Biological Engineering and Characterization of an HIV-1 Envelope-based Genomic Library

DISSERTATION ZUR ERLANGUNG
DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR
BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

Vorgelegt von

Julia Koop

aus

Karaganda, Kasachstan

im Jahr

2018

Biological Engineering and Characterization of an HIV-1 Envelope-based Genomic Library

DISSERTATION ZUR ERLANGUNG
DES DOKTORGRADES DER NATURWISSENSCHAFTEN (DR. RER. NAT.)
DER FAKULTÄT FÜR
BIOLOGIE UND VORKLINISCHE MEDIZIN
DER UNIVERSITÄT REGENSBURG

Vorgelegt von

Julia Koop

aus

Karaganda, Kasachstan

im Jahr

2018

Das Promotionsgesuch wurde eingereicht am:
26.02.2018

Die Arbeit wurde angeleitet von:
Prof. Dr. Ralf Wagner

Meinen Eltern

Table of Contents

Zusammenfassung

Die zahlreichen immunologischen Ausweichstrategien, welche im HIV-1 envelope (Env) Glykoprotein verkörpert sind, stellen für die Entwicklung eines sicheren und effektiven Vakzins weiterhin ein enormes Hindernis dar. Zu den herausfordernsten Ausweichmechanismen zählt die unermessliche genetische Diversität, welche mit der Immundominanz hoch variabler Regionen von Env assoziiert wird. Die derzeitigen Impfstoffansätze zielen darauf hin, breitneutralisierende Antikörper (bnAK) hervorzurufen, von denen einige in der Lage sind mehr als 90% der kursierenden HIV-1 Stämme zu neutralisieren. Allerdings wird die Entwicklung von bnAK aufgrund der komplexen Koevolution von Virus und humoraler Immunantwort erheblich beeinträchtigt. Daher bedarf es neuartiger Env Immunogene sowie innovativer Selektionstechnologien für deren Identifikation, um diesen Prozess zu begünstigen.

Der erste Abschnitt dieser Dissertation beschäftigte sich intensiv mit der biologischen Prozesstechnik einer auf Env basierten sequentiellen Permutationsbibliothek, mit besonderem Schwerpunkt auf Charakterisierung und Qualitätskontrolle der Bibliothek. Jede Position des außenliegenden Env-Bereiches wurde durch 20 natürliche Aminosäuren ersetzt, wodurch eine Bibliothek bestehend aus 658 Unterbibliotheken und schätzungsweise 13.000 Varianten hervorgeht. Gleichzeitig wurden die jeweiligen stabilen Zelllinien durch stabile Transfektion jeder Unterbibliothek in Flp-In™ T-Rex 293 Zellen hergestellt. Das Ziel bestand darin, diese Bibliothek einer Selektionstechnologie beruhend auf einer Zellsortierung zu unterziehen, um Env Varianten mit verbesserter Antigenität zu identifizieren. Sowohl die Plasmid-DNA- (pDNA), als auch die Zelllinien-Bibliothek wurden umfassend auf ihre Qualität kontrolliert. Die eingehende Analyse der pDNA offenbarte Deletionen verschiedenster Länge hauptsächlich in der Env-Region, welche etwa 48% der Bibliothek betreffen. Allerdings traten diese Deletionen in einem kleinen Bruchteil innerhalb der Unterbibliotheken auf, womit die tatsächliche Kontamination jeweils nur zwischen 6-18% lag. Mit dem Schwerpunkt auf der CD4-Bindestelle von Env, wurden Diversität und Verteilung der Aminosäuren der pDNA, sowie der stabilen Zelllinien mittels Next Generation Sequencing (NGS) ermittelt. Während die pDNA eine durchschnittliche Variabilität von 19 Aminosäuren und eine nahezu ideale Verteilung aufwies, zeigten die stabilen Zelllinien sowohl einen etwa 38%-igen Rückgang in der Diversität, als auch eine beträchtliche und zufallsbedingte Ungleichverteilung der Aminosäuren. Es wurde ersichtlich, dass die unzureichende Integration von Env bei der Herstellung der stabilen Zelllinien maßgeblich zu diesem Variabilitätsverlust beitrugen. Dementsprechend wurden einige vielversprechende Ansätze zur Optimierung der Herstellung stabiler Zelllinien eingeleitet, mit dem Ziel eine bessere Diversität und Aminosäureverteilung zu erlangen.

Das zweite Projekt beruhte auf der Identifikation von verbesserten Env-Kandidaten mit vorteilhaftem Antigenitätsprofil. Zu diesem Zweck wurde eine Selektionstechnlogie angewendet, die auf einer Zellsortierung beruht und folgende Vorteile in sich vereinigt: i) Integration einer einzigen Env Variante in eine definierte FRT-Stelle pro Zelle, was eine Kopplung zwischen Geno- und Phänotyp zur Folge hat, ii) induzierbare Env Expression, um

Zytotoxizitätseffekten vorzubeugen, iii) translationale Verknüpfung von GFP und Env zur indirekten Normalisierung auf die induzierte Env Expression und iv) Expression der Env auf Hek293T Zellen, um native Faltung und Säugetierglykosylierung zu gewährleisten. In einem einzelnen Selektionszyklus wurden jeweils 12 Env Varianten mit erhöhter oder verminderter Affinität für den bnAK VRC01 aus der Zelllinien-Bibliothek angereichert. Auffallend dabei war, dass keine der Varianten mit erhöhter, und nur drei Varianten mit erniedrigter Bindungsfähigkeit mittels FACS-basierter Gleichgewichtstitration eindeutig validiert werden konnten. Da die Selektionstechnologie zuvor an einer Bibliothek getestet wurde, welche nur fünf Varianten umfasst, lag es der Vermutung nahe, dass die Methoden für komplexere und größere Bibliotheken weiter ausgebaut und adaptiert werden müssen.

Abstract

The numerous immune evasion strategies embodied in the HIV-1 envelope (Env) glycoprotein still represent a daunting challenge in the development of a safe and effective vaccine. Among the most defying of these evasive mechanisms is the tremendous genetic diversity associated with the immunodominance of highly variable regions of Env. Current vaccine design efforts aim to elicit broadly neutralizing antibodies (bnAb), some of which are able to neutralize more than 90% of circulating HIV-1 strains. However, a complex co-evolution of virus and humoral immune response considerably impairs the development of bnAbs. To facilitate this process, novel Env immunogens as well as innovative selection technologies for their identification are required.

The first part of this thesis concentrated on the biological engineering of an envelope-based sequential permutation library, specifically focusing on characterization and quality control of the library. Each residue in the external part of Env was substituted by 20 natural amino acids, thus creating a library of 658 sublibraries and approximately 13.000 variants. Simultaneously, the respective stable cell lines (SCL) were generated by stably transfecting every sublibrary into Flp-In™ T-Rex 293 cells with the goal to utilize the stable cell line library in a mammalian cell display- and cell sorting-based screening technology to identify Env variants with improved antigenicity. Comprehensive quality controls of plasmid DNA library and the respective stable cell line library were conducted to assess potential limitations. In-depth analysis of the pDNA revealed deletions of various lengths mainly in the Env region affecting about 48% of the library. However, these deletions occurred only in a small fraction within the sublibraries, thus the actual contaminations amounted to 6-18%, respectively, deeming the library still eligible to work on. Focusing on the CD4 binding site (CD4bs) of Env, diversity and amino acid distribution of the pDNA- and the stable cell line-library was analyzed by Next Generation Sequencing (NGS). Whereas pDNA exhibited an average diversity of 19 amino acids in the sublibraries with a nearly ideal distribution, stable cell lines demonstrated a considerable decrease in diversity by approximately 38%, as well as a highly uneven and random distribution of amino acids. It became apparent that particularly insufficient integration of Env during the generation of stable cell lines contributed to this substantial loss of diversity. Accordingly, several promising approaches were tested to optimize the stable cell line generation aimed to improve the diversity and amino acid distribution.

The second project focused on the identification of improved Env candidates with favorable antigenicity from the stable cell line library. For this purpose, a mammalian cell display- and cell sorting-based technology was applied that combines the benefits of i) single integration of Env into a distinct FRT site resulting in the linkage of genotype and phenotype, ii) inducible Env expression to prevent cytotoxicity effects, iii) translational coupling of Env and GFP enabling an indirect normalization for induced Env expression and iv) display on Hek293T cells, thus ensuring native folding and mammalian glycosylation. Using the CD4bs SCL library, twelve Env variants demonstrating increased (gain of binding, GoB) and decreased (loss of binding, LoB) affinity for the bnAb VRC01, respectively, were

selected in a single round of cell sorting procedure. Strikingly, none of the detected GoB variants and merely three LoB candidates could be unequivocally validated by means of a FACS-based equilibrium titration. As the selection technology was previously tested on a five-variant library, there were grounds for supposition, that the methods require further development and adaptation to be utilized for more complex and extensive libraries.

1 Introduction

1.1 Epidemiology of HIV

In June 1981, the U.S. Centers for Disease Control and Prevention (CDC) released a report describing cases of a rare lung infection called Pneumocystis carinii pneumonia (PCP) in five young, previously healthy gay men in Los Angeles [1]. Concurrently, an increased incidence of an unusually aggressive cancer known as Kaposi's Sarcoma was recognized in New York and California [2]. At this point in time, no one established a connection between the two obviously different diseases. It was only two years later that scientists discovered a common thread of impaired cellular immunity that linked these malignancies and other opportunistic infections [3,4]. Eventually, the human immunodeficiency virus (HIV) was identified as causative agent of the substantially increasing cases of severe immune deficiency worldwide. Due to symptoms and progression of an HIV infection, the term acquired immune deficiency syndrome (AIDS) was established by the CDC in 1982. Since then the virus spread globally, causing one of the most debilitating pandemics ever recorded in human history.

Approximately 80% of HIV infections occur during sexual intercourse with an infected partner through direct contact with semen and rectal or vaginal fluids [5]. Blood-to-blood transmissions such as through sharing of needles or contaminated blood transfusions [6], as well as mother-to-child transmissions during pregnancy, childbirth or breastfeeding [7], represent another 20% of all contracted HIV infections.

Natural progression of HIV infection encompasses three stages: an acute phase, followed by an early/clinically latent phase, and finally by the immune collapse/AIDS. The acute or primary phase lasts several months and is characterized by high level viral replication that is reflected in substantial concentrations of virus in plasma and lymphoid tissue. After initial viral decline, concurrent with the appearance of virus-specific $CD8^+$ cytotoxic T cells [8], the plasma viral load usually stabilizes at a steady state. This so-called 'set-point' is the consequence from the equilibrium between the HIV-1 replication and the corresponding immune responses and represents the beginning of the second stage, a long clinical latency. Ultimately, the regenerative $CD4^+$ T cell population slowly diminishes below a crucial threshold rendering the immune system vulnerable to opportunistic infections, thus causing progression to AIDS.

According to the World Health Organization (WHO) more than 70 million people have contracted HIV and an estimated 35 million people have died from AIDS-related illness since the beginning of the pandemic. As of 2015, approximately 36.7 million [34.0–39.8 million] individuals were living with HIV, representing 0.8 % [0.7–0.9 %] of adults aged 15-49 years worldwide (figure 1). Although the prevalence of HIV varies considerably among countries, 70 % of all accounted infections arise in Africa, specifically in the Sub-Saharan regions. While there is currently no cure for HIV, the infection can be

suppressed by a combination of antiretroviral drugs, thus substantially reducing morbidity and mortality. At the moment, approximately 18.2 million [16.1-19.0 million] HIV patients are receiving antiretroviral agents, termed combination antiretroviral treatment (cART). However, the low treatment rate, in addition to severe side effects from the medicaments, drug interactions and resistance demonstrate the importance of discovering a vaccine to finally conquer HIV infections globally.

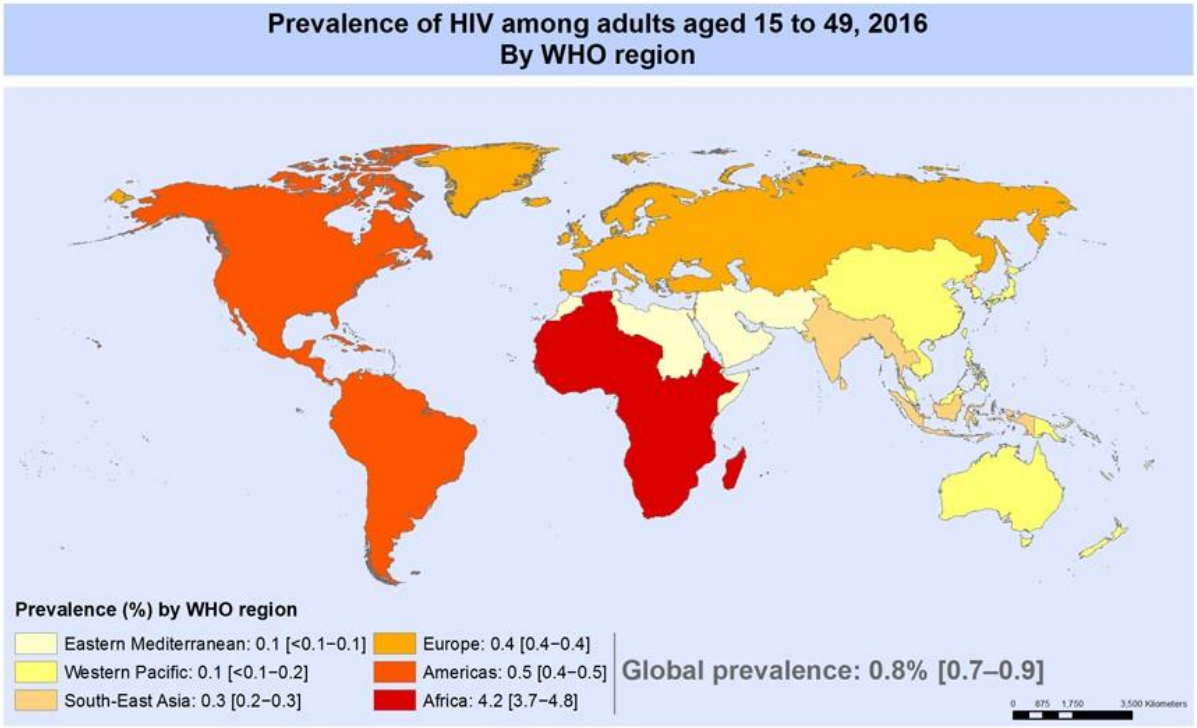


Figure 1 - Global prevalence of HIV in 2016. The illustration shows that an estimated 0.8% [0.7-0.9%] of adults aged 15-49 years worldwide are infected with HIV. Areas that are most severely affected, such as Sub-Saharan Africa, are indicated in dark red. Figure was adapted from WHO Global Health Observatory (GHO) data HIV/AIDS.

1.2 Origin and phylogeny of HIV

HIV appears to have its origin in the simian immunodeficiency virus (SIV) that infects non-humate primates in West and Central Africa. Zoonotic transmission presumably occurred as a consequence of hunting and butchering of primates and keeping of monkeys as pets [9,10]. Two distinct HIV types emerged from the transmissions, HIV type 1 (HIV-1) that descends from chimpanzees [9,11], and HIV type 2 (HIV-2) which is closely related to the SIV of sooty mangabeys [12]. Whereas HIV-2 is relatively uncommon and majorly concentrated in West Africa, HIV-1 represents the predominant virus worldwide.

HIV is characterized by tremendous genetic variability and rapid evolution. Several factors contribute to the extensive heterogeneity, such as the error-prone nature of the HIV-1 reverse transcriptase (RT) [13], host selective immune pressure [14,15], as well as genetic recombination events during replication [16]. Due to this variability, the HIV-1 strains can be classified into four phylogenetic groups, which constitute the groups M (main), O (outlier), N (new or non-M, non-O) and P [17,18]. Among these, group M viruses are globally the most prominent and can be further divided into nine genetically distinct subtypes or clades (A-D, F-H, J and K) [19]. Furthermore, recombination events between strains and groups give rise to an increasing number of circulating recombinant forms (CRFs) [20,21] (figure 2).

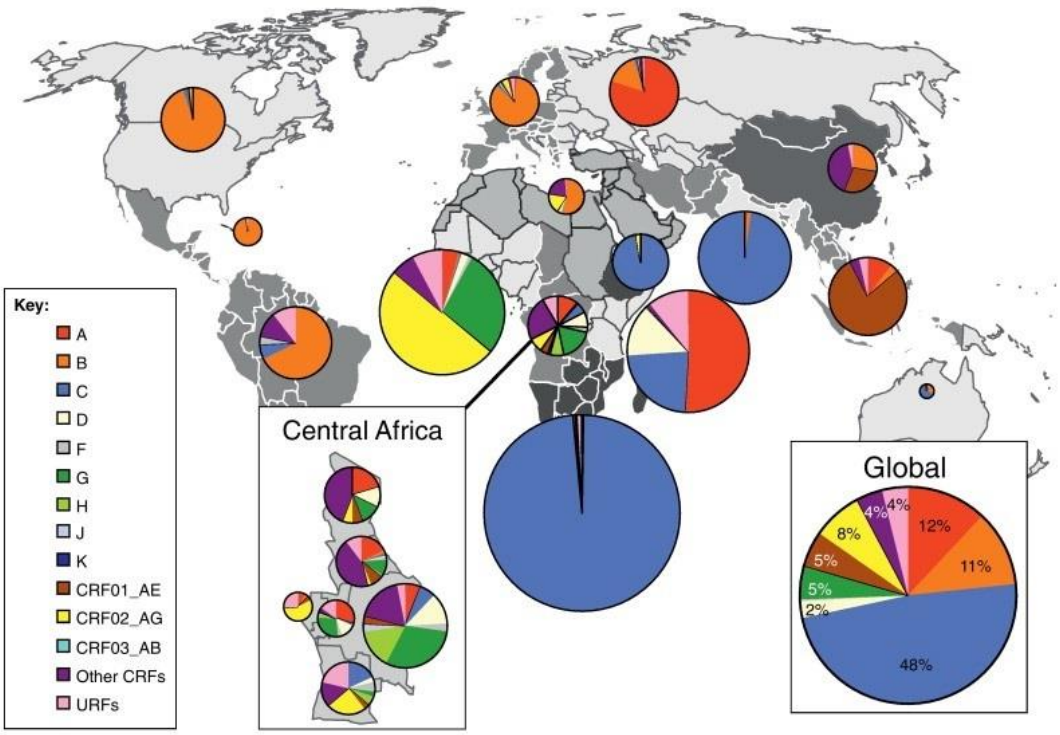


Figure 2 - Global distribution of HIV-1 subtypes and recombinants. Pie charts illustrate the percentage distribution of HIV-1 subtypes represented by different colors in each region. The distribution was calculated according to data gathered from 2004 to 2007. The figure was adapted from [22].

1.3 Genomic organization and structural biology

According to the International Committee on Taxonomy of Viruses (ICTV) the human immunodeficiency virus (HIV) is classified as a *Retrovirus* belonging to the genus *Lentivirus*. It features a roughly spherical morphology with a diameter of about 145 nm [23]. The approximately 10 kb genome is situated in the viral capsid as two non-covalently

linked positive stranded RNA molecules [24,25], and comprises nine open reading frames coding for 15 mature proteins (figure 3) which are divided into three classes [26,27]: i) the major structural proteins, Gag, Pol and Env, ii) the regulatory proteins, Tat and Rev and iii) the accessory proteins, Vpu, Vpr, Vif, and Nef.

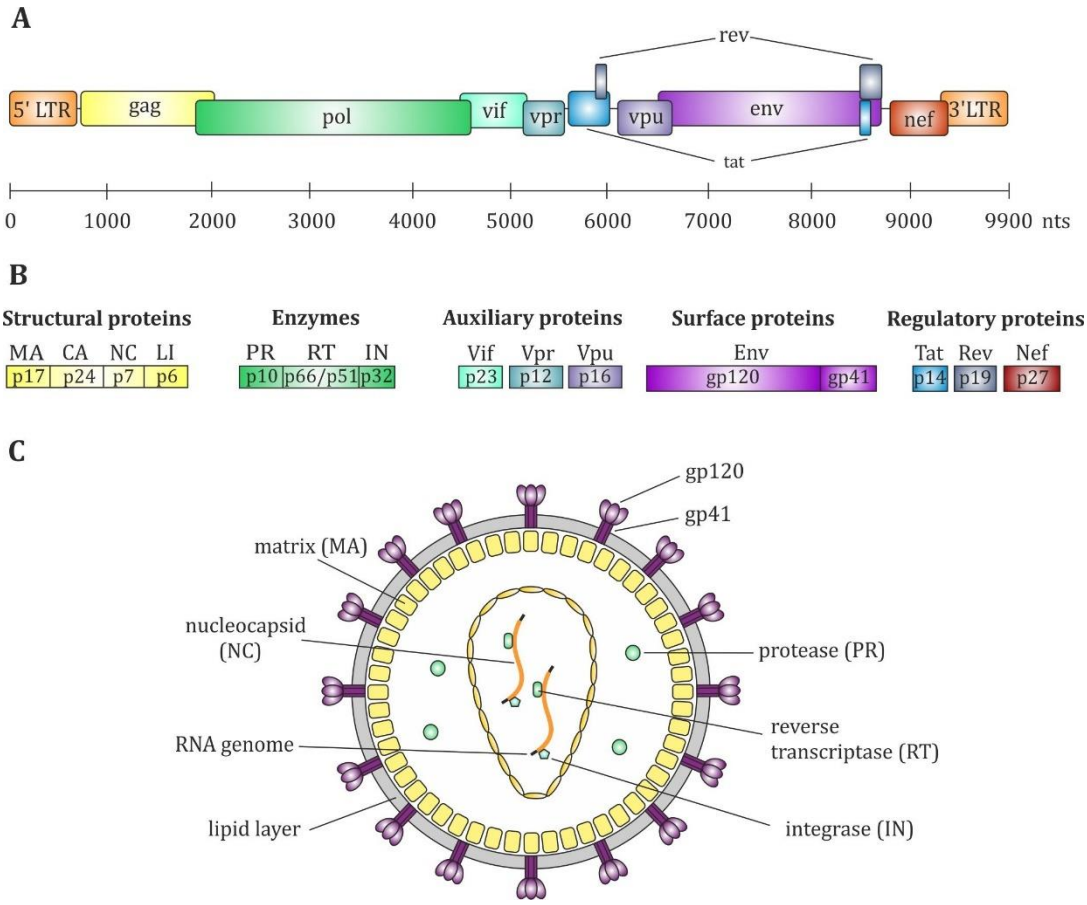


Figure 3 – Genomic organization and structure of HIV. (A) Structure of the RNA genome of HIV-1 that consists of roughly 10.000 nucleotides (nts). Open reading frames of nine genes are shown as rectangles that overlap in some cases. **(B)** The HIV genome encodes 15 proteins that are categorized into enzymes (PR – protease, RT – reverse transcriptase, IN – integrase) as well as structural (MA - matrix, CA – capsid, NC – nucleocapsid), auxiliary, surface (Env – envelope, gp – glycoprotein) and regulatory proteins. Colors of genes and their respective gene products are matched. **(C)** Schematic structure of a mature HIV-1 particle. The enveloped virus features one surface protein, the trimeric envelope glycoprotein (Env) comprising three gp120 and g41 subunits, respectively. Matrix proteins (MA, p17) line the host-derived membrane. The conical capsid (CA, p24) contains two copies of (+)-strand RNA molecules complexed with the nucleocapsid protein (NC, p7). The viral enzymes protease (PR, p10), reverse transcriptase (RT, p66/p51), integrase (IN, p32) are indicated in green, whereas the auxilliary (Vif, Vpr, Vpu) and regulatory proteins (Tat, Rev, Nef) are not shown.

Gag is synthesized as a 55 kDa ($Pr55^{Gag}$) precursor polyprotein on cytosolic ribosomes and contains matrix (MA, p17), capsid (CA, p24), nucleocapsid (NC, p7), p6 domains, as well as two spacer peptides SP1 & SP2, thus, comprising all of the viral elements required

for virus assembly [28–30]. Every Gag protein (MA, CA, NC, p6) performs distinct functions during the viral assembly. The viral genome of HIV-1 is housed within a capsid that assembles into a conical outer shell [31,32]. Matrix proteins are responsible for intracellular trafficking and binding of Gag to the plasma membrane [33,34], as well as directing the incorporation of the sole surface envelope glycoprotein (Env) into virions [33,35]. NC serves as facilitator for viral replication [36] and is a key component of RNA packaging, as well as Gag multimerization [31,37]. Lastly, the p6 domain mediates budding and release of viral particles from the plasma membrane [38,39].
All essential enzymatic functions are provided by the three Gag-Pol proteins ($Pr160^{Gag\text{-}Pol}$), protease (PR), reverse transcriptase (RT) and integrase (IN) [27,40].
HIV-1 entry into host cells is initiated by envelope glycoproteins by mediating virion attachment [41,42], as well as interaction with cellular CD4- and co-receptors [43,44] (see below).
The proteins Tat and Rev assist in essential gene regulatory functions [45,46], while the four accessory proteins Vif, Vpr, Nef and Vpu contribute to infectivity and evasion of immune mechanisms [47–49].

1.4 The HIV-1 life cycle

HIV is able to infect cells which express CD4 molecules on their surface. Primarily, these are macrophages and $CD4^+$ T cells [50,51]. In this context, the HIV-1 envelope (Env) glycoprotein is crucial in the virus replication cycle by mediating the fusion between viral and host cellular membranes during the entry process. After attachment of Env to the cellular surface [41,42,52] and subsequent binding to the CD4 receptor [53–55] (figure 4), a cascade of conformational changes in gp120 and gp41 occurs [56], augmenting its affinity for a co-receptor [57]. The relevant chemokine co-receptors for HIV-1 are CCR5 (R5) and CXCR4 (X4) [58,59]. Upon engagement of gp120 with the co-receptor, additional conformational changes in gp41 trigger a membrane fusion reaction that delivers the viral core into the host cell [60–62]. Subsequently, the viral RNA genome is transcribed into double-stranded DNA by the viral enzyme reverse transcriptase (RT) [63]. Following synthesis, viral DNA is translocated across a nuclear pore in the form of a nucleoprotein complex (pre-integration complex, PIC) into the nucleus and integrated as a provirus into the host cell genome [64], leading to a life-long reservoir of infected $CD4^+$ T cells. The virus-encoded integrase (IN) protein is a component of the PIC that mediates the integration process [65]. After transcription, viral RNAs are transported into the cytoplasm where translation of the viral proteins occurs. At the plasma membrane, virion assembly takes place, wherein newly synthesized proteins as well as two single-stranded copies of viral RNA are packaged and bud from the cell as immature particles [66]. Concomitant with virion release, maturation takes place by proteolytic processing of Gag which leads to a morphological rearrangement within the particle [66–68]. The resulting virus is then able to infect new cells.

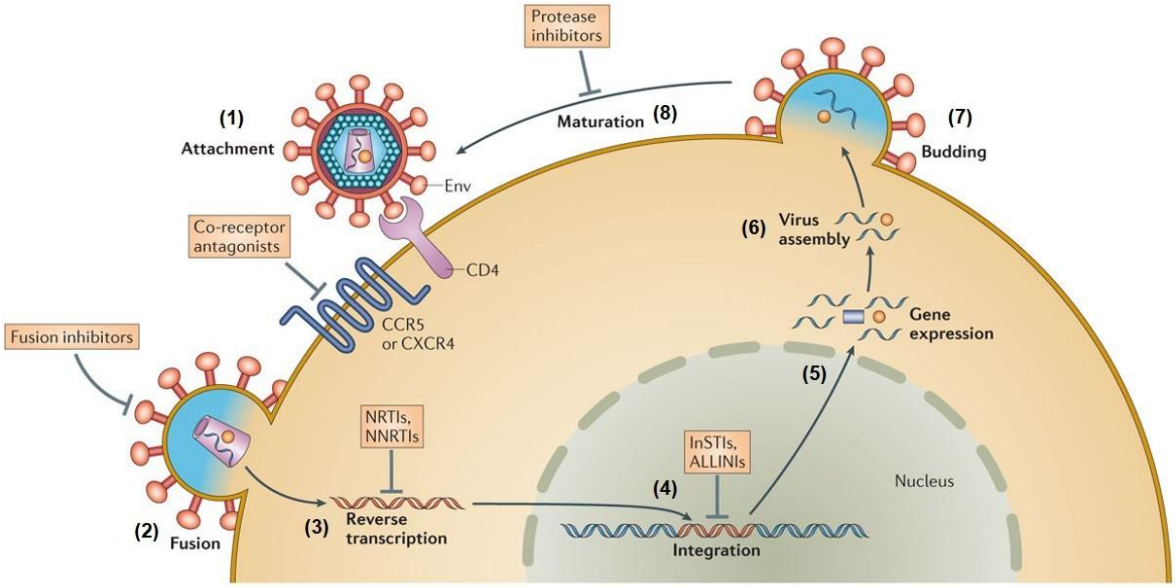


Figure 4 - Schematic illustration of the main steps in the HIV-1 life cycle: (1) After attachment of the viral Env glycoprotein to cell surface proteins CD4 and a co-receptor (CCR5 or CXCR4), fusion of the viral and host cell membranes is mediated (2) enabling entry of the viral capsid into the cell. Once the capsid is uncoated and the viral RNA along with viral proteins are released into the cytoplasm, RNA is reverse transcribed to double stranded DNA (3) and translocated into the cell nucleus. Following successful integration (4), transcription of the provirus takes place resulting in viral RNAs, which are translated into proteins and transported (5) from the nucleus. Upon arrival on the cell surface, viral RNA and proteins are assembled into immature virions (6) that bud from the cell (7). Proteolytic processing of polyproteins initiates maturation (8), resulting in mature virions that are capable of infecting new cells. Many steps of the HIV life cycle can be inhibited by drugs which are displayed in the rectangles. With permission from [69].

1.5 The envelope glycoprotein

The envelope glycoprotein (Env) is one of the most important proteins of HIV as it mediates host cell entry by binding to CD4 receptors. In addition, Env represents the sole target for the host's humoral immune system, and therefore serves as target for HIV-1 neutralizing antibodies [58,70,71]. Thus, Env is the major subject of investigation in respect to vaccine development which focusses particularly on the humoral immune response to the protein.

1.5.1 Env synthesis and trafficking

Env proteins are synthesized as heavily glycosylated gp160 polyprotein precursor molecules from a singly spliced, bicistronic *vpu/env* mRNA on the rough endoplasmic reticulum [72,73]. After folding and oligomerizing [74,75], gp160 is transported to the Golgi

complex where it is subjected to various processing events, such as oligosaccharide modification and proteolytic cleavage [76]. Proteolytic processing is mediated by cellular furin proteases within the *trans*-Golgi network (TGN) to yield the gp120 and the gp41 subunits that are required for viral infection of HIV-1 [77–79]. Three molecules each of gp120 and gp41 assemble into the final heterotrimeric Env spike, held together by meta-stable, non-covalent interactions [80,81]. Following exit from the TGN, the glycoproteins traverse to the plasma membrane [74] where Env either interacts with Gag and gets incorporated into viral particles, or alternatively is endocytosed again [82,83]. In general, an average of 14 to 20 trimers are integrated into virions [84]. Endocytosis of Env or disintegration of the trimeric structure into monomeric gp120 and gp41, also termed 'shedding', as a result of the non-covalent gp120-gp41 interactions can be attributed to the low incorporation events [85,86]. Presumably, low surface spike density serves as an evasion mechanism against the host immune system [87].

1.5.2 Env structure

The envelope glycoprotein is a trimer of heterodimers comprising a complex of trimeric gp120 and gp41, respectively (figure 5C). The gp120 subunit is divided into discontinuous segments of constant and variable regions (figure 5A). Five variable domains (V1-V5) alternate with five relatively constant domains (C1-C5) [37,88–90]. As already indicated by the name, variable regions feature a high degree of sequence and length diversity derived from recombination events, point mutations, insertions and deletions, with the V1V2 domain having the most variation in loop length (50-90 amino acids) and number of glycosylation sites. Typically, the variable regions are arranged in loops which are separated and delimited by disulfide bonds. 18 highly conserved cysteine residues, located throughout gp120 and gp41, form nine intramolecular disulfide bridges that are crucial in establishing the proper tertiary structure of Env [91,92]. However, no disulfide bridge resides between the gp120 and g41 subunits.
Several N-linked glycans, with a small additional contribution of O-linked sugars, are located on the surface of gp120 (figure 5 D) comprising about 50% of its total mass. Importantly these glycans have been shown to protect Env from host immune recognition, to influence Env conformation/oligomerization, as well as to affect viral entry, infectivity and antibody recognition [93].
The gp120 core consists of a highly conserved inner domain (figure 5B), facing the trimer axis, and a heavily glycosylated outer domain, which is mostly exposed on the surface of the trimer [94,95]. One of the most relevant features of gp120 is represented by the CD4 binding site (CD4bs) (figure 5C), which comprises the principal contact sites of CD4. It is arranged in six discontinuous segments , consisting of residues that are highly conserved [92,94,96,97]. Considering the functional conservation among diverse HIV-1

isolates, the CD4 binding site is a favorable target for neutralizing antibodies, and thus also for vaccine design.

Anchored in the viral membrane, the gp41 subunit of Env comprises three major domains: an ectodomain, a transmembrane domain (TM), and a cytoplasmic tail (CT) [98] (figure 5A). All major fusion determinants are located in the ectodomain, including an N-terminal fusion peptide (FP) [99,100], two hydrophobic heptad repeat regions (HR1 and HR2) [101,102] (figure 5B) and a highly conserved tryptophane-rich domain referred to as the membrane-proximal external region (MPER) [103,104]. The gp41 TM anchors Env in the lipid bilayer and is involved in fusion and modulation of immune responses during viral infection [61,105,106]. Last but not least, the cytoplasmic tail mediates intracellular trafficking and incorporation of Env into virions [37,83].

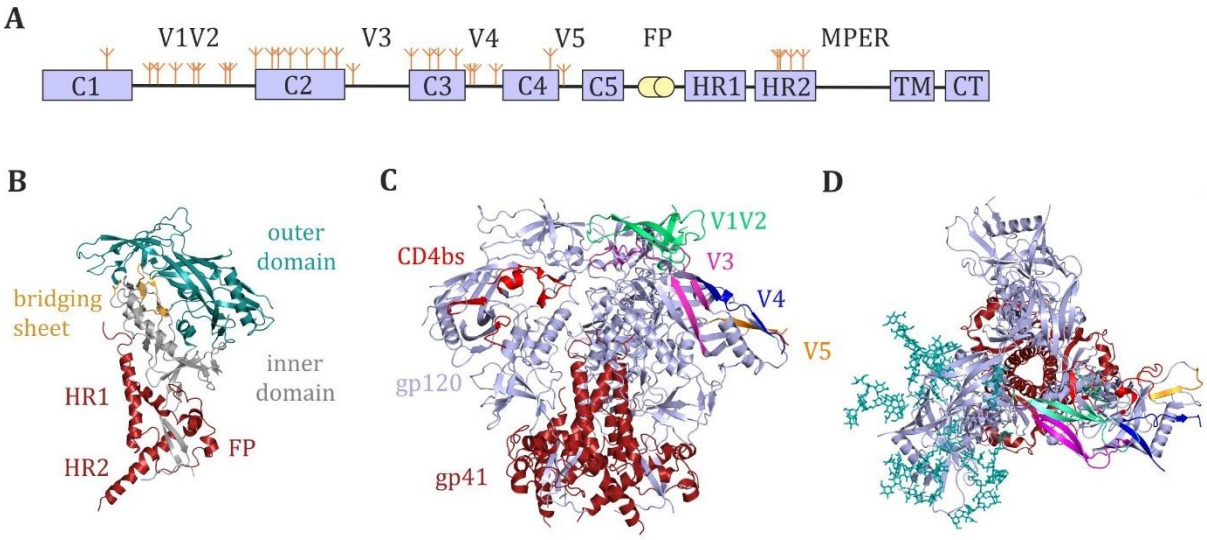


Figure 5 – Structure of the HIV-1 envelope glycoprotein. Structures are based on the BG505 DS SOSIP trimer (PDB 5U1F) **(A)** Schematic representation of the HIV-1 gp160 envelope. The gp120 trimer comprises five constant regions (C1-C5) that are interspersed with five variable regions (V1-V5). The fusion peptide, heptad repeats 1 and 2, membrane proximal external regions, transmembrane domain (TM) and cytoplasmic tail (CT) are located in the gp41 trimer. Glycans are represented by tree-like symbols. **(B)** Structure of an Env protomer consisting of an outer and inner domain that are connected by the bridging sheet. The heptad repeats 1 and 2 (HR1, HR2) are located at the base of gp41, whereas fusion peptide is positioned at the interface of gp120. **(C)** Side and **(D)** top views of the Env trimer. Variable loops (V1-V5) and the CD4 binding site (CD4bs) are shown. Structures of the membrane proximal external region (MPER), transmembrane domain (TM) and cytoplasmic tail (CT) are not included in the illustration since they have not yet been determined. Glycans are shown in teal (only in one protomer). Figure was freely reproduced from [70].

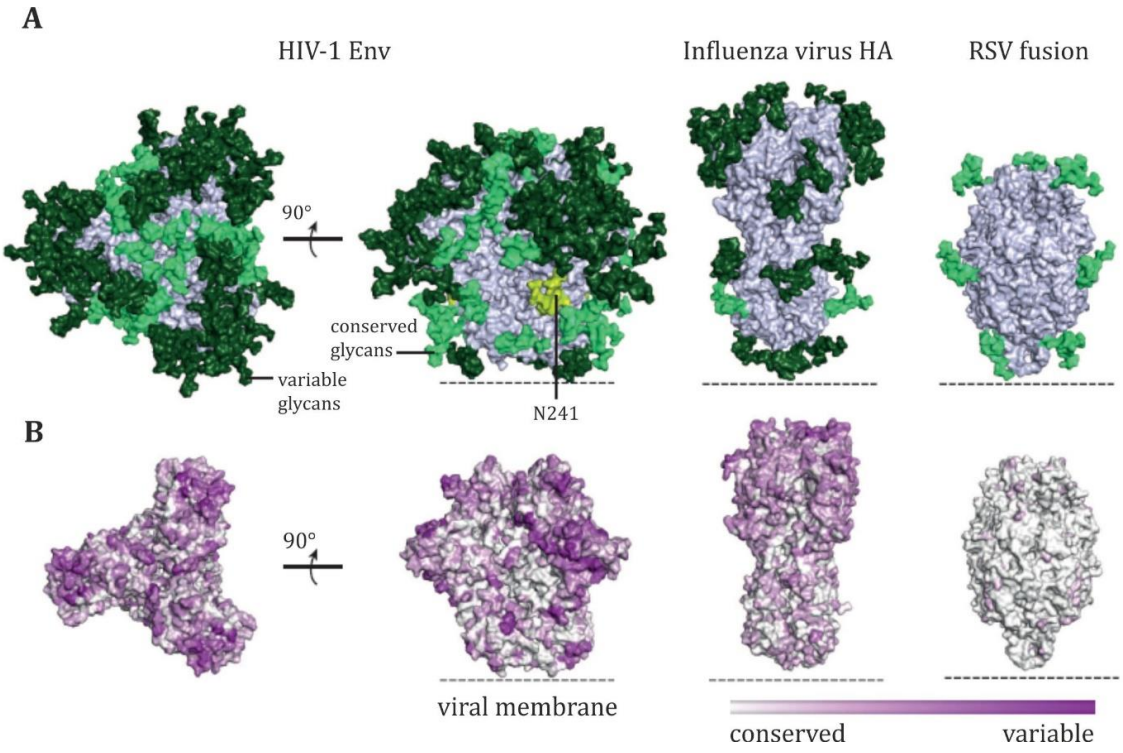
1.5.3 Immune evasion mechanisms of Env

The virus features a multitude of evasion strategies to escape an efficient humoral immune response, most of them embodied in structural properties of the envelope protein. As previously mentioned, low density of viral spikes (14-20) on the surface of HIV virions [107] as well as shedding of Env represent effective evasion strategies [108]. The

tremendous genetic diversity of Env which can exhibit up to 35% sequence variability between subtypes and 20% within a clade [108] is particularly problematic for HIV vaccine design. This diversity is a result of the error-prone nature of the reverse transcriptase and high rates of viral replication [109,110]. Many structures, especially the five variable loops (V1-V5), possess a high-level tolerance for point mutations [111,112], and even insertion and deletion of whole sequence stretches without loss of viral fitness [113,114]. As a consequence, a multitude of escape variants can arise in fast succession, thus continuously evading the host's hummoral immune responses. However, as Env is essential for cell entry, the variability is limited to non-conserved regions in order to maintain its functions.

In addition to the vast sequence diversity of the variable regions, the location and arrangement of structural features of Env lead to conformational masking. This phenomenon describes the capability of certain structures to conceal functionally essential regions of HIV from the immune system. For instance, Env trimer formation results in the burial of neutralizing epitopes within oligomeric interfaces [115,116]. Furthermore, variable loops can successfully occlude conserved regions such as the CD4 binding site, thus restricting access for neutralizing antibodies [117,118]. Extensive glycosylation covering the surface of Env is also able to shield exposed surfaces [119]. In combination with the ability of repositioning of glycans in response to the selection pressure, this 'glycan-shield' (figure 6A) limits immunogenicity and obstructs binding of certain antibodies to Env.

Last but not least, unliganded Env was revealed to be intrinsically dynamic, transitioning between different conformations [120] (figure 6C). During this so-called 'breathing' different non-essential epitopes are presented to the immune system resulting in the generation of non-neutralizing antibodies [121].



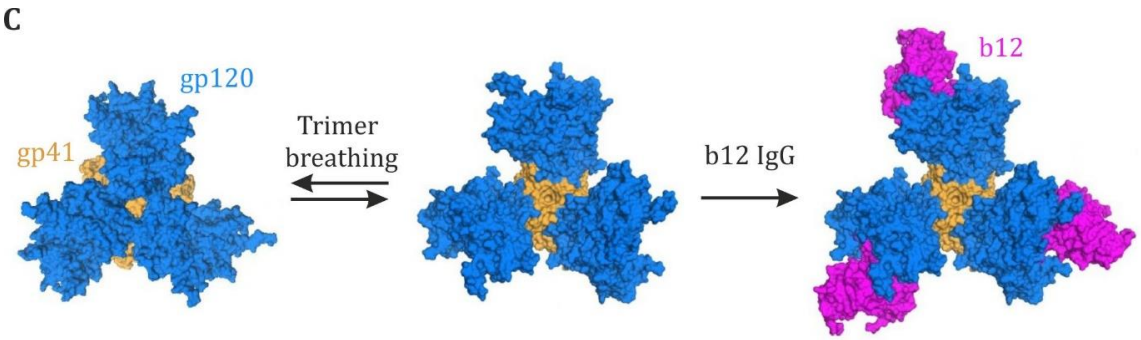


Figure 6 – Evasion mechanisms of HIV-1 Env. (A) N-linked glycosylation and **(B)** sequence variability of Env (left) in comparison with influenza virus H3 haemagglutinin (HA) (middle) and RSV fusion glycoprotein subtype A (right). Conservation of glycans is represented in light green (conserved: > 90% conservation) or dark green (variable: < 90% conservation). Likewise, sequence variability is depicted in light or dark purple (B). Figure was adapted from [122] with minor modifications with permission from Nature Publishing Group. **(C)** Conformational states of Env. The pre-fusion trimer is assumed to be present in various reversible conformations that fluctuate between open and closed states which is referred to as 'trimer breathing'. An antibody-bound state is also shown (right). The conformation remains in a more open state after binding the antibody (in this case b12). Figure was adapted from [123] with minor modifications with permission from Nature Publishing Group.

1.6 Humoral immune response to HIV infection

1.6.1 Ontogeny of the antibody response during HIV infection

Soon after HIV transmission, the B cell branch of the newly infected person's immune system becomes activated. The first antibody response to HIV-1 can be detected within the first week of infection in the form of immunoglobulin (Ig) IgM and IgG antibodies mainly targeting free-floating virions [124]. However, the initial antibody responses do not possess the ability to neutralize the virus. A few days later, circulating anti-gp41 antibodies are generated, followed by anti-gp120 antibodies that are primarily directed against the V3 loop [125]. Even though these antibodies have seemingly no effect on the infecting viral strain, as they are directed mostly against dissociated gp120 and gp41 subunits or aberrantly folded proteins [126], the antibodies seem to be able to convey Fc-mediated effector functions such as antibody-dependent cellular cytotoxicity (ADCC) or antibody-dependent cellular phagocytosis (ADCP) [127]. Several months post infection, the first strain-specific autologous neutralizing antibodies occur which exert selective pressure on the virus leading to the generation of escape mutants [124,128]. After several years of continuing co-evolution of escaping virus and the following adaptation of the humoral immune response, antibodies with increased neutralization breadth and potency can emerge in a small percentage of chronically infected individuals [129]. Some of these antibodies are able to neutralize more than 90% of circulating HIV strains [130]. Notably, the infected patients cannot benefit from the elicited bnAbs, as they acquired escape mutants from said antibodies.

1.6.2 Broadly neutralizing antibodies

Despite the multitude of viral defense mechanisms, approximately 10-30% of the chronically infected individuals develop cross-reactive antibodies that are capable to neutralize various heterologous virus strains [131,132] as a result of the co-evolution between HIV-1 escape variants and antibody affinity maturation. Furthermore, about 1% of the patients are described as 'elite neutralizers', pertaining to HIV-1-infected people with unusually potent cross-reactive neutralizing antibody response against a majority of HIV-1 subtypes [133]. The monoclonal antibodies are referred to as broadly neutralizing antibodies (bnAbs) and target specific key sites of vulnerability on the envelope (figure 7): i) the CD4-binding site [134,135] (e.g. bnAbs VRC01 [136], NIH45-46 [130]), ii) the glycopeptide epitopes of the variable region 1 and 2 [137] (e.g. PG9 and PG16 [136], PGT145 [138]), iii) the glycan-associated variable region 3 [139] (e.g. PGT121-134 [140], iv) the membrane proximal external region (MPER) on gp41 [141,142] (e.g. 4E10 [143], 10E8 [144]) and v) a gp120-gp41 spanning interface [145] (e.g. PGT151 [146], 35O22 [145]). In order to overcome the many viral defenses, bnAbs have acquired one or more unusual characteristics, such as extremely long or short heavy-chain complementarity-determining region 3 loops (HCDR3) [147,148], insertions and/or deletions [149] and polyreactivity [150]. In addition, many bnAbs undergo extensive somatic hypermutation (SHM) [151,152] to achieve neutralization breadth and potency. To accumulate such degree of mutation can require a long time and might explain the unusual duration until cross-neutralizing antibodies occur in HIV-1 infected individuals and why it proved to be challenging to elicit bnAbs so far.

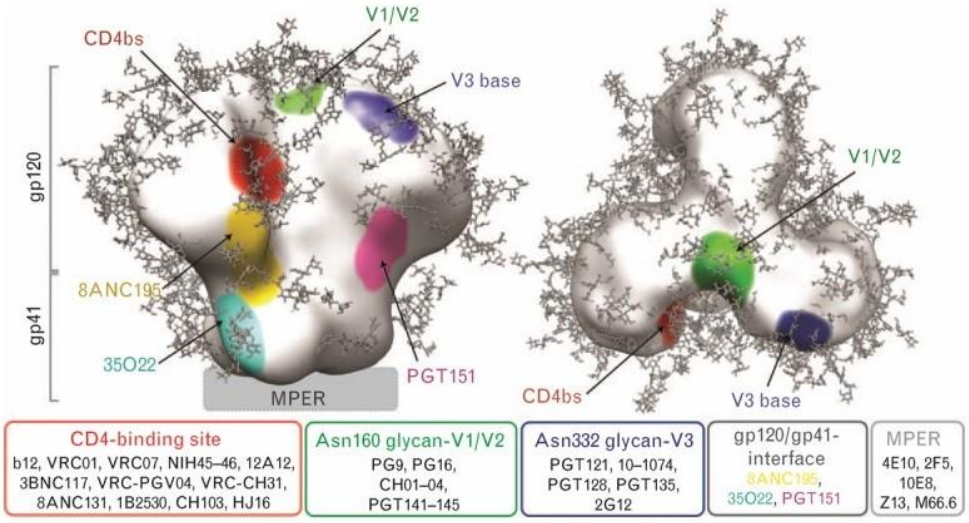


Figure 7 – Location of bnAb epitopes on the HIV-1 Env trimer. So far, five sites of vulnerability were discovered that include the CD4 binding site (CDbs), the trimer apex (V1V2), the glycan-dependent V3 region, the gp120/gp41 interface and the membrane-proximal external region (MPER). Figure was adapted from [153].

1.7 HIV-1 vaccine development

After more than 30 years since the discovery of the HIV-1 pandemic, an effective vaccine for clinical use to prevent infection still remains elusive. Notwithstanding the significant efforts that have been undertaken toward developing an HIV remedy, from over 218 trials only seven vaccines advanced to clinical phase III trials [154–159]. The first studies were performed in the late 1980s and early 1990s and involved the usage of recombinant gp120-based vaccines derived from the isolates MN and B (AIDSVAX B/B') in the VAX004 trial and from clades B/E in the VAX003 trial. The efficacy for both vaccines was estimated at 0.1% and therefore failed to demonstrate protection against HIV-1 infection [160].

Thus, research was redirected toward reduction of viral load setpoints or delay of disease progression by eliciting cytotoxic T-lymphocyte (CTL) responses [161]. Pursuing this strategy, the Step trial (HVTN502 in 2004) and Phambili trial (2007) were conducted by Merck and the HIV Vaccine Trials Network (HVTN), respectively. The vaccine candidate comprised a recombinant Adenovirus 5 (Ad5) vector expressing HIV-1 Gag, Pol and Nef [161]. Although the trials demonstrated $CD8^+$ T-cell responses, both studies were stopped early on the basis of interim data proving futility and due to increased HIV incidence among vaccine recipients [154,162,163].

