CD8+ T cell epitope-enriched HIV-1-Gag antigens with preserved structure and function

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

vorgelegt von

Johannes Peter Meier

aus

Eggenfelden

im Jahr

2016

Das Promotionsgesuch wurde eingereicht am:
22.09.2016

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

Unterschrift:

Johannes Meier

Für meine Eltern.

Contents

Abstract

Control of disease progression in certain HIV-1 infected individuals is often associated with CD8+ T cell responses directed towards Gag-derived epitopes presented on HLA class I molecules. This indicates that such responses play a crucial role in combating virus replication. However, both the large variability of HIV-1 and the diversity of HLA alleles impose a challenge on the elicitation of protective CD8+ T cell responses by vaccination.

To address this problem, an algorithm was conceived to generate Gag antigens enriched with patient-derived CD8+ T cell epitopes. Since the function of Gag to produce virus-like particles (VLPs) was deemed important for priming of an adequate CD8+ T cell response, the program excluded all epitopes with budding-deleterious properties. To achieve this, all amino acid substitutions (AAS) that had been identified in the epitope set through mapping them to a Gag reference sequence, were assessed using a trained classifier that considers structural-energy- and sequence-conservation-based features to predict whether each AAS is compatible with budding. These predictions were validated experimentally for over 100 variants, showing a precision of 100% regarding classification of budding competence. Next, epitopes that contain only budding-retaining AAS were assigned a score that considers various customizable epitope-specific properties, like frequencies of HLA class I molecules presenting the epitope in a given population, subtype affiliation, and conservation status. Using a genetic algorithm, as many compatible epitopes as possible were combined into a novel Gag antigen sequence, aiming to maximize their cumulative score. After each round of antigen generation, all previously integrated epitopes were eliminated from the input data set. Thus, in subsequent rounds only the remaining epitopes were used, which resulted in a set of complementary antigens. To evaluate the performance of the algorithm, a trivalent set of globally applicable CD8+ T cell epitope-enriched Gag antigens (teeGags1-3) was generated and computationally validated in this thesis. It could be shown that the teeGags are superior to any known, naturally found or *in silico* generated Gag sequence from previously published work regarding the number and quality of epitopes, as well as the population coverage, defined as the average number of epitopes presented per person. The shape and size of teeGag VLPs were examined biochemically and wildtype-like characteristics were observed for teeGag1 and teeGag3. teeGag2, however, exhibited some aberrant, tubular structures and slightly larger particles, probably due to a set of mutations within the p2 region of Gag. To characterize the increased immunological breadth of the teeGags, a method to directly identify HLA-class-I-presented epitopes was conceived. For this, the conditioned supernatant from cells that produce soluble forms of HLA (sHLA) was used for HLA-affinity chromatography. Peptides from the isolated sHLA complexes were further purified and employed for sequencing through LC-MS/MS analysis. It was shown in this thesis that this method can be used to identify sHLA-restricted peptides. However, the sensitivity has to be further increased to allow examination of the immunological breadth of antigens.

In conclusion, with the *in silico* validated enhanced immunological breadth and the biochemically verified structural conservation, the presented designer teeGags qualify as next-generation vaccine antigens that potentially elicit superior CD8+ T cell responses.

A Introduction

A.1 Human immunodeficiency virus type-1 (HIV-1)

A.1.1 Epidemiology of HIV-1 and AIDS

HIV-1 was first described in 1983 by Gallo et al.[1] and Barré-Sinoussi et al.[2] as cause of acquired immune deficiency syndrome (AIDS), that had been discovered just two years before[3]. Since then, HIV/AIDS became a global pandemic[4] with a total of 78 million infected people and 36.7 million people living with HIV in 2015[5,6]. Most stricken by HIV/AIDS are Eastern and Southern Africa, which accounted for 46% of all global new HIV infections in 2015[7]. Mainly due to a scale up of antiretroviral therapy, that reached a worldwide coverage of 46% in 2015, the number of AIDS-related deaths continuously declined, from 2 million in 2005 to 1.1 million in 2015. However, the number of new HIV infections among adults remained nearly static in the last years, with 1.9 million in 2015. In Eastern Europe and Central Asia there even has been a steep rise in new infections since 2010[5].

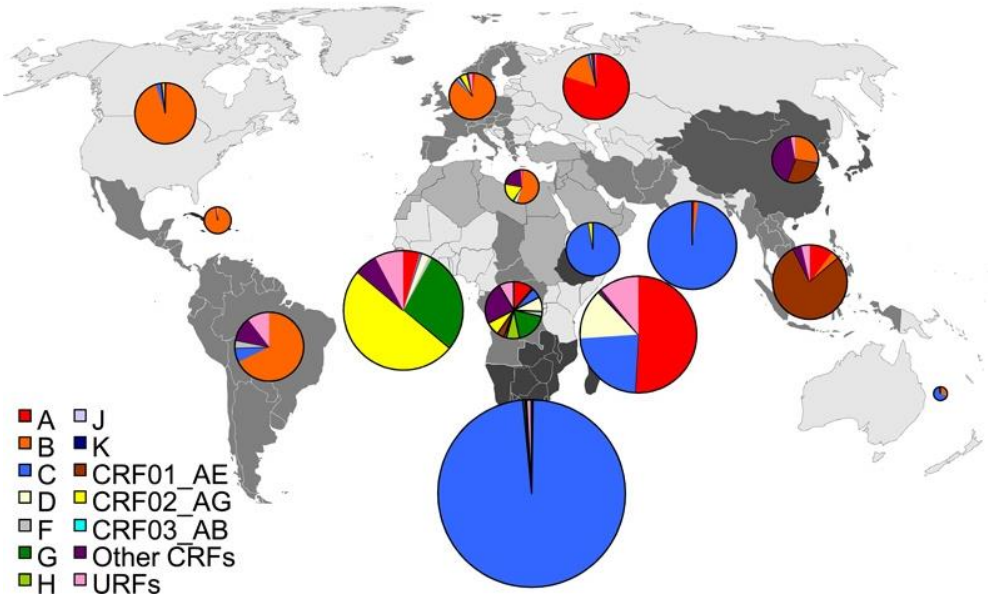


Figure 1. Regional distribution of HIV-1 subtypes, CRFs, and URFs in 2004-2007 (licensed reuse from Hemelaar et al.[8]). For the figure the world was divided into regions, as indicated by differential color shading. Pie-charts symbolize subtype, CRF, and URF frequencies in the respective region, according to the color-coded legend on the lower-left. Areas of the pie-charts are relative to the number of HIV-1 infected people in the regions.

HIV-1 is classified into four groups: M, N, O, and P, which arose at the beginning of the 20th century through at least four independent zoonotic transmissions of the chimpanzee simian immunodeficiency virus (SIV_{CPZ}) to humans[9,10] or for P possibly from gorilla SIV (SIV_{GOR})[11]. Group M is responsible for most infections worldwide, while group O only infected around 10,000 people and groups N and P far less[12]. Group M is further partitioned into 9 phylogenetically linked subtypes, or clades, (A–D, F–H, J, and K). In dually infected persons, recombination can

give rise to circulating or unique recombinant forms (CRFs or URFs) of two or more subtypes. If a recombinant was identified in at least three people with no direct epidemiologic linkage it is called CRF, otherwise URF. As of now, there are 79 HIV-1 CRFs registered[a] and they are becoming more widespread[8]. The regional distribution of subtypes and CRFs is very uneven (Figure 1). The globally most frequent (48% of all people living with HIV; Figure 18) subtype C is prevailing in the epidemic regions of Southern Africa (98%). In Western and Central Europe, as well as North America, in contrast, subtype B is most frequent (85% and 94%, respectively), although only 11% of all worldwide infections are with subtype B viruses. In other regions different subtypes prevail: For example subtype A in Eastern Europe and central Asia (80%) or CRF01_AE in south and southeast Asia (79%)[8].

A.1.2 HIV-1 structure and genome organization

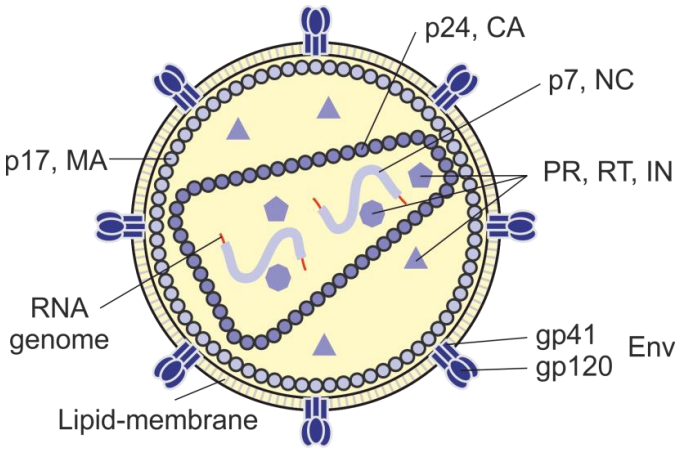


Figure 2. Schematic structure of a mature HIV-1 particle (kindly provided by Benedikt Asbach). HIV-1 proteins are highlighted in different shades of blue. Env proteins gp41 and gp120 are the only viral surface molecules and are anchored in the host-derived lipid-membrane. Gag is processed by PR to p24 (CA) that forms the conical capsid, the lipid-membrane attached p17 (MA), and p7 (NC) that associates with the viral RNA genome (in red). Enzymes RT and IN are stored within the capsid structure

HIV-1 and its close relative HIV-2 are the only two human lentiviruses of the retroviridae family. HIV-1 is roughly spherical with a diameter of 100-120 nm and enveloped by a host-cell-derived membrane[13] (Figure 2). The only HIV-1 protein presented on the particle surface is the envelope protein Env, constituted of two non-covalently linked subunits, gp120 and the membrane-anchored gp41. Three of these heterodimers form the natural trimeric spike structure of Env[14]. Every viral particle displays around 14 such envelope protein trimers[15]. To the other side of the membrane, within the virus, the matrix (MA, p17) protein is bound through an N-terminally attached myristoyl group. A conical capsid, made solely of capsid (CA, p24) protein, contains enzymes reverse transcriptase (RT), integrase (IN), and protease (PR) that are important for the virus replication[16]. Also, the viral genome, comprised of two identical (+)ssRNA copies, which are tightly bound to nucleocapsid (NC, p7) proteins, is stored within the capsid[17]. The genome has a size of about 9,700 to 9,800 nt and consists, besides several structural motifs, of nine genes: the structural genes *gag* (coding for p17, p24, p7, p6, p1, and p2), *pol* (RT, IN, and PR), and *env* (gp120 and gp41), as well as the regulatory and accessory genes *vif*, *vpr*, *tat*, *rev*, *vpu*, and *nef*. The ssRNA genome has a 5' cap and a 3' poly(A) tail, showing characteristics of eukaryotic mRNA. The genes are partially overlapping within the genome and are translated after alternative splicing from different reading frames.

[a] http://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html (last modified April 5 2016)

A.1.3 Replication cycle of HIV-1

HIV-1 replication is a multistep process, starting with the interaction of gp120 with the CD4 surface receptor[18] (Figure 3 - step 1), presented by CD4+ T cells and macrophages. The engagement triggers conformational changes between inner and outer domains of gp120 monomers. A newly formed bridging sheet exposes the binding site of the HIV-1 co-receptor, either chemokine receptor CCR5 or CXCR4[19]. Only after co-receptor binding, gp41 inserts a fusion peptide into the cell membrane, inducing formation of a six-helix hairpin structure that brings the viral and cellular membranes in close proximity, necessary for fusion (2). After uncoating through CA shell disintegration, the heterodimeric RT binds to the ssRNA viral genome[20] (3). The two functionally active sites of RT, an RNA- and DNA-dependent DNA polymerase and an RNase H that digests the RNA strands of RNA-DNA hybrids, transform the ssRNA genome into dsDNA (4). Together with IN and other viral and host proteins, the reverse transcribed dsDNA viral genome forms the pre-integration complex (PIC). This complex enters the nucleus through the nuclear pore complex (5). IN then cleaves the long terminal repeat (LTR) of HIV-1, as well as a random sequence within the host genome, unveiling 5' termini overlaps. Viral DNA ends are ligated by IN to the target DNA 5'phosphates and host enzymes complete the integration by repairing the single-strand gaps, generating a stable integrated provirus genome (6).

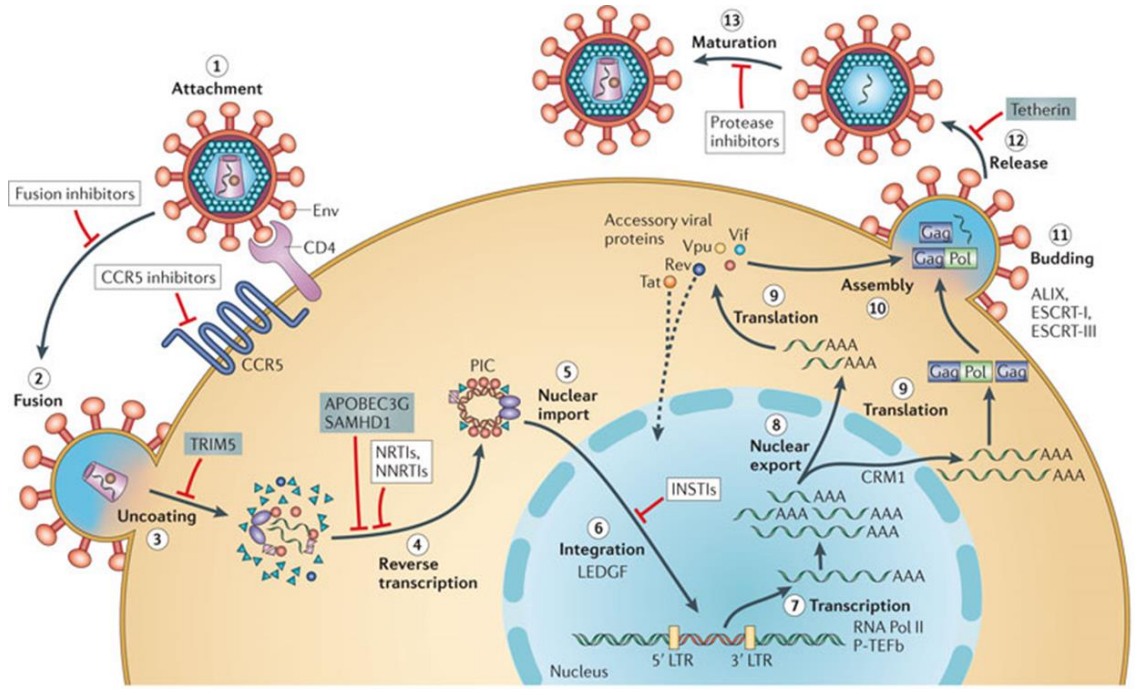


Figure 3. Schematic presentation of HIV-1 replication (Reprinted by permission from Macmillan Publishers Ltd: Nat Rev Microbiol. 10(4):279-90,[21] ©2012). The individual steps of the replication cycle are described in detail in the text (A.1.3). Grey boxes show host restriction factors, inhibiting viral replication and white boxes indicate sites of action of clinical inhibitors.

The post-integration phase of HIV-1 replication focuses on gene expression, as well as assembly and release of new viral particles. The transcription of HIV genes is initiated by a promoter within the upstream LTR. The viral mRNA harbors various splice donor and acceptor sites resulting in a variety of alternatively spliced species (7). Fully spliced mRNAs coding for Tat, Rev, and Nef are exported readily through the Tap-p15 cellular export pathway (8). The proteins Tat and Rev (9) are both transported back into the nucleus. Whereas the viral

transactivator protein Tat increases transcription efficiency manifold, Rev is necessary to export intron-containing, singly spliced (coding for Env, Vif, Vpr, and Vpu) or unspliced (coding for group-specific antigen (Gag) and Pol) viral mRNAs. Rev thereby acts as an adapter to the secondary RNA motif Rev response element (RRE) and the host nuclear export factor CRM1 (8). Env mRNA gets translated directly into the ER and the protein is subsequently transported through the Golgi apparatus to the cell surface. On its way, Env gets highly glycosylated and then cleaved by cellular Furin proteases into its gp41 and gp120 subunits. Gag on the other hand is expressed in the cytoplasm as precursor polyprotein (p55 or $Pr55^{Gag}$), containing the structural proteins p17, p24 p7, p6 and spacers p1 and p2. The GagPol polyprotein is translated in a 1:20 ratio compared to Gag from the same template RNA, due to a ribosomal frameshift event at the slippery site, consisting of six uridines within *gag* (9). By N-terminal myristoylation[22] and conserved basic amino acids within MA[23], Gag and GagPol are anchored at the inner leaflet of the plasma membrane. Gag is sufficient for assembly (10) and budding (11) of viral particles, but in natural HIV-1 replication, unspliced genomic viral ssRNA is encapsidated by binding to the p7 part of Gag[24], and Env is probably included by p17 and gp41 coupling[25]. For release from the cell, the particle must undergo membrane abscission, mediated by the cellular endosomal sorting complexes required for transport (ESCRT) machinery (12). Gag-particle assembly and release is described in detail below (A.3.1). Shortly after, or concomitant to virus budding, the viral protease PR cleaves Gag and GagPol into p17, p24, p7, PR, RT, and IN, resulting in a rearrangement of proteins: p17 remains membrane bound, whereas p7 accumulates at the viral RNA, and p24 forms the conical capsid, embedding the ssRNA genome, RT, and IN (13). Only after this maturation HIV-1 is infectious and ready for the next replication round.

A.1.4 HIV-1 transmission, pathogenesis, and treatment

The main transmission route for HIV-1 is by sexual contact, whereas men who have sex with men, female sex workers, and transgender women are at especially high risk of acquisition. High viral load, measured as HIV-1 RNA level in the plasma, is the most important risk factor that increases probability of transmission[26]. With every $\log_{10}$ elevated viral plasma titer the infection risk increases 2.5 times[27]. Besides sexual transmission, shared needles for injection drug users[28] and mother-to-child transmissions, perinatal or through breastfeeding, are other possible infection routes. Sexual transmission is mostly established by a single founder virus, that is more likely CCR5- than CXCR4-tropic[29]. Main targets after crossing mucosal membranes are activated CD4+ T cells, but other cells expressing CD4 and the appropriate chemokine receptor, like resting CD4+ T cells and macrophages, can be infected as well[26]. After transmission, during the acute phase of the infection, rapidly increased HIV replication causes high viral loads[30] and also a striking immune activation.

At the beginning of the adaptive immune response, mainly CD8+ T-cell-mediated killing of infected cells occurs. Neutralizing antibodies arise only about 3 months after infection. About 20% of all HIV+ people generate broadly neutralizing antibodies (bNAbs)[31], capable of neutralizing a variety of isolates. However, these bNAbs require often years to develop and, because of rapidly established escape-mutations, provide no benefits for patients[26]. During the acute infection, the CD4+ T cell levels are transiently reduced and the virus rapidly disseminates to the lymphoid organs. Over the course of roughly 4-6 weeks the host innate and adaptive immune responses reduce the plasma virus concentration considerably, triggering the virus to

enter the chronic phase and clinical latency. The level at which the viral load stabilizes, the so called viral set point, depends strongly on the quality of the cellular immune response (A.2.1.4).

With establishment of chronic infection and immune-system-mediated viral control, the CD4+ T cell numbers recover initially. The chronic infection is also characterized by a distinct, continuous immune activation that progressively results in immune dysfunction, immune cell depletion and increased virus production. If untreated, the CD4+ T cell levels gradually decrease to about 50-100 cells per µl and the patient develops a high-level viremia, which marks the transition of the clinical latency to AIDS. Due to the dysfunctional immune system, opportunistic diseases arise quickly and without intervention the patient's median survival is only 11 months[32].

However, due to effective treatment, an HIV-1 infection that was formerly considered lethal is now a manageable chronic disease[21]. This antiretroviral treatment (ART) combines nucleoside RT inhibitors with non-nucleoside RT, protease, or integrase inhibitors that act on different steps of viral replication (Figure 3). Until recently, the treatment was normally started, if a limit CD4+ T cell count (200 or 500 per µl) was undercut. The latest guideline on when to start antiretroviral therapy however recommends that "antiretroviral therapy (ART) should be initiated in everyone living with HIV at any CD4 cell count"[33], since early initiation might slow disease progression[34]. Treatment lowers plasma viral load to concentrations below the limit of detection and can restore normal CD4+ T cell levels. Besides near-normal life expectancy in these treated patients[35], the transmission risk is strongly reduced due to their aviremia. However, due to long-living cellular reservoirs, an HIV-1 infection is never eliminated with current treatment concepts. Main HIV-1 reservoirs are latently-infected resting memory CD4+ T cells[36,37] and especially lymph-node-located T_{FH} cells[38]. The latter are able to produce replication-and infection-competent HIV-1 particles even after >14 years of ART treatment, if activated[38]. It was estimated that 70 years of ART are necessary to eradicate all latent reservoirs[39].

As long as no sterilizing (i.e. viral eradication) or at least functional (complete, sustained control of infection without treatment) cure of HIV-1 is possible, prevention of infection is the most important action to control the epidemic. There are different approaches to avoid HIV-1 transmission, like pre-exposure prophylaxis, post-exposure prophylaxis, medical male circumcision, behavior interventions, and vaginal- and rectal-microbicides[26]. Some of these methods showed very promising initial protection in small, selected cohorts, but had problems when used for a broad implementation, due to low adherence[40]. Therefore the easiest and most practicable way to contain and eventually even eliminate the HIV-1 epidemic would be an effective prophylactic vaccine.

A.1.5 HIV-1 vaccine development

Since the discovery of HIV as cause of AIDS, the search for a vaccine had been a top priority among researchers. Against initial expectations that this would be achievable within few years, an effective, prophylactic vaccine remains elusive even today, 33 years later. Main obstacles for researchers are the unusually high genetic variability of HIV-1 (A.3.3), uncertainty about the correlates of protection, the lack of adequate animal models, and the difficulty to develop highly immunogenic antigens. Of the few phase IIb/III clinical studies performed to analyze the efficacy of HIV-1 vaccines concepts, most showed sobering results. The earliest trials, VAX003[41] and VAX004[42], where a mixture of different gp120 proteins was administered, aiming to elicit a protective antibody response, showed no efficacy. The STEP (HVTN 502)[43] and the similar Phambili (HVTN 503)[44] phase IIb studies tested an adenovirus type 5 (Ad5) viral vector-

delivered HIV-1 gag/pol/nef antigen combination as cell-mediated immunity vaccine. However, neither approach caused protection. On the contrary, in Ad5 seropositive and uncircumcised men, the incidence of infection even increased. A following study, HVTN 505, that also employed Ad5 as viral vector was halted at interim-analysis for futility[45]. The only clinical study that showed modest and transient protection from HIV-1 infection was the phase III RV144 (Thai) trial[46]. Through a heterologous prime (recombinant canarypox vector coding for Env, Gag, and PR) and boost (recombinant gp120 protein) regimen a 31.2% protection-efficacy at 42 months (with a p-value of 0.04) as primary endpoint was achieved. The greatest protection was observed 1 year after vaccination (60% efficacy), but the effect decreased over time. Assessing the immune correlates for protection indicated that the risk of HIV-1 infection correlated inversely with humoral immune responses against the V1V2 regions of Env and correlated directly with Env-specific, plasma IgA antibodies[47]. Also, Env-specific CD4+ T cell responses and antibodies eliciting Fc-mediated antibody-dependent cellular cytotoxicity (ADCC) responses to Env have been connected to reduced risk of acquisition[48]. Between vaccinated and unvaccinated persons, who got infected, no difference concerning viral loads was observed, possibly, because nearly no CD8+ T cell responses were elicited through the RV144 vaccination.

A.1.5.1 Broadly neutralizing antibodies

Although not identified as a correlate of protection in the RV144 trial, nowadays most HIV-1 vaccine research focuses on the induction of broadly neutralizing antibodies in vaccinees. In natural infections, these bNAbs evolve in about a fifth of all infected persons, through continuous reciprocal evolution of HIV-1 antibodies and the virus: antibodies against Env drive viral escape through mutations, which in turn induces antibodies against the mutated virus and so on. Eventually, in some cases, antibodies against conserved epitopes are elicited that can neutralize multiple HIV-1 viral strains. Despite their broad specificity, bNAbs shaped during primary infection have little to no effect on disease progression in patients, due to rapid viral Env adaptations that prevent neutralization, or even recognition (viral escape)[49]. However, inclusion of bNAbs in HIV-1 immunotherapy to suppress viremia, generated some promising results in humanized mice[50] and preliminary human studies[51]. Resistant viral strains emerged in humans within 28 days though, showing that this is not yet a therapeutic option and also illustrating how fast escape mutations are generated. The main focus is therefore on a prophylactic vaccine that elicits high titers of bNAbs to eliminate the virus before an infection is established. This hypothesis is mainly based on the fact that passive immunization with monoclonal, broadly neutralizing, Env-specific bNAbs can completely protect non-human primates (NHP) from challenge with pathogenic chimeric simian/human immunodeficiency viruses (SHIV), expressing HIV-1 Env[52,53]. bNAbs often exhibit unusual features, like a long heavy chain complementarity determining region 3 (CDR3), poly-reactivity towards non-HIV-1 proteins, or a high frequency of somatic mutations that often take years to develop and are substantial obstacles to their elicitation through vaccination[54,55]. No vaccine regimen so far has been able to reliably induce the production of such bNAbs. Research therefore aims at improving the immunogen design, like for instance the development of native-like trimers with higher affinity to bNAbs, and mimicking the bNAb maturation process in patients to elicit bNAb responses[56].

A.1.5.2 Vaccine-induced HIV-1 T cell responses

In natural HIV-1 infections, T cell responses are the first adaptive immune responses to occur. Although these *de novo* T cell responses cannot eradicate the virus, especially CD8+ T cell responses help to control the virus replication and reduce the viremia during acute infections[57]. Some HIV-1 infected persons, called long-term nonprogressors (LTNP), can even suppress the virus for up to 30 years without treatment, therefore delaying the development of AIDS. This control in LTNP is often associated with the presence of high-quality CD8+ T cells that are polyfunctional, highly-efficiently kill HIV-1-infected cells and are reactive against more epitopes (=enhanced breadth), as well as exhibit a higher cross-reactivity to epitope variants (=enhanced depth) as compared to disease progressors[58]. A prophylactic vaccine that elicits CD8+ T cell memory (A.2.2.3) could respond to a primary HIV-1 infection with an early, high magnitude and broad CD8+ T cell effector response that might more effectively inhibit viral infection and limit establishment of viral escape and latent reservoirs. Whether T cell memory responses could contribute to a sterilizing immunity against HIV-1 spawned some controversy. For example, HIV-exposed seronegative (HESN) individuals developed robust Gag-specific CD8+ T cells that might have reduced the infection risk[59,60]. There is however no clear evidence that these CD8+ T cell responses are a correlate of protection in these individuals[61]. Archiving sterilizing immunity through T cells seems difficult, since most vaccine-induced T cells require an anamnestic response. Therefore HIV-1 is confronted by the T cells only after the infection is established[62], which helps to reduce viral loads but does not achieve eradication.

The Gag protein is the main target for a cellular HIV-1 vaccine, because of all HIV-1 proteins, mainly Gag specific T cell responses have been associated with a reduced viremia[63]. Additionally, although Gag is not expressed by early genes in a natural infection, the high amount of protein incoming from the viral particle is sufficient for CD8+ T cells to recognize infected cells as early as two hours post-infection, even before proviral genome integration and viral protein synthesis[64]. This quick response could lower the acute phase viremia and control the infection before the appearance of escape mutations. The viral function of Gag and its use as an antigen for vaccination is described in detail in chapter A.3.

A.2 Cell-mediated immune responses

A.2.1 Antigen processing and presentation on MHC class I

To be recognized by host T cells, viral antigens have to be processed and presented on the cell surface by major histocompatibility complex (MHC) class I or class II molecules. In the classical pathways exogenous antigens are presented by MHC class II molecules, which interact with CD4+ T cells, and antigens of intracellular origin by MHC class I molecules, able to activate CD8+ T cells. These interactions are, however, not mutually exclusive, and in APCs, exogenous antigens can be loaded onto MHC class I through cross-presentation (A.2.1.2). Also, endogenous proteins can be presented by MHC class II when they are degraded by autophagy[65]. Since priming of a protective CD8+ T cell response is dependent on efficient antigen processing and MHC class I presentation, those molecular pathways are described in detail in the following chapters.

A.2.1.1 Classical MHC I pathway

MHC class I molecules are expressed by all nucleated cells, in contrast to MHC class II complexes that under normal conditions are only associated with APCs. To be presented through the classical MHC class I pathway, antigens have to be located in the cell cytoplasm or nucleus. If these endogenous proteins reach the end of their functional live or are defective due to erroneous transcription or translation, so called defective ribosomal products (DRiPs)[66], they get degraded by the ubiquitin-dependent proteasome[67]. The decomposition of DRiPs is the reason that antigen presentation kinetics are not necessarily correlated with the protein half-lives[68] and stable viral proteins can be presented immediately after infection and first translation. The proteasome is a protein complex composed of a 20S core barrel with two 19S cap proteins at both ends. In immune cells or under immune activation, subunits within the 20S barrel that form the catalytic center, responsible for peptide bond cleavage, are substituted, altering degradation patterns[69]. This so called immunoproteasome caused a better antigen presentation through selectively generating more immunogenic peptides or simply by a higher turnover, producing larger amounts of peptides[70,71]. Proteasome-degraded peptides that possess a length between 8 to 16 amino acids can be transported by the transporter associated with antigen presentation (TAP) into the ER[72]. TAP is a heterodimeric (consisting of TAP1 and TAP2) ABC transporter that forms a pore across the ER membrane and translocates peptides from the cytosol to the ER lumen, driven by ATP hydrolysis[73,74]. Directly after entering the ER, peptides with a length of 8 to 11 amino acids and a suitable binding motif are loaded onto immature MHC class I complexes. Longer peptides are first further trimmed N-terminally[75] by the two human ER aminopeptidases (ERAP1 and ERAP2) to a length of not less than 8 to 9 amino acids[76,77].

The immature MHC class I complex, consisting of a polymorphic (A.2.1.3) heavy/α-chain and the invariant light chain beta-2-microglobulin (β2m), is only stable after a high affinity peptide has been bound to the peptide-binding groove, which is made up by the α_1 and α_2 domains of the heavy chain. Until appropriate peptide binding, the heavy chain and β2m intermediary complex is stabilized by the cellular chaperone proteins tapasin, calreticulin, and ERp57. All these proteins are associated through tapasin to TAP, and the whole assembly is called the peptide-loading complex (PLC). Through binding of a sufficiently stabilizing peptide the chaperones are released and the mature peptide-MHC class I complexes (further called pMHC) are transported through the Golgi apparatus to the cell surface, presenting the peptide, called epitope, to CD8+ T cells. MHC class I molecules that fail to connect to a suitable peptide are transported back to the cytoplasm through the ERAD (ER associated protein degradation) machinery for proteasomal degradation[78].

A.2.1.2 Cross-presentation

Although all nucleated cells can present viral epitopes on MHC class I molecules, only professional APCs are able to prime naïve antigen-specific CD8+ T cells (A.2.2.1). If these APCs get directly infected, they can readily present epitopes through the classical direct presentation pathway (A.2.1.1). However, if a virus does not target APCs directly, the pathogen has to be taken up exogenously and presented on MHC class I through an alternative pathway called cross-presentation. The APC population mainly involved *in vivo* in cross-presentation are dendritic cells (DCs)[79]. Most molecular and cellular mechanisms associated with cross-presentation are poorly understood, but there seem to be two main pathways, by which DCs can present exogenous antigens on MHC class I molecules: the cytosolic and the vacuolar pathway (Figure 4).

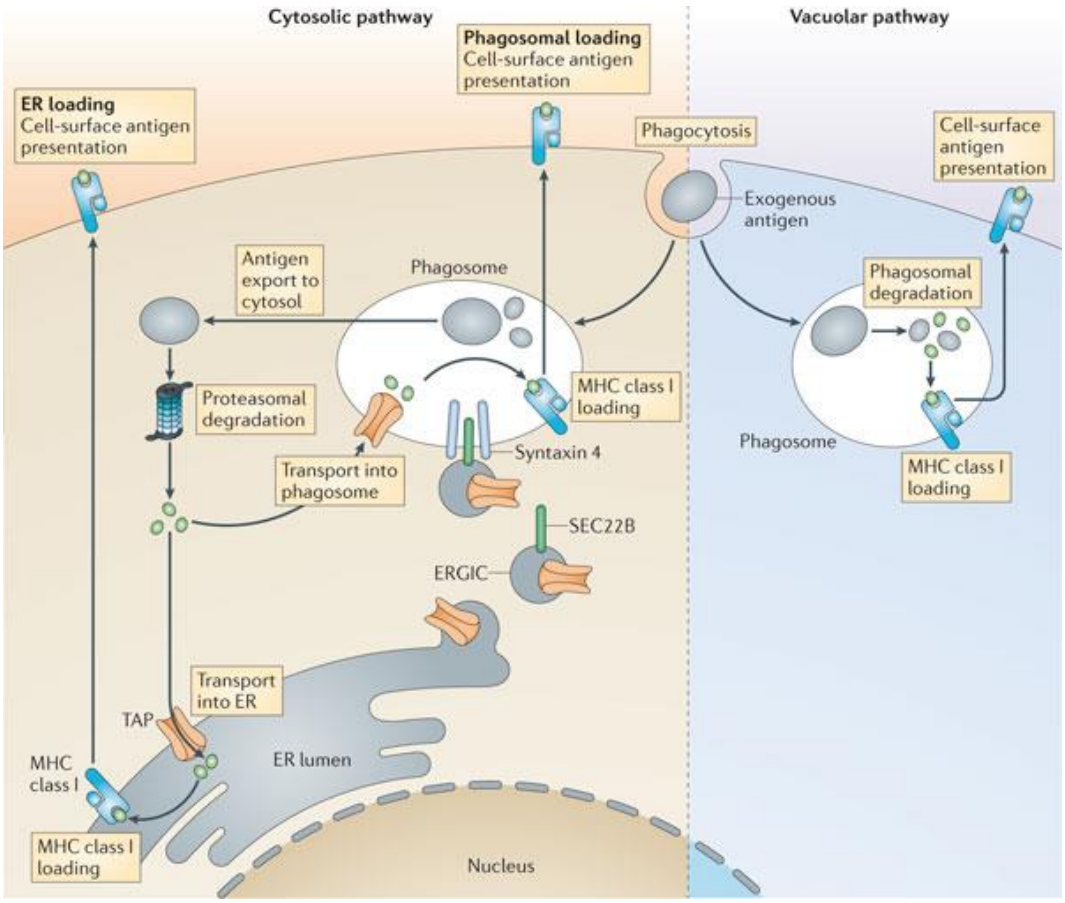


Figure 4. Antigen cross-presentation in dendritic cells (Reprinted by permission from Macmillan Publishers Ltd: Nat Rev Immunol. 12(8):557-69,[80] ©2012). After uptake of exogenous antigen, dendritic cells can cross-present it to MHC class I molecules through two main pathways: The cytosolic pathway (left), which makes use of the classical MHC class I processing machinery, or the vacuolar pathway, where antigen loading takes place in the phagosome (see text A.2.1.2).

In the cytosolic pathway exogenous antigen gets internalized via phagocytosis or macropinocytosis and can then access the cytosol. The transport of antigens from endocytic compartments into the cytoplasm is most probably achieved through the ERAD machinery that is recruited to the phagosomes and endosomes[81]. Pathogen-derived antigens are then degraded by the proteasome and enter the normal direct-presentation pathway as described before (A.2.1.1). Besides normal MHC class I loading by the PLC after transport via TAP into the ER, it has been proposed that degraded antigen peptides can reenter the phagosomes and are loaded there onto class I complexes that are subsequently transported to the cell surface[82]. For this alternative loading in the cytosolic pathway, TAP and the PLC are recruited to phagosomes and endosomes.

The vacuolar cross-presentation pathway is independent of antigen entering the cytoplasm. Antigen uptake, degradation, and pMHC loading all take place in the endocytic compartments. The MHC class I heavy chain and β2m in the phagosome originate thereby either from recycled complexes from the plasma membrane[83] or through recruitment of novel molecules from the ER[84].

DCs seem to be the preferred APC for cross-presentation, because of the low concentration of lysosomal proteases and their reduced activity in the endocytic compartments due to a higher pH value. This limited lysosomal antigen degradation correlates with efficient cross-presentation, since it retains potential MHC class I epitopes[80,85].

A.2.1.3 Human MHC class I polymorphism

In humans, the MHC is called human leukocyte antigen (HLA) and the class I heavy chain genes are encoded by three loci (HLA-A, -B, and –C), located on the short arm of chromosome 6. HLA heavy chains are highly polymorphic and are among the most variable human genes, with already 11,100 different class I alleles being listed[b]. The polymorphic residues cluster especially in the region of the peptide binding grooves, resulting in different epitope-binding motifs for distinct HLA alleles. For most alleles, the motif is defined by two conserved anchor amino acids[87], located at the second (P2) and at the last position of the epitope (PΩ). These anchors, in combination with hydrogen bond interactions of the N- and C-termini[88] of the epitope, are responsible for mediating sufficient binding to the peptide binding groove. Depending on the allele, specific secondary interactions of residues from the epitope with the HLA molecules, so called secondary anchors, also influence binding affinity[89,90]. The remaining amino acid composition of the epitope mostly follows no observable hierarchy.

Since a single person can have up to 6 different class I HLA alleles, the presentation of a broad range of different epitopes is possible, aiding in the eradication of infections. These combinations of HLA alleles, called HLA haplotype, are highly variable, enhancing the population-wide protection by presenting a wide range of epitopes. But this diversity also hampers the development of universal T cell antigens, especially for highly variable pathogens like HIV-1, because the set of cell-surface-presented epitopes is highly diverse from person to person.

On a typical nucleated cell, there are around 1-5 x 10^5 HLA class I complexes presenting around 10^4 distinct epitopes. Epitope lengths range from 8-11 amino acids, but around 95% are 9-mers. The number of one specific epitopes ranges from 1 to 10,000 per cell, with a mean of about 50[91]. The 50 most abundant epitopes, however, already fill 27% of all HLA class I molecules[92].

A.2.1.4 Role of HLA class I presentation during primary HIV-1 infections

In a primary HIV-1 infection an effective CD8+ T cell response is the major determinant for a low viral load set point. Targeting epitopes, where virus escape is accompanied by substantial viral fitness loss, reduces the virus load. The HLA haplotype heterogeneity of the general population results in variable presentation of susceptible epitopes and is therefore a major determinant for HIV-1 control[93].

Certain alleles like HLA B*27[94], where escape is accompanied by a dramatic fitness loss[95], or B*57[96] are frequently found in LTNPs and confer control of virus replication. Contrary, persons with HLA alleles like B*35 or with class I homozygosity, which reduces the breadth of possible epitopes, are susceptible for rapid HIV-1 disease progression to AIDS[97].

To inhibit CD8+ T cell mediated killing, HIV-1 has evolved an immune evasion mechanism that modulates the HLA pathway. The HIV-1 protein Nef diverts the trafficking of HLA-A and -B molecules from the Golgi apparatus to lysosomal compartments for degradation[98], thereby downregulating HLA surface presentation, which inhibits effective epitope display. In contrast, HLA-C and HLA-E presentation is not altered by Nef, to avoid lysis through natural killer cells, which recognize altered or infected cells with reduced HLA expression[99].

[b] http://hla.alleles.org/nomenclature/stats.html[86] - Last updated: 13.07.2016

A.2.1.5 *In vitro* antigen processing and presentation analysis

Since MHC-class-I-presented peptides derive from intracellular proteins, they can be seen as a fingerprint of the cells constitution. Alteration of the cell composition, for example through viral infections or transformation, also changes these fingerprints, leading to the presentation of different peptides on the cell surface. Scrutinizing CD8+ T cell can recognize these alterations and kill the aberrant cells (A.2.2). Knowledge of these MHC-class-I-presented epitopes, referred to as the immunopeptidome, could help to develop efficient targets for therapeutic tumor approaches and provide valuable information for T cell vaccine design. Easy identification of epitopes would also constitute an excellent tool to validated antigen candidates and check, if epitopes are processed and presented efficiently, and whether these are identical to those observed in natural infections. An easy way to identify the presented and immunogenic peptides of an antigen would be to administer it *in vivo* and to read out the primed T-cell responses, for example with synthetic peptides. For antigens that are designed for use in humans, this however, often is not an option, since an adequate animal model is missing. Another approach would be to use epitope-specific CD8+ T cell clones, recognizing individual pMHCs on cells. But this method is limited by the low number of readily accessible clones and the laborious work to identify all presented epitopes.

An elegant way to circumvent these problems is to directly isolate and identify the naturally processed epitopes. Several different approaches for epitope isolation have been published including acidified cell lysates[100,101], elution of peptides from the cell surface by mild acid washing, which destabilizes pMHCs and releases the peptides[102], and immunoaffinity chromatography purification of pMHCs. The immunoaffinity method exhibits the highest sensitivity and is therefore the most prevalent nowadays. This technique exploits antibodies that bind either to specific or a broad spectrum of MHC alleles, to isolate them from a mixture of proteins. The most widely used antibody in human systems is the monoclonal mouse antibody W6/32, which recognizes most HLA A, B, and C alleles[103]. Additionally, this antibody ensures isolation of peptide bound HLA class I complexes, since it binds at the interface of β2m and the heavy chain of the heterodimer that is only stable if also a peptide is bound in the peptide binding groove[104]. The complexes can be either isolated from detergent-solubilized cell lysates[87] or directly from the conditioned medium, if soluble variants of MHC class I molecules (sMHC or for human variants sHLA) are introduced into the cells[105]. For the latter case, the heavy chain is deprived of its transmembrane domain, inhibiting anchorage in the plasma membrane, as infrequently found *in vivo* for the non-classical HLA-G[106] and classical HLA proteins[107,108]. In these cases, after peptide loading in the ER the soluble pMHC complexes get secreted actively[109] into the extracellular environment. Compared to MHC isolation from cell lysates this has the advantage that it can be harvested from the same cells multiple times, by just exchanging the medium, and that it allows to focus on a single MHC allele and not just get a pool of peptides originated from the cells' haplotype. Between immunopeptidomes presented on membrane-bound and soluble HLA molecules, substantial overlap was observed[110,111], indicating that the use of sHLA molecules allows the identification of naturally processed CD8+ T cell epitopes. Additionally, C-terminally added epitope tags to sHLA do not alter the immunopeptidome[112], but allow purification irrespective of a suitable HLA-specific antibody, that in some cases might alter peptide composition.

After isolation, the amino acid sequences of the peptides are determined by tandem mass spectrometry. In this process, the peptides are first separated by liquid chromatography and then detected in a mass spectrometer according to their mass-to-charge ratio. For the sequence

analysis, peaks of the chromatogram are selected for fragmentation in a second mass spectrometry analysis. To identify the epitopes, the fragmentation pattern can be used for a search against possible epitopes from a set of proteins, like all human proteins or the protein of interest. The first use of mass spectrometry to identify MHC-bound peptides was described by Don Hunt and Vic Engelhard[113], and with methodical advances in mass spectrometry it became suitable for broad application. Nowadays the sensitivity is high enough to realize detection of peptides down to attomole levels, which facilitate the theoretical identification of complete immunopeptidomes. Affinity purification combined with mass spectrometry have already been used to identify the endogenous self-epitopes presented in the absence of any alteration[92,110,114], infection-specific epitopes[115–119], tumor-specific epitopes[120], infection- and tumor-induced changes in the presentation of self-epitopes[121–124], and epitopes presented by vaccine candidates[125], partly using soluble HLA proteins[110,119,121,122,124] as peptide source.

A.2.2 CD8+ T cell immune response

In the previous chapter the pathways by which endogenous and exogenous antigen can be loaded onto MHC class I molecules were described. If these pMHC molecules display pathogen-derived epitopes, scrutinizing naïve CD8+ T cells could recognize it, get primed (A.2.2.1) and consequently differentiate to effector T cells that kill infected cells (A.2.2.2) or to memory T cells that prevent the host from a recurrent infection with the same pathogen (A.2.2.3).

A.2.2.1 Priming of naïve T cells

Naïve CD8+ T cells originate in the thymus and are specified by a membrane-bound T cell receptor (TCR). The TCR is a heterodimer, normally consisting of an α and β chain, and is generated through extensive gene rearrangement in the TCR loci. Since the diversity of possible TCR rearrangement combinations is with about 10^{15} far higher than the number of T cells within a person, each cell most likely expresses a unique TCR[126].

