Doehle, et al. (2004). "A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor." Proc Natl Acad Sci U S A **Mar 16;101**(11): 3770-3774.

Borner, K., J. Hermle, et al. (2010). "From experimental setup to bioinformatics: an RNAi screening platform to identify host factors involved in HIV-1 replication." Biotechnol J **5**(1): 39-49.

Borrow, P., H. Lewicki, et al. (1997). "Antiviral pressure exerted by HIV-1-specific cytotoxic T lymphocytes (CTLs) during primary infection demonstrated by rapid selection of CTL escape virus." Nat Med **3**(2): 205-211.

Boser, B., I. Guyon, et al. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory. Pittsburgh, Pennsylvania, United States, ACM.

Brass, A. L., D. M. Dykxhoorn, et al. (2008). "Identification of host proteins required for HIV infection through a functional genomic screen." Science **Feb 15;319**(5865): 921-926.

Breiman, L. (2001). "Random forests." Mach. Learn **45**: 5.

Brumme, Z. L., W. W. Dong, et al. (2004). "Clinical and immunological impact of HIV envelope V3 sequence variation after starting initial triple antiretroviral therapy." AIDS **18**(4): F1-9.

Bushman, F. D., N. Malani, et al. (2009). "Host cell factors in HIV replication: meta-analysis of genome-wide studies." PLoS Pathog **May;5**(5): e1000437.

Cartier, J. L., P. A. Hershberger, et al. (1994). "Suppression of apoptosis in insect cells stably transfected with baculovirus p35: dominant interference by N-terminal sequences p35(1-76)." J Virol **68**(12): 7728-7737.

Cavrois, M., C. De Noronha, et al. (2002). "A sensitive and specific enzyme-based assay detecting HIV-1 virion fusion in primary T lymphocytes." Nat Biotechnol **20**(11): 1151-1154.

Chan, D. C. and P. S. Kim (1998). "HIV entry and its inhibition." Cell **93**(5): 681-684.

Chan, E. Y., W. J. Qian, et al. (2007). "Quantitative analysis of human immunodeficiency virus type 1-infected CD4+ cell proteome: dysregulated cell cycle progression and nuclear transport coincide with robust virus production." J Virol **81**(14): 7571-7583.

Chesebro, B., K. Wehrly, et al. (1992). "Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell-tropic isolates: definition of critical amino acids involved in cell tropism." J Virol **66**(11): 6547-6554.

Chirmule, N., N. Oyaizu, et al. (1994). "Nef protein of HIV-1 has B-cell stimulatory activity." AIDS **Jun;8**(6): 733-734.

Chun, T. W., D. C. Nickle, et al. (2005). "HIV-infected individuals receiving effective antiviral therapy for extended periods of time continually replenish their viral reservoir." J Clin Invest **115**(11): 3250-3255.

Colgrove, R. and A. Japour (1999). "A combinatorial ledge: reverse transcriptase fidelity, total body viral burden, and the implications of multiple-drug HIV therapy for the evolution of antiviral resistance." Antiviral Res **41**(1): 45-56.

Colombo, S., A. Rauch, et al. (2008). "The HCP5 single-nucleotide polymorphism: a simple screening tool for prediction of hypersensitivity reaction to abacavir." J Infect Dis **198**(6): 864-867.

Condra, J. H. (1998). "Resistance to HIV protease inhibitors." Haemophilia **4**(4): 610-615.

Dai, S. J., G. F. Dou, et al. (2005). "Pharmacokinetics of sifuvirtide, a novel anti-HIV-1 peptide, in monkeys and its inhibitory concentration in vitro." Acta Pharmacol Sin **26**(10): 1274-1280.

de Bakker, P. I., G. McVean, et al. (2006). "A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC." Nat Genet **38**(10): 1166-1172.

De Clercq, E., N. Yamamoto, et al. (1992). "Potent and selective inhibition of human immunodeficiency virus (HIV)-1 and HIV-2 replication by a class of bicyclams interacting with a viral uncoating event." Proc Natl Acad Sci U S A **89**(12): 5286-5290.

DeLano, W. "PyMol ".

Deng, H., R. Liu, et al. (1996). "Identification of a major co-receptor for primary isolates of HIV-1." Nature **Jun 20;381**(6584): 661-666.

Dimitriadou, E., K. Hornik, et al. (2005). "e1071: Misc functions of the department of statistics (e1071)." TU Wien.

Doms, R. and J. Moore (1997). "HIV coreceptor use: A molecular window into viral tropism." Human Retroviruses and AIDS **Theoretical Biology and Biophysics, Part III** 1-12.

Dong, F. and H. X. Zhou (2002). "Electrostatic contributions to T4 lysozyme stability: solvent-exposed charges versus semi-buried salt bridges." Biophys J **83**(3): 1341-1347.

Dorr, P., M. Westby, et al. (2005). "Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity." Antimicrob Agents Chemother **49**(11): 4721-4732.

Douek, D. C., L. J. Picker, et al. (2003). "T cell dynamics in HIV-1 infection." Annu Rev Immunol **21**: 265-304.

Drake, J. W. (1993). "Rates of spontaneous mutation among RNA viruses." Proc Natl Acad Sci U S A **90**(9): 4171-4175.

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

Drummond, A. J., A. Rambaut, et al. (2005). "Bayesian coalescent inference of past population dynamics from molecular sequences." Mol Biol Evol **22**(5): 1185-1192.

Dybowski, J. N., D. Heider, et al. (2010). "Prediction of co-receptor usage of HIV-1 from genotype." PLoS Comput Biol **6**(4): e1000743.