In September 2009, the first promising results were reported for the RV144 or Thai trial that was performed by the U.S. Military Research Program in collaboration with several Thai institutions. The vaccination strategy comprised a combination of a recombinant canarypox vector vaccine prime (ALVAC-HIV, Sanofi Pasteur) and the bivalent gp120 protein boost (AIDSVAX B/E) previously utilized in the VAX003 trial. Although only a mediocre efficacy of 31.2% [164] was acknowledged, immune correlates could be identified suggesting that non-neutralizing antibodies directed against V1/V2 may have contributed to protection against HIV infection by elicitation of antibody-dependent cellular cytotoxicity (ADCC) [165].

The HVTN505 trial was designed to evaluate the regimen's effect on viral load. In this respect, a prime/boost vaccination approach was applied which consisted of a DNA prime with Clade B *gag/pol/nef* and multiclad*e env* followed by a boost with Ad5 vector from the Step study. However, the trial was stopped in 2013 due to futility [166].

The HVTN702 trial which started in November 2016 aims to provide greater and more sustained protection than the Thai trial. To achieve this an improved vaccination regimen was adapted to address HIV subtypes that predominate in southern Africa. In this respect, an ALVAC vector containing a clade C Env insert followed by bivalent clade C recombinant gp120 protein were applied in the vaccination regimen [167]. This trial is still ongoing.

1.8 Elicitation of broadly neutralizing antibodies

Results from the RV144 trial led to a general shift in the major focus of research toward an antibody-based HIV-1 vaccine. This concept was additionally substantiated by multiple *in vivo* challenge studies demonstrating that passive administration of bnAbs into humanized mice or non-human primates repeatedly conferred protection against HIV-1 infection [119,168–171]. Thus, there is currently a well consolidated expectation that vaccines able to elicit bnAbs upon immunization would prevent HIV-1 infection.

1.8.1 Engineering of envelope immunogens to induce cross-neutralizing antibody responses

Significant efforts have been undertaken to develop an Env immunogen able to induce broadly neutralizing antibodies. Early strategies were based on the successful Hepatitis B vaccine design and thus involved monomeric gp120 subunits as immunogens. Unfortunately, the antibody responses were relatively weak and were only able to neutralize a very limited spectrum of sensitive viral strains [172,173]. The many failures led to a change in the general mindset of immunogen design and more attention has been focused on candidates that simulate the natural Env structure. A strategy referred to as structure-based reverse vaccinology was devised that first determined the crystallographic structure of a complex between antibody and Env and then reconstructed the respective epitopes [173,174]. So far however, this approach produced only little tangible progress in the elicitation of broadly neutralizing antibodies (reviewed in [175–177]). It is assumed that instability of Env was the major obstacle in the induction of bnAbs [178], thus many efforts have been undertaken to prevent dissociation of the glycoprotein. Approaches encompassed the generation of furin cleavage-deficient Env proteins. This was achieved by mutating the primary recognition sites of the furin protease that ocurr at KAKRRWQR$_{508}$EKR$_{511}$AVGIGALFLGFLGAAG between residues 508 and 509 or between 511 and 512 [179,180]. Accordingly, substitution of the motif REKR with REKS or SEKS resulted in cleavage-incompetent Env proteins which prevented dissociation of the protomers [181–183]. Similarly, cleavage sites were replaced with a glycine-serine peptide linker resulting in native flexibly linked (NFL) envelopes [184]. Another strategy to increase stability of Env was successfully accomplished by introducing appropriately positioned cysteine residues in the ectodomains of Env which created an engineered disulfide bond (SOS) between the two subunits [185,186]. Furthermore, this strategy was combined with the trimer stabilizing mutation I559P [185] (IP), as well as with an improved 6xR furing cleavage site instead of REKR, leading to fully cleaved and well-folded trimers that were referred to as BG505 SOSIP.664 Envs [185].

Many new discoveries corroborated the focus of vaccine research on natural, trimeric immunogens. First and foremost, it was revealed that many bnAb epitopes are strongly (trimer apex, gp120-g41 interface [187]) or partially quaternary-dependent (CD4 binding site [188]). Additionally, several CD4bs-specific bnAbs require contact to a second protomer within the trimer [188,189], thus constraining the presentation of epitopes. Consequently, even though monomeric Env immunogens exhibit the same epitopes, they are not presented in the precise orientation of native spikes [190]. This could potentially complicate the elicitation of bnAbs.

Recent approaches focused on the simulation of the co-evolution between Env and the humoral immune response. For this purpose, directional immunogens were designed to activate germline receptors on B cells that gradually guided the immune system towards the development of mature bnAbs over several intermediate stages. Some success was accomplished in several murine model systems by following this strategy. In a first approach, eight mutations were introduced into an engineered envelope variant (eOD-GT8) with the goal to impart affinity for VRC01 germline antibodies [191]. Sequential immunization of transgenic mice with the germline-targeting constructs, as well as several intermediates and mature envelope (BG505 SOSIP N276D), finally resulted in the elicitation of VRC01-class neutralizing antibodies [192]. Furthermore, the successful development of mature PGT121 bnAbs in mice was demonstrated after following a similar sequential immunization strategy [193]. Ultimately, these approaches lead to the necessity of immunogens exhibiting high affinity towards germline, intermediate or mature antibodies, as well as appropriate methods to identify such improved Env variants.

1.8.2 Advantages of Env-based gene variant libraries

Despite continued advances in the field of protein structure and function, many aspects still cannot be predicted. Therefore, protein engineering according to combinatorial strategies is highly appealing. One fundamental component of such approaches encompasses the construction of protein libraries which usually comprises a nucleic acid library from which the protein library is then translated. This provides the benefit that any protein can be directly identified by DNA sequencing. A wide variety of methods have been devised to generate gene variant libraries. One approach focusses on introducing sequence variation at random positions by physical (*i.e.* UV radiation) and chemical mutagens (*i.e.* alkylating agents), mutator strains (*i.e.* XL1-Red strain) as well as error-prone PCR [194]. Recombination methods can rearrange already existing diverse sequences into novel combinations. These techniques include DNA shuffling [195,196] and the staggered extension process [197]. More controlled randomizations to specific positions can be achieved with direct synthesis of mixtures of DNA molecules and their

subsequent incorporation into genes via PCR or cloning, as in the case of Gibson Assembly [198].

Gene mutant libraries offer beneficial possibilities to study properties, biological functions and structure biology of many proteins simultaneously. As early as 1991, a 15-mer random peptide libraries has been successfully utilized to map epitopes recognized by the antibodies targeting the V3 loop of Env [199,200]. Similarly, critical residues in the epitope of the antibody 4E10 were revealed by screening of a 12-mer peptide library [201]. Another study demonstrated the identification of engineered soluble CD4-inhibitors (sCD4) by combining structure-based design with sequential panning of a large mutant library against different HIV-1 envelopes [201]. The detected CD4-mutants displayed great increase in affinity to gp120, neutralization of pseudoviruses and exhibited significant inhibitory activities in clinical trials. Furthermore, Jardine *et al* (2016) were able to demonstrate the detection of specific Env variants with improved binding affinity for germline antibodies by screening of large mutant libraries [191]. Overall, the positive results from many previous studies consolidated that gene mutant libraries could facilitate the search for Env variants with specific and favorable properties. Thus, a multivariant genomic Env library provided the foundation of this PhD thesis.

2 Objective

The elicitation of broadly neutralizing antibodies still constitutes a major challenge. However, the long-time search for immunogens able to induce bnAbs finally demonstrated the first successes. Recent studies suggested that the complex co-evolution between Env and the immune system can be mimicked by sequential immunization with germline-targeting, intermediate and finally mature variants of Env [192,193]. Such positive results emphasized the demand to develop methods for effective engineering, display and screening of envelope proteins with specific properties, such as improved binding affinities for bnAbs.

The focal point of this PhD thesis was the bioengineering and quality control of a sequential permuation Env library, as well as the implementation of methods that facilitate the work with complex and large libraries. The secondary objective focused on identifying and characterizing trimeric Env immunogens with improved antibody recognition from the library by using a mammalian cell display and cell sorting-based panning approach.

3 Materials and Methods

3.1 Molecular Biology

Unless stated otherwise, all methods were performed in agreement with common protocols of molecular biology [202] or of the respective manufacturers. A detailed list of all oligonucleotides and DNA constructs utilized in this work can be found in section 8.2. Vector backbones for cloning purposes were generated by digestion with the appropriate restriction endonucleases and treatment with CIP to prevent re-ligation. Subsequently, the desired DNA-fragment was isolated from a 0.8-1.0% agarose gel.

Inserts were created according to one of the following methods: 1) Amplification by PCR using eligible oligonucleotides and digestion with suitable restriction enzymes. PCR for analytical purposes was performed with GoTaq Green Master mix (Promega), whereas for preparative applications, Phusion DNA polymerase (NEB) was used. 2) Annealing of complementary oligonucleotides followed by phosphorylation of the 5‘-ends with T4-PNK. 3) Direct recovery from plasmids by restriction digestion and subsequent gel extraction.

Vector DNA was then mixed with a 3-fold molar excess of insert DNA and ligated using the Quick Ligation Kit (NEB). Subsequently, the ligation mixture was used for the transformation of chemically competent *E.coli* DH5α or DH10B according to standard protocols [203].

After cultivation of bacterial cultures in LB or TB medium containing the appropriate antibiotic, plasmid DNA was isolated by alkaline lysis [204] or by usage of the GeneJET Plasmid Miniprep Kit/Plasmid Midi/Maxi Plus Kit. The concentration and purity of the purified DNA was determined spectrophotometrically by measuring absorbance at 260 nm and 280 nm or Agilent 2100 Bioanalyzer (Agilent) using the High Sensitivity DNA Kit (Agilent).

In order to confirm the correctness of the construct, restriction digestion and Sanger sequencing (Seqlab) were performed.

Oligonucleotides	Biomers, Eurofins
Restriction endonucleases	New England Biolabs, Fermentas, Thermo Fisher Scientific
Alkaline Phosphatase, Calf Intenstinal (CIP)	New England Biolabs, M0290L
QIAquick Gel Extraction Kit	Qiagen, 28706
GoTaq® Green Master Mix	Promega, M7123
Phusion ® High-Fidelity DNA Polymerase	New England Biolabs, M0530 L
QuickLigation Kit	New England Biolabs, M2200
T4-PNK	New England Biolabs, M0201
GeneJET Plasmid Miniprep Kit	Thermo Fisher Scientific, K0502

Plasmid Midi/Maxi Plus Kit	Qiagen, 12945, 12965
High Sensitivity DNA Kit	Agilent, 5067-4626

E.coli DH5α	*F- supE44 dlacU169 (fi 80 lacZdM15) hsdR1recA1 endA1 gyrA96 thi-1 relA1*
E.coli DH10B	F– *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) Φ80*lac*ZΔM15 Δ*lac*X74 *rec*A1 *end*A1 *ara*D139 Δ(*ara leu*) 7697 *gal*U *gal*K *rps*L *nup*G λ–
Lysogeny broth (LB medium)	1% Bacto tryptone; 0.5% Bacto yeast extract; 1% NaCl; pH 7.5
Terrific broth medium (TB medium)	1.2% Bacto tryptone; 2.4% Bacto yeast extract; 0.5 % glycerol; 0.17 M KH_2PO_4; 0.72 M K_2HPO_4

3.2 Next Generation Sequencing

3.2.1 Illumina Sequencing by Synthesis Technology

Next Generation Sequencing (NGS) represents a variety of sequencing methods which transcends the capacity of traditional DNA sequencing technologies in respect to cost, speed and data output, thus enabling an in-depth study of biological systems by rapid sequencing of whole genomes. One of the most prevalent NGS technologies was developed by Illumina and is referred to as 'Sequencing-by-synthesis' (SBS) [205]. This method supports massively parallel sequencing proving to be especially beneficial for questions that demand extensive information regarding highly diverse genomic libraries. The sequencing workflow is composed of four basic steps: i) sample preparation, ii) cluster generation, iii) sequencing and iv) data analysis.

In general, NGS sample preparation includes fragmentation of DNA sequences into suitable sizes (~ 300 bps) due to the limitation of reading length of the NGS device (figure 9, (1)) and the annealing of adapters (2) by PCR. The resulting product consists of a sequence of interest flanked 5' and 3' by the adapters P5 and P7 (figure 8), which allow attachment to the surface of a flow cell coated with a lawn composed of complementary adapter oligonucleotides (3). Additional motifs are also introduced during sample preparation, such as the NGS sequencing primer binding site and indices. The indices or barcodes allow distinction among a multitude of samples.

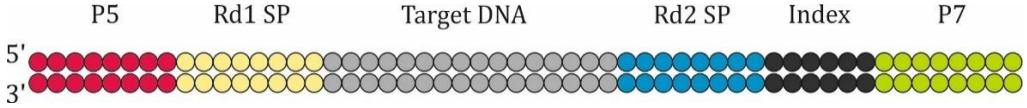


Figure 8 – Schematic illustration of a ready-to-load index library eligible for NGS. The DNA fragment of interest is shown in grey. P5 (red) and P7 (green) indicate the adapters. Rd1 SP (yellow) and Rd2 SP (blue) designate the binding sites for NGS-specific sequencing primers. Barcodes or indices are represented in black.

After binding of samples to the flowcell, the DNA strand folds over and the adaptor region hybridizes to the second type of oligo on the flow cell (4). Polymerases create a complementary strand forming a double stranded bridge (5) which is subsequently denatured, resulting in two single stranded copies of the molecule (6). The process is repeated over and over so that millions of DNA 'clusters' are generated (7). Cluster densities have large impact on sequencing performance in terms of data quality and output.

Sequencing begins with the extension of the first sequencing primer to produce the first read (8). With each cycle, four fluorescently tagged reversible terminator bases compete for incorporation into the growing chain. Only one nucleotide is integrated, based on the sequence of the template while non-incorporated nucleotides are washed away. Subsequently, clusters are excited by a light source and a characteristic fluorescent signal is emitted which is acquired as image by a camera (9). Emission wave length along with the signal intensity, determine the base call. The length of the read is determined by the number of cycles. The entire process generates millions of reads representing all fragments which then can be separated based on the unique indices introduced during the sample preparation. Forward and reverse reads are paired creating contiguous sequences that are aligned back to the reference genome for variant identification. Sequencing coverage describes the average number of reads that align to known reference bases.

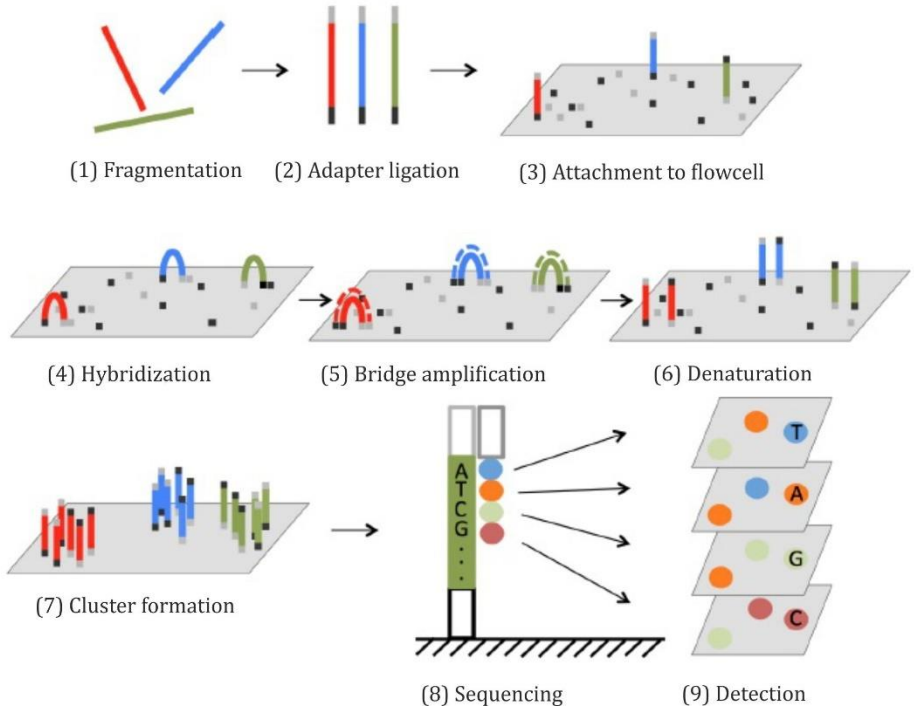


Figure 9 – Outline of Illumina NGS technology. After fragmentation of the samples (1), adapters are annealed to the ends of the sequence (2). Fragments attach to the flowcell (3) by hybridization to oligos complementary to the adapters (4). Subsequently, bridge amplification occurs to produce clusters of fragments (5-7). During each sequencing cycle, one fluorophore-attached nucleotide is added to the

growing strands (8). The fluorophores are then excited by a laser and signals from each fragment cluster are detected and recorded as images (9). Illustration was adapted from [206].

3.2.2 Sequencing library preparation

The sequential permutation (SeqPer) library was the foundation for all steps involving Next Generation Sequencing. A key step in the NGS library sample preparation is generating the input for sequencing. In general, the library preparation was composed of four stages: i) DNA amplification, ii) attachment of oligonucleotide indices and adapters to the ends of the amplified fragments, iii) purification and iv) final library quantification and quality control.

3.2.2.1 Generation of amplicon libraries

In the course of this PhD thesis, polymerase chain reaction was utilized to produce amplicons in the size of approximately 300 bps. These amplicon libraries were generated by two sequential PCRs using specific primers which carry the attachments required for NGS. In the first PCR, NGS-specific primer binding sites were introduced into the target DNA template. Fusion of indices and adapters was achieved by the second PCR (figure 8). The input DNA for the first PCR was derived from the plasmid DNA of the SeqPer library purchased from GeneArt or was isolated from the respective stable cell lines as genomic DNA (see 3.2.2.2).

For the amplification 20 ng plasmid DNA was applied in the first PCR. As the concentration of isolated genomic DNA was always below the detection limit, 10 µL of the extract was utilized for PCR1. Accordingly, 10 µL from the first amplification was applied in the second PCR. The respective primers for PCR1 and PCR2 can be obtained from section 8.2.1. Notably, only the reverse primers for PCR2 varied, whereas the forward primer (ILLUMINAseq_fwd) remained the same. In general, a 50 µL reaction was prepared containing 1 µL dNTPs (10 mM), 2.5 µL forward and reverse primer (10 µM), respectively, 1.5 µL DMSO, 0.5 µL Phusion DNA polymerase, 10 µL 5xHF Phusion buffer, template DNA as mentioned above and nuclease-free water (see table 3). To reduce accumulation of errors from the polymerase, 22 cycles were applied for both amplification steps. The thermocycling conditions can be obtained from table 4. PCR samples were stored temporarily (usually over night) at 4°C if it was not possible to proceed with the experiment.

Table 3 – Composition of PCR 1 and 2 for generation of amplicon libraries.

Component	PCR 1	PCR 2
Nuclease-free water	to 50 µL	to 50 µL
10 mM dNTPs	1 µL	1 µL
10 µM forward primer	2,5 µL	2,5 µL
10 µM reverse primer	2,5 µL	2,5 µL
DMSO	1,5 µL	1,5 µL
Phusion DNA polymerase	0,5 µL	0,5 µL
Template DNA	1 ng pDNA/ 10µL gDNA extract	10 µL from PCR 1
5x Phusion HF buffer	10 µL	10 µL
total	50 µL	50 µL

Table 4 – Thermocycling conditions for PCR 1 and 2.

PCR 1		
98°C	1 min	
98°C	10 sec	22 cycles
68°C	30 sec	
72°C	6 sec	
72°C	5 min	
8°C	∞	

PCR 2		
98°C	1 min	
98°C	10 sec	22 cycles
64°C	30 sec	
72°C	10 sec	
72°C	5 min	
8°C	∞	

Deoxynucleotide (dNTP) Solution Mix	New England Biolabs, N0447L
Dimethyl Sulfoxide (DMSO)	SigmaAldrich, D8418-100ML
Phusion ® High-Fidelity DNA Polymerase	New England Biolabs, M0530 L

3.2.2.2 Extraction of genomic DNA

Extraction of genomic DNA was performed by using the QIAamp DNA Mini Kit according to the protocol 'DNA purification from blood and body fluids'. Unless stated otherwise, $2x10^5$ cells were utilized for gDNA isolation. If less than 10.000 genomic equivalents were present, 10 µg/mL carrier DNA (polyadenylic acid, poly dA) was added to the sample in order to enhance the recovery of DNA and to prevent the small amount of target nucleic acid from being irretrievably bound.

As a first step, 20 µL QIAGEN protease was pipetted into the bottom of a 1.5 mL tube. Subsequently, a 200 µL solution containing $2x10^5$ cells in PBS together with 200 µL AL buffer were added to the tube and mixed by pulse-vortexing for 15 sec. The mixture was then incubated at 56°C for 10 minutes and briefly centrifuged to remove drops from the inside of the lid. 200 µL ethanol (96-100%) were added to the sample, mixed by pulse-vortexing for 15 sec and briefly centrifuged. In the following, the solution was applied to the QIAamp Mini spin column in a 2 mL collection tube and centrifuged at 8.000 rpm for 1 min. After discarding the filtrate, 500 µL AW1 buffer was added to the column and centrifuged for 1 min at 8.000 rpm. A second washing step with 500 µL AW2 buffer was performed at 13.000 rpm for 3 min. To eliminate the chance of possible AW2 buffer carryover, the filtrate was discarded, the spin column was placed in a new collection tube and centrifuged at full speed for 1 min. Finally, gDNA was eluted into a new 1.5 mL tube by adding 30 µL water to the column, incubating at room temperature for 1 min and centrifuging at 8.000 rpm for 1 min. Aliquots of 10 µL extracted DNA were generated and stored at -20°C to reduce damaging of DNA by repeated thawing cycles.

Poly(A), Polyadenylic acid	SigmaAldrich, 10108626001
QIAamp DNA Mini Kit	Qiagen, 51304

3.2.2.3 Generation of stable cell line samples for NGS

The stable cell line library (see 4.2) was analyzed at three stages: i) stock directly after thawing, ii) before induction with doxycycline and iii) after induction to assess potential changes in the quality and amino acid composition. As a first step, cells were thawed (see 3.3.4.2) and resuspended in 1 mL preheated $DMEM_{SCL}$. After determining the cell count, $2x10^5$ cells were withdrawn for gDNA extraction (see 3.2.2.2) and subsequent NGS sample preparation. This sample was referred to as ‘stock’. The remaining cells were transferred to a T75 flask, expanded and split after a confluency of 80% was reached (see 3.3.1). After the third passage (figure 10), $2x10^5$ cells were withdrawn for gDNA extraction and to generate the non-induced sample, whereas the remaining cells were placed back into the T75 flask and induced with 1 µg/mL doxycycline. On the following day, medium was replaced with 15 mL fresh $DMEM_{SCL}$ and the cells were detached from the flask by resuspension. $2x10^5$ cells were then withdrawn for gDNA extraction and the subsequent generation of the ‘induced’ sample for NGS.

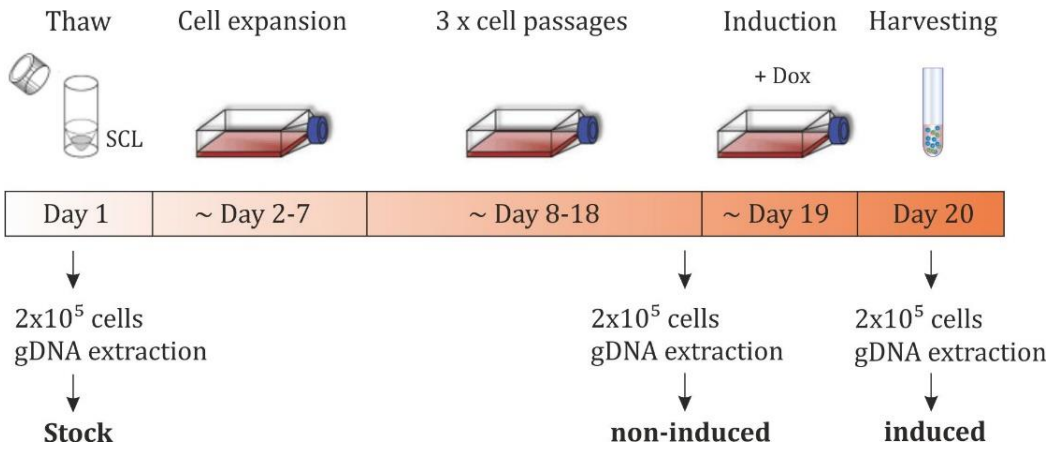


Figure 10 – Workflow for the primary sample preparation of stable cell lines for NGS analysis. After thawing, stable cell lines (SCLs) were expanded and passaged three times before they were induced with doxycycline (Dox). On the next day, cells were harvested and further analyzed. To generate stock, non-induced and induced samples for NGS, $2x10^5$ cells were withdrawn after the thawing process, as well as prior to and post induction with doxycycline, respectively. Subsequently, genomic DNA was extracted from the cells.

3.2.3 Purification of amplicon libraries

Successful Next Generation Sequencing is highly dependent on the quality of the samples used. Therefore, removal of any remaining primer dimer and adapter dimer was essential to prevent their binding to the NGS flow cells.

3.2.3.1 Agarose gel electrophoresis

DNA of amplicon libraries was purified from an agarose gel after PCR1 and PCR2, respectively, to remove primer dimer and adapter dimer. Due to the size of the amplicons and to improve fragment separation (~279 bps from PCR1 and 399 bps from PCR2) a 2% agarose gel was prepared which was run at 160 V for 2 hours. Subsequently, the appropriate band was excised and the DNA was isolated using the QIAquick Gel Extraction Kit according to the manufacturer's instructions.

3.2.3.2 Magnetic beads purification

Amplicons were also purified using the Agencourt AMPure XP magnetic beads (Beckman Coulter) purification protocol after PCR1 and PCR2, respectively. Sample preparation was performed in DNA LoBind tubes (Eppendorf) to maximize DNA yield. Briefly, 1,8x sample volume (typically 30 µL of DNA eluted from gel extraction columns in water) of AMPure magnetic beads was supplemented to the amplified DNA and mixed thoroughly by pipetting. After incubation for five minutes the samples were placed into a magnetic rack until the solution was clear. Subsequently, two washing steps were performed. This was achieved by adding 500 µL of freshly prepared 70% ethanol and incubating for 30 seconds while turning the cups in the magnetic rack two times. After aspirating the ethanol without disturbing the pellet, tubes were centrifuged and the remaining supernatant was removed using a 20 µL pipette. The pellets were air-dried for three minutes on the magnetic rack to remove residual ethanol. Subsequently, DNA was eluted by adding 30 µL nuclease free water and vortexing for 15 seconds. After one minute on the magnetic rack, the supernatant was transferred into DNA LoBind tubes.

Agencourt AMPure XP	Beckman Coulter, A63881
1.5 mL LoBind microcentrifuge tubes	Sigma Aldrich, Z666548

3.2.4 Quantitation of amplicon libraries

3.2.4.1 Quantitation with the Agilent 2100 Bioanalyzer

In addition to the high quality of amplicons, success of NGS is also dependent on an accurate quantitation. Therefore, the concentration and purity of the DNA were determined by the Agilent 2100 Bioanalyzer. Based on concentrations obtained spectrophotometrically by measuring absorbance at 260 nm and 280 nm, all samples were diluted to a final concentration of 500 pg/µL and loaded on a DNA chip according to the instructions provided by the Agilent High Sensitivity DNA Kit as follows:

Prior to analysis, the gel-dye mix was prepared. For this purpose, all reagents were first equilibrated to room temperature for 30 minutes. The High Sensivity DNA dye concentrate was vortexed for 10 seconds, briefly centrifuged and added (15 µL) to the High Sensitivity DNA gel matrix vial. The tube was then vortexed for 10 seconds and the complete gel-dye mix was transferred to the receptacle of a spin filter to be centrifuged for 10 minutes at 6.000 rpm at room temperature.

Before loading of the gel-dye mix, a High Sensivity DNA chip was placed on the chip priming station. In the following, 9 µL gel-dye mix were pipetted to the bottom of the

well marked 'G', the top of the station was closed and the gel was dispensed by pressing down the plunger of the syringe from the priming station until it was held by the clip. After 60 seconds, the plunger was released and slowly pulled back to the 1 mL position of the syringe. Subsequently, reagents were pipetted into the wells as described in the following: i) 9 µL of gel-dye mix in each of the remaining wells marked 'G', ii) 5 µL of High Sensitivity DNA marker into all remaining wells, iii) 1 µL DNA ladder into the well marked with the ladder symbol, and iv) 1 µL of sample into the sample wells. The chip was placed horizontally in the adapter of a IKA vortex mixer and vortexed for 60 seconds at 2.400 rpm. Finally, the chip was inserted into the Bioanalyzer and the dsDNA assay was selected from the file menu.

Agilent High Sensitivity DNA Kit	Agilent Technologies, 5067-4626

3.2.4.2 Generation of library pools and quantification by quantitative PCR

Quantification of samples by bioanalyzer and generation of library pools were performed on the same day to prevent deviation in the concentrations due to freezing and thawing processes. In general, a 20 nM DNA-pool was generated. For this purpose, 12 µL of a 20 nM solution (resuspended in water) for each amplicon library was prepared, of which 10 µL were combined, respectively, to create the library pool used for NGS.

Library pools were quantified by qPCR and bioanalyzer. The latter provided additional important information such as average sequence length of the amplicons and potential contamination with primer dimer. Furthermore, the analysis served as confirmation for results obtained from the qPCR.

As adviced by Illumina, the KAPA Library Quantification Kit (KAPA Biosystems) was utilized for accurate quantification. In brief, a 10-fold dilution series of the library pool in 10 mM Tris/HCl, pH 8.0 was prepared in 1.5 mL LoBind tubes. In this respect, six data points were collected in triplicates comprising a $0.1x10^3$-, $1x10^3$-, $1x10^4$-, $1x10^5$-, $1x10^6$-, and $1x10^7$-fold dilutions. Per 20 µL reaction, 12 µL of KAPA SYBR FAST master mix containing the forward and reverse primers (provided by the manufacturer), 4 µL of each library dilution or DNA standard and 4 µL PCR-grade water were combined in a MicroAmp™ Optical 96-well reaction plate. The DNA standards were provided with the kit and comprised six concentrations in a ten-fold dilution series (ranging from 20 pM to 0.0002 pM). The plate was then sealed with MicroAmp™ Optical 8-cap strips, briefly vortexed and spinned down. Quantification was performed using a standard curve analysis protocol provided by the StepOnePlus device (Applied Biosystems). Cycler conditions can be seen below in table 5. All quantifications were performed on the day of NGS analysis.

Table 5 – Cycler conditions for the quantitative PCR.

Step	Temperature	Duration	Cycles
Initial denaturation	95°C	5 min	1
Denaturation	95°C	30 sec	35
Annealing/Extension	60°C	45 sec	

KAPA Library Quantification Kit	KAPA Biosystems, KK4824
MicroAmp Optical 96-well reaction plate	ThermoFisher Scientific, 4316813
MicroAmp Optical 8-Cap Strips	ThermoFisher Scientific, 4324032

3.2.5 Denaturation and dilution of NGS libraries

Prior to loading of prepared samples on the NGS flow cell, library pools were denatured and diluted. For this purpose, a mixture comprising the library pool and PhiX Control V3 with a final concentration of 2 nM in 10 µL water was created. The composition was calculated by the following formulas (1-3):

$$Volume(Library) = \frac{Cfinal}{Cstart} x\ final\ volume\ \mu L\ x \frac{percentage\ library}{percentage\ total} \quad (1)$$

Cfinal – final concentration of the library for denaturation, 2 nM; Cstart – starting concentration of the quantified library in nM; final volume in µL – 10 µL, percentage library – fraction of library present in the mixture including the PhiX fraction; percentage total – 100%

$$Volume(PhiX) = \frac{Cfinal}{Cstart} x\ final\ volume\ \mu L\ x \frac{percentage\ PhiX}{percentage\ total} \quad (2)$$

Cfinal – final concentration of the PhiX control for denaturation, 2 nM; Cstart – starting concentration of PhiX, 10 nM; final volume in µL – 10 µL; percentage PhiX – fraction of PhiX present in the mixture including the library fraction; percentage total – 100%

$$Volume(H2O) = 10\ \mu L - Volume(Library) - Volume(PhiX) \quad (3)$$

On the example of a 20 nM library and a spike-in of 20% PhiX this results in a library volume of 0.8 µL, a PhiX volume of 0.4 µL and water volume of 8.8 µL. The PhiX control was added due to the low diversity of the utilized library. Generally, a 20 to 30 % PhiX spike-in was used for the NGS experiments.

For denaturation of the samples, 10 µL of freshly prepared 0.2 N NaOH solution diluted in water was prepared and added to the sample. After incubation for five minutes at room temperature, the denatured sample was diluted with prechilled HT1 (provided in the kit) to a final concentration of 8-10 pM in a volume of 1 mL and loaded on the flow cell provided with the MiSeq Reagent Kit v2. The chosen loading concentration depended on the cluster density from preceding NGS runs. Typically, a run yielded between 300,000 and 800,000 clusters/mm^2.

Before sequencing, the sample sheet was created using the following parameter settings in the MiSeq control software: i) FastQ only, and ii) sample prep kit: true seq small RNA After designating the sample names and the respective indices used during sample preparation, cycles were manually adjusted in the sample sheet to 2x151 reads to obtain sequencing from both the 5' and 3' end. NGS was conducted on the MiSeq device (Illumina) at the laboratory of Prof. Dr. Gunter Meister.

MiSeq Reagent Kit v2 (300 cycles)	Illumina, MS-102-2002
PhiX Control v3	Illumina, FC-110-3001

3.2.6 Analysis of NGS data

Analysis of NGS results was performed with the CLC Main Workbench 7. Briefly, after importing the FastQ files into CLC, sequences containing the corresponding forward and reverse reads were paired ('Paired reads' function during importing). Overlapping pairs were merged into one file ('Merge overlapping pairs' function) and subsequently aligned with a reference sequence ('Assemble sequence to reference' function). Finally, results were exported as SAM files and further investigated with an analyzer tool that was programmed for this specific purpose by Dr. Benedikt Asbach. The programme was designed to count every codon at each position and deliver a table with the absolute numbers of amino acids.

3.3 Cell Biology

3.3.1 Cultivation of cell lines

Eukaryotic cell lines were cultivated at 37°C and 5% CO_2 according to common protocols. In general, adherent cells were split in ratio of 1:10 upon reaching a confluency of about 80% by washing with PBS, detaching with Trypsin/EDTA solution

and resuspending of the cells in $DMEM_{FPS}$. The cell count was determined by Trypan Blue staining of the cells. In this respect, the cell sample was diluted in a 0.4% Trypan Blue solution in a 1:1 ratio (usually 50 µL of each). As the cell membrane of dying cells is compromised, Trypan Blue can enter the cells and stain them blue, whereas viable cells remain unstained. After filling a hemocytometer with the mixture, cells were counted under the microscope in four 1x1 mm squares of the hemocytometer chamber.

Expi293F suspension cells were mainly utilized for the expression of antibodies. Cultivation was carried out in 1 L Erlenmeyer flasks at 37°C and 8% CO_2 at up 125 rpm to a maximal density of $2.5x10^6$ cells per mL in Expi medium. The cells were passaged every 3-4 days by harvesting the cells via centrifugation, discarding the supernatant and resuspension in the required volume of fresh medium.

Flp-In™ T-Rex 293 cells with the stable integration of envelope variants in their genome received the same treatment as the adherent cells described above. For their cultivation, $DMEM_{SCL}$ medium was used consisting of DMEM with 5 % FCS, 1% Pen/Strep, hygromycin (100 µg/ml) and blasticidin (15 µg/ml).

HEK293T cells	Ad5-transformed immortalized human kidney fibroblast cell line, stably transfected to express the SV40 large T antigen, ATCC® CRL-11268
Expi293F™ cells	Ad5-transformed immortalized human kidney fibroblast cell line, variant of the 293 cell line, Thermo Fisher Scientific A14527
Flp-In™ T-Rex 293 cells	Ad5-transformed immortalized human kidney cell line containing pFRT/lacZeo and pcDNA™6/TR stably integrated, Thermo Fisher Scientific, R78007
Tetracycline free FCS	Fetal Calf Serum, Biochrom, S0115
Pen/Strep	10000 U/ml penicillin, 10 mg/ml streptomycin, PAN Biotech, P06-07100
Trypsin/EDTA	0.05% trypsin and 0.02% EDTA in PBS, Pan Biotech, P10-023500
DMEM	Dulbecco's Modified Eagle Medium, Thermo Fisher Scientific, 11995-065
$DMEM_{FPS}$	DMEM + 10% FCS, 1% Pen/Strep
$DMEM_{SCL}$	DMEM + 5% FCS, 1% Pen/Strep, 100 µg/mL Hygromycin, 15 µg/mL Blasticidin
Hygromycin B Gold	InvivoGen, ant-hg-5
Blasticidin	InvivoGen, ant-bl5

3.3.2 Transient transfection of mammalian cells

3.3.2.1 Cationic-polymer-mediated transfection

The transient introduction of DNA into HEK293T or Flp-In™ T-Rex 293 cells was conducted mainly by using a standard polyethylenimine (PEI) protocol [207]. For a 6-well transfection, $5x10^5$ cells were seeded in 2.5 mL $DMEM_{FPS}$ into 6-well plates the day

before transfection. On the following day, the cultivation medium was replaced with 1 mL serum-free medium (DMEM without supplements). A total of 2 µg plasmid DNA was diluted in 100 µL DMEM containing 10 µL polyethylenimine (PEI, 1 µg/µL) solution, mixed thoroughly by vortexing and incubated for 10 minutes at room temperature. The DNA-PEI-mixture was then added to the cells. After 6 hours, the medium was replaced with 2 mL $DMEM_{FPS}$. 24 hours post transfection the cells were harvested and analyzed.

For transfection in a 96-well format, $4x10^5$ cells per well were seeded in 200 µL $DMEM_{FPS}$ into a 96-well flat bottom plate. On the following day, medium was exchanged with 30 µL DMEM without supplements. A transfection mixture containing 200 ng DNA, 0.8 µL PEI and 30 µL DMEM with no additives was incubated for ten minutes and subsequently pipetted on the cells. Six hours post transfection, the medium was replaced with $DMEM_{FPS}$.

3.3.2.2 Determination of optimal ratios between DNA and various transfection reagents

Apart from cationic polymer-mediated introduction of DNA by PEI or PEI Max, other transfection methods and reagents were also utilized to test for improved transfection efficiency. For this purpose, the optimal ratio between DNA and transfection reagent needed to be determined before the actual experiments could be performed. In general, DNA to transfection reagent ratios of 1:2, 1:3, 1:4 and 1:5 (µg DNA:µL reagent) were tested while the amount of DNA was kept constant (2 µg per well). A negative control containing only DNA was also prepared ("1:0" ratio). All transfections were carried out following common protocols or the manufacturers' instructions, respectively.

Lipofection was conducted with Lipofectamine 2000 or Lipofectamine 3000. For the lipofection with Lipofectamine 2000, $5x10^5$ adherent cells were seeded in a 6-well flat bottom plate approximately 18-24 hours prior transfection. The next day, five different amounts of Lipofectamine (0, 6, 9, 12, 15 µL) and 2.5 µg DNA were diluted separately in 150 µL Opti-MEM™, respectively. 150 µL of diluted DNA was then resuspended in the Lipofectamine mixture, incubated for five minutes at room temperature and added to the cells. A change of medium after transfection was not required according to the manufacturers' instructions. 48 hours post tranfection, cells were harvested by centrifugation (100xg, 4°C, 5 min), discarding the supernatant and resuspension in 1 mL FACS buffer for analysis of the transfection efficiency by flow cytometry.

The day before lipofection with Lipofectamine 3000, $5x10^5$ cells were seeded in a 6-well flat bottom plate. Five different amounts of Lipofectamine 3000 (0, 5, 7.5, 10, 12.5 µL) and 2.5 µg DNA were diluted separately in 125 µL Opti-MEM™, respectively. Diluted DNA was added to the Lipofectamine mixture and then supplemented with 5 µL P3000™

Reagent (2 µL/µg DNA). After incubating for 15 minutes, 200 µL of the -solution was pipetted on the cells. There was no medium change required according ot the manufacturer. Harvest and analysis was conducted 48 hours post-tranfection.

Additionally, Fugene6-mediated transfection was performed. For this purpose, $5x10^5$ cells were seeded in a 6-well flat bottom the day before transfection. 18 to 24 hours later, medium was replaced with 2.7 mL DMEM without supplements. Five different amounts of Fugene6 reagent (0, 4, 6, 8, 10 µL) and 2 µg DNA were separately diluted in 150 µL DMEM, respectively, and incubated for five minutes at room temperature. Subsequently the DNA solution was added to the transfection mixture, incubated for 15 minutes and pipetted on the cells. Medium was changed 6 hours after transfection.

In comparison to all previously mentioned transfection protocols, all reagents involved in the calcium phosphate-mediated transfection were kept constant while the amount of DNA (0, 2, 3, 4, 5 µg) was altered. 24 hours prior to transfection, $5x10^5$ cells were seeded in a 6-well flat bottom plate. The following day, medium was replaced with DMEM with no supplements one hour before DNA dilutions were added. For this purpose, 5 µg DNA was resuspended in 135 µL H_2O and 15 µL 2.5 M $CaCl_2$. The mixture was then slowly dribbled into a 5 mL corning polystyrene round bottom tube containing 150 µL HeBS while being vortexed simultaneously. After an incubation period of 15 minutes, the DNA solution was gradually added to the cells. Six hours post transfection, medium was replaced with $DMEM_{FPS}$.

Polyethylenimine (PEI)	Polysciences, 23966-2
Polyethylenimine MAX (PEI MAX)	Polysciences, 24765-1
Lipofectamine™ 2000 Transfection Reagent	Thermo Fisher Scientific, 11668030
Lipofecatmine™ 3000 Transfection Reagent	Thermo Fisher Scientific, L3000001
Fugene6 Transfection Reagent	Promega, E2693
2xHeBS	0.28 M NaCl, 1.5 mM $NaHPO_4$, 50 mM Hepes, pH 7.1

3.3.3 Expression of antibodies

In general, the expression of antibodies was done by co-transfecting two plasmids carrying the variable heavy (CMVR VRC01 H) and light chain (CMVR VRC01 L) in a 1:1 ratio into Expi293F™ suspension cells. The transfection was performed by using the Exp Fectamine™ 293 Transfection Kit according to the manufacturers' instructions. Before transfection, two million cells per mL were seeded in 300 mL Expi293 expression medium and cultivated in a 1 L Corning flask for six to eight hours at 125 rpm until a density of 2.5 million cells/mL was reached. 300 µg plasmid DNA (150 µg of CMVR VRC01 H and 150 µg CMVR VRC01 L) and 802.5 µL ExpiFectamine were diluted separately in 15 mL OptiMEM each, and incubated for five minutes at room temperature.

The two solutions were briefly vortexed and incubated for 20-30 minutes to allow the DNA-ExpiFectamine complexes to form, followed by the supplementation of the mixture to the cells. After cultivation for 16-18 hours at 37°C, 8% CO_2 and 125 rpm, 1.5 mL Enhancer I and 15 mL Enhancer II (both enhancers were provided with the Kit) were added. The duration of antibody expression lasted usually five days.

CMVR VRC01 H	NIH AIDS Reagent Program, 12035
CMVR VRC01 L	NIH AIDS Reagent Program, 12036
Expi293F™ cells	Ad5-transformed immortalized human kidney fibroblast cell line, variant from the 293 cell line, Thermo Fisher Scientific A14527
Expi Fectamine ™ 293 Transfection Kit	Thermo Fisher Scientific, A14524
Expi293™ Expression Medium	Thermo Fisher Scientific, A1435102

3.3.4 Generation of stable cell lines

The SeqPer library (see 4.1), previously cloned into the plasmid pQL13, was utilized as foundation for the generation of the respective inducible stable cell line library (SCL library) (see 4.2). This was achieved by targeted transfection of the library into Flp-In™ T-Rex 293 cells. Briefly, 24 hours prior transfection, a density of $5x10^5$ Flp-In™ T-Rex 293 cells resuspended in 2.5 mL cultivation medium ($DMEM_{FlpIn}$) were seeded into 6-well plates. On the following day, medium was replaced with 1 mL DMEM without supplements. Subsequently, 0.4 µg of the helper plasmid pOG44 carrying a Flp recombinase (Thermo Fisher Scientific) was mixed with 1.6 µg of pQL13-SeqPer DNA and 10 µL PEI (1 µg/µl) in 2.5 mL DMEM. After incubation for five minutes at room temperature, Flp-In™ T-Rex 293 cells were supplemented with the DNA-PEI solution. Six hours post transfection, DMEM was replaced with medium containing 5% FCS and 1% Pen/Strep. After 48 hours, medium was aspirated, replaced with 1 mL fresh DMEM and cells were detached by pipetting up and down. Subsequently, cells were transferred into T75 flask and resuspended in 15 mL DMEM containing 5% tetracycline free FCS, 1% Pen/Strep, 15 µg/mL blasticidin and 100 µg/mL hygromycin. Addition of hygromycin started the process of positive selection which was carried out the following 25 days, with the medium being replaced every three days. It should be noted that the majority of stable cell lines was generated by Anja Schütz prior to this thesis.

$DMEM_{FlpIn}$	DMEM + 5% FCS, 1% Pen/Strep, 100 µg/µL zeocin, 15 µg/mL basticidin
Zeocin	InvivoGen, ant-zn
pOG44 Flp-Recombinase Expression Vector	Thermo Fisher Scientific, V600520

3.3.4.1 Cryopreservation and storage of stable cell lines

After the antibiotic selection was completed, stable cell lines were cryopreserved and stored. Prior to storing, a cryoprotective medium (storage medium) was prepared containing 50% DMEM, 30% FCS and 20% DMSO. Upon reaching a confluency of about 80% cells were washed with PBS, detached from the flask with Trypsin/EDTA solution and resuspended in cold $DMEM_{SCL}$. Cell count was then determined and batches of one million cells in 500 µL $DMEM_{SCL}$ were pipetted into cryotubes. Finally, 500 µL of the cryoprotectant were added to the cells, the tubes were placed into a freezing container and stored overnight at -20°C. On the following day, the samples were transferred to a nitrogen tank for longterm storage.

Storage medium	50% DMEM, 30% FCS, 20% DMSO
Dimethyl sulfoxide (DMSO)	SigmaAldrich, C6164-50ML
Mr Frosty™ Freezing Container	ThermoFisher Scientific, 5100-0001

3.3.4.2 Thawing of stable cell lines

Stable cell lines were resuscitated in a quick thawing process to reduce cell death due to the presence of DMSO. For this purpose, cryotubes were placed in a water bath without submersing the top of the tube and incubated in a 37°C water bath for three to five minutes. The cell suspension was then transferred into a falcon containing 10 mL preheated $DMEM_{SCL}$ and centrifuged at 100 rpm for five minutes. Subsequently, supernatant was discarded, the cell pellet was resuspended in 15 mL $DMEM_{SCL}$ and transferred into a T75 flask.

3.3.5 Flow cytometry of mammalian cells

Flow cytometry is an analytical tool for the characterization of cells based on their specific light scattering and fluorescent characteristics. As the cells travel in a thin capillary along a laminar stream of liquid, they are singularized and scanned individually by a laser. The scattering of light provides information on the size and granularity of cells. Simultaneously, fluorescent molecules are excited and their emission spectrum can be measured by various detectors [208].

Affinity of antibodies to Env was analyzed by flow cytometry using the Attune NxT (Applied Biosystems, Thermo Fisher Scientific) cytometer. For titration experiments, a

two-fold serial dilution of screening antibody was prepared with a starting concentration of 400 nM (assuming a molecular weight of 150 kDa) that was diluted in 12 steps to a minimal concentration of 0.2 nM. In general, antibodies carried an Alexa647-conjugate that was attached by using the Alexa Fluor® 647 Protein Labeling Kit as described in section 3.4.2.

Prior to every measurement, $4x10^4$ transiently transfected cells (see 3.2.1) were washed with 200 µL PBE. For this, cells were centrifuged at 500xg for five minutes at 4°C, the supernatant was decanted and cells were resuspended in 200 µL PBE. After removal of the supernatant, cells were incubated for one hour at 4°C with 25 µL PBE containing the bnAb VRC01 at the concentrations given above. Following three washing steps with ice cold 200 µL PBE, cells were resuspended in 200 µL PBE and analyzed with the Attune NxT device.

In respect to titrations performed on stable cell lines, cells were induced 24 hours prior FACS analysis to express GFP and Env by supplementing the medium with 1 µg/mL doxycycline. All following steps were conducted as described above.

In general, cells were gated for living, single cells displaying envelope (signal in APC channel - RL1) and GFP (signal in FITC channel – BL1) expression. As GFP and Env expression were genetically coupled, ratios of the mean fluorescence intensities (MFI) of APC- and FITC-signals were calculated to normalize for varying Env expression levels (relative MFI). The results were then compared with a wildtype control that underwent the same treatment as described above. Negative controls, represented by Env-negative cells that were subjected to the very same antibody titration as the other cells, were substracted to set the starting point of the curves to zero. The resulting titration curves were fitted using the option non-linear least squares regression (hyperbolar one-site binding) in the program GraphPad Prism 5.0.