T cells undergo two rounds of selection before leaving the thymus as mature naïve CD8+ T cells[127]. First, a positive selection, where only CD8+ T cells that show low avidity towards self-peptides presented on MHC class I receive a pro-survival signal, thereby eliminating non-functional TCRs that show no reactivity against MHC molecules. Second, in the negative selection round auto-reactivity is prevented by inducing apoptosis in all T cells that are highly reactive against self-peptides. The mature, naïve CD8+ T cells migrate through peripheral lymphoid tissues and highly efficiently scrutinize pMHCs on APCs. The APCs primarily involved in priming of naïve CD8+ T cells are DCs, which can either get infected with pathogens and thereby directly present pMHCs (A.2.1.1) or take up exogenous antigens and cross-present them[80] (A.2.1.2). Initial contact between APCs and T-cells is mediated by cell-adhesion molecules. Next, the TCR complex, consisting of the variable α- and β- T cell receptor chains and the invariant signaling proteins CD3 and a ζ chain-homodimer, samples the pMHC. If suitable, the variable part of the TCR recognizes the presented antigenic peptide through the highly variable CDR3 loops, as well as the MHC molecule with the more conserved CDR1 and CDR2 loops[126]. The interaction is stabilized by the heavily glycosylated disulfide-linked heterodimeric CD8 co-receptor expressed by CD8+ T cells, that binds to invariant sites of MHC class I molecules. The pMHC-TCR interaction induces cell plasma membrane convergence, which ensues spatial exclusion of the inhibiting CD45 phosphatase and most probably is the reason T cells get activated[128]. In this kinetic segregation model, the kinase Lck at the inner

leaflet of the membrane phosphorylates immunoreceptor tyrosine-based activation motifs (ITAMs) of CD3 and ζ chains, thereby activating the TCR complex.

Importantly, in addition to the binding of the TCR to pMHC, a secondary signal is needed to prime naïve T cells. For CD8+ T cells, this is mainly the co-stimulatory receptor CD28, which gets activated by binding of DC cell surface proteins CD80 or CD86. These molecules are, for example, highly expressed on DCs that matured through active infection, therefore building a bridge to innate immunity and activation of DCs, for example by TLR signaling. If there is no co-stimulus, antigen recognition alone induces functional inactivation and clonal deletion of peripheral T cells. This ensures that T cells are only primed during ongoing infections, and is a fact that has to be considered for any vaccine design. Downstream signaling after activation of the TCR complex and CD28 causes translocation of selected transcription factors to the nucleus. This in turn leads to transcription of genes, most importantly IL-2 that drives T cell proliferation, clonal expansion and differentiation. Besides CD28 activation additional help, provided by CD4+ effector T cells that recognize MHC class II presented antigen on the same APC, is required most of time for thorough priming of CD8+ T cells. On the one hand, the CD40 ligand from CD4+ T cells binds to CD40 on DCs, which increases the expression of CD80 and CD86. On the other hand activated CD4+ T cells directly help to prime CD8+ T cells by producing abundant IL-2.

The complete so-called primary cell-mediated immune response primes the CD8+ T cell to differentiate into effector (A.2.2.2) or memory cells (A.2.2.3), both recognizing the same pMHC as the original naïve T cell[129].

A.2.2.2 CD8+ T-cell-mediated cytotoxicity

Primed naïve CD8+ T cells proliferate and differentiate to cytotoxic effector T cells, also called cytotoxic T lymphocytes (CTL). These effector cells are guided by chemokines and newly expressed adhesion molecules to the site of infection and scrutinize pMHCs. As soon as they encounter the appropriate pathogen-derived pMHC, they respond quickly, and efficiently kill the infected cell without the need of any co-stimulatory signal. The CTL are therefore able to act on any virus-infected cell. The interaction between pMHCs and the TCR complex together with additional nonspecific adhesion molecules forms a tight “immunological synapse” between the CTL and the target cell, which enables highly selective targeting of effector molecules to the infected, antigen bearing cell. Stimulated CTLs release cytokines like IFNγ, which can inhibit viral replication and increase MHC expression as well as antigen processing in the target cell. The main mode of operation is, however, to kill the infected cell by release of the cytotoxins perforin, granulysin, and granzymes that are stored in granulae. Perforin forms a pore in the target cell, which allows granzymes to enter the cytoplasm of the infected cell. Granzymes are a family of serine proteases that induce apoptosis in the target cell. Most notable, granzyme B is responsible for cleavage and activation of caspase 3 and also induces release of cytochrome c from the mitochondria, both events being gateways for programmed cell death. Granulysin has antimicrobial activity and might additionally induce apoptosis. Cells that undergo cell death through the cytotoxic armory of the CTL are rapidly ingested by phagocytic cells. Killing of the infected cells allows CTLs to detach and search for new targets[129].

A.2.2.3 Memory CD8+ T cell response

The process from priming of naïve T cells that induces clonal expansion, differentiation to effector cells, and finally enables eradication of an infection requires several days. Pathogen clearance is accompanied by apoptosis of most effector cells and generation of a small set of memory T cells. Although memory CD8+ T cells seem to require some CD28 co-stimulation to get fully activated[130], they are far quicker and more efficient at killing infected cells in response to an antigen stimulus, than naïve T cells. These memory T cells responses are long-lived with an approximate half-life between 8 and 15 years and are the hallmark of adaptive immunity that protects the host from subsequent challenge by the same pathogen[129].

CD8+ T cells can either differentiate into effector memory (T_{EM}) or central memory T cells (T_{CM})[131]. Effector memory T cells can upon restimulation immediately exert effector functions, thus quickly killing target cells, but they show only limited expansion potential[132]. Contrary to the T_{EM}-mediated protective memory, the central memory T cells represent the reactive memory reservoir[133]. T_{CM} reside in the T cell zones of lymphoid tissues, have no immediate effector function, but are highly sensitive to antigenic stimulation, which readily induces proliferation and differentiation into effector cells. T_{CM} show a high expansion potential[132], but their anamnestic response-driven effector functions are delayed compared to the immediate action of T_{EM}.

Thus, the aim of an HIV-1 cellular-immunity vaccine would be to elicit a broad range of high-quality (A.1.5.1) memory T cells, preferable of effector memory phenotype, to control or even prevent viral infection[62].

A.3 The Gag protein and virus-like particles in vaccinations

A.3.1 HIV-1 Gag VLP assembly and release

The group-specific antigen (Gag) is the main structural protein of HIV-1 and the driving force in the assembly and release of viral particles (Figure 5). Gag is expressed as 55kDa precursor protein that gets processed by the virus PR into the proteins p17, p24 p7, p6 and the spacer peptides p2 and p1 only after viral release (A.1.3). Expression of Gag alone in a cell is sufficient for the formation and release of so called virus-like particles (VLPs)[134]. These VLPs resemble immature viral particles in size and form, but are, due to the absence of a viral genome and any other viral proteins, not infectious[134].

Because of its size and flexible linker between the domains, the structure of the precursor 55 kDa Gag has not been resolved yet. However, atomic-level 3D-structures of the subunits provide valuable information about Gag's functions during assembly of viral particles: The N-terminal p17 matrix domain of the Gag polyprotein folds into a globular structure (composed of five α-helices, a short 3_{10} helical stretch, and a three-stranded mixed β-sheet), with a C-terminal α-helix projecting away, to connect it p17 with the p24 domain[135]. Co-translational covalent attachment of a myristic acid moiety to the N-terminus of p17 targets Gag to the plasma membrane. The p24 capsid domain folds into two distinct, mostly independent domains. The largely helical N-terminal domain (NTD) and C-terminal domain (CTD) are connected by a flexible linker[136,137]. The CTD contains a dimer interface, the main determinant, which drives Gag multimerization, and the major homology region (MHR), which is necessary for particle assembly

and is highly conserved throughout all orthoretroviruses[138]. The p7 nucleocapsid structure contains two zinc finger domains that interact with the viral genomic RNA in natural HIV-1 infections, to recruit the RNA to the particle assembly site[139]. The Gag C-terminal domain p6 is largely unstructured[140] and recruits the endosomal sorting complex required for transport (ESCRT) to the assembly site, which catalyzes the membrane fission and therefore the particle release from the cells.

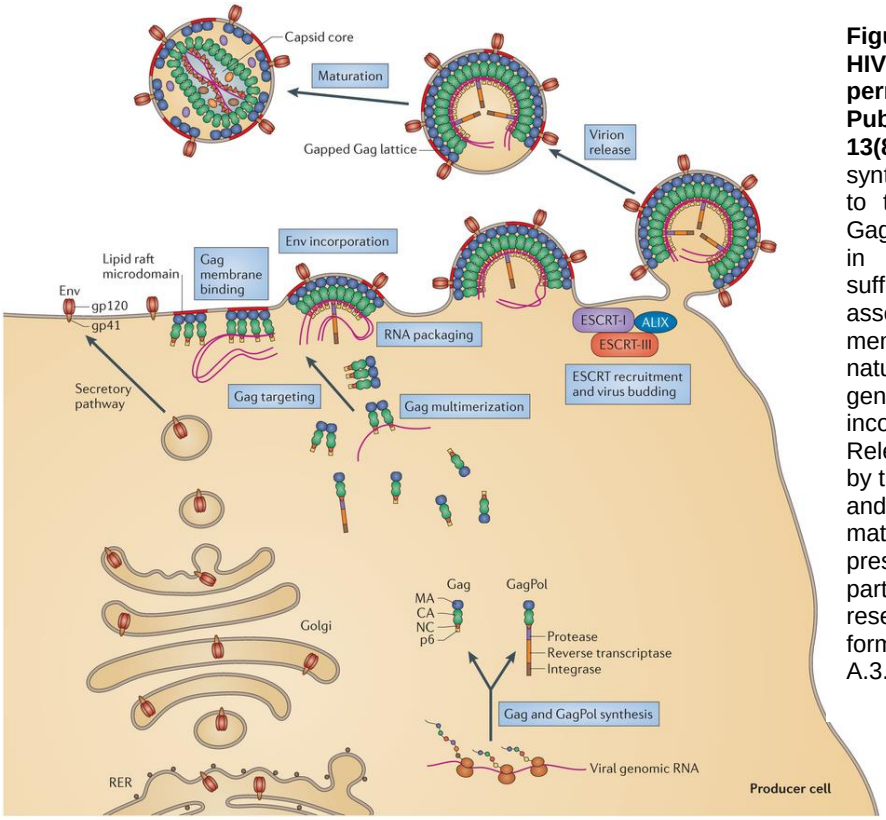


Figure 5. Post-integration phase of HIV-1 replication (Reprinted by permission from Macmillan Publishers Ltd: Nat Rev Microbiol. 13(8): 484-96, ©2015). Env is synthesized in the ER and transported to the plasma membrane, whereas Gag and GagPol synthesis happens in the cytoplasm. Gag alone is sufficient to induce viral particle assembly. It anchors in the plasma membrane and multimerizes there. In natural infections the viral ssRNA genome, GagPol and Env get also incorporated into the viral particle. Release of the particles is facilitated by the host-cellular ESCRT machinery and afterwards PR induces virus maturation by cleaving Gag. In the presence of only Gag virus-like particles bud from the cell, which resemble immature virions in size and form (details in the text of chapters A.3.1 and A.1.3).

For particle assembly the Gag polyprotein associates with the host cell's plasma membrane. The mechanism by which Gag targets only the plasma membrane and no other cellular membranes, is most probably linked to the ability of the myristic acid moiety to adopt two different conformations: a sequestered and an exposed one[141]. In newly translated Gag the myristic acid is concealed, but interaction between p17 and phosphatidylinositol‑4,5‑bisphosphate (PtdIns(4,5)P2), which is highly enriched at the inner leaflet of the plasma membrane triggers a conformation change[142,143]. Thereby, the myristic acid group gets exposed and facilitates interaction with the plasma membrane. A cluster of basic amino acids within p17 additionally strengthens the Gag interaction with the inner leaflet of the plasma membrane[135]. The integrated Gag recruits sphingolipid- and cholesterol-enriched membrane microdomains, so called lipid rafts, which form the platforms for particle assembly[144]. The next step necessary for budding is Gag multimerization at the assembly site. Although Gag-RNA interactions have a supportive role, the main and sufficient determinant for the Gag clustering is the CTD of p24. Cryo-electron tomography revealed that the Gag-multimeric lattice has a hexameric arrangement in VLPs. To form a closed lattice that accounts for the curvature necessary to form particles with a diameter of ~100-120 nm, Gag would also have to generate pentameric structures. But since Gag only forms hexamers, the continuous lattice includes small areas devoid of Gag and one large gap[145],

forming an incomplete, roughly spherical shell (Figure 6). The aligned Gag molecules are packed radially, with the N-terminal p17 located at the membrane and the C-terminus oriented to the particle-center[146]. HIV-1 virions, as well as VLPs, are both reported to contain roughly 2000 copies of the Gag protein[147,148]. To release the viral-particle, late domains (comprised of two distinct motifs: P(T/S)AP and YPXL, as well as one unknown domain[149]), located at the N-terminal region of p6, recruit the host-cellular ESCRT machinery. ESCRT consists of about 20 proteins that form several multiprotein complexes that are important for a range of cellular membrane fission processes. HIV-1 requires only a subset of these proteins, especially ESCRT-I and several subunits of ESCRT-III, for membrane abscission and viral-particle budding from the cell[149]. The precise mechanism by which the ESCRT machinery accomplishes this release is not understood, but ESCRT III might help to constrict the membrane at the budding site by assembling into circular spirals[150] and thereby facilitate release of the immature VLPs. Since PR is missing in Gag-only VLPs, no maturation - characterized by processing of Gag into its subunits and subsequent protein rearrangements - occurs.

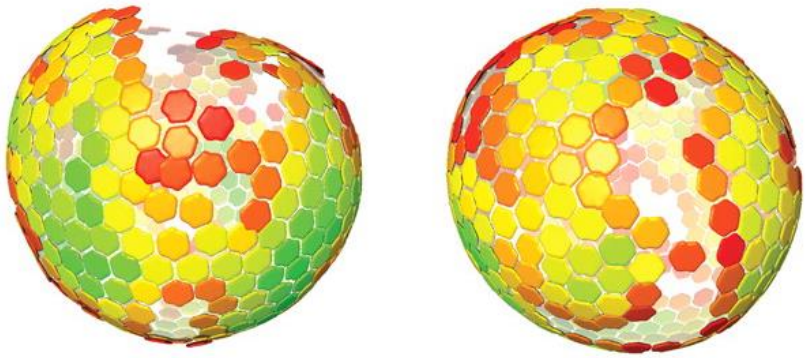


Figure 6. Global lattice map of Gag virus-like particles (adapted from Briggs et al.[145]). Gag protein clusters in a hexameric arrangement during particle assembly, which forms a continuous, but incomplete spherical shell. The picture indicates these Gag-hexamers and the areas void of any Gag.

A.3.2 Gag and HIV-1 VLPs as antigens

In natural infections, control of the viral load in LTNP and elite controller (EC, infected person without detectable viral loads) is often associated with robust cellular immune responses towards Gag[151,152]. Other HIV-1 proteins that induce a strong T cell response, like Nef, are not associated with lowered viremia[153]. Therefore, an optimal cellular-immunity vaccine, centered on Gag, that elicits a broad, poly-functional, and high-magnitude T cell response might control viremia and inhibit disease progression in vaccinees, once an HIV-1 infection is established. Moreover, a rapid memory response might mediate early control at the site of infection and thwart spread from the entry portal, possibly preventing HIV-1 infection from manifesting at all[154], as indicated by robust Gag-specific CD8+ T cell responses in HESN women[60] (A.1.5.2). Moreover in NHPs, the best available animal model to validate HIV-1 vaccine concepts, administering a prime-boost vaccine regimen, based on recombinant vesicular stomatitis virus (VSV) and an alphavirus replicon expressing SIV Gag and SIV Env, led to sterilizing immunity against high-dose challenges with SIVsmE660[155]. If only SIV Env was used for vaccination, no

protection was achieved, indicating that Gag, or the combination of Gag and Env, are required to elicit the sterilizing immunity[156], highlighting the importance of Gag.

As for HIV-1, a strong cellular immunity against Gag, but not Pol- or Env, has been implicated with control of SIV replication in NHP[157]. Such T cell mediated viral inhibition in NHP was most prominently shown when rhesus cytomegalovirus (RhCMV) was applied as viral vector to deliver the SIV proteins Gag, Rev-Tat-Nef and Env[158,159]. About 50% of all RhCMV/SIV-vaccinated monkeys that got intrarectally infected with the highly pathogenic SIVmac developed an early and stringent viral control. After initial viremia, the virus was not detectable any more, with exception of periodical, low level blips. These short periods of measureable viremia, however, gradually waned[158] and control has been stable in all but one of 17 protected monkeys, raising the possibility of ultimately eliminating the SIV infection through continuous immune surveillance[62]. Control was mainly accomplished by CD8+ T cell responses[158] in the absence of neutralizing antibodies[158,159]. In contrast to classical vaccinations, the CD8+ T cell memory responses did not have a T_{CM} phenotype, but the T_{EM} character for persistent pathogens prevalent, allowing a quick non-anamnestic effector response after primary infection[158,159]. Through the deletion of two genes (Rh157.5 and Rh157.4, which are coding for two proteins of the pentameric receptor complex involved in infection of nonfibroblasts[160]) from the RhCMV vector expressing Gag, the elicited CD8+ T cell responses turned out not to be classically MHC-class-I-restricted, but rather MHC-class-II-[161] or MHC-Ib/E-restricted[162].

During primary SIV infection these unconventionally restricted CD8+ T cells do not get efficiently primed, but effector T cells that were previously primed through RhCMV/SIVgag can get reactivated[162]. MHC-class-II- and MHC-E-restricted CD8+ T cells represent therefore an interesting target for vaccination, as they might also circumvent pathogen immune-evasion adaptations, like Nef-mediated MHC class I downregulation (A.2.1.4). As MHC-E is even up-regulated in SIV or HIV infected cells as a result of the evasion from NK-cell-mediated killing, such MHC-E-restricted responses may facilitate control[162]. Since CMV is species-specific, it is currently unknown, whether human CMV would also mediate control and prime such non-MHC-class-I-restricted CD8+ T cells.

Besides the excellent immunological characteristics, also the functional properties of Gag to form VLPs can be put to use. Compared to direct, soluble antigen, such particle-based antigens appear to be favorable for DC uptake via macropinocytosis and endocytosis[163] with subsequent antigen cross-presentation[164] and are capable of inducing better cellular and humoral immune responses[165]. The intake of a single HIV-1 VLP for example feeds about 2000 Gag molecules to a single cell, thereby helping to overcome antigen processing thresholds[70]. Efficient uptake and immune responses of baculovirus- or yeast-derived HIV-1 Gag-VLPs was observed in *ex vivo* experiments[166,167], as well as in mice[168,169] and NHP[170] immunized with Gag-VLP. In NHP, strong, broad, and long-lived CD8+ T cell responses were elicited, in the absence of any adjuvants. HIV-1 Gag VLPs also exhibit some adjuvants properties per se, through activating and maturating DCs[171], which is necessary for efficiently CD8+ T cell priming as outlined above(A.2.2.1).

As an alternative, besides directly applying VLPs as antigen, naked DNA or viral vectors that code for Gag can be administered. The addressed cells are then able to synthesize Gag proteins that can self-assemble into budding VLPs, which then can presumably prime CD8+ T cell responses through cross-presentation. The importance of functional Gag in such antigens has previously been demonstrated in our group. In a first antigen-generation, based on the CRF07_BC 97CN54 isolate[172], the possibility for Gag to form VLPs was, due to safety concerns,

inactivated by a mutation inhibiting addition of the myristic acid moiety (i.e. the G2A mutation). Additionally the natural frameshift within Gag was deleted and Gag was fused to an artificial, safety-optimized PolNef protein (i.e. IN deleted, PR inactivated, and a scrambled Nef replaced the active part of RT, that was put at the C-terminus), resulting in the read-through protein GagPolNef. This construct was administered together with gp120 in a heterologous DNA-prime and NYVAC boost regimen. In mice[173], as well as in NHPs[174], the combination was safe and highly immunogenic. These findings were confirmed in a clinical phase I trial (EuroVacc02), where reliably, poly-functional, and long-lasting T cell responses were induced[175]. However, the T cell responses were, as already observed in the NHP studies, unbalanced and mainly directed against Env and not the preferred Gag antigen.

Therefore, for the next antigen generation a more natural constitution was restored, by allowing myristoylation and therefore budding of Gag-VLPs. Additionally, the frameshift was reinserted, resulting in a Gag to GagPolNef ratio of 20:1, and gp140 was used instead of gp120 in order to elicit better antibody responses. This time, both, the Gag and the gp140, sequences were taken from C clade isolate 96ZM651. In mouse experiments, these combinations showed greatly improved T cell response ratios between Gag and Env[176,177]. In NHPs receiving these second generation antigens in a heterologous DNA prime and NYVAC plus gp120 protein boost[178] a $\log_{10}$ increased T cell response magnitude was obtained as compared to the first generation, with a balanced CD4+ T cell response. Importantly, mainly Gag-targeting CD8+ T cells were observed. These findings emphasize the importance to employ functional Gag in HIV-1 cellular-immunity vaccines.

Another feature that makes VLPs a favorable target for vaccine design is the possibility to incorporate other molecules like viral envelope proteins. The obvious first choice for an additional protein would be the HIV-1 Env. Env incorporated into VLPs resembles the natural HIV-1 virion structure and could be a combinatorial approach, to elicit both, protective broadly neutralizing Env-specific antibodies (A.1.5.1) and also broad, efficient Gag-specific CD8+ T cell responses, to control breakthrough infections. Gag from the VLPs could also contribute to an improved humoral response by intrastructural help of Gag-specific CD4+ T cells[179]. It was already shown that rabbits immunized with native, membrane-expressed Env trimers presented on Gag VLPs were able to elicit autologously Tier 2 HIV-1 neutralizing antibodies[180]. Other surface proteins could be used to broaden the cell tropism of VLPs, like the incorporation of VSV-G that increased Gag T cell responses in mice and lowered viremia in challenged NHPs[181]. Other molecules can be included to enhance the adjuvants properties of VLPs, as shown for GM-CSF and CD40 ligand in SIV Gag based VLPs[182].

In conclusion, there are many reasons to choose Gag as the main target for a cellular HIV-1 vaccine, and its ability to self-assemble into VLPs can be utilized to improve immunogenicity.

A.3.3 T cell antigens to address HIV-1's variability

The biggest obstacle in designing an effective cellular HIV-1 vaccine is the enormous viral diversity. It was shown that the HIV-1 variation in a single infected patient over the course of 6 years is equivalent to the worldwide variation of influenza virus in a year[183]. The high variability of HIV-1 is mainly due to the lack of proofreading functions of the HIV RT and the host cell DNA-dependent RNA polymerase II (RNA pol II), that synthesizes the viral genome. The combined error rate for both polymerases is approximately 2×10^{-5} per nucleotide in every replication cycle[184]. This feature, together with the early onset of HIV-1 replication, the vast number of

replications, influences of the host immune pressure[185,186], and the tolerance of a high degree of genome variability without loss of virus fitness, lead to the enormous diversity among HIV-1 strains and quickly establishes genetically distinct quasispecies in an infected person[12,62]. Amino acid diversity between HIV-1 subtypes (A.1.1) is on average 18% and within subtypes about 12%[187]. The variability is protein-dependent and highest for gp120, while Gag variability within a clade is on average below 10%[188]. This, for HIV-1, comparably low diversity makes Gag a preferable target for T cell vaccines, but nevertheless a potential antigen still has to elicit a broad response to be protective.

The STEP trial failed to induce protective cellular immune responses, but used only an inferior Gag antigen based on a single B clade isolate (CAM-1 - GenBank D10112)[43]. It was stated that a higher magnitude and especially a broader CD8+ T cell response is needed to confer immunity[189]. In the only clinical trial that showed moderate protection (RV144) T cell responses against Gag were not identified as a correlate of protection and no T cell driven sieve effect in HIV-1 breakthrough isolates was noted[190]. However, like for the STEP trial, only a single B clade Gag sequence from a natural isolate (LAI IIIB/Bru - GenBank A04321) was used for vaccination and the Gag T cell responses elicited were very low[46]. Both trials indicate that the T cell responses against Gag were not good enough and next-generation vaccines should aim to improve breadth and magnitude[191].

In the last years, some novel antigens that deal with the viral diversity were developed. One strategy is the use of consensus, ancestral, and center-of-tree HIV-1 sequences (described in detail in B.1.1.3). These sequences are designed to minimize the genetic distance in a given set of diverse viral sequences[192], thereby reducing the abrogating effect of viral variability, with the idea to cover a broad spectrum of viral strains.

A second approach is to focus CD8+ T cell responses only on conserved regions of HIV-1 and excluding highly variable sequences [193,194]. This would reduce the possibility for viral escape, since mutations in these conserved regions are detrimental for viral fitness. In a natural infection, these responses are rarely primed, because of immunodominant responses to variable epitopes, which facilitate viral escape. If CD8+ T cells however are primed against conserved epitopes through a vaccine, it might help to control the virus infection[195]. Conserved regions antigens were shown to elicit broad T cell responses in mice[196] and NHP[197] and vaccination against conserved regions of SIV induces even partial protection against SIVmac251 challenge[198].

A third, promising strategy to address the variability of HIV-1 is the mosaic approach[199]. Therein, a polyvalent set of HIV-1 protein sequences is designed, based on a set of naturally occurring viral sequences in a way to best cover the most frequently occurring 9-mers, which here are used as a surrogate for CD8+ T cell epitopes. In NHP studies mosaic variants of HIV-1 Gag, Pol and Env antigens augmented breadth, depth and magnitude of antigen-specific T cell responses compared to consensus or natural sequence HIV-1 antigens[200]. Since it was shown that a broad Gag-specific cellular immune response correlates with efficient viral control in NHP[201] and humans[63], Gag antigens based on the broadening mosaic approach, seem to be most favorable for a T cell vaccine. The mosaic design, however, does, like all other methods not account for the naturally processed epitopes that are excellent characterized for HIV-1.

A.4 Objective

As outlined in the precious section, there is reason to believe that a broad and potent Gag-specific CD8+ T cell response elicited through vaccination is important to control HIV-1 infection, as it might help to lower the virus set point through an early response[63]. Additionally, such responses might help to prevent establishment of latent infection altogether[154] or at least mediate a functional cure, as was for example shown for RhCMV-vector-induced T_{EM} responses to SIV[159]. This leads to the working hypothesis that a greater breadth and depth of presented HIV-1 Gag epitopes increases the chance of matching epitopes from a transmitted HIV-1 isolate and consequently the ability to control the virus.

Therefore, the objective of this work was to design Gag vaccine-antigens that elicit a broad CD8+ T cell response against immunologically potent Gag epitopes, while simultaneously preserving protein functionality, defined as budding of virus-like particles, which showed beneficial immunological characteristics. The antigens should contain as many of the most potent epitopes as possible. To ensure *in vivo* processing and presentation the set of epitopes should be based only on experimentally validated sequences that have been found in natural infections or vaccination studies.

Due to the high HIV-1 sequence variability, inclusion of epitopes often requires introduction of mutations compared to a natural Gag reference sequence. Since these amino acid substitutions (AAS) might compromise the protein's functionality and therefore reduce the release of VLPs, each of these mutations should be assessed for budding-preservation and be excluded if the AAS turns out to be detrimental in this regard. Since the high number of possible mutations excludes experimental classification into budding-competent and -detrimental ones, this discrimination should be done through an *in silico* implemented multidimensional classifier. Next, a quality score should be assigned to all compatible epitopes that takes various immunological parameters derived from the epitopes' metadata into account. For proof of concept antigens, presented in this thesis, the score of epitopes should be mainly computed based on their ability to benefit a global vaccine approach. Finally, the best epitopes should be combined in a polyvalent antigen set consisting of as few components as possible, since the practicability of a vaccine formulation consisting of more than three components is deemed very low[154]. To highlight the increased immunological potential of the designed Gag antigens they should be analyzed and compared to other antigens regarding number and quality of included epitopes, as well as their potential to cover the worldwide population, as well as the global virus distribution.

Following design, the Gag antigens should also be characterized biochemically for their ability to form VLPs that resemble wildtype particles in structure and size. Additionally, to assess the immunological breadth of these novel CD8+ T cell antigens, an *in vitro* method for direct interrogation of HLA class I presented epitopes through mass spectrometric LC-MS/MS peptide sequencing should be established.

B Datasets and computational methods

B.1 Datasets

B.1.1 HIV-1 Gag sequence sets

B.1.1.1 HIV-1 Gag sequence alignments

Various steps of the antigen generation, as well as in *in silico* validation required HIV-Gag sequence alignments. The Los Alamos National Laboratory (LANL) HIV Database (www.hiv.lanl.gov, Los Alamos, USA) stores the most comprehensive online collection of HIV sequences[202] and provides premade DNA or protein alignments, either of the whole genome or of specific genes or proteins.

Protein Alignment

For this work, the premade “Filtered Web Alignment” for the HIV-1 Gag protein from the year 2013 (Options: Organism: “HIV-1/SIVcpz”; Subtype: “All”; DNA/Protein: “Protein”) was downloaded in FASTA format. The Filtered Web Alignment is a subset of the Web Alignment, excluding sequences with large insertions, high content of ambiguity codes, and multiple frameshifts. The Web Alignment in turn is the most complete set offered by LANL, containing all sequences of the database, except that only one sequence per patient is included and very similar and problematic sequences are removed. For the purpose of this work, the downloaded alignment was manually edited by deletion of all group N, O, and P sequences, all SIV sequences (subtype CPZ), as well as by removal of all entries with unclassified (U) sequence elements.

Nucleotide Alignment

Besides the protein alignment, an alignment of Gag nucleotide sequences was required, too. For this the “Filtered Web Alignment” from the year 2013 was downloaded with the same specifications as for the protein alignment, except DNA/Protein selection was changed to “DNA”.

B.1.1.2 Reference sequence

For the alignment of epitopes and the identification of amino acid substitutions, a Gag amino acid sequence, hereafter called “reference sequence”, was necessary. Best suited for this is the sequence of HIV-1 subtype B isolate HXB2 described in 1985(GenBank accession number K03455)[203], because it is one of the most commonly used strains in many different kinds of functional and structural studies and also all HIV numbering positions in the HIV database are relative to the complete HXB2 genome[204]. The HXB2 Gag protein (GenBank accession number AAB50258.1) consists of 500 amino acids and comprises the domains p17 (amino acids 1-132), p24 (133-363), p2 (364-377), p7 (378-432), p1 (433-448), and p6 (449-500).

B.1.1.3 Other Gag Sequences

As benchmarks to compare newly designed antigens, various other Gag sequences that have been proposed as antigen candidates were collected. These were either specially designed artificial sequences, like consensus, ancestral, mosaic, or center-of-tree, or sequences of natural isolate that were previously included in phase IIb/III clinical trials.

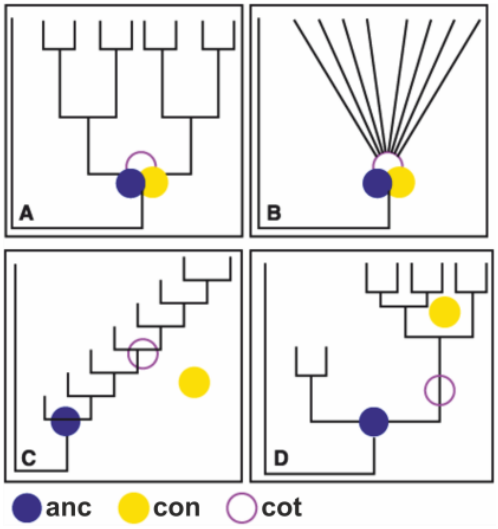


Figure 7. Visualization of different phylogenetic shapes and location of designed antigen sequences. All reconstructed sequences perform well at minimizing the genetic distance for symmetric phylogenies (A+B). For asymmetric phylogenies (C+D) only cot minimizes the distance satisfactorily. (From Science 299(5612):1515-8 ©2003. Reprinted with permission from AAAS).

Consensus (hereafter termed »con«), ancestral (»anc«), and center-of-tree (»cot«) variants of Gag all address the issue of viral diversity by calculating a sequence that minimizes the genetic distance in a given set of viral sequences[192] (Figure 7[205]). Consensus sequences are designed by choosing the most common amino acid at each position of a multiple sequence alignment (MSA). Consensus sequences of Gag based on various HIV-1 subtype sequence sets, and an M group consensus based on all subtype-specific consensus sequences (with same weighting for rare and common clades), were downloaded from the LANL as a premade protein sequence alignment as FASTA file (Options: Alignment type: “Consensus/Ancestral”; Organism: “HIV-1/SIVcpz”; Subtype: “All”; DNA/Protein: “Protein”; Year: “2004”)[206]. Ancestral sequences are computed to represent the most recent common ancestor of an MSA. Subtype-specific ancestral Gag sequences, as well as an M group ancestral sequence were included in the same alignment downloaded for the consensus antigen set[206]. Center-of-tree sequences are calculated by minimizing the phylogenetic distance of viral sequences of a sequence population. The design of a center-of-tree sequences based on subtype B was described before[207,208]. This sequence, together with subtype C and M group center-of-tree sequences, was kindly provided by Dr. James I. Mullins (University of Washington, USA). Mosaic sequences (»mos«) do not aim to minimize the genetic distance but to include as many naturally occurring 9-amino-acid peptide sequences (hereafter called »9-mer«) as possible in a protein sequence. Gag-specific mosaic sequences haven been published for subtype B and C, as well as for the M group[199]. They were designed as sequence sets of different size (combinations of up to six subtype-specific sequences), always aiming for maximal 9-mer coverage. Each set is to be used as an entity. Therefore the mono- and trivalent mosaic-sets were employed for comparison to the here presented antigen-sets that comprise the same numbers of sequences.

Lastly, for evaluation purposes, Gag sequences of natural isolates included in phase IIB/III clinical trials were collected. All of the four trials that included a Gag component in their vaccine used well-characterized subtype B isolates: CAM-1 (GenBank D10112) for the HVTN 502 (STEP) trial[43] and the HVTN 503 (Phambili) trial[44,209], LAI (IIIB/Bru; GenBank A04321) for the RV144 (Thai) trial[46,47], and HXB2 for the HVTN 505 trial[45]. Protein sequences of these isolates were downloaded from the LANL HIV sequence database using the search interface. The earlier conducted VAX004[41] and VAX003[42] clinical phase III trials, solely administered Env as immunogen in their vaccine regimens.

B.1.2 CTL/CD8+ T cell epitope database

In addition to the considerable collection of HIV-1 sequences the LANL HIV database also offers an "Immunology Database"[210], where experimentally validated HIV-1 cytotoxic and helper T cell epitopes, as well as antibody binding sites, from diverse publications are collected with annotations.

By using the "CTL/CD8+ T-cell epitope database" search interface, epitopes of interest for the herein described newly designed antigens were identified. Since there is no web service or interface for direct access, the set of epitopes was downloaded as HTML file and subsequently converted to a text-based comma-separated values (CSV) file. The final epitope list contains the epitope's sequence represented as string, together with other useful information, which are associated with the epitope entry (Table 1).

Table 1. Epitope associated attributes directly extracted from the LANL immunology database. (Attribute description: http://www.hiv.lanl.gov/content/immunology/search_help.html) [a]Attributes that are not directly relevant for antigen design, but are annotated in all entries and allow a more detailed epitope description. [b]Attribute data type: *int* represents an integer and *string* a text based data field.

Attribute	Type[b]	Description
Record Number[a]	int	A unique epitope identification number assigned by the database
Protein	string	Location of the epitope within the Gag polyprotein (three possible values: p17, p24 or p2p7p1p6)
Start	int	Start position in the defined protein
End	int	End position in the defined protein
Author Location[a]	string	Epitope location specified by the author
Epitope	string	Amino acid sequence of the epitope entry
Subtype	string	Subtype under study for the entry (not specified for subtype B)
Species[a]	string	Species in which the immune response to the epitope was stimulated
Immunogen[a]	string	Stimulus of the original immune response
HLA	string	HLA alleles presenting the epitope in question
Donor HLA[a]	string	HLA haplotype of the person in which the immune response was detected
Country[a]	string	Country where the samples were obtained (not specified for studies conducted in the United States)
Experimental Method[a]	string	Methods used to test the immunological response
Keyword[a]	string	Catchwords describing the HIV immunological study of interest
Notes[a]	string	Comments explaining the context in which an immune response to the epitope was found and what was learnt about the epitope in question
References[a]	string	Associated primary publications

Besides the directly accessible features for an epitope, each entry was also examined manually for additional, immunologically valuable information (Table 2). These were either gathered by checking the textual additional notes by the authors, the primary reference or by other epitope metadata as described below. The epitope list was screened for erroneously recorded sequences and false annotations by checking the primary publications given for a questionable database entry. If possible the entry was patched or else deleted.

Table 2. Additional epitope information extracted from the metadata. [a]Attribute data type: *double* represents a floating point number and *Boolean* a Boolean data field that only can be one of two values: false or true.

Attribute	Type[a]	Description
Response in LTNP	Boolean	True if the epitope is associated with responses in LTNP
% response	double	Percentage of persons in the study which show an immune response to the epitope. Normalized to the frequency of the presenting HLA allele (B.2.2.5)
Conservation	double	Conservation score of the epitope (B.2.2.4)

The epitope list had to be further processed due to two special issues (Figure 8):

(1) Because the extracted data from the LANL immunology database is a collection of information gathered from various studies and primary publications, some epitope entries are redundant, i.e. they may describe peptides with identical sequence and location. These database entries were merged into one consensus entry by unifying all available relevant information. In case of information given as a string value (HLA and subtype) all information were summarized in one joint record. For numerical values (%response and conservation), the mean of all entries with non-empty fields for this value was calculated. For the LTNPs Boolean field, a majority voting, i.e. selection of the alternative for which most entries are specified, was performed in case of contradicting information.

(2) In some cases, epitopes (subepitope) are fully embedded in another epitope (superepitope). Since the subepitopes are automatically included if the respective superepitope is incorporated into a new antigen (but not the other way round), all immunologically important characteristics from the subepitopes were added to the superepitopes. This was done in the same way as described for the handling of identical epitopes in (1), except for that the subepitope was kept with its original data as separate entry in the input data set.

Figure 8. Alignment of CD8+ T cell epitopes to the HXB2-Gag reference sequence, duplicate merging, and sub/superepitope attribute transfer. All epitopes of the input data set were aligned to the HXB2-Gag reference amino acid sequence. (1) In case of multiple entries for the same epitope (i.e. identical string sequence and matching start/end positions) epitope-specific attributes were merged into one consensus entry. (2) If an epitope (subepitope) was fully embedded in another epitope (superepitope) all attributes of the subepitope were added to the superepitope, but not the other way round. Both epitopes were kept in the input set.

B.2 The *Optimizer Algorithm*

The first version of the *Optimizer Algorithm* for designing epitope-enriched HIV-1 Gag T cell antigens was developed by Matthias Pfeifer[211]. It consists of four distinct, successive steps (Figure 9): (1) All input epitopes are mapped to a given reference sequence and all amino acid mismatches are identified. (2) Each of these epitope-associated amino acid substitutions (AAS) is then separately analyzed, to determine if the integration of the mutation would result in a functionally impaired protein with defective or reduced VLP budding capacity.

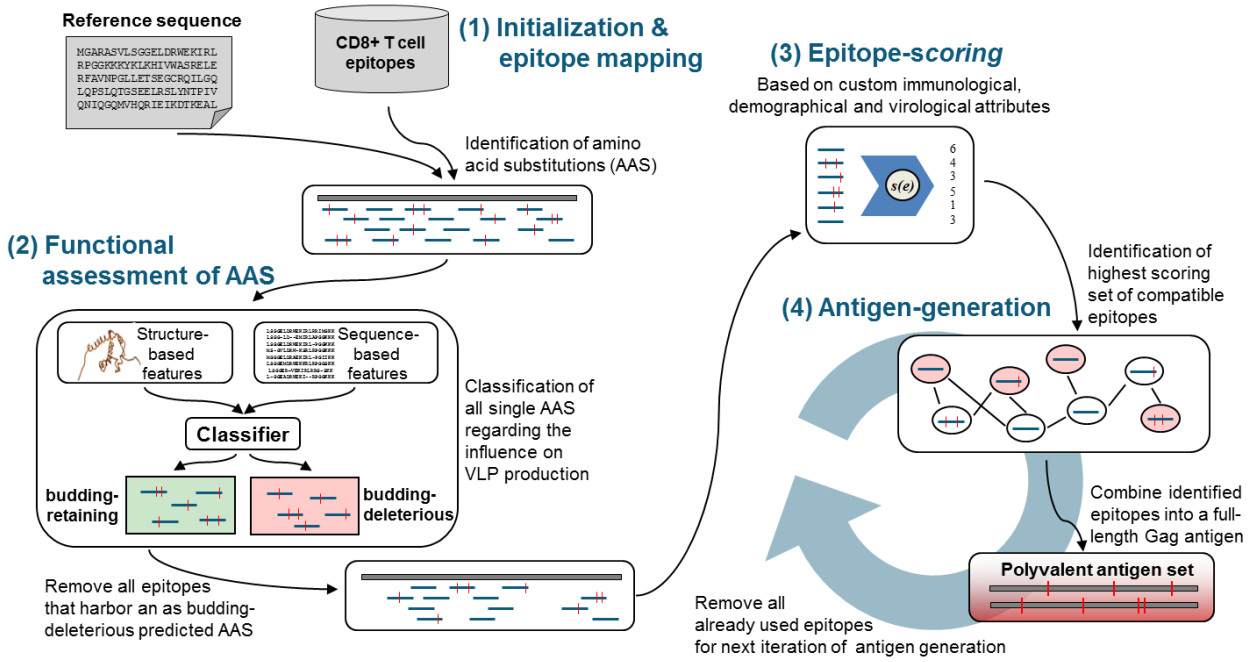


Figure 9. Workflow of the *Optimizer Algorithm* used to design CD8+ T cell epitope-enriched Gag antigens with preserved structure and function (adapted from Matthias Pfeifer[211]). (see B.2)

For the classification of the AAS, structural features as well as sequence conservation properties are taken into account. All epitopes including at least one AAS with budding-deleterious implications are removed from the input data set. (3) In the third part all remaining epitopes get ranked based on a numerical, multi-parametric scoring scheme, which considers weighted immunological, demographical and virological features. (4) The antigen candidate is generated by integrating as many compatible epitopes as possible into the reference sequence in a way that maximizes the overall combined score. For a polyvalent antigen set this last optimization step can be repeated, but all epitopes that are already incorporated, are neglected for each new iteration. This leads to a combination of sequences which complement each other in their overall score and epitope coverage. All these methods are implemented in the *Optimizer Algorithm* for fully-automatically design of T cell epitope-enriched antigens, as described in detail by Matthias Pfeifer[211]. The mode of operation is summarized below, especially pointing out steps of the algorithm that were improved in this thesis.

B.2.1 Functional assessment

Single amino acid substitutions (AAS) in Gag might lead to a partly or completely misfolded protein or influence essential interactions and thereby reduce or inhibit viral particle assembly and budding[212]. To identify epitopes with AAS that would be detrimental for the Gag structure and function, without elaborate experimental validation, a computer-aided screening system was conceived. For this AAS-specific structure- and sequence-based classification features are evaluated by a multidimensional classifier that is based on a supervised machine learning approach, to identify mutations with budding-deleterious effects.

B.2.1.1 Structure-based classification features

Since the protein conformation has crucial influence on the function, differences in the three-dimensional (3D) structures and hence their energetic profiles are indicators to evaluate the effects of amino acid substitutions. To assess such energetic changes each AAS was separately integrated into the reference sequence. Afterwards, the energetic profile of this sequence, mutated at one position, was compared to the profile of the native sequence to identify distinct local or global differences that can be applied as discriminatory features for the classifier.