Dyer, M. D., T. M. Murali, et al. (2008). "The landscape of human proteins interacting with viruses and other pathogens." PLoS Pathog **Feb 8;4**(2): e32.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **Mar 19;32**(5): 1792-1797.

Edwards, R. A. and F. Rohwer (2005). "Viral metagenomics." Nat Rev Microbiol **3**(6): 504-510.

Efstathiou, S. and C. M. Preston (2005). "Towards an understanding of the molecular basis of herpes simplex virus latency." Virus Res **111**(2): 108-119.

Eigen, M. (1996). "On the nature of virus quasispecies." Trends Microbiol **4**(6): 216-218.

Endo, T., K. Ikeo, et al. (1996). "Large-scale search for genes on which positive selection may operate." Mol Biol Evol **May;13**(5): 685-690.

Este, J. A. (2003). "Virus entry as a target for anti-HIV intervention." Curr Med Chem **10**(17): 1617-1632.

Este, J. A. and A. Telenti (2007). "HIV entry inhibitors." Lancet **370**(9581): 81-88.

Eswar, N., B. Webb, et al. (2006). "Comparative protein structure modeling using Modeller." Curr Protoc Bioinformatics **Chapter 5**(5): Unit 5 6.

Fais, S., M. R. Capobianchi, et al. (1995). "Unidirectional budding of HIV-1 at the site of cell-to-cell contact is associated with co-polarization of intercellular adhesion molecules and HIV-1 viral matrix protein." AIDS **Apr;9**(4): 329-335.

Fatkenheuer, G., M. Nelson, et al. (2008). "Subgroup analyses of maraviroc in previously treated R5 HIV-1 infection." N Engl J Med **359**(14): 1442-1455.

Fellay, J., D. Ge, et al. (2009). "Common genetic variation and the control of HIV-1 in humans." PLoS Genet **5**(12): e1000791.

Fellay, J., K. V. Shianna, et al. (2010). "Host genetics and HIV-1: the final phase?" PLoS Pathog **6**(10): e1001033.

Felsenstein, J. (2005). "PHYLIP (Phylogeny Inference Package) version 3.6." Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.

Felsenstein, J. and G. A. Churchill (1996). "A Hidden Markov Model approach to variation among sites in rate of evolution." Mol Biol Evol **Jan;13**(1): 93-104.

Feng, Y., C. C. Broder, et al. (1996). "HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor." Science **272**(5263): 872-877.

Fiebig, E. W., D. J. Wright, et al. (2003). "Dynamics of HIV viremia and antibody seroconversion in plasma donors: implications for diagnosis and staging of primary HIV infection." AIDS **17**(13): 1871-1879.

Fletcher, C. V. (2003). "Enfuvirtide, a new drug for HIV infection." Lancet **361**(9369): 1577-1578.

Forsman, A. and R. A. Weiss (2008). "Why is HIV a pathogen?" Trends Microbiol **16**(12): 555-560.

Fouchier, R. A., M. Groenink, et al. (1992). "Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule." J Virol **66**(5): 3183-3187.

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

Fu, W., B. E. Sanders-Beer, et al. (2009). "Human immunodeficiency virus type 1, human protein interaction database at NCBI." Nucleic Acids Res **Jan;37**(Database issue): D417-422.

Gamble, T. R., S. Yoo, et al. (1997). "Structure of the carboxyl-terminal dimerization domain of the HIV-1 capsid protein." Science **278**(5339): 849-853.

Gao, F., E. Bailes, et al. (1999). "Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes." Nature **Feb 4;397**(6718): 436-441.

Gao, F., L. Yue, et al. (1992). "Human infection by genetically diverse SIVSM-related HIV-2 in west Africa." Nature **Aug 6;358**(6386): 495-499.

Garten, R. J., C. T. Davis, et al. (2009). "Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans." Science **325**(5937): 197-201.

Gattegno, L., A. Ramdani, et al. (1992). "Lectin-carbohydrate interactions and infectivity of human immunodeficiency virus type 1 (HIV-1)." AIDS Res Hum Retroviruses **Jan;8**: 27-37.

Gentleman, R. C., V. J. Carey, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Giri, M. S., M. Nebozyhn, et al. (2009). "Circulating monocytes in HIV-1-infected viremic subjects exhibit an antiapoptosis gene signature and virus- and host-mediated apoptosis resistance." J Immunol **182**(7): 4459-4470.

Gitti, R. K., B. M. Lee, et al. (1996). "Structure of the amino-terminal core domain of the HIV-1 capsid protein." Science **273**(5272): 231-235.

Goonetilleke, N., M. K. Liu, et al. (2009). "The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection." J Exp Med **206**(6): 1253-1272.

Gordon, J. C., J. B. Myers, et al. (2005). "H++: a server for estimating pKas and adding missing hydrogens to macromolecules." Nucleic Acids Res **33**(Web Server issue): W368-371.

Goulder, P. J., A. K. Sewell, et al. (1997). "Patterns of immunodominance in HIV-1-specific cytotoxic T lymphocyte responses in two human histocompatibility leukocyte antigens (HLA)-identical siblings with HLA-A*0201 are influenced by epitope mutation." J Exp Med **185**(8): 1423-1433.

Grenfell, B. T., O. G. Pybus, et al. (2004). "Unifying the epidemiological and evolutionary dynamics of pathogens." Science **303**(5656): 327-332.

Greve, J. M., G. Davis, et al. (1989). "The major human rhinovirus receptor is ICAM-1." Cell **Mar 10;56**(5): 839-847.