PBE	PBS + 2mM EDTA + 0.5% FCS
Alexa Fluor® 647 Protein Labeling Kit	Life Technologies, A20173
Doxycycline	SigmaAldrich, D9891-1G

3.3.6 Cell sorting

Cell sorting is a specialized type of flow cytometry that allows the sorting of a heterogeneous mixture of cells. In the course of this work, this method was utilized to screen the previously generated stable cell line library for Env variants that exhibited enhanced or reduced affinity for the bnAb VRC01.

In this respect, a total of $30x10^6$ cells, composed of equal amounts of the pools of position to be analyzed, were seeded on a 15 cm plate and induced with 1 µg/mL doxycycline 24

hours prior to the sorting. On the following day, cells were detached from the plate with 4 mL chilled PBE by pipetting and washed one time by centrifuging at 250xg for five minutes at 4°C, decanting of the supernatant and resuspending in 4 mL PBE. After another centrifugation step, the cell pellet was resuspended in 1 mL Alexa647-labeled VRC01 solution (10 µg/mL) for 1 hour at 4°C. The cells were washed three times as described above, resuspended in 500 µL PBS and singularized by passing the cells through a 30 µm separation filter for subsequent sorting on the BD FACS Aria III. 50 µL of the sample was retained to serve as an input control.

Cells were sorted using a FACS Aria IIu (BD), with the instrument set to "single cell mode" to discard two-target-events ensuring the most accurate counts of the sorting procedure. The gating strategy was applied according to the following setup: i) FCS-A vs. SSC-A (P1), ii) FSC-H vs. FSC-W (P2), iii) SSC-H vs. SSC-W (P3), and FITC-A vs. APC-A (triangle shaped gates P4 and P5). Triangle shaped gates were chosen to sort cells with the highest or lowest VRC01 signal in relation to the GFP signal, thus selecting cells with the highest or lowest antibody affinity in relation to Env expression. Unless stated otherwise, at least 30.000 cells were sorted per gate into a tube containing 100 µL PBS. Genomic DNA from the input sample and sorted cells was recovered as described above (see 3.2.2.2) and utilized as a template in the sample preparation for subsequent NGS analysis (see sections 3.2.2-3.2.4).

QIAamp DNA Mini Kit	Qiagen, 51304
Doxycycline	SigmaAldrich, D9891-1G
Pre-separation filter	Miltenyi Biotech, 130-041-407

3.4 Protein Biochemistry

3.4.1 Purification of the HIV-1 specific human bnAbs VRC01

Expression of the bnAb VRC01 was conducted by co-transfecting of two plasmids carrying the variable heavy (CMVR VRC01 H, NIH AIDS reagent program) and light chain (CMVR VRC01 L, NIH AIDS reagent program) in a 1:1 ratio into Expi293F™ cells as described in section 3.3.3. All following steps were performed at 4°C.

Five days post transfection, the supernatant was collected by centrifugation of the cell suspension at 1000 rpm for 10 minutes. The purification of antibodies was accomplished by using a 5 mL column that consists of Protein A-Sepharose. Briefly, the column was equilibrated by applying 5 column volumes of Protein A buffer using a peristaltic pump. The flow rate was maintained at 10 mL per hour for all following steps. After loading of the supernatant, the column was washed with 25 mL Protein A buffer to remove unbound molecules and the antibodies were eluted in 1 mL fractions with a 0.1 M

glycine-HCl buffer (pH 2.7). The fractions were tested spectrophotometrically (absorbance at 280 and 650 nm) for their antibody content. All fractions containing antibody were pooled and neutralized by gradually adding 0.1 M NaOH in 5 µL steps until a pH of 6-7 could be detected with pH indicator strips. Typically, four to five mL were added in total. Dialysis of the eluate was carried out in Slide-A-Lyzer™ dialysis cassettes (molecular weight cutoff 20,000) for two days in 1 L PBS that was changed daily. Subsequently, the concentration was measured spectrophotometrically at an absorbance of 280 nm and 650 nm (molar extinction coefficient: 210,000 $M^{-1}cm^{-1}$). Typically, concentrations ranged between 2 and 4 mg/mL. The antibodies were stored at 4°C.

Protein A-Sepharose from *Staphylococcus aureus*	SigmaAldrich, P3391-1.5G
Protein A buffer	20 mM NaH_2PO_4, pH 7.0
Elution buffer	0.1 M glycine-HCl, pH 2.7
pHydrion Insta-Check 0-13	Micro Essential Laboratory, HJ-613
Slide-A-Lyzer™ Dialyse Cassette	Thermo Fisher Scientific, 66012

3.4.2 Labeling of antibodies

Purified antibodies were conjugated with a fluorophore by using the Alexa Fluor® 647 Protein Labeling Kit according to the manufacturers' instructions. Briefly, a 1 M solution of sodium bicarbonate was prepared by adding 1 mL deionized water to the provided vial of sodium bicarbonate and pipetting up and down until fully dissolved. To 0.5 ml of the 2 mg/mL antibody solution (usually in PBS), 50 µL of the bicarbonate was added and mixed by pipetting. Subsequently, the mixture was transferred to the vial of reactive dye which contained a magnetic stir bar and stirred for one hour at room temperature. Meanwhile the purification resin was pipetted into the column (provided with the kit) and allowed to settle. After the incubated mixture was loaded on the column and the solution entered the resin, elution buffer (8 mL of 10-fold diluted elution buffer in water) was added continuously until the labeled antibody had been eluted (typically about 30 minutes). Two colored bands which represented the separation of labeled protein from unincorporated dye could be observed. As the first band contained the labeled antibody, it was collected into a 2 mL tube and instantly centrifuged at 14.000 rpm, 4°C for 20 minutes to remove aggregates. The labeling efficiency and concentration of the antibody was measured spectrophotometrically at an absorbance of 280 and 650 nm. An aliquot of the labeled antibody as well as a 10-fold dilution (in elution buffer) were measured in triplicates. The degree of labeling was calculated with the following formula according to the manufacturers' instructions (4):

$$moles\ dye\ per\ mole\ antibody = \frac{A650\ x\ dilution\ factor}{239{,}000\ x\ protein\ concentration\ [M]} \quad (4)$$

where 239,000 $cm^{-1}M^{-1}$ represented the approximate molar extinction coefficient of Alexa Fluor 647 dye at 650 nm. Typically, three to five moles of fluorophore per mole of antibody could be detected (three to seven moles are optimal).

3.4.3 SDS-PAGE

Antibodies were separated according to their size in a sodiumdodecylsulfate-polyacrylamide gel-electrophoresis (SDS-PAGE). In general, all antibodies were analyzed under reducing and non-reducing conditions. For this purpose, 1 µg of purified antibody diluted in 20 µL PBS was supplemented with 5x Laemmli buffer and incubated at 95°C for five minutes. In respect to non-reducing conditions, 5x Laemmli buffer without β-mercaptoethanol was added instead. Subsequently, the samples were loaded onto a 10% SDS gel and electrophoresis was performed using a PowerPac™ Basic Power Supply (BioRad) at 70V for 20 min, followed by 90V for approximately two hours. PageRuler™ Prestained Protein Ladder (Life Technologies) served as a standard for estimating the size of the proteins assessed.

The SDS gel was first placed in a bowl containing Coomassie staining solution for five minutes and subsequently in destaining solution over night.

5x SDS-PAGE loading buffer	125 mM Tris, 2 % (w/v) SDS, 10 % (v/v) β-mercaptoethanol, 1 mM EDTA, 10 % (w/v) Glycerin, 0,01 % (w/v) Bromphenolblau, pH 6,8
Laemmli buffer	62.5 mM Tris; 1 % (w/v) SDS; 5 % (v/v) β-mercaptoethanol; 0.5 mM EDTA; 5 % (v/v) glycerol; 0.005 % (w/v) bromophenol blue, pH 6.8
PageRuler™ Plus Prestained Protein Ladder	Life Technologies, 26620
Coomassie Brialliant Blue R-250	AppliChem, A1092
Coomassie Staining Solution	0,125 % (w/v) Coomassie Brilliant Blue R-250, 50 % (v/v) ethanol, 7 % (v/v) acetic acid
Destaining solution	7 % (v/v) acetic acid

3.4.4 Envelope ELISA

All purified antibodies were characterized in respect to their affinity towards a soluble HIV-1 BG505 gp140 envelope protein which was previously produced. In addition, the binding affinity of the newly generated antibody was compared with a reference antibody of the same type, as well as a PBS negative control in triplicates, respectively.

High protein-binding ELISA flat bottom plates (Nunc-Maxisorp) were coated with 200 µL PBS containing 1 µg/µL *Galanthus nivalis lectin* (SigmaAldrich) per well at 4°C overnight. All following steps were performed at room temperature. Wash cycles were conducted with PBS/ 0.05 % Tween20 using a Tecan HydroFlex device (HydroControl-Software Version 1.0.). On the next day, wells were blocked for two hours with 200 µl PBS containing 5% skimmed milk powder, 5% heat-inactivated fetal calf serum, and 0.1% Tween 20 and washed subsequently three times with 200 µL wash solution. 100 µL of a solution comprising 1 µg/mL BG505 gp140 envelope protein in PBS was then added and incubated for one hour at room temperature. After six washing steps, wells were incubated for 1 hour with 100 µL monoclonal antibody that was 5-fold serially diluted (starting concentration 20 nM) in PBS containing 1% BSA. The following six washing steps ensured that all unbound antibodies were removed. A secondary HRP-conjugated rabbit anti-human IgG (100 µL/well in PBS) previously diluted in a ratio 1:5.000 in PBS was added to the wells, incubated for one hour and washed ten times. 50 µL of the colorimetric HRP substrate 3,3',5,5'-tetramethylbenzidine (TMB substrate consisting of TMB A : TMB B 20:1) was supplemented into each well until a blue color could be detected. The reaction was stopped with 25 µL of a 1 M H_2SO_4 solution followed by the measurment of the optical density at 450 nm in triplicates on a microplate reader. All curves were corrected by subtracting both the negative control signals obtained without addition of envelope gp140 and primary antibody. Furthermore, curves were further analyzed with GraphPad Prism 5.0 by fitting with the non-linear least squares regression (hyperbolar one site binding).

NUNC Maxisorp ELISA plates	Thermo Fisher Scientifc, 442404
Galanthus nivalis lectin	SigmaAldrich, L8275-5MG
Fetal Calf Serum	Merck, S0115
Tween20	SigmaAldrich, P9416-100ML
Bovine Serum Albumin	Biomol, 01400.100
Polyclonal Rabbit Anti-Human IgG/HRP	Dako, P0214
TMB A	30 mM potassium citrate, pH 4.2
TMB B	10 mM tetramethylbenzidine (Roth 6350.2), 10 % (v/v) acetone, 90 % (v/v) ethanol, 80 mM H_2O_2

4 Results

4.1 Overview of the sequential permutation library (SeqPer)

Mutant libraries representing a vast diversity of protein variants are increasingly utilized to identify structurally critical residues for protein function prediction. Considering the benefits of this technology it was decided to apply such a library for the selection of HIV envelope variants with improved antigenicity. For this purpose, a trimeric Env sequential permuation library was conceived.

Design and implementation of the library was conducted by Dr. Alexander Kliche in cooperation with the International AIDS Vaccine Initiative (IAVI). The trimeric Env library was based on the C-clade isolate, 16055 (accession code EF117268) to address the globally most prevalent subtype C. Truncation of the cytoplasmic tail (gp145 Env trimer) and adaptation of the isolate sequence to human codon usage were implemented to ensure efficient Env expression and surface presentation. Additionally, a substitution at the furin-cleavage site from REKR to REKS was introduced to prevent processing of the precursor protein into the non-covalently linked gp120-gp41 complex prone to shedding. Every position of the external envelope was sequentially permuted, i.e. by creating a separate plasmid-pool for each position (referred to as pool-of-position). The generation was performed by GeneArt using degenerate NNK codons (N = equal molar mix of A, C, G and T, and K = equal molar mix of G and T) within the synthesis process to allow diversification of DNA sequences. Notably, the first 30 amino acid positions representing the signal peptide as well as the cleavage site were excluded from the sequential permutation. In accordance with contractual stipulations, the minimal requirements regarding the diversity of the library constituted at least 16 amino acids per position resulting in a total of 658 sub-libraries, also termed as “pools-of-position”. Thus, the whole library comprises approximately 10.528-13.160 individual Env variants (figure 11). The library will be henceforth referred to as SeqPer (sequential permutation) library.

The library was inserted into the novel vector system, pQL13 (see 8.2.2), which was developed by Dr. Tim-Henrik Bruun during his PhD thesis [209]. A unique characteristic of this plasmid was represented in the translational coupling of tetracycline inducible Env expression with GFP production via a TaV 2A peptide, thus allowing normalization of envelope expression by means of GFP fluorescence. Additionally, inherent FRT (flippase recombination target) sites allowed the stable insertion of Env variants into Flp-In™ T-Rex 293 cells via site-directed recombination, enabling the generation of stable cell lines (see 3.3.4).

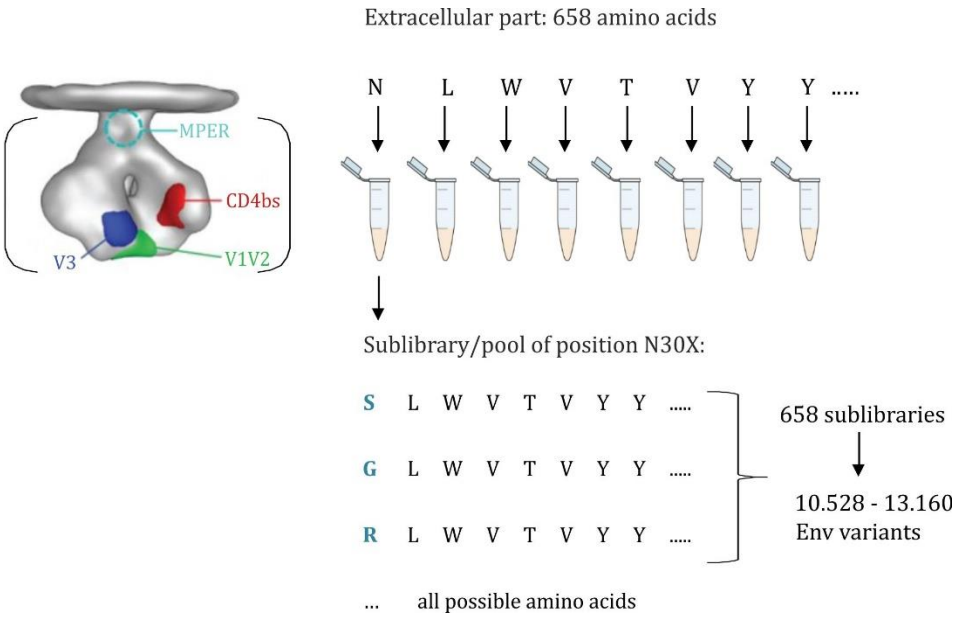


Figure 11 – Schematic illustration of the SeqPer library. The library was based on the extracellular part (indicated by brackets) of a trimeric C-clade envelope protein. Tubes represent the wild type amino acid position of the Env that was substituted by the 20 natural amino acids with the possible composition of the sub-library (pool-of-position N30X) illustrated below. Nomenclature of the pools of position was defined as follows: first letter (N) represents the wild type (WT) amino acid at the respective position in the sequence according to HXB2 numbering (30) and X indicating the randomization of this position, thus constituting a sub-library. Depending on the amount of amino acids occurring at every position (16-20) a total of 10.528 to 13.160 diverse variants are assumed in the 658 sublibraries. Figure of Env was adapted from the Nature Publishing Group [210] with minor modifications.

4.2 Generation of the stable cell line SeqPer library

Stable cell lines were created by stable transfection of Flp-In™ T-Rex 293 cells with a plasmid-mixture coding for an individual pool-of-position of the SeqPer (see 3.3.4). This event was mediated by the constitutively expressed Flp recombinase situated on the plasmid pOG44 which was supplemented in conjunction with the Env library during transfection. Upon co-transfection, the Flp recombinase mediated homologous recombination between the FRT sites in the plasmid and the cell's genome leading to the insertion of the-Env-containing plasmid into the Flp-In™ T-Rex 293 cell genome (figure 12A). As consequence, the *lacZ* Zeocin™ fusion gene shifts out of the reading frame and is inactivated. The resulting stable cell lines exhibited hygromycin resistance that served as antibiotic selection marker for cell with successful integration events. The sole integration locus at a predefined FRT site ensures the insertion of only one Env variant per cell. In addition, all stable cell lines display a translational coupling of eGFP and Env expression that results from a so-called Tav 2A peptide (TaVp2A) present between the two genes. The peptide originates from the insect virus *Thosea asigna* and is responsible for the impairment of peptide bond formation by a ribosomal skip mechanism between

the 2A glycine and 2B proline of the consensus motif Asp-Val/Ile-Glu-X-Asn-Pro-Gly (2A); Pro (2B) [211]. As a consequence, an equal expression level for eGFP and Env should occur which enables normalization for Env expression in respect to eGFP fluorescence. Due to the cytotoxicity of Env [212], a mechanism to regulate the expression was implemented comprising a combination of the Tet repressor in Flp-In™ T-Rex 293 cells and the doxycycline-inducible chimeric CMV promotor/TetO of the pQL13 plasmid. Generally, the functionality of all newly generated stable cell lines was analyzed by assessing their ability to express Env and GFP. Expression rates of both proteins were measured via flow cytometry after induction of the cells with doxycycline and staining with the antibody VRC01 conjugated with an Alexa647 fluorophore (see 3.3.5). The relative mean fluorescence intensities were calculated and compared with the WT 16055 stable cell line. In general, the detected MFIs were comparable to the WT signals indicating proper expression of Env and GFP, respectively. The entire stable cell line library generation and subsequent analysis of the expression was performed by Anja Schütz.

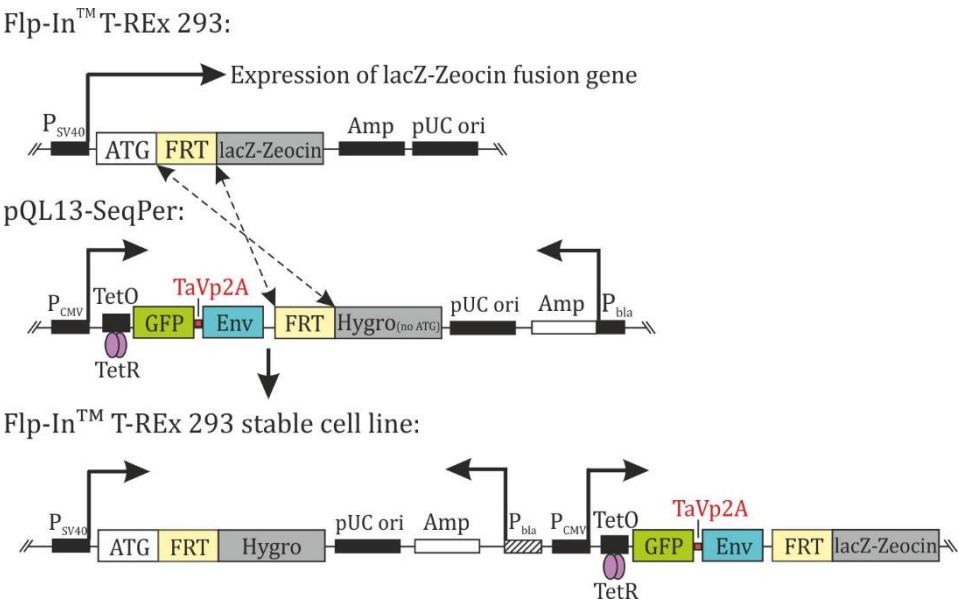


Figure 12 – Schematic illustration of the generation of stable cell lines. Linearized genomic organization of Flp-In™ T-Rex 293 cells and the pQL13 vector are shown before and after stable integration at Flp Recombination target sites (FRT). As a consequence, the hygromycin gene without start codon shifted into a reading frame with a start codon (ATG), leading to the acquisition of a hygromycin (Hygro) resistance whereas the zeocin resistance disappeared due to the loss of the start codon. Linkage between eGFP and Env expression was achieved by a TaV 2A peptide (TaVp2A) in pQL13. Expression of eGFP and Env was regulated by the inducible Tet operator/repressor (TetO/TetR) system. Figure was freely adapted from [213].

4.3 Quality control of the SeqPer library

The sequential permutation library was the basis of the screening for improved immunogens. Prior to this thesis, the library has already been analyzed partially by Christian Ziegler during his master's project [214]. Unfortunately, his data demonstrated several quality-related problems as he could confirm the existence deletions of various sizes in the V3 region of the pDNA library. Similarly, Dr. Alexander Kliche revealed a considerable loss in the amino acid diversity in the data from the NGS analysis of generated stable cell lines (data not shown). Considering the concerning previous results, especially in light of the complexity and dimension of the library, a detailed quality control should be performed to understand the characteristics and limitations of this type of library, as well as to interpret screening results correctly. Therefore, an in-depth evaluation of the quality was conducted on the level of the plasmid DNA that was originally obtained from the gene synthesis provider, and the generated stable cell lines at various stages of the generation and screening procedures.

4.3.1 Quality of plasmid DNA

Plasmid DNA was obtained by preparation (see 3.1) performed from the DNA stocks that were forwarded by GeneArt.

4.3.1.1 Purity of plasmid DNA

The general purity of plasmid DNAs was assessed spectrophotometrically. An average of 1.95±0.03 for the absorbance ratio at 260 and 280 nm (table S1) could be detected for all 658 library positions. The results were within the admissible range in regard of the optimal values ($OD_{260/280}$ ~ 1.8-1.9) for pure DNA according to literature [202].

4.3.1.2 Restriction enzyme assay

A restriction enzyme assay was performed on all 658 pools-of-position to evaluate potential errors (i.e. larger deletions, insertions, etc.) during the generation and the cloning procedure of the library. For this purpose, 1 µg plasmid DNA of every library

position was digested with the restriction endonuclease NcoI-HF (NEB) for 2 hours at 37°C and visualized on a 0.8 % agarose gel. As control, an equal amount of native pDNA was also analyzed to assess the content of supercoiled DNA. This was important for the generation of stable cell lines as transfection of supercoiled DNA is more efficient than open circular or linear DNA [215].

The expected fragment profile was calculated *in silico* with the 'Restriction Site Analysis' function of the CLC Workbench program, resulting in a band of 7937 base pairs for the native pDNA and four fragments with the sizes of 3744, 2353, 1519 and 321 base pairs for the digested samples, respectively.

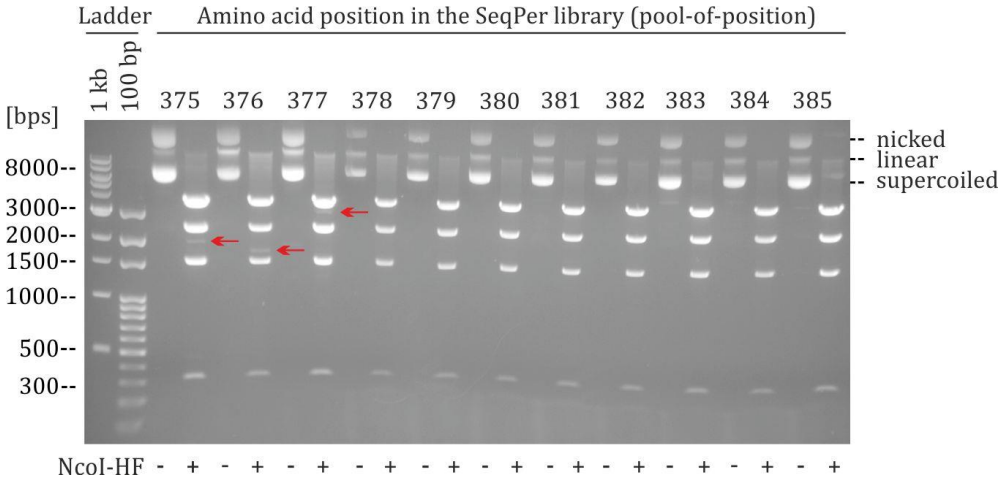


Figure 13 – Agarose gel electrophoresis of native and digested plasmid DNA. A representative restriction assay of various pools-of-position from the SeqPer library is shown. Non-digested DNA (NcoI-HF) migrated in three forms: i) nicked, ii) linear and iii) supercoiled DNA at the length of 7937 bps, with supercoiled DNA representing the biggest fraction. In respect to the digested DNA (NcoI-HF +), four bands at 3744, 2353, 1519 and 321 bps were expected. Aberrant fragments are indicated by red arrows.

For the majority of the digestions, the resulting patterns displayed the expected fragment sizes. However, in 48.3% (318 of 658) of the library, band patterns with additional fragments could be detected (figure 13, S1). All aberrant bands however, manifested in substantially lower intensities than then wildtype fragments, indicating that merely a fractional amount of the respective positions is erroneous. Furthermore, the sublibraries 522-532 (figure S2) demonstrated only one distinct band at 3744 bp with a multitude of additional smaller low-intensity fragments resulting in a smear-like appearance. New preparation of DNA from the stocks or after re-transformation of bacteria with the pDNA did not improve the quality of the digestion which indicated general issues during the production of these positions. When analyzing non-digested DNA, three bands could be detected in most cases: i) nicked, ii) linear and iii) supercoiled DNA. Expectedly, the supercoiled pDNA represented the largest fraction in the library.

4.3.1.3 Densitometric analysis of aberrational phenotypes

Densitometric analysis of digital gel images provided a cost and time effective alternative to analyze relative DNA quantities. In this context, the percentage of visually detected aberrations was determined by comparing the band intensities of the wildtype fragments with the erroneous fragments to estimate their occurrence in the SeqPer library. The results are displayed in figure 14 and revealed an average rate of aberration of 6-18% at the 318 visually erroneous sublibraries, with pool-of-position 530 displaying the highest error rate with about 40%. Furthermore, an increased accumulation of aberrations could be identified at two major hot spots located in the stretch between the variable loop 5 (V5) and heptad-repeat region 1 (HR1), as well as in the center of the gp41 region, indicating poor quality at these library positions.

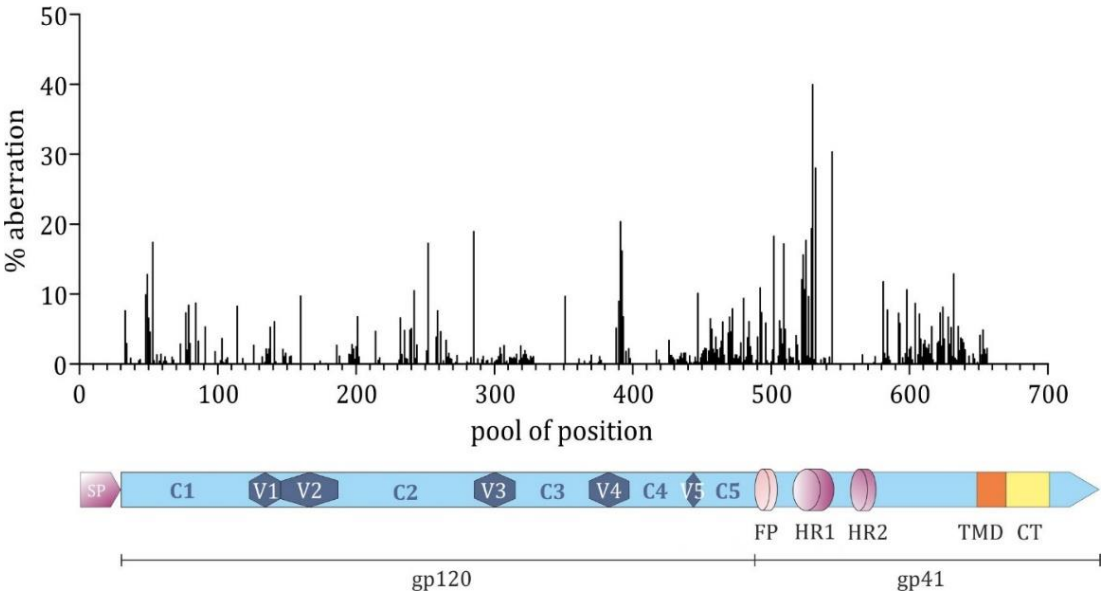


Figure 14 – Depiction of errors in the SeqPer library. Percentage of detected aberrational fragments plotted against the respective 318 erroneous pools of position in the SeqPer library. The location of the deviating positions on Env is indicated below. Schematic illustration of Env was freely adapted from [37].

4.3.1.4 Characterization of selected aberrational phenotypes

In order to analyze the origin of the anomalies, three random sub-libraries (pool-of-positions 50, 136, 316) exhibiting deviating fragments (figure S2) were selected for an in-depth evaluation. Briefly, 100 ng of pool DNA was used for the transformation of chemically competent *E. coli* DH10B cells. The resulting colony forming units represented individual Env variants from the pool-of-position, respectively. Plasmid DNA from 40 colonies per pool-of-position was extracted and characterized by a

restriction enzyme assay. The fragment profile of native and NcoI-HF-treated pDNA (2h, 37°C) was visualized on a 0.8 % agarose gel.

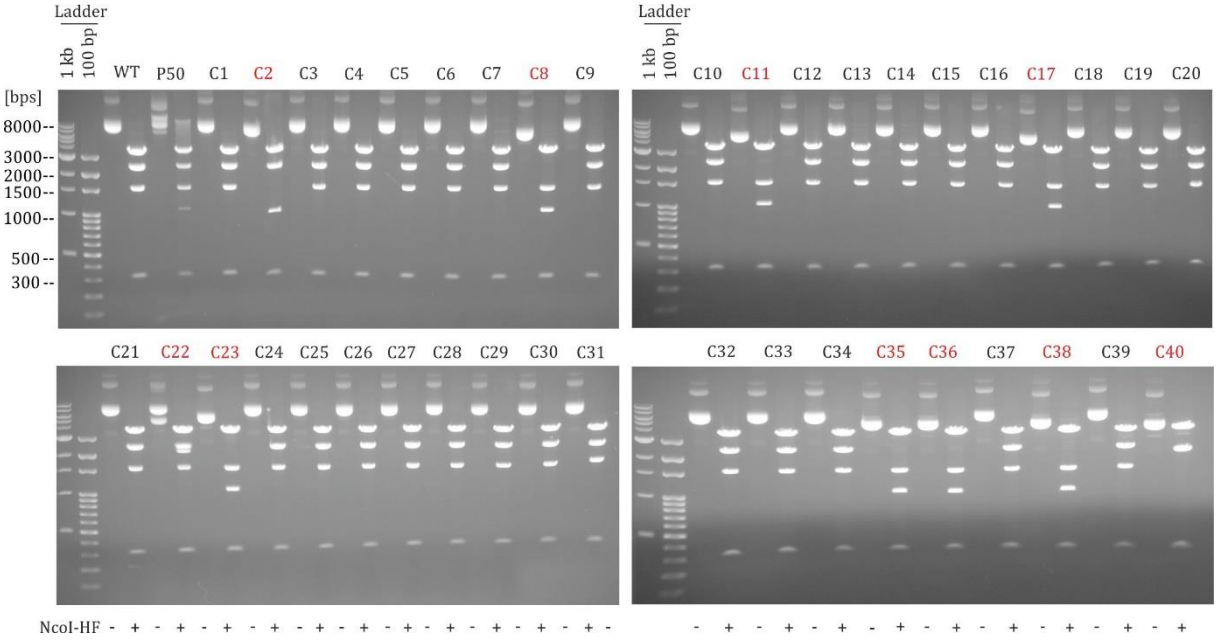


Figure 15 – Restriction enzyme assay profile for sub-library 50. Plasmid DNA of 40 random colony-forming units (C1-C40) from sub-library 50 in the SeqPer library was analyzed. 1 µg native and digested (NcoI-HF, 2h, 37°C) plasmid DNA was visualized on a 0.8 % agarose gel. The wildtype restriction pattern from the 16055 gp145 Env mastergene (WT) and the pool-of-position 50 (P50) are also shown. Aberrational colony forming units are highlighted in red.

The digestion patterns from the individual clones of the three selected positions demonstrated mainly the correct phenotype. However, a minority of the analyzed colonies proved to be erroneous which is consistent with the lower band intensities of the detected aberrations from the restriction enzyme assays of the plasmid pools of sublibraries (figure 13, S2). At positions 50 and 316 ten out of 40 colonies displayed aberrations, whereas position 136 demonstrated only four faulty clones (figure 15, S3). On closer examination of the restriction profile of sub-library 50 (figure 15, P50), an additional faint fragment of approximately 1 kbs can be detected, representing a contamination of about 6.5% (densitometric analysis, 4.3.1.3) at the whole library position. Sanger sequencing of 40 random colony-forming units revealed that eight out of ten faulty clones (C2, C8, C11, C17, C23, C35-36 and C38) carried deletions of about 1.3 kbs within the Env region. This lead to an aberrational digestion pattern of single colonies and subsequently caused the emergence of a faint additional fragment that is visible for pool-of-position 50 (P50). Furthermore, mutations and deletions within the vector occurred, as in the case of clones 22 and 40, however, they were not visible in the restriction enzyme assay probably due to their under-representation. Similarly, Sanger

sequencing of faulty clones from positions 136 and 316 demonstrated deletions of varying sizes and insertions of additional restriction sites in the vector as well as the envelope region (figure S3 A and B). This indicated technical limitations during the generation of the library or cloning procedure.

4.3.2 Establishment and validation of the Next Generation Sequencing library sample procedure

In the following, all samples involved in the NGS analysis are termed as 'NGS library' or 'NGS amplicons' to prevent confusion with the SeqPer library. Understanding the NGS library preparation is important to ensure the highest quality sequencing data. Therefore, all relevant conditions regarding library preparation for NGS analysis had to be optimized and standardized for quality control of the SeqPer library and for future utilization.

4.3.2.1 Determination of PCR conditions for NGS library preparation

A successful and efficient NGS analysis strongly relies on high quality NGS amplicons. Sufficient quantities of DNA for the denaturation step are just as important as pure and faultless samples. Due to the error-prone nature of DNA polymerases, a careful balance between the optimal amount of input DNA and PCR cycles had to be determined first. Here, the Phusion DNA polymerase was used that has a 3'-5'-exonuclease activity for proofreading, for which an error rate of $4.4x10^{-7}$ in Phusion HF buffer is given by the manufacturer (NEB). Since genomic DNA was often a limiting factor, DNA from different amounts of cells (500, 2500 & 12.500) was extracted and amplified with 19, 22 or 25 cycles, respectively, to establish the conditions for optimal yield of PCR products. In general, 10 µL isolated gDNA were utilized for PCR1 as concentrations were below the detection limit. Similarly, 10 µL of purified PCR1 product (see 3.2.3) was applied for PCR2. Assuming that the extraction of gDNA worked optimally and no DNA degradation took place, 10 µL isolated gDNA from 500, 2500 and 12.500 cells, resembling 0.004, 0.02 and 0.1 pg, respectively, were utilized in the initial PCR.

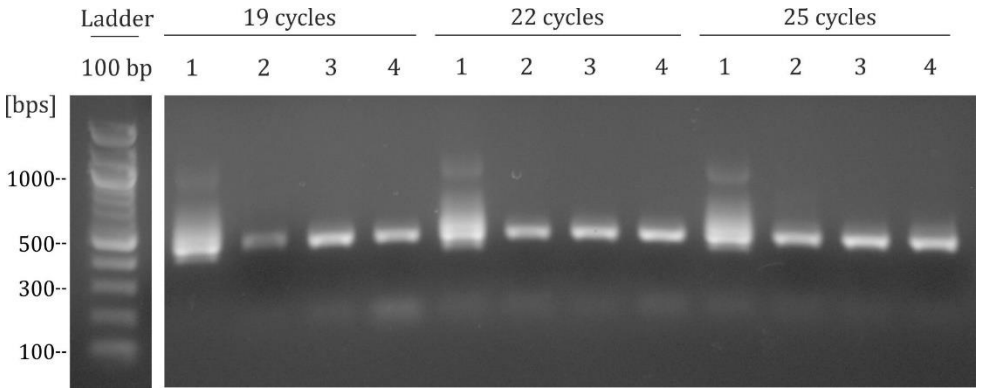


Figure 16 - Analysis of PCR conditions for NGS sample preparation. Sample preparation comprised two amplification steps. In the first PCR, NGS-specific primer binding sites were introduced into the target DNA template. Fusion of indices and adapters was achieved by the second PCR. To determine optimal conditions for NGS sample preparation 19, 22 and 25 cycles were tested in combination with different amounts of input DNA. In respect to plasmid DNA, 1 ng was utilized in the initial PCR (1), whereas 0.004 (2), 0.02 (3) and 0.1 (4) pg genomic DNA previously extracted from 500, 2.500 and 12.500 cells (WT 16055 gp145 stable cell line) were applied, respectively. Amplification products shown here were visualized after the second PCR by agarose gel electrophoresis.

Sample	Conc. [ng/µL]	$A_{260/280}$
19_1	76.4	2.0
19_2	7.5	3.0
19_3	20.2	2.1
19_4	34.3	2.2
22_1	67.5	1.9
22_2	29.9	2.1
22_3	28.9	2.1
22_4	39.2	2.1
25_1	48.8	2.1
25_2	21.9	2.1
25_3	35.3	2.1
25_4	36.5	2.0

Table 5 – Quality control of amplification products. General purity of PCR products was determined spectrophotometrically by measuring the absorbance ratios $A_{260/280}$ and $A_{260/230}$. Sample names correspond to figure 16 and describe the amount of cycles and input DNA utilized in the initial PCR (i.e. 19_1 relates to 19 cycles and 1 ng pDNA).

As expected, a higher number of cycles and amount of input material resulted in a greater yield of final DNA after gel extraction of the corresponding bands. Whereas genomic DNA exhibited distinct bands in the gel electrophoresis, the resulting fragment from the plasmid DNA displayed a smear-like appearance which indicated unspecific binding of primers, probably due to an excess of DNA template applied in the PCR. However, neither the number of cycles nor the amount of input material seemed to have a direct influence on the quality of the resulting product. The only exception was represented in the sample 19_2 which displayed poor quality compared to the other specimens indicating potential contamination. Considering the error-prone nature of DNA polymerases, less PCR cycles in combination with higher amounts of input DNA

were considered for all further experiments, specifically 22 cycles and target DNA template amounts of 1 ng for pDNA and at least 0.1 pg for genomic DNA.

4.3.2.2 Adjustment of the purification of amplicons

Although the initial NGS runs demonstrated satisfactory results, in several of the ensuing experiments the NGS analysis was terminated by the MiSeq device. The reason for the failed sequencing was presumed to be an overloading of the flow cell which can occur in cases of extreme overclustering. In this respect, a cluster density of over 1.000 K/mm^2 (between 600 and 800 K clusters per mm^2 are optimal) was detected, indicating an excessive overclustering. As optimal NGS strongly relies on accurate quantification, the sample composition was analyzed to assess probable sources of error. This was achieved by closer inspection of the Bioanalyzer data of individual samples obtained after the purification procedure with magnetic beads. In addition to the expected amplicon in the size of approximately 400 bps, an additional peak at about 130 bps was detected with the Bioanalyzer in most of the samples (figure 17 A). Notably, the band was also faintly visible in the agarose gel (figure 16). This impurity might have led to inaccurate quantification by qPCR and consequently, to overloading of the flow cell. The additional peak was attributed to an accumulation of adapter or primer dimers, as the forward and reverse primers utilized in the second PCR amplification during NGS sample preparation featured a length of 63 bps, respectively. Since magnetic beads are unable to remove fragments longer than 100 bps, the purification approach seemed inefficient. To ensure the removal of the dimers, correct bands were excised after agarose gel electrophoresis and the DNA was extracted as described above (see 3.2.3.1). The subsequent assessment of the samples with the Bioanalyzer demonstrated a single peak at the appropriate length (figure 17 B). Since the problem of overloading did not occur in the following NGS experiments any more, the purification procedure was adjusted correspondingly to include extraction from an agarose gel.

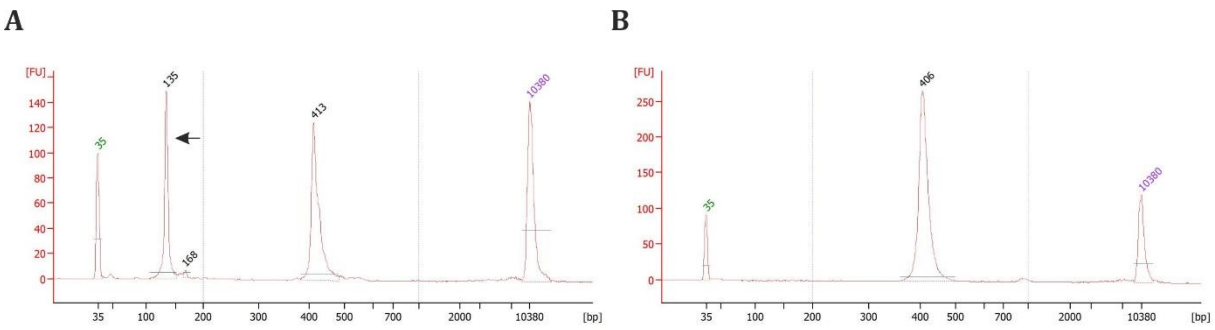


Figure 17 – Bioanalyzer data of an exemplary library amplicon after purification via **(A)** magnetic beads and **(B)** extraction from agarose gel. The x-axis shows the size of the detected samples, whereby 35 and

10.380 bps represent the lower and upper markers, respectively. Fluorescence units (FU) indicate the intensity of the samples. The arrow denotes the accumulated adapter or primer dimers (approximately 130 bps). Notably, the dimers are completely removed from the sample after excision of the appropriate band and purification from the agarose gel.

4.3.2.3 NGS background determination

Before the *de facto* deep sequencing of samples, a proper consideration of the necessary coverage and cluster densities is required since both have large impact on sequencing performance in terms of data quality and output. In order to prevent an over- or under-clustering which can lead to poor run performance or even sequencing failure, an accurate estimation of sample amounts to be loaded is essential. While the necessary coverage can be easily calculated given a known degree of expected diversity, the background level originating from the NGS device and sample preparation represented the sole unknown factor. The Illumina platform suffers from numerous biases due to imperfect chemistry and sensors [216]. The main problems occur in the form of phasing and pre-phasing. Phasing refers to failed incorporation of a base in a given cycle leading to sequencing that lags behind, whereas pre-phasing refers to synthesis of multiple bases in a single cycle [217]. This results in base-calling errors which increase with each cycle, subsequently, limiting the overall read length. Thus, it was important to determine the total number of errors that are introduced during sample preparation and by the MiSeq device. Therefore, samples were generated that were used to address both potential sources of error.

The investigation of intrinsic MiSeq device errors was performed on a defined and error-free reference template. The reference sample was generated by PCR amplification using the primers that introduce the required adapters and indices (see 3.2.2), followed by blunt end cloning into a pUC18-MlyI plasmid. This vector contained a MlyI-cassette that was inserted previously over inherent SmaI restriction sites (see 6.2.3). Correctness of the insert was confirmed by Sanger sequencing. Subsequently, the DNA insert was extracted by MlyI digestion, purified using AMPure magnetic beads (3.2.3) and could directly be used for the NGS sequencing. Thus, the error rate should correspond to the inaccuracy of the plasmid replication in *E. coli* ($5x10^{-10}$ per base pair/round of replication) [218].

In order to assess errors pertaining to amplicon generation by PCR, the samples from the experiment described above (see 4.3.2.1) were compared with the error-free reference template in the same NGS run.

The results are depicted in figure 18 and exhibit the maximal error probability from the MiSeq as well as the sample preparation. As expected, the error rate from the NGS device (figure 18, sample 1) was relatively low, displaying about 12 mutations per 10^5 reads.

Accuracy of the Phusion DNA Polymerase was apparently only little compromised with increased amounts of PCR cycles during the amplification, as no distinct difference in the error rates between the various cycles could be identified. Conspicuous, was the influence of the utilized DNA quantity. The lower the initial amount of DNA applied in the PCR, the more errors were detected, exemplified in the sample with the lowest input of genomic DNA (0.004 pg, samples 5-7) displaying the most errors. All in all, the maximal error rate (referring to sample 7) originating from the NGS device and Phusion DNA Polymerase amounted to a total of 0.0004% per base for a 300 bp long amplicon.

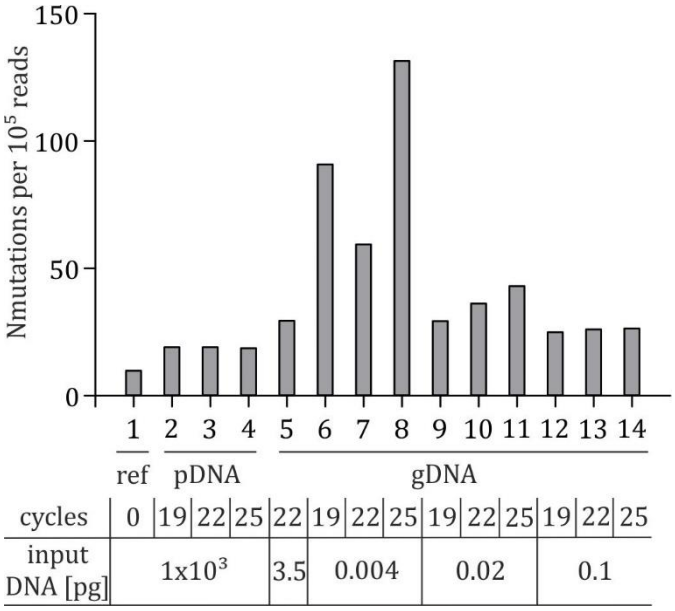


Figure 18 – Total error rate expected during an NGS run. The number of detected mutations ($N_{mutations}$) per 10^5 reads are shown. To determine the PCR conditions with the lowest error-rate, various amounts of cycles (19, 22 and 25) and input DNA were analyzed. Errors originating from the NGS device were assessed on an error-free reference sample (ref, 1). 50% PhiX was utilized for the NGS run.

In conclusion, all further NGS samples were generated by amplification with 22 cycles for each PCR, respectively. In terms of initial DNA quantity, at least 0.1 pg genomic DNA extracted from >12.500 cells were generally utilized, depending on availability. This amount was chosen due to the seemingly reduced error susceptibility when gDNA from more cells were used in the amplification. Based on the above data, a background mutation rate of 0.0004% per base can be estimated under these conditions.

4.3.3 Determination of the diversity of the SeqPer library on the example of the CD4 binding site

Apart from the general DNA quality of the SeqPer library, quality controls also comprised analysis of the diversity. Specifically, this includes the amount of amino acids

and the percentage amino acid distribution at every pool-of-position. Ideally, all 20 amino acids should be detectable for each position. Comprehensive analysis of the codon composition of the pDNA library was performed by NGS. Subsequently, the codons were translated into amino acids by using a programme that counts every codon at each position and delivers the absolute numbers of amino acids (see 3.2.6). In the following, the translated codon composition is referred to as the (decoded) amino acid composition of pDNA or stable cell lines. Similiary, the respective stable cell lines were also analyzed to assess the diversity of the pDNA and stable cell line library. This information was important for both the generation of stable cell lines as well as the assessment of results from the screening for improved Env immunogens. Low amino acid diversity could lead to reduced variability in the stable cell lines, thus, limiting the pool from which immunogens are selected. Due to the considerable size and complexity of the library, the diversity was determined by Next Generation Sequencing. Over the course of this PhD thesis, the analysis of the diversity focused exclusively on the CD4 binding site region of the SeqPer library, as concentrating solely on this region offered various advantages. Firstly, the CD4bs constitutes a small and manageable area that demonstrated a good quality regarding the plasmid DNA from quality controls (see 4.3.1). Additionally, with 46 positions the size was appropriate for the establishment, standardization and validation of most of the applied methods (e.g. NGS sample preparation and subsequent analysis, FACS-based screening, etc.) while keeping the high expenses of NGS to a minimum. Also, the CD4bs represents one of the most conserved and thus important regions of HIV-1 bringing forth the most potent bnAbs during infection. Last but not least, these CD4bs-specific bnAbs were easily accessible due to the established in-house production which proved to be favorable as high quantities of antibody that were required for multiple experiments.

4.3.3.1 Diversity of the CD4 binding site on the level of plasmid DNA

The CD4bs consists of 46 amino acids that are arranged over a stretch of 579 base pairs into the following four discontinuous segments on the envelope: loop D, CD4 binding loop, bridging sheet (bs) and variable loop 5 (V5) [219,220]. Over the course of this thesis, the entire CD4 binding site (CD4bs) of the SeqPer library was analyzed by NGS regarding the available amino acid distribution in the plasmid DNA as well as in the corresponding stable cell lines. In consequence of the disjointed regions and the limited read length of the MiSeq reagent kit v2 (300 bps), the amino acid positions were divided by means of their functional segments into the four following DNA regions: i) loop D, ii) CD4 binding loop and iii) bridging sheet and iv) variable loop 5. For NGS analysis the segment pools were generated by pooling every amino acid position of the respective region in an equal ratio.