Homology modeling and Discrete Optimized Protein Energy

For the structural validation of AAS, tertiary structures of the reference, as well as of the mutated sequence with the integrated AAS were required. To generate these 3D structures homology modeling was used. This computational approach utilizes the biological observation, that tertiary structural folds and motifs are far more conserved than their underlying primary amino acid sequence[213,214]. Referring to this principle, a high degree of identity between two sequences is mostly also a sign of similarities on a structural level. Homology modeling predicts the tertiary structure of a protein target sequence primarily based on its alignment to at least one similar template sequence with known structure[215,216]. One of the most widely used programs for homology modeling is *MODELLER*[215], which can automatically calculate a structural model by optimally satisfying spatial restraints derived from the target-template-alignment[217,218]. Additionally, the tool also implements some auxiliary tasks, like aligning sequences and structures as well as model evaluation by energy-based calculations. One of these intrinsic evaluation methods is the Discrete Optimized Protein Energy (DOPE)[219], an atomic-distance-dependent statistical potential, grounded entirely in the probability theory. For all homology modeling of structures and conformational energy predictions via DOPE calculations the *MODELLER* v9.7 was utilized.

Three dimensional structures of the reference sequence

The reference sequences, needed for modeling sequences with single AAS, as well as for the energy landscape comparisons, were also generated via homology modeling. The required template sequences were identified with a BLASTP (Protein Basic Local Alignment Search Tool, http://blast.ncbi.nlm.nih.gov) [220,221] search against the Protein Data Bank (PDB, www.rcsb.org)[222]. From the BLASTP hits with a sequence identity of at least 90%, the templates were selected in descending order of query coverage until the entire target was covered by at least one template structure. Target-template alignment and model building were performed with the fully-automated tools implemented in MODELLER.

Table 3. Templates for homology modelling of the reference sequence. Possible templates were identified in a BLASTP search against the PDB. Templates were chosen in descending order of query coverage of all search hits with a sequence identity greater than 90%. Selection was proceeded until the complete target sequence was covered by at least one template.

Protein	Template [PDB ID]	Identity [%]	Coverage [%]
p17	1L6N	93.89	99.24
	1TAM	99.17	90.01
p24	3H47	97.51	87.01
	3GV2	97.26	94.81
	2K1C	95.89	31.60
p2	1U57	100.00	100.00
p7	1AAF	94.55	100.00
p1	no suitable structures found		
p6	2C55	90.38	100.00

Identification of suitable templates (Table 3) and structure modeling was done before by Matthias Pfeifer[211], except for p1, because there was no sequence available fulfilling the predefined criteria.

Protein free energy changes as classification features

Analysis of AAS-induced energy changes compared to the native structure can provide evidence for conformational perturbations. Such differences can lead to malfunctions and might influence the viral budding in the context of Gag.

To analyze the energy landscape alterations accompanied by the introduction of AAS, sequences based on the reference sequence with only one single AAS at a time were generated. For each mutated sequence, five 3D-structures were generated by homology modeling using the *automodel* class of MODELLER with the previously designed reference structures as template. The corresponding structure with the lowest *DOPE Model Score* was chosen for calculating a residue-by-residue energy profile based on the DOPE energy function. This gives a position-specific energy e_k for every residue of a protein, with an amino acid sequence length *n*. As potential features to classify the AAS, the following four different structural energy parameters were computed.

$f_{1,DOPE}$: The first feature is the DOPE potential of the mutated structure (*mut*) at the residue (*k*) with the AAS itself.

$$f_{1,DOPE} = e_{mut,k}$$

$f_{2,\Delta DOPE}$: This feature represents the differences in DOPE at the mutation site compared to the reference (*ref*) at the same position.

$$f_{2,\Delta DOPE} = e_{ref,k} - e_{mut,k}$$

$f_{3,\Delta DOPE\text{-}global}$: Global DOPE differences compared to the reference are used for feature three. This value was calculated by analyzing the energy profile differences at all *n* amino acid positions separately and summing them up.

$$f_{3,\Delta DOPE-global} = \sum_{i=1}^{n} (e_{ref,i} - e_{mut,i})$$

$f_{4,\Delta DOPE\text{-}local}$: As last feature local energy differences in the structure were analyzed, by only comparing the residues in close proximity to the AAS to those of the native reference structure. The optimal window size of 12 amino acids, 6 in N-terminal and 5 in C-terminal direction from *k,* to analyze these local energy drifts was determined before[211].

$$f_{4,\Delta DOPE-local} = \sum_{i=k-6}^{k+5} (e_{ref,i} - e_{mut,i})$$

B.2.1.2 Sequence-based classification feature

In addition to the structural-energy-based features, the sequence-position-specific conservation of each AAS was calculated as a potential criterion to classify budding-deleterious mutations. The underlying premise is that functionally and structurally important sites of proteins are highly conserved and mutations here would be detrimental for the protein[223,224].

The amino acid substitutions were analyzed with a position-specific substitution matrix (PSSM)[225]. The procedure first uses a premade Gag protein alignment (B.1.1.1) to calculate the frequency (count) of each amino acid for each position separately. Since the alignment is only an incomplete representation of all naturally occurring sequences the counts were then broadened by pseudo-counts[226,227]. These pseudo-counts were determined on the basis of the natural frequencies with which different amino acids substitute for each other[228] and were calculated for each position separately according to its overall diversity[229], with conserved positions getting less pseudo-counts than variable ones. As sequence-based feature for the classifier, the PSSM-score was calculated as logarithm of the odds-ratio of the probability p_{ca} of finding an amino acid *a* on the observed position *c* in the alignment to the probability p_a of observing this amino acid by chance alone (i.e. the random distribution of the amino acid in proteins[c]).

$$f_{5,seq} = \log\left(\frac{p_{ca}}{p_a}\right)$$

B.2.1.3 A supervised machine learning AAS classifier

To discriminate between AAS with no effect on the budding of VLPs and those detrimental to it, a Fisher's Linear Discriminant Analysis (FLD)[230] was performed. This machine learning algorithm reads the entries of a training-set consisting of experimentally verified AAS[212], which were sorted into two different classes, namely budding-deleterious and budding-retaining mutations (Table 4). The classifier features of this set were used by the FLD to calculate a one-dimensional hyperplane, in a way that the two classes of AAS are best separated. Considering this hyperplane, AAS with unknown effect on budding can be classified using only the structural and sequence-based discriminatory features of this mutation. Finally all epitopes containing at least one budding-deleterious AAS were removed from the input data set.

Table 4. Training-set of classified AAS based on data from Freed et al (1994). The publication describes the introduction of 37 single AAS into p17 of the HIV-1 infectious molecular clone NL4-3. All AAS integrated into the molecular clone that were described to show no detectable virus production or delayed kinetics relative to the wild type, as determined by reverse transcriptase activity, were classified as budding-deleterious mutation for the training-set.

budding-retaining AAS		budding-deleterious AAS	
A5D	S72I	G2A	Y86G
G10E	N80G	S6I	C87D
R20L	K98G	V7R	C87S
L21E	E106V	A37E	V88E
P23E	K113T	L50D	H89G
K26T	K114T	G56E	D96L
K27T	A120E	C57D	A100E
K32A	D121L	C57S	
R43A	V128E	I60E	
E52G	Q130G	L85R	

[c] obtained from http://www.pseudogene.org/composition/index.cgi

To review the quality of the classifier without laborious experiments and also to find the optimal combination of the five classification features, a repeated *k*-fold cross-validation was performed[231]. For this, the training-set was randomly subdivided into *k* equally sized disjoint subsets (or folds). All but one of the k subsets were then applied to train the classifier. The remaining subset containing AAS with known effects was used to test the accuracy and precision of the created respective classifier (Table 5). This was repeated *k* times, until the predictions had been calculated for every fold as test-set. The final quality is presented as mean of all k-tests. Since this procedure is highly dependent on the split of the training-set, the *k*-fold cross-validation was repeated multiple times with different divisions of the data set. For calculating the k-fold cross-validation output, budding-retaining AAS were defined as positive regarding condition and test-outcome (Table 5).

Table 5. Contingency table of binary classification outcomes. To assess the quality of the FLD predicted classification the experimentally validated training-set AAS ("Gold standard") were tested in a *k*-fold cross-validation. Maintenance of budding was thereby defined as positive event. True positive (tp), false positive (fp), false negative (fn), and true negative (tn) classifications of the "Gold standard" were applied to calculate accuracy, sensitivity, specificity, positive predictive value (PPV, precision), and negative predictive value (NPV) as described in the table.

		Condition (experimentally validated training-set = "Gold standard")		
		Budding-retaining	Budding-deleterious	
Test outcome (FLD classification)	Budding-retaining	**True positives** (tp)	**False positives** (fp)	PPV (precision) = tp/(tp+fp)
	Budding-deleterious	**False negatives** (fn)	**True negatives** (tn)	NPV = tn/(fn+tn)
Accuracy = (tp+tn)/(tp+fp+fn+tn)		Sensitivity = tp/(tp+fn)	Specificity = tn/(fp+tn)	

B.2.2 Epitope scoring

After the functional analysis the epitopes get ranked based on a score calculated by considering immunological, demographical and virological characteristics. The higher the score the more favorable the epitope would be for integration into the newly designed antigens. For the calculations, multiple entries for the same epitope, as well as the assessment of sub- and superepitopes, were handled as described above (B.1.2).

In this thesis, the frequencies of HLA class I molecules presenting the epitope, the subtype affiliation, the association with LTNPs, the conservation status, and the expected population-wide immune response against the epitope were applied as attributes to calculate the specific score *s(e)* for each epitope *e*. However, the algorithm is implemented to work with any number of Boolean or positive numerical values for score considerations. With positive weighting parameters $w_a \in \mathbb{N}_0$ for each attribute *a* the epitope ranking can be further modified according to user preferences.

$$s(e) = \sum_{a \in A} \left(w_a \cdot a(e) \right)$$

In the next chapters the different attributes contributing to the epitope-specific score are explained in detail.

B.2.2.1 HLA score

The HLA class I alleles represent some of the most variable human genes and the haplotype of a person greatly influences which epitopes can be presented to CD8+ T cells. The frequency variations of HLA class I alleles across different human populations is quite severe. Since the aim for a vaccine would be to protect as many people of the target population as possible, epitopes presented by alleles with higher frequencies in the respective population should be included more favorably.

To rank the epitopes an HLA class I score $a_{HLA}(e)$ was calculated by summing up the frequencies of all alleles *h* from the complete set of HLA class I alleles *H* that are specified in the database entry as being able to present the epitope *e*. If more than one population is of interest, the mean HLA class I frequency $\bar{f}$ was used to compute the score

$$a_{HLA}(e) = \sum_{h \in H} \left(r(e,h) \cdot \bar{f}(h) \right); \quad with\, r(e,h) = \begin{cases} 1 & if\ e\ is\ presented\ by\ h \\ 0 & otherwise \end{cases}$$

The HLA score was normalized to the interval [0, 1], with 1 being the highest and 0 the lowest calculated score.

B.2.2.2 Subtype score

Due to the high variability of HIV-1, many different subtypes emerged. However these subtypes are not equally distributed and are highly region-specific. The higher the epitope-affiliated subtype-frequency in the target region is, the more favorably it should be included in the newly designed antigens.

To account for the targeted subtype-diversity a subtype score with an adjustable weighting parameter $w_s \in \mathbb{N}_0$ for each subtype *s* of all specified subtypes *S* was conceived. Each epitope *e* subtype score $a_{Subtype}(e)$ was computed by summing up the weights of all subtypes *s* that are specified for the epitope in its LANL HIV database entry.

$$a_{Subtype}(e) = \sum_{s \in S} (r(e,s) \cdot w_s); \quad with\, r(e,h) = \begin{cases} 1 & if\ e\ is\ affiliated\ with\ s \\ 0 & otherwise \end{cases}$$

As for the HLA score, the subtype score was also normalized to the interval [0, 1], with 1 being the highest and 0 the lowest computed score.

B.2.2.3 LTNP Score

If an epitope is associated with LTNPs it might have a beneficial impact on protection, if it is integrated in an antigen. Therefore the notes section of each database entry was checked, if affiliation with LTNPs had been reported. Since this is a Boolean question the LTNP score a_{LTNP} can only be 1, if it is true (i.e. the epitope *e* is associated with LTNPs) or otherwise 0.

$$a_{LTNP}(e) = \begin{cases} 1 & if\ e\ is\ associated\ with\ LTNP \\ 0 & otherwise \end{cases}$$

B.2.2.4 Conservation score

Another feature for evaluating epitopes was based on their conservation among different HIV-1 subtypes. The higher the conservation, the more favorably a sequence should be rated for inclusion in the newly designed antigens, since a broader spectrum of different viral clades can be addressed with one epitope. To assess and compare the epitope set (B.1.2), a conservation score was calculated for each entry. For this, a reference set of proteins sequences with specified subtype affiliation, the epitope entries, and a list of all subtypes *S* with a customizable weighting were needed. At first, the frequency of every epitope *e* in all reference sequences with subtype *s* was determined. This frequency *f(e,s)* was calculated by dividing the number of subtype s sequences, which include the epitope *e* [*sequences(e,s)*], by the count of all subtype *s* sequences in the reference set [sequences(s)].

$$f(e,s) = \frac{sequences(e,s)}{sequences(s)}$$

Each subtype was assigned a customizable positive weighting parameter $w_s \in \mathbb{N}_0$, by which the subtypes were ranked so that their effect on score calculations could be varied. The epitope-*e*-specific conservation score a_{Cons} was then computed as the sum of all w_s weighted subtype $s \in S$ epitope frequencies *f(e,s)*.

$$a_{Cons}(e) = \sum_{s \in S} \left(w_s \cdot f(e,s)\right)$$

The score was finally normalized to the interval [0, 1], with 0 meaning epitope *e* was found in none of the reference sequences (i.e. conservation of 0%) and 1 that it was found in every sequence (i.e. conservation of 100%).

B.2.2.5 Expected immune response score

Many of the CD8+ T cell epitopes in the LANL immunology database were identified in studies with large cohorts. For these entries, the percentage of people responding to the specified epitope was retrieved from the metadata. Depending on the type of the underlying study the data was applied differently to calculate an expected immune response score $a_{\%Resp}$.

(1) If the study was based on a large population *p* with random HLA class I haplotype distribution the score was the fraction of people responding to the epitope *p(e)*.

$$a_{\%Resp} = \frac{p(e)}{p}$$

(2) If the target population was a selection of people with one specific HLA class I allele *h*, shown to present the epitope *e*, the response score was computed by taking the allele frequency $\bar{f}(h)$ (B.2.2.1) into account. The population-fraction possessing the allele *h* was calculated by considering the diploid human chromosomes ($2 \cdot \bar{f}(h)$) and also homozygote persons (subtracting $\bar{f}(h)^2$). This term was then multiplied by the fraction of the study population *p* that had responded to the epitope *p(e)*.

$$a_{\%Resp} = \frac{p(e)}{p} \cdot \left(2 \cdot \bar{f}(h) - \bar{f}(h)^2\right)$$

B.2.3 Antigen generation

After removing all epitopes with functionally detrimental properties and scoring of the remaining ones, the highest-scoring epitopes were to be combined and integrated into the reference sequence to generate the optimal antigen candidate. But, since the epitopes are derived from a broad spectrum of variable HIV-1 isolates, many have non-matching overlaps to other epitopes and their incorporation into the reference sequence is mutually exclusive (see examples in Table 6).

Table 6. Compatible and incompatible epitopes. Many epitopes show an overlap among each other, when aligned to the reference sequence. (A) If the shared sequence of the overlapping epitopes is identical (marked in green), both epitopes can be incorporated into the antigen. (B) However if there are differences in the overlap (marked in red), the integration of the epitopes is mutually exclusive

A. Compatibility	B. Incompatibility
MGARA**SVLS**	MGARA**SVLS**
SVLSGGELD	**SILS**GGELD

To represent these incompatibilities, a graph theoretic approach was conceived. Therein, an undirected incompatibility graph *G* illustrates all epitopes as vertexes with edges between two epitopes representing their incompatibility. The connected elements were identified with a straightforward depth-first search[232]. Based on this graph, an independent set of epitopes, i.e. a combination of vertexes connected by no edge, with the maximal overall epitope score (or weight) should be selected. Since the computing of this so-called maximum weight independent set (MWIS) of an arbitrary graph is an NP-hard problem[233], meaning that the computational run time increases exponentially with the input size, already the calculation of the optimal solution for a small input set is infeasible. To calculate a set near the MWIS, heuristics have to be applied. Many modern heuristic algorithms are inspired by natural phenomena[234], like the genetic algorithm used here, which uses evolutionary principles, like mutations and survival of the fittest, to select an optimum near the MWIS[235]. Compared to others, like sequential greedy heuristics[233], a genetic algorithm has the advantage that it does not aim to find just a local, but a global maximum to a given problem.

The implemented genetic algorithm for constructing the best combination of high scoring epitopes consists of various steps. Initially, a parent population of 70 randomly selected independent sets of epitopes, called chromosomes, is generated. The chromosomes are encoded as bit strings of length E for all epitopes, with each bit representing one epitope of the input graph. If the epitope e_i is included in the set, the bit x_i gets assigned the value 1, otherwise 0. All chromosomes get ranked by their fitness, which is calculated by a function *f(x)* which sums up all scores *s* of epitopes included in the set and that also accounts for the number of incorporated sub-(n_{sub}) and superepitopes (n_{sup}) in the solution. The differential weighting of those three factors was empirically determined by testing several parameters and choosing the best-performing combination.

$$f(x) = \left(\sum_{i=1}^{E}(x_i \cdot s(e_i))\right)^2 + 20 \cdot n_{sup}(x) + n_{sub}(x);\ with\ x_i = \begin{cases} 1 & if\ e_i\ is\ included\ in\ the\ set \\ 0 & otherwise \end{cases}$$

The fittest chromosome of the newly generated parent population is initially selected as the solution representative, i.e. the best combination of compatible epitopes found by the algorithm. The evolutionary procedure starts by choosing two of the parent chromosomes. The first one is selected with a probability proportional to its fitness. The second one is selected the same way or, with a probability of 0.4 a randomly generated new chromosome is utilized in order to

maintain genetic variability. By a three-point crossover of the parents a “child” chromosome is created that inherits bit strings, i.e. epitopes from both parents. The sites for the crossover are chosen randomly but are restricted by the fitness of the parent chromosomes, i.e. the fitter the parent the more bits are passed on to the child. During each evolutionary step, subpopulation pool containing 24 sequences is filled with child chromosomes, or with a chance of 0.05, with a randomly initialized chromosome, again to maintain genetic variability. If one of the subpopulation chromosomes is identical to one in the parent population, it is rejected. Otherwise its fitness is compared to four randomly selected parent chromosomes. If the fitness is higher than any of them, it replaces the lowest scoring one in the parent pool and if it has also a higher fitness than the current solution, it also substitutes this one. In each evolutionary round all chromosomes undergo spontaneous mutation. Through this, the inclusion of each epitope *e* in the respective set is changed with the probability 0.01, i.e. the bit value is switched between 0 and 1. Since both three-point crossover and spontaneous mutations could introduce new epitope incompatibilities, a repair method is implemented after both procedures. For this, the epitopes with the most incompatibilities are subsequently removed until the set is independent again. After 2000 evolutionary generations, the MWIS search is terminated and the solution representative is used as antigen candidate. After excluding all epitopes already represented in the solution set, the next round of the genetic algorithm can be performed, resulting in a polyvalent set of complementary CD8+ T cell epitope-enriched antigens.

B.3 Antigen evaluation

B.3.1 Phylogenetic tree calculations

For phylogenetic classification, the Gag protein sequence reference set[236] from the year 2010, comprising of all group M subtypes and CRFs, was downloaded from the LANL HIV sequence database[d]. The set was manually modified by removing all CRFs except CRF01_AE and CRF02_AG and adding antigen sequences of interest. Alignment of all sequences and phylogenetic tree reconstruction was performed with standard implementations of *ClustalX 2.1*[237]. The tree was visualized using the *Rainbow Tree*[238,239] tool from the LANL HIV database[e].

B.3.2 Antigen score

To address the quality of an antigen *a* (or a polyvalent combination of antigens), the score of all epitopes from the complete set of CD8+ T cell epitopes *E* (B.1.2) that are incorporated in the respective antigen *a* was summed up to an antigen-specific score (Figure 10). The epitope-*e*-specific score *s(e)* is calculated based on the same formulas as described above (B.2.2). Since superepitopes include all attributes (B.1.2) from their associated subepitopes, the antigen score calculations were only based on epitopes that were not also represented by a superepitope in antigen *a*. Applying this restriction the epitope-*e*-associated score *s(e,a)* for a specific antigen *a* can be either *s(e)* or 0, depending on the incorporation into antigen *a* (*r(e,a*)) and the presence of superepitopes (*z(e,a)*).

[d] http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html#ref
[e] http://www.hiv.lanl.gov/content/sequence/RAINBOWTREE/rainbowtree.html

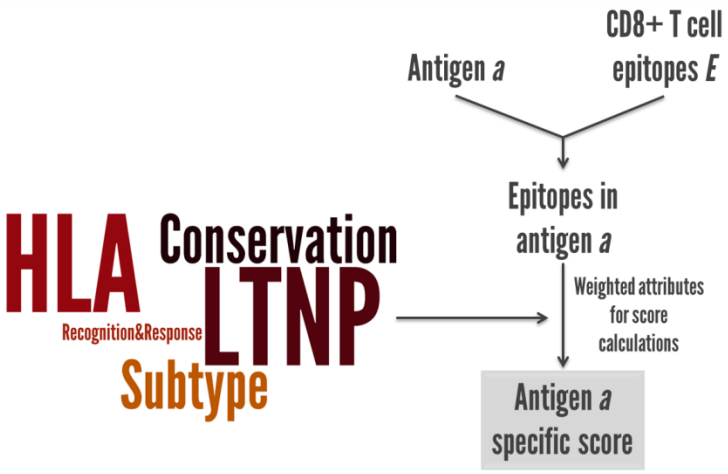


Figure 10. Antigen score. To evaluate and compare antigens an antigen score was calculated. For this first the antigen *a* was compared against the complete CD8+ T cell epitope set. The score of all epitopes found in antigen *a* was calculated as described above (B.2.2) based on the frequencies of HLA class I molecules presenting the epitope, subtype affiliation, the association with LTNPs, the conservation status, and the expected population-wide immune response against the epitope. The text size in the word cloud symbolizes the weighting of the different attributes. The antigen score was calculated by summing up the scores of all epitopes that are included in the antigen without those already being represented by a superepitope in the antigen *a*.

$$s(e,a) = r(e,a) \cdot z(e,a) \cdot s(e)$$

$$with\ r(e,a) = \begin{cases} 1 & if\ e\ is\ included\ in\ the\ antigen\ a \\ 0 & otherwise \end{cases}$$

$$and\ z(e,a) = \begin{cases} 0 & if\ e\ has\ a\ superepitope\ in\ the\ antigen\ a \\ 1 & otherwise \end{cases}$$

The score of antigen *a* (or a polyvalent combination of antigens) was finally calculated as sum of the corresponding scores *s(e,a)* of all epitopes $e \in E$. Additionally, the number of incorporated epitopes was captured for each antigen *a*.

$$AntigenScore(a) = \sum_{i=1}^{E} \left(s(e_i, a)\right)$$

B.3.3 Population coverage

HLA alleles are distributed quite heterogeneously across the human population. An efficient vaccine candidate should be able to present a broad range of immunologically valuable epitopes on various HLA alleles of the particular target population

To assess the potential breadth of an immune response against antigen *a* (i.e. the number of different epitopes from *a* targeted by CTLs) an *in silico* analysis tool (called "*Population coverage*", Figure 11) was implemented. In a first step a target population of size *P*, specified by different HLA haplotypes *Y* was generated. The haplotype allele selection was done separately for each class I HLA gene locus, HLA-A, -B, or -C. Due to the diploid nature of the human genome, two alleles were randomly chosen for each locus with a probability proportional to the allele frequency $\bar{f}$ (B.2.2.1) within the target population. Thus a test-person *p* of the target population *P* is defined by its HLA haplotype Y_p, consisting of six alleles.

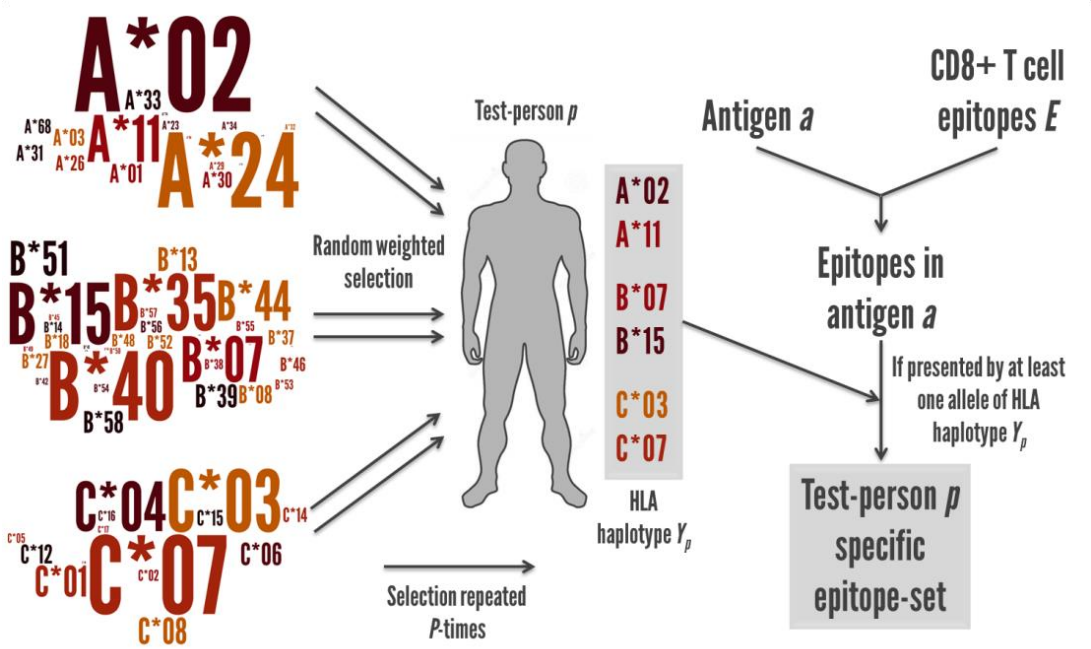


Figure 11. Population coverage. To assess how well an antigen fits to its target-population, a test-population of size *P*, defined by different HLA haplotypes *Y* was generated. For each test-person *p*∈*P* two HLA-A, -B, and -C alleles were randomly selected proportionally to the allele frequency for each gene locus (as symbolized by the font size in the word clouds). For population coverage calculation of an antigen *a* all epitopes of a given set *E* found in *a* and also presented by at least one allele of the test-person *p*'s haplotype *Yp* got included in the test-person-specific epitope-set. The score of this set was calculated as described before (B.3.2). The population coverage of *a* was determined for every *p*∈*P.*

In the next step, all in antigen *a* (or sets of polyvalent antigens) incorporated epitopes from a given CD8+ T cell epitope list *E* were identified. These antigen-*a*-specific epitopes were then tested against each test person *p*. If an epitope *e* can be presented on at least one HLA allele (i.e. if the HLA is denoted in the epitope's database entry) of the test-person-*p*-specific haplotype Y_p it was included into this test person's epitope set, otherwise it was rejected. Each of these P test-person-specific epitope sets were finally analyzed for the number of epitopes in the set and their cumulative epitope scores *s(e,a)* (B.3.2):

$$PopulationScore(p,a) = \sum_{i=1}^{E} \left(r(e_i, Y_p) \cdot s(e_i, a) \right)$$

$$with\ r(e_i, Y_p) = \begin{cases} 1 & if\ e_i\ is\ presented\ by\ at\ least\ one\ allele\ of\ HLA\ haplotype\ Y_p \\ 0 & otherwise \end{cases}$$

B.3.4 Pathogen coverage

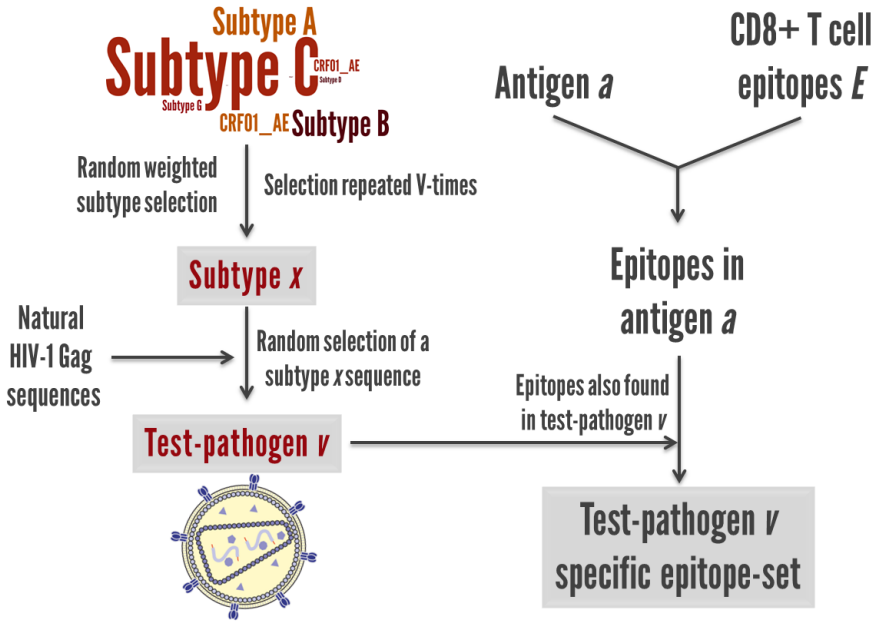


Figure 12. Pathogen coverage. To validate how well an antigen *a* covers a target-region's subtype distribution, first a test-pathogen set of size *V* was generated. The subtype *x* of each test-virus *v* was selected by a random selection proportional to the denoted clade distribution (symbolized by the font size in the subtype word cloud). The test-pathogen v was then initialized by selecting a natural HIV-1 Gag sequence of subtype *x*. For pathogen coverage calculation of an antigen *a*, all epitopes of a given set *E* found in *a* and also in the test-virus-*v*-associated Gag sequence, got included in the test-pathogen-specific epitope-set. The score of this set was calculated as described above (B.3.2). The pathogen coverage of *a* was determined for all *v*∈*V*.

In addition to a broad targeting of the target population, an antigen should also be able to cover as many different HIV-1 isolates as possible. To assess the degree with which an antigen covers the pathogen distribution in a specific region of the world another computational tool (called "*Pathogen Coverage*", Figure 12) was developed. This tool first generates a test-pathogen set of size *V*. The subtype *x* of each pathogen *v*∈*V* was based on a random selection with a probability proportional to the given clade frequency in the target region. A Gag sequence of this subtype *x* was then randomly picked from the set of natural Gag sequences (i.e. the curated filtered web alignment of Gag - B.1.1.1) and assigned to the pathogen *v*. Then, an antigen-*a*-specific epitope set was determined again (B.3.3), based on a given antigen *a* and a CD8+ T cell epitope list *E*. All epitopes included in antigen *a* that could also be found in the Gag sequence of the respective test pathogen *v*, were added to the test-pathogen's epitope set. All *V* test pathogen sets were reviewed for number of epitopes and the cumulative epitope scores *s(e,a)* (B.3.2).

$$PathogenScore(v,a) = \sum_{i=1}^{E} \left(r(e_i, v) \cdot s(e_i, a) \right)$$

$$with\ r(e_i, v) = \begin{cases} 1 & if\ e_i\ is\ included\ in\ pathogen\ v \\ 0 & otherwise \end{cases}$$

B.4 Codon adaptation

The genetic code is degenerated and up to six codons code for the same amino acid. Depending on the organism, these codons are used with different frequencies[240]. For an efficient heterologous gene expression system, an adaption to the host's codon usage is highly beneficial[241,242].

B.4.1 Human codon adaptation

For expression in human cell lines, the HXB2-Gag and the newly designed T cell epitope-enriched Gag antigens were optimized with the online tool "*Geneart Geneoptimizer*"[243,244] (Thermo Fisher) based on the *homo sapiens* codon usage.

B.4.2 HIV-1-Gag-specific codon adaptation

Besides the individual expression in human cell lines, the newly designed teeGags were also incorporated into an HIV-1 molecular clone. For this, the Gag sequences had to be adapted to the unusually high adenine and low cytosine content in HIV-1[245]. However there are subtle differences in base composition across the HIV-1 genome, with the adenine content for example ranging from 25.0% (LTR R3-U-R5) to 38.6% (Vpu-ORF) and with the Gag-ORF (36.9%) located fairly close to the maximum percentage[246]. This instance and the fact that the viral replication is regulated by conserved, mostly uncharacterized genomic RNA structures[247] prohibited a random codon-optimization based on HIV codon frequencies, which would probably not result in a functional virus. To circumvent this problem the natural Gag nucleotide sequence of the reference HXB2 (GenBank K03455) was employed as scaffold to integrate the mutations accumulated during the optimization process.

Based on a Gag nucleotide alignment (B.1.1.1), including the reference sequence, a position-specific codon usage matrix was generated. This was an n×64 matrix, where n is the number of codons of the query sequence (for HXB2-Gag n=500). For each position of the matrix, the counts were determined by summing up the respective codons in the alignment. For the teeGags, the HXB2 nucleotide sequences was used as backbone and for each AAS compared to the HXB2 reference, the most common codon for this mutation at the respective position was incorporated into the scaffold.

C Material and experimental methods

C.1 Material

C.1.1 DNA

C.1.1.1 Oligonucleotides

Oligonucleotides were ordered at Eurofins (Luxembourg) or at Biomers (Ulm, Germany)

Name	Sequence (5'→3')	Description
pc31 fwd	CTATATAAGCAGAGCTCTCTGGC	Standard cloning and sequencing primer for pcDNA3.1. Binds at the 3' end of the CMV promoter.
pc31 rev	GCAACTAGAAGGCACAGTCG	Standard cloning and sequencing primer for pcDNA3.1. Binds at the 5' end of the BGH polyA site.
teeGag-His rev	GACTCTCGAGTCATCAGTGGTGATGGTG GTGGTGGCTGCCTCTCTGGCTGCTGGG GT	Primer for C-terminal addition of a 6x-His epitope tag (RGSHHHHHH) to teeGag1-3, HXB2-Gag, and HXB2-Gag^{Myr-}. Contains an XhoI restriction site for cloning.
teeGag1-GL9-GS fwd	GGCCCAAGCCACAAAGCCAGAGTGCTG	Forward binding mutation primer for fusion PCR. Introduces a G to S mutation in teeGag1 to change its GL9 epitope sequence to GP**S**HKARVL.
teeGag1-GL9-GS rev	TTTGTGGCTTGGGCCTCCCACG	Reverse binding mutation primer for fusion PCR. Introduces a G to S mutation in teeGag1 to change its GL9 epitope sequence to GP**S**HKARVL.
teeGag3-GL9-IV fwd	GCCAGAGTGCTGGCCGAGGCCA	Forward binding mutation primer for fusion PCR. Introduces an I to V mutation in teeGag3 to change its GL9 epitope sequence to GPSHKAR**V**L.
teeGag3-GL9-IV rev	GGCCAGCACTCTGGCCTTGTGAGAAGG	Reverse binding mutation primer for fusion PCR. Introduces an I to V mutation in teeGag3 to change its GL9 epitope sequence to GPSHKAR**V**L.
HXB2-Gag-GL9-GS fwd	CGGACCTAGCCACAAAGCCAGAGTGCTG	Forward binding mutation primer for fusion PCR. Introduces a G to S mutation in the HXB2-Gag reference sequence to change its GL9 epitope sequence to GP**S**HKARVL.
HXB2-Gag-GL9-GS rev	GCTTTGTGGCTAGGTCCGCCCACG	Reverse binding mutation primer for fusion PCR. Introduces a G to S mutation in the HXB2-Gag reference sequence to change its GL9 epitope sequence to GP**S**HKARVL.
teeGag-AL-fwd	CAGCGCGCACGGCAA	Forward binding primer used for teeGag proviral cloning. Binds at the BssHII restriction site in the 5' UTR.
teeGag-AL-rev	TAAAAAATTGGCTTGTGACGAGGGGTCG	Reverse binding primer used for teeGag proviral cloning. Binds at the 3' end of all p6 and harbors an overlap to p6* of the pNL4-3_AL vector.
N4-3-AL-fwd	CCTCGTCACAAGCCAATTTTTTAGGGTAG ATCTGGC	Forward binding primer used for teeGag proviral cloning. Binds at the 5' end of p6* and harbors an overlap to p6 of teeGag1-3 and HXB2-Gag.
NL4-3-AL-rev	ACCCTGCAGGATGTGGTATTCC	Reverse binding primer used for teeGag proviral cloning. Binds at the SbfI restriction site within the pol reading frame of pNL4-3_AL.
HXB2-AL fwd	GCCTAGTTATAAAGGGCGCCCGGGTAAC TTCCTCCAGAGCAGACCAGAGCCAAC	Forward binding primer used for teeGag proviral cloning. Binds in front of the HXB2-Gag slippery site and Introduces mutations to destroy the Gag-Pol frameshift.
HXB2-AL rev	CGCCCTTTATAACTAGGCCAAATTTTACC CAGGAAATTAGCCTGTCTCTCAGTACAAT CTTTC	Reverse binding primer used for teeGag proviral cloning. Binds after the HXB2-Gag slippery site and Introduces mutations to destroy the Gag-Pol frameshift.

G2A-AL fwd	AGAGATGGCAGCGAGAGCGTCAGTATTAAGCG	Forward binding mutation primer for fusion PCR. Introduces the G2S mutation in HXB2-Gag to change generate the budding deficient HXB-Gag^{Myr-}.
G2A-AL rev	CGCTCTCGCTGCCATCTCTCTCCTTCTAGCCTCCG	Reverse binding mutation primer for fusion PCR. Introduces the G2S mutation in HXB2-Gag to change generate the budding deficient HXB-Gag^{Myr-}.
sHLA A*02:01 fwd	CTAAAGCTTATGGCCGTCATGGCG	Forward binding cloning primer for sHLA A*02:01 construction. Binds directly at the start codon and introduces a 5' HindIII restriction site.
sHLA B*07:02 fwd	CTAAAGCTTATGCTGGTCATGGCGC	Forward binding cloning primer for sHLA B*07:02 construction. Binds directly at the start codon and introduces a 5' HindIII restriction site.
sHLA-exon4+His rev	CTAGCTCGAGTCAGTGGTGATGGTGGTGGTGGCTGCCTCTCCATCTCAGGGTGAGGGG	Reverse binding cloning primer for sHLA construction. Binds at 3' end of HLA exon 4 and introduces a C-terminal 6x-His epitope tag (RGSHHHHHH) and an XhoI restriction site for subcloning.

Primers used for AAS single site mutations are listed in Extended Data Table 6 and Extended Data Table 7.

C.1.1.2 Vectors

Description	Selectable marker	Specification (Supplier - Product no.)
p5'	Amp^R	The 5' half of HIV-1 from pHXB2gpt[248] was cloned into the standard cloning vector pGEM-3Z as a 5978 bp fragment (NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: p5' from Drs. Dean Winslow and Lee Bacheler[249])
pBluescript KS (-) CMV-SEAP	Amp^R	Bacterial expression vector. pBluescript backbone with integration of a CMV promotor for mammalian expression of secreted alkaline phosphatase (SEAP) (Addgene, Cambridge, USA - plasmid #24595)
pcDNA3.1(+)	Amp^R, Neo^R	Mammalian expression vector (Thermo Fisher, Waltham, USA - V79020)
pcDNA5/FRT	Amp^R, Hygro^R (no ATG)	Mammalian expression vector. Designed for generation of stable cell lines with the Flp-In^{TM} system. Hygromycin resistance gene is only expressed after integration at the flippase recognition target (FRT) site (Thermo Fisher - V601020).
pcDNA5/FRT/TO	Amp^R, Hygro^R (no ATG)	Inducible Mammalian expression vector. Designed for generation of stable cell lines with the Flp-In^{TM} TREx^{TM} system. Hygromycin resistance gene is only expressed after integration at the FRT site (Thermo Fisher - V652020).
pNL4-3_AL	Amp^R	Molecular HIV-1 clone. For production of infectious isolate NL4-3 HIV-1 particle. Derived from NL4-3_(Nef-R71)[250] with shifted slippery site to uncouple the overlapping part of *gag* and *pol* reading frames[251] (in-house).
pOG44	Amp^R	Mammalian vector for transient expression of the Flp recombinase[252] (Thermo Fisher - V600520)
VRC-8400	Kan^R	Mammalian expression vector (G. Nabel, Vaccine Research Center, NIAID, Bethesda, USA).

C.1.2 Antibodies

Description	Conjugate	Clone	Supplier (Cat #)	Specifications (Application[f]/Dilution)
α-6x-His Epitope Tag	Biotin	HIS.H8	Thermo Fisher (MA1-21315-BTIN)	Monoclonal mouse IgG2b anti-6x-His synthetic peptide antibody (WB/1:2,000; ELISA/1:1,000; SLOT: 1:1,000)
α-β2m	HRP	-	Thermo Fisher (PA1-29662)	Polyclonal rabbit IgG anti-beta-2-microglobulin antibodies (WB/1:500)
α-CD1a	PE	HI149	BD, East Rutherford, USA (555807)	Monoclonal mouse IgG1, κ anti-human-CD1a antibody (FC/1:66)
α-CD8α	FITC	RPA-T8	BioLegend, San Diego, USA (301006)	Monoclonal mouse IgG1, κ anti-human-CD8a antibody (FC/1:100)
α-CD14	FITC	MφP9	BD (345784)	Monoclonal mouse IgG2a, κ anti-human-CD14 antibody (FC/1:33)
α-HLA-ABC	-	EMR8-5	Abcam, Cambridge, UK (ab70328)	Monoclonal mouse IgG1, κ anti-human-pan-HLA ABC antibody (WB/1:3,000)
α-HLA-ABC	-	W6/32	BioLegend (311402)	Monoclonal mouse IgG2a, κ anti-human-pan--HLA ABC antibody (ELISA/1:500)
α-IFN-γ	APC	4S.B3	BioLegend (502512)	Monoclonal mouse IgG1, κ anti-human-IFN-γ antibody (FC/1:100)
α-mouse Ig	HRP	-	Dako, Agilent, Santa Clara, USA (P0260)	Polyclonal rabbit anti-mouse Ig purified Ig fraction (WB 1:2,000)
α-p24/p55	-	37G12	Polymun, Klosterneuburg, Austria (AB005)	Monoclonal human anti-HIV-1-p24 antibody. Also reactive against HIV-1 p55. Epitope unknown (ELISA/Lot dependent)
α-p24/p55	-	CB-13/5	Hybridoma cell line	Monoclonal mouse IgG1, κ anti-HIV-1-p24 antibody. Also reactive against HIV-1 p55. Epitope described as (VHQAISPRTL)NAWVK$^{253–255}$. (WB/1:1,000)
α-p24/p55	-	M01-16/4/1	Polymun (AB006)	Monoclonal mouse IgG2a, κ anti-HIV-1-p24 antibody. Also reactive against HIV-1 p55. Epitope described by supplier as GATPQDLNTML. (ELISA/Lot dependent)
α-p24/p55	RD1	KC57	Beckman Coulter, Brea, USA (41116015)	Monoclonal mouse IgG1 anti-HIV-1-p24 antibody. Also reactive against HIV-1 p55. Epitope unknown. (FC/1:200)
α-rabbit Ig	HRP	-	Dako (P0448)	Polyclonal goat anti-rabbit Ig antibodies (WB 1:1,000)
α-VSV-G	-	P5D4	Sigma Aldrich, St. Louis, USA (V5507)	Monoclonal mouse IgG1 anti- Vesicular Stomatitis Virus Glycoprotein (VSV-G) antibody. (WB/1:10,000)
Isotype control	PE	MOPC-21	BioLegend (400114)	Monoclonal mouse IgG1, κ isotype control (FC/assay dependent)

[f] WB = Western Blotting; FC = Flow Cytometry; ELISA = Enzyme-linked Immunosorbent Assay; SLOT = Slot Blotting

C.1.3 Peptides

Description	Specifications	Supplier
GL9	Sequence: GPGHKARVL Origin: HIV-1 Gag p24 Residues: 223-231 Purity: >91.20%	Centic Biotec, Heidelberg, Germany
GL9-GS	Sequence: GPSHKARVL Origin: HIV-1 Gag p24 Residues: 223-231 Purity: >96.24%	Centic Biotec

C.2 Experimental methods

C.2.1 Microbiological techniques

C.2.1.1 Cultivation and selection of bacterial cultures

Strain	Genotype
DH5α	F- supE44 ΔlacU169 (φ80 lacZΔM15) hsdR1 recA1 endA1 gyrA96 thi-1 relA1 [256]
DH10B	F- mcrA Δ(mrr-hsdRMS-mcrBC) Φ80dlacZΔM15 ΔlacX74 endA1 recA1 deoR Δ(ara,leu)7697 araD139 galU galK nupG rpsL λ- [257]

Cultivation of *E. coli* was performed overnight at 37°C in lysogeny broth medium (LB-medium) on an orbital shaker at 220 rpm. For selection of transformed cells, antibiotics (100 µg/ml ampicillin or 50 µg/ml kanamycin) were added to the medium.