Grutter, C., C. Briand, et al. (2006). "Structure of the PRYSPRY-domain: implications for autoinflammatory diseases." FEBS Lett **580**(1): 99-106.

Guidotti, L. G. and F. V. Chisari (2006). "Immunobiology and pathogenesis of viral hepatitis." Annu Rev Pathol **1**: 23-61.

Guyon, I., J. Weston, et al. (2002). "Gene selection for cancer classification using support vector machines." Mach. Learn **46**: 389-422.

Hahn, B. H., M. A. Gonda, et al. (1985). "Genomic diversity of the acquired immune deficiency syndrome virus HTLV-III: different viruses exhibit greatest divergence in their envelope genes." Proc Natl Acad Sci U S A **Jul;82**(14): 4813-4817.

Hahn, B. H., G. M. Shaw, et al. (2000). "AIDS as a zoonosis: scientific and public health implications." Science **Jan 28;287**(5453): 607-614.

Hastie, T., R. Tibshirani, et al. (2001). The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York., Springer.

Heeney, J. L., A. G. Dalgleish, et al. (2006). "Origins of HIV and the evolution of resistance to AIDS." Science **Jul 28;313**(5786): 462-466.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." Proc Natl Acad Sci U S A **89**(22): 10915-10919.

Hildebrandt, A., R. Blossey, et al. (2007). "Electrostatic potentials of proteins in water: a structured continuum approach." Bioinformatics **23**(2): e99-103.

Hirsch, V. M., R. A. Olmsted, et al. (1989). "An African primate lentivirus (SIVsm) closely related to HIV-2." Nature **Jun 1;339**(6223): 389-392.

Hoffman, N. G., F. Seillier-Moiseiwitsch, et al. (2002). "Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop." J Virol **76**(8): 3852-3864.

Hoffmann, C., N. Minkah, et al. (2007). "DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations." Nucleic Acids Res **35**(13): e91.

Holmes, E. C., L. Q. Zhang, et al. (1992). "Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient." Proc Natl Acad Sci U S A **89**(11): 4835-4839.

Huang, C. C., S. N. Lam, et al. (2007). "Structures of the CCR5 N terminus and of a tyrosine-sulfated antibody with HIV-1 gp120 and CD4." Science **317**(5846): 1930-1934.

Huang, C. C., M. Tang, et al. (2005). "Structure of a V3-containing HIV-1 gp120 core." Science **310**(5750): 1025-1028.

Huang, Y., W. A. Paxton, et al. (1996). "The role of a mutant CCR5 allele in HIV-1 transmission and disease progression." Nat Med **2**(11): 1240-1243.

Huson, D. H. (1998). "SplitsTree: analyzing and visualizing evolutionary data." Bioinformatics **14**(1): 68-73.

Hyrcza, M. D., C. Kovacs, et al. (2007). "Distinct transcriptional profiles in ex vivo CD4+ and CD8+ T cells are established early in human immunodeficiency virus type 1 infection and are characterized by a chronic interferon response as well as extensive transcriptional changes in CD8+ T cells." J Virol **81**(7): 3477-3486.

James, L. C., A. H. Keeble, et al. (2007). "Structural basis for PRYSPRY-mediated tripartite motif (TRIM) protein function." Proc Natl Acad Sci U S A **104**(15): 6200-6205.

Jeang, K. T. (2010). "HTLV-1 and adult T-cell leukemia: insights into viral transformation of cells 30 years after virus discovery." J Formos Med Assoc **109**(10): 688-693.

Jensen, M. A. and A. B. van 't Wout (2003). "Predicting HIV-1 coreceptor usage with sequence analysis." AIDS Rev **5**(2): 104-112.

Jia, B., R. Serra-Moreno, et al. (2009). "Species-specific activity of SIV Nef and HIV-1 Vpu in overcoming restriction by tetherin/BST2." PLoS Pathog **May;5**(5): e1000429.

Johnson, J. A., J. F. Li, et al. (2008). "Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy." PLoS Med **5**(7): e158.

Johnston, S. H., M. A. Lobritz, et al. (2009). "A quantitative affinity-profiling system that reveals distinct CD4/CCR5 usage patterns among human immunodeficiency virus type 1 and simian immunodeficiency virus strains." J Virol **83**(21): 11016-11026.

Karolchik, D., R. M. Kuhn, et al. (2008). "The UCSC Genome Browser Database: 2008 update." Nucleic Acids Res **Jan;36**(Database issue): D773-779.

Katze, M. G., J. L. Fornek, et al. (2008). "Innate immune modulation by RNA viruses: emerging insights from functional genomics." Nat Rev Immunol **Aug;8**(8): 644-654.

Kawashima, S., H. Ogata, et al. (1999). "AAindex: Amino Acid Index Database." Nucleic Acids Res **27**(1): 368-369.

Keeble, A. H., Z. Khan, et al. (2008). "TRIM21 is an IgG receptor that is structurally, thermodynamically, and kinetically conserved." Proc Natl Acad Sci U S A **105**(16): 6045-6050.

Keele, B. F., J. H. Jones, et al. (2009). "Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz." Nature **Jul 23;460**(7254): 515-519.

Kirchhoff, F. (2009). "Is the high virulence of HIV-1 an unfortunate coincidence of primate lentiviral evolution?" Nat Rev Microbiol. **Jun;7**(6): 467-476.

Knipe, D., P. Howley, et al. (2006). Fields Virology, Lippincott Williams & Wilkins.