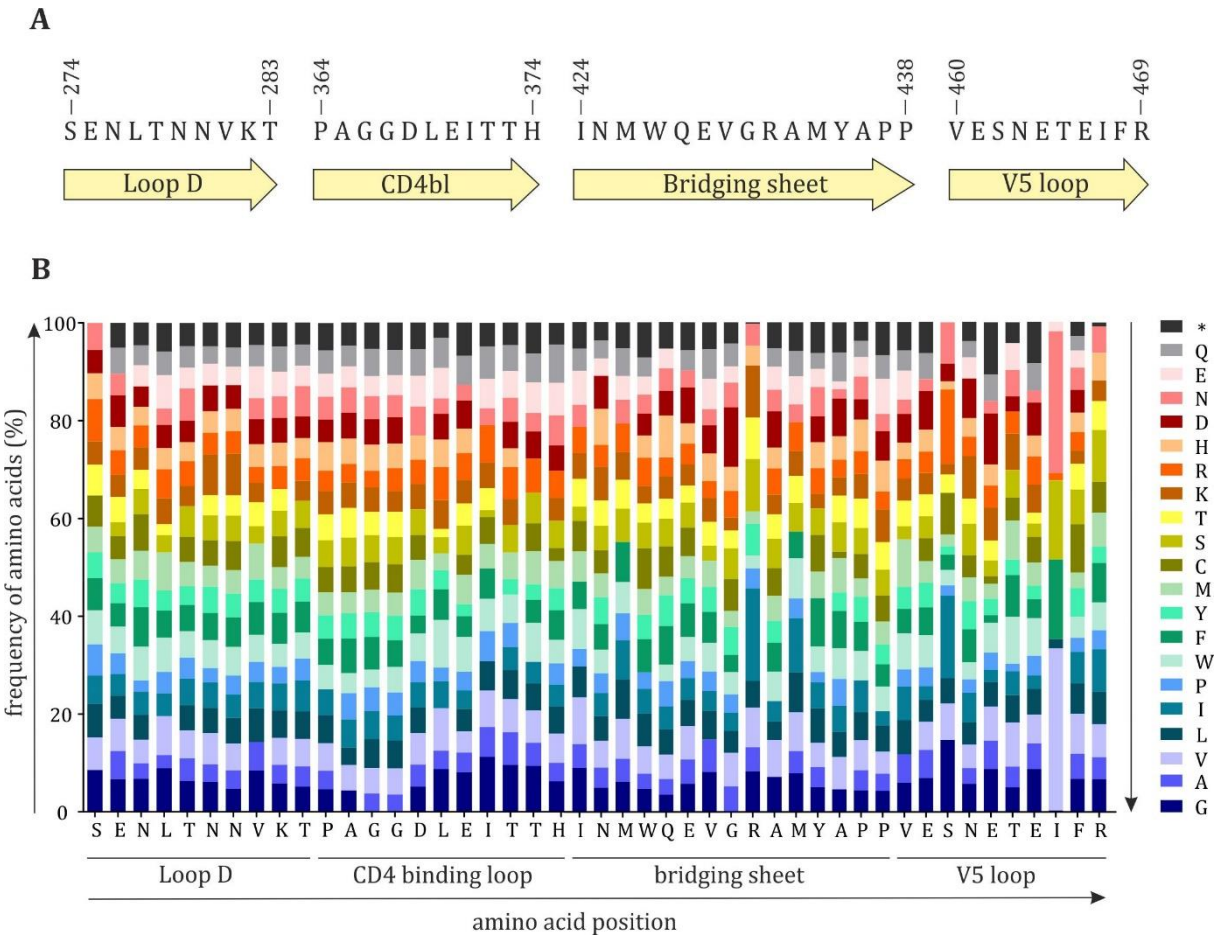


Figure 19 – (A) Schematic illustration of the CD4 binding site. The CD4bs consists of the four discontinuous segments Loop D, CD4 binding loop, bridging sheet and V5 loop. HXB2 numbering was applied. **(B) Amino acid distribution of the complete CD4 binding site on the level of pDNA.** Heights of the stacked bars (y-axis) represent the detected amino acids in percent at every pool-of-position in the SeqPer library (x-axis). To provide a clearer overview, the respective wildtype amino acid at every pool-of-position was excluded from the data set.

As the variable positions of the library were generated by using NNK oligonucleotides which encompasses 32 codons, optimally, every codon should be represented with 3.1%. This scenario changes slightly in the case of amino acids due to the degeneracy of codons. For instance, the amino acids Arg, Leu and Ser are encoded by six different codons. Consequently, these amino acids should occur with a frequency of 6x3.1%, whereas the amino acids encoded by one codon, such as Met and Trp, should be represented with 3.1%. Irrespective of the degeneracy of codons, the optimal pool-of-position should consist of 20 amino acids with a frequency of 3.1% for every amino acid, respectively.

NGS analysis of the CD4bs on the level of plasmid DNA demonstrated a high variability at almost every pool-of-position with an average of 19 amino acids (figure 20 A) while maintaining nearly optimal distribution of amino acids. The even distribution of amino

acids is represented in figure 19 B by the almost equal heights of the stacked bars. In general, the wildtype amino acid for every pool-of-position was highly overrepresented in the results which was a consequence from the pooling of multiple sub-libraries that display variability only in one position, whereas the rest remains constant. Therefore, to provide a clearer overview of the frequency for the non-wildtype amino acids, the wildtype amino acid data was excluded from the depiction. Only a small minority of the pool-of-positions displayed poor diversity. These include sub-libraries S274X (loop D), I371X (CD4 binding loop), R432X and M434X (bridging sheet), S462X, I467X and R469X (V5 loop), all displaying a lower amount (between 10 to 17 aas) as well as suboptimal distribution of amino acids.

4.3.3.2 Diversity of the CD4 binding site on the level of stable cell lines

In addition to the diversity of the plasmid DNA, the diversity of the generated stable cell lines was analyzed. In this regard, stable the cell lines were investigated in all major stages of the overall screening procedure to assess potential changes in the amino acid distribution during sample preparation: i) stock directly after thawing, ii) before induction with doxycycline and iii) after induction (see 3.2.2.3). Thus, in the following, the stable cell line samples analyzed by NGS were termed ‘stock’, ‘non-induced’ and ‘induced’.

The results from the NGS analysis demonstrated a significant decrease of diversity by approximately 32% in the stable cell lines compared to the plasmid DNA (figure 20 A) that was initially utilized for the generation of the stable cell lines. Whereas an average of 19 decoded amino acids were detected in the pDNA, 14 were present in the stock and non-induced stable and only 12 in the induced stable cell lines (figure 20 A and B). The CD4 binding loop exhibited the largest loss in diversity. Furthermore, a general trend towards a gradual reduction in diversity could be detected from pDNA>stock/non-induced>induced cell lines (figure 20 A, table 6). No loss in diversity between stock and non-induced stable cell lines was apparent. Strikingly, the nearly optimal decoded amino acid distribution in the plasmid DNA underwent an extreme shift towards specific variants in all stable cell lines (figure 20 C, S4). The results indicated that the loss of diversity occurred during the generation of stable cell lines.

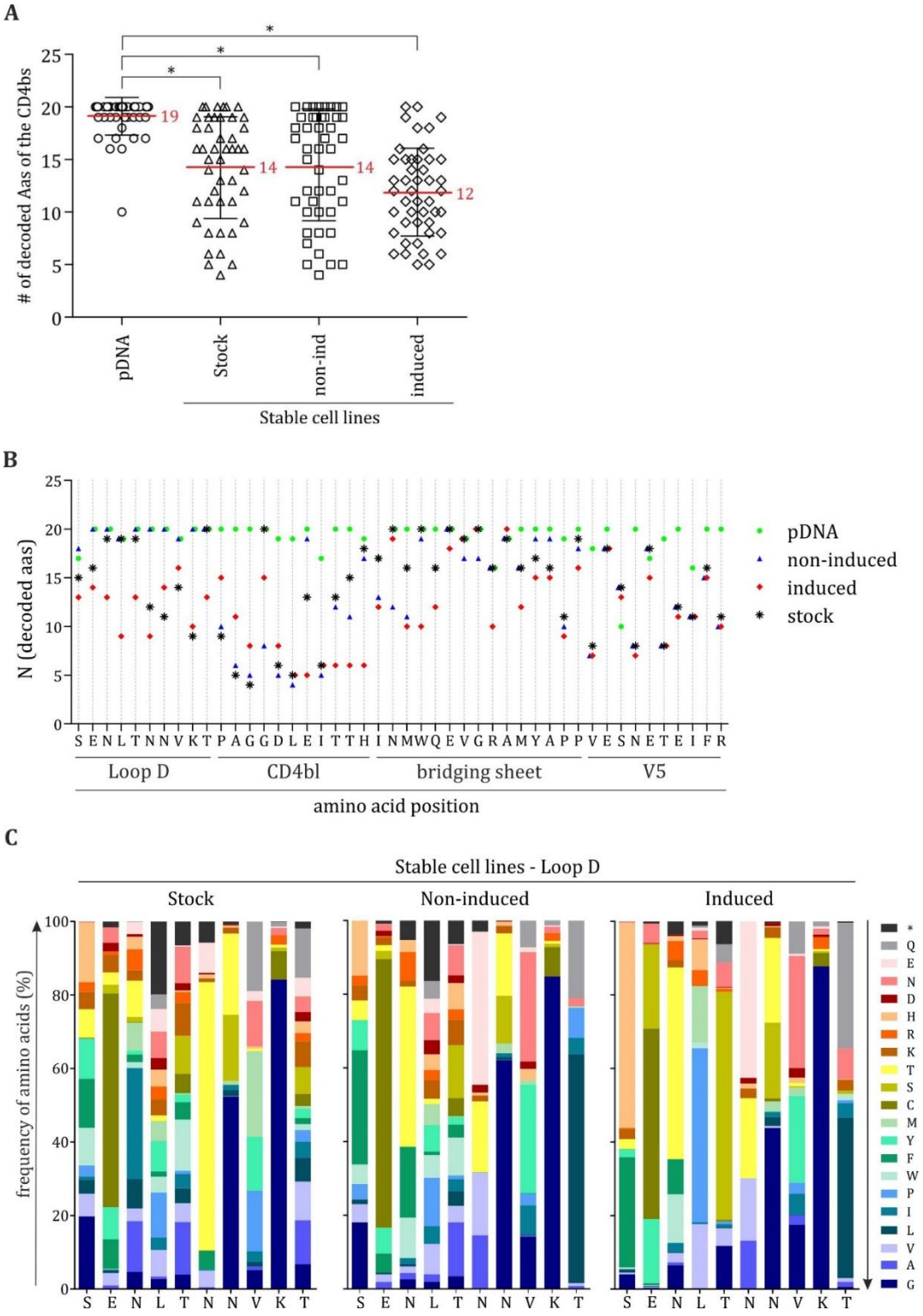


Figure 20 – (A) Total number of decoded amino acids in pDNA and stable cell lines. The detected amino acids in pDNA and stock, non-induced and induced stable cell lines of the whole CD4bs are shown. Red lines and numbers denote the mean of the detected amino acids. Statistical significance ($p<0.0001$ for all samples) is indicated by asterisks and was calculated with Graphpad Prism using a Wilcoxon signed rank test. **(B) Summary of the variability of the whole CD4 binding site on the level of pDNA and the stable cell line (SCL) counterpart**. Analysis of SCLs was performed on stock, non-induced and induced pools of position. The x-axis denotes the wild type amino acid sequence of the whole CD4bs, y-axis displays the amount (N) of the decoded amino acids at every position in the CD4bs library. **(C) Amino acid distribution of stock (left), non-induced (middle) and induced Loop D (right).** Heights of the stacked

bars represent the detected amino acids in percent for every pool-of-position in the SeqPer library. To provide a clearer overview, the respective wildtype amino acid for every pool-of-position was excluded from the data set.

4.3.3.3 Quantitative representation of amino acid diversity

Diversity describes a statistical measure of sequence conservation or variation. In the case of the SeqPer library, diversity represents the amount of the 20 possible amino acids that are observed at a given position. Thus, if the pool-of-position is completely conserved (i.e. only one amino acid is present), the diversity is 0.05 (1/20); if the optimal position comprises all amino acids, the diversity is 1.0 (20/20) [221]. However, possible differences in the distribution among the amino acids are not taken into account in this representation of diversity, leading to a loss of information. Here, a quantitative measure of amino acid variability is introduced in the form of a diversity coefficient that gives an estimation on the frequency of amino acids at a given position. Calculation of the diversity was deduced from the Wu-Kabat variability coefficient which is defined as [222]:

$$variability = \frac{N * k}{n} \quad (5)$$

where N describes the number of sequences, k the number of different amino acids at a given position and n the frequency of the most common amino acid at that position [223]. In general, the NGS output data from all experiments was transcribed into a table listing the number of sequences that were assigned a specific amino acid. Since these sequences were subsequently converted to represent the percentage amino acid distribution, N was set to 1, resulting in the new equation (6):

$$diversity\ (D) = \frac{k}{n} \quad (6)$$

with k representing the number of amino acids at a given position and n the frequency of the most common amino acid at that position. Notably, the calculation refers solely to non-wildtype amino acids. Consequently, the relative diversity D_{rel} was calculated with the equation (7):

$$Drel = \frac{detected\ diversity}{maximal\ diversity} \quad (7)$$

From the formula (7), it follows that $0 \leq D_{rel} \leq 1$, with D values near 1 representing high variability and those near 0 almost no diversity.

Table 6 – Relative diversity (D_{rel}) of the whole CD4 binding site. Diversity coefficients of pDNA and stable cell lines were calculated according to equation 3. Intensities of the heatmap denote the extent of the variability at every pool-of-position, with 0.0 representing the minimal and 1.0 the maximal relative diversity.

relative Diversity $D_{rel.}$ [a)]

		pDNA	stable cell lines					pDNA	stable cell lines		
	Aa pos.		Stock	Non-ind	Ind		Aa pos.		Stock	Non-ind	Ind
Loop D	S	0,49	0,19	0,10	0,07	Bridging	I	0,45	0,40	0,06	0,15
	E	0,75	0,07	0,05	0,06	sheet	N	0,68	0,33	0,04	0,12
	N	0,62	0,16	0,07	0,07		M	0,61	0,19	0,06	0,13
	L	0,53	0,24	0,29	0,08		W	0,60	0,55	0,39	0,04
	T	0,76	0,33	0,34	0,06		Q	0,58	0,20	0,20	0,08
	N	0,60	0,04	0,05	0,05		E	0,69	0,24	0,28	0,28
	N	0,59	0,05	0,04	0,08		V	0,58	0,20	0,11	0,12
	V	0,56	0,15	0,12	0,13		G	0,41	0,18	0,16	0,10
	K	0,78	0,03	0,03	0,03		R	0,23	0,19	0,11	0,06
	T	0,79	0,37	0,04	0,07		A	0,67	0,36	0,29	0,11
							M	0,36	0,31	0,20	0,08
CD4bl	P	0,86	0,06	0,05	0,12		Y	0,48	0,25	0,19	0,08
	A	0,70	0,02	0,04	0,04		A	0,65	0,24	0,31	0,30
	G	0,75	0,02	0,02	0,04		P	0,77	0,07	0,08	0,09
	G	0,87	0,03	0,02	0,08		P	0,70	0,34	0,26	0,24
	D	0,74	0,02	0,02	0,03						
	L	0,49	0,01	0,01	0,02	V5	V	0,49	0,04	0,05	0,03
	E	0,62	0,19	0,17	0,04		E	0,64	0,14	0,17	0,13
	I	0,38	0,03	0,02	0,05		S	0,24	0,20	0,15	0,17
	T	0,52	0,09	0,07	0,05		N	0,57	0,05	0,04	0,04
	T	0,53	0,06	0,05	0,03		E	0,47	0,14	0,26	0,19
	H	0,61	0,22	0,16	0,02		T	0,47	0,04	0,04	0,03
							E	0,57	0,10	0,10	0,09
							I	0,08	0,13	0,12	0,12
							F	0,51	0,18	0,17	0,16
							R	0,40	0,06	0,05	0,05

min. 0.0 — 1.0 max.
diversity

a) calculations were made according to equation: $relative\ diversity\ D_{rel.} = \frac{D_{detected}}{D_{max}}$ and $diversity\ D = \frac{k}{n}$

k = amount of non-WT aas, n = frequency of most represented aa

Strikingly, a significant reduction in the relative diversity in the stable cell lines could be demonstrated confirming the emergence of an extensive bias in the amino acid distribution (table 6, figure 21). In addition, an incremental decrease of diversity, from pDNA>stock>non-induced>induced stable cell lines, was detected for most of the pools-of-position which is consistent with the reduced amount of amino acids in the stable cell lines (figure 20 A and figure 21).

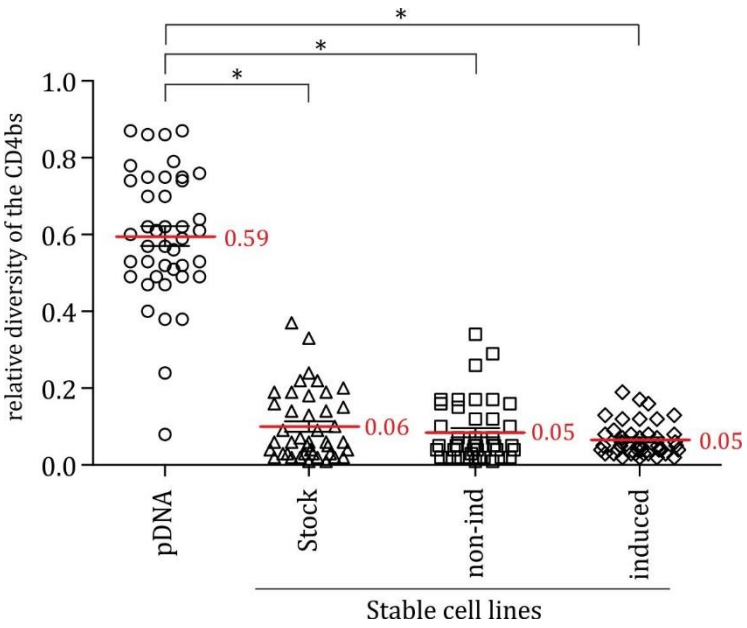


Figure 21 – Relative diversity of pDNA and the respective stable cell lines. The calculated relative diversity of pDNA and stock, non-induced and induced stable cell lines of the whole CD4bs are shown. Red lines and numbers denote the mean of the detected amino acids. Statistical significance ($p<0.0001$ for all samples) is indicated by asterisks and was calculated with Graphpad Prism using a Wilcoxon signed rank test.

As the observed reduction in diversity is not satisfactory, the steps at which the bias was introduced during stable cell line generation, as well as appropriate solutions to prevent suboptimal outcomes, should be identified.

4.3.4 Optimization of stable cell line generation

The stable cell line library provides the basis for the planned screening technology. Consequently, any limitations or defects as for example in the case of decreased variability could be detrimental during the selection, and thus compromise the final results. Hence, an optimization of the stable cell line generation was the most obvious step to improve screening results.

4.3.4.1 Reproducibility during stable cell line generation

Successful integration as in the case of stable cell line generation is a rare and random event. Therefore, it is reasonable to assume that the aforementioned bias could be the result of insufficient integration events in combination with inadequate statistical coverage of all amino acids.

To analyze if the bias is randomly introduced during stable cell line generation or the result of a natural growth advantage or disadvantage of certain Env variants, three different SCLs (N276, L277, N280) were produced in triplicates under identical conditions. The stable cell lines were selected due to their distinct shift in bias in the three stages of the overall screening procedure. Subsequently, the respective amino acid distribution in stock, non-induced and induced samples was determined via NGS as described above (see 4.3.3.2).

Even though an average variability of 19 amino acids was detected in most of the generated SCLs in all three stages, the frequencies of occurring amino acids varied considerably within all replicates (figure 22, S6). Once again, a striking bias towards certain variants could be determined which displayed a slight shift in stock, non-induced and induced samples, except for replicate N276-1 that exhibited barely any change in the amino acid distribution over time. For instance, in replicate 1, histidine was the dominantly encoded amino acid at the variable position, whereas in the other replicates histidine was present at low levels, while arginine and stop, or aspartic acid and phenylalanine were dominant in replicates 2 and 3, respectively. The results confirmed that despite identical conditions, the reproducibility of uniform stable cell lines was not possible with the standard procedure used for their generation, most likely due to the completely random integration. Thus, the likelihood of growth advantage or disadvantage of certain Env variants as cause for the emerging bias is rather low, suggesting insufficient statistical coverage instead.

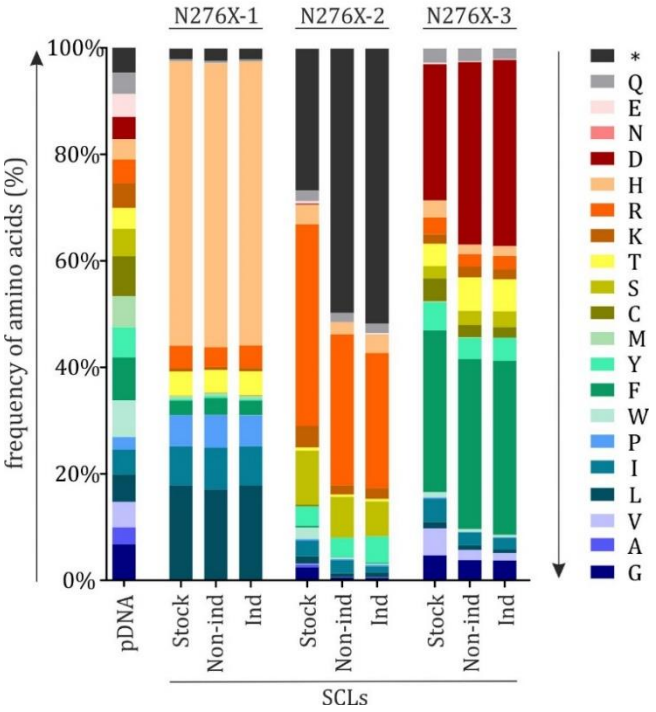


Figure 22 – Amino acid distribution of stable cell line N276X. Stock, non-induced and induced samples of SCL triplicates compared to the pDNA utilized during the generation of the stable cell lines. Heights of the stacked bars represent detected amino acids in percent.

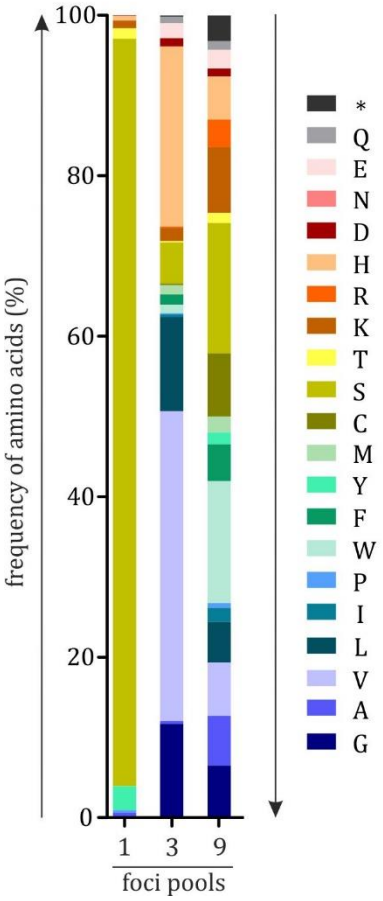
4.3.4.2 Impact of the number of integration events

To establish an adequate statistical coverage of all amino acids, it was presumed that the frequency of integrations must be increased significantly. In general, about 100-200 foci or colony forming units (cfu) could be detected after the completed antibiotic selection. However, the 200 foci did not necessarily represent 200 individual integration events. The actual number of integration events is likely to be far lower due to one cell division cycle that takes place in the time during transfection and begin of antibiotic selection (see 3.3.4). In this respect, it was important to determine the required number of foci to prevent occurrence of bias and to obtain an approximately even distribution of amino acids in a stable cell line.

An increased yield of foci was suggested to be achieved in two ways: i) upscaling of transfection and/or ii) generating and pooling multiple SCLs of the same position. In respect to upscaling process, volumes of all transfection reagents and DNA amounts were increased ~8-fold to adjust for a transfection from a 6-well flat bottom well (9 cm^2) to a T75 flask (75 cm^2). As cells were usually transferred from 6-well to T75 flask 48 hours post transfection according to the protocol, for this experiment the cells were split equally into eight T75 flask. In the second setting, multiple SCLs (1, 3 and 9) were generated based on the library position N276X. After the antibiotic selection procedure (25 days), the amounts of foci from one, three or nine T75 flasks were counted and pooled. The diversity was assessed by NGS as described above (see 4.3.3).

The results from the stable cell line generation demonstrated particularly striking fluctuations in the yield of foci from the individual 6-well transfections, that ranged from four to 305 counted colony forming units (figure 23 B). Differences in the number of foci from the upscaled transfection could also be detected (116-404), even though they were less substantial than from the 6-well transfections. When looking at the diversity of the individual stable cell lines, the expected reduced diversity (11 amino acids) in combination with strong bias in the amino acid composition was observed (figure 23 A+B). Strikingly, the analysis of pooled foci from three (total of 346 foci) and nine (total of 1154 foci) transfections revealed an improvement in the relative diversity. This was apparent especially in the pool from nine 6-well transfections which displayed a relative diversity of 0.31, thus confirming that an increased amount of foci could reduce the bias introduced during the generation of stable cell lines.

A



B

6-well transfection

N(replicates)	$N(foci)_{total}$	Mean	N(Aa)	D_{rel}
1	139	139	11	0.03
3	346	115.3 ± 70	20	0.13
9	1154	128.2 ± 108	20	0.31

T75 transfection

N(replicates)	$N(foci)_{total}$	Mean	N(Aa)	D_{rel}
1	1836	229.5 ± 109	20	NA

Figure 23 – (A) Amino acid distribution of SCL replicates. One, three and nine 6-well transfections were performed under identical conditions and the amino acid frequency of pooled generated stable cell lines was analyzed by NGS. Heights of the stacked bars represent detected amino acids in percent. **(B) Summary of results from the SCL generation.** Yield of manually counted foci ($N(foci)_{total}$) from the 6-well and T75 flask transfections, mean and standard deviation of counted foci from each tranfection (mean), amount of differing amino acids detected via NGS (N(Aa)), and calculated relative diversity (Drel) from pooled SCLs is shown. Unfortunately, no data could be acquired (NA) for the pool from the T75 transfection due to insufficient reads from the NGS run.

Upscaling of the transfection by a factor of eight resulted in higher foci yields, however the amount of 1836 detected colony forming units was lower than anticipated, which

ideally should be eight-fold higher. Unfortunately, no data from the T75-pool could be acquired due insufficient reads obtained from the NGS run.

The minimal amount of foci required for an ideal amino acid distribution could not be determined since the relative diversity of the pools neither had optimal amino acid distribution, nor approached the nearly ideal diversity of the pDNA samples that showed an average relative diversity of 0.57 (figure 19 B and table 6). Nevertheless, when comparing the relative diversities of pDNA samples and the pool comprising nine replicated stable cell lines (D_{rel} = 0.31), a distinct improvement in the diversity could be detected. Consequently, the hypothesis that insufficient integration events were responsible for the strong bias in the amino acid composition was confirmed. However, the bias could be reduced by increasing the numbers of foci.

4.3.4.3 Improvement of the transfection efficiency

Another possibility to increase the number of foci could be achieved by increasing the transfection efficiency. In this respect, the first approach was to examine the effectiveness of six different transfection reagents in an identical experimental setup as for the usual generation of stable cell lines, i.e. transfection in a 6-well with$5x10^5$ cells. As a preliminary step, optimal ratios of transfection reagent and DNA amounts were determined for the transfection of Flp-In™ T-Rex 293 cells (see section 3.3.2).

Analysis of the transfected cells did not reveal substantial differences for the different ratios of transfection reagent and DNA utilized, since mean fluorescence intensities and GFP positive cells mostly remained on the same level even at higher amounts of transfection reagents (figure 24). An exception could be seen for calcium phosphate where higher amounts of DNA applied in the transfection seemed to increase efficiency. In addition, a reduction of MFIs with increasing reagent volume could be noted for Lipofectamine 2000. Among the reagents, Lipofectamine 3000 demonstrated the best transfection efficiency, as cells displayed the highest mean fluorescence intensities, as well as the largest fraction of GFP positive cells (about 40%). Approximately 37% transfected cells were achieved with PEI and PEI Max and no difference could be seen between them. Lipofectamine 2000 and Fugene 6 exhibited mediocre results (~ 22%), whereas calcium phosphate displayed the worst transection efficiency. Considering the results, a 3:1 ratio of transfection reagent volume and DNA amount was utilized during the generation of stable cell lines in further experiments. In respect to calcium phosphate, 5 µg DNA was applied.

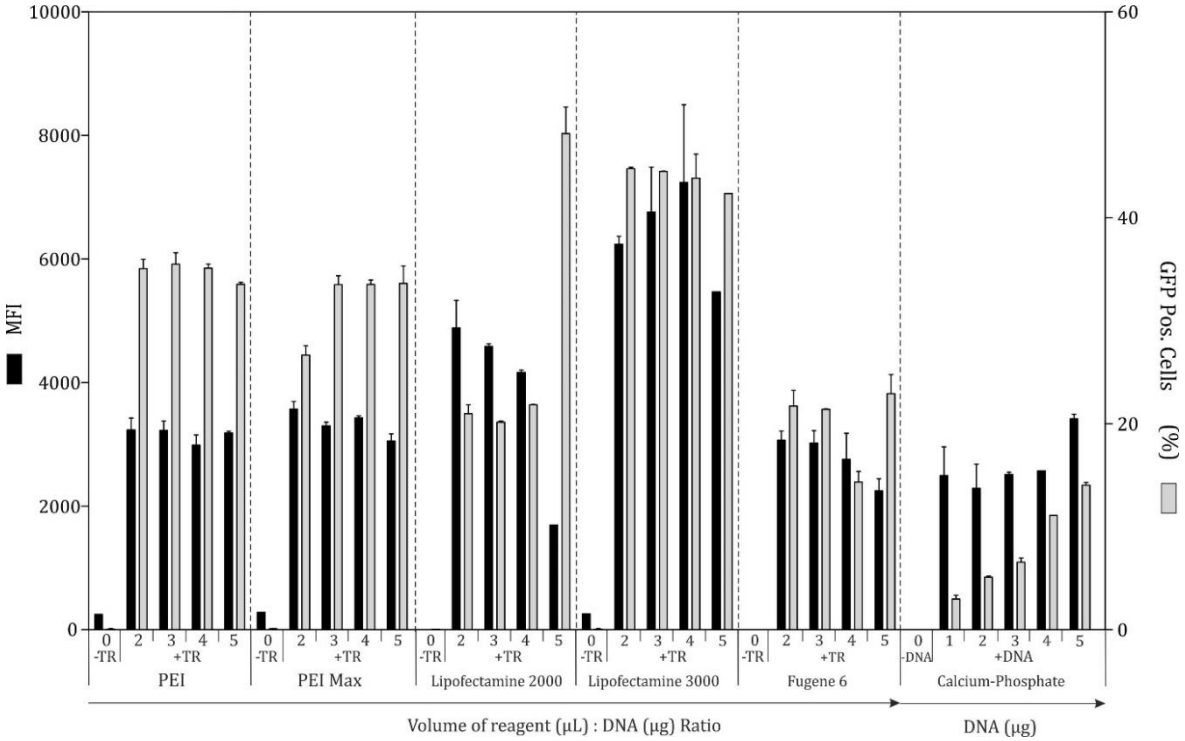


Figure 24 – Efficiency of different transfection reagents. Flp-In™ T-Rex 293 cells were transiently transfected under identical conditions utilizing six different transfection reagents (PEI, PEI Max, Lipofectamine 2000 & 3000, Fugene 6 and calcium phosphate). Different ratios of transfection reagent (+TR) were tested. For of calcium phosphate, the amount of DNA was adjusted while keeping the volumes of reagents constant. Negative controls (-TR) were included for every tested transfection reagent by replacing the reagent with water (or DNA as in the case of calcium phosphate transfection). 48 hours post transfection, mean fluorescence intensities (black bars) and the percentage of GFP positive cells (grey bars) were measured via flow cytometry.

Transfection reagent	N(replicates)	N(foci)	Mean
PEI	1	229	284 ± 79
	2	248	
	3	375	
PEI Max	1	156	156 ± 27
	2	129	
	3	182	
Lipofectamine 2000	1	521	473 ± 42
	2	449	
	3	448	
Lipofectamine 3000	1	44	112 ± 93
	2	75	
	3	218	
Fugene 6	1	301	245 ± 59
	2	252	
	3	183	
Calcium-Phosphate	1	355	251 ± 143
	2	311	
	3	88	

Table 7 – Yield of foci obtained from different transfection reagents. Triplicates of stable cell lines were generated with the six different transfection reagents PEI, PEI Max, Lipofectamine 2000 & 3000, Fugene 6 and calcium phosphate. After the antibiotic selection process, the amount of foci was counted manually.

Strikingly, findings from the preliminary determination of transfection efficiency did not correspond with the final results acquired after the selection process of stable cell lines (table 7). Contrary to the expectations, Lipofectamine 3000 led to the lowest foci yield although it demonstrated the best transfection efficiency before, whereas the previously mediocre results of Lipofectamine 2000 produced the highest numbers of foci (448-521). Interestingly, calcium phosphate also displayed high foci yields while the transfection efficiency had been below 20% in the preliminary experiment. However, since the efficiency was tested on transiently and not stably transfected cells, this could explain the striking differences. Since the limiting steps in the generation of stable cell lines involve not only transfection efficiency of cells but also the integration rate, the yield of foci could probably be further improved by adjusting transfection conditions according to resulting numbers of foci. Nevertheless, Lipofectamine 2000 should be considered as transfection reagent for the generation of stable cell lines that require an ideal amino acid composition since it leads to about two-fold higher foci yields than cells transfected with the commonly used PEI. A combination of upscaled transfection and Lipofectamine 2000 could lead to even higher diversity among the stable cell lines than observed so far.

4.4 High-throughput screening of a stable cell line library to identify improved HIV-1 antigen candidates

It is a generally accepted concept that the elicitation of broadly neutralizing antibodies could evoke protection against HIV-1 infection. This theory has been confirmed by various immunization studies in macaques demonstrating protection against infection after passive administration of bnAbs [119,168–171]. However, the search for immunogens able to induce an efficient humoral immune response still represents a major challenge. In addition to the multitude of evasion strategies embodied in the virus, a lack of innovative screening technologies able to identify beneficial vaccine candidates are among the reasons to the prevailing absence of a HIV-1 vaccine. A promising concept in the search for favorable immunogens might be a mammalian cell display and screening platform. The basic concept has been developed previously by Dr. Tim-Henrik Bruun as part of his PhD thesis [209,224,225]. The technique presumably allows the selection of Env variants with specific properties, such as higher affinity towards given bnAbs. In the course of this PhD thesis, this screening system should be applied to the significantly larger SeqPer library, thus enabling the isolation of Env variants with beneficial antibody affinity from a vast pool of mutants. Since the technology has not been tested on libraries in the size of the SeqPer, the selection process was performed initially on a defined section of the generated SCL library, the CD4 binding site, and the well-characterized human monoclonal bnAb, VRC01, that potently neutralized HIV via binding to the CD4 binding site on Env. Cells expressing envelopes with increased or decreased antibody

affinity were selected by flow cytometry-based cell sorting, identified via NGS and the detected variants were validated by flow cytometric analysis of equilibrium titrations.

4.4.1 Overview of the mammalian cell-display-based screening technology

The generated stable cell lines provided the basis for the mammalian cell display-based screening technology allowing the identification of favorable Env immunogens from the SeqPer library (figure 25, 1-2). The stable cell line library featured a major benefit: due to the sole FRT site in the Flp-In™ T-Rex 293 cells, only one Env variant can be integrated into the genome, therefore, establishing a linkage between geno- and phenotype regarding the proteins on the cell surface. After induction and following expression of the Env library (3), the screening procedure could be commenced. In this process, surface envelopes were stained with a screening antibody (4) and subsequently, variants displaying the highest and lowest affinities towards the antibody were selected by cell sorting (5). In this respect, variants with the highest or lowest antibody affinity in relation to the eGFP expression were enriched. Since a correlation between eGFP and Env expression was observed (according to data from [213]), eGFP was utilized as an indirect marker to normalize for Env expression. In order to identify the selected envelope variants, genomic DNA was recovered, amplified and characterized by Next Generation Sequencing (6-7).

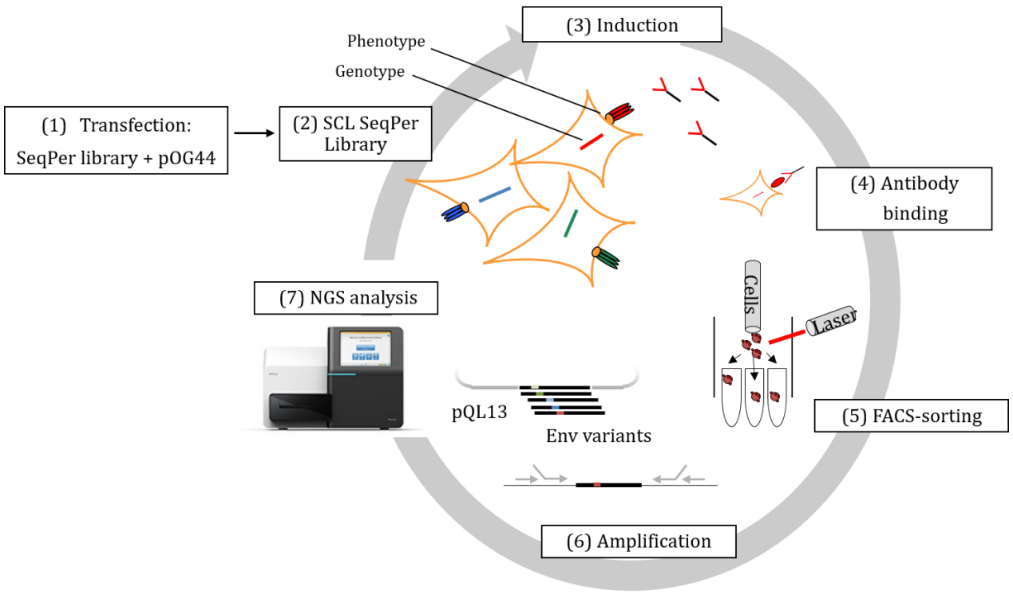


Figure 25 – Schematic workflow of the mammalian cell-display and screening platform. The technology consists of seven steps. Firstly, a stable cell line library (SCL library) is generated based on the SeqPer plasmid library by targeted transfection of Flp-In™ T-Rex 293 cells with a pQL13-Env sub-library and the helper plasmid pOG44 (1+2). Due to the single integration locus in the cells, only one Env variant

is inserted into the host genome, therefore, resulting to a linkage between pheno- and genotype. Hygromycin resistance serves as antibiotic selection marker for successful integration. After induction of Env expression by addition of doxycycline (3), surface proteins were stained with the screening antibody VRC01 (4). Subsequently, the cells underwent flow cytometry-based sorting to select Env variants with the desired phenotype (5). Genomic DNA of sorted cells and input controls was recovered for the ensuing amplification by PCR (6). Env variants were then analyzed and identified by Next Generation Sequencing (7). Figure was freely adapted from [224].

4.4.2 Purification and validation of the bnAb VRC01

The CD4bs-specific broadly neutralizing antibody VRC01, utilized in the screening of the SeqPer library, was generated by transient transfection of the variable heavy and light chains into Expi293F™ suspension cells (see 3.3.3). Purification of 80 mL supernatant of transfected cells via a Protein A-Sepharose column (see 3.4.1) resulted in a total antibody yield of approximately 19 mg. Proper assembly of the antibody was assessed by SDS-PAGE (see 3.4.3). The expected fragment sizes for heavy and light chain could be detected at approximately 50 and 25 kDa, respectively (figure 26 A). Analysis of the VRC01 affinity for a soluble BG505 gp140 envelope protein by ELISA demonstrated a slightly improved binding for Env compared to a VRC01 reference batch (figure 26 B) obtained from the NIH AIDS Reagents Programme (#12033).

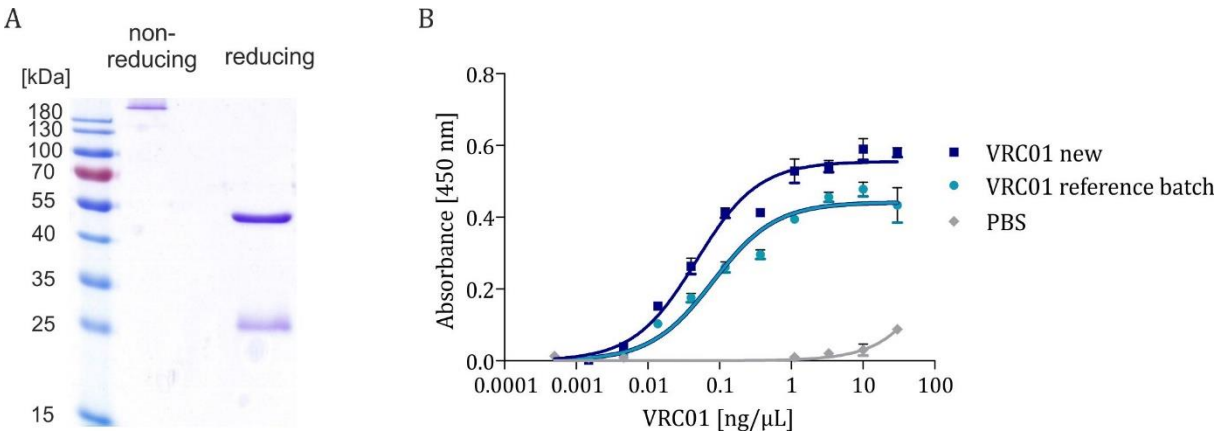


Figure 26 – (A) SDS-PAGE analysis of the monoclonal bnAb VRC01 under non-reducing and reducing conditions. 2 µg VRC01 previously purified via Sepharose-A affinity chromatography was utilized in the electrophoresis. The left lane represents non-reduced, whole VRC01 (~ 150 kDa). Bands at 50 and 25 kDa indicate the variable heavy and light chains of the antibody, respectively (reducing conditions, right lane). **(B) Envelope ELISA of purified VRC01.** Binding affinity of newly generated antibody toward a soluble HIV-1 BG505 gp140 envelope protein was compared to a VRC01 reference batch. PBS served as negative control to determine unspecific binding of the antibody.

4.4.3 Identification of Env variants with increased or decreased binding affinity for the bnAb VRC01

The screening for Env variants displaying increased or decreased binding affinity toward the bnAb VRC01 was performed on batches of induced stable cell lines that comprises the variants of the residues forming the CD4 binding site. To determine potential limitations of the mammalian screening technology, the four different cell batches with increasing numbers of pools-of-position belonging to the CD4bs were generated: i) one pool-of-position (T278X), ii) five positions (first part of loop D), iii) 15 positions (whole loop D + first part of the CD4 binding loop) and iv) 46 positions (whole CD4bs). Notably, the T278X position was part of each batch and served as a positive control since increased affinity of VRC01 towards the variants T278A and T278H (HXB2 numbering) has already been confirmed independently by Veronika Grassmann [213] and Christina Schmalzl [226] in similar experiments.

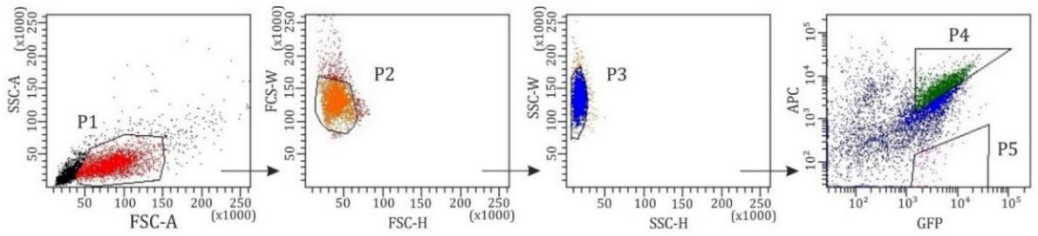
The induced stable cell line batches (i-iv) were sorted individually on the same day with identical settings. The gating strategy ensured that only singlets and living cells were selected (figure 27 A). In addition, triangular shaped gates were chosen to sort cells with the highest or lowest VRC01 signal in relation to the GFP signal, thus selecting cells with the presumably highest or lowest antibody affinity in relation to GFP expression (figure 27 A). Selected variants were subsequently identified by Next Generation Sequencing following the procedures described above (see 3.2). Through comparison of the sorted samples with input material that was drawn prior to the screening procedure, a fold enrichment or depletion of specific Env variants could be calculated [227] (figure 27 B, S4-6). Mutations demonstrating increased (high affinity gate P4) or decreased affinity (low affinity gate P5) toward VRC01 are referred to as 'gain-of-binding' (GoB) or 'loss-of-binding' (LoB) variants, respectively.

The sort was performed with $30x10^6$ cells of which 30.000 cells were sorted into the high affinity gate. Sorting of the low affinity gate was stopped after the required number of cells in the high affinity gate was collected (usually $2\text{-}3x10^5$ cells). When analyzing the results from all sorts, a total of 79 individual GoB and 73 LoB candidates were detected which exhibited a >1.5-$\log_2$-fold enrichment in the high or low affinity gates (figure 27 B, S4-6). A wide array of the variants that displayed enrichment in a specific gate originated from T278, highlighting it as a particularly sensitive and relevant position for VRC01 recognition. Notably, mutations T278A and T278H could be detected as GoBs in most of the sorts.

Positions 276 and 278 were of specific interest in the analysis, since they are part of the N276-X277-T278 motif, a crucial N-glycosylation site at the CD4 binding site. This glycan was previously reported to prevent interaction between germline VRC01 and Env due to a steric clash. Thus, a removal of this glycan by substitution of either N276 or T278 was confirmed to improve accessibility of the CD4 binding site for VRC01 [191]. Strikingly, the sorts resulted in many substitutions at position 278 which displayed increased

affinity for mature VRC01, whereas no mutation at position 276 exhibited an influence on antibody affinity. Similar results were obtained by Veronika Grassmann which indicated that the beneficial effect of glycan removal at N276 could supposedly be abolished by a negative structural effect of a mutation on the CD4bs, as mutation showed significant reduction of sCD4 binding [213].

A



B

N(Aa position)	Fold enrichment in high affinity gate P4 [a]					
	1	5				
Position	278	274	275	276	277	278
Aa	**T**	**S**	**E**	**N**	**L**	**T**
G	-1,01	-3,61	0,03	-3,65	0,82	0,31
A	1,58	0,09	0,24	-2,12	n.d.	2,77
V	-1,35	n.d.	-2,04	-3,11	-2,58	0,41
L	-2,28	-2,74	0,73	-2,49	WT	n.d.
I	-1,60	-0,49	0,04	-2,67	0,31	1,22
P	0,53	1,18	0,13	0,24	-1,77	0,19
W	-0,29	n.d.	n.d.	n.d.	n.d.	n.d.
F	-3,46	-2,78	n.d.	-2,37	0,15	-1,01
Y	-0,22	-3,74	-2,84	-0,08	n.d.	n.d.
M	n.d.	0,22	0,10	1,61	-2,14	0,01
C	-4,03	n.d.	-2,72	0,25	0,09	0,08
S	0,14	WT	-2,82	n.d.	n.d.	3,08
T	n.d.	-3,01	n.d.	-2,83	n.d.	WT
K	-0,83	-1,56	0,32	-0,81	n.d.	n.d.
R	-0,18	-0,18	0,35	-2,18	-5,03	3,11
H	2,18	-3,37	0,21	0,12	-2,11	3,88
D	-0,84	0,12	0,77	-3,15	0,09	2,10
N	-0,51	-1,59	-1,21	WT	n.d.	-2,23
E	-5,75	0,09	WT	0,31	n.d.	-0,13
Q	0,82	0,12	0,08	0,26	1,31	3,88

log$_2$-fold enrichment

depletion

C

N(pools of position)	GoB	LoB
1	T278A	T278E
	T278H	T278K
		T278N
5	T278Q	E275F
	T278R	E275T
	T278S	L277K
15	E275K	L277Q
	G366D	V281E
	G367C	D368E
46	E275N	A365D
	T373I	P437N
	T373N	E464I
	H374Y	

[a] calculations were performed according to the formula from [204]:

$$\textit{fold enrichment} = \left(\frac{n\,(G)\,\textit{detected}}{n\,(G)\,\textit{reference}}\right) / \left(\frac{n\,(\textit{non-G})\,\textit{detected}}{n\,(\textit{non-G})\,\textit{reference}}\right)$$ (on the example of glycine enrichment)

Figure 27 – (A) Gating strategy of the FACS-based screening procedure. A representative sorting experiment is displayed (30.000 events of 30x10^6 sorted cells shown). First, living, single cells were gated (P1-3). The cells separated into populations of uninduced (low GFP/low APC) and induced cells displaying high GFP and either high or low APC signals depending on the affinity of VRC01 to the Env variants. Induced cells were gated for highest (P4) or lowest (P5) APC signals (VRC01) in relation to GFP, thus resulting in triangular-shaped gates. **(B) Enrichment rates from the sorting of one and five pools of position (gate P4).** Enrichment is represented as log$_2$-fold change in the number of the respective NGS reads for each variant compared to the input sample drawn prior to the sorting. The extent of enrichment (green) or depletion (red) is indicated by the intensity of the respective color. Columns denote the wild type amino acid and its position in the library, whereas rows denote the enriched amino acids from the respective pool-of-position. Wild type and variants that could not be detected in the sorting were referred to as WT or n.d. (not detected), respectively. **(C) Summary of variants demonstrating the highest fold**

change increase or decrease in affinity towards VRC01. With exception of the screening of one pool-of-position, three mutations (four from the GoB sort iv) from each sorting were selected for further characterization. HXB2 numbering was applied.

Simultaneously, specifically the variants T278A and T278H were confirmed as gain of bindings during similar screenings from another available library (alanine scanning library) that were performed by Veronika Grassmann [213] and Christina Schmalzl [226].