LB-medium	0.5% (w/v) yeast extract, 1% (w/v) tryptone, 1% (w/v) NaCl

C.2.1.2 Transformation of chemically competent bacteria

For *E. coli* transformation, the bacterial cells were made chemically competent with the $RbCl_2$ method[256,258] and stored at -80°C until further use. To introduce heterologous DNA, 100 µl of these competent cells were thawed on ice and then incubated with either 1 µg plasmid DNA or a ligation mix (C.2.2.1) for 30 min on ice. To induce the DNA uptake, the cells were treated with a 45 s heat shock in a 42°C water bath, followed by 2 min incubation on ice. For cell rescue, 0.9 ml LB-medium were added and the tubes were subsequently incubated for 1 h at 37°C and 300 rpm in an orbital shaker. For selection of positive clones, the *E. coli* bacteria were plated on LB agar plates with an appropriate antibiotic.

LB-agar	LB-medium, 1.5% (w/v) agar, antibiotic (100 µg/ml ampicillin or 50 µg/ml kanamycin)

C.2.2 Molecular biology techniques

C.2.2.1 Standard cloning procedure

For cloning of genes, standard cloning techniques[259] were applied. Vector backbones were digested using restriction endonucleases, dephosphorylated with CIP and purified from a 1% agarose gel employing the QIAquick Gel Extraction Kit. Inserts for cloning were generated using (1) digestion of plasmids with suitable restriction endonucleases or (2) PCR amplification, applying the proof-reading Phusion High-Fidelity DNA Polymerase with subsequent digestion with restriction endonucleases. The inserts were also extracted from a 1% agarose gel with the QIAquick Gel Extraction Kit. Restriction sites and protein tags were introduced with primer extension PCR.

For ligation of insert and vector with matching cohesive ends the Quick Ligation™ Kit was utilized, according to the manufacture's protocol. The complete ligation mix was applied to transform chemically competent *E. coli* (C.2.1.2). LB medium with the respective antibiotic was inoculated with single colonies. After cultivation, the plasmid DNA was purified using alkaline lysis with subsequent isopropanol precipitation[260]. To verify the correct cloning, the DNA was sequenced (SeqLab, Göttingen, Germany).

Alkaline Phosphatase, Calf Intestinal (CIP)	New England Biolabs, Ipswich, USA (M0290)
Phusion® High-Fidelity DNA Polymerase	New England Biolabs (M0530)
QIAquick Gel Extraction Kit	Qiagen, Hilden, Germany (28706)
Quick Ligation™ Kit	New England Biolabs (M2200)
Restriction endonucleases	New England Biolabs (various)

C.2.2.2 Fusion PCR for site-specific mutations

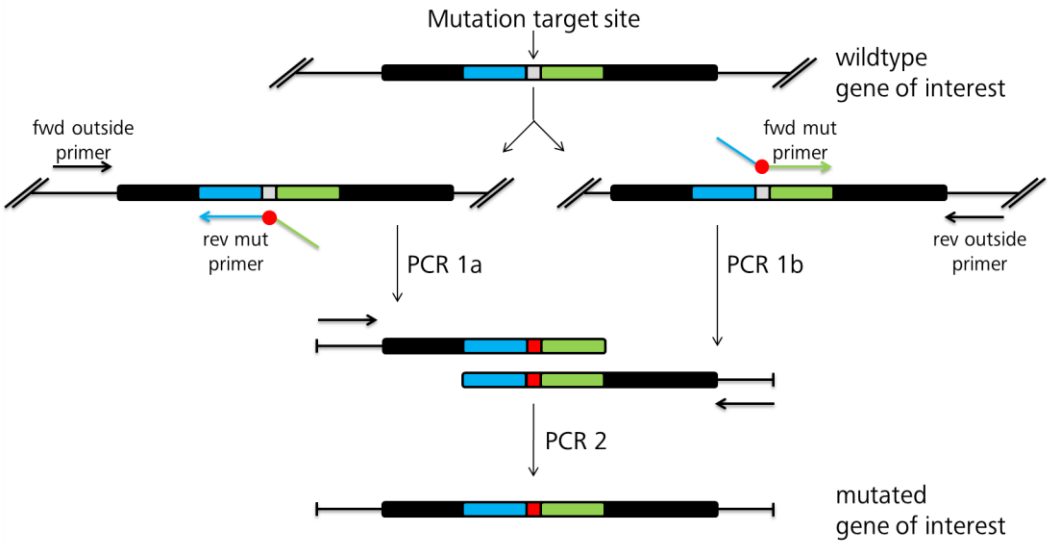


Figure 13. Fusion PCR for site-specific mutations. For site-specific sequence mutations, primers binding at the site of interest and harboring the mutation (fwd/rev mut primer) were combined in a first PCR (1a+b) with primers binding outside the gene of interest (fwd/rev outside primer). Due to a designed overlap, the two PCR products could be combined in a second reaction (PCR 2) applying the outside binding primers again.

For the introduction of site-specific mutations, a PCR fusion-based approach was employed (Figure 13). In a first PCR amplification (PCR 1a) step, a forward primer, binding outside the gene of interest (GOI), was combined with a reverse primer binding in front of the mutation site, harboring the desired mutation, and a 5' overhang complementary to the GOI. In the same way, a forward primer, with the respective mutation and an overhang binding after the mutation site was combined with a reverse primer binding outside the GOI in a second PCR reaction (PCR 1b). Both PCR products, which share the introduced overlap and mutation, were combined as template in an equimolar ratio in a third PCR amplification (PCR 2) with the same outside binding primers as before. The amplicon was then cloned using the standard cloning techniques (C.2.2.1).

C.2.2.3 Purification of plasmid DNA for transfections

Depending on the needed amount, the QIAprep Spin Miniprep Kit or the QIAGEN Plasmid Midi/Maxi/Mega Kit was used for purification of plasmid DNA.

QIAprep Spin Miniprep Kit	Qiagen (27106)
QIAGEN Plasmid Midi/Maxi/Mega Kit	Qiagen (12143/12163/12183)

C.2.3 Cell culture techniques

C.2.3.1 Cell line cultivation

Table 7. Cell lines

Cell Line	Description
CB-13/5 hybridoma	Mouse hybridoma cell line producing the monoclonal mouse anti-p24/p55 antibody CB-13/5[261].
Flp-In™ T-Rex™ 293 sus	Derived from the HEK293 cell line. Contains a single stably integrated FRT site for rapid generation of stable cell lines (Thermo Fisher - R78007). Adapted for suspension (sus) growth in serum-free medium[262].
FreeStyle™ 293-F	Derived from the HEK293 cell line and adapted for suspension growth in serum-free medium (Thermo Fisher - R79007)
GL9-CTL clone	CD8+ T cell clone isolated from an HLA B*07:02 HIV+ patient, specific for the p24-GL9 (GPGHKARVL) peptide (kind gift from Cornelis Melief, Leiden University Medical Center, Leiden, Netherlands)
HEK293T	Ad5-transformed human embryonic kidney cell line[263], expressing the "SV40 large T-antigen"[264]
LCL 554	EBV-transformed human lymphoblastoid cell line from an HLA B*07:02 positive donor (Vaecgene Biotech, Munich, Germany)
W6/32 hybridoma	Mouse hybridoma cell line producing the monoclonal mouse anti-HLA-ABC antibody W6/32 [103] (Sigma Aldrich - 84112003)

Table 8. Cell culture media and additives

Description	Ingredients
2-Mercaptoethanol (50 mM)	Thermo Fisher (31350-010)
CTL freezing medium	IMDM, 40% (v/v) FBS, 10% (v/v) DMSO
Dimethyl sulfoxide (DMSO)	Sigma Aldrich (D8418)
DMEM (=DMEM-0)	Dulbecco's Modified Eagle Medium (Thermo Fisher - 41966052)
DMEM-10	DMEM, 10% (v/v) FBS, 100 U/ml penicillin, 0.1 mg/ml streptomycin
DMEM-PS	DMEM, 100 U/ml penicillin, 0.1 mg/ml streptomycin
Fetal Bovine Serum (FBS)	Sigma Aldrich (F7524 – Lot: 123M3398)
FreeStyle medium	FreeStyle™ *293 Expression Medium* (Thermo Fisher - 12338026)
FreeStyle-0.5 medium	FreeStyle Medium, 50 U/ml penicillin, 0.05 mg/ml streptomycin
L-Glutamine (200 mM)	PAN Biotech, Aidenbach, Germany (P04-80100)
MEM Non-Essential Amino Acids Solution (100X)	Thermo Fisher (11140-035)
MEM Vitamin Solution (100X)	Thermo Fisher (11120-052)
Penicillin(10,000 U/ml) / Streptomycin (10 mg/ml)	PAN Biotech (P06-07100)
RPMI 1640	Roswell Park Memorial Institute 1640 medium (PAN Biotech - P04-16500)
RPMI-10	RPMI 1640, 10% (v/v) FBS, 100 U/ml penicillin, 0.1 mg/ml streptomycin
Sodium Pyruvate (100 mM)	Thermo Fisher (11360-039)

FreeStyle™ 293-F and Flp-In™ T-Rex™ 293 sus expression cells were cultured at 37°C, 8% CO_2 and 90 rpm using 1 L polycarbonate or 1 L borosilicate glass Erlenmeyer flasks containing 300 ml Freestyle-0.5 medium. By subculturing, the density was maintained between 0.2 - 2×10^6 cells/ml.

All other cell lines were maintained at 37°C and 5% CO_2. Adherent HEK293T cells were grown in DMEM-10 medium and subcultured in a 1:10 ratio twice a week. Hybridoma suspension cells were maintained between 3 - 9×10^5 cells/ml in RPMI-10. LCL 554 cells were subcultured in a 1:5 ratio twice a week in RPMI-10.

Dulbecco's Phosphate Buffered Saline (DPBS)	Sigma Aldrich (D8537)
Trypsin 0.05 %/EDTA 0.02 % in PBS	PAN Biotech (P10-023100)
Corning® Erlenmeyer cell culture flasks (capacity: 1 L; material: polycarbonate)	Sigma Aldrich (CLS431147)
Duran® Erlenmeyer narrow-neck flasks (capacity: 1 L; material: borosilicate glass)	Sigma Aldrich (Z232858)

C.2.3.2 Transient transfection

HEK293T cells

For transient transfection of plasmid DNA, HEK293T cells were cultured in DMEM-10 to a confluency of 80%. Afterwards the medium was aspirated and changed to DMEM without any additives (DMEM-0). The DNA was diluted in DMEM-0 and then mixed with the cationic polymer PEI (1 mg/ml in PEI-buffer, pH 7.5) in a 1:5 (w/w) ratio (Table 9). This transfection mix was incubated at room temperature for 10 min to allow the formation of DNA/PEI complexes. The transfection mix was subsequently added dropwise onto the HEK293T cells to a final concentration of 2 µg/ml DNA and 10 µg/ml PEI. After incubation at 37°C and 5% CO_2 for 6 h, the medium was changed again to DMEM-10.

Table 9. Transfection mixtures for different cell culture cultivation scales

Per reaction	Transfection mix volume [µl]	DNA [µg]	PEI [µg]	Final cell culture volume [ml]
2.2 cm dish (12 well)	50	1	5	0.5
3.5 cm dish (6 well)	100	2	10	1
15 cm dish	1800	36	180	18

FreeStyle™ 293-F cells and Flp-In™ T-Rex™ 293 sus expression cells

For transient transfection of suspension cells the concentration was adjusted to 1×10^6 cells/ml in 300 ml FreeStyle medium. 300 µg DNA were diluted in 6.5 ml DMEM-0 and 1.2 ml PEI were also diluted in 6.5 ml DMEM-0. After 5 min pre-incubation at room temperature the diluted PEI was added to the DNA (never the other way round). This transfection mix was incubated at room temperature for another 20 min to allow DNA/PEI complex formation and was subsequently added dropwise to the cells. After a 6 h incubation period at 37°C, 8% CO_2 and 90 rpm the medium of the cells was changed to FreeStyle-0.5.

PEI-buffer	25 mM HEPES, 150 mM NaCl
Polyethylenimine (PEI), linear, MW 25,000	Polysciences, Warrington, USA (23966)

C.2.3.3 Generation of stable expression cell lines

For generating stable mammalian expression cell lines, the Flp-recombinase-mediated integration system from Thermo Fisher was utilized. As target, Flp-In™ T-Rex™ 293 sus cells that are adapted for growth in suspension[262] were employed. These cells stably express a *lacZ*-Zeocin™ fusion gene under the control of an SV40 early promotor and harbor a single Flp-In target site (FRT) directly after the start-codon of the fusion gene. Co-transfecting the Flp-In™ T-Rex™ 293 with the Flp recombinase expression plasmid, pOG44, and the gene of interest cloned into a pcDNA5/FRT vector, mediates an integration of the pcDNA5/FRT expression construct through FRT site-specific DNA recombination in some cells[252,265]. Thereby, the Zeocin™ resistance gene is disrupted and the hygromycin resistance gene from the pcDNA5/FRT vector, which lacked a promotor and the ATG initiation-codon, gets activated by in-frame-insertion downstream of the SV40 early promoter and the start-codon (Figure 14). By adding hygromycin B to the medium, one can screen for successfully generated stable transfectants.

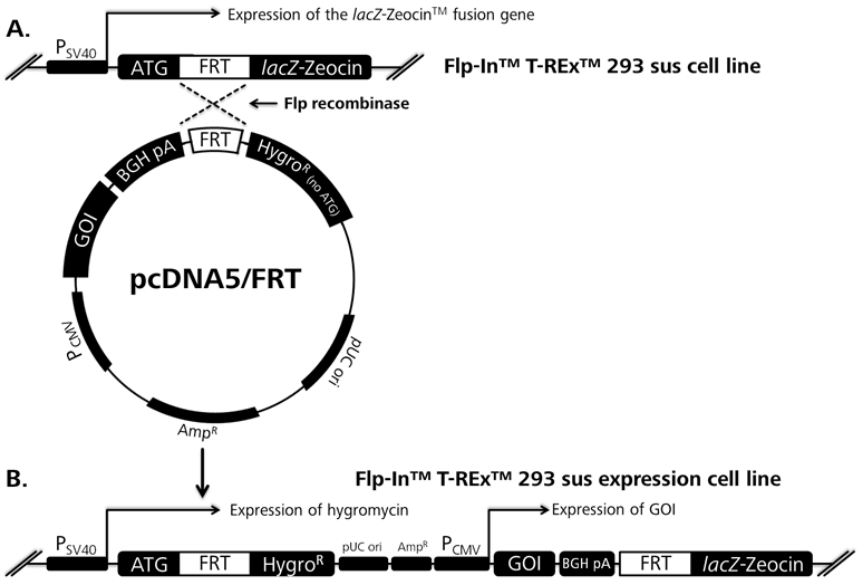


Figure 14. Generation of stable Flp-In™ T-Rex™ 293 sus expression cell lines. (A) The Zeocin™-resistant (by stable *lacZ*-Zeocin™ fusion gene expression) Flp-In™ T-Rex 293 sus cells were co-transfected with pcDNA5/FRT containing the gene of interest (GOI) and pOG44. The pOG44-expressed Flp recombinase catalyzes a homologous recombination (indicated by the dashed line) between the FRT sites in the cells and the pcDNA5/FRT vector. (B) Integration of the expression vector downstream of the start codon of the lacZ-Zeocin™ fusion gene confers Zeocin™ sensitivity and hygromycin resistance, while allowing CMV promoter driven GOI expression.

To start the stable cell line generation, 5×10^5 Flp-In™ T-Rex™ 293 cells were plated in a 3.5 cm dish (6 well plate) with 2 ml medium (FreeStyle medium, 5% (v/v) FBS, 50 U/ml penicillin, 0.05 mg/ml streptomycin, 100 µg/ml Zeocin™) and grown as adherent cells. The next day, the medium was changed to FreeStyle medium without any additives. Afterwards the cells were transiently transfected (C.2.3.2) with a 4:1 (w/w) mixture of the pcDNA5/FRT vector carrying the GOI and pOG44. After 6 h the medium was changed to FreeStyle medium with 5% (v/v) FBS, 50 U/ml penicillin and 0.05 mg/ml streptomycin. After 5 days, the cells were transferred into a 75 cm^2 cell culture flask. 6 h later; hygromycin B was added to a final concentration of 100 µg/ml, thereby starting the selection process for successful stable integration. Hygromycin-B-resistant isogenic foci were pooled and expanded to a cell count of 6×10^7 cells. These cells were than detached with trypsin/EDTA and transferred into a 1 L polycarbonate Erlenmeyer flask containing 300 ml FreeStyle-0.5 supplemented with 100 µg/ml hygromycin B and henceforth cultured as suspension cells at 37°C, 8% CO_2 and 90 rpm. By cloning the GOI into the tetracycline-independent expression vector pcDNA5/FRT a constitutive expression was achieved, bypassing the inducible expression characteristic of the Flp-In™ T-Rex™ 293 cells.

Hygromycin B	Invivogen, San Diego, USA (ant-hg)
Zeocin™	Invivogen (ant-zn)
Other cell culture media and additives	See Table 8

C.2.3.4 Generation of monocyte-derived dendritic cells

For the differentiation of monocyte-derived dendritic cells (mdDCs), a modified protocol as described by Thurner et al.[266] was applied. First human peripheral blood mononuclear cells (PBMCs) were isolated from 60 ml heparinized blood, drawn from healthy donors using Ficoll-Paque™ Plus density gradient centrifugation as follows: In two 50 ml tubes, 30 ml blood were

layered on top of 15 ml Ficoll-Paque™ Plus each, and centrifuged at 1000 g for 15 min without active breaking. The buffy coat was transferred to a new 50 ml tube and added up to 50 ml with P2 buffer.

Afterwards, CD14+ monocytes were isolated from the PBMCs via the magnetic-activated cell sorting (MACS) technology from Miltenyi Biotech. The PBMCs were pelleted (600 g, 10 min) and resuspended in 800 µl MACS buffer. 50 µl of human CD14 MicroBeads were added for positive selection of monocytes. After incubation at 4°C for 15 min, cells were washed two times (pelleted each time by centrifugation at 600 g for 10 min) with 5 ml MACS buffer to remove unbound MicroBeads and were finally resuspended in 500 µl MACS buffer. For magnetic separation a MS column was placed in the magnetic field of an OctoMACS Separator and equilibrated with 500 µl MACS buffer. The PBMCs were applied onto the column and washed three times with 500 µl MACS buffer to remove unlabeled cells. The column was then unplugged from the magnetic field and positively selected CD14+ cells were flushed out with 1 ml MACS buffer. The purity of the isolated cells was determined by assessing CD14 expression via flow cytometry (C.2.6.1).

For differentiation of the CD14+ monocytes to mdDCs, the eluted cells were pelleted (500 g, 10 min) and cultivated in mdDC medium at a concentration of 1×10^6 cells/ml over a six-day period. To promote mdDC differentiation, 1000 U/ml GM-CSF and 1000 U/ml IL-4 were added to the medium. After two days the medium was renewed. The final quality of the mdDCs was evaluated in a flow cytometric analysis (C.2.6.1) with surface staining for the dendritic cell marker CD1a.

CD14 MicroBeads, human	Miltenyi Biotec, Bergisch Gladbach, Germany (130-050-201)
Ficoll-Paque™ PLUS	GE Healthcare, Little Chalfont, UK (17-1440-02)
Heparin	Heparin-Natrium-5000-ratiopharm, Ulm Germany
MACS buffer	DPBS, 1% (v/v) FBS, 20 mM EDTA, degassed
mdDC medium	RPMI 1640, 10% (v/v) FBS, 100 U/ml penicillin, 0.1 mg/ml streptomycin, 1x non-essential amino acids solution, 1x vitamin solution, 2 mM L-glutamine, 1 mM sodium pyruvate, 100 µM 2-Mercaptoethanol, 1,000 U/ml IL-4, 1,000 U/ml GM-CSF
MS columns	Miltenyi Biotec (130-042-201)
OctoMACS™ Separator	Miltenyi Biotec (130-042-109)
P2 buffer	DPBS, 1% (v/v) FBS
Recombinant human GM-CSF	Miltenyi Biotech (130-093-865)
Recombinant human IL-4	Miltenyi Biotech (130-093-921)
Other cell culture media and additives	See Table 8

C.2.3.5 Proliferation of CTL clones

For p24-GL9-peptide-specific (C.1.3) cytotoxic T cell (CTL) clone proliferation, the cells were stimulated with peptide-pulsed HLA-B*07:02 expressing LCL 554 feeder cells.

For this, an aliquot of the previously expanded GL9 CTL clone was thawed and adjusted to a concentration of $1x10^5$ cells/ml in IMDM-8+. 100 µl/well of this cell suspension were seeded in a 96 round bottom well plate and cultivated at 37°C and 5% CO_2. Two days later, LCL cells were adjusted in RPMI-10 to a concentration of 1×10^7 cells/ml, pulsed with 30 µg synthetic GL9-peptide per 1×10^7 cells and incubated for 3 h at 37°C. The feeder cells were subsequently

irradiated in a gamma-ray device with 60 gy and afterwards adjusted with IMDM-8+ to $1x10^6$ cells/ml. Co-cultivation was started by adding 100 µl of LCL cells to each well of the previously seeded CTL clones, resulting in a 10-times excess of LCL cells over CTLs. Every 3 days 70 µl of conditioned medium were carefully aspirated and replaced with 100 µl fresh medium. After 14 days, the proliferating CTLs were pooled, adjusted to 1×10^5 cells/ml in IMDM-8+, and again stimulated with pulsed LCL cells, as described above. After another 10 days of cultivation, the cells were pooled, pelleted through centrifugation (200 g, 10 min), adjusted to 1×10^7 cells/ml in CTL freezing medium and frozen in 1 ml aliquots with assistance of the Ice Cube 14S computer-controlled freezing device (SY-LAB), and stored in a liquid nitrogen cryopreservation system.

IMDM	Iscove's Modified Dulbecco's Medium (Thermo Fisher - 21980032)
IMDM-8+	IMDM, 10% (v/v) FBS, 8% (v/v) TCGF, 10 ng/ml IL-15, 50 U/ml penicillin, 0.05 mg/ml streptomycin
Natural human interleukin-2 (IL-2) / T cell growth factor (TCGF) (500 BRMP units/ml)	Helvetica Health Care, Geneva, Switzerland (0801017)
Recombinant human IL-15	CellGenix, Freiburg im Breisgau, Germany (1413-010)
Other cell culture media and additives	See Table 8

C.2.3.6 VLP-mdDC co-cultivation

The protocol for VLP co-cultivation was optimized together with Tanja Stief[267]. At first, 5×10^5 mdDCs/well were seeded in a 96 round bottom well plate in 100 µl mdDC medium (C.2.3.4). After a 3 h cultivation period, varying amounts of VLPs were directly added to the mdDCs. Another two hours later, the cells were pelleted (500g, 5min), the medium aspirated, the cells washed with 200 µl DPBS and finally resuspended in 100 µl RPMI-10. After another 4 h incubation period, the cells were employed for CTL restimulation experiments (C.2.3.8).

Cell culture media and additives	See Table 8

C.2.3.7 Peptide pulsing

For peptide pulsing experiments, 5×10^5 mdDCs/well were seeded into a 96 round bottom well plate in 100 µl mdDC medium (C.2.3.4). After a 3 h incubation period, 1 µl of varying peptide (C.1.3) concentrations, dissolved and diluted in DPBS, was directly pipetted to the mdDCs. Another 2 hours later the cells were pelleted (500 g, 5 min), the medium aspirated, the cells washed two times with 100 µl DPBS and finally resuspended in 100 µl RPMI-10. These cells were used for CTL restimulation assays (C.2.3.8) instantaneously.

Cell culture media and additives	See Table 8

C.2.3.8 mdDC and CTL mixed leukocyte reaction

Liquid-nitrogen-frozen CTLs (C.2.3.5) were quickly thawed and adjusted to 5×10^5 cells/ml with RPMI-10 3 hours prior to starting the co-cultivation. For the mixed leukocyte reaction, 100 µl of CTL cells were added to already seeded mdDCs (C.2.3.6 and C.2.3.7), resulting in an effector-to-target ratio of 1:1. To inhibit the secretion of cytokines, Brefeldin A was added to a final

concentration of 1 µg/ml. The mixed leukocyte reaction (MLA) was performed for 6 h at 37°C and 5% CO_2. Cells were then directly employed for flow cytometry analysis (C.2.6).

Brefeldin A (BFA)	Sigma Aldrich (B6542)
Other cell culture media and additives	See Table 8

C.2.4 Virological techniques

C.2.4.1 HIV-1 production

To produce HIV-1 viral particles, HEK-293T cells were transiently transfected in 6-well plates (C.2.3.2) with molecular clones of full-length, replication- and infection-competent HIV-1. 24, 48 and 72 h after transfection the virus-containing supernatant was completely harvested and replaced with fresh medium. The conditioned medium was cleared by centrifugation (1000 g, 5 min). For virus inactivation and solubilization of Gag, the virus containing, cell-free supernatant was incubated with Triton X-100 in a final concentration of 1% (v/v) for a minimum of 4 h[268,269].

Triton™ X-100	Sigma Aldrich (T8787)

C.2.4.2 RT-Assay for virus quantification

To determine the amount of virus produced, the RT activity was quantified with the RetroSys™ RT Activity Kit[270]. The assay procedure consists of two steps. First, the RT in the virus-containing medium synthesizes a DNA strand using an immobilized polyA template and an oligo-dT primer. Bromo-deoxyuridine triphosphate (BrdUTP), which is added to the reaction, gets incorporated into the DNA, and, in the second step, gets quantified with a BrdU-binding antibody conjugated to alkaline phosphatase. The activity of the alkaline phosphatase is measured in a colorimetric reaction and is proportional to the RT activity of the sample.

The assay was performed according to the manufacture's protocol for quantification of RT activity, with the exception that the RT reaction step was always carried out overnight.

RetroSys™ RT Activity Kit	Innovagen, Lund, Sweden (RT-001 - discontinued)

C.2.5 Protein biochemistry techniques

C.2.5.1 Antibody biotinylation

Biotinylation of antibodies is an excellent way to detect them with high specificity. For the Gag ELISA (C.2.5.2), the detection-antibody 37G12 had to be biotinylated. The protocol for it was based on personal communication with the distributer of the antibody (Polymun Scientific, Klosterneuburg, Austria). First the buffer of the antibody was exchanged to DPBS using NAP-5 gel filtration columns according to the manufacturer's instructions. The antibody was eluted from the column stepwise with 100 µl DPBS. Protein containing fractions, as determined spectrophotometrically with the NanoDrop ND-1000 at 280 nm, were pooled. Next, a 10 mM solution of the biotinylation reagent EZ-Link-NHS-LC-Biotin in dimethylformamide was prepared. The biotin reagent was added in a 75 molar excess to the antibody in PBS (40.5 µl 10 mM EZ-

Link-NHS-LC-Biotin per 1 mg antibody). After an incubation period of 1 h at room temperature, the buffer was changed to DPBS by NAP-5 column chromatography eluting the antibody stepwise with 100 µl DPBS. Antibody-containing fractions were pooled and stabilized by adding BSA and NaN_3 to a final concentration of 0.1% (w/v) and 0.2% (w/v), respectively.

Dulbecco's Phosphate Buffered Saline (DPBS)	Sigma Aldrich (D8537)
EZ-Link-NHS-LC-Biotin	Thermo Fisher (21336)
NAP-5 gelfiltration columns	GE Healthcare Life Sciences (17-0853-01)
Antibodies	See C.1.2

C.2.5.2 Gag-ELISA

Gag concentrations in cell culture supernatants were determined in a quantitative p24-sandwich-ELISA[271]. Both antibodies used in this assay bind to the HIV-1 capsid protein p24, but also recognize the full-length Gag protein, which makes it applicable for VLP-, as well as for virus-quantification.

First, a MaxiSorp 96 flat bottom well plate was coated with 0.25 µg/well of the M01-16/4/1 antibody in 100 µl coating buffer and incubated at 4°C overnight. The next day, the plate was washed with 3 well volumes of PBS-T using the HydroFlex plate washer (Tecan, Männedorf, Switzerland). Afterwards, 100 µl/well of Triton-treated (0.5% (v/v) final concentration, incubated for 1 h at room temperature) conditioned medium that had been cleared by centrifugation (two subsequent centrifugations at 500 g and 1000 g for 5 min each), and different dilutions of recombinant p24 protein in medium as standard were added to the plate and incubated for 1 h at 37°C. Following another washing step with 6 well volumes PBS-T, 6.8 ng of the biotinylated (C.2.5.1) detection antibody 37G12 in 100 µl PBS with 1% (w/v) BSA were added per well and incubated for 1 h at room temperature. The plate was then washed with 10 well volumes of PBS-T. Subsequently 5 mU horseradish peroxidase conjugated streptavidin, which binds specific and with high affinity to biotin, in 100 µl PBS with 1% (w/v) BSA were added per well. The plate was incubated for 30 min at room temperature and then washed with 10 well volumes of PBS-T. The colorimetric quantification-reaction was started by adding 100 µl/well of freshly prepared TMB substrate and stopped after 2 min by pipetting 50 µl of 1 M H_2SO_4 into each well. The colorimetric reaction was measured at 450 nm with a microplate reader (Model 680 - Bio-Rad Laboratories, Hercules, USA) and the Gag concentration in the samples was determined by reference to the recombinant p24 protein as standard in a nonlinear regression analysis with *GraphPad Prism 5* (variable slope (four parameter) equation).

Coating buffer	0.1 M carbonate, pH 9.5
Nunc MaxiSorp® flat-bottom 96 well plate	NeoLab, Schwabing-Freimann, Germany (104342404)
PBS	137 mM NaCl, 2.7 mM KCl, 10 mM Na_2PO_4, 1.8 mM KH_2PO_4, adjusted to pH 7.4 with HCl
PBS-T	PBS, 0.05% (v/v) Tween-20
Recombinant HIV-1 p24 protein	Abcam (ab43037)
Streptavidin-POD	Sigma Aldrich (11089153001)
3,3',5,5'-Tetramethylbenzidin (TMB)	Carl Roth, Karlsruhe, Germany(R6350)
TMB A	30 mM tri-potassium citrate-monohydrate adjusted to pH 4.1 with 10% (w/v) citric acid
TMB B	0.24% (w/v) TMB, 10% (v/v) acetone, 90% (v/v) ethanol, 80 mM H_2O_2
TMB substrate	TMB A and TMB 2 mixed in a 20:1 ratio

Triton™ X-100	Sigma Aldrich (T8787)
Tween-20	Merck Millipore, Darmstadt, Germany (822184)
Antibodies	See C.1.2

C.2.5.3 sHLA ELISA

Relative concentrations of intact sHLA complexes in conditioned medium were determined in a quaternary-structure-dependent sandwich ELISA. This ELISA utilizes the characteristic of the monoclonal α-pan-HLA antibody[103] W6/32 that binds only intact HLA/sHLA complexes, consisting of the HLA α-chain, β2m, and a peptide bound to the epitope binding groove[104].

First, a MaxiSorp 96 well plate was coated with 0.1 µg/well W6/32 antibody in 100 µl coating buffer and incubated overnight at 4°C. After washing with 3 well volumes of PBS-T, free binding sites of the plate were blocked with 3% (w/v) BSA in PBS overnight at 4°C. The next day, each well was washed again with 3 well volumes of PBS-T and 100 µl of centrifugation-cleared (500 g, 5 min) sample-supernatant were pipetted into each well. After incubation for 1 h at 37°C, all wells were washed with 6 well volumes PBS-T. For detection of bound sHLA complexes, 0.1 µg biotinylated α-6x-His Epitope Tag antibody in 100 µl PBS with 1% (w/v) BSA were added per well, incubated for 1 h at room temperature, and subsequently washed with 6 well volumes PBS-T. Finally, 5 mU horseradish-peroxidase-conjugated streptavidin in 100 µl PBS with 1% (w/v) BSA were added per well. The plate was incubated for 30 min at room temperature and then washed with 10 well volumes of PBS-T. The colorimetric quantification-reaction was started by adding 100 µl/well of freshly prepared TMB substrate and stopped after about 10-15 min, depending on the experiment, by pipetting 50 µl of 1 M H_2SO_4 into each well. The colorimetric reaction was measured at 450 nm in a microplate reader (Model 680 - Bio-Rad Laboratories).

Buffers and substrates	See C.2.5.2
Antibodies	See C.1.2

C.2.5.4 SEAP Assay

To account for variables like transfection efficiency or varying cell numbers in transfection experiments, normalization based on the expression of the reporter gene secreted embryonic alkaline phosphatase (SEAP) was established previously[272,273]. SEAP gets secreted by the cells and can be quantified directly in the supernatant using a colorimetric assay[274].

For the assay, the pBluescript KS(-) CMV-SEAP vector, which expresses SEAP under control of the CMV promoter, was co-transfected (B.4.3.2) with the plasmid expressing the GOI in a 1:20 (w/w) ratio. The supernatant was harvested after the indicated time and clarified by two subsequent centrifugations (500 g and 1000 g for 5 min each). The samples were then diluted with DMEM-PS if needed. After heating at 65°C for 10 min, 10µl of the samples were mixed with 90 µl dH_2O and 100 µl 2xSEAP assay buffer. The heating and the L-homoarginine in the buffer inhibit endogenous alkaline phosphatase activities[275]. The samples were then pre-warmed to 37°C for 10 min in a 96-flat bottom well plate. To start the colorimetric reaction, 20 µl of 120 mM para-nitrophenyl phosphate dissolved in 1xSEAP assay buffer were added to each well. The hydrolysis of the substrate by SEAP accompanies an increased in absorbance at 405 nm over time[276], which was measured with a microplate reader (Model 680 - Bio-Rad Laboratories) at 37°C. To quantify the measurements, a relative SEAP reference standard was included in every experiment and analyzed in a linear regression analysis (Excel 2010).

L-homoarginine hydrochloride	MP Biomedicals, Santa Ana, USA (05221003)
Para-nitrophenyl phosphate disodium salt $6H_2O$	AppliChem (A1442)
SEAP buffer (1x)	1 M Diethanolamine, 0.5 mM $MgCl_2$,10 mM L-homoarginine, adjusted to pH 9.8 with HCl
Other cell culture media and additives	See Table 8

C.2.5.5 Bradford Assay

To quantify protein samples via the method by Bradford[277] the "Bio-Rad Protein Assay" was applied according to the manufacturer's instructions, using BSA in different concentrations as protein standard.

Bovine serum albumin (BSA), fraction V	Biomol, Hamburg, Germany (01400)
Bio-Rad Protein Assay Kit	Bio-Rad, Hercules, USA (5000001)

C.2.5.6 VLP production

For VLP production, varying quantities of HEK293T cells, depending on the experiment, were transiently transfected (C.2.3.2) with Gag-expression-plasmids. For VSV-G pseudotyped VLPs Gag DNA was mixed with the VSG-G plasmid 1:1 (w/w). The VLP-containing supernatant was harvested 48 h post transfection and clarified by centrifugation (3000 g, 15 min, and 4°C). To concentrate the VLPs, 30 ml of the supernatant were pipetted onto a 5 ml 30% (w/w) sucrose (in DPBS) cushion in 38.5 ml centrifuge tubes and centrifuged at 100,000 g for 2 h at 4°C in an ultracentrifuge (Optima L-90K, Beckman Coulter with rotor SW 32 Ti). Afterwards the pelleted VLPs were resuspended in an appropriate amount of DPBS.

Dulbecco's Phosphate Buffered Saline (DPBS)	Sigma Aldrich (D8537)
Centrifuge tubes (38.5 ml, 25x89 mm)	Beckman Coulter (326823)

C.2.5.7 Sucrose density gradient centrifugation

To purify VLPs according to their density, the concentrated VLPs (C.2.5.6), or the clarified supernatant directly, were loaded onto a 10-50% (w/w) sucrose gradient (2 ml of 10, 20, 30, 40 and 50% each) in 14 ml centrifuge tubes and centrifuged for 2.5 h at 100,000 g and 4°C (Optima L-90K, Beckman Coulter with rotor SW 40 Ti). Then, fractions of 550 µl each were collected and analyzed for the presence of Gag-VLPs by Gag-ELISA (C.2.5.2) and by immunodetection (C.2.5.14) after Western blotting (C.2.5.12). The density of each fraction was determined by weighing on a high-precision scale (Sartorius analytic).

Centrifuge tubes (14 ml, 14x95 mm)	Beckman Coulter (344060)

C.2.5.8 Transmission and scanning electron microscopy

VLP-containing sucrose gradient fractions (C.2.5.7) were pooled, transferred to a 38.5 ml centrifuge tube and filled up with DPBS. After an ultracentrifuge run (Optima L-90K, Beckman Coulter with rotor SW 32 Ti) at 100,000 g for 2 h at 4°C, the pelleted VLPs were resuspended in 200 µl DPBS. To fixate the samples for electron microscopy 180 µl of the VLPs in DPBS were mixed with 20 µl of a 25% (w/w) glutaraldehyde solution. The 2.5%-glutaraldehyde-fixed samples were analyzed via transmission electron microscopy (TEM) and field emission scanning electron microscopy (FESEM) by Professor Gerhard Wanner at the department for ultrastructure research of the Ludwig-Maximilian University of Munich, as described briefly below.

For TEM, the fixed samples were deposited on carbon-coated copper grids and negatively stained with 2.0% (w/v) phosphotungstic acid (PTA), pH 7.0. Transmission electron micrographs of samples were taken with an EM 912 electron microscope (Zeiss, Oberkochen, Germany) equipped with an integrated OMEGA energy filter operated at 80 kV in the zero loss mode.

For FESEM, drops of the fixed sample were placed onto a glass slide, covered with a coverslip, and rapidly frozen with liquid nitrogen. The coverslip was removed with a razor blade and the glass slide was immediately fixed with 2.5% (w/v) glutaraldehyde in fixative buffer, postfixed with 1% (w/v) osmium tetroxide in fixative buffer, dehydrated in a graded series of acetone solutions, and critical-point dried from liquid CO_2. Specimens were mounted on stubs, coated with 3 nm platinum using a magnetron sputter coater, and examined with a Zeiss Auriga scanning electron microscope operated at 1 kV.

Centrifuge tubes (38.5 ml, 25x89 mm)	Beckman Coulter (326823)
Dulbecco's Phosphate Buffered Saline (DPBS)	Sigma Aldrich (D8537)
Fixative buffer	50 mM cacodylate buffer, pH 7.0
Glutaraldehyde solution, 25% in H_2O	Sigma Aldrich (G5882)

C.2.5.9 Dynamic light scattering

Virus-like particle size analysis was performed using dynamic light scattering (DLS). For this, sample preparation was identical as described for electron microscopy imaging (C.2.5.8) except that no fixation was performed. The purified VLPs were diluted in DPBS as required and examined with a High Performance Particle Sizer (Malvern Instruments, Malvern, UK) at the Biochemistry I department of the University of Regensburg using a predefined standard protocol for particles solved in PBS.

Dulbecco's Phosphate Buffered Saline (DPBS)	Sigma Aldrich (D8537)

C.2.5.10 SDS-PAGE

Proteins were separated according to their size via sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)[278]. Protein-containing samples were mixed with 2x Laemmli buffer and boiled at 95°C for 5 min and subsequently shortly centrifuged (30 s, 5000 g). Samples were then loaded together with a prestained protein ladder for reference on a 10-15% (depending on

the respective experiment) SDS gel and electrophoresis was performed at 30 mA using the Mighty small II (8x7cm) system (Serva Electrophoresis; Heidelberg, Germany).

Laemmli buffer (1x)	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	Thermo Fisher (26620)

C.2.5.11 Coomassie staining

For direct visualization of all proteins, SDS gels were incubated in Coomassie staining solution for 15 min and afterwards destained in dH_2O. Approximate protein quantification was achieved by densitometrical analysis with comparison to BSA reference samples that were also loaded on the same gel.

Coomassie Brilliant Blue R-250	AppliChem (A1092)
Coomassie staining solution	1.25% (w/v) Coomassie Brilliant Blue R-250, 50% (v/v) ethanol, 7% (v/v) acetic acid

C.2.5.12 Western blot

For Western blotting of the SDS-PAGE-separated proteins (C.2.5.10) onto a nitrocellulose membrane (pore size 0.2 µm or 0.45 µm, depending on the protein of interest's size) the BlueFlash-L semi-dry blotting unit (Serva Electrophoresis) was utilized according to the manufacturer's instructions (with 2.5 mA/cm^2 gel for 1.5 h). Uniform loading and blotting was checked by reversible staining of the membrane with Ponceau S staining solution for 1 min and subsequent destaining with H_2O.

Nitrocellulose Membrane, Amersham Protran (pore size 0.2 µm/0.45 µm)	GE Healthcare Life Sciences (10600001 / 10600003)
Ponceau S staining solution	0.2% (w/v) Ponceau S, 1% (v/v) acetic acid
Whatman™ 3MM Chr Blotting Paper	GE Healthcare Life Sciences (3030-917)

C.2.5.13 Slot blot

To transfer protein solutions directly to nitrocellulose membranes, the Bio-Dot® SF microfiltration apparatus (Bio-Rad, Hercules, USA) was used. The membrane (9x12 cm) was pre-incubated in transfer buffer for 10 min and then placed in the Bio-Dot apparatus on top of five transfer-buffer-moistened Whatman blotting papers (7.7x11.3 cm). Gentle vacuum was applied and the apparatus was equilibrated with 250 µl transfer buffer per well. Afterwards, 10 µl of the protein sample dilutions in 200 µl PBS were loaded per well and then washed with 250 µl transfer buffer.

Nitrocellulose Membrane, Amersham Protran (pore size 0.2 µm/0.45 µm)	GE Healthcare Life Sciences (10600001 / 10600003)
PBS	137 mM NaCl, 2.7 mM KCl, 10 mM Na_2PO_4, 1.8 mM KH_2PO_4, adjusted to pH 7.4 with HCl
Transfer buffer	25mM Tris, 150mM glycine, 10% (v/v) methanol
Whatman™ 3MM Chr Blotting Paper	GE Healthcare Life Sciences (3030-917)

C.2.5.14 Immunodetection on membrane

For immunodetection of specific proteins after Western or slot blots, the membrane was blocked overnight at 4°C in TBS-M. After washing 3 times with TBS-T and incubation with the primary antibody in TBS (for antibody concentrations see C.1.2) for 1 h at room temperature, the membrane was washed again 3 times with TBS-T. The secondary HRP-conjugated antibody or streptavidin-POD (diluted 1:2,000) in TBS were added and again incubated for 1 h. Last, the membrane was subjected to 3 washing steps with TBS-T and incubated with ECL substrate solution. After 5 min, the substrate was removed and the chemiluminescence emitted by the peroxidase activity was detected with a ChemiluxPro device (Intas, Göttingen, Germany). If necessary, to reach higher sensitivity the SuperSignal™ Femto Substrate was used instead of the standard ECL substrate.

In order to reuse a membrane for detection of another protein, the primary and secondary antibodies were removed by incubating with stripping buffer for 30 min at 50°C. After two washing steps with TBS-T, the membrane was blocked with TBS-M and subjected to another immunodetection round.

ECL substrate solution	100 mM Tris/HCl pH 8.5, 12.5 mM luminol, 1.98 mM coumaric acid, 305 µl/l 30% (v/v) H_2O_2
Streptavidin-POD	Sigma Aldrich (11089153001)
Stripping buffer	62.5 mM Tris pH6.6, 100 mM β-mercaptoethanol, 2% (w/v) SDS
SuperSignal™ West Femto Maximum Sensitivity Substrate	Thermo Fisher (34095)
Tris buffered saline (TBS)	50 mM Tris/HCl, 150 mM NaCl, adjusted to pH 7.5 with HCl
TBS-M	TBS, 5% (w/v) skimmed powdered milk
TBS-T	TBS, 0.05% (v/v) Tween-20
Antibodies	See C.1.2

C.2.6 Flow-cytometry

Flow-cytometric multi-parametric analysis of cells was performed with the FACSCanto™ II cell analyzer (BD, Franklin Lakes, USA) along with the provided FACSDiva™ software. For the individual antibody concentrations used, see C.1.2.