Komiyama, T., C. A. Ray, et al. (1994). "Inhibition of interleukin-1 beta converting enzyme by the cowpox virus serpin CrmA. An example of cross-class inhibition." J Biol Chem **269**(30): 19331-19337.

König, R., Y. Zhou, et al. (2008). "Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication." Cell **Oct 3;135**(1): 49-60.

Kono, K., H. Song, et al. (2008). "Comparison of anti-viral activity of rhesus monkey and cynomolgus monkey TRIM5alphas against human immunodeficiency virus type 2 infection." Virology **373**(2): 447-456.

Kono, K., H. Song, et al. (2010). "Multiple sites in the N-terminal half of simian immunodeficiency virus capsid protein contribute to evasion from rhesus monkey TRIM5alpha-mediated restriction." Retrovirology **7**: 72.

Korber, B., B. Foley, et al. (1998). "Numbering positions in HIV relative to HXB2CG." Human Retrovirus and AIDS **III**: 102-111.

Korber, B., B. Gaschen, et al. (2001). "Evolutionary and immunological implications of contemporary HIV-1 variation." Br Med Bull **58**: 19-42.

Korber, B., M. Muldoon, et al. (2000). "Timing the ancestor of the HIV-1 pandemic strains." Science **Jun 9;288**(5472): 1789-1796.

Korber, B., J. Theiler, et al. (1998). "Limitations of a molecular clock applied to considerations of the origin of HIV-1." Science **280**(5371): 1868-1871.

Kortemme, T. and D. Baker (2002). "A simple physical model for binding energy hot spots in protein-protein complexes." Proc Natl Acad Sci U S A **99**(22): 14116-14121.

Kowalski, M., J. Potz, et al. (1987). "Functional regions of the envelope glycoprotein of human immunodeficiency virus type 1." Science **237**(4820): 1351-1355.

Kräusslich, H. G., R. H. Ingraham, et al. (1989). "Activity of purified biosynthetic proteinase of human immunodeficiency virus on natural substrates and synthetic peptides." Proc Natl Acad Sci U S A **86**(3): 807-811.

Kuroishi, A., K. Bozek, et al. (2010). "A single amino acid substitution of the human immunodeficiency virus type 1 capsid protein affects viral sensitivity to TRIM5 alpha." Retrovirology **7**: 58.

Kuroishi, A., A. Saito, et al. (2009). "Modification of a loop sequence between alpha-helices 6 and 7 of virus capsid (CA) protein in a human immunodeficiency virus type 1 (HIV-1) derivative that has simian immunodeficiency virus (SIVmac239) vif and CA alpha-helices 4 and 5 loop improves replication in cynomolgus monkey cells." Retrovirology **6**(70): 70.

Lafrenie, R. M., L. M. Wahl, et al. (1996). "HIV-1-Tat modulates the function of monocytes and alters their interactions with microvessel endothelial cells. A mechanism of HIV pathogenesis." J Immunol **Feb 15;156**(4): 1638-1645.

Larder, B., D. Wang, et al. (2007). "The development of artificial neural networks to predict virological response to combination HIV therapy." Antivir Ther **12**(1): 15-24.

Lengauer, T. and T. Sing (2006). "Bioinformatics-assisted anti-HIV therapy." Nat Rev Microbiol **4**(10): 790-797.

Little, S. J., A. R. McLean, et al. (1999). "Viral dynamics of acute HIV-1 infection." J Exp Med **190**(6): 841-850.

Liu, R., W. A. Paxton, et al. (1996). "Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection." Cell **86**(3): 367-377.

Locher, C. P., S. W. Barnett, et al. (1998). "Human immunodeficiency virus-2 infection in baboons is an animal model for human immunodeficiency virus pathogenesis in humans." Arch Pathol Lab Med **122**(6): 523-533.

Locher, C. P., S. A. Witt, et al. (2001). "Baboons as an animal model for human immunodeficiency virus pathogenesis and vaccine development." Immunol Rev **183**: 127-140.

Low, A. J., D. Marchant, et al. (2008). "CD4-dependent characteristics of coreceptor use and HIV type 1 V3 sequence in a large population of therapy-naive individuals." AIDS Res Hum Retroviruses **24**(2): 219-228.

Lu, M., S. C. Blacklow, et al. (1995). "A trimeric structural domain of the HIV-1 transmembrane glycoprotein." Nat Struct Biol **2**(12): 1075-1082.

Lukashov, V. V., C. L. Kuiken, et al. (1995). "Intrahost human immunodeficiency virus type 1 evolution is related to length of the immunocompetent period." J Virol **69**(11): 6911-6916.

Maddon, P. J., A. G. Dalgleish, et al. (1986). "The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain." Cell **47**(3): 333-348.

Malim, M. H. and M. Emerman (2008). "HIV-1 accessory proteins--ensuring viral survival in a hostile environment." Cell Host Microbe **3**(6): 388-398.

Mandl, J. N., A. P. Barry, et al. (2008). "Divergent TLR7 and TLR9 signaling and type I interferon production distinguish pathogenic and nonpathogenic AIDS virus infections." Nat Med **Oct;14**(10): 1077-1087.

Marmor, M., K. Hertzmark, et al. (2006). "Resistance to HIV infection." J Urban Health **83**(1): 5-17.

Martin, M. P. and M. Carrington (2005). "Immunogenetics of viral infections." Curr Opin Immunol **17**(5): 510-516.

Masso, M. and Vaisman, II (2010). "Accurate and efficient gp120 V3 loop structure based models for the determination of HIV-1 co-receptor usage." BMC Bioinformatics **11**: 494.