Overall, unfortunately, the results proved to be ambiguous since only five GoBs or LoBs were identical in the individual sorts from the different cell batches (i.e. GoBs T278Q/T278R from sort i, ii and iv, but not in iii) and none of them could be found in all sorts. Furthermore, certain variants were detected as 'gain'- as well as 'loss-of-binding' variants, as in the case of T278A (sorts i and ii) and T278H (sorts i, iii and iv) (figure 27, S4-6). The inconclusive results indicated potential problems either in the gating strategy or limitations of the screening technology (see 5.4).

To check the screening results, three variants with the highest fold change from every sort (i-iv) were selected for further characterization (figure 27 C)

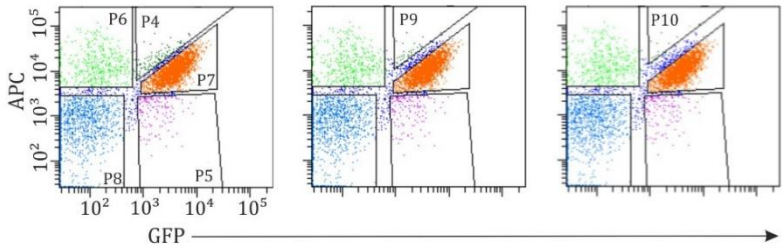
4.4.4 Optimization of the gating strategy

One possible cause for the inconsistent sorting results could be an unfavorable gating strategy. To obtain more information from cell sorting, a new gating strategy was applied to address several points of interest pertaining to: i) analysis of variants enriched in all detected populations, ii) reproducibility of results originating from the same gate and iii) stringency of placed gates. For this purpose, all populations of the induced stable cell line T278X were arranged into specifically set gates (figure 28 A) and analyzed via NGS (figure 28 B). Reproducibility was addressed by sorting the cell sample from one gate in triplicates (P4), whereas stringency was attained by shifting gate P4 two times upwards (P9, P10) to obtain variants with the highest measurable binding signal for VRC01 in relation to GFP. With a starting population of $90x10^6$ cell, in the following, $3x10^4$ cells were acquired from each gate. As there were insufficient cells at the end of the sorting, $15x10^3$ cells from the reproducibility gate (P4) and 5.000 cells from gates P9 and P10 were collected.

Notably, even though cells separated into three predominant, easily distinguishable populations during the sorting (i.e. cells with high APC/high GFP (gates P4/P5/P7/P9/P10), high APC/low GFP (gate P6) and low APC/low GFP signal (gate P8)), predominantly the two variants T278A and T278H were unexpectedly enriched

irrespective of the placement of the gates. The results could be the cause of too closely situated gates in combination with low sensitivity of the sorting device. In respect to reproducibility, the sorting displayed very similar enrichment rates among the triplicates. More stringent gating seemingly increased the sensitivity for detecting specific variants which was noticable in the lower amounts of identified variants and the slightly elevated enrichment rates of T278A and T278H. Consequently, the increased selectivity of gates P9 and P10 highlighs the necessity of a more stringent gating strategy.

A



B

Fold change enrichment in respective gates

	Whole population (i)					Reproducibility (ii)			Stringency (iii)	
Aa	**P4**	**P5**	**P6**	**P7**	**P8**	**P4-1**	**P4-2**	**P4-3**	**P9**	**P10**
G	-0,73	2,04	0,36	0,26	-0,20	-1,93	-1,33	-0,92	n.d.	0,26
A	3,18	2,96	4,08	3,55	6,68	4,92	4,85	3,98	7,04	7,47
V	-1,07	1,79	-1,09	-0,05	-0,07	-2,42	-0,50	n.d.	n.d.	n.d.
L	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
I	0,57	0,43	-1,06	0,76	-0,25	-0,22	-4,40	-0,24	-2,64	-0,23
P	-0,07	0,18	-0,05	-0,07	0,97	0,04	0,31	-0,32	-0,11	0,07
W	-1,17	0,12	n.d.	-0,75	-2,00	n.d.	-2,67	-1,26	n.d.	n.d.
F	1,17	-1,22	-0,09	0,32	1,16	n.d.	n.d.	n.d.	n.d.	n.d.
Y	0,15	3,75	-2,92	0,11	-1,49	n.d.	-0,79	0,36	-2,06	n.d.
M	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
C	-6,97	3,57	-0,11	-1,25	0,23	n.d.	n.d.	n.d.	n.d.	-3,08
S	-0,01	-2,49	-0,01	-0,02	-0,07	0,00	-0,01	0,00	-0,08	-0,13
T	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT
K	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
R	-0,18	-0,08	-0,21	0,17	-1,21	-1,24	-0,27	-0,35	-1,66	-0,42
H	4,65	4,71	4,45	5,37	5,57	5,10	6,08	5,91	6,57	6,49
D	0,67	0,43	-1,82	1,19	-1,65	-5,04	-0,44	-3,48	n.d.	-4,46
N	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
E	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
Q	0,11	-1,63	0,01	0,35	-1,82	-1,51	-0,50	-1,01	-1,32	-2,84

$\log_2$-fold enrichment

depletion

Figure 28 – (A) Different gating strategies for the FACS-based screening procedure. Living, single cells were gated as before (figure 25 A). Gates were chosen to address the whole population (left), reproducibility (P4 was sorted in triplicates) and stringency (middle +right). **(B) Enrichment rates from the sorting of pool-of-position T276X according to the new gating strategy.** Enrichment is represented as $\log_2$-fold change in binding compared to the input sample withdrawn prior to the sorting. Extent of enrichment (green) or depletion (red) is indicated by the intensity of the respective color. Wild type and variants that could not be detected in the sorting were referred to as WT or n.d. (not detected),

respectively. Notably, mutations T278A & H were predominantly detected from every sort irrespective of the gate placement (indicated by red boxes).

4.4.5 Validation of the detected GoB and LoB variants

Three variants of every screening having the highest fold change increase or decrease in affinity towards VRC01 were selected to verify if the screening procedure and if the results yielded valid results. For this purpose, single mutation variants as well as the respective stable cell lines (figure 27 C) were generated and their affinity profiles for VRC01 were analyzed by flow cytometric equilibrium titration. This was achieved by staining the transfected cells or induced stable cell lines with a dilution series of VRC01-Alexa647 (see 3.3.5).

Titration experiments were intended to determine the dissociation constant (K_D) of the VRC01 antibody and an Env protein. However, this was not possible in this case since none of the binding curves reached saturation levels even at very high antibody concentrations (figure S8). However, a slight indication toward saturation could be recognized for the binding curves of stable cell lines. Therefore, the area under the curve was calculated and used for the analysis

In general, the titration results between the transiently transfected and stable cell line variants did not differ considerably, with the exception that stable cell lines always displayed lower mean fluorescence intensities than transiently transfected cells. Notwithstanding, this did not influence the outcome of the titrations.

The overall picture of results for transiently transfected LoB-Env-variants and their stable cell line counterparts proved to be almost identical, with six variants displaying a loss in affinity towards VRC01 (figure 29 C+D). Notably, variants E275F, E275T and D368E demonstrated statistical significance (figure 29 C+D, table S1) in both, transiently transfected cells and stable cell lines, whereas L277Q and A365D displayed significance only in transiently transfected variants. Although detected as 'loss of bindings' during the selection procedure, a slight if not significant increase in affinity for VRC01 was visible in four variants, with three of them involving position T278. These results were in accordance with the findings of Veronika Grassmann, that confirmed a generally increased affinity for VRC01 for most amino acid substitutions at position T278 [213]. Additionally, analysis of the GoB-variants demonstrated increased binding for five variants at the same position, even though they were apparent only in the stable cell lines and did not exhibit statistical significance (figure 29 A+B, table S2). Furthermore, transiently transfected GoB-Env-variants did not display a distinct increase of binding, but rather a majorly wild type-like affinity. Assumedly, the results might be attributed to the high standard deviation of the WT. Many variants unexpectedly even showed a

severe loss of binding for VRC01 that could be reproduced in the stable cell lines. In this respect, the GoB-variant G367C (p = 0.0004) was highly statistically significant.

In addition, there seemed to be no connection between the 'number of amino acid positions' (one, five, 15 or 46 positions) applied in the sorts and the selected GoB- or LoB-variants which could have explained the inconsistencies in the selection procedure.

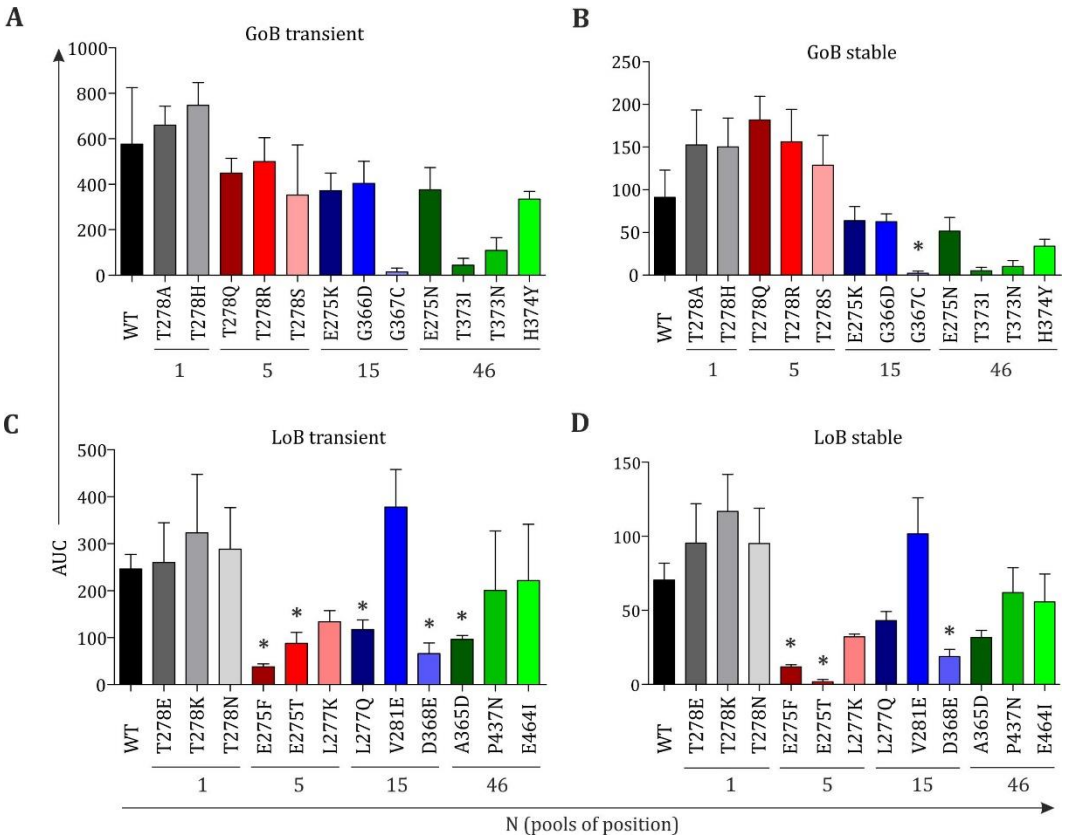


Figure 29 – FACS equilibrium titration of GoB- and LoB-variants with the bnAb VRC01. Three variants with the highest fold change increase in affinity from one, five, 15 and 46 pool-of-position (N) sorts were analyzed. Areas under the curve (AUC) of **(A+C)** transiently transfected variants and their respective **(B+D)** induced stable cell lines were analyzed and compared with the 16055 gp145 wild type. Statistical significance was determined with Excel using a two-sided, homoscedastic t-test and corrected according to Bonferroni to adjust for multiple testing. Variants that exhibited p-values <0.0042 were considered significant and are indicated by asterisks.

5 Discussion

The focus of this PhD thesis was the characterization and optimization of a sequential Env permutation library. In a first step, the general quality of the library was analyzed on the level of pDNA (see 4.3.1 and 4.3.3). Plasmid DNA was then utilized to generate the respective stable cell lines that were characterized regarding their diversity in Env variants and their amino acid distribution. As the stable cell lines displayed reduced amounts of Env variants which featured also a strongly biased amino acid composition, various optimization approaches for the generation of stable cell lines were tested (see 4.3.4). A novel flow cytometry-based technology was devised to screen the library for Env variants with increased or decreased affinity for bnAbs (see 4.4.3). Due to the size of the library, only the CD4 binding site was used in the selection procedure as a proof of concept and to assess efficiency and possible limitations of the system. Subsequently, selected variants that showed a substantial impact on the affinity of the bnAb VRC01 were validated by flow cytometric analysis of equilibrium titrations (see 4.4.5).

5.1 Evaluation of the SeqPer library

5.1.1 Advantages of the sequential permutation library

The design of the SeqPer library provides several benefits. One of the advantages was represented in the utilization of the C-clade strain 16055 as it addresses the globally most prevalent HIV-1 subtype. A special feature of the library was the sequential substitution of every residue in the external part of Env with all 20 aas, resulting in 658 sublibraries and approximately 10.000-13.000 different Env variants in total. The considerable size of the library renders the individual characterization rather difficult, thus, a high-throughput screening system is required. In addition, the many different Env variants create a vast pool from which mutant proteins with desired properties can be selected. In respect to this thesis, the Env library should be utilized to identify potential vaccine candidates with specific antigenic profiles.

Due to the considerable size of the library, comprehensive quality controls have been conducted to asses whether the intended diversity was achieved for the plasmid DNAs obtained from the gene synthesis provider, and whether this diversity is maintained over all steps of stable cell line generation and screening (see figure 25).

5.1.2 Quality of the pDNA and the possible implications

In-depth analysis of the pDNA of the SeqPer library revealed deletions that occurred majorly in the Env region and to a lesser extent in the vector region as the main problem (section 4.3.1). The detected errors affected approximately 48.3% of the library and occurred most likely during the DNA manufacturing procedure. Although many of the library positions displayed errors, these errors did not afflict the whole sublibrary as the plasmid mixture contatined predominantly the correct plasmid and only 6-18% actual errors were detected within the affected positions (see 4.3.1.3). Although unfortunate, these findings were not too surprising as gene synthesis and the resulting products unavoidably contain errors such as deletions, insertions, or base substitutions due to mistakes in chemical oligonucleotide synthesis and subsequent enzymatic gene assembly procedures [228]. However, as only a minor part was affected, the library was deemed acceptable for further experiments. Nonetheless, one must be aware that flaws in the quality of the library delineated the first line of errors and could have created implications for the subsequent experiments. It cannot be ruled out that the generation of stable cell lines was affected by deletions in the Env regions since incomplete Env sequences might have been integrated into the host genome, thus leading to either lacking protein expression or expression of misfolded, non-functional Env variants on the cell surface. However, it is unlikely that aberrational and normal variants competed for integration as erroneous plasmids represented only a small percentage. Hence, there should be merely a minor impact for stable cell line generation.

Strikingly, the sublibraries demonstrated a high diversity with an average of 19 decoded amino acids on the level of pDNA, as well as near optimal amino acid distribution. Unfortunately, several sublibraries did not display the ideal composition of 20 amino acids. A possible solution might be to generate the positions with the lowest diversity anew or if only few variants are missing in the pools, single variants could be generated by cloning and supplemented to the respective pools-of-position. In respect to the stable cell lines, however, the latter solution might be inefficient, as too many Env variants are missing within the pools. Generating new stable cell lines might be less time- and cost-consuming.

5.1.3 Impact of stable cell line quality on cell-display-based screening

The genomic integration of the library vector pQL13 at a single, distinct FRT site, allowed the generation of a stable cell line library with linked genotype and phenotype of the respective *env* gene and the corresponding protein per cell. This library was utilized in the screening for improved Env immunogens. Unfortunately, a considerable reduction in the amino acid diversity on average by 38% alongside a strongly distorted amino acid

composition was detected in all analyzed stable cell lines (4.3.3.2). In particular, the CD4 binding loop was the most affected regarding the loss of diversity (figure 20 B). The reason for that could be due to daily experimental fluctuations, as usually batches of 15 to 20 stable cell lines were generated simultaneously, which might explain why 11 positions in a row demonstrated such a low diversity. To improve the diversity of the CD4 binding loop, the respective stable cell lines should be created anew for further experiments.

As the SCL library represented the basis for cell display and cell sorting-based selection for improved immunogens, it can be reasonably assumed that any obtained data reflected the stable cell line quality. Reduced variability signified a smaller pool for the selection procedure and thus, not the whole spectrum of possible Env candidates was available during the screening. In addition, it was highly likely that the strong bias in the amino acid composition would result in inaccurate enrichment of immunogens favoring amino acids which were strongly overrepresented. In accordance with this hypothesis, the data obtained from the cell sorting and flow cytometry experiments revealed several discrepancies. This was shown specifically in the enrichment of the variants T278A and T278H which were detected as GoB-variants, as well as LoB-variants in some of the sorting experiments (see 4.4.3). Also, as every sorting experiment enriched different variants, no consensus between the sorted variants was found. T278A and T278H represented an exception since they were enriched in almost all sorts (see 4.4.3, S5-S6). In addition, although several Env candidates with high or low affinity for VRC01 were identified in the sorting experiments, from the 24 tested variants (12 GoB- and 12 LoB-variants with the highest enrichment factor in the high or low affinity gate) no GoB- and five LoB-variants (E275F, E275T, L277Q, D368E and A365D) could be truly validated by FACS equilibrium titrations (see 4.4.5). Although unknown limitations of the screening technology cannot be ruled out as possible reasons for the outcome of the screening, it stands to reason that the quality of stable cell lines was one of the main factors influencing the experiments. Thus, optimizing the generation of stable cell lines is imperative before accurate data can be obtained.

5.2 Improvement of stable cell line generation

5.2.1 Identification of factors influencing SCL generation

Whereas transient transfection is beneficial for quick analysis of genes and protein production on smaller scales, stable transfection ensures long-term, reproducible and defined gene expression. As described above, stable transfection of multi-variant sublibraries were utilized to create cell lines expressing a single Env protein on the surface of mammalian cells. Analysis of stable cell lines demonstrated a loss in diversity

of amino acids, as well as an excessive shift in the amino acid distribution which assumably occurred due to insufficient integration events. However, the generation of stable cell lines encompasses many steps as the plasmid with the gene of interest needs to be introduced first into the cell, then into the nucleus and finally it needs to be integrated into the host's genome. Accordingly, there are many possible sources that could lead to such a bias as described above (see 4.3.4).

Since stable integration of foreign DNA into the genome is a relatively rare event that is aggravated by possible inactivation via epigenetic mechanisms [229], it was hypothesized that it represented the most influencing and limiting factor during cell line generation. Unfortunately, tracking of integration events proved to be a difficult endeavor, as evidence of successful insertion was provided only by the amount of foci or colony forming units after the selection process. Furthermore, the resulting amount of foci can be misleading because not every colony necessarily represented a different Env variant. In addition, it was assumed that the diversity of the resulting foci is further reduced during the stable cell line generation. As cells remain in the 6-well for approximately 48 hours post transfection before they are transferred into a T75 flask, one cell division cycle occurs in this period. This fact was revealed by manually counting the amounts of cells 24 and 48 hours after transfection for samples that were specifically prepared for that purpose (amounts doubled during that time from $5x10^5$ to $1x10^6$ cells). Therefore, the diversity of the resulting stable cell lines might be reduced by a factor of two.

Eventually, it could be confirmed that the bias was introduced during the generation of stable cell lines due to inefficient and insufficient integration of Env (4.3.4). Assumably, a considerable increase of foci was hypothesized to reduce the occurrence of bias. This supposition was confirmed by the incrementally improved relative diversty of 0.03, 0.13 and 0.31 after pooling the cells from one, three and nine transfections, respectively. In particular, the results from nine transfections indicated that the integration events from one transfection were not sufficient to generate a satisfactory amino acid diversity. Unfortunately, an ideal aa composition was not achieved with 1154 foci (total number of foci from nine transfections) (see 4.3.4.2) and the total number of foci to ensure optimal distribution could not be determined. In this respect, only a hypothesis can be made according to basic mathematical calculations. If 1154 foci generate a relative diversity of 0.31, then it might be reasonable to assume that approximately 4000 foci might lead to the ideal relative diversity of 1.0 which represents an even amino acid distribution. However, standard transfection protocols and utilization of alternative reagents failed to create high enough foci yields from a single transfection, although several transfection reagents such as Lipofectamine 2000 generated two-fold higher numbers of foci. In this respect, only upscaling of the transfection provided a slight improvement in the yield (1836 foci from an eight-fold upscaled transfection). If the mathematical connection mentioned above can be applied here, two T75-transfections might be sufficient to generate the assumedly 4000 required foci. Accordingly, in order to further reduce the bias, a combination of upscaled transfection and optimal reagents

should be employed during generation of such stable cell lines. Further suggestions for optimization are discussed in section 5.2.2.

There are also many additional factors that could influence transfection efficiency which require careful consideration. These include: i) quality of transfected DNA, ii) ratios between transfection reagent and DNA, iii) viability of cells and iv) variations in the transfection efficiency between different Env variants. It cannot be ruled out that additional unknown determinants are involved.

5.2.2 Possible optimization approaches

The initial step of stable cell line generation involves the transfer of library constructs into the cells. Thus, it stands to reason that improvement of the transfection conditions (i.e. optimal ratio of reagent and DNA, ideal DNA quality and cell viability, etc.) and efficiency might increase the chance of integration events. Choosing the right transfection reagent for the selected cell type is as important as the method of transfection to minimize unnecessary damaging of cells. Lipofectamine reagents are widely accepted as 'gold-standard' for the safe and gentle delivery of DNA into cells due to their high transfection efficiency across a broad range of cell lines [230]. The increased yield of foci that was obtained from transfections performed with Liopfetamine2000 corroborated this notion. Therefore, transfection protocols should be based on Lipofectamine-mediated transfection.

The advantageous effect of Lipofectamine can be supported by synchronization of the cell cycle phase as it has been repeatedly reported that gene transfer into cells is dependent on the cell cycle [231–233]. Cells arrested in their S- or G2-phase by treatment with mimosine, aphidicolin or thymidine demonstrated an at least three-fold increase in the percentage of transfected cells compared to untreated cells [231].

Additionally, it has been reported that nuclear targeting and entry of plasmid DNA can be enhanced, for instance, by using a 366 bp sequence of DNA containing the SV40 origin of replication and early promoter known to bind to a number of general transcription factors. This sequence was reported to favorably promote nuclear import [55,234]. Thus, implementation of the sequence into the library vector could be a promising approach to facilitate transfer of plasmid DNA into the nucleus of cells irrespective of their cell cycle phase.

Tracking and active selection of successful integration events could prove very helpful in optimizing the generation of stable cell lines. In this respect, the availability of eGFP in the described constructs might be a reliable reporter. Cells with successful insertion can be directly identified and isolated by FACS cell sorting according to GFP

fluorescence [229]. However, this approach relies on a sterile environment during and after the cell sorting, as the selected cells would be further cultivated.

Since the number of foci is assumedly inaccurate to quantify the integration events (see 5.2.1), insertion of reporter plasmids carrying a distinct index (sufficiently long NNNN-sequence) might be a possible solution. Due to the fact that every cell integrates a different sequence, the total number of integrations that took place during stable cell line generation can later be determined by NGS analysis.

It also might be worth considering creating single mutation stable cell lines and pool them equally to create a defined and controlled amino acid composition. However, this is not applicable for a library of the size of the SeqPer as this approach is time-and cost-intensive. Although it can be utilized for a limited amount of interesting positions as in the case of the Loop D of Env where the major contact sites for VRC01 reside.

5.3 Adaptation of NGS sample preparation for library applications

Due to the considerable size of the SeqPer library, gathering of important data by Next Generation Sequencing became indispensable, as it allowed the simultaneous analysis of thousands of different sequences. In this PhD thesis, the Illumina MiSeq NGS sample preparation was successfully adapted for applications regarding genomic libraries.

Generally, the core steps in preparing DNA for NGS analysis are: i) fragmentation and/or sizing of target sequences to a desired length, ii) attachment of oligonucleotide adapters 5' and 3' of fragments and iii) quantitation of the final library product for sequencing [235].

Although there are several approaches available to fragment DNA sequences, such as physical (i.e. sonication, hydrodynamic shearing), chemical (heat digestion with divalent metal cation), and enzymatic (non-specific nuclease, transposase, DNaseI, restriction endonuclease) methods, the latter in the form of amplification of DNA sequences by PCR was favored in the course of this PhD thesis. The resulting library products had a homogeneous and defined fragment size of approximately 300 bps. This was beneficial, since the *env* gene comprising approximately 2.1 kbs cannot be analyzed in its whole length with the Illumina technology. Therefore, a smaller fragment was chosen as a model that encompassed 300 bps which could be easily obtained by PCR amplification and which was eligible for the commonly used MiSeq Reagent Kit v2 (300 cycles). In this way, the whole *env* gene could be covered by creating eight overlappying 300 bp long fragments via PCR amplification. Simultaneously, oligonucleotide adapters as well as indices could be attached during the amplification steps, therefore bypassing expensive adapter ligation kits. Notably, adapter sequences were chosen to provide a well-balanced AT- and GC- content.

Although purification of the PCR samples with magnetic beads (see 3.2.3.2) showed initially good results, several NGS runs were aborted by the MiSeq, presumably, due to overloading. On closer inspection of the bioanalyzer data after magnetic beads purification, the results demonstrated an accumulation of adapter or primer dimers at the length of approximately 130 bps (figure 17 A). Since magnetic beads fail to remove DNA fragments with >100 bps, it stands to reason that the removal of dimers by magnetic bead purification seemed inefficient. In addition, the dimers might be able to bind to the NGS flow cell. As a consequence, the flow cell capacity is reduced which could have resulted in an overloading of the NGS chip. To prevent this outcome, gel purification after the second PCR to remove adapter dimers (see 3.2.3.1) proved to be successful (figure 17 B), since the problem of overloading did not occur in the following NGS experiments.

As amplification can be susceptible to a bias resulting from the error-prone nature of polymerases, careful selection of enzymes for PCR, thermocycling conditions as well as the proper amount of starting material is necessary. A systematic analysis of error rates from PCR-related sources obtained under various conditions demonstrated that 22 cycles were best in terms of yield and the occurring error rate (see 4.3.2). Although the amount of DNA has to be chosen according to availability, however, not less than 1 pg is advisable regarding genomic DNA, as an increase in errors was associated with lower quantities of input genetic material.

Moreover, an accurate library quantification is required to prevent over- and underloading of the NGS flowcell. In this respect, the Kapa Library Quantification Kit was favored for qPCR due to the provided highly stable standards and the polymerase that was specifically engineered to amplify diverse DNA fragments with similar efficiencies°[236].

A total of 9-10 pM sample was usually loaded on the flow cell. Higher amounts lead to overloading and to abortion of the NGS run, whereas lower amounts resulted in decreased generation of clusters as well as reads.

In general, a spike-in of 5-10% of a PhiX library is performed into the prepared library samples to provide a quality control for cluster generation. However, this amount is only applicable for samples with an evenly balanced AT and GC content. This was not the case for the SeqPer library as merely one position of the whole *env* gene was substituted in each sublibrary. Thus, the rest of env sequence remained constant and resulted in an unbalanced base composition of NGS samples. Unfortunately, the Illumina system has been reported to be sensitive regarding extreme base compositions (i.e. GC-poor or GC-rich sequences), often leading to an uneven coverage or no coverage of reads across the genome due to phasing [237–240] (see 4.3.2.3 for more details). To prevent such an outcome, 20-30% PhiX library were usually added to the NGS samples.

5.4 Analysis of a mammalian cell-display-based screening technology

5.4.1 Advantages of the mammalian cell-display technique

The herein described mammalian cell-display technology provided several advantages compared to other currently used methods. In contrast to yeast [241,242] or phage display [243,244], Env proteins are expressed in their trimeric conformation on widely used Hek293 cells which ensure mammalian glycosolyation as well as native folding. This is particularly beneficial since glycan-dependent epitopes [138,140] on Env are preserved. To prevent shedding of gp120, the furin cleavage site was mutated from REKR to REKS. Unfortunately, it was recently reported that the mutation of cleavage sites disturbs the quaternary structure of Env [245] and thus, analysis with certain structure-dependent antibodies such as PG145 or 35O22 is excluded. Nevertheless, antibodies recognizing monomeric Env, as in the case of VRC01 and PG9, are not affected by this conformational change and are eligible for the screening technology. In the case of conformational-dependent antibodies, however, Env proteins with the correct conformation are required which is achieved either by efficient cleavage complemented by additional stabilization as in the case of the SOSIP-variants (see 1.8.1).

In addition, the library vector pQL13 featured several benefits. Firstly, translational coupling of Env and eGFP was achieved by insertion of a TaV 2A peptide [211] between the genes. Therefore, GFP represented an independent component that could be utilized to normalize deviations in expression levels of different Env variants. This bypasses normalization via co-staining with secondary reagents and the problems that could arise due to sterical constraints caused by simultaneous staining with two antibodies (i.e. screening- and reference antibody). The described normalization approach has been successfully demonstrated previously by Dr. Tim-Henrik Bruun [209,224,225]. Due to regulation of the eGFP and Env expression by an inducible Tat operator/repressor system, Env cytotoxicity effects during cell cultivation could be eliminated.

5.4.2 Evaluation of the screening technology

Env variants with increased and decreased affinity for bnAbs were selected with a FACS cell sorting-based screening technology which was described above (see 3.3.6 and 4.4.1). In an initial sorting experiment, an enrichment of several Env candidates with the desired characteristics was demonstrated by screening the CD4bs using the CD4bs-directed antibody VRC01. To assess possible limitations of the selection system, one, five, 15 and 46 permutated positions of the CD4bs, i.e. approximately 14, 70, 210 and 644 different variants (numbers were based on the average of 14 detected amino acids

in the stable cell lines) were used in the screening, respectively. Among the enriched Env mutants, the GoB-variants T278A and T278H proved to be the most promising candidates with increased affinity to the VRC01 antibody as they were also identified independently by Veronika Grassmann and Christina Schmalzl [226] with another library using the same screening technology. Corresponding with their data, T278H was slightly superior to the T278A mutant (one to three-fold increased affinity), as it showed a slightly higher affinity for VRC01 which ranged between two to four-fold increase depending on the sort. The superiority of T278H also correlated with the structural analysis and is discussed in section 5.4.3. Unexpectedly, there was only a statistically significant difference in the antibody binding in flow cytometric equilibrium titration experiments for three of the LoB-variants and none for the 12 GoB-variants as compared to the wildype Env. Whereas loss in binding can easily occur due to structurally detrimental mutations, gain of binding is more difficult to achieve. It has been suggested previously that several mutations are necessary to evoke a distinctly increased affinity for VRC01 [191,246]. Thus, a single point mutation is unlikely to accomplish enhanced binding for the antibody.

Furthermore, results were often not reproducible as none of the variants were detected in all screened pools. Also, several substitutions were enriched as both, GoB and LoB variants, such as in the case of T278A and H. The incongruent data indicated limitations or problems with the screening technology, probably in combination with an unfavorable gating strategy. The latter was demonstrated by the same enriched variants (T278A & H) irrespective of the gate placement, so even in the LoB gate (see 4.4.4).

Additionally, the unequal amino acid composition and diversity of stable cell lines might largely contribute to the inconclusive results. In this respect, measurements should be repeated with newly generated and improved stable cell lines to assess in what way an extreme composition in stable cell lines might influence data from screening procedures. The selection technology was developed and tested by Tim-Henrik Bruun and Veronika Grassmann on a small and highly defined library containing only five Env variants where it yielded good results. For instance, the equimolar mixture of five Env variants (the wildtype variable loop 3 was replaced with the V3 regions of isolates MN, RF, CDC42, HXB2 or SF33) that previously showed a differential binding to the antibody 447-52D [139] were subjected to the flow cytometry-based cell sorting and NGS analysis. The results demonstrated a statistically significant enrichment of high affinity Env/V3 variant MN from an equal distribution of all variants for each analysis which corresponded to the previous results [139] and confirmed the validity of the selection technology for the five-variant library. However, it it possible that the screening still requires further improvement when using more complex libraries and thus, warrants specific optimizations in respect to ideal duration of induction, amounts of sorted cells, gating strategy as well as analysis approaches.

5.4.3 Structural analysis of envelope interactions with the bnAb VRC01

Structural analysis of Env interactions with VRC01 was performed by using Pymol in order to assess structural impact of amino acid substitutions.

According to the data obtained from the sorting experiments, T278 seems to be a key residue of Env for VRC01 binding as many substitutions lead to increased affinity for the antibody. Two effects might have contributed to this phenomenon: i) direct contact to the light chain of VRC01 via the residue Y91 (figure 28 B) and ii) removal of the glycan by disrupting the NxT_{278} sequon. As many substitutions lead to improved if not statistically significant affinity for VRC01 (see 4.4.5), it appears that knock-out of the N-glycosylation sites was most likely responsible for the increase in binding. However, due to several inconsistencies in the data, this could not be confirmed. For instance, removal of the glycan seemed not to be the unequivocal reason for beneficial effects on VRC01 binding, as substitutions at N276 would also eliminate the glycosylation. Yet, no favorable mutations were enriched at this position during the screening procedure. There is also the possibility that position N276 represents an important contact residue irrespective of the glycan, or that other amino acids disrupt the structure leading to reduced antibody binding. Thus, it was not possible to determine unequivocally if glycan removal improves affinity for VRC01 according to the acquired data from flowcytometric equilibrium titrations and structural analysis.

The GoB-variants T278H and T278R represented the most promising variants exhibiting improved antibody binding, though the difference to the wildtype was not significant (figure 29 A+B). The benefit might be the result of a cation-π interaction formed between VRC01 Y91 and the histidine/arginine at position 278 (figure 30 B). This is a strong, noncovalent binding interaction that can arise between phenylalanine, tyrosine or tryptophan as the π component and lysine, arginine or histidine as the cation. In this respect, the π component provides a surface of negative electrostatic potential that can bind to positively charged residues through a predominantly electrostatic interaction [247]. Furthermore, even though the T278K substitution was detected as a LoB-variant in the screening, flowcytometric equilibrium titrations revealed a slight increase in VRC01 binding, thus supporting the cation-π interaction as possible reason for higher antibody affinity.

It stands to reason that the substitution of hydrophobic amino acids with ones carrying a charge would lead to detrimental structural effects. Such a situation presumably occurred with the LoB variants L277K and A368D. The replacement of hydrophobic leucine with positively charged lysine probably lead to rearrangement of loop D, and thus resulted in loss of affinity for VRC01. Similarly, decreased binding of the variant A368D might be attributed to a possible disruption of the CD4 binding loop due to the introduction of the negative charge of aspartate. A beneficial case of charged amino acids was assumed for the GoB-variant E275K, where the positive charge of lysine could lead

to a slight rearrangement in the structure and subsequently the formation of a salt bridge with the D99 (figure 30 C).

A

	Residue	Contact
GoB	T278A	*
	T278H	*
	T278Q	*
	T278R	*
	T278S	*
	E275K	●
	G366D	○
	G367C	○
	E275N	●
	T373I	-
	T373N	-
	H374Y	-
LoB	T278E	*
	T278K	*
	T278N	*
	E275F	●
	E275T	●
	L277K	*
	L277Q	*
	V281E	●
	D368E	●
	A365D	●
	P437N	-
	E464I	*

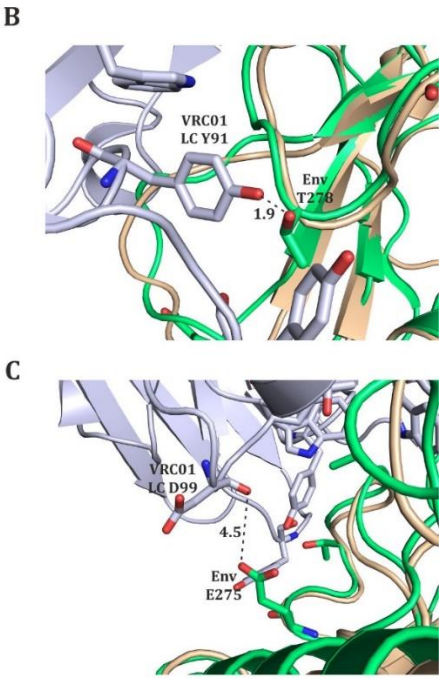


Figure 30 – Structural analysis of Env interactions with VRC01. (A) Contact residues between VRC01 and GoB/LoB variants according to published data from [249]. Amino acid numbering of mutants is based on the HIV-1 HXB2 sequence. Contact residues for VRC01 identified in the crystal structure are denoted with open circles (○) designating gp120 main-chain-only contacts, filled circles (●) representing both main-chain and side-chain contacts, and asterisks (*) denoting gp120 side-chain-only contacts. Residues that have no direct contact with VRC01 are indicated by (-). Environment of residues **(B)** T278 and **(C)** E275/V281 in the co-crystal of VRC01 and ZM176.66 gp120 (PDB 4LST). As there is no co-crystallization of VRC01 with 16055 available, the gp120 structure of 16055 (PDB 5UM8) was superimposed (green) on the ZM176.66 model (light orange). Black dashed lines illustrate the distance between residues in Angström. Structures were generated with Pymol.

Even though G367C has been detected as a GoB-variant in the screening procedure, flowcytometric equilibrium titration experiments claimed the opposite since a statistically significant loss in VRC01 affinity (p= 0.0004) could be confirmed. As cysteines usually serve an important structural role in many proteins by forming disulfide-bridges, introduction of this amino acid most likely disrupted the Env structure and therefore lead to reduced VRC01 binding. A similar case is represented in the LoB-variant P437N. Proline is known to play a key role in the rate-determining steps of protein folding [248], hence a replacement with asparagine supposedly had a destabilizing effect and resulted in decreased antibody affinity.

6 Summary and conclusions

This PhD thesis focused on the characterization and improvement of an Env-based sequential permuation library. In the first part, the general quality of the SeqPer library was assessed with different methods (see 4.3.1 and 4.3.3). The quality of the pDNA demonstrated errors in the form of deletions mainly in the Env region in 48% of the library (see 4.3.1.2). However, when looking at the sublibraries individually, the majority of the pools-of-position showed the correct pDNA and only a small portion of the sublibrary (6-18%) was actually affected by the deletions (see 4.3.1.3 and 4.3.1.4). Thus, the quality of the pDNA was deemed acceptable for further experiments.

In the following, the diversity and general amino acid composition of the CD4 binding site of the SeqPer library were analyzed on the level of the pDNA and the generated respective stable cell lines (see 4.3.3). The pDNA demonstrated a nearly even decoded amino acid composition with an average of 19 detected amino acids in every sublibrary. However, a closer examination of the stable cell lines revealed a significant loss in diversity by 32% (from an average of 19 to 14 amino acids in the stock/non-induced and 12 aas in the induced stable cell lines). In addition, a considerable bias in the amino acid composition was detected in every stable cell line (see 4.3.3). These results could be attributed to limitations in the procedure of the generation of stable cell lines due to insufficient integration events. As the quality of stable cell lines proved to be unsatisfactory, various approaches were undertaken to optimize their production. It was shown that the relative diversity of an average of 0.05 (maximal relative diversity 1.0) could be improved to 0.31 by upscaling of the transfection to generate more foci (see 4.3.4). Although the diversity did not reach the value from the pDNA (D_{rel}=0.59), the distinct improvement indicated that further optimizations for the generation of stable cell lines can be achieved (discussed in section 5.2).

The second part of this dissertation focused on the application of the CD4bs of the SeqPer library in the described cell sorting-based screening technology (see 4.4) to identify Env variants with increased or decreased binding affinity for the bnAb VRC01. The results showed that the library could be successfully utilized in the screening procedure, as an enrichment of the GoB-variants T278A and T278H was demonstrated. These variants in particular were previously reported to increase affinity for the bnAb VRC01 [213,226]. Unexpectedly, no GoB- and only three LoB-variants proved to be statistically significant and additionally, the selection of individual variants was not reproducible in every cell-sorting. For instance, T278A was enriched in two cell sorting experiments, whereas T278H was found in three. Additionaly, T278A and T278H could be detected as GoB- and LoB-variants. This indicated limitations of the screening technology which previously was tested on only a five-variant library [209,213]. Therefore, the screening procedure and gating approach probably require adaptations for usage of more complex and large libraries.

7 Perspective

The usage of the SeqPer library was limited due to suboptimal amino acid composition of stable cell lines thus, the next steps should address the improvement of stable cell line generation. As part of Benjamin Zimmers PhD thesis, single variant stable cell lines representing a BG505 SOSIP gp145 library are currently generated. Variants will be pooled equally to create the desired amino acid composition. The library consists of 56 variants in which all glycosylation motifs (NxT/NxS) are substituted by alanine. This concept was designed to assess glycan-dependence of antibodies and to identify germline-targeting Env candidates.

The herein described mammalian cell display and FACS-based cell sorting technology was tested on the small and defined CD4 binding site of the library to assess possible limitations. With improvement of stable cell lines, the screening should eventually be applicable on the complete SeqPer library of approximately 13.000 variants. In addition, screening antibodies could be expanded to identify germline-targeting Env variants that can be utilized in sequential immunization approaches. As a fact, a well-folded BG505 SOSIP library is currently under development which is intended to partake in projects within the EHVA (European HIV Vaccine Association) consortium.

As it is unlikely that a singular mutation leads to elicitation of bnAbs, detected favorable substitutions could be combined within a Env immunogen in order to utilize possible synergistic effects [191,246]. This was successfully achieved by Veronika Grassman with membrane-bound gp145 and soluble gp140 Env proteins containing L11A and T278A/H mutations. The variants not only demonstratd increased affinity for bnAbs, but also reduced recognition of non-neutralzing antibodies and improved Env trimerization. Currently, 16055 L111A/T278H Env immunogens, alongside 16055 WT, 16055 SOSIP WT and 16055 SOSIP L111A/T278H proteins, are tested in a prime boost immunization study with white rabbits.

8 Appendix

8.1 List of Abbreviations

The respective IUPAC (International Union of Pure and Applied Chemistry) single letter codes were used for designation of amino acids [250] and nucleotides [251].

Aa	Amino acids
Ad5	Adenovirus 5
ADCC	Antibody-Dependent Cellular Cytotoxicity
AIDS	Acquired Immune Deficiency Syndrome
ART	Anti-Retroviral Therapy
AUC	Area under the Curve
bnAb	Broadly Neutralizing Antibody
bp	Base pair
bs	Bridging sheet
BSA	Bovine Serum Albumine
C1-C5	Constant Regions of HIV-1 Envelope
CA	Capsid
CCR5	C-C Chemokine Receptor Type 5
CDC	Center for Disease Control and Prevention
CD4	Cluster of Differentiation 4 (receptor)
CD4bs	CD4 binding site
CD4bl	CD4 binding loop
CDR H3	Complementarity Determining Region 3 of the Antibody Heavy Chain
Cfu	Colony forming unit
CIP	Calf Intestinal Phosphatase
CMV	Cytomegalovirus
CRF	Circulating Recombinant Form
CT	Cytoplasmic Tail of HIV-1 Envelope
CTL	Cytotoxic T-Lymphocytes
CXCR4	C-X-C Chemokine Receptor Type 4
DMEM	Dulbecco`s Modified Eagle Medium
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic Acid
D_{rel}	Relative Diversity
E. coli	Escherichia Coli
EDTA	Ethylenediaminetetraacetic acid
Env	HIV-1 Envelope protein
FACS	Fluorescence-Activated Cell Sorting
FCS	Fetal Calf Serum
FP	Fusion Peptide
FRT	Flippase Recombination Target
gDNA	genomic Deoxyribonucleic Acid
GFP	Green Fluorescent Protein
GoB	Gain of Binding
gp	Glycoprotein
HAART	Highly Active Anti-Retroviral Therapy
HEK	Human Embryonic Kidney
HVTN	HIV Vaccine Trials Network
IAVI	International AIDS Vaccine Initiative
ICTV	International Committee on Taxonomy of Viruses
Ig	Immunoglobulin
IN	Integrase
kb	Kilobases
K_D	Dissociation Constant
kDa	Kilo Dalton
LB	Lysogeny Broth
LoB	Loss of Binding
MA	Matrix
MFI	Mean Fluorescence Intensity
MPER	Membrane Proximal External Region
NC	Nucleocapsid
NGS	Next Generation Sequencing
NIH	National Institute of Health
OD	Optical Density
PAGE	Polyacrylamide Gel Electrophoresis
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PCP	Pneumocystic Carinii Pneumonia
pDNA	Plasmid Deoxyribonucleic Acid
PEI	Polyethylenimine
Pen/Strep	Penicillin/Streptomycin
PIC	Pre-Integration Complex
PNK	Polynucleotide Kinase
Poly dA	Polyadenylic Acid
PR	Protease
RNA	Ribonucleic Acid
RT	Reverse Transcriptase
SBS	Sequencing by Synthesis
SCL	Stable Cell Line
SDS	Sodium Dodecyl Sulfate
SeqPer	Sequential Permutation Library
SHM	Somatic Hypermutation
SIV	Simian Immunodeficiency Virus
SP	Spacer Peptide
TB	Terrific Broth
TGN	Trans -Golgi Network
TMB	Tetramethylbenzidine
TMD	Transmembrane Domain
V1-V5	Variable Loops of HIV-1 Envelope
w/v	weight per volume

HIV-1	Human Immunodeficiency Virus type 1	**WHO**	World Health Organization
HR1/2	Heptad Repeats / Helical Regions 1 and 2 of HIV-1 Envelope	**WT**	Wildtype
HRP	Horse-Raddish Peroxidase		

8.2 DNA constructs

8.2.1 Oligonucleotides

Oligonucleotides for Cloning and Sequencing

Name	Sequence 5‘ → 3‘	Name	Sequence 5‘ → 3‘
fw_Mly-link-pUC	AGCTTGAGTCCGCCCGGGCGGACTCG	**16055-QL-For**	ATAATAGCTAGCCGTCTCCCTAGCATGAGAGTGCGGGGCATCCTGCG
rev_Mly-link-pUC	AATTCGAGTCCGCCCGGGCGGACTCA	**16055-gp145-QL-Rev**	ATATATACTCGAGCGTCTCGTCGAGTCATCAGCTGTATCCCTGCCGC
MlyI-Linker-pUC18	AGCTTGAGCCGCCCGGGCGGACTG	**#108_DP**	TCACTCATTAGGCACCCCA
pC-CMV-For	GTAGGCGTGTACGGTGGGAGG	**#109_DP**	AGAAAATACCGCATCAGGC
pC-BGH-Rev	GCAACTAGAAGGCACAGTCGAGG	**9F1**	ATATATACTCGAGCGTCTCGTCGAGTCATCAGCTGTATCCCTGCCGC
THB-8D9-seq-f	CATGGTCCTGCTGGAGTTCGTG	**9F3**	ATAATAGCTAGCCGTCTCCCTAGCATGAGAGTGCGGGGCATCCTGCG
T276S fwd	TCATCAGAAGCGAGAACCTGAGCAACAACGTGAAAACCATCGT	**T276N_fwd_LoB1p**	TCATCAGAAGCGAGAACCTGAACAACAACGTGAAAACCATCAT
T276S rev	ACGATGGTTTTCACGTTGTTGCTCAGGTTCTCGCTTCTGATGA	**T276N_rev_LoB1p**	ATGATGGTTTTCACGTTGTTGTTCAGGTTCTCGCTTCTGATGA
T276R fwd	TCATCAGAAGCGAGAACCTGAGAAACAACGTGAAAACCATCGT	**T276E_fwd_LoB1p**	TCATCAGAAGCGAGAACCTGGAGAACAACGTGAAAACCATCAT
T276R rev	ACGATGGTTTTCACGTTGTTTCTCAGGTTCTCGCTTCTGATGA	**T276_rev_LoB1p**	ATGATGGTTTTCACGTTGTTCTCCAGGTTCTCGCTTCTGATGA
T276Q fwd	TCATCAGAAGCGAGAACCTGCAGAACAACGTGAAAACCATCGT	**E273T_fwd_LoB5p**	GCGAGATCATCATCAGAAGCAACAACCTGACCAACAACGTGAA
T276Q rev	ACGATGGTTTTCACGTTGTTCTGCAGGTTCTCGCTTCTGATGA	**E273T_rev_LoB5p**	TTCACGTTGTTGGTCAGGTTGTTGCTTCTGATGATGATCTCGC
T369I fwd	CGGCGACCTGGAAATCACCATCCACAGCTTCAACTGCA	**E273F_fwd_LoB5p**	GCGAGATCATCATCAGAAGCTTCAACCTGACCAACAACGTGAA
T369I rev	TGCAGTTGAAGCTGTGGATGGTGATTTCCAGGTCGCCG	**E273F_rev_LoB5p**	TTCACGTTGTTGGTCAGGTTGAAGCTTCTGATGATGATCTCGC
T369N fwd	CGGCGACCTGGAAATCACCAACCACAGCTTCAACTGCA	**L275K_fwd_LoB5p**	TCATCATCAGAAGCGAGAACAAGACCAACAACGTGAAAACCAT
T369N rev	TGCAGTTGAAGCTGTGGTTGGTGATTTCCAGGTCGCCG	**L275K_rev_LoB5p**	ATGGTTTTCACGTTGTTGGTCTTGTTCTCGCTTCTGATGATGA
H370Y fwd3	GCGACCTGGAAATCACCACCTACAGCTTCAACTGCAGAGGCGA	**V279E_fwd_LoB15p**	GCGAGAACCTGACCAACAACGAGAAAACCATCATCGTGCACCT