C.2.6.1 Surface marker staining

To stain for expression of cell surface markers, 5×10^5 cells were washed 2 times by adding 100 µl FC buffer, followed by centrifugation (500 g, 5 min, 4°C) and aspiration of the buffer. The cells were then incubated with a fluorophore conjugated antibody (diluted in 50 µl FC buffer) for 25 min at 4°C. To remove unbound antibody, the cells were washed 3 times with FC buffer, as described above, and finally resuspended in 100 µl FC buffer.

FC buffer	PBS, 1% (v/v) FBS
Antibodies	See C.1.2

C.2.6.2 Intracellular staining (ICS)

Staining for intracellularly expressed molecules was achieved by washing 5×10^5 cells 2 times by adding 100 µl PBS, followed by centrifugation (500 g, 5 min, 4°C) and aspiration of the buffer. Afterward, permeabilization and fixation of the cells was achieved by incubation with 50 µl cytofix/cytoperm for 20 min at 4°C. After 3 washing steps with 100 µl perm/wash buffer, as described above, the cells were incubated with the fluorophore-conjugated antibody (diluted in 50 µl perm/wash buffer) for 25 min at 4°C. After another 3 wash steps with 100 µl perm/wash buffer, the cells were resuspended in 100 µl FC buffer.

Cytofix/cytoperm	4% (w/v) paraformaldehyde, 1% (w/v) saponin
FC buffer	PBS, 1% (v/v) FBS
Paraformaldehyde	Merck Millipore (104005)
Perm/wash buffer	PBS, 0.1% (w/v) Saponin
PBS	137 mM NaCl, 2.7 mM KCl, 10 mM Na_2PO_4, 1.8 mM KH_2PO_4, adjusted to pH 7.4 with HCl
Saponin from quillaja bark	Sigma Aldrich (S4521)
Antibodies	See C.1.2

C.2.7 Epitope sequencing

To avoid contaminations and sample loss during preparation steps, only protein low-bind or low-retention plastic-ware was used. Solutions and buffers were prepared in clean glass ware.

C.2.7.1 W6/32 affinity matrix preparation

sHLA purification from conditioned cell medium was based on affinity chromatography techniques. The affinity matrix was composed of the α-pan-HLA W6/32 antibody (C.1.2) crosslinked to Protein A-Sepharose[279,280].

First, for every 1 l of conditioned sHLA containing medium, 100 mg of the lyophilized Protein A-Sepharose were swollen in 2 ml borate buffer for 30 min at room temperature. The resin was subsequently poured onto the porous polyethylene disc of a 5 ml disposable column and washed with 5 column volumes (cv) of borate buffer. Afterwards, 2 mg of W6/32 antibody per 100 mg Protein A were applied to the column. Protein A binds the Fc portion of the mouse IgG2a antibody W6/32 with strong affinity, while leaving the antigen-specific sites for sHLA affinity chromatography untouched. As source for the antibody, the conditioned W6/32 hybridoma cell supernatant (antibody concentration was lot dependent, but always about 40µg/ml, as determined by immunodetection after Western blotting and comparison to purified W6/32 antibody as standard) was employed. The medium was cleared by centrifugation (1000 g, 10 min), mixed 1:1 with borate buffer, loaded onto the column with a peristaltic pump, and allowed to empty by gravity flow. Residual unspecific proteins were removed by washing with 10 cv borate buffer. To avoid elution of the antibody in later steps, crosslinking to Protein A was performed. For this, the column was equilibrated with 5 cv of crosslinking coupling buffer (CCB) and then washed with 2 cv 40 mM dimethyl pimelimidate (DMP), a bifunctional coupling reagent, in CCB. Following, the bottom cap was secured on the column tip and the Protein A-Sepharose was resuspended in 2 cv 40 mM DMP in CCB. After 1h at room temperature, the bottom cap was removed, the column was allowed to empty by gravity flow, and then was washed with 5 cv CCB. The crosslinking reaction was quenched by washing with 2 cv of ice-cold crosslinking

termination buffer (CTB) and subsequent incubation with 2 cv CTB for 10 min (performed as described above for DMP incubation). After subsequent washing with 5 cv CTB and 5 cv borate buffer, all not covalently bound antibody was eliminated by washing with 5 cv acid wash buffer. After a final washing step with 10 cv borate buffer the column was stored in borate buffer with 0.02% (w/v) NaN_3 at 4°C until further use.

Acid wash buffer	0.58% (v/v) acetic acid, 150 mM NaCl
Borate-buffer	0.1M Boric acid, adjusted to pH 8.2 with NaOH
Crosslinking coupling buffer (CCB)	0.2 M triethanolamine, adjusted to pH 8.2 with HCl
Crosslinking termination buffer (CTB)	0.1 M ethanolamine, adjusted to pH 8.2 with HCl
Disposable Columns, 5 mL, Pierce™	Thermo Fisher (29922)
Dimethyl pimelimidate dihydrochloride (DMP)	Sigma Aldrich (D8388)
Ethanolamine	Carl Roth (0342)
Protein A-Sepharose CL-4B	Sigma Aldrich (P3391)
Triethanolamine	Carl Roth (6300)

C.2.7.2 Affinity chromatography of peptide bound sHLA complexes

The antibody W6/32 has the unusual quality to only bind intact complexes consisting of the HLA α-chain, β2m and a binding groove attached peptide. This property was harnessed to isolate only peptide-bound sHLA complexes by affinity chromatography of sHLA containing cell supernatant.

The pre-cleared (centrifugation at 1000 g, 10 min) and subsequently filtered supernatant was mixed in a 3:1 ratio with borate buffer and applied onto the prepared affinity column (C.2.7.1) using a peristaltic pump (Pump P1, Pharmacia, New Jersey, USA) and allowed to empty by gravity flow at 4°C. The column was subsequently washed with 15 cv borate buffer and 5 cv dH_2O. The sHLA complexes were eluted stepwise with 1 cv 10% acetic acid each, until the eluate reached a baseline absorbance at 280 nm, determined at a NanoDrop ND-1000.

Acetic acid	Sigma Aldrich (33209-1L-GL)
Borate-buffer	0.1M Boric acid, adjusted to pH 8.2 with NaOH
Steritop-GP vacuum filtration system (0.22µm)	Merck Millipore (SCGPT02RE)

C.2.7.3 Peptide purification

To dissociate peptides from the sHLA-heavy chain and β2m, the protein-containing eluate fractions (identified by 280 nm absorbance at a NanoDrop ND-1000) from the affinity chromatography (C.2.7.2) were heated to 78°C for 10 min[119]. Following this “acid boil” step, the peptides were purified by centrifugation using a centrifugal filter with a nominal molecular weight limit cutoff of 3 kDa (3 kDa cut-off filter) according to the manufacturer’s instructions. The concentrate containing all proteins bigger than 3 kDa was saved for analytical purposes. The filtrate that contained the dissociated peptides was dried to complete dryness in a vacuum concentrator (Bachofer).

Centrifugal filter; Amicon® Ultra-0.5 - 3 kDa	Sigma Aldrich (Z740169)

C.2.7.4 HPLC and LC-MS/MS

Purified peptide samples (C.2.7.3) were examined and sequenced at the proteomics core facility of the “Helmholtz Zentrum München” as follows. SpeedVac-dried samples were resuspended in 2% acetonitrile (ACN) in 0.5% trifluoroacetic acid (TFA) and LC-MS/MS analysis was performed as described previously using an Ultimate 3000 nano-RSLC (Dionex, Sunnyvale, USA) coupled to a LTQ-Orbitrap XL mass spectrometer (Thermo Fisher Scientific)[281]. Before loading, the samples were centrifuged for 5 min at 4°C. Every sample was automatically injected and loaded onto the trap column (Acclaim PepMap100, C18, 5 µm, 100 Å, 300 µm i.d. x 5 mm, Dionex) at a flow rate of 30 µl/min. After 5 min, the peptides were eluted from the trap column and separated on the analytical column (Acclaim PepMap RSLC C18, 2 µm, 100 Å, 75 µm i.d. x 25 cm, Dionex) by a 60 min gradient from 5 to 30% of ACN in 0.1% formic acid (FA) at 300 nl/min flow rate followed by a short gradient from 30 to 73% ACN in 0.1% FA in 5 min. Between each sample, the gradient was set back to 5% ACN in 0.1% FA and left to equilibrate for 20 min. From the MS prescan, the 10 most abundant peptide ions were selected for fragmentation in the linear ion trap if they exceeded an intensity of at least 200 counts and if they were of charge states +1 to +3, with a dynamic exclusion of 30 sec. During fragment analysis, a high-resolution (60,000 full-width half maximum) MS spectrum was acquired in the Orbitrap with a mass range from 300 to 1500 Da.

C.2.7.5 Spectrum analysis and Databases

The acquired spectra were loaded to the MaxQuant software with the corresponding search engine Andromeda[282,283]. For identification in the Ensembl Human protein database[284] (release 72, 40,047,703 residues, 105,287 sequences) the following search settings were used: charges of +1 to +3 were allowed, length 7-15 amino acids, cleavage unspecific, MS tolerance 10 ppm, MSMS tolerance 0.6 Da, 1% false discovery rate (FDR) on peptide and protein level, methionine oxidation and asparagine or glutamine deamidation were allowed as variable modifications.

C.2.8 Software and Statistics

Band intensities of Coomassie gels, Western-, and Slot Blots were analyzed with the Java-based *ImageJ* (version 1.47, NIH, USA) program[285].

Prediction of peptide binding to HLA molecules were performed at the web-interface (http://www.cbs.dtu.dk/services/NetMHC/) of the artificial neural networks server *NetMHC 4.0*[286,287].

Sequence logos[288], as a more detailed alternative to consensus sequences, were created with the online accessible tool *WebLogo 3.4*[289] (http://weblogo.threeplusone.com/).

GraphPad Prism 5 (GraphPad Software, Inc., La Jolla, USA) was used for most statistical analyses, regression analyses, and 2D graphing. Significance levels α were set to 0.05 (indicated by *), 0.01(**), and 0.001(***). For multiple comparisons Bonferroni corrected significance levels[290] were calculated by dividing α by the number of tested hypotheses.

The *Optimizer Algorithm* (B.2) and all *in silico* antigen evaluation programs (B.3.2 to B.3.4) were implemented in *Java 1.8* (Oracle, Redwood City, USA), using the integrated development environment *Eclipse SDK* (version 4.2.1, Eclipse Foundation, Ottawa, Canada). Python (version 2.6, Python Software Foundation, Delaware, USA) was used for communications with *MODELLER v9.7*.

D Results

D.1 Design of epitope-enriched HIV-1-Gag antigens

The objective of this thesis was to design breadth-optimized Gag antigens, by combining as many potent CD8+ T cell epitopes as possible in maximal three full-length Gag sequences, while also preserving the Gag function to release VLPs. The *Optimizer Algorithm* (B.2) was used to create these antigens using patient-derived epitopes. In this chapter the data retrieved as input for the algorithm (D.1.1), the use of the FLD to preserve Gag structure and function (D.1.2), the ranking of CD8+ T cell epitopes (D.1.3), and finally the generation of novel proof of concept CD8+ T cell epitope-enriched Gag antigens (teeGags) (D.1.4) is described.

D.1.1 Analysis of input data sets

The *Optimizer Algorithm* requires a variety of input data, as described in B.2. In addition to an alignment of natural Gag sequences (D.1.1.1) and CD8+ T cell epitopes (D.1.1.2), also HLA class I allele- (D.1.1.3) and HIV-1 subtype-frequencies (D.1.1.4) in the target population have to be specified. The data used for the proof of concept teeGags, designed for worldwide application, are described below.

D.1.1.1 HIV-1 Gag alignments

As representation of the genetic universe of Gag sequences, the Filtered Web Alignment from the LANL HIV database was downloaded and curated (B.1.1.1). From the originally 5085 sequences, 84 were deleted, due to their HIV-1 group- or subtype-association (Table 10), leaving 5001 sequences in the final alignment.

Table 10. Sequences omitted from the Gag alignment. Number and affiliation of non-group M or unclassified Gag sequences that were removed from the LANL Filtered Web Alignment (B.1.1.1)

HIV-1 group or subtype	Number of deleted sequences
Group N	8
Group O	26
Group P	2
CPZ	16
U (unclassified)	32
Overall	84

D.1.1.2 CD8+ T cell epitopes

CD8+ T cell epitopes were downloaded (B.1.2) from the LANL molecular immunology database on April 24, 2015. In the search mask, the parameters p17, p24, and p2p7p1p6 (this nomenclature for the N-terminal domains p2, p7, p1, and p6 was retained in the *Optimizer Algorithm*) were specified for the HIV proteins, selecting human as species, and keeping the remaining options on default. The 2688 matching records found were downloaded and converted into a CSV file. All database entries were subjected to a manual quality control to remove or correct questionable and erroneous sequences and annotations. Epitope entries with HIV-2-based sequence, unclear sequence (marked by “?” after epitope sequence), a sequence longer than 20 amino acids, and those described as only computer predicted, without any experimental validation, were deleted. Additionally, all epitopes, which could not be found at

least once in any natural Gag sequence (resulting in a conservation score of 0; see B.2.2.4) of the 2013 Filtered Web Alignment (D.1.1.1) or in a subsequent search in the complete, unfiltered LANL sequence set (using the online sequence search interface[g]), were further evaluated by double-checking the primary publication given for the respective database entry. If possible the epitope sequence of the entry was patched, or otherwise deleted. Furthermore, in some cases the epitope annotations, like HLA-association, virus subtype, or epitope start- and end-positions had to be changed. For example all subtype "A1" annotations were changed to "A", because in many cases only "A" is specified, allowing no differentiation between A1 and A2. Overall, 99 database entries were removed and 154 modified. All epitope database modifications performed are summarized in Table 11 and a detailed changelog can be found in Extended Data Table 1.

Table 11. Overview of deleted and modified epitope database entries. The 2688 Gag-specific human epitope entries downloaded from the LANL molecular immunology database were curated for erroneous or improper entries. If feasible, the entries were modified (right column), or otherwise omitted (left column).

Omitted epitope entries		Modified epitope entries	
Reason	**Count**	**Changed field**	**Count**
Unclear sequence ("?")	4	Epitope sequence	37
Sequence >20aa	19	HLA	15
Computer prediction	3	Start-end annotation	12
HIV-2 sequence	31	Subtype	93
Conservation score = 0	39		
Other	3		

The residual 2589 epitope entries were analyzed for entries describing the identical epitope sequence at the same protein position. Such multiple records were merged into one consensus entry as described in B.1.2. This measure reduced the input data to a set of 691 unique epitopes that are listed, together with all relevant database information, in Extended Data Table 2 .

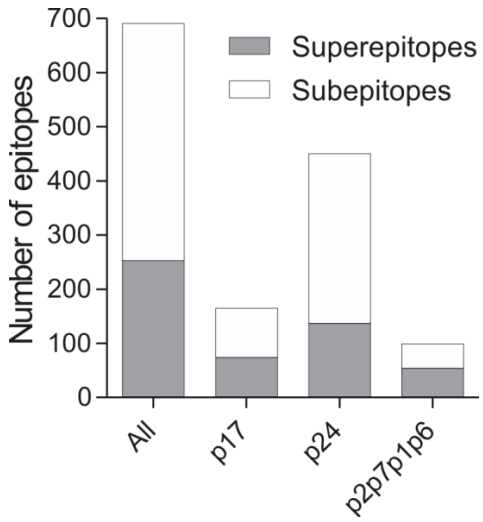


Figure 15. Total number of epitopes, super-, and subepitopes in the different Gag proteins. The graph displays the absolute numbers of superepitopes (grey bars) and subepitopes (white bars) for the complete Gag (All), or the domains p17, p24, and p2p7p1p6. Epitopes located on the interface of two protein domains were counted as epitope for both proteins (e.g. an epitope spanning over the p17-p24 boundary was counted as p17-, as well as p24-located epitope).

[g] http://www.hiv.lanl.gov/components/sequence/HIV/search/search.html

According to B.1.2, the unique entries were partitioned into 253 super- and 438 subepitopes (Figure 15). Most epitopes (427 total epitopes, 125 superepitopes, and 302 subepitopes) were located in the capsid p24 protein, followed by matrix p17 protein (157 total, 70 super-, and 87 subepitopes), and finally, with the least epitopes, the joined N-terminal domains p2p7p1p6 (84 total, 46 super-, and 38 subepitopes). Some epitopes spanned over the defined amino acid boundary of p17 to p24 (8 total, 4 super-, and 4 subepitopes) or p24 to p2p7p1p6 (15 total, 8 super-, and 7 subepitopes). These epitopes were for subsequent analysis counted as epitope for both proteins (Figure 15). Normalized to the protein length, p17, p24, and p2p7p1p6 were represented by 1.3 (0.6 superepitopes and 0.7 subepitopes), 1.9 (0.6 and 1.4), and 0.7 (0.4 and 0.3) total epitopes per amino acid, respectively.

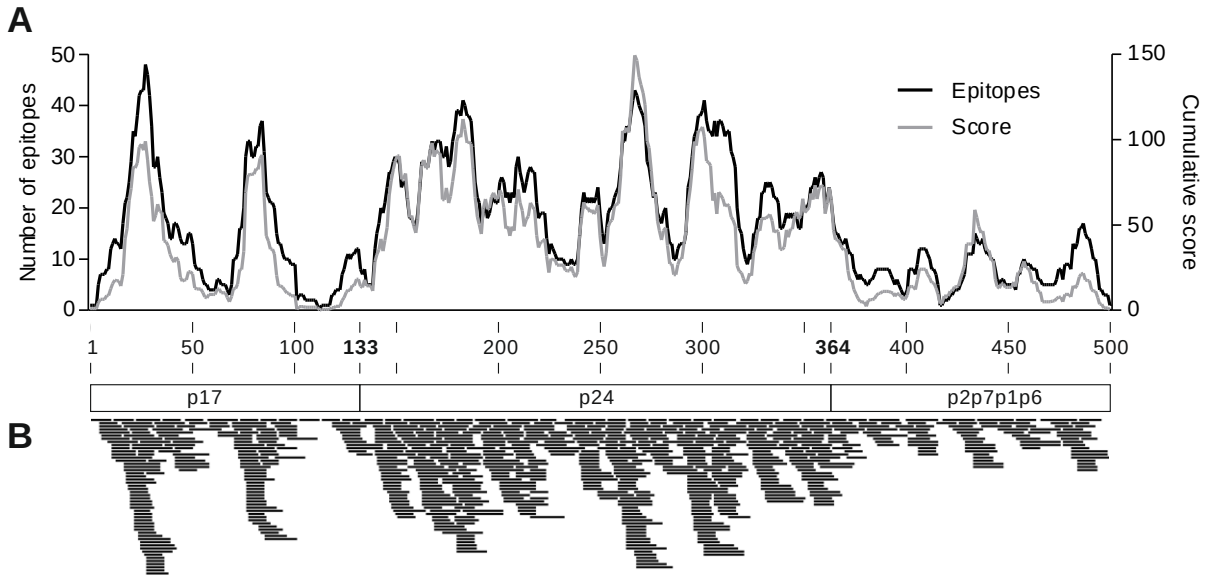


Figure 16. Epitope Distribution across the HXB2-Gag sequence. (A) Coverage of each amino acid position in the epitope set alignment to the HXB2-Gag sequence (back line). The score at each position (grey line) was calculated by adding up the scores (B.2.2) of all epitopes that contain the respective positon. (B) The proportional location of all unique epitopes within HXB2-Gag that was divided into p17, p24, and the joined N-terminal domains p2p7p1p6. Each line represents one epitope at its exact location drawn to scale.

The complete set of unique epitopes was aligned to the HXB2-Gag reference sequence (Figure 16). The distribution of epitopes and hence the amino acid position coverage was not uniform. The p17 matrix protein, for examples, comprises two epitope “hotspots” around HXB2-Gag numbering positions 28 and 85, with over 30 epitopes covering these locations. Between these hotspots the coverage declined rapidly and position 113 was even covered by none of the 691 epitopes. For the capsid protein, the epitope distribution was more equal, but also here the coverage ranges from 5 to 43 epitopes per amino acid position. The epitope distribution for p2p7p1p6 is on a similar, low level for the complete, jointed protein length.

To identify in the epitopes variable and conserved Gag positions, all 691 unique epitopes were aligned with the reference HXB2-Gag sequence. Most Gag amino acid positions (372) were conserved and epitopes harbored no mutations compared to the reference. However 128 variable positions were identified in the epitope-to-reference-alignment. In most cases (102) only one amino acid substitution was observed at the variable positions, but 22 times 2 different and 4 times 3 different amino acids were identified. Summed up, 158 unique AAS were found in the epitopes compared to the reference sequence. Each unique epitope contained on average 0.74

AAS, with a range from 0 to 10 (Figure 17). However, most epitopes (380 epitopes, 55% of complete set) were identical to the HXB2-Gag reference sequence. Epitopes with mutations as compared to the reference mostly only harbored one (188, 27%) or two (83, 12%) AAS. Epitopes with many mutations are mostly due to amino acid insertions or deletions, which prevent a gap-free alignment to HXB2.

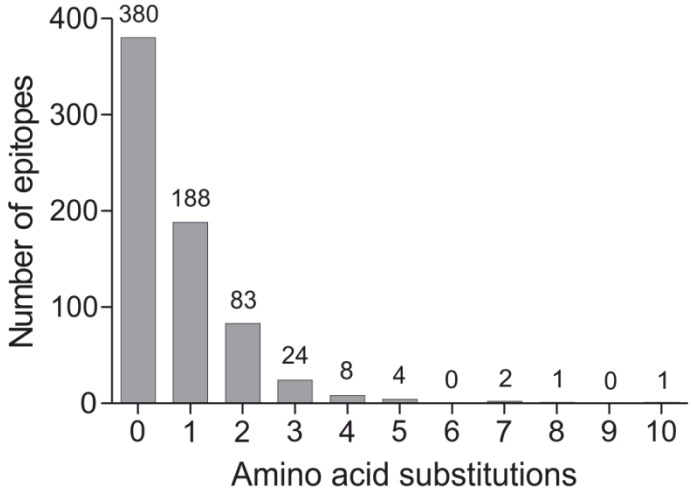


Figure 17. Number of AAS per epitopes identified in an alignment with the reference HXB2-Gag sequence.

D.1.1.3 HLA allele frequencies

For a proof of concept first generation of the newly-designed teeGags, it was decided to aim for a global vaccine approach. Due to this, information about the worldwide HLA class I allele frequency distribution was required for epitope scoring (B.2.2.1) and population score calculations (B.3.3).

The most comprehensive publication, summarizing broad global patterns of allelic differentiation was published by Solberg et al.[291]. Therein, the HLA allele frequencies of 497 population studies and approximately 66,800 individuals from throughout the world were reviewed in a meta-analysis. The overall average HLA allele frequencies for loci A, B, and C, across all population samples (data available in the online supporting material[h]) were used in this thesis as representation of the worldwide allele distribution. Since some primary references of the LANL epitope entries describe the HLA alleles only in a low resolution (e.g. A*02 instead of A*02:01 or A*02:02), the high resolution published by Solberg et al. had to be adapted. Hence, all specific HLA protein frequencies were summed up in their corresponding allele group frequency (e.g. A*02 consists of the cumulative frequencies of A*02:01, A*02:02,…). The complete list of the calculated HLA class I allele group frequencies is given in the Extended Data Table 3.

D.1.1.4 HIV-1 subtype weighting

For the immunological scoring of epitopes (B.2.2.2), as well as for the validation of antigens (pathogen coverage B.3.4), a ranking of the circulating HIV-1 subtypes was required. Since a holistic approach should be implemented for the first generation teeGags, the subtype weighting was oriented at the global distribution (Figure 18), as published by Hemelaar et al.[8]. Therein, country-specific epidemiology data from 2004-2007 were combined with estimated numbers of

[h] www.pypop.org/popdata/

HIV-infected people in the respective countries. The weighting of subtypes and circulating recombinant forms (CRFs) was chosen directly proportional to the described global frequencies (Table 12). Rare CRFs and unique recombinant forms (URFs) were assigned a weight of 0. For the scoring, this had no distinct impact, since except one primary reference, reporting epitopes derived from B/C recombinant viruses[292], none of the LANL entries were associated with any of the rare CRFs or URFs.

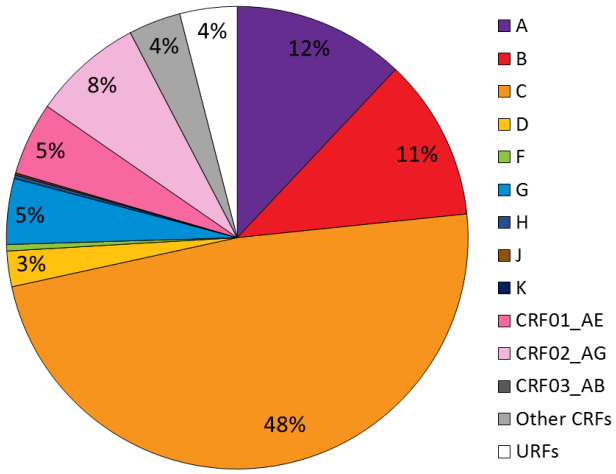


Figure 18. Global HIV-1 subtype, CRF, and URF frequencies as described by Hemelaar et al.[8]. In the publication rare CRFs (Other CRFs) and unique recombinant forms (URFs) were summed up to one entry each.

Table 12. Subtype and recombinant weighting proportional to the global frequency. All rare CRFs and URFs were assigned a weight of 0.

Subtype or recombinant	Global frequency [%]	Weight
A	12.03	1203
B	11.33	1133
C	48.23	4823
D	2.49	249
F	0.45	45
G	4.60	460
H	0.26	26
J	0.12	12
K	0.01	1
CRF01_AE	5.09	509
CRF02_AG	7.73	773
CRF03_AB	0.00	0
Other CRFs	3.65	0
URFs	4.01	0

D.1.2 Functional assessment of AAS

As a main feature of the *Optimizer Algorithm*, all AAS with negative influence on the release of virus-like particles should be excluded for teeGag design. To determine, if an AAS is compatible the budding of VLPs or detrimental to its function, a computational multidimensional classifier was implemented. This method employs combinations of structural and sequence-based properties (D.1.2.1), to classify unknown AAS (D.1.2.2 and D.1.2.3), based on a training-set of mutations with experimentally determined phenotype and is in detail described in B.2.1.

D.1.2.1 Classifier feature selection

To assess the effect of AAS on VLP budding, structural features that interpret energetic landscape alterations of protein structure models, introduced by the AAS, were defined (B.2.1.1). For this, 3D-structures and the energy profile as DOPE (B.2.1.1, section “Homology modeling and Discrete Optimized Protein Energy”) for the p17-, p24-, p2-, p7-, and p6-protein of the reference HXB2-Gag, as well as for Gag-domains with integrated single AAS, were calculated using homology modelling, except for p1, because no suitable template could be identified.

Four different DOPE-based energetic observations were tested as classification feature, namely (I) the DOPE at the AAS position ($f_{1,DOPE}$), (II) DOPE difference between the reference and the mutated structure at the AAS position ($f_{2,\Delta DOPE}$), (III) the sum of all DOPE difference values between the reference and mutated 3D-model at all amino acids of the Gag protein where the

AAS was located ($f_{3,\Delta DOPE\text{-}global}$), and (IV) the sum of DOPE value differences between reference and mutated structure for all amino acids in a window of six in N- and five in C-terminal direction around the AAS ($f_{4,\Delta DOPE\text{-}local}$). A fifth feature ($f_{5,seq}$) originated from sequence-specific characteristics and is a measure of sequence conservation at the respective positon, displayed as PSSM score (B.2.1.2). To assess the performance of the five features, to discriminate between budding-deleterious and budding-retaining mutations, a training-set with known phenotypes was analyzed (Table 4). Five entries of this set were removed beforehand: the G2A mutations, since the negative effect is not based on structural properties, but on inhibition of a posttranslational myristoylation and four additional AAS (A5D, S6I, V128E, and Q130G), which are positioned too close to the beginning or end of the sequence, prohibiting $f_{4,\Delta DOPE\text{-}local}$ calculations. Statistical analysis of the budding-deleterious (-) and budding-retaining (+) groups of the training-set with the nonparametric Mann-Whitney test revealed significant group separation for all features, except $f_{1,DOPE}$ (Figure 19). On its own $f_{1,DOPE}$ therefore allows no discrimination between the two classes. Better, but with a p-value of 0.031 the second-worst, was the PSSM score displayed by $f_{5,seq}$. The best features concerning significance were $f_{2,\Delta DOPE}$, $f_{3,\Delta DOPE\text{-}global}$, and $f_{4,\Delta DOPE\text{-}local}$. All three rely on energetic comparison between mutated and reference structures.

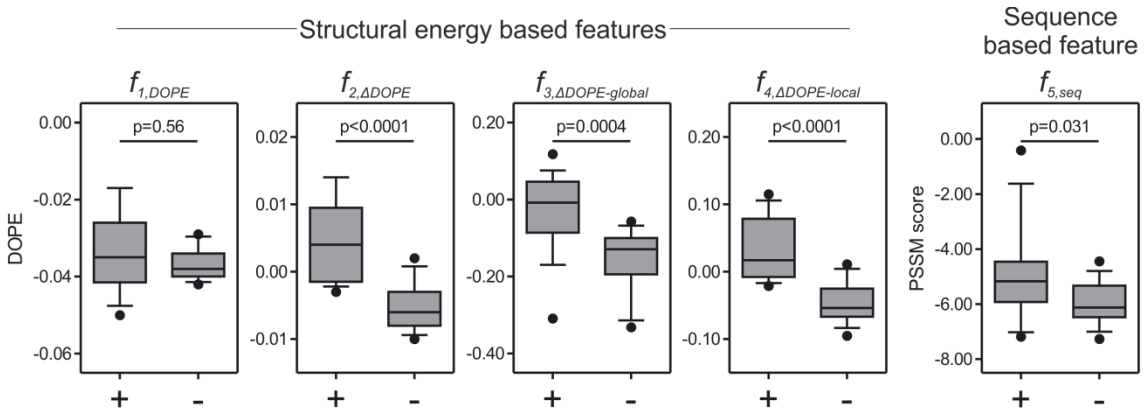


Figure 19. Group separation of training-set data for all five classification features. AAS with experimentally determined phenotype were classified in budding-retaining (marked "+") and budding-deleterious ("-") mutations and their respective feature scores are depicted. p-values of group separation of the four structural-energy- and the sequence-based features were computed with the non-parametric Mann-Whitney test. Boxes represent the median and 50 % quartiles, whiskers the 10-90 % percentile, and black dots the outliers.

However, even for the features with highly significant group separation, there was still substantial overlap between AAS with different effects on budding. To improve the differentiation, it was tested to base predictions on combinations of selected features. The training-set values of feature combinations were used in a Fisher's Linear Discriminant Analysis (B.2.1.3). During the FLD, a vector *d* was calculated that projects the multidimensional feature data combinations onto a 1D-subspace *t* (Figure 20). The reduction onto the 1D-space was done in a way that best separates the two classes (i.e. budding-retaining and budding-deleterious AAS) of the training-set data. To discriminate between the groups, a hyperplane *Z* between projections of the means of both classes was computed. This allowed sorting of AAS with unknown function into the budding-retaining or budding-deleterious group, solely based on its structural and sequence-based features.

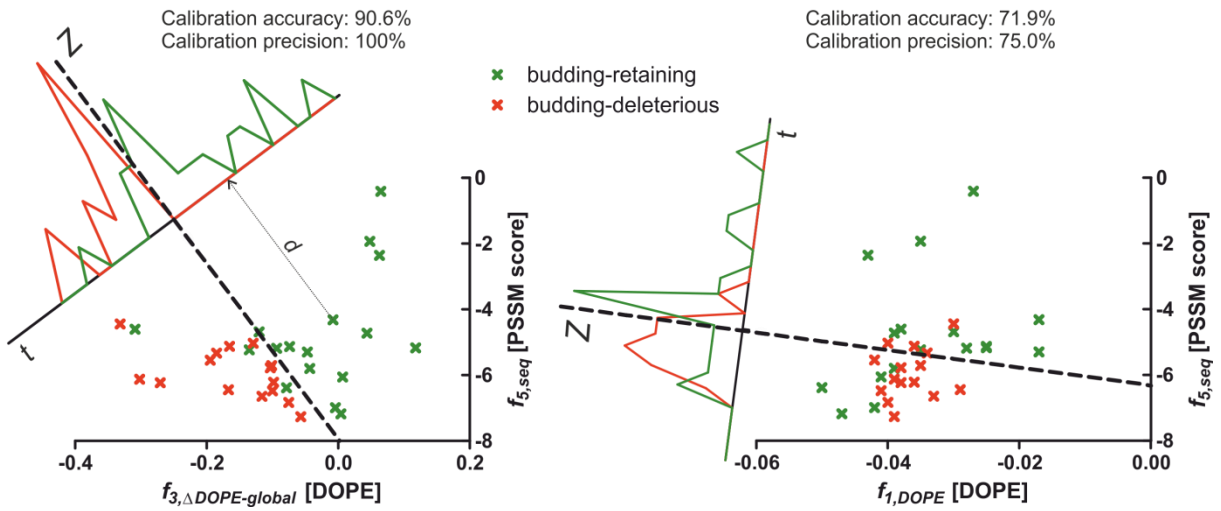


Figure 20. Fisher's Linear Discriminant Analysis of the training-set data. The experimentally validated training-set entries were classified into the budding-retaining (green) or budding-deleterious (red) group and all possible combinations of discriminatory features were analyzed in a FLD. Combining $f_{3,\Delta DOPE\text{-}global}$ and $f_{5,seq}$ (left graph) resulted in the highest 10-fold cross-validation (100 repeats) precision (98%, Extended Data Table 4) and had a calibration precision of 100%. The worst double combination was $f_{1,DOPE}$ with $f_{5,seq}$ (right graph). Both depicted graphs show all training-set entries, located according to their respective feature data. In both cases, all AASs were projected with the vector *d* onto the 1D-space *t*. Frequency distribution above *t* indicates the achieved group separation. Hyperplane Z represents the actual computed group discrimination rule.

To identify the best discriminatory combination for the classifier, all possible feature permutations were tested using the training-set in a 10-fold-crossvalidation (B.2.1.3) with 100 repeats each. To rank all possible combinations, the accuracy, sensitivity, specificity, PPV (precision), and NPV (Table 5) of the binary classification predictions were computed. As preservation of budding-competence was considered mandatory, for the design of the T cell epitope-enriched Gag antigens, it was decided on a high PPV as most important quality characteristic for the classifier. This value is calculated as percentage of the true positives among all positives, i.e. as budding-retaining predicted AAS. A high PPV engenders only few false positives, which had to be avoided, since incorporation of a single false positive AAS could destroy the whole Gag functionality.

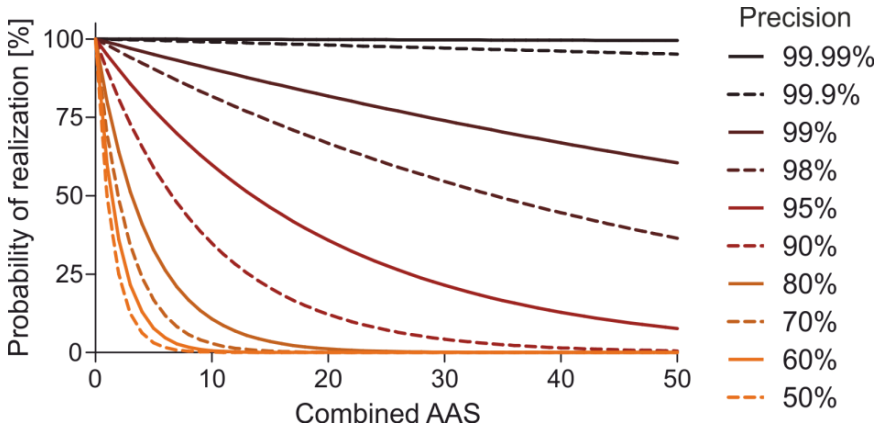


Figure 21. Probability of realization. Depending on the precision of classifying budding-retaining AAS, the probability of including only true positive predictions, therefore resulting in functional Gag, diminishes in different rates with increasing numbers of AAS combined in the same sequence (x-axis). The graph shows theses probabilities (y-axis) of realization based on various precision.

The importance of a high precision is exemplified by the probability of realization (Figure 21). This is the probability of being able to combine *x* different AAS in a single Gag protein, without incorporation of a single false-positive predicted AAS, which probably would inhibit VLP production. It is calculated as the precision to the power of *x*. A high precision of 99% for example, would mean, that only 1 of 100 budding-retaining classifications would be a false positive. Because of this many mutations can be combined, without risk of including false positives that potentially would inhibit viral release (e.g. combination of 30 AAS would still have a realization probability of 74%). For lower precisions the realization quickly diminishes. Combining 30 AAS at a precision of 80%, for example, would only have a realization probability of 12%.

The complete list of feature permutations sorted first by PPV and then accuracy is summarized in Extended Data Table 4. The top position was occupied by the combination of feature $f_{3,\Delta DOPE\text{-}global}$ and $f_{5,seq}$ with a precision of 98% and an accuracy of 86%. The calibration accuracy (90.6%) and calibration precision (100%), calculated on basis of the actual classification of all AAS, within the entire training-set, were even higher (Figure 20, therein compared to the worst combination of two features - $f_{1,DOPE}$ and $f_{5,seq}$). The worst discrimination was determined for $f_{1,DOPE}$ alone, with an accuracy of only 58% (precision: 64%), which means that it is just slightly better than a classification by chance. This was however expected, since this feature was the only one that showed no significant separation of the two groups (Figure 19). This entails also that almost every time when $f_{1,DOPE}$ was included, the classification result was worse than for the same feature combination without $f_{1,DOPE}$ (e.g. $f_{1,DOPE}$+ $f_{2,\Delta DOPE}$ + $f_{3,\Delta DOPE\text{-}global}$ had an accuracy of 71% and a precision 83%, but without $f_{1,DOPE}$ the values increased to 75% and 85%, respectively). Vice versa, inclusion of $f_{5,seq}$, despite being the second-worst single feature (accuracy: 72%, precision: 76%), increased the accuracy and precision of the prediction almost in every combination, compared to the same feature set without $f_{5,seq}$. Although $f_{4,\Delta DOPE\text{-}local}$ was the best single feature (accuracy: 80%, precision: 86%), no combination with any other one had a beneficial effect on the classification.

D.1.2.2 Classification of unknown AAS from epitope set

After determining the optimal combination of classification features ($f_{3,\Delta DOPE\text{-}global}$ and $f_{5,seq}$), the group affiliation of all unique AAS, identified in the Gag CD8+ T cell epitope set, was examined. This fully implemented classification is illustrated in Figure 22. All AAS with unknown effect were plotted according to their $f_{3,\Delta DOPE\text{-}global}$ and $f_{5,seq}$ values. Appling decision boundary *Z*, calculated before by reference to the training-set (Figure 20), the AAS were grouped into budding-retaining and budding-deleterious mutations (Figure 22).

The 158 AAS identified in the epitope input data set were distributed quite homogenously across p17, p24, and p2p7p1p6 with 48, 60, and 50 occurrences, respectively (Figure 23). Normalized to the domain length, p24 has the fewest with one AAS every 3.85 amino acids, as compared to 2.75 and 2.74 for p17 and p2p7p1p6. Of all AAS identified in the epitope input dataset, 18 (11.4%) were classified as budding deleterious. For one AAS (Y441H) no classification was possible, because it was located in the p1-domain, of which no suitable template structure for homology modelling was available. To prevent incorporation of this unclassified mutation, it was precautionally classified as budding-deleterious.

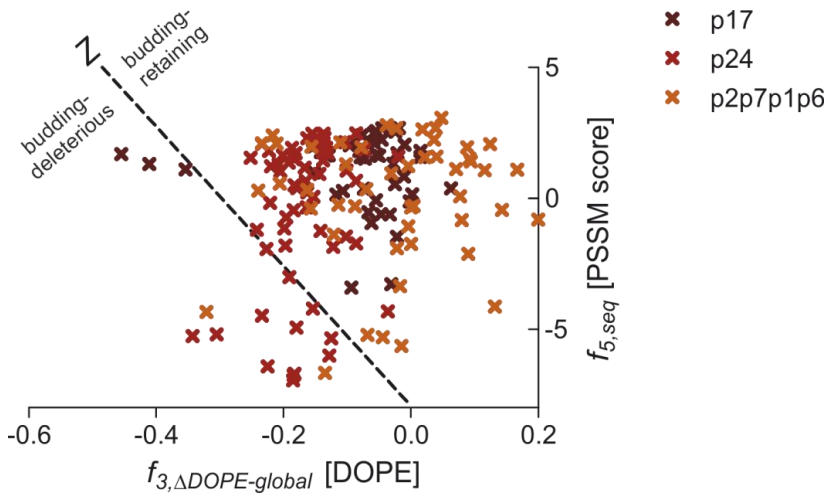


Figure 22. Classification of unknown natural occurring AAS into budding-deleterious and budding-retaining mutations. All 158 AAS, identified from the alignment of all unique CD8+ T cell epitope to the HXB2-Gag reference sequence, were plotted according to their respective $f_{3,\Delta DOPE\text{-}global}$ and $f_{5,seq}$ values. Using the hyperplane *Z*, calculated on the basis of the training-set, the unknown AAS were classified either in the budding-deleterious or the budding-retaining class.

Most AAS with negative effect on budding were located in p24 (13 = 21.7% of all AAS in the protein domain) with a hotspot between positions 267 to 273. In this area, 7 as budding-deleterious predicted AAS accumulated (E.2.1). For p17, only 3 (6.3%), and for p2p7p1p6, 2 (4% - with p1 3 (6%)) mutations were classified as functionally detrimental, respectively. From the 691 unique epitopes in the input data, 21 entries harbor at least one of the 19 (including the p1 located mutation) as budding-deleterious classified AAS, and were therefore removed for all subsequent processes. Thereby, 670 unique epitopes remained for the design of the T cell epitope-enriched Gag antigens.

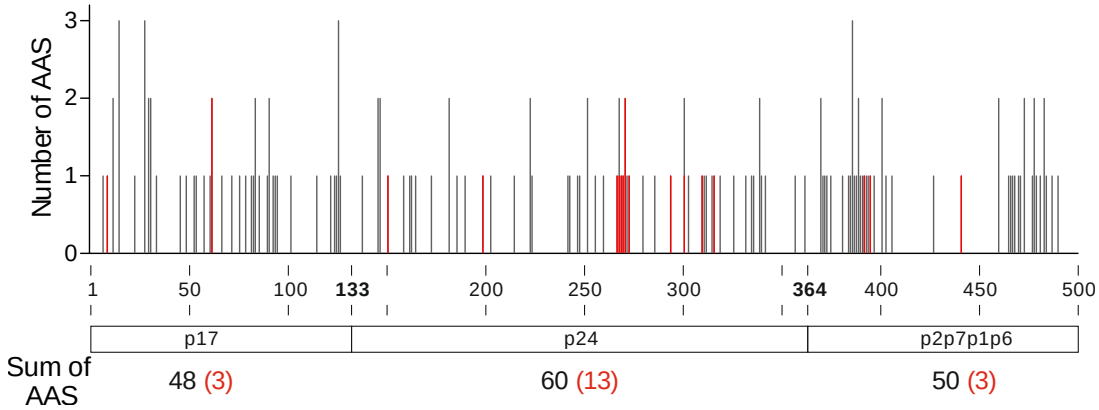


Figure 23. Distribution of naturally occurring AAS per position of HXB2 reference sequence. Gag was divided in p17, p24, and p2p7p1p6 protein domains. AAS classified as budding-deleterious are highlighted in red. The respective sum of all identified AAS and the number budding-negative grouped mutations (in red) is stated below each protein domain.

D.1.2.3 Classification of all possible Gag AAS

Besides classifying all naturally occurring Gag AAS, all theoretically possible mutations were analyzed. For this, each amino acid position of the reference HXB2-Gag was permutated to all other 19 amino acids. For all these sequences, containing a single AAS each, a 3D-model was

generated using homology modelling. Afterwards, as above, the decision boundary *Z,* generated via the FLD and the features $f_{3,\Delta DOPE\text{-}global}$ and $f_{5,seq}$ of the training-set, was applied to classify all possible 9,196 AAS in p17, p24, and p2p7(p1)p6. Again, all amino acids located in p1 were excluded from this permutation analysis, due to the lack of a template structure for homology modeling.