McNatt, M. W., T. Zang, et al. (2009). "Species-specific activity of HIV-1 Vpu and positive selection of tetherin transmembrane domain variants." PLoS Pathog **Feb;5**(2): e1000300.

McNearney, T., Z. Hornickova, et al. (1992). "Relationship of human immunodeficiency virus type 1 sequence heterogeneity to stage of disease." Proc Natl Acad Sci U S A **89**(21): 10247-10251.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Miedema, F., L. Meyaard, et al. (1994). "Changing virus-host interactions in the course of HIV-1 infection." Immunol Rev **140**: 35-72.

Milich, L., B. Margolin, et al. (1993). "V3 loop of the human immunodeficiency virus type 1 Env protein: interpreting sequence variability." J Virol **67**(9): 5623-5634.

Montagnier, L., J. Gruest, et al. (1984). "Adaptation of lymphadenopathy associated virus (LAV) to replication in EBV-transformed B lymphoblastoid cell lines." Science **225**(4657): 63-66.

Moore, J. P. (1997). "Coreceptors: implications for HIV pathogenesis and therapy." Science **276**(5309): 51-52.

Muller-Trutwin, M. C., S. Corbet, et al. (1996). "The evolutionary rate of nonpathogenic simian immunodeficiency virus (SIVagm) is in agreement with a rapid and continuous replication in vivo." Virology **223**(1): 89-102.

Murphy, P. M. (1993). "Molecular mimicry and the generation of host defense protein diversity." Cell. **Mar 26;72**(6): 823-826.

Nakayama, E. E., H. Miyoshi, et al. (2005). "A specific region of 37 amino acid residues in the SPRY (B30.2) domain of African green monkey TRIM5alpha determines species-specific restriction of simian immunodeficiency virus SIVmac infection." J Virol **79**(14): 8870-8877.

Nasirudeen, A. M. and D. X. Liu (2009). "Gene expression profiling by microarray analysis reveals an important role for caspase-1 in dengue virus-induced p53-mediated apoptosis." J Med Virol **81**(6): 1069-1081.

Neil, S. and P. Bieniasz (2009). "Human immunodeficiency virus, restriction factors, and interferon." J Interferon Cytokine Res **29**(9): 569-580.

Neil, S. J., T. Zang, et al. (2008). "Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu." Nature **451**(7177): 425-430.

Nelson, J. A., S. A. Fiscus, et al. (1997). "Evolutionary variants of the human immunodeficiency virus type 1 V3 region characterized by using a heteroduplex tracking assay." J Virol **71**(11): 8750-8758.

Nichol, S. T., C. F. Spiropoulou, et al. (1993). "Genetic identification of a hantavirus associated with an outbreak of acute respiratory illness." Science **Nov 5;262**(5135): 914-917.

Nicholson, J. K., S. W. Browning, et al. (2001). "CCR5 and CXCR4 expression on memory and naive T cells in HIV-1 infection and response to highly active antiretroviral therapy." J Acquir Immune Defic Syndr **27**(2): 105-115.

Nicol, M. R. and A. D. Kashuba (2010). "Pharmacologic opportunities for HIV prevention." Clin Pharmacol Ther **88**(5): 598-609.

O'Brien, S. J. and G. W. Nelson (2004). "Human genes that limit AIDS." Nat Genet **36**(6): 565-574.

Ohkura, S., M. W. Yap, et al. (2006). "All three variable regions of the TRIM5alpha B30.2 domain can contribute to the specificity of retrovirus restriction." J Virol **80**(17): 8554-8565.

Ohnimus, H., M. Heinkelein, et al. (1997). "Apoptotic cell death upon contact of CD4+ T lymphocytes with HIV glycoprotein-expressing cells is mediated by caspases but bypasses CD95 (Fas/Apo-1) and TNF receptor 1." J Immunol **159**(11): 5246-5252.

Ortiz, M., N. Guex, et al. (2009). "Evolutionary trajectories of primate genes involved in HIV pathogenesis." Mol Biol Evol **Dec;26**(12): 2865-2875.

Pandrea, I., D. L. Sodora, et al. (2008). "Into the wild: simian immunodeficiency virus (SIV) infection in natural hosts." Trends Immunol **Sep;29**(9): 419-428.

Perelson, A. S., P. Essunger, et al. (1997). "Decay characteristics of HIV-1-infected compartments during combination therapy." Nature **387**(6629): 188-191.

Perez-Caballero, D., T. Hatziioannou, et al. (2005). "Human tripartite motif 5alpha domains responsible for retrovirus restriction activity and specificity." J Virol **79**(14): 8969-8978.

Perney, P., C. Turriere, et al. (2009). "CXCR3 expression on peripheral CD4+ T cells as a predictive marker of response to treatment in chronic hepatitis C." Clin Immunol **Jul;132**(1): 55-62.

Perron, M. J., M. Stremlau, et al. (2006). "Two surface-exposed elements of the B30.2/SPRY domain as potency determinants of N-tropic murine leukemia virus restriction by human TRIM5alpha." J Virol **80**(11): 5631-5636.

Petit, F., J. Corbeil, et al. (2001). "Role of CD95-activated caspase-1 processing of IL-1beta in TCR-mediated proliferation of HIV-infected CD4(+) T cells." Eur J Immunol **Dec;31**(12): 3513-3524.

Pierson, T. C. and R. W. Doms (2003). "HIV-1 entry and its inhibition." Curr Top Microbiol Immunol **281**: 1-27.

Pillai, S., B. Good, et al. (2003). "A new perspective on V3 phenotype prediction." AIDS Res Hum Retroviruses **19**(2): 145-149.