H370Y rev3	TCGCCTCTGCAGTTGAAGCTGTAG GTGGTGATTTCCAGGTCGC	**V279E_rev_LoB 15p**	AGGTGCACGATGATGGTTTTCTCG TTGTTGGTCAGGTTCTCGC
E273K fwd	GCGAGATCATCATCAGAAGCAAAA ACCTGACCAACAACGTGAA	**L275Q_fwd_LoB 15p**	TCATCATCAGAAGCGAGAACCAGA CCAACAACGTGAAAACCAT
E273K rev	TTCACGTTGTTGGTCAGGTTTTTG CTTCTGATGATGATCTCGC	**L275Q_rev_LoB 15p**	ATGGTTTTCACGTTGTTGGTCTGG TTCTCGCTTCTGATGATGA
G362D fwd	TCAACTTCACCAGCCCCGCTGATG GCGACCTGGAAATCAC	**D364E_fwd_LoB 15p**	TCACCAGCCCCGCTGGCGGCGAGC TGGAAATCACCACCCACAG
G362D rev	GTGATTTCCAGGTCGCCATCAGCG GGGCTGGTGAAGTTGA	**D364E_rev_LoB 15p**	CTGTGGGTGGTGATTTCCAGCTCG CCGCCAGCGGGGCTGGTGA
G363C fwd	ACTTCACCAGCCCCGCTGGCTGCG ACCTGGAAATCACCA	**A361D_fwd_LoB 46p**	TCATCAACTTCACCAGCCCCGACG GCGGCGACCTGGAAATCAC
G363C rev	TGGTGATTTCCAGGTCGCAGCCAG CGGGGCTGGTGAAGT	**A361D_rev_LoB 46p**	GTGATTTCCAGGTCGCCGCCGTCG GGGCTGGTGAAGTTGATGA
E273N fwd	GCGAGATCATCATCAGAAGCAACA ACCTGACCAACAACGTGAA	**E459I_fwd_LoB 46p**	ACGGCGGCGTGGAAAGCAACATCA CAGAGATCTTCAGACCCGG
E273N rev	TTCACGTTGTTGGTCAGGTTGTTG CTTCTGATGATGATCTCGC	**E459I_rev_LoB46p**	CCGGGTCTGAAGATCTCTGTGATG TTGCTTTCCACGCCGCCGT
T276K_fwd_LoB 1p	TCATCAGAAGCGAGAACCTGAAGA ACAACGTGAAAACCATCAT	**P432N_fwd_LoB 46p**	TGGGCAGAGCTATGTACGCCAACC CCATCGAGGGCAACATCAC
T276K_rev_LoB 1p	ATGATGGTTTTCACGTTGTTCTTC AGGTTCTCGCTTCTGATGA	**P432N_rev_LoB 46p**	GTGATGTTGCCCTCGATGGGGTTG GCGTACATAGCTCTGCCCA

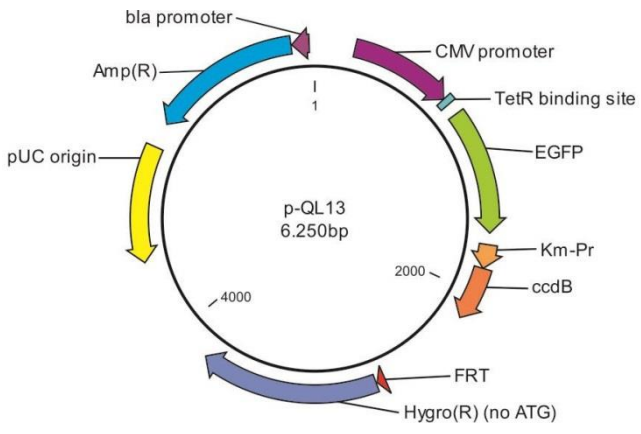
Oligonucleotides for NGS library preparation

Name	Sequence 5‘ → 3‘	Name	Sequence 5‘ → 3‘
B2#3 fwd+ILL	AGTTCTACAGTCCGACGATCCAAG GCCTACGAGAAAGAGG	**B9 fwd+ILL**	AGTTCTACAGTCCGACGATCGCAA CAACAAGACCTTCAACG
B2#3 rev+ILL	CTTGGCACCCGAGAATTCCACGTT GGTGGTGTTGACCTG	**B9 rev+ILL**	CTTGGCACCCGAGAATTCCAATTT CCACGCTCTCGTTCAG
B4fwd+ILL	AGTTCTACAGTCCGACGATCTCTA TCGGCTGCTGGAAGAT	**B10_fwd+ILL**	AGTTCTACAGTCCGACGATCGAGG GCGAGATCATCATCAG
B4rev+ILL	CTTGGCACCCGAGAATTCCAACTC TGTTCACCACGCTCAG	**B10_rev+ILL**	CTTGGCACCCGAGAATTCCAGATG ATCCGTCTGGGGAAGT
B5fwd+ILL	AGTTCTACAGTCCGACGATCGGCC ATCGAGAGATACCTGA	**B11 fwd +ILL**	AGTTCTACAGTCCGACGATCCGGA CCCTGCAGAGAGTG
B5rev+ILL	CTTGGCACCCGAGAATTCCATCTC GTTCTGTTCTTGCTGGT	**B11 rev+ILL**	CTTGGCACCCGAGAATTCCACTGT TGCTGTTGGTGTCGTT
B6 fwd+ILL	AGTTCTACAGTCCGACGATCCTGT GCGTGACCCTGGAAT	**B12 fwd+ILL**	AGTTCTACAGTCCGACGATCAAAT CACCACCCACAGCTTC
B6 rev+ILL	CTTGGCACCCGAGAATTCCAGCAG TTGATCAGCCGGTACT	**B12rev+ILL**	CTTGGCACCCGAGAATTCCAGCTC TTGCATGTGATGTTGC
B7fwd+ILL	AGTTCTACAGTCCGACGATCCAAT GCCACCACCGAGAT	**B13fwd+ILL**	AGTTCTACAGTCCGACGATCCAGC CTGGACATCACCATC
B7rev+ILL	CTTGGCACCCGAGAATTCCAGCCG TTGAAGGTCTTGTTGT	**B14fwd+ILL**	AGTTCTACAGTCCGACGATCTGCA AGAGCAACATCACC

B15fwd+ILL	AGTTCTACAGTCCGACGATCACAAGGTGGTGGAAATCAAG	**B14rev+ILL**	CTTGGCACCCGAGAATTCCACAGAAAGCCGAAGATCACG
B15rev+ILL	CTTGGCACCCGAGAATTCCACCTTCAGCAGGTTGCTCTG	**B15fwd+ILL**	AGTTCTACAGTCCGACGATCACAAGGTGGTGGAAATCAAG
B16fwd+ILL	AGTTCTACAGTCCGACGATCGTGATCTTCGGCTTTCTGG	**B15rev+ILL**	CTTGGCACCCGAGAATTCCACCTTCAGCAGGTTGCTCTG
B16rev+ILL	CTTGGCACCCGAGAATTCCAGAGCTGCTGATCCTTCAGGT	**B16fwd+ILL**	AGTTCTACAGTCCGACGATCGTGATCTTCGGCTTTCTGG
B2#1_fwd+ILL	AGTTCTACAGTCCGACGATCCTTCTGGGTGCTGATGATCT	**B16rev+ILL**	CTTGGCACCCGAGAATTCCAGAGCTGCTGATCCTTCAGGT
V3Pool_f2+ILL	AGTTCTACAGTCCGACGATCGAGGGCGAGATCATCATCAG	**B17fwd+ILL**	AGTTCTACAGTCCGACGATCATCGAGGCCCAGCAGCAT
V3Pool_r2+ILL	CTTGGCACCCGAGAATTCCACAGCTTCTTGCCCACTCTCT	**B17rev+ILL**	CTTGGCACCCGAGAATTCCAATCTCGTCGTGGCTCTTGTT
B2#1_fwd+ILL	AGTTCTACAGTCCGACGATCCTTCTGGGTGCTGATGATCT	**B18fwd+ILL**	AGTTCTACAGTCCGACGATCGGATCAGCAGCTCCTGGGCATCT
B2#1_rev+ILL	CTTGGCACCCGAGAATTCCAACGTTTTCCAGCACCATTTC	**B18rev+ILL**	CTTGGCACCCGAGAATTCCACTACACCAACACCATCTATCG
ILLUMINAseq_fwd	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGACGATC	**ILLUMINAseq_rev_Index20**	CAAGCAGAAGACGGCATACGAGATGGCCACGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index21_ILL_rev**	CAAGCAGAAGACGGCATACGAGATCGAAACGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCC	**Index22_ILL_rev**	CAAGCAGAAGACGGCATACGAGATCGTACGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index23_ILL_rev**	CAAGCAGAAGACGGCATACGAGATCCACTCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index24_ILL_rev**	CAAGCAGAAGACGGCATACGAGATGCTACCGTGACTGGAGTTCC
ILLUMINAseq_rev_Index5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index25_ILL_rev**	CAAGCAGAAGACGGCATACGAGATATCAGTGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index26_ILL_rev**	CAAGCAGAAGACGGCATACGAGATGCTCATGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index7	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCC	**Index27_ILL_rev**	CAAGCAGAAGACGGCATACGAGATAGGAATGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index8	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index28_ILL_rev**	CAAGCAGAAGACGGCATACGAGATCTTTTGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index9	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCC	**Index29_ILL_rev**	CAAGCAGAAGACGGCATACGAGATTAGTTGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index10	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index30_ILL_rev**	CAAGCAGAAGACGGCATACGAGATCCGGTGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA

ILLUMINAseq_rev_Index11	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index31_ILL_rev**	CAAGCAGAAGACGGCATACGAGATATCGTGGTGACTGGAGTTCC
ILLUMINAseq_rev_Index12	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index32_ILL_rev**	CAAGCAGAAGACGGCATACGAGATTGAGTGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index13	CAAGCAGAAGACGGCATACGAGATTTGACTGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index33_ILL_rev**	CAAGCAGAAGACGGCATACGAGATCGCCTGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index14	CAAGCAGAAGACGGCATACGAGATGGAACTGTGACTGGAGTTCC	**Index34_ILL_rev**	CAAGCAGAAGACGGCATACGAGATGCCATGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index15	CAAGCAGAAGACGGCATACGAGATTGACATGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index35_ILL_rev**	CAAGCAGAAGACGGCATACGAGATAAAATGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index16	CAAGCAGAAGACGGCATACGAGATGGACGGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index36_ILL_rev**	CAAGCAGAAGACGGCATACGAGATTGTTGGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index17	CAAGCAGAAGACGGCATACGAGATCTCTACGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index37_ILL_rev**	CAAGCAGAAGACGGCATACGAGATATTCCGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
ILLUMINAseq_rev_Index18	CAAGCAGAAGACGGCATACGAGATGCGGACGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index38_ILL_rev**	CAAGCAGAAGACGGCATACGAGATAGCTAGGTGACTGGAGTTCC
ILLUMINAseq_rev_Index19	CAAGCAGAAGACGGCATACGAGATTTTCACGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA	**Index39_ILL_rev**	CAAGCAGAAGACGGCATACGAGATGTATAGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA

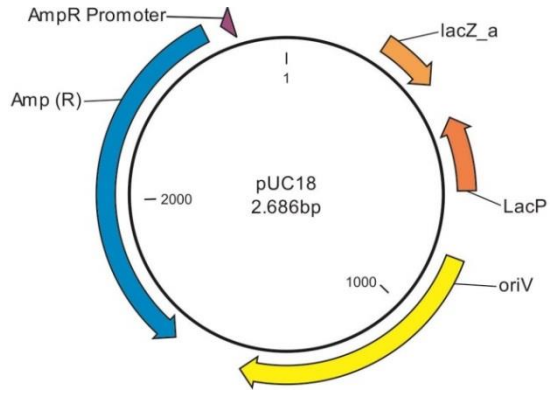
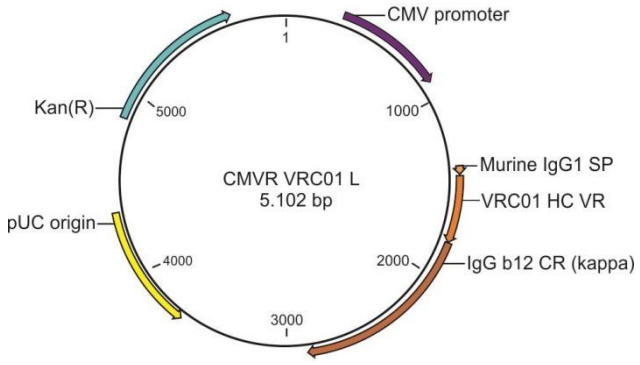
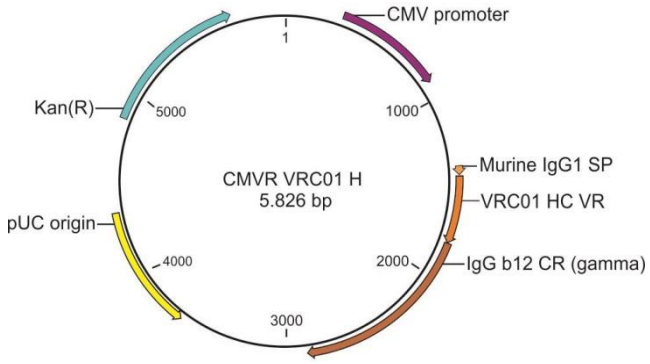
8.2.2 Plasmids



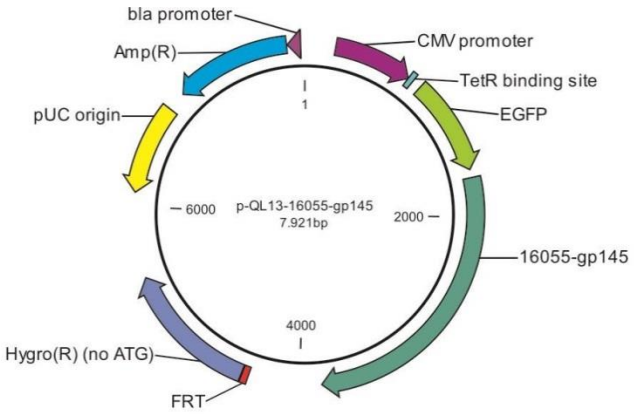
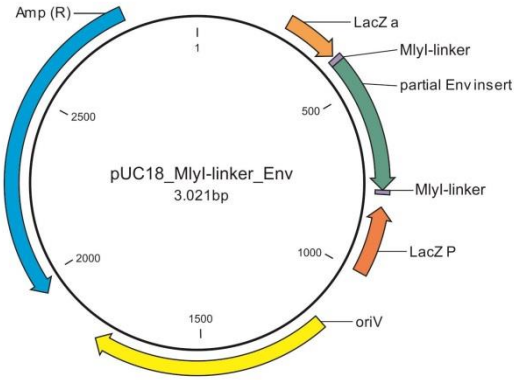
Name: pQL13

The plasmid was developed by Dr. Tim-Henrik Bruun during his PhD thesis.

The vector is eligible for stable cell line generation due to the inherent FRT site. Insertion of target DNA can be performed by using BsmBI restriction sites that frame the toxic ccdB cassette.

	Name: pUC18 Commercially available plasmid acquired from GenScript, Catalog number SD1162. Plasmid cloning vector is isolated from *E. coli* strain DH5α by standard procedures.
	Name: CMVR VRC01 L Commercially available plasmid at NIH AIDS Reagent Program, Catalog number 12036 Expression of leader, variable & constant regions of VRC01 light chain under the control of the CMV promoter.
	Name: CMVR VRC01 H Commercially available plasmid at NIH AIDS Reagent Program, Catalog number 12035 Expression of leader, variable & constant regions of VRC01 heavy chain under the control of the CMV promoter

8.2.3 Cloning Constructs

	Name: pQL13-16055-gp145 The pQL13 plasmid was developed by Dr. Tim-Henrik Bruun during his PhD thesis. Cloning of Env was performed by using BsmBI restriction sites leading to removal of the toxic ccdB cassette.
	Name: pUC18_MlyI-linker Insertion of a cassette framed by MlyI restriction sites. Env was inserted into the MlyI cassette by using SmaI which was part of the cassette. The construct was generated specifically to produce an error-free blunt end Env-fragment for the background determination of the NGS device (4.3.2.2).

8.3 Supplemental Material

Table S1 – Purity of the pDNA SeqPer library. Absorbance ratios at 260 and 280 ($OD_{260/280}$) as well as 260 and 230 ($OD_{260/230}$) at each pool-of-position were measured with the NanoDrop 1000 spectrophotometer to assess probable contaminations.

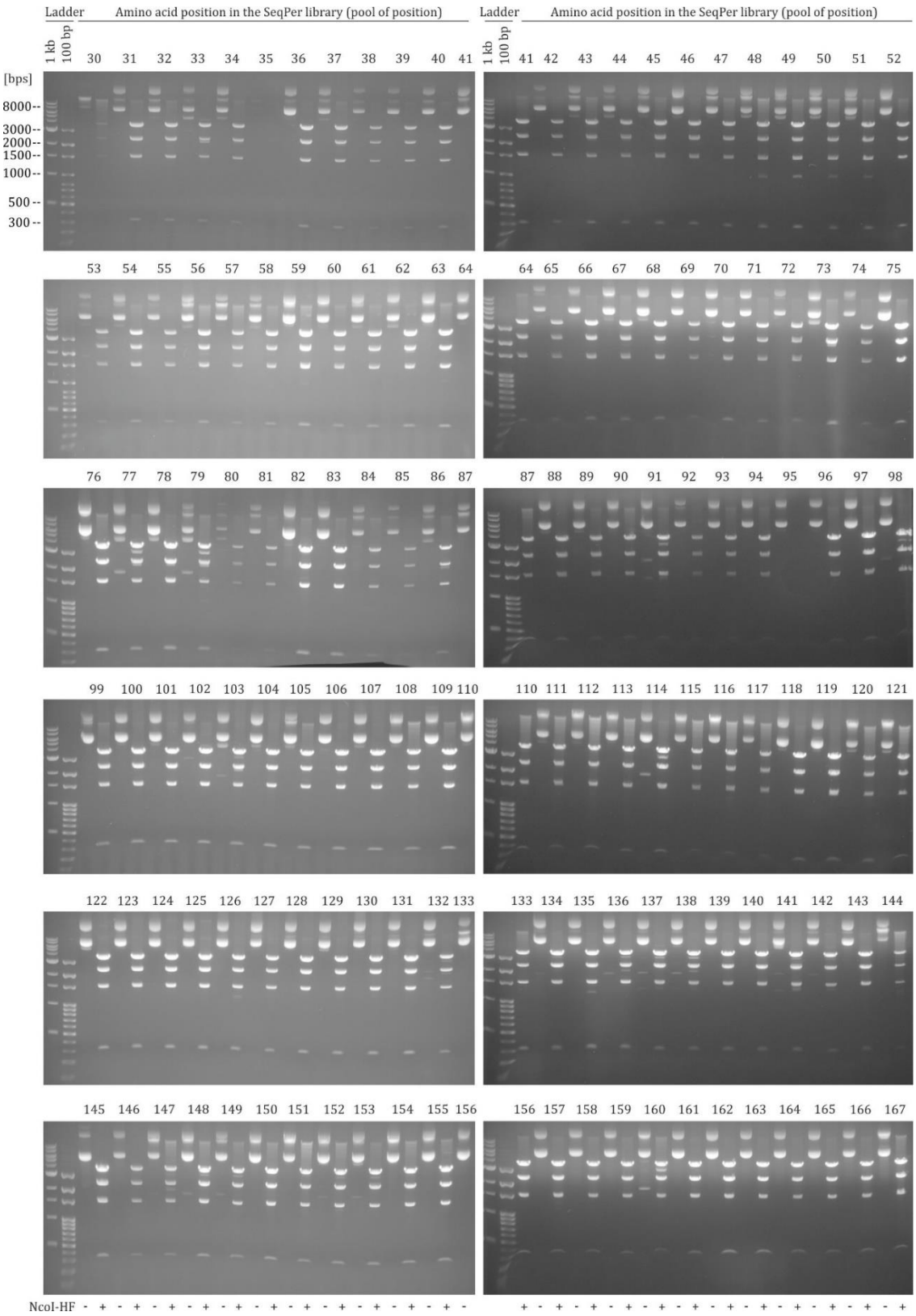
Position	$OD_{260/280}$	Position	$OD_{260/280}$	Position	$OD_{260/280}$
N30X	1,94	V240X	1,93	V450X	1,94
L31X	1,97	S241X	1,93	R451X	1,93
W32X	1,95	T242X	1,93	D452X	1,95
V33X	1,97	V243X	1,94	G453X	1,96
T34X	1,95	Q244X	1,93	G454X	1,96
V35X	2,00	C245X	1,94	V455X	1,98
Y36X	1,88	T246X	1,97	E456X	1,96
Y37X	1,95	H247X	1,96	S457X	1,94
G38X	1,97	G248X	1,95	N458X	1,97
V39X	1,97	I249X	1,95	E459X	1,95
P40X	1,97	K250X	1,97	T460X	1,95
V41X	1,94	P251X	1,88	E461X	1,96
W42X	1,94	V252X	1,97	I462X	1,96
K43X	1,89	V253X	1,95	F463X	1,95
E44X	1,96	S254X	1,97	R464X	1,93
A45X	1,91	T255X	1,95	P465X	1,95
K46X	1,95	Q256X	1,90	G466X	1,87

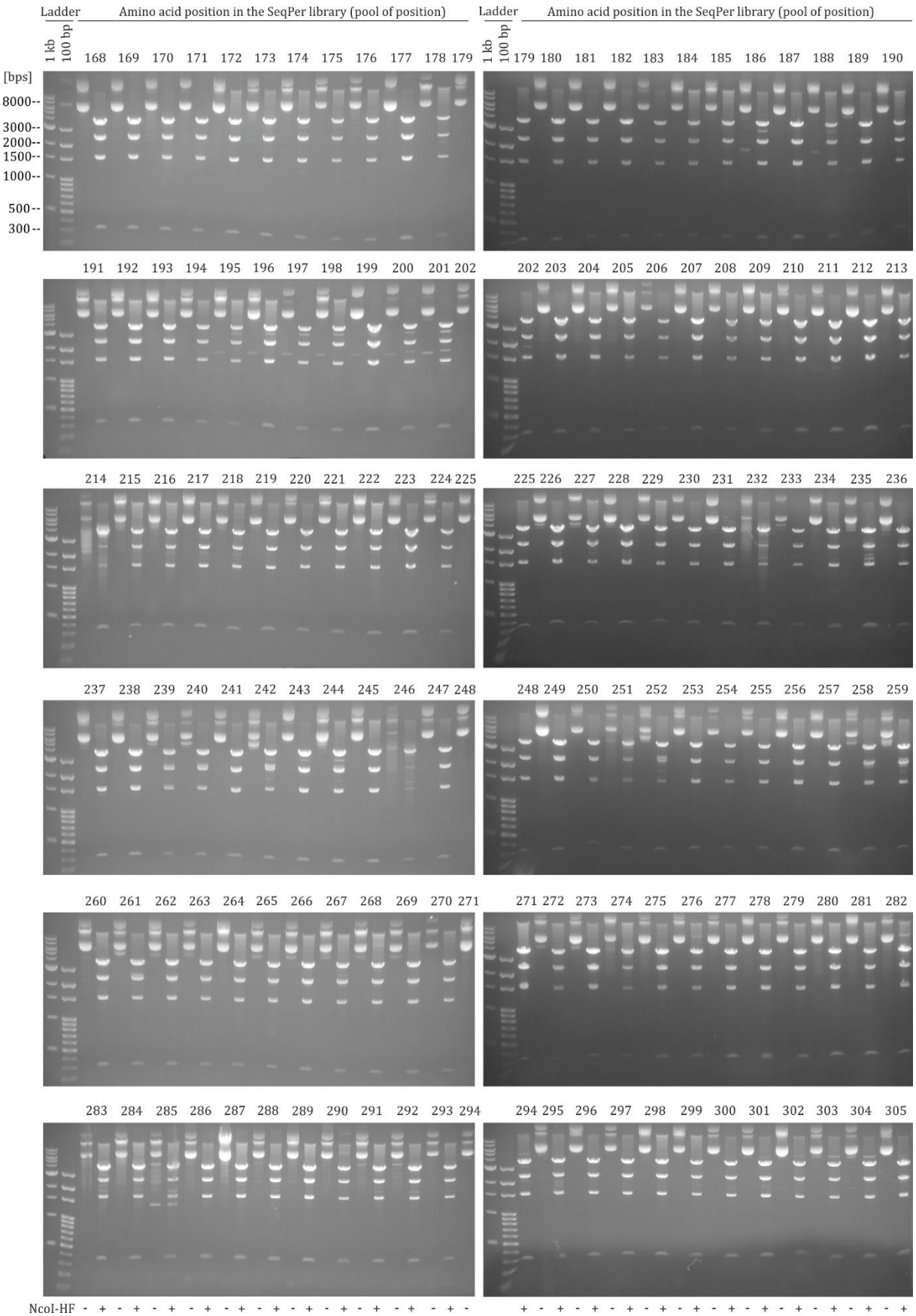
T47X	1,90	L257X	1,86	G467X	1,89
T48X	1,95	L258X	1,98	G468X	1,98
L49X	1,96	L259X	1,88	D469X	1,92
F50X	1,97	N260X	1,92	M470X	2,00
C51X	1,96	G261X	1,94	R471X	1,99
A52X	1,94	S262X	1,88	N472X	2,01
S53X	1,98	L263X	1,91	N473X	1,98
D54X	1,96	A264X	1,93	W474X	1,97
A55X	1,97	E265X	1,92	R475X	1,95
K56X	1,94	G266X	1,90	S476X	1,96
A57X	1,96	E267X	1,90	E477X	1,98
Y58X	1,95	I268X	1,97	L478X	1,89
E59X	1,97	I269X	2,00	Y479X	1,90
K60X	1,97	I270X	1,99	K480X	1,96
E61X	1,96	R271X	1,91	Y481X	2,00
V62X	1,97	S272X	1,97	K482X	1,98
H63X	1,95	E273X	1,91	V483X	1,96
N64X	1,97	N274X	1,99	V484X	1,97
V65X	1,96	L275X	1,94	E485X	1,98
W66X	1,96	T276X	1,95	I486X	2,00
A67X	1,93	N277X	1,93	K487X	2,00
T68X	1,97	N278X	1,90	P488X	2,01
H69X	1,97	V279X	1,96	L489X	2,01
A70X	1,96	K280X	1,95	G490X	1,97
C71X	1,97	T281X	1,91	I491X	1,97
V72X	1,97	I282X	1,93	A492X	2,00
P73X	1,94	I283X	1,91	P493X	1,94
T74X	1,97	V284X	1,90	T494X	2,00
D75X	1,93	H285X	1,98	A495X	1,96
P76X	1,94	L286X	1,95	A496X	1,98
N77X	1,91	N287X	1,90	K497X	1,97
P78X	1,95	E288X	1,90	R498X	1,96
Q79X	1,96	S289X	1,96	R499X	1,99
E80X	1,99	V290X	1,95	V500X	1,98
M81X	1,97	E291X	1,97	V501X	1,96
V82X	1,93	I292X	1,99	E502X	1,97
L83X	1,94	V293X	2,00	R503X	1,98
E84X	2,00	C294X	1,96	E504X	1,98
N85X	1,99	T295X	1,93	K505X	1,95
V86X	1,97	R296X	2,00	S506X	1,98
T87X	1,96	P297X	2,01	A507X	1,96
E88X	1,97	N298X	1,93	V508X	2,00
N89X	1,97	N299X	1,99	G509X	1,96
F90X	1,98	N300X	1,98	L510X	1,97
N91X	1,98	T301X	1,97	G511X	1,96
M92X	1,98	R302X	1,97	A512X	1,96
W93X	1,97	K303X	1,98	V513X	1,98
K94X	1,99	S304X	1,97	I514X	1,95
N95X	1,98	I305X	1,96	F515X	1,96
D96X	1,97	R306X	1,95	G516X	1,96
M97X	1,92	I307X	1,94	F517X	1,94
V98X	1,89	G308X	1,94	L518X	1,96
E99X	1,95	P309X	1,93	G519X	1,99
Q100X	1,91	G310X	1,95	A520X	1,95
M101X	1,91	Q311X	1,94	A521X	1,95

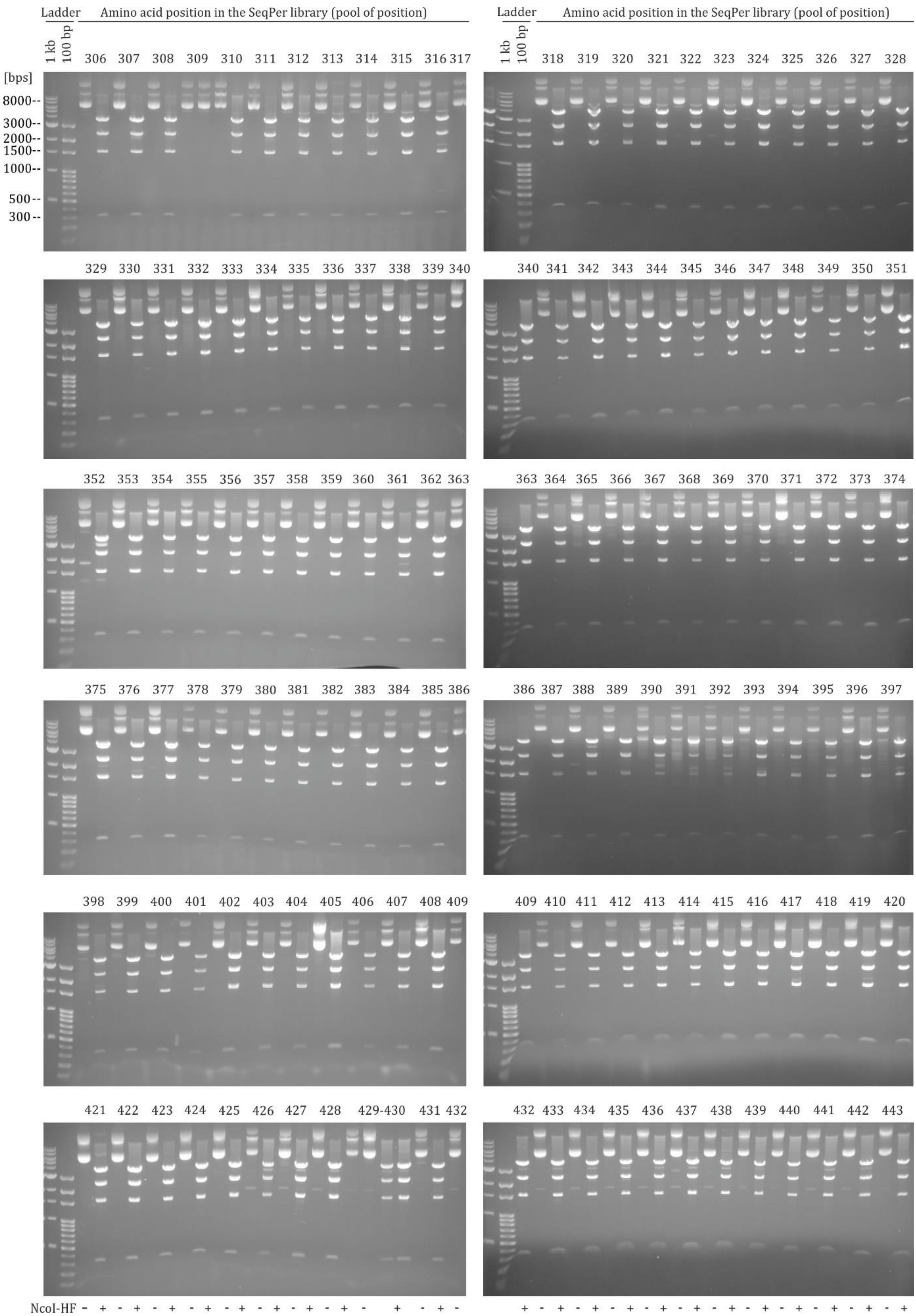
H102X	1,96	T312X	1,93	G522X	1,97
E103X	1,91	F313X	1,94	S523X	1,97
D104X	1,94	Y314X	1,93	T524X	1,99
V105X	1,95	A315X	1,95	M525X	1,96
I106X	1,90	T316X	1,94	G526X	1,97
S107X	1,89	G317X	1,94	A527X	1,95
L108X	1,98	D318X	1,95	A528X	1,96
W109X	1,95	I319X	1,94	S529X	2,00
D110X	1,95	I320X	1,95	I530X	1,90
Q111X	2,00	G321X	1,93	T531X	1,96
S112X	1,92	N322X	1,93	L532X	1,99
L113X	1,88	I323X	1,96	T533X	1,96
K114X	1,93	R324X	1,94	V534X	1,95
P115X	1,91	Q325X	2,02	Q535X	1,96
C116X	1,96	A326X	1,99	A536X	1,95
V117X	1,96	Y327X	1,99	R537X	1,98
K118X	1,93	C328X	1,97	Q538X	1,94
L119X	1,89	N329X	1,97	L539X	1,95
T120X	1,95	I330X	1,99	L540X	1,96
P121X	1,94	K331X	1,96	S541X	1,94
L122X	1,93	K332X	1,97	G542X	1,96
C123X	1,94	D333X	1,95	I543X	1,97
V124X	1,94	D334X	1,93	V544X	1,97
T125X	1,93	W335X	1,99	Q545X	1,97
L126X	1,95	I336X	1,93	Q546X	1,96
E127X	1,94	R337X	1,96	Q547X	1,98
C128X	1,88	T338X	1,94	S548X	1,97
R129X	1,95	L339X	1,99	N549X	1,96
Q130X	1,86	Q340X	1,71	L550X	1,95
V131X	1,87	R341X	1,97	L551X	1,95
N132X	1,93	V342X	1,96	K552X	1,97
T133X	1,94	G343X	1,96	A553X	1,98
T134X	1,85	K344X	1,97	I554X	1,97
N135X	1,96	K345X	1,96	E555X	1,97
A136X	1,88	L346X	1,95	A556X	1,95
T137X	1,89	A347X	1,95	Q557X	1,97
S138X	1,89	E348X	1,96	Q558X	1,97
S139X	1,97	H349X	1,96	H559X	1,98
V140X	1,96	F350X	1,97	L560X	2,00
N141X	1,86	P351X	1,94	L561X	1,98
V142X	1,90	R352X	1,95	Q562X	2,02
T143X	1,89	R353X	1,91	L563X	1,99
N144X	1,92	I354X	1,98	T564X	1,98
G145X	1,93	I355X	1,99	V565X	1,99
E146X	1,98	N356X	2,00	W566X	1,93
E147X	1,96	F357X	1,98	G567X	1,98
I148X	1,89	T358X	1,94	I568X	1,95
K149X	1,99	S359X	1,92	K569X	2,00
N150X	1,93	P360X	1,99	Q570X	1,96
C151X	1,94	A361X	1,94	L571X	1,94
S152X	1,89	G362X	1,90	Q572X	1,95
F153X	1,97	G363X	1,94	T573X	1,94
N154X	1,90	D364X	2,09	R574X	1,97
A155X	1,94	L365X	1,99	V575X	1,93
T156X	1,94	E366X	1,95	L576X	2,02

T157X	1,91	I367X	1,97	A577X	1,93
E158X	1,90	T368X	1,96	I578X	1,92
I159X	1,88	T369X	2,05	E579X	2,02
R160X	1,90	H370X	1,97	R580X	1,95
D161X	1,91	S371X	1,98	Y581X	1,93
K162X	1,90	F372X	1,95	L582X	1,99
K163X	1,91	N373X	1,99	K583X	1,90
Q164X	2,00	C374X	1,95	D584X	2,02
K165X	1,98	R375X	1,94	Q585X	1,95
V166X	1,99	G376X	1,94	Q586X	1,98
Y167X	1,93	E377X	1,95	L587X	1,97
A168X	1,94	F378X	1,98	L588X	1,94
L169X	1,89	F379X	1,97	G589X	1,96
F170X	1,89	Y380X	1,97	I590X	2,00
Y171X	2,04	C381X	1,97	W591X	1,97
R172X	1,92	N382X	1,97	G592X	1,97
L173X	1,89	T383X	1,98	C593X	1,99
D174X	1,95	S384X	1,98	S594X	1,99
I175X	1,94	S385X	1,97	G595X	1,92
V176X	1,94	L386X	1,97	K596X	1,94
P177X	1,89	F387X	1,96	L597X	1,98
L178X	1,99	N388X	1,97	I598X	1,98
E179X	1,94	S389X	1,96	C599X	1,97
E180X	1,93	T390X	1,97	T600X	1,91
E181X	1,96	Y391X	1,98	T601X	1,99
R182X	1,96	N392X	1,99	A602X	1,98
K183X	2,00	P393X	1,97	V603X	1,97
G184X	1,93	N394X	1,99	P604X	2,00
N185X	1,93	D395X	1,98	W605X	1,95
S186X	1,91	T396X	1,99	N606X	1,93
S187X	1,90	N397X	1,99	S607X	1,95
K188X	1,94	S398X	1,97	S608X	1,94
Y189X	1,93	N399X	1,95	W609X	1,94
R190X	1,92	S400X	1,99	S610X	1,94
L191X	1,91	S401X	1,98	N611X	1,95
I192X	1,91	S402X	1,97	K612X	1,94
N193X	1,92	S403X	1,97	S613X	1,96
C194X	1,91	N404X	1,97	H614X	1,88
N195X	1,92	S405X	1,93	D615X	1,97
T196X	1,95	S406X	1,99	E616X	1,96
S197X	1,93	L407X	1,98	I617X	1,95
A198X	1,91	D408X	1,93	W618X	1,93
I199X	1,92	I409X	1,98	G619X	1,95
T200X	1,94	T410X	2,00	N620X	1,95
Q201X	1,93	I411X	1,97	M621X	1,89
A202X	1,93	P412X	1,98	T622X	1,97
C203X	1,93	C413X	1,98	W623X	1,94
P204X	1,92	R414X	1,95	M624X	1,99
K205X	1,94	I415X	1,93	Q625X	1,95
V206X	1,94	K416X	1,94	W626X	1,96
T207X	1,92	Q417X	1,95	D627X	1,96
F208X	1,96	I418X	1,95	R628X	1,93
D209X	1,92	I419X	1,94	E629X	1,91
P210X	1,92	N420X	1,94	I630X	1,94
I211X	1,90	M421X	1,97	S631X	1,93

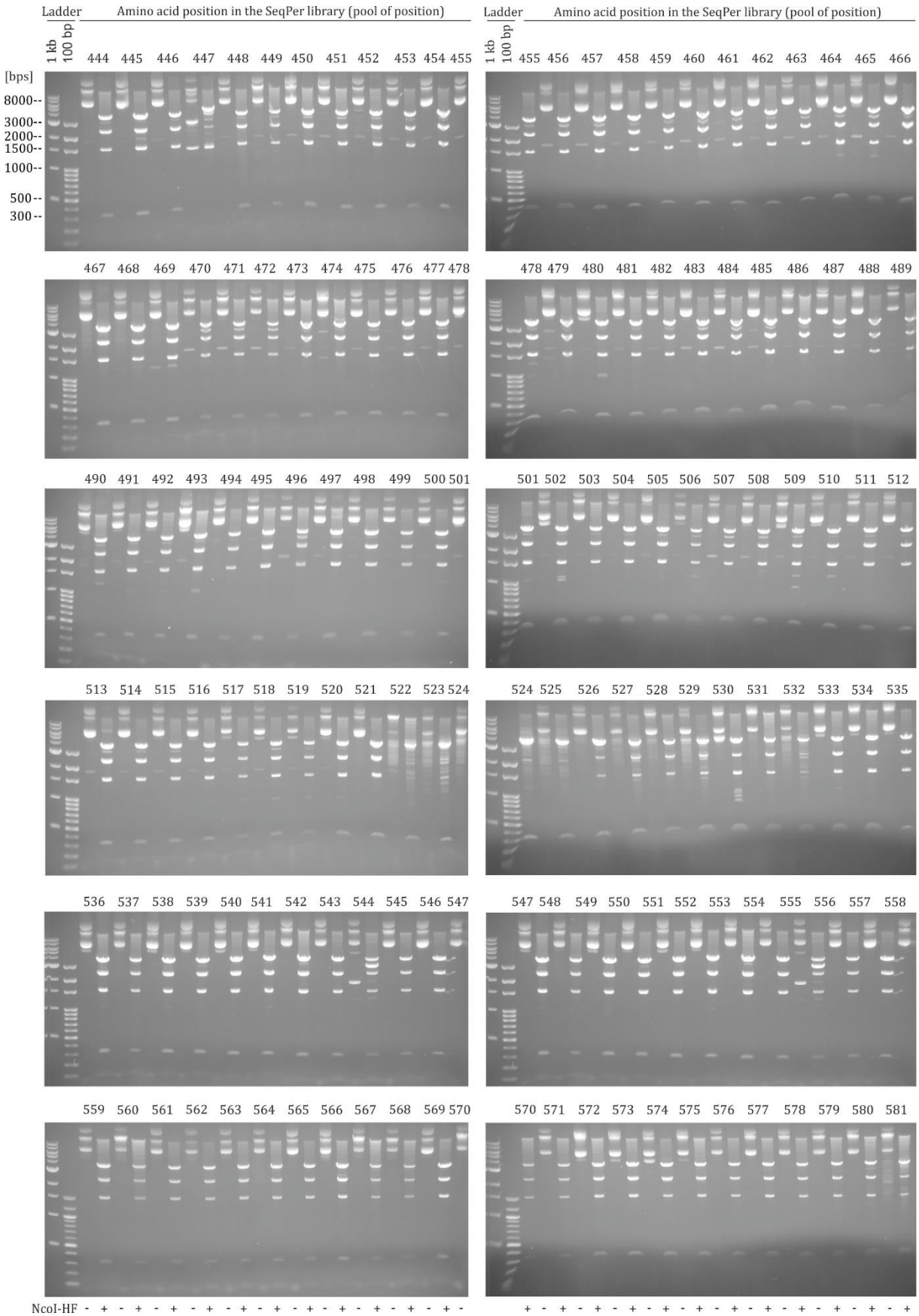
P212X	1,91	W422X	2,00	N632X	1,92
I213X	1,89	Q423X	2,00	Y633X	1,93
H214X	1,96	E424X	2,01	T634X	1,94
Y215X	1,91	V425X	1,98	N635X	1,95
C216X	1,95	G426X	1,98	T636X	1,93
A217X	1,89	R427X	1,98	I637X	1,93
P218X	1,94	A428X	1,96	Y638X	1,94
A219X	1,93	M429X	1,92	R639X	1,93
G220X	1,94	Y430X	1,99	L640X	1,93
Y221X	1,94	A431X	2,04	L641X	1,94
A222X	1,88	P432X	1,88	E642X	1,95
I223X	1,89	P433X	1,91	D643X	1,93
L224X	1,95	I434X	1,91	S644X	1,93
K225X	1,94	E435X	1,89	Q645X	1,95
C226X	1,94	G436X	1,91	N646X	1,94
N227X	1,92	N437X	1,99	Q647X	1,93
N228X	1,97	I438X	1,99	Q648X	1,94
K229X	1,93	T439X	1,91	E649X	1,97
T230X	1,95	C440X	2,05	Q650X	1,94
F231X	1,92	K441X	1,89	N651X	1,92
N232X	1,94	S442X	1,89	E652X	1,95
G233X	1,90	N443X	2,07	K653X	1,93
T234X	1,93	I444X	1,92	D654X	1,93
G235X	1,92	T445X	1,93	L655X	1,93
P236X	1,92	G446X	1,94	L656X	1,92
C237X	1,93	L447X	1,94	A657X	1,93
N238X	1,92	L448X	1,95	L658X	1,92
N239X	1,93	L449X	1,96		


Ladder
Amino acid position in the SeqPer library (pool of position)
1 kb
100 bp
[bps]
8000
3000
2000
1500
1000
500
300
30 31 32 33 34 35 36 37 38 39 40 41
41 42 43 44 45 46 47 48 49 50 51 52
53 54 55 56 57 58 59 60 61 62 63 64
64 65 66 67 68 69 70 71 72 73 74 75
76 77 78 79 80 81 82 83 84 85 86 87
87 88 89 90 91 92 93 94 95 96 97 98
99 100 101 102 103 104 105 106 107 108 109 110
110 111 112 113 114 115 116 117 118 119 120 121
122 123 124 125 126 127 128 129 130 131 132 133
133 134 135 136 137 138 139 140 141 142 143 144
145 146 147 148 149 150 151 152 153 154 155 156
156 157 158 159 160 161 162 163 164 165 166 167
NcoI-HF - + - + - + - + - + - + - + - + - + - + - + - +
+ - + - + - + - + - + - + - + - + - + - + - + - +


Ladder
Amino acid position in the SeqPer library (pool of position)
1 kb
100 bp
[bps]
8000--
3000--
2000--
1500--
1000--
500--
300--
168 169 170 171 172 173 174 175 176 177 178 179
179 180 181 182 183 184 185 186 187 188 189 190
191 192 193 194 195 196 197 198 199 200 201 202
202 203 204 205 206 207 208 209 210 211 212 213
214 215 216 217 218 219 220 221 222 223 224 225
225 226 227 228 229 230 231 232 233 234 235 236
237 238 239 240 241 242 243 244 245 246 247 248
248 249 250 251 252 253 254 255 256 257 258 259
260 261 262 263 264 265 266 267 268 269 270 271
271 272 273 274 275 276 277 278 279 280 281 282
283 284 285 286 287 288 289 290 291 292 293 294
294 295 296 297 298 299 300 301 302 303 304 305
NcoI-HF - + - + - + - + - + - + - + - + - + - + - +
+ - + - + - + - + - + - + - + - + - + - + - + - +



Ladder
Amino acid position in the SeqPer library (pool of position)
1 kb
100 bp
306 307 308 309 310 311 312 313 314 315 316 317
318 319 320 321 322 323 324 325 326 327 328
[bps]
8000--
3000--
2000--
1500--
1000--
500--
300--
329 330 331 332 333 334 335 336 337 338 339 340
340 341 342 343 344 345 346 347 348 349 350 351
352 353 354 355 356 357 358 359 360 361 362 363
363 364 365 366 367 368 369 370 371 372 373 374
375 376 377 378 379 380 381 382 383 384 385 386
386 387 388 389 390 391 392 393 394 395 396 397
398 399 400 401 402 403 404 405 406 407 408 409
409 410 411 412 413 414 415 416 417 418 419 420
421 422 423 424 425 426 427 428 429-430 431 432
432 433 434 435 436 437 438 439 440 441 442 443
NcoI-HF - + - + - + - + - + - + - + - + - + - + -
+ - + - + - + - + - + - + - + - + - + - + - + - +


Ladder
Amino acid position in the SeqPer library (pool of position)
1 kb
100 bp
[bps]
8000--
3000--
2000--
1500--
1000--
500--
300--
444 445 446 447 448 449 450 451 452 453 454 455
455 456 457 458 459 460 461 462 463 464 465 466
467 468 469 470 471 472 473 474 475 476 477 478
478 479 480 481 482 483 484 485 486 487 488 489
490 491 492 493 494 495 496 497 498 499 500 501
501 502 503 504 505 506 507 508 509 510 511 512
513 514 515 516 517 518 519 520 521 522 523 524
524 525 526 527 528 529 530 531 532 533 534 535
536 537 538 539 540 541 542 543 544 545 546 547
547 548 549 550 551 552 553 554 555 556 557 558
559 560 561 562 563 564 565 566 567 568 569 570
570 571 572 573 574 575 576 577 578 579 580 581
NcoI-HF - + - + - + - + - + - + - + - + - + - + - +
+ - + - + - + - + - + - + - + - + - + - + - +

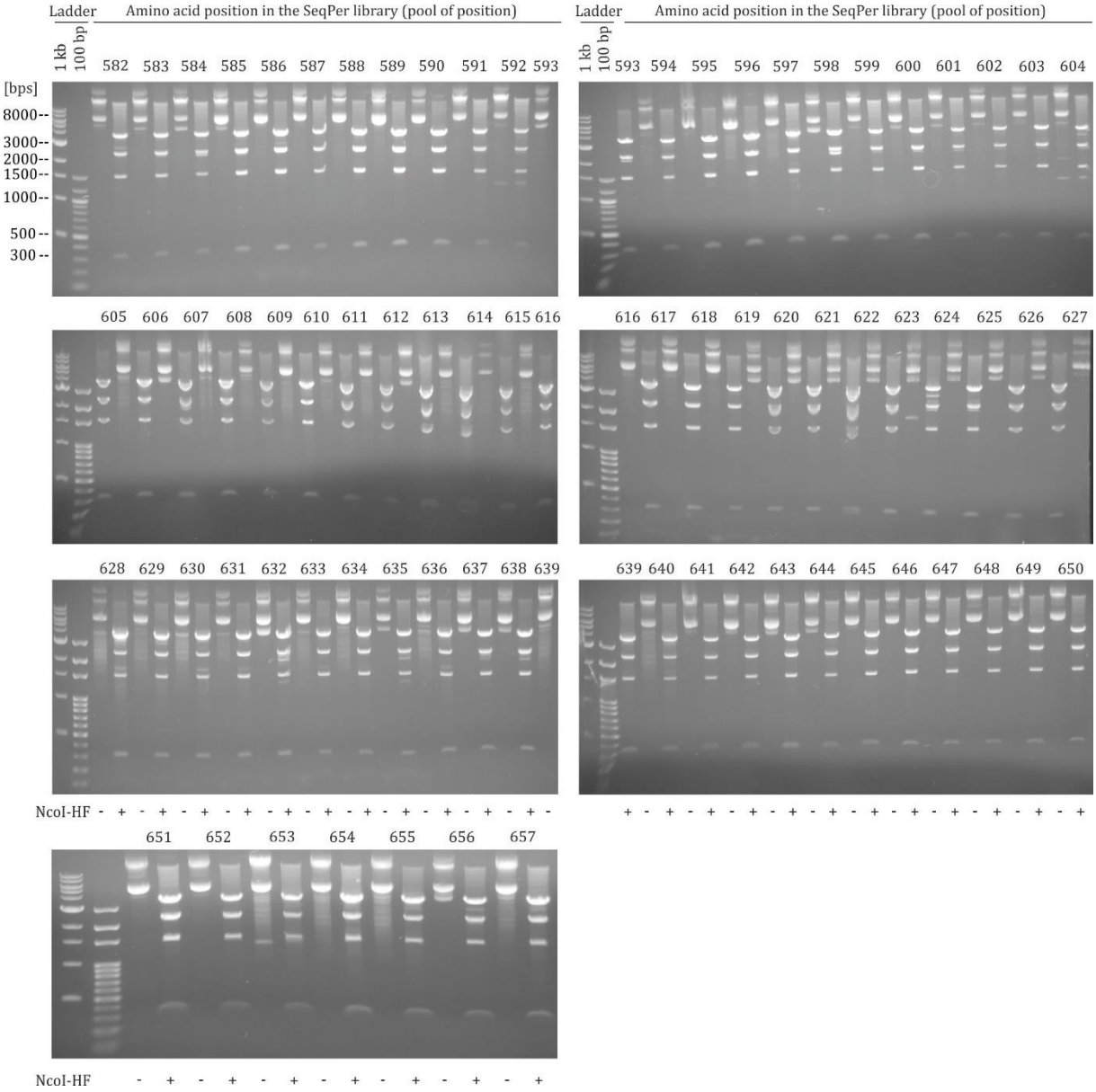


Figure S2 - Agarose gel electrophoresis of native and digested plasmid DNA. Restriction assay profiles of all pools-of-position from the SeqPer library are shown. One band at the length of 7937 bps for the native (NcoI-HF -) and four bands at 3744, 2353, 1519 and 321 base pairs for the digested pDNA (NcoI-HF +) were expected.