The three defined protein domains of Gag were examined separately (Figure 24). The relatively conserved p24 had the highest rate of budding-deleterious predicted AAS (83.6% of all 4,389 in p24 located AAS). This is mainly due to the fact that the $f_{3,\Delta DOPE\text{-}global}$ values shifted into a more negative range and hence across the decision boundary *Z* (Figure 24 middle panel). For p17 and p2p7(p1)p6, in contrast, most AAS (66.9% of 2,508 and 85.3% of 2,299) were classified to have no negative influence on Gag functionality. For the complete Gag sequence, 4,839 AAS (52.6%) were predicted as budding-deleterious.

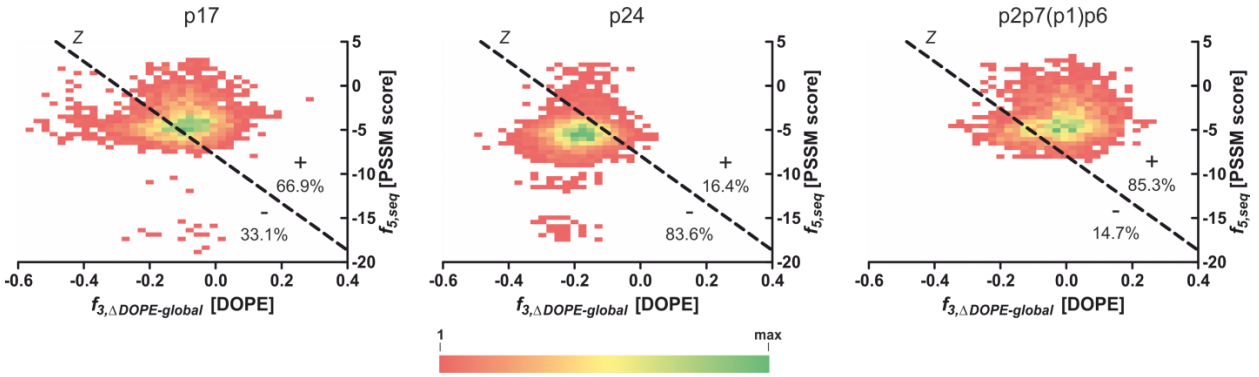


Figure 24. FDA of all possible permuted Gag AAS. The 9,196 AAS, generated by per-muting each amino acid position of HXB2-Gag separately to all other 19 possible amino acids, were partitioned into p17, p24, and p2p7(p1)p6. Each AAS is plotted according to their $f_{3,\Delta DOPE\text{-}global}$ and $f_{5,seq}$ values and classified using the hyperplane *Z*. The three depicted graphs show the location of all AAS as heatmap. For the generation of the heatmap, data points were condensed with a bin width of 0.5 for the PSSM score and a bin width of 0.025 for the DOPE value. The percentage of budding-retaining (indicated by "+") and budding-deleterious ("-") AAS is given for each protein domain.

Counting the number of as budding-retaining or as budding-deleterious predicted AAS for each Gag amino acid position, highlights the difference between the protein domains (Figure 25). The data depicts clearly that the rate of as budding-deleterious predicted AAS is far higher for p24 compared to p17 or p2p7(p1)p6. As a measure of variation, the Shannon entropy at each position was calculated using the LANL web-tool "Entropy-one"[i]. Shannon entropy is a quantitative measure of uncertainty at an alignment position, i.e. it describes how heterogeneous the position is in a set of defined sequences. As sample input, the curated filtered web alignment was used. The computed values illustrate the relative variation at the different amino acid positions. A high Shannon entropy implies therefore a high variability at the respective position.

The entropy values correlate significantly with the number of as budding-retaining predicted AAS (Figure 25) with a Pearson product-moment correlation coefficient of r=0.56 and a p-value <0.0001 (data not shown). This indicates that variable positions (i.e. high entropy) tolerate more mutations, without destroying the Gag functionality. For conserved positions (i.e. low entropy), on the other hand, the probability that a mutation leads to a non-functional Gag is elevated. This suggests, that the conserved positions have important function during assembles and release of VLPs.

[i] http://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_one.html

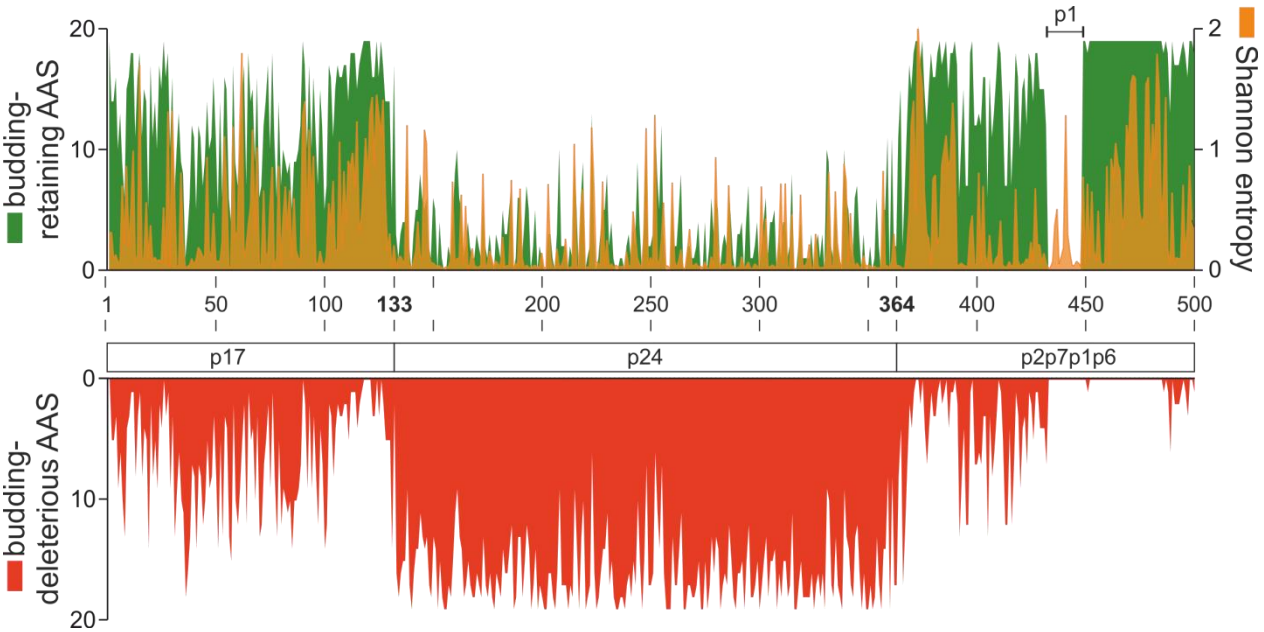


Figure 25. Amino-acid-position-specific distribution of budding-retaining and budding-deleterious AAS. For each amino acid position of the reference HXB2-Gag (partitioned in p17, p24, and p2p7p1p6), all 19 permutations were classified. The numbers of budding-retaining AAS are depicted in the upper graph (green) and the budding-deleterious AAS in the lower graph (red). p1 was excluded from the analysis, because of a missing template for homology modelling. Additionally, the position-specific Shannon entropy, as a measure of the sequence variability, is plotted in the upper graph (orange).

D.1.3 Epitope score evaluation

After the functional assessment of all naturally occurring AAS, and the subsequent removal of all epitopes with at least one as budding-deleterious predicted AAS, the remaining epitopes were ranked according to their immunological, demographical and virological characteristics. Each epitope *e* was assigned a specific score *s(e)*, which was calculated based on multiple, differently weighted attributes *a*.

Table 13. Attribute weighting for epitope score *s(e)* calculations

Attribute *a*	Weighting w_a
a_{HLA}	6
$a_{Subtype}$	3
a_{LTNP}	6
a_{Cons}	3
$a_{\%Resp}$	1

For our proof of concept T cell epitope-enriched antigens, (i) the frequencies of class I HLA molecules presenting the epitope (a_{HLA}), (ii) the subtype affiliation ($a_{Subtype}$), (iii) the association with LTNPs (a_{LTNP}), (iv) the conservation status (a_{Cons}), and (v) the expected population-wide immune response to the respective epitope ($a_{\%Resp}$) were used as scoring attributes. The influence of each attribute *a* on the s(e) was further specified with a weighting parameter w_a (Table 13). For score calculations, all attribute data of subepitopes were added to their associated superepitopes, as described in the methods (B.1.2).

As one of the most important features the presentation of epitopes on frequent HLA class I alleles was defined, thus it was assigned a high weight (w_a=6). The HLA score a_{HLA} was calculated as the sum of frequencies of all defined HLA allele groups able to present the epitope in question (B.2.2.1). Of the 72 annotated HLA allele groups (Extended Data Table 3), 49 were specified by at least one epitope entry of the complete set. In detail, 14 of all 22 HLA-A allele groups, 22 of the 36 HLA-B allele groups, and 13 of the 14 HLA-C allele groups known today

were covered. Adding up the frequencies of all groups represented at least once in the epitope set, 94.9%, 88.2%, and 95.6% of the HLA-A,-B, and -C diversity were covered, respectively. Most epitopes (205; 29.7%) were assigned to exactly one allele group. However, 364 (52.7%) epitopes were associated with at least two different HLA allele groups with a maximum of 15 different groups (Figure 26 A). For the residual 122 (17.7%), the epitope entry annotation was incomplete and the presenting HLA allele was not stated. The number of epitopes associated with HLA allele groups correlates significantly with the global frequency of the respective group ($r=0.61$, $p<0.0001$ Figure 26 B). This can be seen as indication that the epitope set is representative of the target population (i.e., here, the worldwide population, due to the HLA frequencies used - D.1.1.3). In a frequency distribution of HLA scores a_{HLA} (normalized to the interval [0, 1] - bin width 0.05; Figure 26 C) the highest peak is around 0 due to those epitopes without an annotated HLA allele group (Figure 26 A). Most other epitopes have an HLA score between 0 and 0.25. a_{HLA} higher than 0.5 was observed rarely (only in 9.4% of all epitopes).

As second feature, with a lower attribute weight ($w_a=3$), the affiliation with different HIV-1 subtypes was applied to derive the subtype score $a_{Subtype}$. It was calculated by summing up all frequency-weighted (D.1.1.4) subtypes that are reported to be associated with the particular epitope, and finally normalizing to the interval [0, 1]. Due to the fact that, if the clade affiliation was not specified in the LANL epitope entry, the standard subtype B was allocated, every epitope had at least one subtype listed (Figure 26 E). Partly because of this correction, most epitopes (37.6%) were assigned to exactly one subtype. However, there are various entries with more than one subtype (up to 9 of the overall 13 subtypes stated in the epitope set). Although the worldwide frequency of subtypes (Figure 18) correlates significantly with the respective amounts of epitopes affiliated with it ($r=0.56$, $p=0.039$; data not shown), there are some major imbalances. Most prominently the subtype B is largely overrepresented in the epitope set, whereas the globally by far most frequent subtype C (48%) is only on the third position (Figure 26 D). Due to the high number of epitopes, which are only B-clade-associated, the highest peak in the frequency distribution (Figure 26 F) of the subtype score clusters at a low value (~0.12). The score of the remaining epitopes was distributed quite homogenously between 0 and 1.

To estimate the fraction of the target population that is expected to prime a T cell response against a specific epitope, the expected immune response $a_{\%Resp}$ was computed (described in B.2.2.5). For 53.7% (371 of 691) of the epitopes this value was denoted, partly due to the data inheritance from sub- to superepitopes. Since the response is normalized to the associated HLA allele group frequency, most values were below 0.1 (Figure 26 G). Because this feature is derived from the epitope entry metadata and was not available for a large proportion of the entries, a low weight for was assigned to $a_{\%Resp}$ ($w_a=1$).

Association with LTNP-status was defined as Boolean value. Hence, the high-weighted ($w_a=6$) LTNP score a_{LTNP} could only attain 1, if defined as connected to an LTNP, or no value otherwise. In a distribution of LTNP-associated epitopes across the Gag protein, six different peaks can be observed (Figure 26 H). Most (5) of those are located in the capsid protein region. The importance of p24 in LTNPs was expected, since it is the most conserved domain of Gag and mutations in conserved regions are associated with high loss of viral fitness[58]. Therefore, CD8+ T cells addressing these vulnerable sites can confer long-lasting viral control, since rapid escape of the virus is less likely.

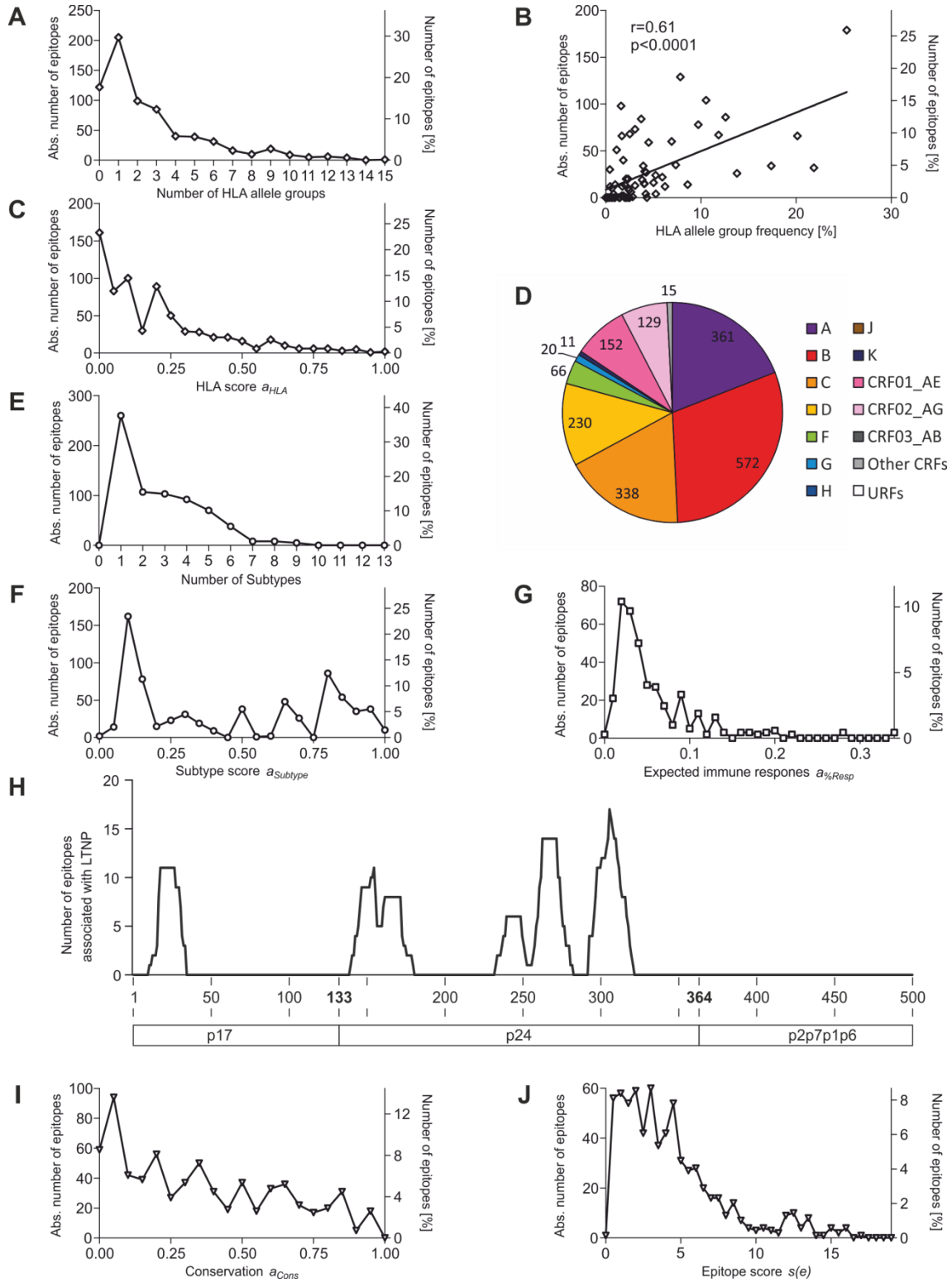


Figure 26. Attributes and epitope scores. (A) Number of HLA allele groups registered per epitope. (B) Correlation between the numbers of epitopes specified to be presented by an HLA allele group and the global frequency of the respective HLA allele group. (C) HLA score a_{HLA} frequency distribution (bin width 0.05) of all epitopes. (D) Pie chart of the number of epitopes assigned to each subtype or CRF. (E) Number of subtypes listed per epitope. (F) Subtype score $a_{Subtype}$ frequency distribution (bin width 0.05) of all epitopes. (G) Expected immune response $a_{\%Resp}$ frequency distribution (bin width 0.01) of all epitopes. (H) Location of epitopes registered as associated with LTNPs. (I) Conservation score a_{Cons} frequency distribution (bin width 0.05) of all epitopes. (J) Frequency distribution (bin width 0.5) of all calculated epitope scores *s(e)*.

As last scoring attribute a_{Cons} (w_a=3), the conservation of an epitope was calculated as the proportion of sequences from the filtered web alignment that include the epitope, while also accounting for the global frequency of the subtypes (B.2.2.4). The computed conservation scores were spread quite homogeneously between 0 (i.e. epitope found in no sequence) and 1 (i.e. epitope found in every sequence) (Figure 26 I).

By virtue of their design, each scoring attribute attained a value between 0 and 1, before the calculation of the epitope score *s(e)*. To compute s(e), the five different weighed attributes were summed up according to the scoring function (B.2.2). Due to the pre-defined weighting of the attributes (Table 13), the theoretical maximum score of an epitope was 19. The highest-scoring epitopes with a value 17.1 was fairly close to the maximum. Most epitopes however cluster between a value of 1 to 5 (Figure 26 J). Calculating the score at each Gag position (Figure 16 A), by adding up the scores of all epitopes that are located at the respective position, matched the absolute number of epitopes significantly (r=0.94, p<0.0001; data not shown).

D.1.4 Generation of T cell epitope-enriched Gag antigens

In the final step of the design algorithm for the teeGags, as many high scoring epitopes as possible were combined into a new sequence, aiming for the highest possible overall score. For this, incompatibilities between overlapping sequences were identified and a genetic algorithm (B.2.3) was applied to find the set of compatible epitopes that best satisfy the fitness function *f(x)*.

Multiple iterations of this antigen generation were performed and after each of these “optimization” rounds, all epitopes included in the solution antigen of the current round were excluded from the input data. By this, the next iteration had to incorporate new epitopes, resulting in a complementary, polyvalent set of teeGags. By virtue of their design, teeGags of higher iterations should never be employed alone, but only in combination with all the teeGags generated in previous rounds. For instance teeGag1, as the first antigen generated, was conceived so that it can be applied alone, however, teeGag2 was designed solely to improve teeGag1, and therefore may only be administered in combination with teeGag1.

After 11 iterations, all of the 670 epitopes that were predicted to have no negative effect on Gag functionality were incorporated into teeGags. However, already after three iterations (teeGag1, teeGag2, and teeGag3; amino acid sequences are displayed in Extended Data Sequences F.2.3.1) 87.8% of all epitopes and 91.0% of the maximal possible score were covered, compared to a mean 58.4% and 59.5% in random selections (100,000 repeats) of three natural Gag sequences from the filtered web alignment (Figure 27 C+D). Any further generated teeGag only contributed marginally to the coverage of epitopes and score.

This is further clarified by the fact that the number of AAS compared to the reference rapidly diminished after iteration 3 (Figure 27 A). In the last two rounds, only one AAS and one new epitope were added. Nonetheless, any combination of iterative complementary teeGags was far superior to any same-sized selection of randomly picked natural sequences (Figure 27 C+D). The fact that in each iteration the highest scoring epitopes were included can be seen by the declining mean score of the epitopes in the antigens (B.3.2, Figure 27 B).

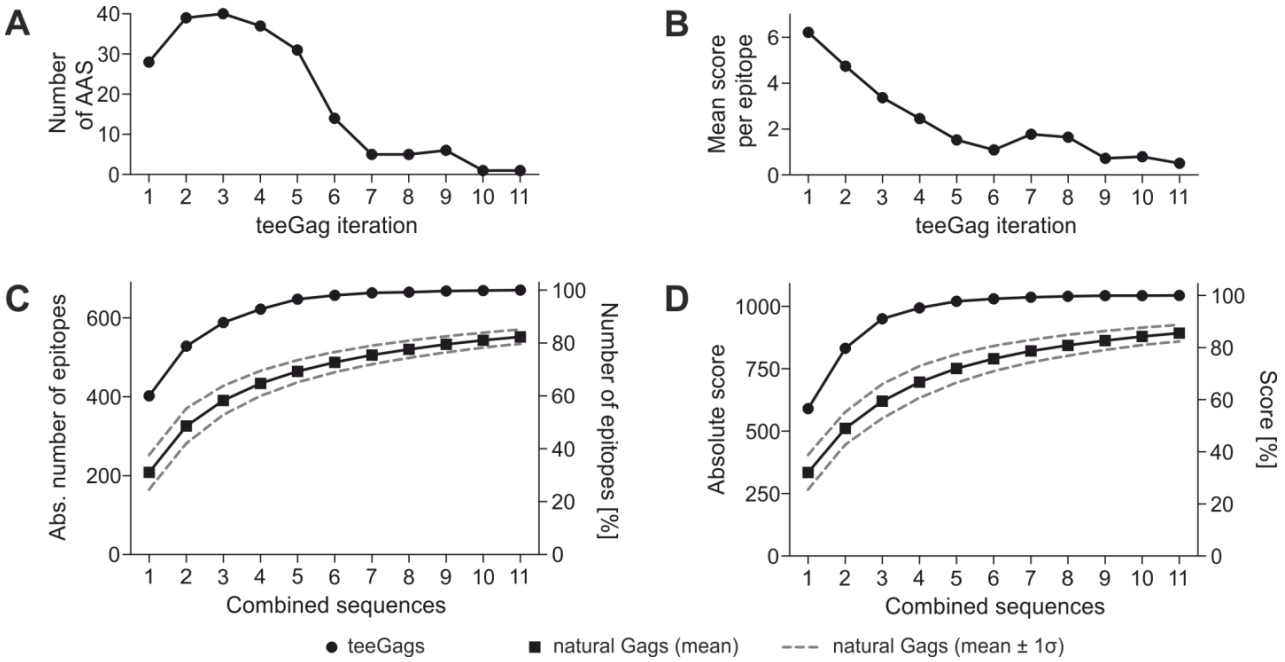


Figure 27. teeGag generation. (A) Number of AAS compared to the HXB2-Gag reference sequence in every teeGag iteration. (B) Mean score of every epitope that was used to calculate the score in every teeGag iteration. (C+D) Progress of total epitope (C) and score (D) coverage of complementary teeGags and random combinations of natural sequences from the filtered web alignment. The iteratively generated teeGags were combined as indicated on the x-axis and the overall number of epitopes or score are displayed. For comparison, the corresponding values for all natural 5001 Gag sequences from the filtered web alignment (round 1) or 100,000 combinations of 2 to 11 randomly selected natural sequences are shown as mean with 1-σ-interval.

D.2 *In silico* analysis of optimized Gag antigens

After the design of the teeGags, the newly conceived candidate antigens were evaluated in various *in silico* regarding virological properties (D.2.1) and enhanced immunological breadth (D.2.2 - D.2.5). The setup of the computational tools used is described in detail in B.3.

D.2.1 Phylogenetic classification of teeGags

For phylogenetic tree reconstruction (B.3.1), teeGag1-3, conM, ancM, cotM, and mosM1.1 were aligned to a curated set of 45 carefully selected Gag reference sequences[236]. In this reference set, each HIV-1 group M subtype, as well as CRF01_AE and CRF02_AG were represented by 2 to 4 sequences. Location of the antigens in the final phylogenetic tree (Figure 28) gives some indication of the subtype affiliation. As expected due to the underlying design principle, conM and ancM were located close to the center of the phylogenetic tree. However, cotM and mosM1.1 that were also expected near the center of the unrooted tree, clustered with the subtype B and C reference sequences, respectively. All analyzed teeGag sequences were located near clade B, but whereas teeGag1 was in the midst of the subtype B sequences, teeGag2+3 were genetically more distant.

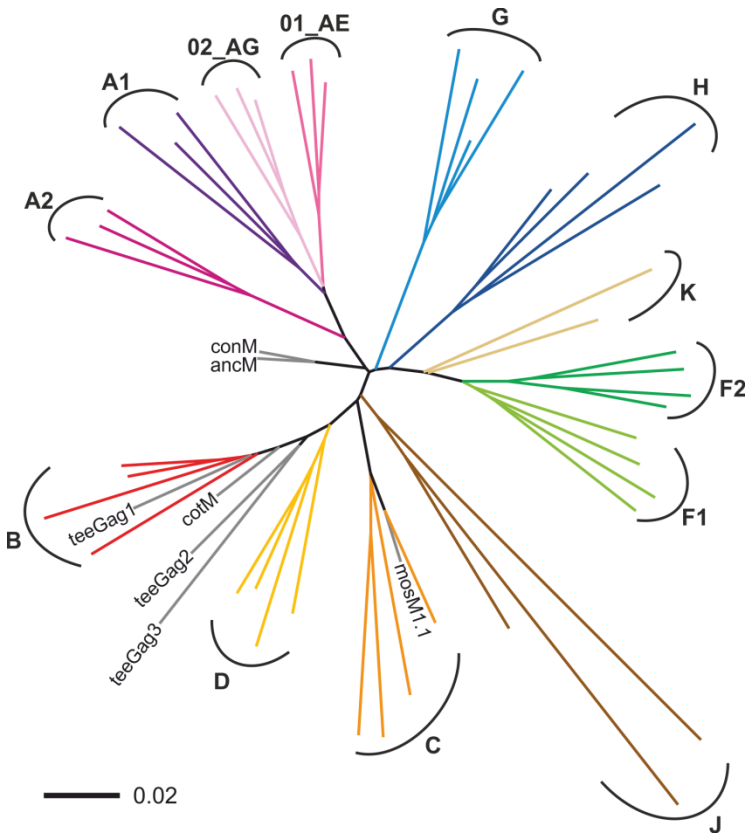


Figure 28. Phylogenetic tree reconstruction of artificial Gag antigens among a set of subtype reference Gag sequences. Sequence alignment and tree reconstructions were performed with ClustalX 2.1. Phylogenetic classification of subtypes (color-coded and lettered accordingly) and antigens of interest (indicated in grey) was visualized with Rainbow Tree. The unit of branch length indicates the amino acid substitutions per site.

D.2.2 Antigen score

To assess the quality of the newly designed teeGags, the number of unique epitopes from the input set included, and the correspondingly calculated antigen score (B.3.2), were determined. In a vaccine including only a single Gag antigen (monovalent regimen), teeGag1 would, by virtue of the design, be the most potent of the herein described novel T cell epitope-enriched variants, since it includes the highest scoring epitopes (Figure 27). Regarding the number of incorporated epitopes, as well as the overall antigen score, teeGag1 (Epitopes e: 402; Score s: 590) was clearly superior to all 5001 natural Gag variants from the filtered web alignment (Figure 29 A). Due to the overrepresentation of B-clade-derived epitopes, nearly all of the natural sequences with a high number of epitopes were B-clade-associated. As a consequence of the different weighting of the subtypes for the scoring function, this B clade bias was equalized to some extent for the antigen score (Figure 29 A - highlighted in dark grey). In addition to natural Gag sequences, teeGag1 was also compared to other *in silico* computed antigens and Gags that had been administered in clinical phase IIb/III studies so far. teeGag1 surpassed them all regarding the number of epitopes and the score (Extended Data Table 5). In Figure 29, the group M variants for the consensus (conM; e: 283; s: 424), ancestral (ancM; e: 293; s: 426), center-of-tree (cotM; e: 311; s: 450), and mosaic Gag sequences (mosM1.1; e: 277; s: 466) are presented, since those are the most comparable sequences to our proof of concept teeGags that were designed as pan-clade vaccine candidates.

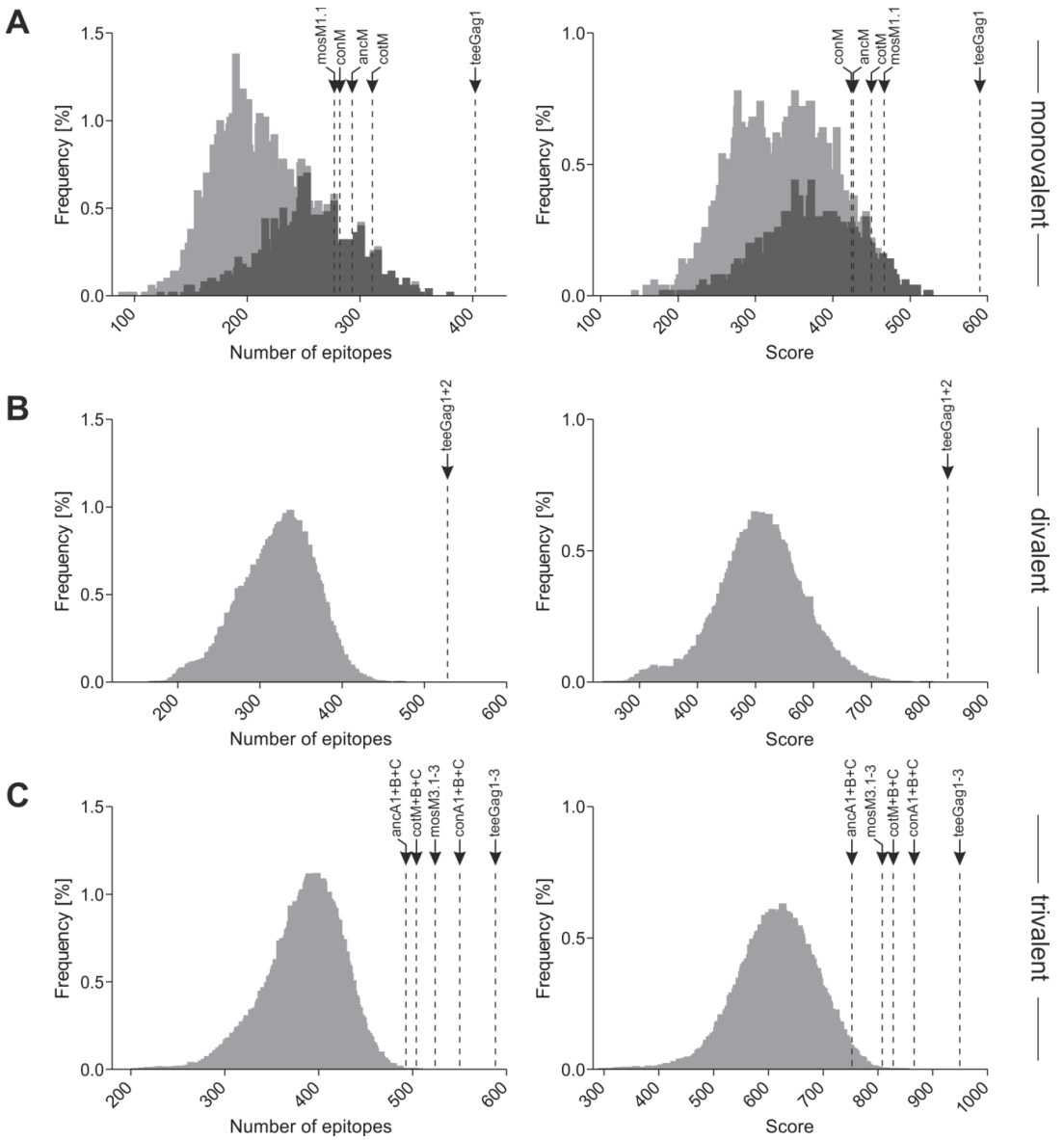


Figure 29. Epitope and antigen score coverage of mono-, di-, and trivalent sequence sets of natural or artificially designed Gag sequences. (A) Frequency of all 5001 natural Gag sequences that include the respective number of epitopes (left panel) or attain the respective antigen score (right panel) as indicated on the x-axis. Values of all B clade sequences from the natural set are highlighted in dark grey. Number of epitopes and scores of the examined Gag antigen designs are indicated by dashed lines. (B) Number of epitopes and score frequencies of 100,000 combinations of two randomly selected natural Gag sequences (grey). The values of the divalent teeGag1+2 are indicated by the dashed lines. (C) Frequencies regarding number of epitopes and score of a 100,000 trivalent, randomly combined natural Gag sequences (grey) and trivalent artificial Gag antigens (dashed lines).

Combinations of Gag variants derived from the first two or three iterations of the algorithm (teeGag1+2 or teeGag1-3), in a presumable di- or trivalent vaccine regimen, have a calculated epitope coverage and overall score that is far better than of any of 100,000 combinations created of two or three randomly picked natural Gag sequences (Figure 29 B+C). Due to their design principle, there can only be one consensus, ancestral, and center-of-tree sequence for each set of sequences (i.e. there are no polyvalent sets for these designs). Hence, the trivalent teeGag set was compared to the combination of the consensus and ancestral sequences of the three most frequent subtypes (conA1+B+C and ancA1+B+C) and to the center-of-tree sequences for group M, subtype B, and subtype C (cotM+B+C). Since the mosaic antigens were

conceived as polyvalent vaccine candidates, there was an M-group-specific trivalent set available (mosM3.1-3) for comparison. Of all those trivalent antigen designs, teeGag1-3 (e: 588; s: 949) was the best concerning epitope coverage and score followed by conA1+B+C (e: 551; s: 867), mosM3.1-3 (e: 524; s: 808), cotM+B+C (e: 504; s: 828), and finally ancA1+B+C (e: 493; s: 752) (Figure 29 C).

D.2.3 Population coverage

By accounting for the allele group frequencies of HLA molecules during epitope scoring (D.1.3), the teeGags were created for optimal coverage of the target population, which for our proof of concept antigens described here, would be the general population world-wide, thus the global allele frequencies were applied as target population attributes. To review the actual achieved population coverage of the teeGags, a target population with 1,000 individuals was simulated *in silico* (B.3.3). Each person of this population is specified by an HLA haplotype consisting of two HLA-A, -B and -C allele groups each. The allele groups were chosen randomly, with reference to the reported target population frequencies. Next, each epitope of the set was analyzed, whether it was included in the antigen and also could be presented by at least one of the test person's corresponding HLA alleles. If this was the case, the epitope got included in the test person's epitope set, or otherwise it was rejected. This was repeated for all 1,000 test persons. The test-person-specific epitope sets were finally analyzed for the number of epitopes and the overall score.

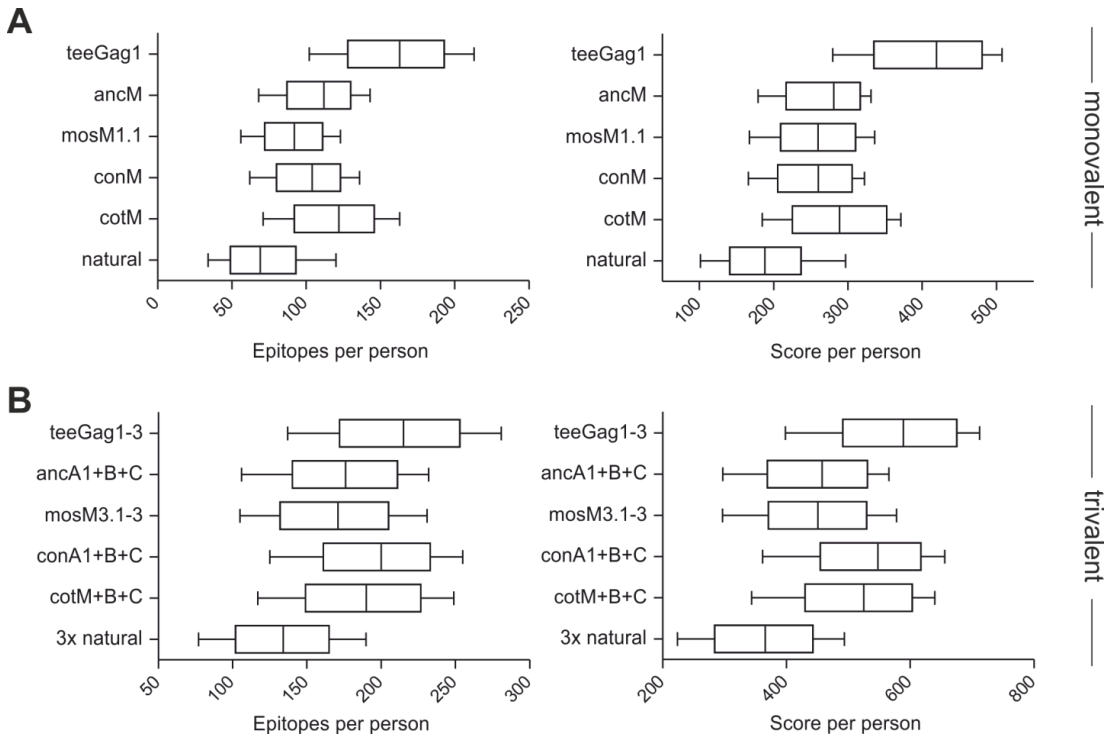


Figure 30. Population coverage of mono- and trivalent antigen sets. For each antigen or combination of antigens, a population of 1,000 persons, defined by their HLA haplotypes that were randomly selected according to the allele frequencies, was generated. The epitopes (left panel) and score (right panel) covered per person are given for (A) monovalent or (B) trivalent Gag antigens. Boxes represent the median with 50 % quartiles, whiskers the 10 and 90 % percentiles. Natural sequences (monovalent) or combinations of three natural sequences (trivalent) were randomly selected from the Gag curated filtered web alignment.

Applying this population coverage calculation, values of cotM (median epitopes: 122; median score: 289), ancM (e: 112; s: 281), mosM1.1 (e: 92; s: 260), and conM (e: 104; s: 260) exceed those of natural Gag sequences (e: 69; s: 188) that were randomly chosen from the Gag sequence alignment for each test-person. However, teeGag1 clearly surpassed all of the monovalent antigen designs, regarding both the number of epitopes and the score (e: 163; s: 419; Figure 30 A). In a trivalent setting, the complementary teeGag1-3 were still the best (e: 215; s: 589), but the difference to the other antigen designs, especially conA1+B+C (e: 200; s: 548), decreased. The combination of three randomly picked natural Gag sequences was clearly inferior (e: 134; s: 366; Figure 30 B). The examined population coverages for a larger set of Gag sequences are displayed in Extended Data Figure 1.

D.2.4 Pathogen coverage

In addition of optimized targeting of the target population, teeGags were conceived to cover the prevailing viruses, as defined by the HIV subtype distribution of the target area, in case of the herein described proof of concept teeGags the worldwide clade distribution. For computational validation of this pathogen coverage (B.3.4), a test pathogen set, comprised of various Gag sequences, was generated. First, for each virus of this set, an HIV subtype was randomly selected proportionally to the natural subtype frequencies in the target region of the conceived vaccine.

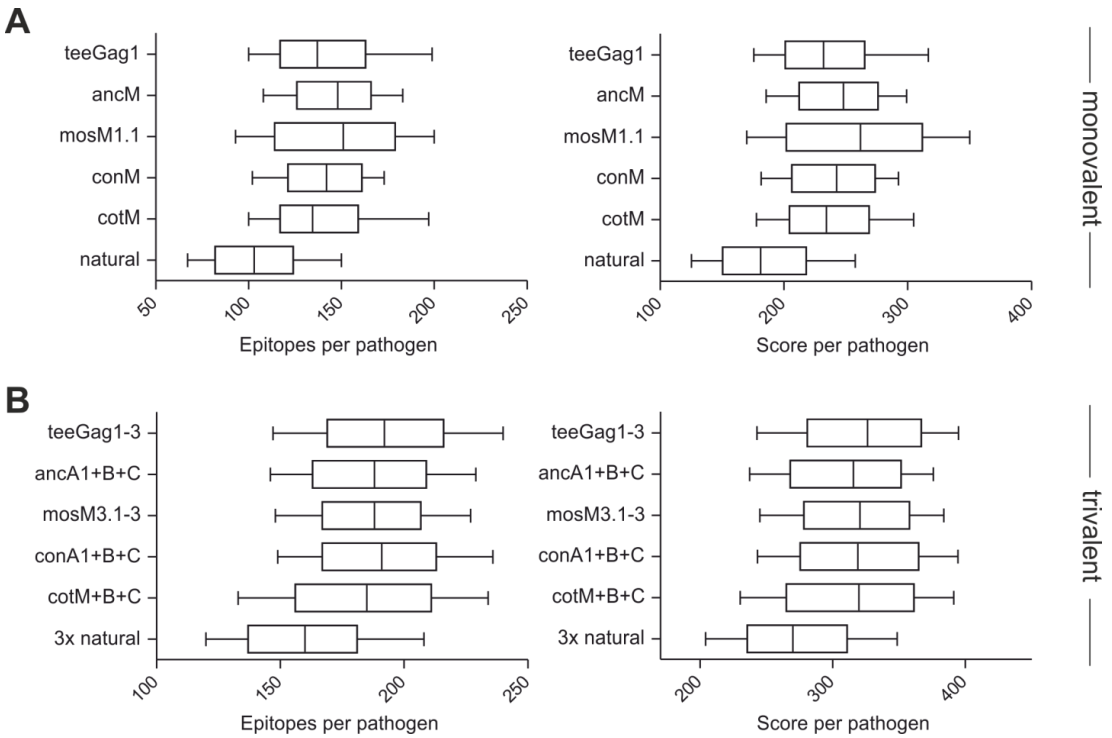


Figure 31. Pathogen coverage of mono- and trivalent antigen sets. For each antigen or combination of antigens, a set of 1,000 test viruses (“pathogens”) was generated. Each pathogen represents a Gag sequence that was randomly selected from the filtered web alignment, with regard to the global subtype frequencies. The number of epitopes (left panel) and the score (right panel) covered per pathogen are given for (A) monovalent or (B) trivalent Gag antigens. Boxes represent the median with 50 % quartiles, whiskers the 10 and 90 % percentiles. Natural antigen sequences (monovalent) or combinations of three natural sequences (trivalent) were randomly selected from the Gag curated filtered web alignment.

Since the teeGag design was based on the global subtype frequencies (D.1.1.4), the same values were initially applied for the assembly of the test pathogen set. After picking the virus subtype, a Gag sequence from the 5001 natural sequences with this subtype was randomly selected as the test pathogens sequence. Next, for each epitope of the input data set, it was checked, if it was included in the antigen of interest, as well as in the test pathogen. If this was the case, the epitope was added to the test-pathogen-specific epitope set, or otherwise it was rejected. The test pathogen selection and computation of the associated epitope set was repeated 1,000 times. For each of these sets, the number of included epitopes and the overall score was calculated.

The pathogen coverage, expressed as the number of epitopes and their score per pathogen, of teeGag1 was far superior to randomly selected natural Gag sequences. However, teeGag1 was slightly inferior to ancM, mosM1.1, and conM and on the same level as cotM (Figure 31 A). This seems to be due to the fact that the epitopes' subtype distribution (Figure 32 C - pie chart) in the database does not resemble the global HIV-1 clade frequencies (Figure 18). This is in line with the phylogenetic tree reconstruction, where teeGag1 and also cotM clustered with the B clade reference sequences.