Poggi, A., R. Carosio, et al. (2004). "Migration of V delta 1 and V delta 2 T cells in response to CXCR3 and CXCR4 ligands in healthy donors and HIV-1-infected patients: competition by HIV-1 Tat." Blood **Mar 15;103**(6): 2205-2213.

Pöhlmann, S., J. Zhang, et al. (2003). "Hepatitis C virus glycoproteins interact with DC-SIGN and DC-SIGNR." J Virol **Apr;77**(7): 4070-4080.

Poignard, P., R. Sabbe, et al. (1999). "Neutralizing antibodies have limited effects on the control of established HIV-1 infection in vivo." Immunity **10**(4): 431-438.

Poon, A. F., L. C. Swenson, et al. (2009). "Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1." Mol Biol Evol **27**(4): 819-832.

Price, D. A., P. J. Goulder, et al. (1997). "Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection." Proc Natl Acad Sci U S A **94**(5): 1890-1895.

Profy, A. T., P. A. Salinas, et al. (1990). "Epitopes recognized by the neutralizing antibodies of an HIV-1-infected individual." J Immunol **144**(12): 4641-4647.

Ptak, R. G., W. Fu, et al. (2008). "Cataloguing the HIV type 1 human protein interaction network." AIDS Res Hum Retroviruses **Dec;24**(12): 1497-1502.

Resch, W., N. Hoffman, et al. (2001). "Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks." Virology **288**(1): 51-62.

Rey-Cuille, M. A., J. L. Berthier, et al. (1998). "Simian immunodeficiency virus replicates to high levels in sooty mangabeys without inducing disease." J Virol **72**(5): 3872-3886.

Reymond, A., G. Meroni, et al. (2001). "The tripartite motif family identifies cell compartments." Embo J **20**(9): 2140-2151.

Ringrose, J. H., R. E. Jeeninga, et al. (2008). "Proteomic studies reveal coordinated changes in T-cell expression patterns upon infection with human immunodeficiency virus type 1." J Virol **82**(9): 4320-4330.

Rotger, M., K. K. Dang, et al. (2010). "Genome-wide mRNA expression correlates of viral control in CD4+ T-cells from HIV-1-infected individuals." PLoS Pathog **6**(2): e1000781.

Rousseeuw, P. (1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." J. Comput. Appl. Math **20**: 53-65.

Ryan, K. and C. Ray (2004). Sherris Medical Microbiology, McGraw Hill.

Sabeti, P. C., P. Varilly, et al. (2007). "Genome-wide detection and characterization of positive selection in human populations." Nature **Oct 18;449**(7164): 913-918.

Sadler, A. J. and B. R. Williams (2008). "Interferon-inducible antiviral effectors." Nat Rev Immunol **Jul;8**(7): 559-568.

Sala, M. and S. Wain-Hobson (2000). "Are RNA viruses adapting or merely changing?" J Mol Evol **51**(1): 12-20.

Samson, M., F. Libert, et al. (1996). "Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene." Nature **382**(6593): 722-725.

Sander, O., T. Sing, et al. (2007). "Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage." PLoS Comput Biol **3**(3): e58.

Sattentau, Q. J. and J. P. Moore (1995). "Human immunodeficiency virus type 1 neutralization is determined by epitope exposure on the gp120 oligomer." J Exp Med **182**(1): 185-196.

Sawyer, S. L., M. Emerman, et al. (2004). "Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G." PLoS Biol **Sep;2**(9): E275.

Sawyer, S. L., L. I. Wu, et al. (2005). "Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain." Proc Natl Acad Sci U S A **Feb 22;102**(8): 2832-2837.

Schindler, M., J. Münch, et al. (2006). "Nef-mediated suppression of T cell activation was lost in a lentiviral lineage that gave rise to HIV-1." Cell **Jun 16;125**(6): 1055-1067.

Schindler, M., J. Schmökel, et al. (2008). "Inefficient Nef-mediated downmodulation of CD3 and MHC-I correlates with loss of CD4+T cells in natural SIV infection." PLoS Pathog **Jul 18;4**(7): e1000107.

Schuitemaker, H., N. A. Kootstra, et al. (1991). "Monocytotropic human immunodeficiency virus type 1 (HIV-1) variants detectable in all stages of HIV-1 infection lack T-cell line tropism and syncytium-inducing ability in primary T-cell culture." J Virol **65**(1): 356-363.

Sebastian, S. and J. Luban (2005). "TRIM5alpha selectively binds a restriction-sensitive retroviral capsid." Retrovirology **2**(40): 40.

Shankarappa, R., J. B. Margolick, et al. (1999). "Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection." J Virol **73**(12): 10489-10502.

Shannon, P., A. Markiel, et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome Res **Nov;13**(11): 2498-2504.

Sharp, P. M., E. Bailes, et al. (2000). "Origins and evolution of AIDS viruses: estimating the time-scale." Biochem Soc Trans **Feb;28**(2): 275-282.

Sheehy, A. M., N. C. Gaddis, et al. (2002). "Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein." Nature **Aug 8;418**(6898): 646-650.

Shepherd, J. C., L. P. Jacobson, et al. (2008). "Emergence and persistence of CXCR4-tropic HIV-1 in a population of men from the multicenter AIDS cohort study." J Infect Dis **198**(8): 1104-1112.

Shioda, T., J. A. Levy, et al. (1992). "Small amino acid changes in the V3 hypervariable region of gp120 can affect the T-cell-line and macrophage tropism of human immunodeficiency virus type 1." Proc Natl Acad Sci U S A **89**(20): 9434-9438.