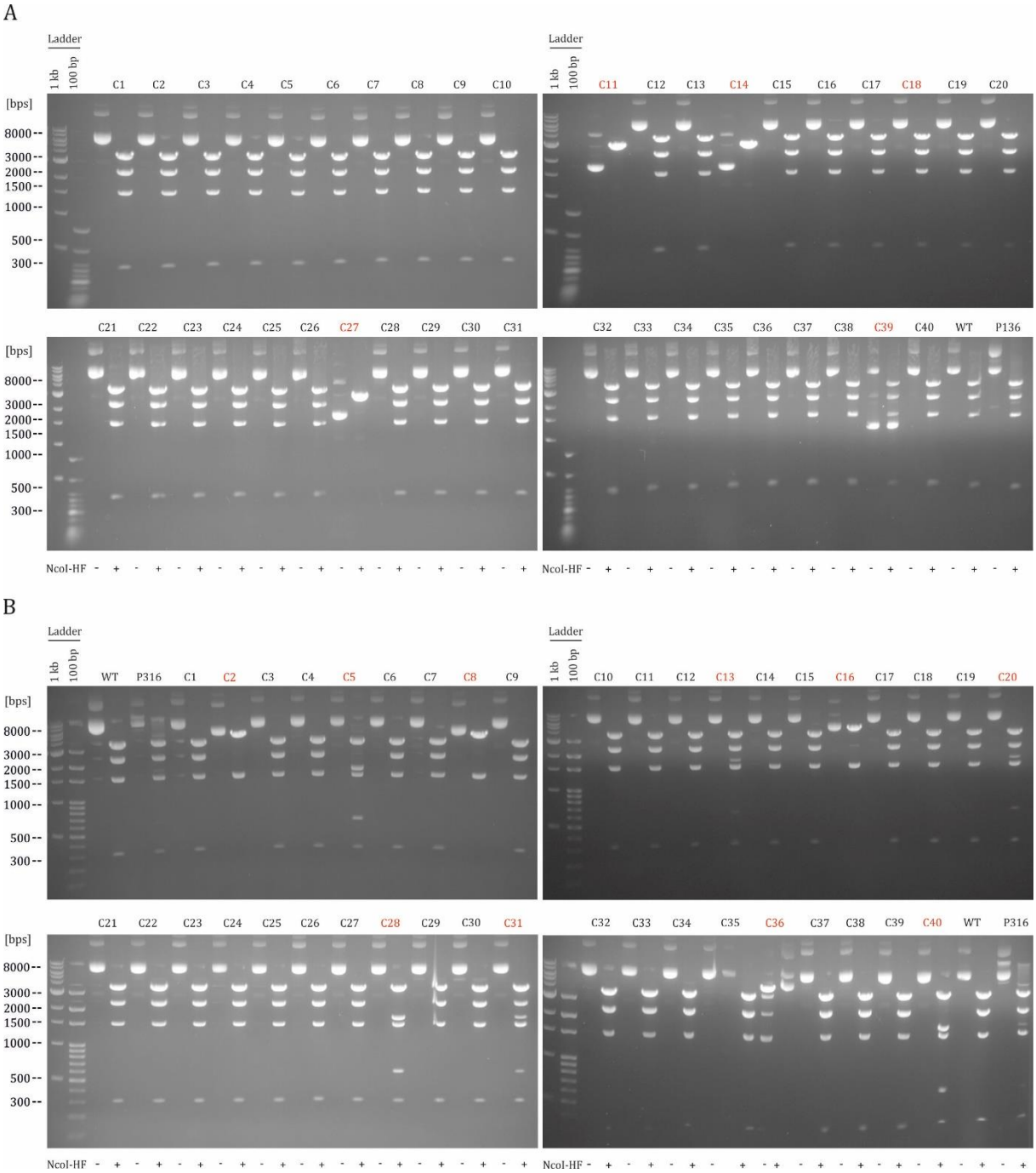
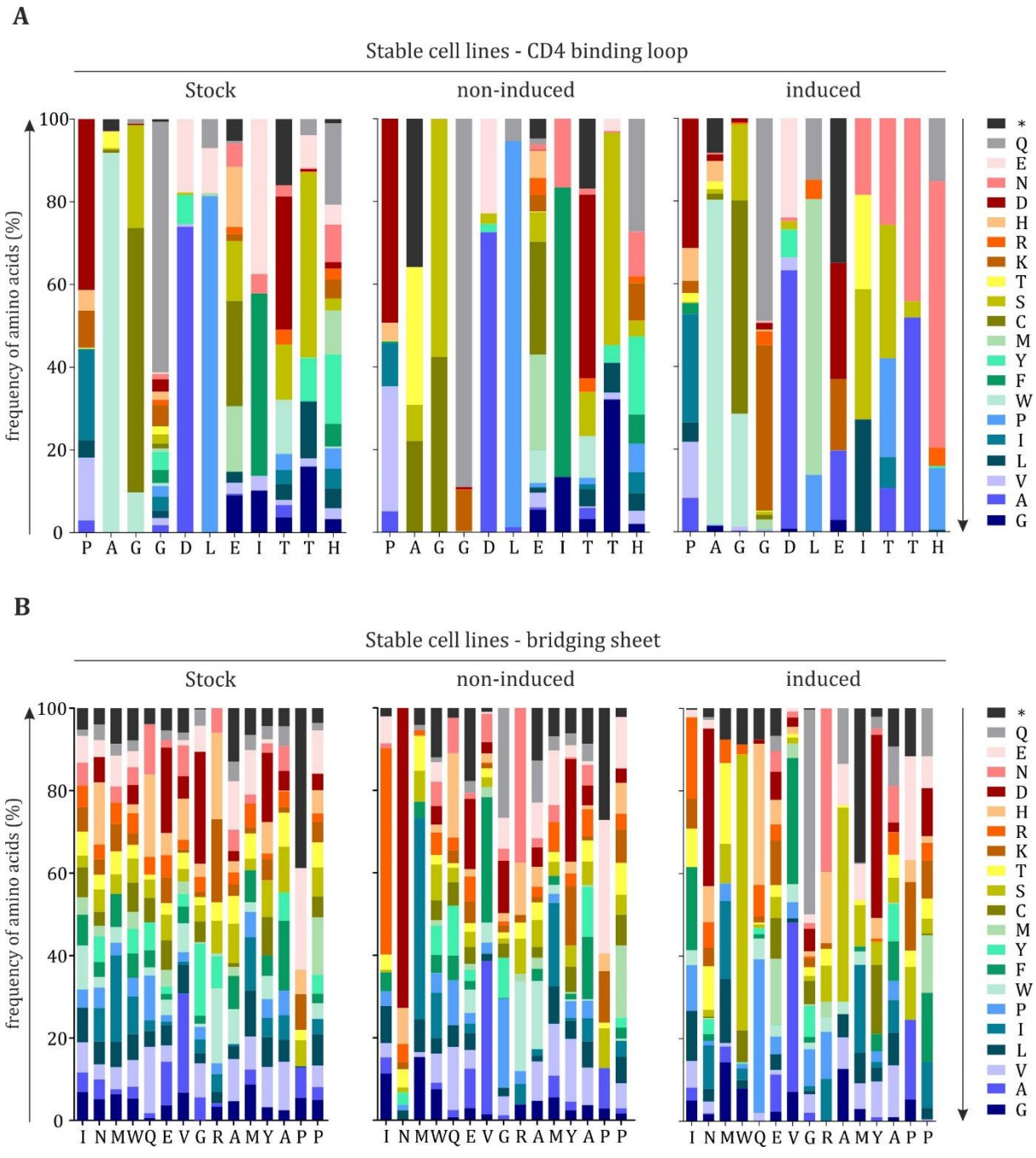


Figure S3 - Restriction enzyme assay profile for (A) sub-library 136 and (B) sub-library 316. Plasmid DNA of 40 random colony-forming units from (C1-C40) sub-library 136 and 316 of the SeqPer library was analyzed. 1 µg native and digested (NcoI-HF, 2h, 37°C) plasmid DNA was visualized on a 0.8 % agarose gel. The wildtype restriction pattern from the 16055 gp145 Env (WT) and the pools-of-position 135 (P136) and 316 (P316) are also shown. Aberrational colony forming units are denoted in red.


A
Stable cell lines - CD4 binding loop
Stock
non-induced
induced
frequency of amino acids (%)
100
80
60
40
20
0
P A G G D L E I T T H
* Q E N D H R K T S C M Y F W P I L V A G
B
Stable cell lines - bridging sheet
Stock
non-induced
induced
frequency of amino acids (%)
I N M W Q E V G R A M Y A P P

C

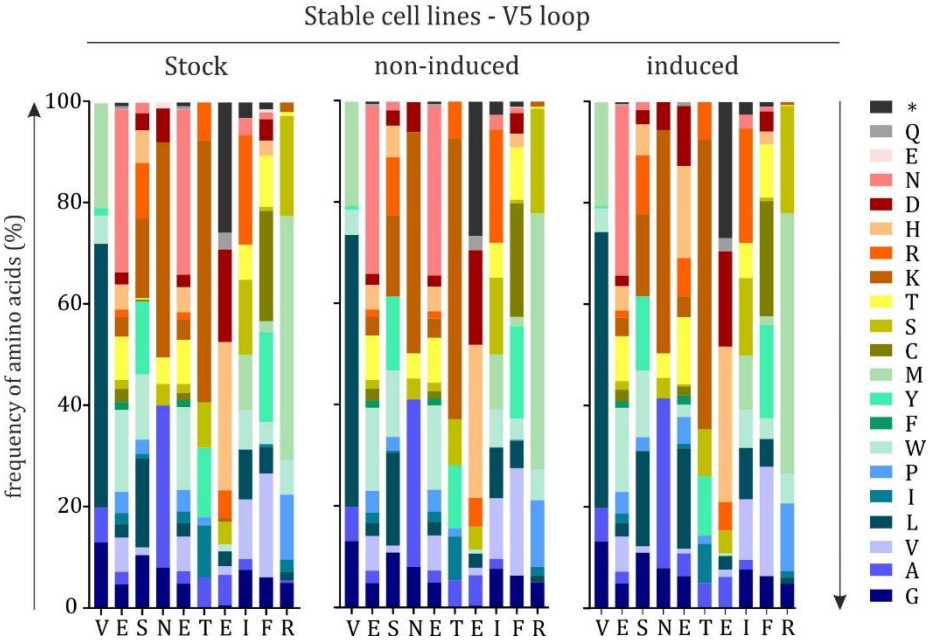


Figure S4 – Amino acid distribution in stable cell lines. Amino acid distribution of stock, non-induced and induced stable cell lines of the **(A)** CD4 binding loop, **(B)** bridging sheet and **(C)** V5 loop is shown. Height of the stacked bars (y-axis) represent the detected amino acids in percent at every pool-of-position (x-axis) in the SeqPer library. To provide a clearer overview, the respective wildtype amino acid at every pool-of-position was excluded from the data set.

A

Fold enrichment in high affinity gate P4 [a)]

Position	274	275	276	277	278	279	280	281	282	283	364	365	366	367	368
Aa	**S**	**E**	**N**	**L**	**T**	**N**	**N**	**V**	**K**	**T**	**P**	**A**	**G**	**G**	**D**
G	0,11	3,36	0,17	n.d.	0,49	n.d.	-1,64	-3,41	-0,06	n.d.	n.d.	n.d.	WT	WT	n.d.
A	n.d.	1,94	4,87	n.d.	1,26	0,74	n.d.	1,16	n.d.	1,34	-0,43	WT	n.d.	n.d.	0,00
V	n.d.	n.d.	n.d.	-1,96	-0,30	-0,80	n.d.	WT	n.d.	n.d.	1,06	n.d.	n.d.	n.d.	n.d.
L	n.d.	n.d.	n.d.	WT	n.d.	n.d.	n.d.	2,18	n.d.	-0,37	0,04	n.d.	n.d.	n.d.	n.d.
I	3,08	n.d.	n.d.	n.d.	-0,70	n.d.	n.d.	-1,52	n.d.	-0,83	-0,82	n.d.	n.d.	n.d.	n.d.
P	n.d.	n.d.	n.d.	0,52	-2,60	n.d.	n.d.	n.d.	n.d.	4,10	WT	n.d.	n.d.	n.d.	n.d.
W	n.d.	n.d.	n.d.	n.d.	3,06	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,00	-1,75	n.d.	n.d.
F	-0,09	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
Y	n.d.	-0,59	n.d.	n.d.	n.d.	n.d.	n.d.	-0,92	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,00
M	n.d.	n.d.	n.d.	-1,16	n.d.	n.d.	-1,68	3,74	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
C	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,83	n.d.	n.d.	0,00	0,55	5,70	n.d.
S	WT	0,65	n.d.	n.d.	-0,03	n.d.	0,61	n.d.	-0,58	n.d.	n.d.	0,00	-1,01	3,10	n.d.
T	1,89	n.d.	-2,25	n.d.	WT	-1,03	0,85	n.d.	2,66	WT	n.d.	0,00	n.d.	n.d.	n.d.
K	n.d.	5,40	-0,60	n.d.	n.d.	3,94	1,33	n.d.	WT	n.d.	0,00	n.d.	n.d.	-0,85	n.d.
R	2,64	n.d.	1,57	-2,89	0,92	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-0,01	n.d.
H	-0,37	n.d.	2,27	n.d.	1,41	4,26	n.d.	3,84	n.d.	n.d.	2,68	n.d.	n.d.	n.d.	n.d.
D	n.d.	0,81	n.d.	n.d.	0,17	n.d.	n.d.	2,44	n.d.	n.d.	-2,21	0,00	6,41	3,29	WT
N	-0,61	n.d.	WT	n.d.	-2,97	WT	WT	-1,83	1,91	0,13	n.d.	n.d.	n.d.	n.d.	n.d.
E	n.d.	WT	n.d.	n.d.	n.d.	n.d.	n.d.	4,56	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,00
Q	n.d.	n.d.	n.d.	1,77	-0,81	n.d.	n.d.	-0,32	2,01	-0,33	n.d.	n.d.	n.d.	-0,12	n.d.

B

Fold enrichment in high affinity gate P4[a)]

Position	274	275	276	277	278	279	280	281	282	283	364	365	366	367	368	369	370	371	372	373
Aa	S	E	N	L	T	N	N	V	K	T	P	A	G	G	D	L	E	I	T	T
G	0,85	n.d.	n.d.	n.d.	0,95	n.d.	-0,31	-4,21	0,05	n.d.	n.d.	n.d.	WT	WT	n.d.	n.d.	1,38	n.d.	n.d.	n.d.
A	n.d.	-0,08	n.d.	n.d.	-4,71	n.d.	n.d.	-0,24	n.d.	n.d.	n.d.	WT	n.d.	n.d.	-0,16	n.d.	n.d.	n.d.	4,50	3,56
V	n.d.	n.d.	n.d.	-0,50	1,74	0,87	n.d.	WT	n.d.	n.d.	0,35	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
L	n.d.	n.d.	n.d.	WT	n.d.	n.d.	-1,08	-0,14	n.d.	-0,55	n.d.	n.d.	n.d.	n.d.	n.d.	WT	n.d.	n.d.	n.d.	n.d.
I	1,66	n.d.	n.d.	n.d.	0,09	n.d.	n.d.	1,93	n.d.	-0,05	0,72	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	WT	4,59	5,69
P	n.d.	n.d.	n.d.	0,29	-3,91	n.d.	n.d.	-0,84	n.d.	0,27	WT	n.d.	n.d.	n.d.	n.d.	0,62	n.d.	n.d.	-1,92	1,48
W	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,00	-1,73	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
F	0,01	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
Y	n.d.	-0,82	n.d.	n.d.	n.d.	n.d.	n.d.	-0,03	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	1,22	n.d.	n.d.	n.d.	n.d.	n.d.
M	n.d.	n.d.	n.d.	0,41	n.d.	n.d.	-2,32	-0,22	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	1,21	n.d.	n.d.	n.d.	n.d.
C	n.d.	-0,41	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,21	n.d.	n.d.	0,00	0,39	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
S	WT	0,38	n.d.	n.d.	2,32	n.d.	0,74	n.d.	0,70	n.d.	-0,25	0,00	-0,62	0,78	n.d.	n.d.	n.d.	n.d.	-1,18	n.d.
T	-1,78	n.d.	-0,39	n.d.	WT	0,41	0,35	n.d.	0,07	WT	3,40	0,00	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	WT	WT
K	n.d.	-0,31	1,59	n.d.	-4,25	-0,67	1,44	n.d.	WT	n.d.	n.d.	n.d.	n.d.	-0,11	n.d.	n.d.	3,64	n.d.	n.d.	n.d.
R	1,25	n.d.	1,24	0,73	5,77	n.d.	n.d.	n.d.	-0,19	n.d.	n.d.	n.d.	n.d.	0,38	n.d.	1,60	n.d.	n.d.	n.d.	n.d.
H	0,52	n.d.	0,34	-0,81	3,31	2,38	0,41	n.d.	n.d.	n.d.	-0,12	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
D	n.d.	3,78	n.d.	n.d.	n.d.	n.d.	n.d.	4,92	n.d.	n.d.	-2,38	0,00	4,39	0,77	WT	n.d.	n.d.	n.d.	n.d.	n.d.
N	0,75	6,46	WT	n.d.	-5,27	WT	WT	-1,32	-0,17	0,28	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	4,96	5,21
E	n.d.	WT	n.d.	n.d.	n.d.	0,56	n.d.	0,46	0,75	n.d.	n.d.	n.d.	n.d.	n.d.	-1,09	n.d.	WT	n.d.	n.d.	n.d.
Q	n.d.	n.d.	n.d.	1,27	6,56	n.d.	n.d.	2,15	-0,28	1,86	n.d.	n.d.	n.d.	-0,14	n.d.	-1,42	2,16	n.d.	n.d.	n.d.

log$_2$-fold enrichment

depletion

...

Fold enrichment in high affinity gate P4

374	424	425	426	427	428	429	430	431	432	433	434	435	436	438	439	460	461	462	463	464	465	466	467	468	469
H	**I**	**N**	**M**	**W**	**Q**	**E**	**V**	**G**	**R**	**A**	**M**	**Y**	**A**	**P**	**P**	**V**	**E**	**S**	**N**	**E**	**T**	**E**	**I**	**F**	**R**
n.d.	n.d.	n.d.	-0,88	1,39	n.d.	-0,46	1,18	WT	n.d.	n.d.	0,36	n.d.	3,66	-0,22	n.d.	-0,01	0,01	0,01	-0,01	0,03	n.d.	-0,64	0,00	0,00	-0,03
n.d.	n.d.	n.d.	-0,09	n.d.	n.d.	0,24	-0,25	0,50	n.d.	WT	n.d.	n.d.	WT	-0,87	n.d.	-0,02	-0,01	n.d.	0,01	-0,57	-0,10	0,00	-0,06	n.d.	n.d.
n.d.	-2,63	n.d.	2,00	n.d.	n.d.	1,60	WT	0,06	n.d.	0,54	-4,59	-1,92	-1,04	n.d.	n.d.	WT	0,01	-0,10	n.d.	3,39	n.d.	-0,12	0,00	0,01	n.d.
n.d.	1,09	-1,25	-0,26	1,49	n.d.	-0,55	1,17	n.d.	n.d.	-0,67	0,06	-0,54	-2,19	n.d.	-3,74	0,01	0,00	0,01	n.d.	-2,52	n.d.	-0,02	0,00	0,01	-0,13
n.d.	WT	-4,76	0,48	1,75	n.d.	n.d.	n.d.	n.d.	0,97	n.d.	0,66	2,17	0,27	n.d.	-0,94	n.d.	-0,07	n.d.	n.d.	1,32	-0,17	n.d.	WT	n.d.	-0,50
3,95	n.d.	0,62	0,16	2,59	-0,69	-0,35	-0,30	-0,99	-1,30	n.d.	n.d.	-0,02	-1,44	WT	WT	n.d.	0,00	-0,04	n.d.	0,10	0,00	n.d.	n.d.	n.d.	0,01
n.d.	n.d.	n.d.	n.d.	WT	0,21	1,03	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-0,09	0,01	0,00	n.d.	3,19	n.d.	-1,09	-0,02	-0,04	-0,07
n.d.	1,08	n.d.	n.d.	n.d.	n.d.	n.d.	-0,08	n.d.	n.d.	n.d.	n.d.	-5,67	0,08	n.d.	-2,03	n.d.	-0,12	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	WT	n.d.
6,28	n.d.	-0,19	n.d.	n.d.	n.d.	0,70	n.d.	0,66	n.d.	n.d.	n.d.	WT	0,01	n.d.	n.d.	-2,05	n.d.	0,00	n.d.	n.d.	-0,08	n.d.	n.d.	0,01	n.d.
n.d.	n.d.	n.d.	WT	n.d.	n.d.	-1,55	-0,39	n.d.	n.d.	n.d.	WT	n.d.	4,14	n.d.	-1,28	-0,01	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-0,01	-0,12	0,02
n.d.	n.d.	n.d.	n.d.	1,36	n.d.	-1,12	n.d.	0,80	n.d.	n.d.	n.d.	-0,49	n.d.	n.d.	n.d.	n.d.	-0,06	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,01	n.d.
n.d.	n.d.	0,60	-2,72	-0,86	n.d.	-1,38	n.d.	1,22	0,11	0,22	0,42	1,05	0,99	-0,47	0,35	n.d.	-0,01	WT	-0,03	2,43	0,01	0,00	0,01	-0,10	0,02
n.d.	-2,96	0,26	0,83	n.d.	n.d.	-0,21	n.d.	-1,64	0,13	1,80	2,20	n.d.	0,17	0,84	1,21	n.d.	0,01	n.d.	-0,04	-0,17	WT	n.d.	-0,01	0,01	n.d.
n.d.	0,39	n.d.	n.d.	n.d.	1,50	0,52	n.d.	-0,24	0,29	n.d.	n.d.	n.d.	n.d.	-0,52	-0,39	n.d.	-0,02	0,00	0,01	0,20	0,05	n.d.	n.d.	n.d.	-1,54
3,16	0,13	-0,57	n.d.	1,52	-1,57	1,21	n.d.	2,97	WT	n.d.	2,56	n.d.	0,45	n.d.	-0,68	n.d.	-0,02	0,01	n.d.	-2,07	0,00	0,01	0,01	n.d.	WT
WT	n.d.	0,39	n.d.	n.d.	0,94	-0,01	0,89	0,90	0,69	n.d.	n.d.	1,67	n.d.	0,70	2,61	n.d.	-0,01	-0,03	n.d.	-1,55	n.d.	0,02	n.d.	-0,07	n.d.
n.d.	n.d.	0,89	n.d.	n.d.	2,59	0,51	n.d.	0,93	n.d.	n.d.	n.d.	-0,63	n.d.	n.d.	n.d.	n.d.	-0,06	-0,09	-0,09	-2,16	n.d.	0,01	n.d.	-0,04	n.d.
3,54	n.d.	WT	n.d.	n.d.	n.d.	-0,50	n.d.	n.d.	-0,57	n.d.	n.d.	0,89	-0,64	n.d.	n.d.	n.d.	0,02	-0,19	WT	n.d.	n.d.	n.d.	-0,09	-0,26	n.d.
n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	WT	1,32	-1,35	n.d.	n.d.	0,97	n.d.	n.d.	1,26	0,64	n.d.	WT	n.d.	n.d.	WT	n.d.	WT	n.d.	n.d.	n.d.
n.d.	n.d.	n.d.	n.d.	n.d.	WT	0,28	n.d.	-0,27	n.d.	-0,03	n.d.	n.d.	-0,77	n.d.	-1,67	n.d.	-0,95	n.d.	n.d.	-2,20	n.d.	-0,13	n.d.	n.d.	n.d.

Figure S5 - Enrichment rates from gate P4 (gain of binding). (A) 15 amino acid positions and **(B)** the whole CD4 binding site (46 pools of position) were analyzed by FACS sorting and subsequent NGS. The enrichment is represented as log2-fold change compared to the input sample that was withdrawn prior to the selection. The extent of enrichment or depletion is represented in the intensity of green (enrichment) or red (depletion) Wild type (WT) was excluded from the calculation. Variants that could not be detected in the sorting were referred to as n.d. (not detected).

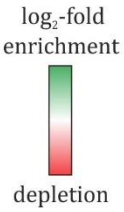
a) calculations were performed according to the formula from [204]:

$$\textit{fold enrichment} = \left(\frac{n\ (G)\ \textit{detected}}{n\ (G)\ \textit{reference}}\right) / \left(\frac{n\ (\textit{non-}G)\ \textit{detected}}{n\ (\textit{non-}G)\ \textit{reference}}\right)$$ (on the example of glycine enrichment)

A

Fold enrichment in low affinity gate P5 [a)]

N(Aa position)	1	5				
Position	278	274	275	276	277	278
Aa	**T**	**S**	**E**	**N**	**L**	**T**
G	-0,38	1,88	-3,32	-0,02	4,99	-0,51
A	1,59	n.d.	0,95	3,27	1,38	3,33
V	-0,17	3,79	3,91	0,85	-0,80	1,10
L	2,04	2,31	2,33	2,64	WT	n.d.
I	0,31	n.d.	n.d.	3,63	n.d.	0,85
P	-0,02	n.d.	n.d.	n.d.	-1,52	0,35
W	-1,45	5,22	n.d.	-3,06	1,50	n.d.
F	n.d.	-1,51	8,37	-2,51	n.d.	n.d.
Y	-0,01	1,79	-0,36	n.d.	5,11	n.d.
M	n.d.	n.d.	n.d.	n.d.	-1,16	n.d.
C	2,24	n.d.	0,50	n.d.	n.d.	4,95
S	-0,91	WT	-2,97	4,35	n.d.	-1,91
T	WT	0,68	5,59	-2,17	3,66	WT
K	3,76	0,26	5,99	0,74	5,75	n.d.
R	-0,76	3,52	n.d.	-0,07	0,06	1,07
H	4,21	-2,26	n.d.	1,61	-0,26	-3,38
D	0,86	n.d.	3,87	-0,59	n.d.	0,32
N	4,00	n.d.	0,21	WT	4,02	1,53
E	3,78	n.d.	WT	n.d.	4,53	n.d.
Q	-2,22	n.d.	n.d.	n.d.	n.d.	n.d.



B

Fold enrichment in low affinity gate P5 [a)]

Position	274	275	276	277	278	279	280	281	282	283	364	365	366	367	368
Aa	**S**	**E**	**N**	**L**	**T**	**N**	**N**	**V**	**K**	**T**	**P**	**A**	**G**	**G**	**D**
G	-0,35	-2,73	0,17	n.d.	n.d.	n.d.	0,52	-0,07	-0,03	n.d.	n.d.	n.d.	WT	WT	n.d.
A	n.d.	-3,64	n.d.	n.d.	1,10	-2,23	n.d.	-0,46	n.d.	-0,72	-1,54	WT	n.d.	n.d.	0,14
V	n.d.	n.d.	n.d.	-0,15	n.d.	1,55	0,54	WT	n.d.	n.d.	-0,65	n.d.	n.d.	n.d.	n.d.
L	-1,34	-0,16	-1,60	WT	n.d.	n.d.	0,60	-1,13	n.d.	0,25	0,10	n.d.	n.d.	n.d.	n.d.
I	n.d.	n.d.	0,72	n.d.	n.d.	n.d.	-0,27	-3,59	n.d.	-0,05	0,33	n.d.	n.d.	n.d.	n.d.
P	n.d.	n.d.	n.d.	-0,38	-0,87	n.d.	n.d.	-1,08	n.d.	-0,69	WT	n.d.	n.d.	n.d.	n.d.
W	n.d.	n.d.	-1,46	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	1,11	n.d.	0,00	-1,38	n.d.	n.d.
F	-0,54	2,23	0,86	n.d.	n.d.	n.d.	2,28	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
Y	0,82	0,58	n.d.	n.d.	n.d.	n.d.	1,14	0,34	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-1,75
M	n.d.	n.d.	n.d.	1,33	n.d.	n.d.	1,15	2,01	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.
C	n.d.	0,99	n.d.	n.d.	n.d.	n.d.	-2,17	n.d.	0,69	n.d.	n.d.	0,00	0,17	n.d.	n.d.
S	WT	-2,05	n.d.	n.d.	-2,68	n.d.	-0,80	n.d.	1,32	-3,15	n.d.	0,00	0,34	-0,62	n.d.
T	0,77	n.d.	-0,15	n.d.	WT	-0,51	-1,62	n.d.	-1,51	WT	n.d.	0,00	n.d.	n.d.	n.d.
K	0,24	0,90	-1,20	n.d.	n.d.	-0,08	0,44	n.d.	WT	0,83	-1,13	n.d.	n.d.	-0,09	n.d.
R	0,65	n.d.	-0,56	-0,16	n.d.	n.d.	n.d.	n.d.	0,24	n.d.	n.d.	n.d.	n.d.	-0,50	n.d.
H	0,07	n.d.	n.d.	-0,10	3,59	n.d.	n.d.	n.d.	n.d.	n.d.	1,60	n.d.	n.d.	n.d.	n.d.
D	n.d.	0,75	-5,41	n.d.	-3,07	2,30	0,65	n.d.	n.d.	n.d.	-0,17	0,00	1,55	-0,76	WT
N	1,68	-2,73	WT	-3,38	1,20	WT	WT	0,16	0,69	-1,77	n.d.	n.d.	n.d.	n.d.	n.d.
E	n.d.	WT	n.d.	0,77	n.d.	-2,71	n.d.	3,69	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	2,59
Q	n.d.	n.d.	n.d.	3,62	n.d.	n.d.	n.d.	-1,33	-2,37	-0,78	n.d.	n.d.	n.d.	0,09	n.d.

[a)] calculations were performed according to the formula from [204]:

$$\textit{fold enrichment} = \left(\frac{n\,(G)\,\textit{detected}}{n\,(G)\,\textit{reference}}\right) / \left(\frac{n\,(\textit{non-G})\,\textit{detected}}{n\,(\textit{non-G})\,\textit{reference}}\right)$$ (on the example of glycine enrichment)

C

Fold enrichment in low affinity gate P5 [a)]

Position	274	275	276	277	278	279	280	281	282	283	364	365	366	367	368	369	370	371	372	373	374
Aa	**S**	**E**	**N**	**L**	**T**	**N**	**N**	**V**	**K**	**T**	**P**	**A**	**G**	**G**	**D**	**L**	**E**	**I**	**T**	**T**	**H**
G	-3,23	n.d.	n.d.	n.d.	n.d.	n.d.	0,31	-0,71	-0,11	n.d.	n.d.	n.d.	WT	WT	n.d.	n.d.	-0,42	1,25	-0,63	1,31	n.d.
A	n.d.	-2,52	n.d.	n.d.	-2,22	n.d.	n.d.	-0,16	n.d.	n.d.	n.d.	WT	n.d.	n.d.	-0,19	-2,01	n.d.	n.d.	0,67	1,05	n.d.
V	n.d.	n.d.	n.d.	-2,67	n.d.	1,16	n.d.	WT	n.d.	2,47	-0,46	n.d.	n.d.	n.d.	n.d.	n.d.	-0,74	n.d.	-0,78	-2,27	2,42
L	-2,12	n.d.	n.d.	WT	n.d.	n.d.	-0,06	n.d.	n.d.	0,18	n.d.	n.d.	n.d.	n.d.	n.d.	WT	-1,44	n.d.	-2,94	-0,14	-2,50
I	n.d.	n.d.	n.d.	n.d.	-0,65	n.d.	n.d.	n.d.	n.d.	0,42	-0,32	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	WT	1,28	0,32	1,49
P	n.d.	n.d.	n.d.	-0,35	-2,58	n.d.	n.d.	n.d.	n.d.	-7,10	WT	n.d.	n.d.	n.d.	n.d.	1,09	-3,41	n.d.	-0,37	n.d.	1,68
W	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-1,89	-1,96	n.d.	n.d.	n.d.	0,47	n.d.	-0,47	n.d.	n.d.
F	-0,66	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-0,93	-1,39	n.d.	-2,61
Y	n.d.	0,10	n.d.	n.d.	n.d.	n.d.	n.d.	0,57	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,86	n.d.	n.d.	n.d.	n.d.	-2,37	3,63
M	n.d.	n.d.	n.d.	1,55	n.d.	n.d.	-1,97	0,13	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-0,54	-0,97	n.d.	n.d.	n.d.	n.d.
C	n.d.	1,26	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-3,17	n.d.	n.d.	n.d.	0,20	n.d.	n.d.	n.d.	1,32	n.d.	-1,45	n.d.	n.d.
S	WT	-2,48	n.d.	n.d.	-2,80	n.d.	-0,11	n.d.	n.d.	n.d.	n.d.	2,03	0,46	-0,43	1,73	n.d.	-2,31	n.d.	-1,62	-0,37	-2,95
T	-3,73	n.d.	0,34	n.d.	WT	-1,86	-1,00	n.d.	n.d.	WT	3,11	2,63	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	WT	WT	n.d.
K	n.d.	-0,92	0,18	n.d.	n.d.	-0,37	-3,34	n.d.	WT	n.d.	n.d.	n.d.	n.d.	0,33	n.d.	n.d.	-1,78	n.d.	n.d.	n.d.	0,60
R	1,51	n.d.	n.d.	-0,34	n.d.	n.d.	n.d.	n.d.	1,90	n.d.	n.d.	n.d.	n.d.	0,13	n.d.	-1,98	-0,43	n.d.	-1,08	n.d.	1,89
H	0,62	n.d.	n.d.	n.d.	5,32	n.d.	-0,43	n.d.	n.d.	n.d.	0,98	n.d.	n.d.	n.d.	n.d.	n.d.	-1,16	n.d.	n.d.	n.d.	WT
D	n.d.	0,98	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-0,52	6,20	1,54	-0,63	WT	n.d.	n.d.	n.d.	0,56	n.d.	n.d.
N	3,23	n.d.	WT	n.d.	1,69	WT	WT	0,61	0,85	0,97	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,42	1,47	-1,77
E	n.d.	WT	n.d.	n.d.	n.d.	-0,28	n.d.	0,80	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,63	n.d.	WT	n.d.	n.d.	1,79	n.d.
Q	n.d.	n.d.	n.d.	2,57	n.d.	n.d.	n.d.	-1,22	-1,57	-0,71	n.d.	n.d.	n.d.	-0,12	n.d.	1,52	-0,85	n.d.	n.d.	n.d.	-0,38

log$_2$-fold enrichment ... depletion

Fold enrichment in low affinity gate P5 [a)]

424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	460	461	462	463	464	465	466	467	468	469
I	**N**	**M**	**W**	**Q**	**E**	**V**	**G**	**R**	**A**	**M**	**Y**	**A**	**P**	**P**	**V**	**E**	**S**	**N**	**E**	**T**	**E**	**I**	**F**	**R**
n.d.	-0,37	-0,54	-1,61	n.d.	0,44	-0,21	WT	n.d.	-1,60	-3,71	n.d.	n.d.	-0,95	n.d.	-0,97	1,33	0,06	n.d.	1,36	n.d.	3,66	-0,22	0,94	-0,93
n.d.	0,08	n.d.	n.d.	n.d.	1,54	-0,58	-1,04	n.d.	WT	n.d.	-1,83	WT	-1,99	n.d.	0,86	-1,56	n.d.	-2,36	-2,11	0,07	-4,54	-3,23	n.d.	n.d.
-1,69	-0,65	2,74	n.d.	-4,78	-0,23	WT	1,35	n.d.	0,02	-1,77	-0,61	0,36	n.d.	n.d.	WT	0,83	n.d.	n.d.	4,22	n.d.	n.d.	-1,14	-1,93	n.d.
0,54	1,73	-0,74	0,24	n.d.	-2,47	2,65	-0,73	n.d.	-1,45	-0,81	0,55	-0,07	n.d.	0,97	-1,49	2,05	-1,21	n.d.	-0,47	n.d.	-2,85	-0,92	-1,37	0,79
WT	1,19	1,71	-6,20	n.d.	n.d.	n.d.	0,38	2,27	n.d.	0,36	n.d.	0,35	n.d.	n.d.	n.d.	4,09	n.d.	n.d.	5,48	1,41	n.d.	WT	n.d.	-1,80
n.d.	2,01	n.d.	0,53	0,36	-0,89	-2,26	0,32	-1,37	n.d.	n.d.	n.d.	-0,13	WT	WT	n.d.	-2,70	1,19	n.d.	-2,61	-1,75	n.d.	n.d.	n.d.	-0,64
n.d.	n.d.	n.d.	WT	n.d.	-0,48	-3,55	n.d.	n.d.	2,00	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-3,07	n.d.	n.d.	0,11	n.d.	n.d.	-0,09	1,83	1,24
-0,22	0,68	n.d.	n.d.	n.d.	n.d.	0,27	n.d.	n.d.	n.d.	n.d.	0,71	0,09	n.d.	-5,36	n.d.	0,72	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	WT	n.d.
n.d.	0,85	n.d.	n.d.	n.d.	n.d.	n.d.	-0,48	n.d.	n.d.	n.d.	WT	-0,56	n.d.	n.d.	n.d.	n.d.	-0,22	n.d.	n.d.	n.d.	n.d.	n.d.	-2,68	n.d.
n.d.	2,32	WT	n.d.	n.d.	-6,49	-0,60	n.d.	n.d.	n.d.	WT	n.d.	3,76	n.d.	-0,79	2,66	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-0,20
n.d.	2,66	n.d.	1,15	n.d.	-2,61	n.d.	1,48	n.d.	n.d.	n.d.	0,52	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	0,74	n.d.
n.d.	0,68	-0,86	-0,07	n.d.	-0,22	-1,18	1,62	1,04	-0,60	-1,79	1,00	-0,49	-0,75	-0,14	n.d.	-1,49	WT	1,08	0,96	1,19	n.d.	1,20	1,31	0,20
0,29	-1,26	-0,49	n.d.	1,38	-3,63	-4,42	n.d.	0,36	3,73	n.d.	n.d.	0,18	1,88	1,97	n.d.	-2,65	n.d.	0,53	-2,84	WT	n.d.	-1,31	0,84	3,47
n.d.	0,05	n.d.	n.d.	-0,50	-0,95	n.d.	-2,79	n.d.	n.d.	n.d.	n.d.	n.d.	0,18	-0,67	n.d.	-1,37	1,69	1,06	-1,14	0,02	n.d.	n.d.	n.d.	n.d.
0,11	-0,43	-1,52	-0,21	-0,06	-0,62	n.d.	1,13	WT	n.d.	1,99	0,25	0,17	n.d.	0,20	n.d.	-2,59	0,51	n.d.	-4,65	n.d.	1,47	-0,75	n.d.	WT
n.d.	0,54	n.d.	n.d.	-0,32	-1,53	n.d.	0,26	-0,74	n.d.	n.d.	-0,18	n.d.	1,92	2,35	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-1,22	n.d.	0,72	n.d.
n.d.	-2,29	n.d.	n.d.	n.d.	0,46	0,43	1,91	n.d.	n.d.	n.d.	-0,80	n.d.	n.d.	-0,58	n.d.	-2,89	-0,67	-4,50	-4,99	n.d.	0,48	n.d.	-1,23	n.d.
n.d.	WT	n.d.	n.d.	n.d.	-0,46	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-0,31	7,87	n.d.	n.d.	-1,77	n.d.	WT	n.d.	n.d.	n.d.	2,82	n.d.	n.d.
n.d.	n.d.	n.d.	n.d.	n.d.	WT	4,17	0,93	n.d.	-4,72	-2,58	n.d.	n.d.	0,16	-0,69	n.d.	WT	n.d.	n.d.	WT	n.d.	WT	n.d.	n.d.	n.d.
n.d.	-4,50	n.d.	n.d.	WT	0,12	n.d.	-2,18	n.d.	2,24	n.d.	-0,36	0,09	n.d.	-1,81	n.d.	n.d.	n.d.	n.d.	n.d.	n.d.	-1,01	n.d.	n.d.	n.d.

Figure S6 - Enrichment rates from low affinity gate P5 . (A) One & Five, **(B)** 15 and **(C)** the whole CD4 binding site (46 pools of position) were analyzed by FACS sorting and subsequent NGS. The enrichment is represented as log2-fold change compared to the input sample that was withdrawn prior to the selection. The extent of enrichment or depletion is represented in the intensity of green (enrichment) or red (depletion) Wild type (WT) was excluded from the calculation. Variants that could not be detected in the sorting were referred to as n.d. (not detected).

a) calculations were performed according to the formula from [204]:

$$\textit{fold enrichment} = \left(\frac{n\,(G)\ \textit{detected}}{n\,(G)\ \textit{reference}}\right) / \left(\frac{n\,(\textit{non-G})\ \textit{detected}}{n\,(\textit{non-G})\ \textit{reference}}\right)$$ (on the example of glycine enrichment)

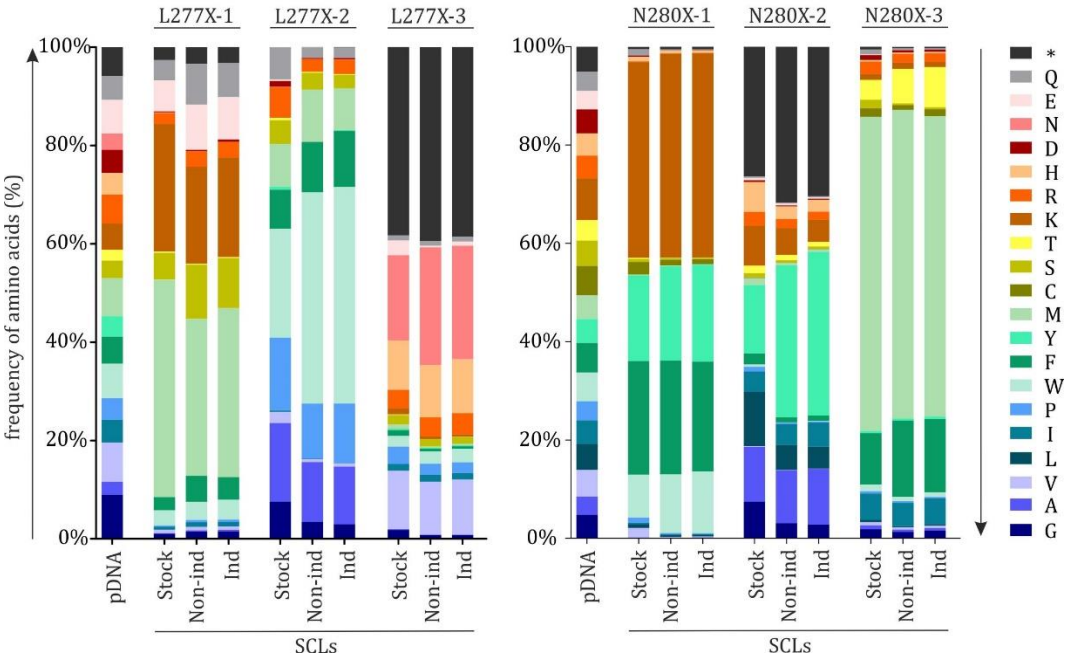


Figure S7 –Amino acid distribution of stable cell lines L275X and N280X. Stock, non-induced and induced samples of SCL triplicates were compared to the respective pDNA utilized during the generation of the stable cell lines. Height of the stacked bars represents detected amino acids in percent (y-axis).

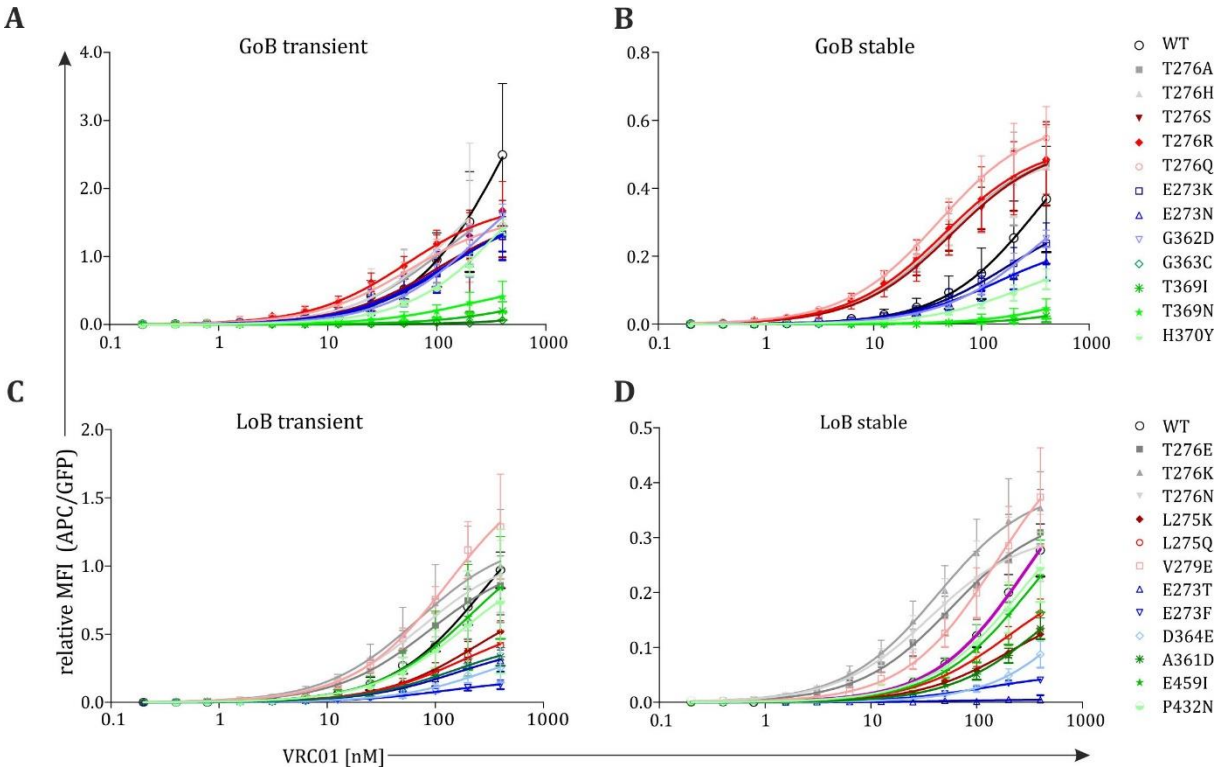


Figure S8 – FACS equilibrium titration curves of GoB and LoB variants. Relative mean fluorescence intensities of **(A+C)** transiently transfected variants and their respective **(B+D)** induced stable cell lines

were compared with 16055 gp145 WT. Signals were normalized on the fluorescence of GFP. Notably, none of the variants displayed distinct saturation.

Table S2 – Statistics of FACS equilibrium titrations. Statistical significance of transiently transfected GoB and LoB variants and their respective stable cell lines was determined with excel using a two-sided, homoscedastic t-test and corrected according to Bonferroni to adjust for multiple testing. Variants that exhibited p-values <0.0042 were considered significant and are indicated by asterisks.