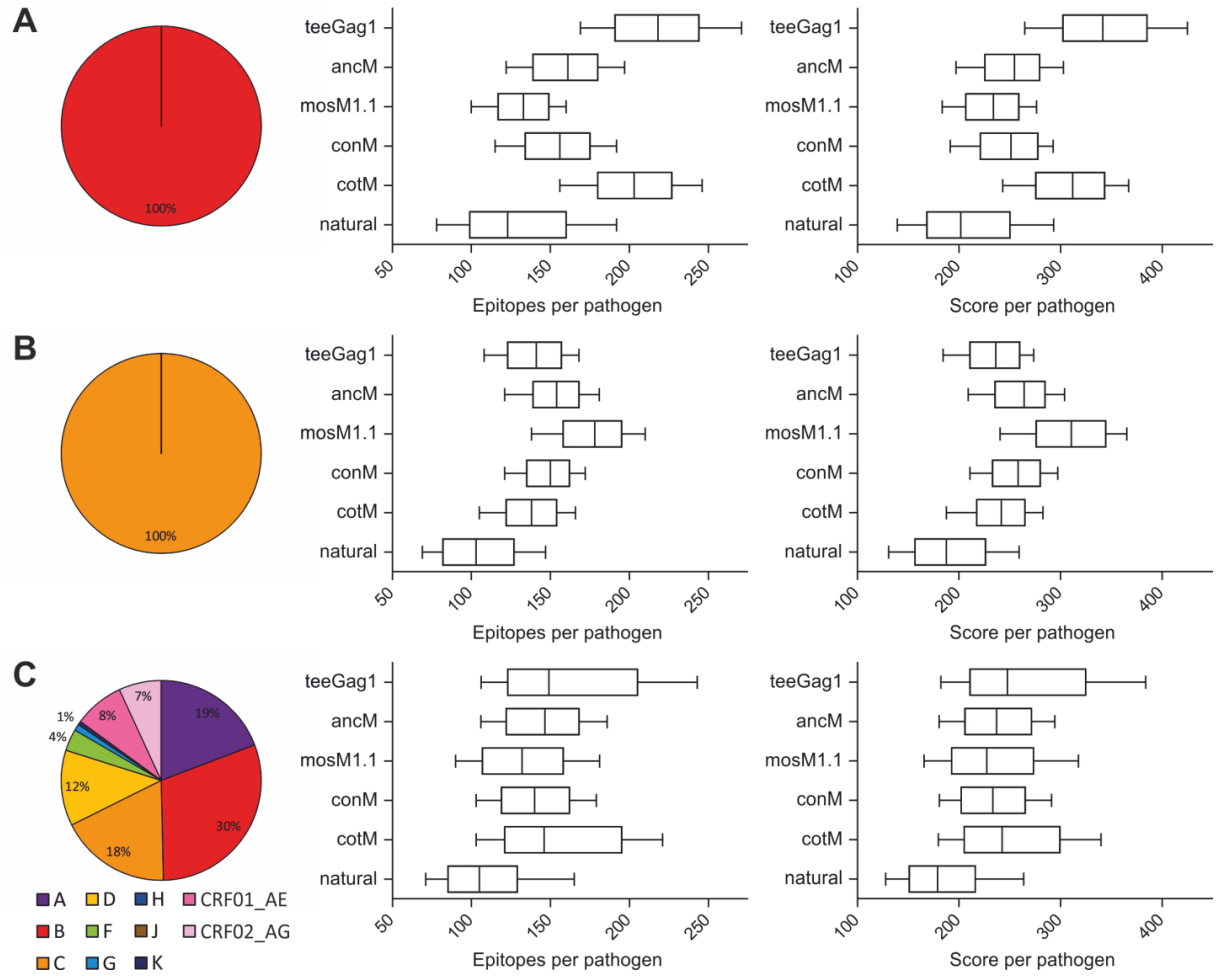


Figure 32. Pathogen coverage for monovalent antigens with differently weighted subtype selections. Pathogen selection was done considering only (A) subtype B or (B) subtype C entries or (C) proportional to the abundance of each subtype in the epitope dataset. For each approach the pathogen set subtype frequencies (left panel) and the calculated pathogen coverage expressed as number of epitopes (middle panel) and overall score (right panel) per pathogen are given. Boxes represent the median with 50 % quartiles, whiskers the 10 and 90 % percentiles.

This B clade bias could be confirmed, when the pathogen coverage was calculated considering only subtype B sequences (Figure 32 A) or based on the abundance of each subtype in the epitope database (Figure 32 C). In both cases, teeGag1, and also cotM, showed a higher epitope and score coverage per pathogen compared to conM, ancM, mosM1.1, or random natural Gag sequences. On the other hand, if only subtype C viruses were considered for the pathogen set, teeGag1 was inferior to conM, ancM, and mosM1.1, yet still superior to natural sequences (Figure 32 B). By far the best design in covering clade C viruses was mosM1.1. This was not surprising, since mosM1.1 was located near the subtype C reference sequences in the phylogenetic classification, unlike conM and ancM that were both located near the center of the unrooted tree (Figure 28). Hence, it seems that the pathogen score highly depends on that the antigen subtype matches the composition of the tested pathogen set.

There is nearly no difference in pathogen coverage concerning the number of epitopes and score between trivalent teeGag1-3, ancA1+B+C, mosM3.1-3, and conA1+B+C, although teeGag1-3 had the highest median value in both cases (Figure 31 B). Random selection and combination of three natural Gag sequences resulted in a far inferior pathogen coverage. Coverage scores for the complete set of antigens analyzed are depicted in Extended Data Figure 2.

D.2.5 Mosaic analysis tools

The LANL HIV-1 database offers, besides the possibility to design mosaic (mos) sequences[199], a suite of web-based tools[293] to calculate the coverage of *k*-mers, representing potential T cell epitopes (PTEs). PTEs are not based on any experimentally validated or through *in silico* processing and HLA binding predicted epitope sequences, but are only defined as possible *k*-mers from a set of viral sequences. To analyze the teeGags according to the criteria used for the mosaic design, their published evaluation tools to calculate the coverage of PTEs (i.e. *k*-mers) was used. For this, the programs *Epicover* and *Posicover* were applied to compare the coverage of the curated filtered web alignment by different antigen sets. For *Epicover,* the nominal epitope length *k* for computing the coverage was set to 9, since this is the most frequent length of HLA-class-I-presented epitopes[294]. Due to the size of the filtered web alignment the threshold for rare sequences was as recommended increased (set to 5), the remaining options were kept on default. As before, the mono- and trivalent combinations of anc, mos, con, and cot sequences were compared to the newly designed teeGags. As output, the fraction of 9-mers shared with the respective antigen set is given as per-sequence mean of all sequences in the Gag alignment. The coverage is divided in exact (i.e. 9 of 9 amino acids match; Off-by-0), off-by-1 (i.e. at least 8/9 match), and off-by-2 (i.e. at least 7/9 match) fits. In a monovalent setting (Figure 33 A - left) there was no difference (exact match always between 0.40 and 0.42) between the five different antigens concerning 9-mer coverage of the curated filtered web alignment. Since mosM3.1-3, was optimized to cover naturally occurring 9-mers it exhibited, as expected, the highest coverage of all trivalent antigen sets (Off-by-0: 0.66). Second best was conA1+B+C (0.62) followed by the similar ancA1+B+C (0.59), teeGag1-3 (0.58), and cotM+B+C (0.57) (Figure 33 A - right).

With the *Posicover* algorithm, the fraction of antigen-covered 9-mers from the curated filtered web alignment was calculated for each position in the alignment. A sliding window of length 9 was applied for the complete length of Gag, generating up to 9 unique epitopes for each

position. Next, the number of these unique 9-mers covered by the antigen set was calculated for each position of each Gag alignment entry.

For each antigen set, the results are given as fraction of matched 9-mers of the Gag alignment. In Figure 33 B the alignment positions are sorted by coverage. Like for the overall coverage (Figure 33 A), there were no differences between the monovalent Gag antigens (Figure 33 B left panels) regarding coverage per position. Only cotM had a slightly higher percentage in the middle fraction range of matched 9-mers.

For the trivalent antigen sets mosM3.1-3 (Figure 33 B right panels) was the best option across most positions. This was expected because it had by far the best overall 9-mer coverage (Figure 33 A). Keeping this in mind, it was not surprising that conA1+B+C showed the next best coverage by position followed by the similar teeGag1-3, ancA1+B+C, and cotM+B+C.

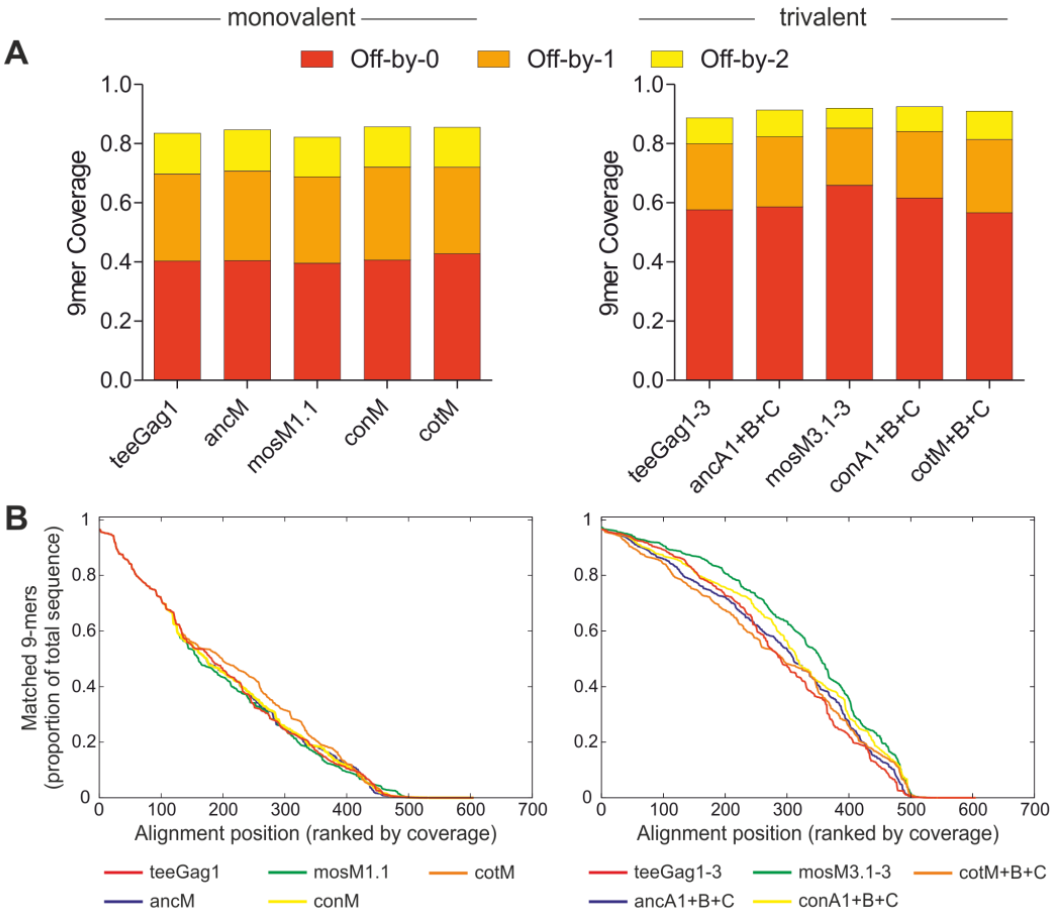


Figure 33. 9-mer coverage of mono- and trivalent antigen sets. For each antigen set, the covered fraction of potential 9-mers from the curated filtered web alignment was calculated. (A) Per-sequence coverage mean by monovalent (left) or trivalent (right) antigen sets as calculated by Epicover. The color code highlights the exact (off-by-0, red), off-by-1 (orange), and off-by-2 (yellow) covered 9-mers. (B+C) Posicover-computed, alignment-position-specific 9-mer coverage rates. Positons on the x-axis were ranked by coverage. Analysis was spilt into (B) matched and (C) missed 9-mers for monovalent (left) and trivalent (right) antigens.

In Extended Data Figure 3, the coverage was mapped on the filtered web-alignment, indicating high and low fractions of matched 9-mers by position of all individual sequences. The Gag alignment sequences were partitioned according to their subtypes. As for the phylogenetic tree reconstruction and the pathogen coverage, here, again the B clade bias for teeGag1 was

apparent, since the highest 9-mer coverage was observed for subtype B sequences. cotM coverage was similar to teeGag1, also with best coverage for subtype B, whereas mosM1.1 best matched the 9-mers of subtype C Gag sequences. For ancM and conM no clade preference could be observed.

For the trivalent antigen combinations the coverage was more homogeneous across all HIV-1 subtypes. By displaying the antigen-matched 9-mers not by coverage rank, but by natural alignment position, differences between the Gag protein domains became obvious (Extended Data Figure 4). For the most conserved region, p24, the 9-mer coverage by all antigens was highest. This is in good accordance with the number of as budding-retaining classified AAS and the position-specific Shannon entropy that both were also distinctly influenced by the low variability of p24 (Figure 25).

D.3 Validation of functional conservation

After the *Optimizer Algorithm* was shown in computational analyses (D.2) of being able to design antigens with an enhanced breadth of immunologically potent CD8+ T cell epitopes, the second feature of the program, the ability to preserve Gag functionality was addressed experimentally. First, the FLD classified AAS were examined separately to validate the predictions (D.3.1). Afterwards, the novel teeGags were biochemically characterized (D.3.2) to show that the incorporation of up to 40 AAS did not alter their ability to release VLPs.

D.3.1 Effects of single amino acid substitutions on VLP budding

To determine the quality of the FLD classification predictions, all AAS present in teeGag1, 2, or 3 (84 different AAS; Extended Data Table 6) and all as negative predicted AAS (18 AAS; Extended Data Table 7) were validated experimentally. As template, the human codon-optimized HXB2-Gag was ordered from GeneArt (Regensburg, Germany) and cloned using the KpnI and XhoI restrictions sites into pcDNA3.1(+). Afterwards, AAS were separately introduced into the reference sequence, resulting in Gag sequences with single site mutations compared to HXB-Gag. The mutations were introduced by fusion PCR (C.2.2.2) using mutagenesis oligonucleotides (with pc31 fwd/rev as outside binding primers), to change the codon of the original amino acid to the most frequent human codon for the desired mutation. A first batch of single AAS Gag sequences was provided directly by GeneArt and was also inserted into pcDNA3.1(+). Cloning of all 102 variants was verified by sequencing. DNA for transfection experiments was prepared with the “QIAprep Spin Miniprep Kit”.

To quantify the VLP release of the single AAS Gag variants, HEK293T cells were co-transfected in a 12-well plate with 1 µg of the Gag-plasmid and 0.05 µg pBluescript KS(-) CMV-SEAP vector using PEI (C.2.3.2). After 6 h, the medium was changed to DMEM-0. Conditioned supernatant was harvested after 48 h and cleared by two subsequent centrifugation steps. The cell-free supernatant was treated for 1 h with 0.5 % Triton X-100. Budding capacity was examined using a Gag-ELISA (C.2.5.2). Additionally, the SEAP concentration was determined in an enzymatic activity test to normalize for differences in transfection efficiency (C.2.5.4). The relative budding (RB) of all single AAS Gag proteins was calculated by dividing the Gag to SEAP concentration ratios of the mutated variant (*AAS*) and the HXB2-Gag wildtype protein (*wt*):

$$Relative\ budding\ (RB) = \frac{\frac{c(Gag)_{AAS}}{c(SEAP)_{AAS}}}{\frac{c(Gag)_{wt}}{c(SEAP)_{wt}}}$$

Figure 34 shows the RB for all 84 Gag sequences with a positively predicted (+) AAS, as well as the entire set of 18 negatively predicted (-) AAS sequences (individual RB values in Extended Data Table 6 and Extended Data Table 7). Analyzing the two groups with the Mann-Whitney test revealed a significant (p=0.0007) reduced RB in the group of as budding-deleterious predicted AAS. Further, when analyzing all individual AAS in a Bonferroni corrected t-test, 9 of the 18 budding-deleterious predicted AAS showed significantly reduced budding (marked in red) compared to the overall mean budding. These were hence categorized as true negative AAS.

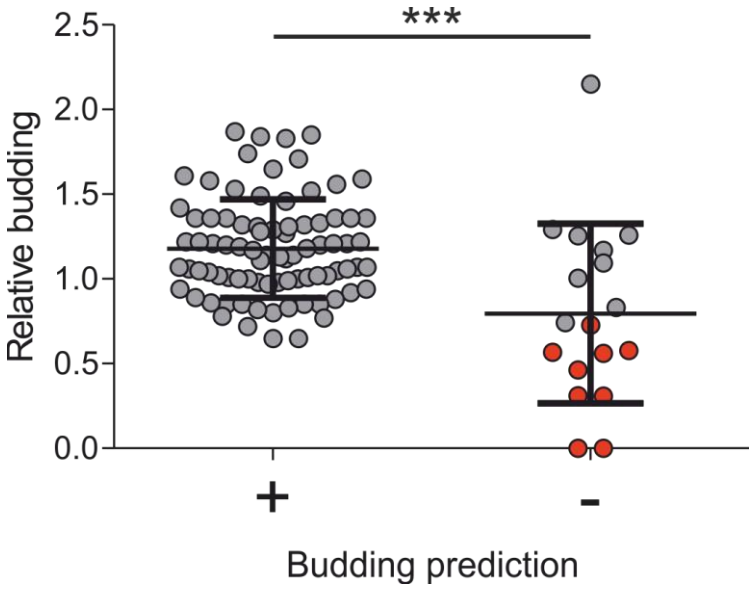


Figure 34. Relative budding (RB) of Gag proteins harboring a single AAS. AAS were grouped into budding-retaining (+) and as budding-deleterious (-) classified mutations. Each data point in the graph represents the RB of one AAS, as determined by at least six independent experiments. Significant reduced budding for single AAS Gags compared to the mean budding of all tested proteins was defined in a Bonferroni corrected two-sided t-test with unequal variance and are highlighted in red. Inter-group (represented as mean ± SD) differences were analyzed with a Mann-Whitney test (p=0.0007).

The experimentally determined accuracy was 91% (Table 14) and thus higher than that calculated using the computational 10-fold cross-validation (86%; Extended Data Table 4). Due to the fact that no as budding-retaining predicted AAS showed significantly reduced budding, the specificity, as well as the PPV, were both 100% and therefore even higher than initially predicted (cross-validation: 99% and 98%). The experimentally determined sensitivity was with 90% also higher than expected (75%), which is partly due to the huge amount of budding-retaining AAS tested. Only the NPV, with just 50% was clearly reduced in the experimental setting, compared to the 10-fold cross-validation predictions (78%). This means that only half of the AAS predicted to be budding-deleterious result in a significantly reduced budding, if they are included in the HXB2-Gag reference sequence. On the other hand, a PPV of 100% shows that all as budding-retaining predicted AAS indeed have no negative influence on budding, which had been declared as the most important feature of the classifier beforehand.

Table 14. Contingency table of experimentally validated binary classification predictions. 84 as budding-retaining and all 18 as budding-deleterious predicted AAS were validated experimentally regarding their effect on VLP budding. Accuracy, sensitivity, specificity, positive predictive value (PPV, precision), and negative predictive value (NPV) were calculated as described in Table 5.

		Experimentally validated data			
		Budding-retaining	Budding-deleterious	Sum	
FLD classification	Budding-retaining	84	0	84	PPV (precision) 100%
	Budding-deleterious	9	9	18	NPV: 50.0%
Sum		93	9	102	
Accuracy: 91.2%		Sensitivity: 90.3%	Specificity: 100%		

D.3.2 Biochemical characterization of teeGags

D.3.2.1 Budding capacity of teeGags1-3

After validating the quality of the classifier predictions, which showed that the *Optimizer Algorithm* can successfully exclude epitope with negative effect on VLP production, the teeGags were biochemically examined. Since the Gag variants designed in the first three iterations of the *Optimizer Algorithm* already contained 88% of all possible epitopes that account for 91% of the maximum score, only teeGag1, teeGag2, and teeGag3 were characterized in detail.

The antigens were ordered as human-codon-optimized nucleotide sequences from GeneArt. Because of the incorporation of up to 40 mutations in the teeGags as compared to the HXB2-Gag reference sequence, alterations of the antibody binding affinity for ELISA quantification could not be excluded. Therefore, to address the budding capacity in a sequence-independent analysis, a 6x-His epitope tag was added to the C-terminus of the teeGags, the HXB2-Gag and the budding-incompetent G2A-mutated reference HXB2-Gag^{Myr-}. The tag was appended to the sequences with primer extension PCR (primer teeGag-His rev) and cloned into pcDNA3.1(+) using KpnI and XhoI restriction enzymes. HEK293T cells were co-transfected with the pcDNA3.1(+)-Gag variant and pBluescript KS(-) CMV-SEAP as described above (D.3.1). The conditioned medium was harvested 48 h later as described above (D.3.1). Dilutions of the samples and a reference standard were transferred to a 0.45 µm nitrocellulose membrane in a slot blot assay (C.2.5.13). The membrane was blocked overnight at 4°C with TBS-M. Gag protein was detected using a biotinylated α-6x-His Epitope Tag antibody followed by incubation with peroxidase conjugated streptavidin. ECL substrate was added and the amount of His-tagged Gag was measured densitometrically. Relative budding was again described as the SEAP normalized ratios of the variant Gag and the wildtype reference Gag concentrations (D.3.1).

Although all three teeGag variants examined were able to release VLPs into the supernatant (Figure 35), teeGag2 had a slightly, but statistically significant (as determined in a Bonferroni-corrected t-test) reduced RB (0.78). teeGag1 and teeGag3, however, exhibited wildtype like behavior. As expected, almost no Gag could be detected in the supernatant of HXB2-Gag^{Myr-}. As all proof of concept teeGags showed sufficient VLP release, they were further characterized.

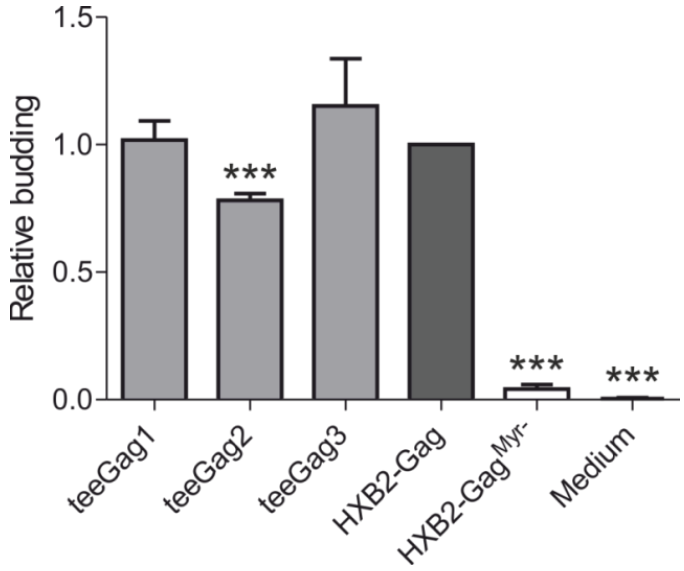


Figure 35. Budding-capacity of novel T cell epitope-enriched Gag antigens: The densitometrically determined, SEAP-normalized RB is plotted on the y-axis. For each 6xHis-epitope-tagged Gag protein, the mean ± SEM of six independent experiments is given. Statistical significance compared to the reference HXB2-Gag was calculated with a Bonferroni-corrected two-sided t-test with equal variance.

D.3.2.2 Sucrose gradient ultracentrifugation for teeGag particle size analysis

Gag virus-like particles have a rather defined size of 100-120 nm and can be purified in a sucrose gradient due to their characteristic density. Wildtype HXB2-Gag-like density would be a good quality criterion for the teeGags, indicating normally-sized particles. To assess the density plasmids encoding all three teeGags, HXB2-Gag, and HXB2-Gag^{Myr-} (all with C-terminal 6x-His epitope tag) were transfected into HEK293T cells in a 6-well plate. The conditioned supernatants were harvested after 48 h and pre-cleared by centrifugation. Afterwards, 2 ml of the supernatant were directly loaded onto a 10-50% sucrose gradient (C.2.5.7). Particles were separated according to their density in an ultracentrifugation run. 21 fraction of 550 µl each were taken by pipetting from the top and their density was determined by weighing. Gag-containing fractions were identified via immunoblotting using the α-p24/p55 antibody CB-13/5 and an HRP-conjugated α-mouse secondary antibody. The identified fractions, and those in close vicinity, were further analyzed in a Gag-ELISA. The concentrations are presented as percent, relative to the maximum value of each gradient in Figure 36.

Density, immunoblot and ELISA data for all Gags are summarized in Figure 36. After the ultracentrifugation run, each gradient showed a rather linear density increase across the 21 fractions ranging from ρ=0.96g/ml to ρ=1.24g/ml. As expected, no immunoblot signal could be detected for the budding deficient HXB2-Gag^{Myr-}. The reference protein HXB2-Gag, as well as all three teeGags on the other hands showed clear bands, restricted to fractions 14 and 15. These findings were supported by the Gag-ELISA that also peaked at fraction 15. The density of the fractions with the highest Gag concentration was, with 1.16 g/ml, identical for teeGag1, teeGag3, and HXB2-Gag. Only for teeGag2, the density of the fraction with maximal Gag content was slightly higher (ρ=1.18 g/ml). In conclusion, all Gags, except the negative control, displayed the expected gradient distribution, indicating well-constructed virus-like particles.

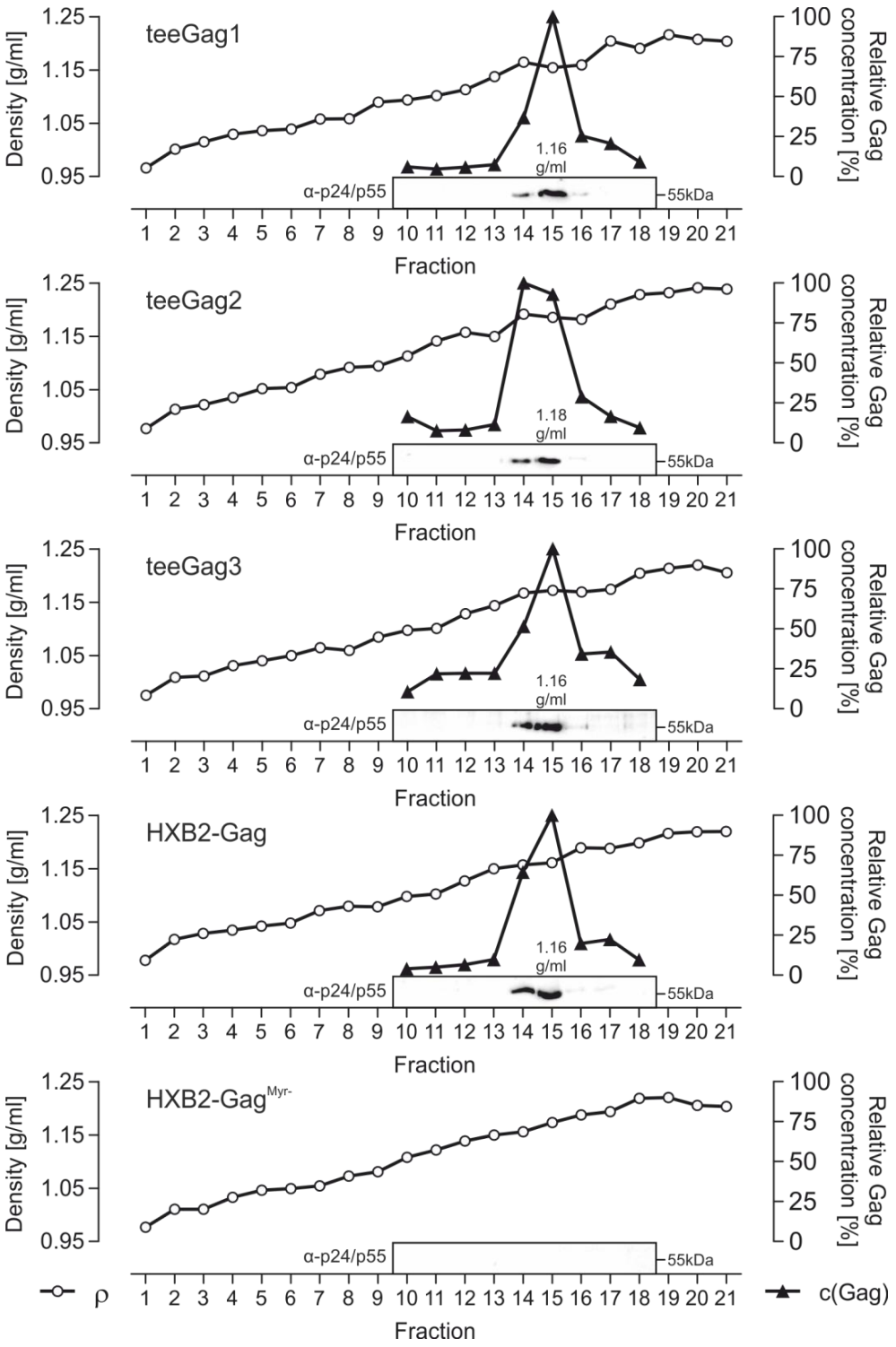


Figure 36. Sucrose gradient ultracentrifugation analysis. VLP containing conditioned supernatants were laid onto a 10-50% sucrose gradient and particles were separated, according to their density, in an ultracentrifugation run. After the run, fractions of 550 µl each were taken, their density (ρ, indicated by white circles) measured, and screened for Gag protein via immunoblot (with α-p24/p55 antibody CB-13/5 and HRP-conjugated α-mouse secondary antibody) and a Gag-ELISA (black triangles, relative to the maximum concentration for each Gag). For each Gag variant, the graph shows the density of all fractions combined with immunoblot and ELISA data for a selected range. The density of the most prominent immunoblot band is stated above.

D.3.2.3 Particle morphologies revealed through electron microscopy

For a more direct way to examine their structure, purified VLPs of all three teeGags, as well as the wildtype HXB2-Gag were visualized through electron microscopy. For each Gag variant (with 6x-His epitope tag), three 15 cm plates of HEK293T cells were transfected. After 48 h, the centrifugally cleared, conditioned medium was ultracentrifuged over a 30% sucrose cushion. The pelleted VLPs were resuspended in PBS and further purified in a 10-50% sucrose gradient centrifugation. The Gag VLP containing fractions, visible to the naked eye (Figure 37 A; confirmed by SDS-PAGE and Coomassie staining, data not shown) were pooled, diluted with DPBS, and again pelleted by ultracentrifugation.

The purified VLPs were resuspended in DPBS, fixed in 2.5% glutaraldehyde and analyzed at the department of ultrastructure research of the Ludwig-Maximilian University of Munich using transmission and scanning electron microscopy technologies (C.2.5.8). As quality control, the Gag content of VLPs was also verified in α-6x-His epitope tag- (with biotinylated α-6x-His Epitope Tag antibody and streptavidin-POD) and α-p24/p55-immunoblots (with α-p24/p55 antibody CB-13/5 and HRP-conjugated α-mouse secondary antibody) (Figure 37 B).

Field emission scanning electron microscopy (FESEM) showed mostly spherical particles for all four different Gag proteins (Figure 38 A+B). The reference HXB2-Gag, teeGag1, and teeGag3 were comparable in size, with a median diameter, determined by image analysis, of about 100 nm. teeGag2, however, was significantly ($p=3.6x10^{-15}$) larger compared to the HXB2-Gag reference with a median diameter of 145 nm (Figure 37 C). Also some aberrant, tubular structures were observed uniquely for teeGag2. Negatively stained spherical VLPs were visualized using transmission electron microscopy (TEM). The spherical particles appeared as electron-dense cores surrounded by a lighter lipid envelope for all Gags (Figure 38 C). For teeGag2 aberrant structures were observed, again.

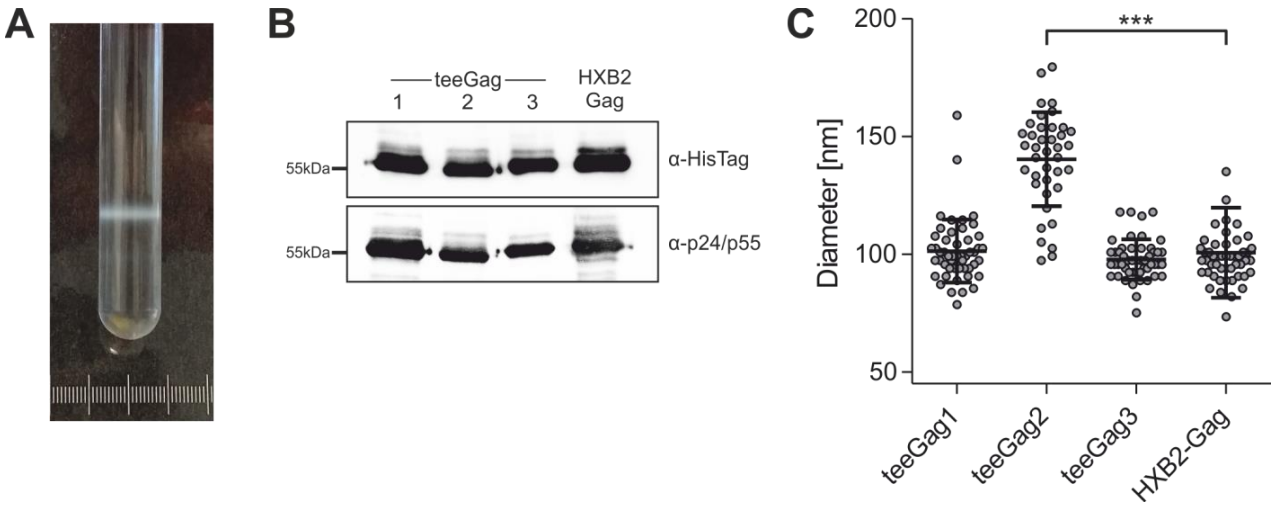


Figure 37. (A) Ring of concentrated VLPs after density ultracentrifugation over a 10-50% sucrose gradient. (B) Immunoblots to validate the presence of Gag in the final samples used for EM analysis. The upper blot was treated with biotinylated α-6x-His Epitope tag Antibody (1:1,000) followed by streptavidin-POD (1:2,000) and the lower blot with α-p24/p55 antibody CB-13/5 (1:1,000) with subsequent HRP-conjugated α-mouse secondary antibody (1:2,000). (C) VLP diameter (y-axis) as determined by analysis of SEM images. Every dot represents the size of an individually measured VLP of the Gag variant, as indicated on the x-axis. For each Gag variant about 50 particles were analyzed. Significant difference to HXB2 was determined by a two-sided t-test with equal variance.

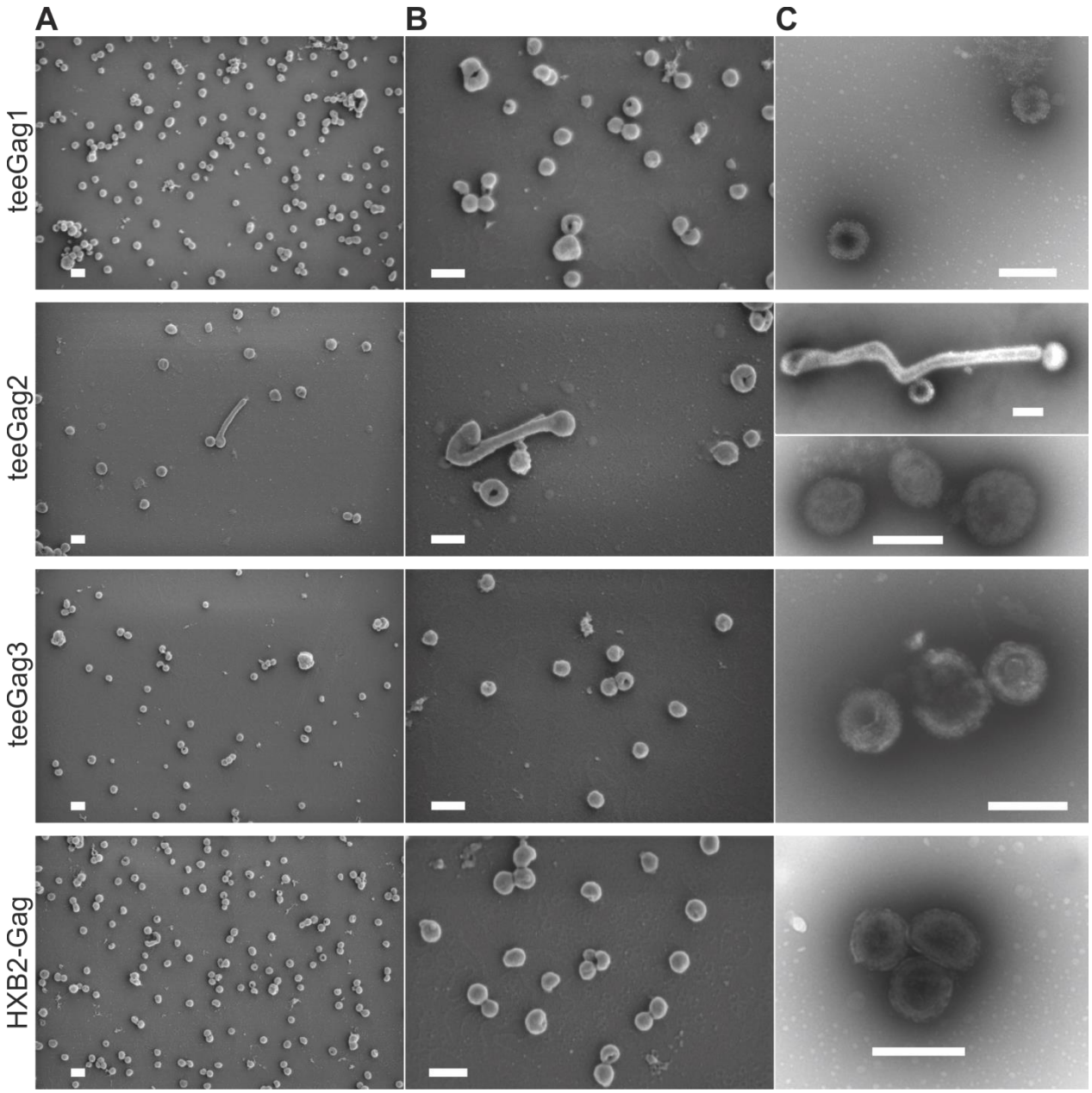


Figure 38. FESEM and TEM images of HIV-1 Gag VLPs. (A+B) FESEM images of VLPs from the respective Gag variant indicated on the left of each row in (A) 20,000x and (B) 50,000x magnification, examined with a Zeiss Auriga scanning electron microscope operated at 1 kV. (C) TEM images taken with an EM 912 electron microscope (Zeiss) in various magnifications, representing the spherical VLPs with electro-dense cores. For teeGag2 the observed aberrant structure was also resolved in a TEM image. The white scale bar always represents 200 nm.

D.3.2.4 Dynamic light scattering for particle size comparison

In addition to analysis of electron microscopy images, the size of Gag particles was also compared via dynamic light scattering (DLS) analysis. For this, the VLPs were purified as described before (D.3.2.3), but without glutaraldehyde fixation, and measured with a Malvern High Performance Particle Sizer (C.2.5.9). Although the particle diameters, determined using DLS, were distinctly greater than in the EM analysis (E.2.2.1), teeGag1 and HXB2-Gag were again identical, when comparing their size distributions (Figure 39 A) and median diameters (Figure 39 B). The size of teeGag3 was slightly, but significantly ($p=0.04$) increased and, as

observed before, teeGag2 was distinctly (p=0.03) larger than the reference VLPs (Figure 39 B). The differences of teeGag2 and teeGag3 to the reference can also be observed in a shift to the right (higher diameter) in the size distribution histograms (Figure 39 A).

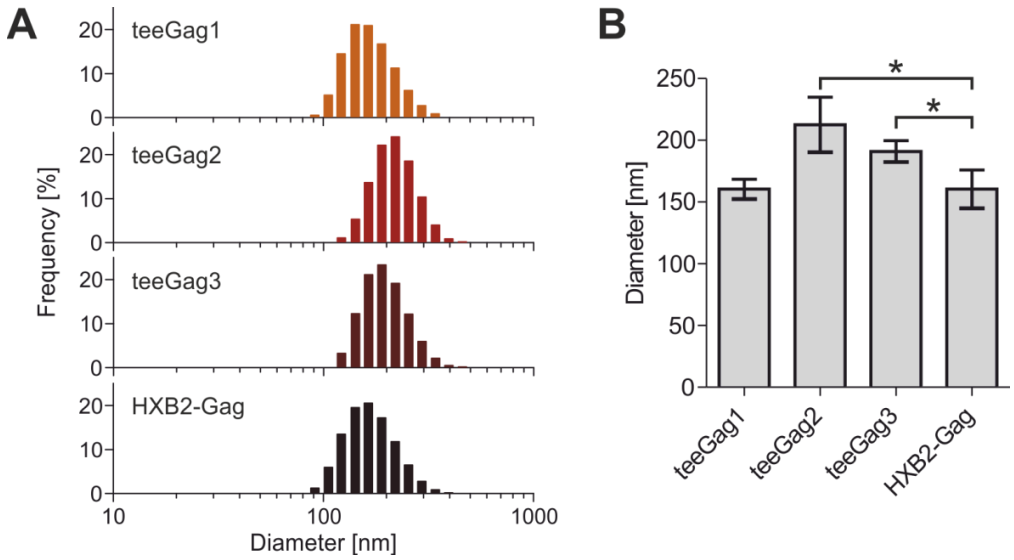


Figure 39. Dynamic light scattering analysis of HIV-1 Gag VLPs. Purified VLPs from Gag proteins were examined in a High Performance Particle Sizer in three replicates. Determined particle diameters are given as (A) size distribution histogram or (B) mean ± SD of the centers of nonlinear regression analysis of the log-normally distributed particle diameters. Significant difference to the reference was calculated with a two-sided t-test with equal variance.

D.3.3 Initial immunological characterization of teeGags-VLPs

After the biochemical characterization, some initial immunological analyses were performed to assess if VLPs of the newly designed teeGags had a similar immunological potential than the wildtype HXB2-Gag VLPs. For this, the capacity of the Gag particles to restimulate a Gag-specific CTL clone was examined. A protocol had been established by Tanja Stief[267], who showed that, for efficient *ex vivo* restimulation, the VLPs had to be pseudotyped with VSV-G. The CTL clone used herein is HLA B*07:02-restricted and recognizes the GL9 peptide at HXB2-Gag numbering positons 355 to 363, at the C-terminal end of p24. During the generation of the teeGags mutations had been introduced within this epitope sequence in teeGag2 (G637S) and teeGag3 (G637S and V362I). Such mutations could potentially reduce or abolish the binding to HLA B*07:02 or the recognition by the GL9-CTL clone. For an unbiased comparison of the immunological potential of Gag particles, the epitope sequence of teeGag1-3 and HXB2-Gag had to be unified. The optimal epitope sequence was determined by using synthetic peptides of naturally occurring GL9 variants for restimulation of the CTL clone. Complemented by *in silico* HLA B*07:02 binding predictions the most promising sequence was GP**S**HKARVL (data not shown). This epitope harbors a glycine to serine mutation at the third position compared to the HXB2-Gag reference and can, for example, be found in the 96ZM651 C clade isolate. This epitope was harmonized in all His-tagged teeGags (teeGag2 already had the appropriate sequence) and HXB2-Gag using fusion PCR (C.2.2.2) and subcloned again into pcDNA3.1(+) making use of the KpnI and XhoI restrictions sites. VSV-G pseudotyped VLPs were produced by co-transfecting HEK293T cells with the Gag vectors and pcDNA5/FRT/TO coding for VSV-G in a 1:1 ratio. For each Gag variant, one 15 cm plate was transfected, and after 48 h the VLPs were

concentrated from conditioned, cell-free medium by ultracentrifugation over a 30% sucrose cushion. The pellet was resuspended in DPBS and frozen at -80°C until further use. The Gag concentration of the VLP samples was quantified in a slot blot as described before (D.3.2.1). As reference for absolute quantification, HXB2-Gag protein of known concentration, as determined by a Gag-ELISA, was used. Monocytes were isolated from an HLA B*07:02 positive donor by MACS sorting with CD14 microbeads and differentiated to mdDCs by cultivation with 1000 U/ml IL-4 and GM-CSF over a 6-day period (C.2.3.4). The quality of differentiation was assessed by surface staining and flow cytometry analysis with the DC-specific α-CD1a antibody (Figure 40 A). Monocyte differentiation to mdDCs was sufficient, with 72.6% of the cells positive for CD1a expression (Figure 40 A). Analysis of the VLPs in α-VSV-G and α-p24/p55 immunoblots of the concentrated VLPs (Figure 40 B) exhibited the expected bands. The minor height differences of the bands, as determined in the α-p24/p55 immunoblot are probably due to the different amino acid composition of the Gag variants. For the co-cultivation assay, mdDCs (5×10^5 each in a 96-U-bottom well plate) were pulsed with various amounts of VLPs, washed 2 h later and incubated for another 4 h (C.2.3.6). Subsequently, BFA-treated CTLs were added to the mdDCs in a 1:1 effector to target ratio (C.2.3.8). This mixed leukocyte reaction (MLR) was continued for 6 h. Co-cultivation was stopped by treatment with cytofix/cytoperm, followed by immune-staining for CD8+ IFN-γ-producing cells and analysis with a FACS Canto II flow cytometer.