Shrake, A. and J. A. Rupley (1973). "Environment and exposure to solvent of protein atoms. Lysozyme and insulin." J Mol Biol **79**(2): 351-371.

Silvestri, G. (2005). "Naturally SIV-infected sooty mangabeys: are we closer to understanding why they do not develop AIDS?" J Med Primatol **Oct;34**(5-6): 243-252.

Silvestri, G., M. Paiardini, et al. (2007). "Understanding the benign nature of SIV infection in natural hosts." J Clin Invest **Nov;117**(11): 3148-3154.

Simen, B. B., J. F. Simons, et al. (2009). "Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes." J Infect Dis **199**(5): 693-701.

Simmonds, P., P. Balfe, et al. (1990). "Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1." J Virol **64**(12): 5840-5850.

Sing, T., N. Beerenwinkel, et al. (2004). Learning mixtures of localized rules by maximizing the area under the ROC curve. 1st International Workshop on ROC Analysis in Artificial Intelligence, Valencia, Spain, IOS Press.

Sing, T., A. J. Low, et al. (2007). "Predicting HIV coreceptor usage on the basis of genetic and clinical covariates." Antivir Ther **12**(7): 1097-1106.

Sing, T., O. Sander, et al. (2005). "ROCR: visualizing classifier performance in R." Bioinformatics **21**(20): 3940-3941.

Song, H., E. E. Nakayama, et al. (2007). "A single amino acid of the human immunodeficiency virus type 2 capsid affects its replication in the presence of cynomolgus monkey and human TRIM5alphas." J Virol **81**(13): 7280-7285.

Speck, R. F., K. Wehrly, et al. (1997). "Selective employment of chemokine receptors as human immunodeficiency virus type 1 coreceptors determined by individual amino acids within the envelope V3 loop." J Virol **71**(9): 7136-7139.

Srivastava, V., M. Manchanda, et al. (2009). "Toll-like receptor 2 and DC-SIGNR1 differentially regulate suppressors of cytokine signaling 1 in dendritic cells during Mycobacterium tuberculosis infection." J Biol Chem **Sep 18;284**: 25532-25541.

Stewart, S. A., B. Poon, et al. (2000). "Human immunodeficiency virus type 1 vpr induces apoptosis through caspase activation." J Virol **74**(7): 3105-3111.

Stranger, B. E., M. S. Forrest, et al. (2005). "Genome-wide associations of gene expression variation in humans." PLoS Genet **1**(6): e78.

Stremlau, M., C. M. Owens, et al. (2004). "The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys." Nature **Feb 26;427**(6977): 848-853.

Stremlau, M., M. Perron, et al. (2006). "Specific recognition and accelerated uncoating of retroviral capsids by the TRIM5alpha restriction factor." Proc Natl Acad Sci U S A **103**(14): 5514-5519.

Stremlau, M., M. Perron, et al. (2005). "Species-specific variation in the B30.2(SPRY) domain of TRIM5alpha determines the potency of human immunodeficiency virus restriction." J Virol **79**(5): 3139-3145.

Suttle, C. A. (2005). "Viruses in the sea." Nature **437**(7057): 356-361.

Tang, C., Y. Ndassa, et al. (2002). "Structure of the N-terminal 283-residue fragment of the immature HIV-1 Gag polyprotein." Nat Struct Biol **9**(7): 537-543.

Tardif, M. R. and M. J. Tremblay (2003). "Presence of host ICAM-1 in human immunodeficiency virus type 1 virions increases productive infection of CD4+ T lymphocytes by favoring cytosolic delivery of viral material." J Virol **Nov;77**(22): 12299-12309.

Telenti, A. (2009). "HIV-1 host interactions: integration of large-scale datasets." F1000 Biol Rep **1**.

Thapa, M. and D. J. Carr (2009). "CXCR3 deficiency increases susceptibility to genital herpes simplex virus type 2 infection: Uncoupling of CD8+ T-cell effector function but not migration." J Virol **Sep;83**(18): 9486-9501.

Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-4680.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." J. R. Statist. Soc **B 58**: 267-288.

Tolosi, L. and T. Lengauer (2011). "Classification with correlated features: unreliability of feature ranking and solutions." Bioinformatics **May 16**.

Towers, G. J. (2005). "Control of viral infectivity by tripartite motif proteins." Hum Gene Ther **Oct;16**(10): 1125-1132.

Tozzi, V., R. Libertone, et al. (2008). "HIV pharmacogenetics in clinical practice: recent achievements and future challenges." Curr HIV Res **6**(6): 544-554.

Trautmann, L., L. Janbazian, et al. (2006). "Upregulation of PD-1 expression on HIV-specific CD8+ T cells leads to reversible immune dysfunction." Nat Med **12**(10): 1198-1202.

Trkola, A., T. Dragic, et al. (1996). "CD4-dependent, antibody-sensitive interactions between HIV-1 and its co-receptor CCR-5." Nature **384**(6605): 184-187.

Trkola, A., S. E. Kuhmann, et al. (2002). "HIV-1 escape from a small molecule, CCR5-specific entry inhibitor does not involve CXCR4 use." Proc Natl Acad Sci U S A **99**(1): 395-400.

Turelli, P. and D. Trono (2005). "Editing at the crossroad of innate and adaptive immunity." Science **Feb 18;307**(5712): 1061-1065.

Turville, S. G., P. U. Cameron, et al. (2002). "Diversity of receptors binding HIV on dendritic cell subsets." Nat Immunol **3**(10): 975-983.