GoBs	T276A	T276H	T276Q	T276R	T276S	E273K	G362D	G363C	E273N	T369I	T369N	H370Y
transient	0,61	0,33	0,44	0,65	0,31	0,25	0,33	0,02	0,26	0,02	0,03	0,17
stable	0,14	0,94	0,02	0,40	0,41	0,04	0,92	0,0004*	0,01	0,01	0,35	0,02

LoBs	T276E	T276K	T276N	E273F	E273T	L275K	L275Q	V279E	D364E	A361D	P432N	E459I
transient	0,80	0,36	0,48	0,0003*	0,0021*	0,01	0,004 *	0,06	0,0012*	0,0012*	0,58	0,75
stable	0,21	0,04	0,18	0,001 *	0,0005*	0,004	0,02	0,11	0,002 *	0,01	0,51	0,31

8.4 References

1. M.S. Gottlieb, M.D., Robert Schroff, Ph.D., Howard M. Schanker, M.D., Joel D. Weisman, D.O., Peng Thim Fan, M.D., Robert A. Wolf, M.D., and Andrew Saxon, M. D. Pneumocystis carinii Pneumonia and Mucosal Candidiasis in Previously Healthy Homosexual Men — Evidence of a New Acquired Cellular Immunodeficiency. *N Engl J Med* 1425–1431 (1981). doi:10.1056/NEJM198112103052401

2. Hymes, K. *et al.* Kaposi'S Sarcoma in Homosexual Men—a Report of Eight Cases. *Lancet* **318,** 598–600 (1981).

3. Jeffrey Vieira, M.D., Elliot Frank, M.D., Thomas J. Spira, M.D., and Sheldon H. Landesman, M. D. Acquired Immune Deficiency in Haitians — Opportunistic Infections in Previously Healthy Haitian Immigrants. *N Engl J Med* **308,** 125–129 (1983).

4. Small, C. B. *et al.* Community-acquired opportunistic infections and defective cellular immunity in heterosexual drug abusers and homosexual men. *Am. J. Med.* **74,** 433–441 (1983).

5. Gill, S. K., Loveday, C. & Gilson, R. J. Transmission of HIV-1 infection by oroanal intercourse. *Genitourin. Med.* **68,** 254–7 (1992).

6. Busch, M. *et al.* Risk of human immunodeficiency virus (HIV) transmission by blood transfusions before the implementation of HIV-1 antibody screening. The Transfusion Safety Study Group. *Transfusion* **31,** 4–11 (1991).

7. Baggaley, R. F., Boily, M.-C., White, R. G., Alary, M. & Baggaley, R. Risk of HIV-1 transmission for parenteral exposure and blood transfusion: a systematic review and meta-analysis. (2006).

8. Koup, R. A. *et al.* Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J. Virol.* **68,** 4650–5 (1994).

9. Hahn, B. H., Shaw, G. M., De Cock, K. M. & Sharp, P. M. AIDS as a zoonosis: scientific and public health implications. *Science* **287,** 607–14 (2000).

10. de Groot, N. G. & Bontrop, R. E. The HIV-1 pandemic: does the selective sweep in chimpanzees mirror humankind's future? *Retrovirology* **10,** 53 (2013).

11. Sharp, P., Shaw, G. & Hahn, B. Simian Immunodeficiency Virus Infection of Chimpanzees. *J. Virol.* **79,** 3891–3902 (2005).

12. Lemey, P. *et al.* Tracing the origin and history of the HIV-2 epidemic. *Proc. Natl. Acad. Sci.* **100,** 6588–6592 (2003).

13. Hizi, A. Fidelity of the reverse transcriptase of human immunodeficiency virus type 2. *Bakhanashvili M, Hizi A.* **306,** 151–156 (1992).

14. Richman, D. D., Wrin, T., Little, S. J. & Petropoulos, C. J. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 4144–9 (2003).

15. Barouch, D. H. *et al.* Eventual AIDS vaccine failure in a rhesus monkey by viral escape from cytotoxic T lymphocytes. *Nature*

415, 335–339 (2002).

16. Rhodes, T. D., Nikolaitchik, O., Chen, J., Powell, D. & Hu, W.-S. Genetic recombination of human immunodeficiency virus type 1 in one round of viral replication: effects of genetic distance, target cells, accessory genes, and lack of high negative interference in crossover events. *J. Virol.* **79,** 1666–77 (2005).

17. Gao, F. *et al.* Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature* **397,** 436–441 (1999).

18. Vallari, A. *et al.* Confirmation of putative HIV-1 group P in Cameroon. *J. Virol.* **85,** 1403–7 (2011).

19. Moore, J. P., Parren, P. W. H. I. & Burton, D. R. Genetic Subtypes , Humoral Immunity , and Human Immunodeficiency Virus Type 1 Vaccine Development. *J. Virol.* **75,** 5721–5729 (2001).

20. Peeters, M. *et al.* Characterization of a highly replicative intergroup M/O human immunodeficiency virus type 1 recombinant isolated from a Cameroonian patient. *J. Virol.* **73,** 7368–75 (1999).

21. Rousseau, C. M. *et al.* Extensive intrasubtype recombination in South African human immunodeficiency virus type 1 subtype C infections. *J. Virol.* **81,** 4492–4500 (2007).

22. Hemelaar, J. The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* **18,** 182–192 (2012).

23. Briggs, J. A. G., Wilk, T., Welker, R., Kräusslich, H. G. & Fuller, S. D. Structural organization of authentic, mature HIV-1 virions and cores. *EMBO J.* **22,** 1707–1715 (2003).

24. Lole, K. S. *et al.* Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J. Virol.* **73,** 152–60 (1999).

25. Sükösd, Z. *et al.* Full-length RNA structure prediction of the HIV-1 genome reveals a conserved core domain. *Nucleic Acids Res.* **43,** gkv1039 (2015).

26. Watts, J. M. *et al.* Genome. *Nature* **460,** 711–716 (2009).

27. Frankel, A. D. & Young, J. A. T. HIV-1: Fifteen Proteins and an RNA. *Annu. Rev. Biochem* **67,** 1–25 (1998).

28. Freed, E. O. HIV-1 assembly, release and maturation. *Nat Rev Microbiol* **13,** 484–496 (2015).

29. Saad, J. S. *et al.* Structural basis for targeting HIV-1 Gag proteins to the plasma membrane for virus assembly. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 11364–9 (2006).

30. Freed, E. O. HIV-1 gag proteins: diverse functions in the virus life cycle. *Virology* **251,** 1–15 (1998).

31. Campbell, E. M. & Hope, T. J. HIV-1 Capsid : The Multifaceted Key Player in HIV-1 infection. **13,** 471–483 (2016).

32. Chalovich, J. M. & Eisenberg, E. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature. 2013 May 30; 497(7451) 643–646.* **257,** 2432–2437 (2005).

33. Fiorentini, S., Giagulli, C., Caccuri, F., Magiera, A. K. & Caruso, A. HIV-1 matrix protein p17: A candidate antigen for therapeutic vaccines against AIDS. *Pharmacol. Ther.* **128,** 433–444 (2010).

34. Wu, Z. *et al.* Total chemical synthesis of N-myristoylated HIV-1 matrix protein p17: structural and mechanistic implications of p17 myristoylation. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 11587–92 (2004).

35. Davis, M. R. *et al.* A Mutation in the Human Immunodeficiency Virus Type 1 Gag Protein Destabilizes the Interaction of the Envelope Protein Subunits gp120 and gp41. *Society* **80,** 2405–2417 (2006).

36. Thomas, J. A. & Gorelick, R. J. Nucleocapsid protein function in early infection processes. *Sci. York* **134,** 39–63 (2009).

37. Checkley, M. A., Luttge, B. G. & Freed, E. O. HIV-1 envelope glycoprotein biosynthesis, trafficking, and incorporation. **410,** 582–608 (2012).

38. Demirov, D. G., Orenstein, J. M. & Freed, E. O. The Late Domain of Human Immunodeficiency Virus Type 1 p6 Promotes Virus Release in a Cell Type-Dependent Manner The Late Domain of Human Immunodeficiency Virus Type 1 p6 Promotes Virus Release in a Cell Type-Dependent Manner. *J. Virol.* **76,** 105–117 (2002).

39. Solbak, S. M. ??ie *et al.* HIV-1 p6 - A structured to flexible multifunctional membrane-interacting protein. *Biochim. Biophys. Acta - Biomembr.* **1828,** 816–823 (2013).

40. Hill, M. K., Hooker, C. W., Harrich, D., Crowe, S. M. & Mak, J. Gag-Pol Supplied in trans Is Efficiently Packaged and Supports Viral Function in Human Immunodeficiency Virus Type 1 Gag-Pol Supplied in trans Is Efficiently Packaged and Supports Viral Function in Human Immunodeficiency Virus Type 1. *Society* **75,** 6835–6840 (2001).

41. Arthos, J. *et al.* HIV-1 envelope protein binds to and signals through integrin alpha4beta7, the gut mucosal homing receptor for peripheral T cells. *Nat. Immunol.* **9,** 301–309 (2008).

42. Geijtenbeek, T. B. *et al.* DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. *Cell* **100,** 587–597 (2000).

43. Maddon, P. J. *et al.* The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain. *Cell*

47, 333–348 (1986).

44. Tran, E. E. H. *et al.* Structural mechanism of trimeric HIV-1 envelope glycoprotein activation. *PLoS Pathog.* **8,** 37 (2012).

45. Karn, J. & Stoltzfus, C. M. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb. Perspect. Med.* **2,** 1–17 (2012).

46. Das, A. T., Harwig, A. & Berkhout, B. The HIV-1 Tat protein has a versatile role in activating viral transcription. *J. Virol.* **85,** 9506–16 (2011).

47. Bour, S. & Strebel, K. The HIV-1 Vpu protein: A multifunctional enhancer of viral particle release. *Microbes Infect.* **5,** 1029–1039 (2003).

48. Le Noury, D. A., Mosebi, S., Papathanasopoulos, M. A. & Hewer, R. Functional roles of HIV-1 Vpu and CD74: Details and implications of the Vpu-CD74 interaction. *Cell. Immunol.* **298,** 25–32 (2015).

49. Ru, E., Grivel, J., Mu, J., Kirchhoff, F. & Margolis, L. Vpr and Vpu Are Important for Efficient Human Immunodeficiency Virus Type 1 Replication and CD4 ¿ T-Cell Depletion in Human Lymphoid Tissue Ex Vivo. *Society* **78,** 12689–12693 (2004).

50. Pitcher, C. J. *et al.* HIV-1-specific CD4+ T cells are detectable in most individuals with active HIV-1 infection, but decline with prolonged viral suppression. *Nat. Med.* **5,** 518–525 (1999).

51. Douek, D. C. *et al.* HIV preferentially infects HIV-specific CD4+ T cells. *Nature* **417,** 95–98 (2002).

52. Saphire, A. C. S. *et al.* Syndecans Serve as Attachment Receptors for Human Immunodeficiency Virus Type 1 on Macrophages. *J. Virol.* **75,** 9187–9200 (2001).

53. McDougal, J. ., Kennedy, M. ., Sligh, J.M, Cort, S. ., Mawle, A. & Nocholson, J. K. . Binding of HTLV-III/IAV to T4+ T Cells by a Complex of tie 11OK Vira Pmten and the T4 Molecule J. *Science (80-.).* **231,** 382–385 (1986).

54. Popik, W., Hesselgesser, J. E. & Pitha, P. M. Binding of human immunodeficiency virus type 1 to CD4 and CXCR4 receptors differentially regulates expression of inflammatory genes and activates the MEK/ERK signaling pathway. *J Virol* **72,** 6406–6413 (1998).

55. Vacik, J., Dean, B. S., Zimmer, W. E. & Dean, D. A. Cell-specific nuclear import of plasmid DNA. *Gene Ther.* **6,** 1006–1014 (1999).

56. Sullivan, N. *et al.* CD4-Induced conformational changes in the human immunodeficiency virus type 1 gp120 glycoprotein: consequences for virus entry and neutralization. *J. Virol.* **72,** 4694–703 (1998).

57. Sattentau, Q. J., Moore, J. P., Vignaux, F., Traincard, F. & Poignard, P. Conformational changes induced in the envelope glycoproteins of human and simian immunodeficiency virus by soluble receptor binding. *J. Virol.* **64,** 7383 (1993).

58. Choe, H. *et al.* The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. *Cell* **85,** 1135–1148 (1996).

59. Cashin, K. *et al.* Linkages between HIV-1 specificity for CCR5 or CXCR4 and in vitro usage of alternative coreceptors during progressive HIV-1 subtype C infection. *Retrovirology* **10,** 98 (2013).

60. Weissenhorn, W. *et al.* Atomic structure of the ectodomain from HIV-1 gp41. *Nature* **387,** 426–430 (1997).

61. Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. Core Structure of gp41 from the HIV Envelope Glycoprotein. *Cell* **89,** 263–273 (1997).

62. Buzon, V. *et al.* Crystal structure of HIV-1 gp41 including both fusion peptide and membrane proximal external regions. *PLoS Pathog.* **6,** 1–7 (2010).

63. Lori, F. *et al.* Viral DNA carried by human immunodeficiency virus type 1 virions. *J Virol* **66,** 5067–5074 (1992).

64. Bukrinsky, M. I. *et al.* Active nuclear import of human immunodeficiency virus type 1 preintegration complexes. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 6580–6584 (1992).

65. Gallay, P., Hope, T., Chin, D. & Trono, D. HIV-1 infection of nondividing cells through the recognition of integrase by the importin/karyopherin pathway. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 9825–30 (1997).

66. Ivanchenko, S. *et al.* Dynamics of HIV-1 assembly and release. *PLoS Pathog.* **5,** (2009).

67. Murakami, T., Ablan, S., Freed, E. O. & Tanaka, Y. Regulation of human immunodeficiency virus type 1 Env-mediated membrane fusion by viral protease activity. *J. Virol.* **78,** 1026–1031 (2004).

68. Wyma, D. J. *et al.* Coupling of Human Immunodeficiency Virus Type 1 Fusion to Virion Maturation : a Novel Role of the gp41 Cytoplasmic Tail Coupling of Human Immunodeficiency Virus Type 1 Fusion to Virion Maturation : a Novel Role of the gp41 Cytoplasmic Tail. *J. Virol.* **78,** 3429–3435 (2004).

69. Laskey, S. B. & Siliciano, R. F. A mechanistic theory to explain the efficacy of antiretroviral therapy. *Nat. Rev. Microbiol.* **12,** 772–780 (2014).

70. Ward, A. B. & Wilson, I. A. Insights into the trimeric HIV-1 envelope glycoprotein structure. *Trends Biochem. Sci.* **40,** 101–

107 (2015).

71. Merk, A. & Subramaniam, S. HIV-1 envelope glycoprotein structure. *Curr. Opin. Struct. Biol.* **23,** 268–276 (2013).

72. Schwartz, S., Felber, B. K., Fenyö, E. M. & Pavlakis, G. N. Env and Vpu proteins of human immunodeficiency virus type 1 are produced from multiple bicistronic mRNAs. *J. Virol.* **64,** 5448–56 (1990).

73. Krummheuer, J. *et al.* A minimal uORF within the HIV-1 vpu leader allows efficient translation initiation at the downstream env AUG. *Virology* **363,** 261–271 (2007).

74. Earl, P. L., Moss, B. & Doms, R. W. Folding, interaction with GRP78-BiP, assembly, and transport of the human immunodeficiency virus type 1 envelope protein. *J. Virol.* **65,** 2047–55 (1991).

75. Poumbourios, P., Wilson, K. a, Center, R. J., El Ahmar, W. & Kemp, B. E. Human immunodeficiency virus type 1 envelope glycoprotein oligomerization requires the gp41 amphipathic alpha-helical/leucine zipper-like sequence. *J. Virol.* **71,** 2041–2049 (1997).

76. Pal, R., Hoke, G. M. & Sarngadharan, M. G. Role of oligosaccharides in the processing and maturation of envelope glycoproteins of human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. U. S. A.* **86,** 3384–8 (1989).

77. Bosch, V. & Pawlita, M. Mutational analysis of the human immunodeficiency virus type 1 env gene product proteolytic cleavage site. *J Virol* **64,** 2337–2344 (1990).

78. Haim, H., Salas, I. & Sodroski, J. Proteolytic Processing of the Human Immunodeficiency Virus Envelope Glycoprotein Precursor Decreases Conformational Flexibility. *J. Virol.* **87,** 1884–1889 (2012).

79. Herrera, C. *et al.* The impact of envelope glycoprotein cleavage on the antigenicity, infectivity, and neutralization sensitivity of Env-pseudotyped human immunodeficiency virus type 1 particles. *Virology* **338,** 154–172 (2005).

80. Munro, J. B. & Mothes, W. Structure and Dynamics of the Native HIV-1 Env Trimer. *J. Virol.* **89,** 5752–5 (2015).

81. Khayat, R. *et al.* Structural characterization of cleaved, soluble HIV-1 envelope glycoprotein trimers. *J. Virol.* **87,** 9865–72 (2013).

82. J.F. Rowell, A.L. Ruff, F.G. Guarnieri, K. Staveley-O'Carroll, X. Lin, J. Tang, J. T. A. and R. F. S. Lysosome-associated membrane protein-1-mediated targeting of the HIV-1 envelope protein to an endosomal/lysosomal compartment enhances its presentation to MHC class II-restricted T cells. *J Immunol* **155,** 1818–1828 (1995).

83. Groppelli, E., Len, A. C., Granger, L. A. & Jolly, C. Retromer Regulates HIV-1 Envelope Glycoprotein Trafficking and Incorporation into Virions. *PLoS Pathog.* **10,** (2014).

84. Zhu, P. *et al.* Electron tomography analysis of envelope glycoprotein trimers on HIV and simian immunodeficiency virus virions. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 15812–7 (2003).

85. Josef Schneider, Oskar Kaaden, Terry D. Copeland, Steven Oroszlan, G. Shedding and Interspecies Type Sero-reactivity of the Envelope Glycopolypeptide gp120 of the Human Immunodeficiency Virus. *J. Gen. Virol.* **67,** 2533–2538 (1986).

86. Chertova, E. *et al.* Envelope glycoprotein incorporation, not shedding of surface envelope glycoprotein (gp120/SU), Is the primary determinant of SU content of purified human immunodeficiency virus type 1 and simian immunodeficiency virus. *J. Virol.* **76,** 5315–25 (2002).

87. Klein, J. S. & Bjorkman, P. J. Few and far between: How HIV may be evading antibody avidity. *PLoS Pathog.* **6,** 1–6 (2010).

88. Hoffman, N. G., Seillier-Moiseiwitsch, F., Ahn, J., Walker, J. M. & Swanstrom, R. Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop. *J. Virol.* **76,** 3852–3864 (2002).

89. Modrow, S. *et al.* Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. *J. Virol.* **61,** 570–578 (1987).

90. Starcich, B. R. *et al.* Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. *Cell* **45,** 637–648 (1986).

91. Leonard, K., Spellman, W., Harris, R. J. & Thomas, N. Assignment of Intrachain Disulfide Bonds and Characterization of Potential Glycosylation Sites of the Type 1 Recombinant Human Immunodeficiency Virus Envelope Glycoprotein (gp120) Expressed in Chinese Hamster Ovary Cells *. **265,** 10373–10382 (1990).

92. Go, E. P., Zhang, Y., Menon, S. & Desaire, H. Analysis of the disulfide bond arrangement of the HIV-1 envelope protein CON-S gp140 delta-cFI shows variability in the V1 and V2 regions. *J. Proteome Res.* **10,** 578–591 (2011).

93. Wang, W. *et al.* A systematic study of the N-glycosylation sites of HIV-1 envelope protein on infectivity and antibody-mediated neutralization. *Retrovirology* **10,** 14 (2013).

94. Kwong, P. D. *et al.* Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature* **393,** 648–659 (1998).

95. Sterjovski, J. *et al.* CD4-binding site alterations in CCR5-using HIV-1 envelopes influencing gp120-CD4 interactions and fusogenicity. *Virology* **410,** 418–428 (2011).

96. Lasky, L. A. *et al.* Delineation of a region of the human immunodeficiency virus type 1 gp120 glycoprotein critical for interaction with the CD4 receptor. *Cell* **50,** 975–985 (1987).

97. Wyatt, R. *et al.* The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **393,** 705–11 (1998).

98. Wyatt, R. & Sodroski, J. The HIV-1 Envelope Glycoproteins: Fusogens,Antigens, and Immunogens. *Science (80-.).* **280,** 1884–8 (1998).

99. Bosch, M. L. *et al.* Identification of the fusion peptide of primate immunodeficiency viruses. *Science* **244,** 694–7 (1989).

100. Kong, R. *et al.* Fusion peptide of HIV-1 as a site of vulnerability to neutralizing antibody. **352,** 828–833 (2016).

101. Dubay, J. W., Roberts, S. J., Brody, B. & Hunter, E. Mutations in the leucine zipper of the human immunodeficiency virus type 1 transmembrane glycoprotein affect fusion and infectivity. *J. Virol.* **66,** 4748–56 (1992).

102. Mische, C. C. *et al.* An alternative conformation of the gp41 heptad repeat 1 region coiled coil exists in the human immunodeficiency virus (HIV-1) envelope glycoprotein precursor. *Virology* **338,** 133–143 (2005).

103. Muñoz-Barroso, I., Salzwedel, K., Hunter, E. & Blumenthal, R. Role of the membrane-proximal domain in the initial stages of human immunodeficiency virus type 1 envelope glycoprotein-mediated membrane fusion. *J. Virol.* **73,** 6089–92 (1999).

104. Mcmaster, M. L., Kristinsson, S. Y., Turesson, I., Bjorkholm, M. & Landgren, O. Structure of the HIV-1 gp41 Membrane-Proximal Ectodomain Region in a Putative Prefusion Conformation. *Biochemistry* **9,** 19–22 (2010).

105. Kim, J. H., Hartley, T. L., Curran, A. R. & Engelman, D. M. Molecular dynamics studies of the transmembrane domain of gp41 from HIV-1. *Biochim. Biophys. Acta - Biomembr.* **1788,** 1804–1812 (2009).

106. Apellániz, B. *et al.* The atomic structure of the HIV-1 gp41 transmembrane domain and its connection to the immunogenic membrane-proximal external region. *J. Biol. Chem.* **290,** 12999–13015 (2015).

107. Zhu, P. *et al.* Distribution and three-dimensional structure of AIDS virus envelope spikes. *Nature* **441,** 847–852 (2006).

108. Lynch, R. M., Shen, T., Gnanakaran, S. & Derdeyn, C. a. Appreciating HIV type 1 diversity: subtype differences in Env. *AIDS Res. Hum. Retroviruses* **25,** 237–248 (2009).

109. Mansky, L. M. & Temin, H. M. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.* **69,** 5087–94 (1995).

110. Merluzzi, V. J. *et al.* Structure and Function oh HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition. *Science (80-.).* **250,** 1411–1413 (2010).

111. Wei, X. *et al.* Antibody neutralization and escape by HIV-1. *Nature* **422,** 307–312 (2003).

112. Kalia, V., Sarkar, S., Gupta, P. & Montelaro, R. C. Antibody neutralization escape mediated by point mutations in the intracytoplasmic tail of human immunodeficiency virus type 1 gp41. *J. Virol.* **79,** 2097–107 (2005).

113. Sagar, M., Wu, X., Lee, S. & Overbaugh, J. Human immunodeficiency virus type 1 V1-V2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. *J. Virol.* **80,** 9586–98 (2006).

114. Wood, N. *et al.* HIV evolution in early infection: Selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS Pathog.* **5,** (2009).

115. Kwong, P. D. *et al.* HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* **420,** 678–682 (2002).

116. Susan Zolla-Pazner and Timothy Cardozo. Structure-function relationships of HIV-1 Envelope sequence-variable regions provide a paradigm for vaccine design. *Nat rev immunol* **10,** 527–535 (2010).

117. Rusert, P. *et al.* Interaction of the gp120 V1V2 loop with a neighboring gp120 unit shields the HIV envelope trimer against cross-neutralizing antibodies. *J. Exp. Med.* **208,** 1419–1433 (2011).

118. Mclellan, J. S. & Pancera, M. Structure of HIV-1 gp120 V1/V2 domain with broadly neutralizing antibody PG9. *Nature* **480,** 336–343 (2012).

119. Moldt, B. *et al.* Highly potent HIV-specific antibody neutralization in vitro translates into effective protection against mucosal SHIV challenge in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 18921–5 (2012).

120. James B. Munro, Jason Gorman, Xiaochu Ma, Zhou Zhou, James Arthos, Dennis R. Burton, Wayne C. Koff, Joel R. Courter, Amos B. Smith III, Peter D. Kwong, Scott C. Blanchard, W. M. Conformational dynamics of single HIV-1 envelope trimers on the surface of native virions. *October* **454,** 42–54 (2007).

121. McCaffrey, R. A., Saunders, C., Hensel, M. & Stamatatos, L. N-linked glycosylation of the V3 loop and the immunologically silent face of gp120 protects human immunodeficiency virus type 1 SF162 from neutralization by anti-gp120 and anti-gp41 antibodies. *J. Virol.* **78,** 3279–95 (2004).

122. Pancera, M. *et al.* Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature* **514,** 455–61 (2014).

123. Ozorowski, G. *et al.* Open and closed structures reveal allostery and pliability in the HIV-1 envelope spike. *Nature* (2017). doi:10.1038/nature23010

124. Tomaras, G. D. *et al.* Initial B-cell responses to transmitted human immunodeficiency virus type 1: virion-binding immunoglobulin M (IgM) and IgG antibodies followed by plasma anti-gp41 antibodies with ineffective control of initial viremia. *J. Virol.* **82,** 12449–63 (2008).

125. Overbaugh, J. & Morris, L. The antibody response against HIV-1. *Cold Spring Harb. Perspect. Med.* **2,** 1–17 (2012).

126. Burton, D. R. Antibodies, viruses and vaccines. *Nat. Rev. Immunol.* **2,** 706–713 (2002).

127. Dennis R. Burton and John R Mascola. Antibody responses to envelope glycoproteins in HIV-1 infection. *Nat Immunol.* **175,** 777–783 (2015).

128. Mouquet, H. Antibody B cell responses in HIV-1 infection. *Trends Immunol.* **35,** 549–561 (2014).

129. Hraber, P. *et al.* Prevalence of broadly neutralizing antibody responses during chronic HIV-1 infection. *Aids* **28,** 163–169 (2014).

130. Florian Klein, Hugo Mouquet, Pia Dosenovic, Johannes Scheid, L. S. and M. C. N. Antibodies in HIV-1 Vaccine Development and Therapy. *Science (80-.).* **341,** 1199–1204 (2013).

131. Sather, D. N. *et al.* Factors associated with the development of cross-reactive neutralizing antibodies during human immunodeficiency virus type 1 infection. *J. Virol.* **83,** 757–69 (2009).

132. Binley, J. M. *et al.* Profiling the specificity of neutralizing antibodies in a large panel of plasmas from patients chronically infected with human immunodeficiency virus type 1 subtypes B and C. *J. Virol.* **82,** 11651–68 (2008).

133. Euler, Z. *et al.* Longitudinal analysis of early HIV-1-specific neutralizing activity in an elite neutralizer and in five patients who developed cross-reactive neutralizing activity. *J. Virol.* **86,** 2045–55 (2012).

134. Barbas, C. F. *et al.* Recombinant human Fab fragments neutralize human type 1 immunodeficiency virus in vitro. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 9339–9343 (1992).

135. Burton, D. R. *et al.* Efficient neutralization of primary isolates of HIV-1 by a recombinant human monoclonal antibody. *Science* **266,** 1024–7 (1994).

136. Kwon, Y. Do *et al.* Structural Basis for Broad and Potent Neutralization of HIV-1 by Antibody VRC01. **329,** 811–817 (2011).

137. Trkola, a *et al.* Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of human immunodeficiency virus type 1. *J. Virol.* **70,** 1100–1108 (1996).

138. Bivona, A. E., Cerny, N., Alberti, A. S., Cazorla, S. I. & Malchiodi, E. L. Identification of common features in prototype broadly neutralizing antibodies to HIV envelope V2 apex to facilitate vaccine design. *Immunity* **43,** 959–973 (2015).

139. Gorny, M. K. *et al.* Neutralization of diverse human immunodeficiency virus type 1 variants by an anti-V3 human monoclonal antibody. *J. Virol.* **66,** 7538–42 (1992).

140. Julien, J. *et al.* Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* **477,** 466–470 (2011).

141. Muster, T. *et al.* A conserved neutralizing epitope on gp41 of human immunodeficiency virus type 1. *J. Virol.* **67,** 6642–6647 (1993).

142. Zwick, M. B. *et al.* Broadly Neutralizing Antibodies Targeted to the Membrane-Proximal External Region of Human Immunodeficiency Virus Type 1. *J. Virol.* **75,** 10892–10905 (2001).

143. Frey, G. *et al.* Distinct conformational states of HIV-1 gp41 are recognized by neutralizing and non-neutralizing antibodies. *Nat Struct Mol Biol.* **17,** 1486–1491 (2010).

144. Ruscio, A. Di *et al.* Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. *Nature* **491,** 406–412 (2012).

145. Huang, J. *et al.* Broad and potent HIV-1 neutralization by a human antibody that binds the gp41-gp120 interface. *Nature* **515,** 138–142 (2014).

146. Mcelroy, M. *et al.* Structural delineation of a quaternary, cleavage-dependent epitope at the gp41-gp120 interface on intact HIV-1 Env trimers. *Immunity* **40,** 669–680 (2014).

147. Willis, J. R. *et al.* Long antibody HCDR3s from HIV-naïve donors presented on a PG9 neutralizing antibody background mediate HIV neutralization. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 1518405113- (2016).

148. Yu, L. & Guan, Y. Immunologic basis for long HCDR3s in broadly neutralizing antibodies against HIV-1. *Front. Immunol.* **5,** 28–32 (2014).

149. Kepler, T. B. *et al.* Immunoglobulin gene insertions and deletions in the affinity maturation of HIV-1 broadly reactive neutralizing antibodies. *Cell Press* **16,** 304–313 (2015).

150. Liu, M. *et al.* Polyreactivity and autoreactivity among HIV-1 antibodies. *J. Virol.* **89,** 784–98 (2015).

151. Corti, D. *et al.* Analysis of memory B cell responses and isolation of novel monoclonal antibodies with neutralizing breadth from HIV-1-infected individuals. *PLoS One* **5,** (2010).

152. Bonsignori, M. *et al.* Analysis of a clonal lineage of HIV-1 envelope V2/V3 conformational epitope-specific broadly neutralizing antibodies and their inferred unmutated common ancestors. *J. Virol.* **85,** 9998–10009 (2011).

153. Sievers, S. a., Scharf, L., West, A. P. & Bjorkman, P. J. Antibody engineering for increased potency, breadth and half-life. *Curr. Opin. HIV AIDS* 1 (2015). doi:10.1097/COH.0000000000000148

154. Janeiro, R. De *et al.* Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. **372,** 1881–1893 (2009).

155. Badalà, F., Nouri-mahdavi, K. & Raoof, D. A. NIH Public AccessMagnitude and breadth of a non-protective neutralizing antibody response in an efficacy trial of a candidate HIV-1 gp120 vaccine (AIDSVAX™ B/B). *J Infect Dis* **144,** 724–732 (2008).

156. Chalovich, J. M. & Eisenberg, E. Safety and efficacy of the HVTN 503/Phambili study of a clade-B-based HIV-1 vaccine in South Africa: a double-blind, randomised, placebo-controlled test-of-concept phase 2b study. *Lancet Infect Dis.* **257,** 2432–2437 (2005).

157. Hammer, S. M. *et al.* Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. *N Engl J Med.* **369,** 2083–2092 (2014).

158. Pitisuttithum, P. *et al.* Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. *J. Infect. Dis.* **194,** 1661–1671 (2006).

159. Pisano, E. D. *et al.* Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N Engl J Med.* **361,** 2209–20 (2005).

160. O'Connell, R. J., Kim, J. H., Corey, L. & Michael, N. L. Human immunodeficiency virus vaccine trials. *Cold Spring Harb. Perspect. Med.* **2,** (2012).

161. Day, T. A. & Kublin, J. G. Lessons learned from HIV vaccine clinical efficacy trials. *Curr. HIV Res.* **11,** 441–9 (2013).

162. Szabo, F. K. & Hoffman, G. E. Safety and efficacy assessment of the HVTN 503/Phambili Study: A double-blind randomized placebo-controlled test-of-concept study of a Clade B-based HIV-1 vaccine in South Africa. *Lancet Infect Dis.* **37,** 62–70 (2012).

163. Duerr, A. *et al.* Extended follow-up confirms early vaccine-enhanced risk of HIV acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus HIV vaccine (Step Study). *J. Infect. Dis.* **206,** 258–266 (2012).

164. Sandham, J. D., Hull, M. D. R. D., Brant, M. B. B. S. R. F. & D, P. Vaccination with ALVAC and AIDSVAX to Prevent HIV-1 Infection in Thailand. *N Engl J Med* **361,** 2209–2220 (2009).

165. Kim, J. H., Excler, J.-L. & Michael, N. L. Lessons from the RV144 Thai Phase III HIV-1 Vaccine Trial and the Search for Correlates of Protection. *Annu. Rev. Med.* **66,** 423–437 (2015).

166. Hammer, S. M. *et al.* Efficacy Trial of a DNA/rAd5 HIV-1 Preventive Vaccine. *N. Engl. J. Med.* **369,** 2083–2092 (2013).

167. Sausen, M. *et al.* Approaches to Preventative and Therapeutic HIV vaccines. *Curr Opin Virol* **17,** 104–109 (2016).

168. Mascola, J. R. *et al.* Protection of macaques against vaginal transmission of a pathogenic HIV-1/SIV chimeric virus by passive infusion of neutralizing antibodies. *Nat. Med.* **6,** 207–210 (2000).

169. Hessell, A. J. *et al.* Broadly neutralizing human anti-HIV antibody 2G12 is effective in protection against mucosal SHIV challenge even at low serum neutralizing titers. *PLoS Pathog.* **5,** (2009).

170. Pegu, A. *et al.* Neutralizing antibodies to HIV-1 envelope protect more effectively in vivo than those to the CD4 receptor. *Sci. Transl. Med.* **6,** 243ra88 (2014).

171. Klein, F. *et al.* HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature* **492,** 1–14 (2013).

172. Peter Gilbert, Maggie Wang, T. W. et al. Magnitude and breadth of a non-protective neutralizing antibody response in an efficacy trial of a candidate HIV-1 gp120 vaccine (AIDSVAX™ B/B). *J Infect Dis* **202,** 595–605 (2010).

173. Kovacs, J. M. *et al.* HIV-1 envelope trimer elicits more potent neutralizing antibody responses than monomeric gp120. *Proc. Natl. Acad. Sci.* **109,** 12111–6 (2012).

174. Walker, L. M. & Burton, D. R. Rational Antibody-based HIV-1 Vaccine Design: Current Approaches and Future Directions. *Curr Opin Immunol* **22,** 358–366 (2010).

175. Wilson, R. P. and I. A. Structure-based vaccine design in HIV: blind men and the elephant? *Curr Pharm Des* **16,** 3744–3753 (2010).

176. Sattentau, L. K. and Q. J. Antigenicity and Immunogenicity in HIV-1 Antibody-Based Vaccine Design. *Cell Host Microbe* **12,** 396–407 (2012).

177. Denis R. Burton, Rafi Ahmed, D. H. B. et al. A blueprint for HIV vaccine discovery. **12,** 396–407 (2012).

178. Moore PL *et al.* Nature of nonfunctional envelope proteins on the surface of human immunodeficiency virus type 1. *J. Virol.* **80,** 2515–2528 (2006).

179. Decroly, E. *et al.* The convertases furin and PC1 can both cleave the human immunodeficiency virus (HIV)-1 envelope glycoprotein gp160 into gp120 (HIV-I SU) and gp41 (HIV-I TM). *J. Biol. Chem.* **269,** 12240–12247 (1994).

180. Oliva, R. *et al.* Structural investigation of the HIV-1 envelope glycoprotein gp160 cleavage site, 2: Relevance of an N-terminal helix. *ChemBioChem* **4,** 727–733 (2003).

181. Heyndrickx, L. *et al.* Selected HIV-1 Env trimeric formulations act as potent immunogens in a rabbit vaccination model. *PLoS One* **8,** e74552 (2013).

182. Perdiguero, B. *et al.* Virological and immunological characterization of novel NYVAC-based HIV/AIDS vaccine candidates expressing clade C trimeric soluble gp140(ZM96) and Gag(ZM96)-Pol-Nef(CN54) as virus-like particles. *J. Virol.* **89,** 970–88 (2015).

183. Binley, J. M. *et al.* Enhancing the proteolytic maturation of human immunodeficiency virus type 1 envelope gGlycoproteins. *J. Virol.* **76,** 2606–2616 (2002).

184. Ruscio, A. Di *et al.* Cleavage-independent HIV-1 Env trimers engineered as soluble native spike mimetics for vaccine design. *Cell Rep* **11,** 539–550 (2015).

185. Sanders, R. W. *et al.* A Next-Generation Cleaved, Soluble HIV-1 Env Trimer, BG505 SOSIP.664 gp140, Expresses Multiple Epitopes for Broadly Neutralizing but Not Non-Neutralizing Antibodies. *PLoS Pathog.* **9,** (2013).

186. Binley, J. M. *et al.* A recombinant human immunodeficiency virus type 1 envelope glycoprotein complex stabilized by an intermolecular disulfide bond between the gp120 and gp41 subunits is an antigenic mimic of the trimeric virion-associated structure. *J. Virol.* **74,** 627–43 (2000).

187. Steven W. de Taeye, J. P. M. and R. W. S. HIV-1 envelope trimer design and immunization strategies to induce broadly neutralizing antibodies. *Trends Immunol* **37,** 221–232 (2016).

188. Zhao, Y. *et al.* Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 Env trimer. *Science (80-.).* **342,** 1484–1490 (2013).

189. Medina-Ramírez, M., Sanders, R. W. & Sattentau, Q. J. Stabilized HIV-1 envelope glycoprotein trimers for vaccine use. *Curr. Opin. HIV AIDS* **12,** 241–249 (2017).

190. Sanders, R. W. & Moore, J. P. Native-like Env trimers as a platform for HIV-1 vaccine design. *Immunol. Rev.* **275,** 161–182 (2017).

191. Jardine, J. *et al.* Rational HIV immunogen design to target specific germline B cell receptors. *Science* **340,** 711–6 (2013).

192. Briney, B. *et al.* Tailored Immunogens Direct Affinity Maturation toward HIV Neutralizing Antibodies. *Cell* **166,** 1459–1470.e11 (2016).

193. Escolano, A. *et al.* Sequential Immunization Elicits Broadly Neutralizing Anti-HIV-1 Antibodies in Ig Knockin Mice. *Cell* **166,** 1445–1458.e12 (2016).

194. Neylon, C. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: Library construction methods for directed evolution. *Nucleic Acids Res.* **32,** 1448–1459 (2004).

195. Stemmer, W. P. Rapid evolution of a protein in vitro by DNA shuffling. *Nature* **370,** 389–391 (1994).

196. Stemmer, W. P. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 10747–10751 (1994).

197. Merchant, a M. *et al.* Molecular evolution by staggered extension process (StEP) in vitro recombination. *Group* **16,** 291–294 (1998).

198. Thomas, S., Maynard, N. D. & Gill, J. DNA library construction using Gibson Assembly regions. *Nat. Publ. Gr.* **12,** i–ii (2015).

199. Gorny, M. K. *et al.* Production of site-selected neutralizing human monoclonal antibodies against the third variable domain of the human immunodeficiency virus type 1 envelope glycoprotein. *Immunology* **88,** 3238–3242 (1991).

200. Boots, L. J. *et al.* Anti-human immunodeficiency virus type 1 human monoclonal antibodies that bind discontinuous epitopes in the viral glycoproteins can identify mimotopes from recombinant phage peptide display libraries. *AIDS Res Hum Retroviruses* **13,** 1549–1559 (1997).

201. Dieltjens, T. *et al.* Evolution of antibody landscape and viral envelope escape in an HIV-1 CRF02_AG infected patient with 4E10-like antibodies. *Retrovirology* **6,** 113 (2009).

202. Green, MR and Sambrook, J. *Molecular Cloning - A Laboratory Manual 4th Edition*. *Cold Spring Harbor Laboratory Press* (2012). doi:10.3724/SP.J.1141.2012.01075

203. Hanahan, D. Studies on transformation of Escherichia coli with plasmids. *J. Mol. Biol.* **166,** 557–580 (1983).

204. Bimboim, H. C. & Doly, J. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* **7,** 1513–1523 (1979).

205. Shin, J., Ming, G. & Song, H. Decoding neural transcriptomes and epigenomes via high-throughput sequencing. *Nat Neurosci* **17,** 1463–1475 (2014).

206. Sap, K. A. & Demmers, J. A. A. Next Generation Sequencing in Aquatic Models. *Intech* **6,** 111–133 (2016).

207. Boussif, O. *et al.* A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine. *Proc. Natl. Acad. Sci.* **92,** 7297–7301 (1995).

208. Perfetto, S. P., Chattopadhyay, P. K. & Roederer, M. Seventeen-colour flow cytometry: unravelling the immune system. *Nat. Rev. Immunol.* **4,** 648–55 (2004).

209. Bruun, T.-H. Development of mammalian cell display and panning techniques for the selection of HIV-1 vaccine candidates. *Dissertation* (2012).

210. Doria-Rose, N. A. *et al.* Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature* **509,** 55–62 (2014).

211. Szymczak, A. L. *et al.* Correction of multi-gene deficiency in vivo using a single 'self-cleaving' 2A peptide–based retroviral vector. *Nat. Biotechnol.* **22,** (2004).

212. Chakrabarti, B. K. *et al.* Modifications of the Human Immunodeficiency Virus Envelope Glycoprotein Enhance Immunogenicity for Genetic Immunization Modifications of the Human Immunodeficiency Virus Envelope Glycoprotein Enhance Immunogenicity for Genetic Immunization. *J. Virol.* **76,** 5357–5368 (2002).

213. Grassmann, V. Development and application of high- throughput screening methods to generate novel HIV-1 envelope immunogens. *Dissertation* (2017).

214. Ziegler, C. Identification and characterization of trimeric HIV envelope protein variants with increased thermostability. *Master thesis* (2014).

215. Konishi, M., Kawamoto, K., Izumikawa, M., Kuriyama, H. & Yamashita, T. Gene transfer into guinea pig cochlea using adeno-associated virus vectors. *J. Gene Med.* **10,** 610–618 (2008).

216. Ledergerber, C. & Dessimoz, C. Base-calling for next-generation sequencing platforms. *Brief. Bioinform.* **12,** 489–497 (2011).

217. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43,** (2015).

218. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* **148,** 1667–1686 (1998).

219. Chen, B. *et al.* Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature* **433,** 834–41 (2005).

220. Davenport, Y. W., West, A. P. & Bjorkman, P. J. Structure of an HIV-2 gp120 in Complex with CD4. *J. Virol.* **90,** 2112–2118 (2015).

221. Rodi, D. J., Mandava, S. & Makowski, L. DIVAA: Analysis of amino acid diversity in multiple aligned protein sequences. *Bioinformatics* **20,** 3481–3489 (2004).

222. Jores, R., Alzari, P. M. & Meo, T. Resolution of hypervariable regions in T-cell receptor beta chains by a modified Wu-Kabat index of amino acid diversity. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 9138–9142 (1990).

223. Garcia-Boronat, M., Diez-Rivero, C. M., Reinherz, E. L. & Reche, P. A. PVS: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery. *Nucleic Acids Res.* **36,** 35–41 (2008).

224. Bruun, T. H., Mühlbauer, K., Benen, T., Kliche, A. & Wagner, R. A Mammalian Cell Based FACS-Panning Platform for the Selection of HIV-1 Envelopes for Vaccine Development. *PLoS One* **9,** (2014).

225. Bruun, T.-H. *et al.* Mammalian cell surface display for monoclonal antibody-based FACS selection of viral envelope proteins. *MAbs* **0,** 0–0 (2017).

226. Schmalzl, C. Generation and biochemical characterization of HIV-1 envelope immunogens with increased stability and affinity to broadly neutralizing antibodies. *Master thesis* (2016).

227. Khare, P. D., Rosales, A. G., Bailey, K. R., Russell, S. J. & Federspiel, M. J. Epitope selection from an uncensored peptide library displayed on avian leukosis virus. *Virology* **315,** 313–321 (2003).

228. Siying Ma, I. S. and J. T. Error correction in gene synthesis technology. *Trends Biochem. Sci.* **30,** 147–154 (2012).

229. Büssow, K. Stable mammalian producer cell lines for structural biology. *Curr. Opin. Struct. Biol.* **32,** 81–90 (2015).

230. Cardarelli, F. *et al.* The intracellular trafficking mechanism of Lipofectamine-based transfection reagents and its implication for gene delivery. *Sci. Rep.* **6,** 25879 (2016).

231. Pichardo, S., Togtema, M., Jackson, R., Zehbe, I. & Curiel, L. Influence of cell line and cell cycle phase on sonoporation transfection efficiency in cervical carcinoma cells under the same physical conditions. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **60,** 432–435 (2013).

232. Brunner, S. *et al.* Cell cycle dependence of gene transfer by lipoplex, polyplex and recombinant adenovirus. *Gene Ther.* **7,** 401–407 (2000).

233. Golzio, M., Teissié, J. & Rols, M. P. Cell synchronization effect on mammalian cell permeabilization and gene delivery by electric field. *Biochim. Biophys. Acta - Biomembr.* **1563,** 23–28 (2002).

234. Dean, D. a. Import of plasmid DNA into the nucleus is sequence specific. *Exp. Cell Res.* **230,** 293–302 (1997).

235. Head, S., Komori, K., LaMere, S., Whisenant, T. & et al. Library construction for next-generation sequencing: Overviews and challenges Steven. *Biotechniques* **56,** (2015).

236. Description, P. Technical Data Sheet KAPA Library Quantification Kits For Illumina sequencing platforms KAPA Library Quantification Kits – Illumina sequencing platforms. 2–7 (2014).

237. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36,** (2008).

238. Bentley, D. R. *et al.* Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. *Nature* **456,** 53–59 (2009).

239. Quail, M. A. *et al.* A large genome centre' s improvements to the Illumina sequencing system. *Nat. Methods* **5,** 1005–1010 (2009).

240. Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y. & Hwang, C. C. Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoS One* **8,** (2013).

241. Grimm, S. K., Battles, M. B. & Ackerman, M. E. Directed evolution of a yeast-displayed HIV-1 SOSIP gp140 spike protein toward improved expression and affinity for conformational antibodies. *PLoS One* **10,** 1–20 (2015).

242. Cherf, G. M. & Cochran, J. R. Applications of yeast surface display for protein engineering. *Methods Mol Biol.* **1319,** 155–175 (2015).

243. De Berardinis, P. *et al.* Phage display of peptide epitopes from HIV-1 elicits strong cytolytic responses. *Nat. Biotechnol.* **18,** 873–876 (2000).

244. Delhalle, S., Schmit, J. C. & Chevign??, A. *Phages and HIV-1: From display to interplay*. *International Journal of Molecular Sciences* **13,** (2012).

245. Kwong, P. D. & Wilson, I. A. HIV-1 and influenza antibodies: seeing antigens in new ways. *Nat. Immunol.* **10,** 573–578 (2009).

246. Steichen, J. M. *et al.* HIV Vaccine Design to Target Germline Precursors of Glycan-Dependent Broadly Neutralizing Antibodies. *Immunity* **45,** 483–496 (2016).

247. Dougherty, D. Cation-p Interactions Involving Aromatic. *J. Nutr.* **137,** 1504–1508 (2007).

248. Wedemeyer, W. J., Welker, E. & Scheraga, H. a. Proline Cis - Trans Isomerization and Protein Folding. *Biochemistry* **41,** 14637–14644 (2002).

249. Li, Y. *et al.* Mechanism of Neutralization by the Broadly Neutralizing HIV-1 Monoclonal Antibody VRC01. *J. Virol.* **85,** 8954–8967 (2011).

250. IUPAC-IUB Commission on Biochemical Nomenclature. A One-Letter Notation for Amino Acid Sequences. *Eur. J. Biochem.* **5,** 151–153 (1968).

251. Baumlein, H., Wobus, U., Pustell, J. & Kafatosl, F. C. A One-Letter Notation for Amino Acid Sequences. *Nucleic Acids Res.* **14,** 2707–2720 (1986).

Acknowledgements

Undertaking this doctoral study has been a truly life-changing experience for me and I would like to convey my heartfelt gratitude to all who supported me throughout this long journey and inspired me to persevere. This thesis would not have been possible without the assistance of many people who I am indebted to.

First and foremost, I would like to express my sincerest appreciation to Prof. Dr. Ralf Wagner for giving me the opportunity to join his group. Your enlightening guidance and inspiring instructions throughout the thesis allowed me to expand and shape my expertise as a scientist. Furthermore, I would like to thank my mentors Prof. Dr. Falk Nimmerjahn and Prof. Dr. Dr. Hans Robert Kalbitzer for their constructive advice and profound suggestions.

I am highly appreciative to Prof. Dr. Gunter Meister and his study group who allowed me to utilize their MiSeq device for my Next Generation Sequencing analysis. A special gratitude goes to Norbert Eichner who was very generous in imparting his extensive knowledge of NGS to me and who spent many valuable hours of his time to teach me this elaborate technique.

I would like to express my utmost gratitude to Dr. David Peterhoff for sharing his vast scientific expertise and for his guidance of my laboratory work. Your incessant attention to detail drove me to finally learn how to approach a problem by systematic thinking and data-driven decision making. My sincere appreciation should also go to Dr. Benedikt Asbach for his dedicated advice, his knowledgeable skills in programming and his patience in proof-reading this thesis. Furthermore, I thank Dr. Alexander Kliche for his supervision throughout my first project.

There are not enough words to express my deep appreciation of my colleagues who genially accepted me into their fold. Dear Alexandra, Anh, Anja, Christina, Malin, Melanie, Miriam, Vroni, Ali, Benni, Jogi, Richie and Tom, thank you so much for all your scientific support and personal help that kept me moving forward. Your friendship and warmth has always made me feel at ease. I will always cherish and fondly remember our coffee time banter and laughter, our beach volleyball games, parties, the many dinners and gaming nights we had together.

Finally, I would like to express my deepest love to my Mom and Dad for their abiding affection, their unfailing support and for always encouraging me to follow my dreams.