Tychonoff, A. (1943). "On the stability of inverse problems." Doklady Akademii Nauk SSSR **39**(5): 195-198.

Van Heuverswyn, F., Y. Li, et al. (2007). "Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon." Virology **368**(1): 155-171.

Van Heuverswyn, F. and M. Peeters (2007). "The Origins of HIV and Implications for the Global Epidemic." Curr Infect Dis Rep **Jul;9**(4): 338-346.

van Marle, G., S. Henry, et al. (2004). "Human immunodeficiency virus type 1 Nef protein mediates neural cell death: a neurotoxic role for IP-10." Virology **Nov 24;329**(2): 302-318.

van Rij, R. P., M. D. Hazenberg, et al. (2003). "Early viral load and CD4+ T cell count, but not percentage of CCR5+ or CXCR4+ CD4+ T cells, are associated with R5-to-X4 HIV type 1 virus evolution." AIDS Res Hum Retroviruses **19**(5): 389-398.

Voight, B. F., S. Kudaravalli, et al. (2006). "A map of recent positive selection in the human genome." PLoS Biol **Mar;4**(3): e72.

Vranken, W. F., M. Budesinsky, et al. (1995). "The complete Consensus V3 loop peptide of the envelope protein gp120 of HIV-1 shows pronounced helical character in solution." FEBS Lett **374**(1): 117-121.

Wang, C., Y. Mitsuya, et al. (2007). "Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance." Genome Res **17**(8): 1195-1201.

Wang, D., A. Urisman, et al. (2003). "Viral discovery and sequence recovery using DNA microarrays." PLoS Biol **1**(2): E2.

Waters, L., S. Mandalia, et al. (2008). "The impact of HIV tropism on decreases in CD4 cell count, clinical progression, and subsequent response to a first antiretroviral therapy regimen." Clin Infect Dis **46**(10): 1617-1623.

Weiss, R. A. (2001). "Gulliver's travels in HIVland." Nature **410**(6831): 963-967.

Wertheim, J. O. and M. Worobey (2009). "Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2." PLoS Comput Biol **May;5**(5): e1000377.

Westby, M., M. Lewis, et al. (2006). "Emergence of CXCR4-using human immunodeficiency virus type 1 (HIV-1) variants in a minority of HIV-1-infected patients following treatment with the CCR5 antagonist maraviroc is from a pretreatment CXCR4-using virus reservoir." J Virol **80**(10): 4909-4920.

Westby, M., C. Smith-Burchnell, et al. (2007). "Reduced maximal inhibition in phenotypic susceptibility assays indicates that viral strains resistant to the CCR5 antagonist maraviroc utilize inhibitor-bound receptor for entry." J Virol **81**(5): 2359-2371.

Westby, M. and E. van der Ryst (2005). "CCR5 antagonists: host-targeted antivirals for the treatment of HIV infection." Antivir Chem Chemother **16**(6): 339-354.

Whitcomb, J. M., W. Huang, et al. (2007). "Development and characterization of a novel single-cycle recombinant-virus assay to determine human immunodeficiency virus type 1 coreceptor tropism." Antimicrob Agents Chemother **51**(2): 566-575.

Wodarz, D. and C. R. Bangham (2000). "Evolutionary dynamics of HTLV-I." J Mol Evol **50**(5): 448-455.

Wolinsky, S. M., B. T. Korber, et al. (1996). "Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection." Science **272**(5261): 537-542.

Woolhouse, M. E., J. P. Webster, et al. (2002). "Biological and biomedical implications of the co-evolution of pathogens and their hosts." Nat Genet **Dec;32**(4): 569-577.

Word, J. M., S. C. Lovell, et al. (1999). "Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms." J Mol Biol **285**(4): 1711-1733.

Yamaguchi-Kabata, Y. and T. Gojobori (2000). "Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes." J Virol **74**(9): 4335-4350.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **Aug;24**(8): 1586-1591.

Yang, Z. and R. Nielsen (2000). "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models." Mol Biol Evol **Jan;17**(1): 32-43.

Yang, Z., R. Nielsen, et al. (2000). "Codon-substitution models for heterogeneous selection pressure at amino acid sites." Genetics **155**(1): 431-449.

Yang, Z., W. S. Wong, et al. (2005). "Bayes empirical bayes inference of amino acid sites under positive selection." Mol Biol Evol **Apr;22**(4): 1107-1118.

Yap, M. W., S. Nisole, et al. (2005). "A single amino acid change in the SPRY domain of human Trim5alpha leads to HIV-1 restriction." Curr Biol **15**(1): 73-78.

Zanotto, P. M., E. G. Kallas, et al. (1999). "Genealogical evidence for positive selection in the nef gene of HIV-1." Genetics **153**(3): 1077-1089.

Zhang, J. and D. M. Webb (2004). "Rapid evolution of primate antiviral enzyme APOBEC3G." Hum Mol Genet **Aug 15;13**(16): 1785-1791.

Zhang, L., R. S. Diaz, et al. (1997). "Host-specific driving force in human immunodeficiency virus type 1 evolution in vivo." J Virol **71**(3): 2555-2561.

Zhang, L. Q., P. MacKenzie, et al. (1993). "Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection." J Virol **67**(6): 3345-3356.

Zhou, H., M. Xu, et al. (2008). "Genome-scale RNAi screen for host factors required for HIV replication." Cell Host Microbe **Nov 13;4**(5): 495-504.

Zufferey, R., T. Dull, et al. (1998). "Self-inactivating lentivirus vector for safe and efficient in vivo gene delivery." J Virol **72**(12): 9873-9880.