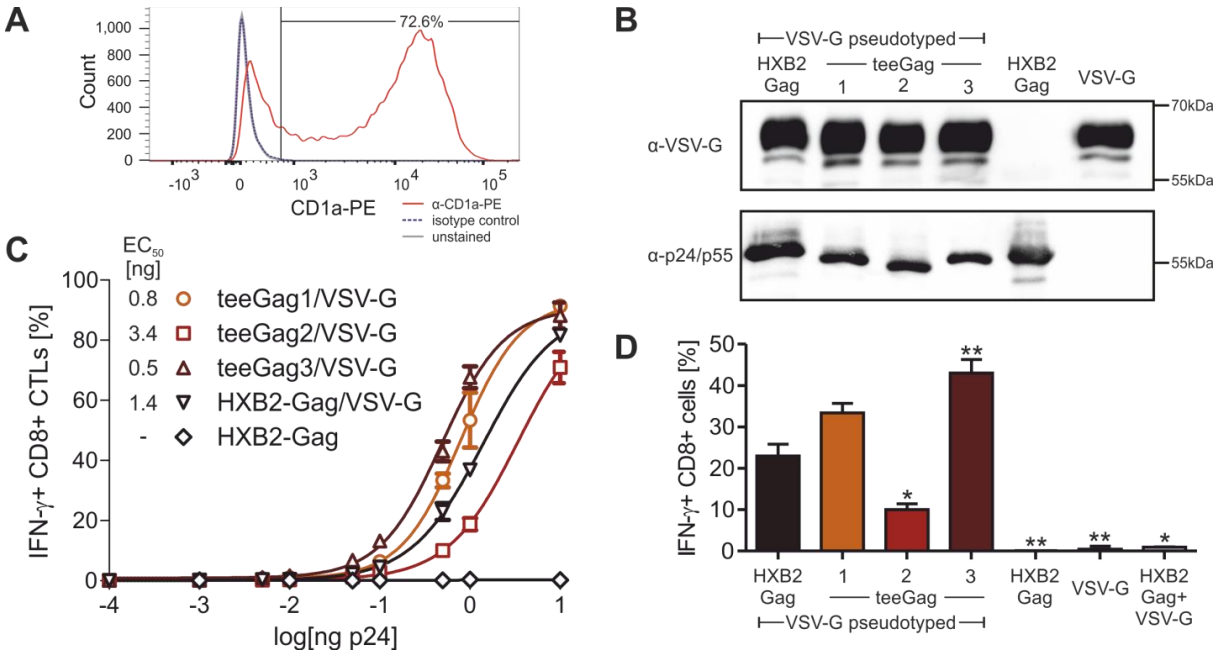


Figure 40. CTL restimulation rates in a mixed leukocyte reaction with mdDCs pulsed with VSV-G pseudotyped VLP. (A) Quality control of the differentiation of MACS-isolated monocytes to mdDCs. Cells were stained with a PE-conjugated α-CD1a antibody or an isotype control antibody and analyzed by flow cytometry. The histogram plot shows the percentage of CD1a expressing cells. (B) Immunoblots of samples concentrated by ultracentrifugation through a sucrose cushion, detecting the VSV-G protein (upper blot), by α-VSV-G and HRP-conjugated α-mouse Ig antibodies and detecting Gag (lower blot), via α-Gag CB-13/5 and HRP-conjugated α-mouse Ig antibodies. (C) mdDCs were pulsed with various Gag-VLP amounts, as indicated on the x-axis, and employed in a mixed leukocyte reaction with GL9-peptide-specific CTLs. As measure of restimulation, the percentages of CD8+ IFN-γ producing cells, determined via flow cytometry analysis, are plotted on the y-axis as mean ± SD of triplicates. EC_{50} values calculated in a variable slope nonlinear regression model are given. (D) Restimulation rates for 0.5 ng p24 equivalents of Gag VLPs used for pulsing with extended controls panel. For the sample "HXB2-Gag+VSV-G" both proteins were purified separately and mixed only before mdDC pulsing. Statistical significance was calculated in a Bonferroni-corrected two-sided t-test with equal variance with reference to the VSV-G pseudotyped HXB2-Gag VLPs.

In the MLR, increasing concentrations of VSV-G pseudotyped VLPs induced higher CTL restimulation, shown as percentage of IFN-γ producing CD8 positive cells. Analyzing the restimulation rates of the various Gag variants in a variable slope nonlinear regression model, EC_{50} values ranging from 0.5 to 3.4 ng p24 (Figure 40 C) were calculated. The least restimulation was determined for the pseudotyped teeGag2, followed by the reference HXB2-Gag and with the highest EC_{50} teeGag1 and teeGag3. HXB2-Gag VLPs that were not pseudotyped with VSV-G, did not cause restimulation of the CTL clone as demonstrated by the absence of IFN-γ-producing cells. CTL restimulation with 0.5 ng p24 protein equivalents of VLPs was analyzed in more detail (Figure 40 D), showing significantly higher rates for teeGag3 compared to the pseudotyped HXB2-Gag VLPs. On the other side, the percentage of IFN-γ producing T cells was clearly reduced for teeGag2, and absent for the controls, that consisted of non-pseudotyped HXB2-Gag VLPs, purified VSV-G, or the combination of both mixed just before pulsing. Summarizing the immunological characterization, it was shown that all VLPs based on teeGags are similar immunogenic as HXB2-Gag reference VLPs. This further highlights that the in the teeGags incorporated mutations did not alter Gag functionality, the VLPs are still able to incorporate viral surface proteins (here VSV-G), and the VLPs get normally processed in mdDCs.

D.3.4 Virological characterization of teeGags integrated in a HIV-1 molecular clone

The biochemical characterization of the teeGags showed that their capability to produce VLPs was comparable to the reference HXB2-Gag. However, these experiments were based on human-codon-optimized genes and produced via mammalian expression vectors. Whether teeGags would still be able to form viral particles in a natural virus setting was unknown. To address this, HIV-codon-optimized (B.4.2) teeGag sequences with uncoupled *gag* and *pol* reading frames as described by Leiherer et al.[251] were ordered from GeneArt. These ranged from the unique 5' BssHII restriction site located in the UTR up to the end of Gag. HXB2-Gag was not synthesized, but amplified per PCR from the p5' vector, which contains the 5' half of an HXB2 molecular clone (kindly provided by Prof. Barbara Schmidt, University of Regensburg). By using suitable primers the gag and pol reading frames of HXB2-Gag were uncoupled (by removing the slippery site and destabilizing the following stem loop) in a fusion PCR reaction (C.2.2.2, with HXB2-AL fwd/rev as mutation primer and teeGag-AL fwd/rev as outside binding primer). HXB2-Gag and teeGag1-3 (amplified with primer teeGag-AL fwd/rev) were fused to a fragment of pNL4-3_AL (amplified with primer pNL4-3AL fwd/rev) ranging from p6* to the unique SbfI restriction site by overlap extension PCR (primer teeGag-AL fwd and NL4-3-AL rev). Budding-deficient HXB2-Gag^{Myr-} was constructed by introducing the G2A mutation via fusion PCR (C.2.2.2; mutation primer: G2A-AL rev/fwd; outside binding primer: teeGag-AL fwd and NL4-3-AL rev). The frameshift-depleted versions of teeGag1-3, HXB2-Gag, and HXB2-Gag^{Myr-} were subcloned into the pNL4-3_AL vector, using BssHII and SbfI restriction endonucleases. Cloning was verified by restriction enzyme digestion and sequencing. HEK293T cells were transfected in a 6-well plate with the pNL4-3_AL plasmid containing the NL4-3 HIV genome with either wildtype NL4-3-AL Gag or teeGag1-3, HXB2-Gag, or HXB2-Gag^{Myr-}. The supernatants were harvested and renewed every 24 h, for three days, and analyzed by Gag-ELISA, reverse transcriptase assay and by immunoblot.

Comparing the wildtype pNL4-3-Gag to HXB2-Gag 48 h after transfection, no difference between the Gag concentrations in the supernatant, as determined by a Gag-ELISA (C.2.5.2), was detected (Figure 41 A). This showed that the exchange of Gag had no implication on budding of the artificial molecular clone. HXB2-Gag was therefore used as sole reference construct hereafter. For HXB2-Gag^{Myr-}, included as negative control, the amount released into the medium was, as expected, strongly reduced. Next, the supernatants of cells transfected with teeGag1-3 were compared to the reference HXB2-Gag and HXB2-Gag^{Myr-} in a Gag immunoblot (Figure 41 B). teeGag1 showed rather reference-like behavior with a strong band at 24 kDa representing the processed p24 and two other bands for the partially cleaved p41 and the full-length p55 Gag at approximately 41 and 55 kDa. teeGag3, on the other hand had a pattern comparable to the budding-incompetent HXB-Gag^{Myr-}. Only a weak band for p24 and nearly no signal for p41 and p55 was observable. teeGag2 showed characteristics of both other variants with a band for p24, but weaker as for teeGag1 and also a signal representing p41. These initial immunoblot results were further analyzed by a Gag-ELISA and an RT-assay (C.2.4.2) of the conditioned supernatants (Figure 41 C+D). In both assays, teeGag1 exhibited a similar behavior as HXB2-Gag, with distinct Gag and RT concentrations in the supernatant, both peaking after 48 h. For teeGag2 and teeGag3 however nearly no Gag or RT-activity was measured in the supernatants. Both variants showed similarities to HXB2-Gag^{Myr-}, as all three had significantly reduced Gag and RT concentrations in comparison to the HXB2-Gag reference, indicating defects in viral release.

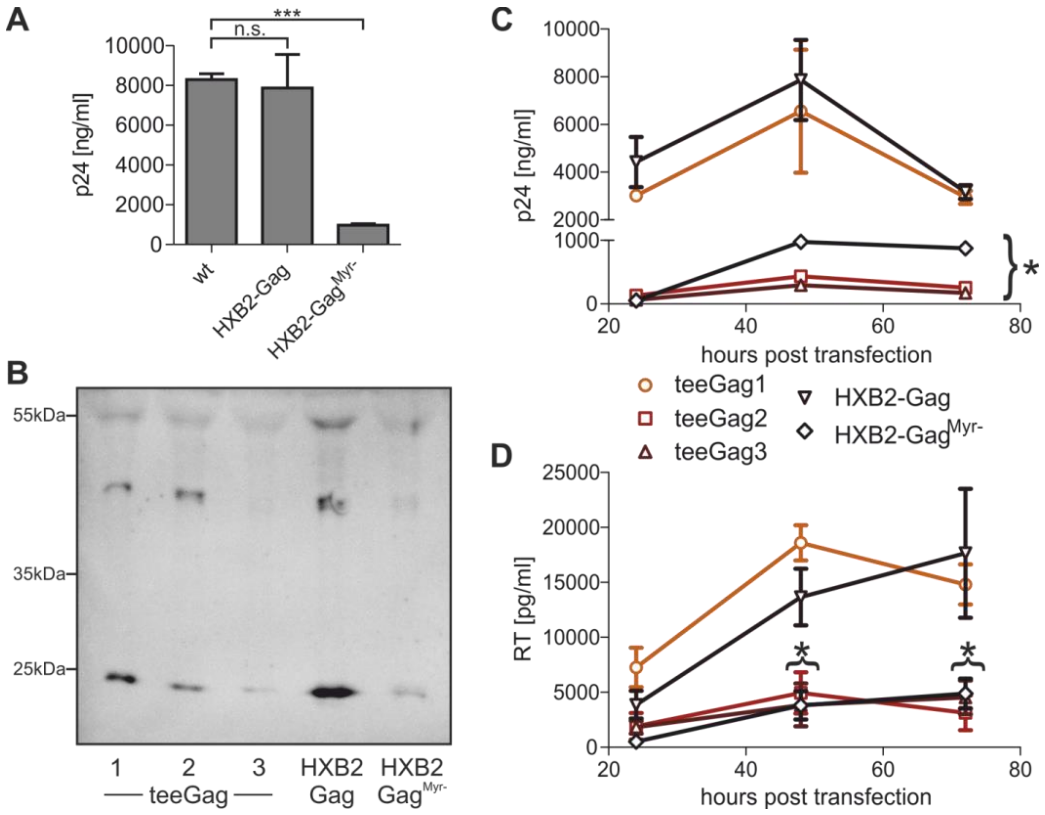


Figure 41. Viral particle release capacity of teeGags integrated into an HIV molecular clone. The wildtype Gag protein of the HIV molecular clone pNL4-3-AL was substituted with frameshift-deleted teeGag1-3, HXB2-Gag, or HXB2-Gag^{Myr-}. HEK293T cells were transfected with these plasmids and the supernatants were harvested every 24 h for 3 days. (A) pNL4-3-AL wildtype Gag concentration in the supernatant 48 hpt, as determined by Gag-ELISA, compared to HXB2-Gag and HXB2-Gag^{Myr-}. (B) Immunoblot of conditioned medium 48 hpt. Gag was detected by α-Gag CB-13/5 and α-mouse Ig antibodies. (C) Gag ELISA and (D) RT-assay data of the conditioned media. Statistical comparison to reference controls was performed with as two-sided t-test with equal variance.

D.4 Method development for *in vitro* assessment of immunological breadth

In addition to the functional conservation, the other main feature of the teeGags was the incorporation of as many immunologically potent, patient-derived CTL epitopes as possible. Experimental *in vivo* validation that teeGags induce a broader T cell response is quite difficult, since animal models are not applicable, due to the fact that the design targets only human HLAs and their associated epitopes. To bypass this limitation and analyze the immunological breadth of antigens, an *in vitro* analysis for epitope presentation was conceived. This method employs cells that produce soluble variants of HLA molecules (sHLA). When these cells also express the antigens of interest, the proteins should be processed through the classical HLA I machinery. Antigen-derived peptides should then be capable of forming a complex with the sHLA protein and the endogenously provided β2m. Since these complexes get secreted to the supernatant, they can be isolated directly from the conditioned medium. After complex destabilization and further purification of the peptides, they can eventually directly be sequenced using tandem mass spectrometry. Through this method HLA-presented epitopes can be identified. Since the presence of a suitable TCR is also required (A.2.2.1), not all identified epitopes are necessary able to a prime CD8+ cell response. Additionally, as long as no complete sequencing of the immunopeptidome can be ensured, it is not possible to conclude that not identified epitopes do not get presented.

D.4.1 Soluble HLA class I molecule design

For establishing epitope sequencing of sHLA-presented peptides via mass spectrometry, two model HLA class I alleles, A*02:01 and B*07:02, were chosen. Both are among the globally most frequent alleles in the respective gene locus (Extended Data Table 3). HLA B*07:02 was of special interest, since the p24 CTL clone recognizing the GL9 peptide used beforehand (D.3.3), is HLA B*07:02-restricted. To generate sHLA expression plasmids, RNA of HEK293T cells, that have a reported HLA haplotype consisting of A*02:01/A*03:01 and B*07:02/ B*07:02[295] was isolated with the RNeasy Plus Mini Kit (Qiagen). The HEK293T mRNA was used in a first-strand cDNA synthesis with random hexamer primers applying the SuperScript™ III Reverse Transcriptase kit (Thermo Fisher). The cDNA was added as template to a PCR reaction amplifying HLA A*02:01 (primers sHLA A*02:01 fwd and sHLA-exon4+His rev) and HLA B*07:02 (primers sHLA B*07:02 fwd and sHLA-exon4+His rev) fragments, each ranging from the start codon till after exon 4, therefore cutting off the transmembrane region encoded in exon 5 and the following cytoplasmic part. These sHLA proteins hence comprise the signal peptide coded by exon 1 and the three alpha subunits of the HLA molecule, coded by exons 2-4. A C-terminal His-tag (introduced by reverse binding primer sHLA-exon4+His rev) that was shown to not distort peptide specificity[112], was added for analytical purposes and as an alternative purification possibility. Both sHLA variants were cloned into the expression plasmids pcDNA3.1(+) and pcDNA5/FRT using HindIII and XhoI restriction sites that were introduced previously by PCR. With the pcDNA5/FRT vectors, coding for sHLA A*02:01 (sA02) and sHLA B*07:02 (sB07), stable expression Flp-In™ T-Rex™ 293 sus cells (C.2.3.3) were generated. Expression of

soluble HLA complexes was validated in a sHLA-specific ELISA (C.2.5.3). Compared to transient transfection with pcDNA3.1(+), the stable cell lines had the advantage that all cells produce the sHLA protein and therefore a significant higher concentration in the supernatant could be achieved (data not shown). Moreover, using stable cell lines is far less laborious, since large amounts of plasmid DNA had to be prepared for transient transfection.

D.4.2 Affinity chromatography of sHLA-peptide complexes

Isolation of sHLA-epitope complexes was performed by affinity chromatography with the murine pan-HLA class I reactive W6/32 antibody. The special feature of this antibody is that it binds to the interface of the heavy HLA chain and the β2m. The interaction of β2m and heavy HLA chain is only stable, if there is also an peptide bound in the peptide binding groove of the HLA molecule[296] and also a prerequisite for the complex to leave the ER. Because of this, the W6/32 antibody allows a specific isolation of complete HLA complexes consisting in this case of the sHLA chain, β2m, and a bound peptide[104]. The W6/32 antibody was produced by a hybridoma cell line and the conditioned medium was run over a self-packed Protein A-Sepharose column that binds the Fc part of antibodies with high affinity (C.2.7.1). Next, the supernatant containing the sHLA complexes was applied to the column using gravity flow. At the end, the sHLA complexes were eluted with 10% acetic acid (C.2.7.2). Crosslinking of the antibody to the Protein A beads before affinity chromatography resulted in a cleaner elution of the HLA complexes, since the antibody remained bound to the beads (data not shown). Additionally, the column could then be reused multiple times. The eluted sHLA complexes were heated to 78°C for 10 min. This acid boil step destabilizes the sHLA complexes and they disaggregate into their components. Peptides were further purified over a 3 kDa cut-off filter (C.2.7.3). The filtrate (i.e. the fraction with substances smaller than 3 kDa), containing the peptides, was used for CTL restimulation experiments and mass spectrometry sequencing, whereas the concentrate (i.e. fraction bigger 3 kDa) was employed for quality controls.

D.4.3 CTL restimulation with isolated peptides

As a first proof of concept experiment, peptides were isolated from the conditioned medium of the cell lines stably expressing sB07 or sA02 that were additionally transiently transfected either with a Gag expression or a control plasmid. The isolated peptides were used for peptide pulsing experiments (C.2.3.7) and subsequent reactivation of the HLA B*07:02-restricted GL9 CTL clone (C.2.3.8). The Gag sequence of the 96ZM651 isolate was chosen, because in previous processing and presentation experiments it exhibited the best GL9 CTL restimulation capacity of all tested natural isolates (data not shown). For each test $6x10^8$ stably sHLA (sA02 or sB07) expressing cells were transfected (600 ml scale) either with pcDNA3.1(+) coding for the human-codon-optimized 96ZM651-Gag or, as negative control, the pcDNA3.1(+) plasmid without any insert. Production of proteins was confirmed via flow-cytometry after α-p24/p55 ICS (with KC57-RD1 antibody) and sHLA-ELISA (data not shown). After 48 h, the conditioned medium was harvested and cleared by centrifugation. Until further use the cleared supernatant, containing the sHLA complexes, was stored at -80°C. For column preparations 400 mg Protein A Sepharose beads were swollen in borate buffer and afterwards loaded with 8 mg W6/32 antibody in a disposable column. After DMP-crosslinking and an acid wash step to remove unbound antibody,

the beads were equally divided onto four columns, one for each test. Next, the sHLA-complex-containing supernatant was thawed and filtered through a 0.22 µm pore sized membrane. The cleared supernatant was mixed with borate buffer and then applied to the W6/32 affinity matrix. After washing, sHLA complexes were eluted stepwise with 500 µl 10% acetic acid until the baseline at 280 nm was reached (measured by UV-absorption at 280 nm of regularly taken aliquots). After acid boiling of the protein-containing fractions, peptides were purified using a 3 kDa cut off filter. The peptides-containing flow-through was dried in a SpeedVac run and reconstituted in 50 µl DPBS. The fraction bigger 3 kDa contained the sHLA protein, β2m, and residual W6/32 antibody. For quality control, Protein A beads before (Bb) and after (Ba) elution, as well as the concentrate of the filtration were separated by SDS-PAGE and stained with Coomassie (Figure 42 A). For all four tests, the beads before the elution and the eluate showed a distinct band for the sHLA protein at about 35kDa.

The protein identity was verified using an α-pan-HLA-ABC and an α-6x-His epitope tag immunoblot (Figure 42 A). Since the murine α-pan-HLA-ABC was detected with a polyclonal α-mouse Ig antibody, also the murine W6/32 antibody used for affinity chromatography that was bound to the beads was visualized. The α-His epitope tag immunoblot, however, showed only one specific band for the C-terminal 6x-His-epitope-tagged sHLA protein. The isolation of complete HLA complexes was finally checked in an α-β2m immunoblot. Again, only for the beads before elution, and the concentrate, a β2m-signal was observed. In conclusion, using the W6/32 antibody for affinity chromatography, results in the isolation of intact sHLA complexes consisting of the sHLA heavy chain and β2m. Due to the crosslinking, nearly no affinity antibody is eluted from the Protein A Sepharose, resulting in a rather pure eluate.

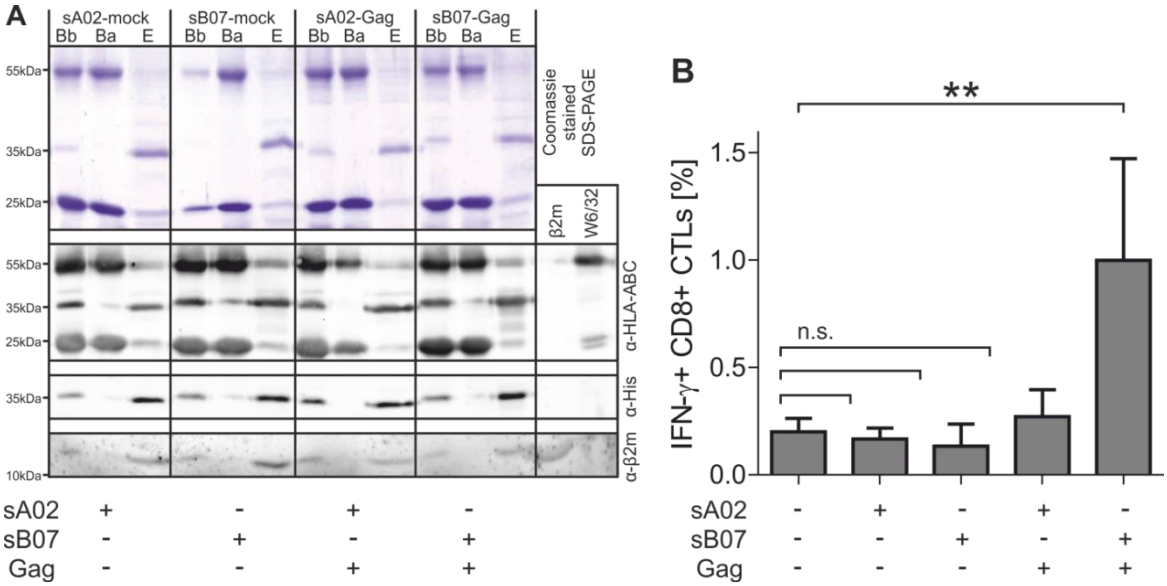


Figure 42. HLA-specific epitope purification. (A) Protein A Sepharose beads before elution (Bb) and after elution (Ba), as well as the concentrate (i.e. fraction >3 kDa after filtration) of the eluate were separated using an SDS-PAGE. For unspecific visualization of all proteins the Gel was stained using Coomassie Brilliant Blue (top row). For identification of protein-specific signals, three different immunoblots were performed: Visualizing the HLA heavy chain (second row; detected with α-HLA-ABC EMR8-5 and HRP-conjugated α-mouse Ig antibodies), the 6x-His epitope tag (third row; detected with biotinylated α-6x-His Epitope Tag antibody and streptavidin-POD), or β2m (bottom row; detected with polyclonal α-β2m and HRP-conjugated α-rabbit Ig antibodies). (B) mdDCs from an HLA B*07:02/A*02:01 positive donor were pulsed with 1.5 µg affinity chromatography purified peptides or PBS as negative control and examined for their capacity to restimulate the HLA B*07:02-restricted GL9-CTL clone. Restimulation was determined as percentage of IFNγ producing CD8+ cells, as determined by flow cytometry and is given as mean with standard deviation of 6 replicates. Significant difference to the negative control (PBS) was calculated in a two-sided t-test with equal variance.

To test the GL9-CTL-restimulation capacity of the isolated peptides, monocytes of an HLA B*07:02 and HLA A*02:01 positive donor were isolated and differentiated to mdDCs. $5x10^4$ mdDCs were seeded in each well of a 96-round-bottom well plate. Next, for each test 1.5 µg of the reconstituted peptides, as determined by UV-absorption at 280 nm, were added to the mdDCs. Two hours after the peptide pulsing, GL9-specific, BFA-treated CTLs were added in a 1:1 target-effector cell ratio. After 6 h, the MLR was stopped by fixation and permeabilization. After staining with appropriate antibodies (α-IFN-γ-APC and α-CD8-FITC), the percentage of IFN-γ producing CD8+ cells was determined in a flow cytometry analysis. Only the peptides isolated from the conditioned medium of sB07 stably transfected cells also expressing Gag (sB07-Gag), induced a CTL restimulation that was significant higher than the background level (only PBS added) (Figure 42 B). Since the HIV-1-Gag-specific GL9-CTL clone is HLA B*07:02 restricted, as expected, all other isolated peptides (sA02-mock, sB07-mock, and sA02-Gag) showed no restimulation. This was taken as evidence that through sHLA affinity chromatography HLA-specific Gag epitopes can be isolated.

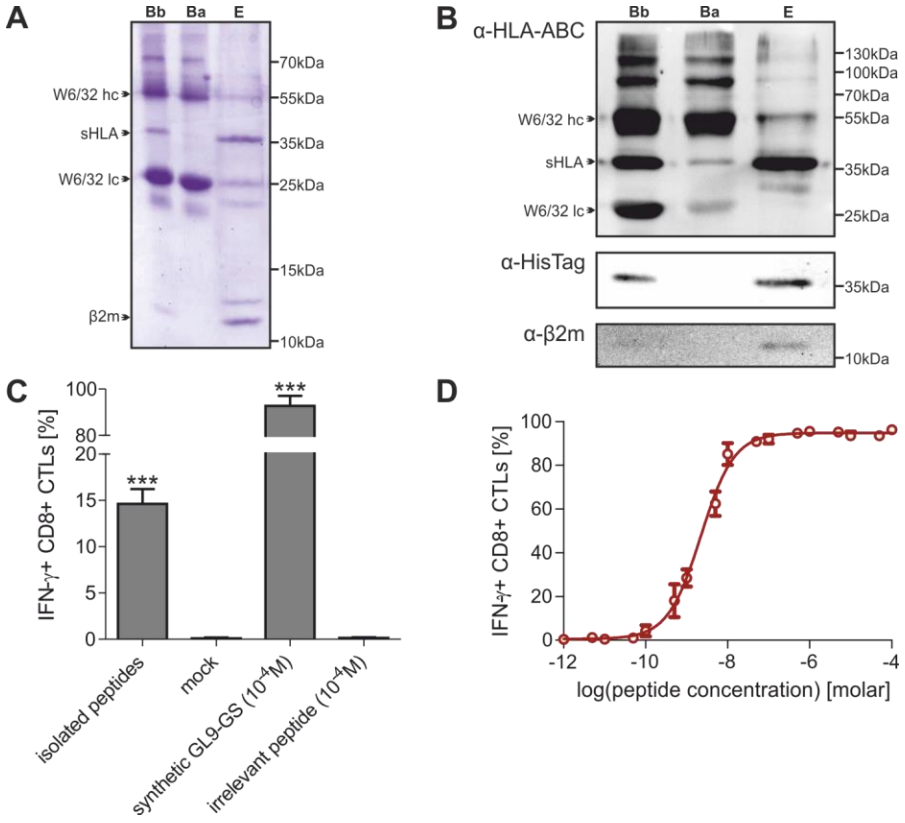


Figure 43. Scaled-up HLA B*07:02-restricted peptide purification for CTL restimulation. Peptides were purified from the supernatant of a stable sB07 expression cell line that was transiently transfected with pcDNA3.1(+)-96ZM651-Gag. (A) Coomassie stained SDS-PAGE as affinity chromatography quality control. Protein A Sepharose beads before elution (Bb), beads after elution (Ba), and the 3 kDa cut off filtration concentrate of the eluate (E) were loaded. (B) Immunoblots detecting the HLA heavy chain (top row; detected with α-HLA-ABC EMR8-5 and HRP-conjugated α-mouse Ig antibodies), the 6x-His epitope tag (second row; detected with biotinylated α-6x-His Epitope Tag antibody and streptavidin-POD), or β2m (bottom row; detected with polyclonal α-β2m and HRP-conjugated α-rabbit Ig antibodies). (C) mdDCs derived from an HLA B*07:02 positive donor were pulsed with affinity chromatography-purified peptide, a mock control (PBS), synthetic GL9-GS peptide (10^{-4} M) or the same concentration of an irrelevant peptide. The graph shows the mean ± SD (from triplicates) percentage of IFNγ producing CD8+ cells after a MLR with peptide-pulsed mdDCs and HLA B*07:02-restricted GL9 CTL clone, as determined by flow cytometry. Significance was determined by calculating p-values in a two-sided t-test with equal variance. (D) Exemplified calibration curve generated by pulsing HLA B*07:02+ mdDCs with different concentrations of synthetic GL-GS peptide and determining the percentage of IFNγ producing CD8+ cells after the MLR with GL9-CTLs. Curve fitting was computed in variable slope nonlinear regression analysis.

Another CTL restimulation experiment aimed at quantifying the amount of GL9 peptide that can be purified. For this, the volume of stable sB07 transfected cells, that were additionally transiently transfected with 96ZM651-Gag, was scaled up. Overall 3.3 l with a concentration of $1x10^6$ sB07 expressing cells/ml were used for the transfection with pcDNA3.1(+)-96ZM651-Gag. The conditioned supernatant was harvested 48 h post transfection, cleared by centrifugation, and stored at -80°C. 300 mg of Protein A Sepharose was swollen in borate buffer, transferred to a disposable column, and loaded with 6 mg W6/32 antibody that was subsequently crosslinked to the beads. Afterwards the conditioned sHLA-containing medium was thawed, filtered through a 0.22 µm membrane and applied to the W6/32-Protein A resin column using a peristaltic pump. After extensive washing steps, the sHLA complexes were eluted stepwise in 1 ml fractions with 10% acetic acid until the baseline at 280 nm was reached. Protein-containing eluate fractions were pooled and, after acid boiling, the peptides were purified using a 3 kDa cut-off filter. The filtrate was dried in a SpeedVac and reconstituted in 100 µl DPBS.

Analyzing the fraction bigger than 3 kDa in a Coomassie stained SDS-PAGE (Figure 43 A), a clear band for the sHLA protein again could be observed for the beads before elution (Bb) and the filtration-concentrated eluate fraction (E). Elution with 10% acetic acid removes nearly all sHLA protein, while the crosslinked W6/32 antibody remains bound to the beads (Ba). This could also be verified in immunoblots (Figure 43 B) against the sHLA protein, the sHLA C-terminally added 6x-His epitope tag, and β2m, where a signal could be observed in the lanes for the beads before elution and the concentrate of the eluate

For the restimulation experiments, mdDCs were generated from monocytes of an HLA B*07:02 positive donor. $5x10^4$ mdDCs were seeded per well of a 96-round-bottom plate and pulsed with 1% or 10% of the affinity-chromatography-purified peptides or known concentrations of the synthetic GL9-GS peptide. After 2 h, GL9-specific CTLs were added in a 1:1 ratio to the mdDCs and incubated for another 6 h. The mixed leukocyte reaction was stopped by fixation and permeabilization of the cells, followed by immunostaining (with α-IFN-γ-APC and α-CD8-FITC antibodies) and flow cytometric analysis for IFN-γ producing CD8+ cells.

10% of the final peptide preparation that contains a mixture of all isolated peptides was applied for the mdDC pulsing and CTL restimulation. This led to a restimulation of 14.6% of all CD8+ cells, as measured by IFN-γ-production (Figure 43 C). Using a calibration curve (exemplified in Figure 43 D), where mdDCs were pulsed with known amounts of the synthetic GL9-GS peptide, the overall amount of GL9-GS in the isolated peptides was determined as 1.3 pmol. Each isolated sHLA molecule should consist of one sHLA heavy chain, one β2m protein, and one peptide. Hence the amount of sHLA could be used to estimate the overall amount of peptide isolated by affinity chromatography.

Of the overall 292 µg protein in the concentrate after filtration, as determined by a Bradford assay, about 89 µg were sHLA protein, as calculated by image analysis of the Coomassie stained SDS-PAGE. Hence, in a very rough estimation, 2.0 nmol of sHLA complexes and therefore the same amount of peptides were isolated. This means that the GL9-GS peptide (1.3 pmol) represent 0.07% of the overall quantity of isolated peptides. Since a cell presents about 10^4 distinct epitopes[91], the determined concentration for the GL9-GS epitope seems to be in a reasonable range. Since tandem mass spectrometry can identify peptides in a femto- down to attomole range[297], the amount of peptides isolated using the presented method were expected to be sufficient for identification via LC-MS/MS.

D.4.4 Peptide identification using LC-MS/MS *de novo* sequencing

After the promising proof of concept experiments, that used affinity-chromatography-isolated peptides to restimulate a Gag-specific CTL clone, the peptides should be sequenced employing tandem mass spectrometry. Three different peptide isolation experiments were performed. Two of them were performed as described above with stable sB07 expression cell lines transiently transfected one time with 96ZM651-Gag (#1) and one time with teeGag1 (#2). In the third experiment, HEK293F cells were transiently co-transfected with pcDNA3.1(+)-sB07 and pcDNA3.1(+)-teeGag1 (#3). For all three approaches the affinity chromatography and peptide purification was identical to the procedure described above (900 ml scale). The purified, SpeedVac-dried peptides were analyzed at the proteomics core facility of the "Helmholtz Zentrum München" in an LC-MS/MS peptide sequencing experiment (C.2.7.4). The obtained spectra were loaded to the *MaxQuant* software with the corresponding search engine *Andromeda*[282,283] to identify endogenous human-protein-derived peptides as well as Gag-specific epitopes (C.2.7.5).

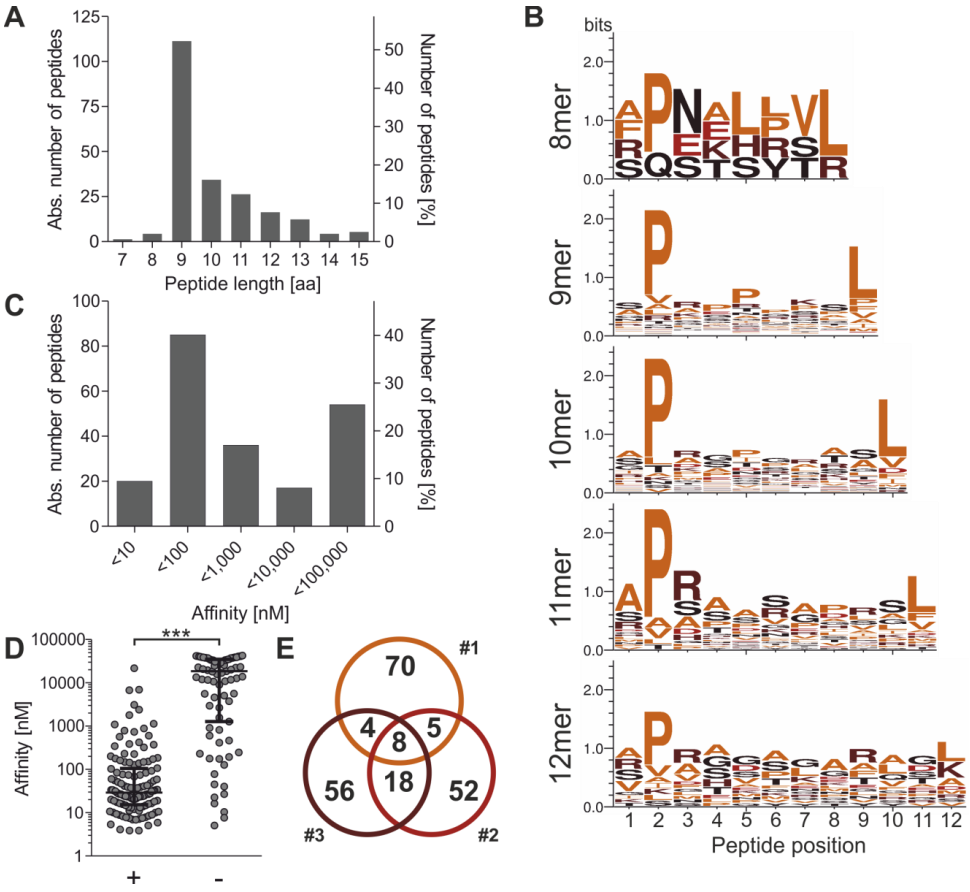


Figure 44: Identification of HLA B*07:02 restricted epitopes using LC-MS/MS *de novo* sequencing. (A) Length distribution of all 213 identified peptides given in absolute numbers (left y-axis) or percent (right y-axis). (B) Amino acid distribution of all eluted 8-12mer peptides pooled from all three independent experiments visualized as sequence logos (*WebLogo 3.4*). The height of symbols within the stack is scaled to the relative frequency of the corresponding amino acid at the indicated position. (C) *In silico* calculated (NetMHC 4.0) binding affinities to HLA B*07:02 for all identified epitopes, given as frequency distribution in absolute numbers (left y-axis) or percent (right y-axis). (D) Calculated binding affinities separated by sequences with matching HLA B*07:02 binding motif (+) and those without (-). Line and whiskers represent median with interquartile range. Significance analysis was performed with the non-parametric Mann-Whitney test. (E) Venn diagram of the numbers of all eluted peptides partitioned according to the experiment (#1-#3) in which they were identified. Numbers in overlapping areas symbolize sequences found in two, or all three experiments.

Although in no experiment any Gag epitope could be observed, a total number of 213 peptides sequences were identified (about 90 per experiment). Of those, 35 sequences were identified in at least two independent experiments (Figure 44 E). Analyzing the length distribution (Figure 44 A) of the epitopes showed that most epitopes (~50%) have a length of 9 amino acids, followed by lengths of 10 and 11, which is characteristic for CD8+ T cell epitopes. Epitopes of one HLA allele can be quite diverse, but the anchor positions at P2 (i.e. second amino acid) and PΩ (i.e. last amino acid) are highly conserved[298,299]. Therefore, the quality of sequenced epitopes can be validated based on their anchor positions. For HLA B*07:02 the amino acid at P2 is nearly always a proline. At PΩ, most nonpolar amino acids (L, F, M, A, I, and V) are tolerated[j]. 67% of the identified peptides had a proline at P2 and 79% one of the tolerated nonpolar amino acids at the last position. Combination of the appropriate residues at P2 and PΩ was observed for 61% of all peptides. Sequence logos of all 8 to 12mer epitopes, displaying the amino acid distribution, are shown in Figure 44 B. Based on the amino acid compositions of human proteins[301,302], the probability of observing a random peptide with the HLA B*07:02 binding motif would be only 2.1%. Hence the enrichment of HLA B*07:02 epitopes with the affinity chromatography procedure performed here was highly significant ($p<0.0001$), as determined in a Fisher's exact test, showing that these are indeed HLA B*07:02-restricted CD8+ T cell epitopes.

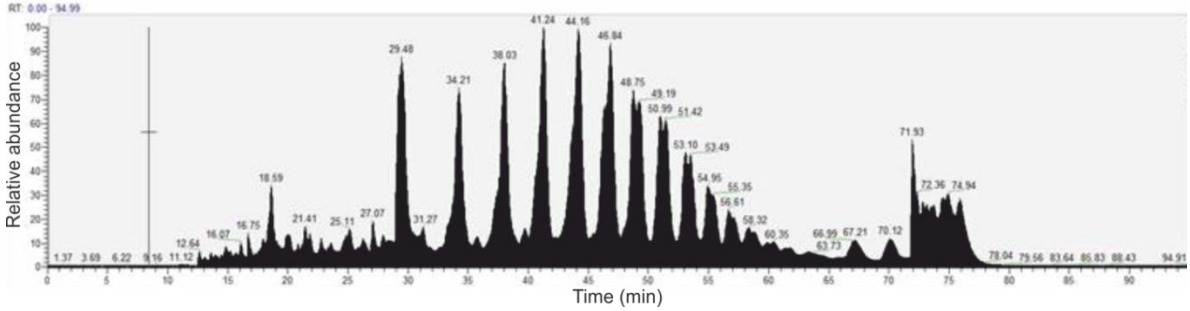


Figure 45. Total ion chromatograms of an LC-MS/MS *de novo* sequencing experiment. The TIC was created by summing up intensities (shown as relative abundance on the y-axis) of mass spectral peaks at every point (i.e. retention time as plotted on the x-axis) of the analysis for the same scan. The peptide signals are superimposed by PEG contamination signal, prohibiting sensitive epitope identification.

Using *NetMHC* to predict the binding affinities of all isolated peptides to HLA B*07:02, showed that 50% had a very high affinity (<100 nM) and the top 66% had at least a medium affinity (<1000 nM) (Figure 44 C). Analyzing only peptides with the correct HLA B*07:02 anchor positions, a median affinity of 29 nM was calculated. This was a distinctly ($p<0.0001$ as determined with Mann-Whitney test) lower affinity as for the remaining epitopes (with a median affinity of 18.5 µM), indicating that those peptides were most likely not HLA B*07:02-associated peptides (Figure 44 D). In conclusion, the LC-MS/MS results show that mainly HLA B*07:02 restricted epitopes were identified. However, the sensitivity has to be improved, since no Gag epitopes were found, although the CTL restimulation experiments had proven the presence of at least the p24-GL9 epitope in high concentration. The low sensitivity is likely due to the presence of impurities in the sample. The total ion chromatograms (TIC) displayed a polyethylenglycol (PEG) contamination that superimposes most peptide signals (Figure 45). Avoiding, or at least reducing, this contamination should result in a far higher amount of identifiable peptides and allow to retrieve Gag-specific epitopes.

[j] Immune epitope database[300] published HLA B*07:02 binding motif: http://www.iedb.org/MHCalleleId/251

E Discussion

Three decades after the discovery of HIV-1 as cause of AIDS, a protective prophylactic vaccine still remains elusive. Even encouraging advances, like the moderate and transient protection observed in the RV144 trial, are only first steps towards a worldwide solution. Nowadays, development of a protective vaccine mainly focuses on the elicitation of broadly neutralizing antibodies in vaccinees. There is, however, broad consent that efficient control of the HIV-1 epidemic relies on both, a strong humoral bNAb response to reduce infection rates as well as a robust and particularly a broad cellular immune response[154]. Cellular immunity, mainly mediated by CD8+ T cells, could control breakthrough infections and thereby improve disease prognosis and reduce transmission. All phase IIb/III HIV-1 vaccine trials so far failed to induce a broad cellular response, like in the RV144[46], where T cell responses were very low and no correlate of protection[190], or in the STEP study[189], where the breadth of CD8+ T cell responses was extremely limited, with a median of only one epitope per antigen protein[303]. These failures illustrate the lack of a vaccine strategy that elicits a greater breadth of high-quality (A.1.5.2) CD8+ T cell responses to account for the enormous variability of HIV-1.

The subject of this thesis was to address HIV-1's extraordinarily high sequence diversity by designing T cell epitope-enriched Gag antigens (teeGags) containing a maximized amount of naturally occurring CD8+ T cell epitopes (D.1). Since the formation of virus-like particles (VLPs) by Gag is considered favorable for the induction of immune responses, retaining this important function was also ensured through appropriate computational assessment. Functional predictions and antigen design features were confirmed by *in silico* evaluations (D.2), as well as through biochemical *in vitro* characterization (D.3). Additionally, to examine the enhanced breadth of the novel teeGags, a method to directly identify presented CD8+ T cell epitopes through mass spectrometry was established (D.4).

E.1 Designing global T cell epitope-enriched antigens

A potential protective HIV-1 vaccine would ideally be globally applicable. There are four different strategies described to achieve this[304]: (1) designing antigens tailored to each geographic region, based on the prevailing circulating strains, (2) eliciting bNAbs capable of neutralizing all HIV-1 subtypes, (3) focusing cellular immunity on highly conserved HIV-1 sequences, or (4) eliciting an HIV-1-specific immune response that is highly diverse, matching the viral variability.

The algorithm presented in this work is highly flexible, allowing design of antigens for a specific geographic region, by altering the epitope scoring weights or changing the input data set. The design can also be focused on conserved regions, by excluding undesired epitopes in variable regions, where viral escape is achieved more easily. For this thesis however, the proof of concept teeGags were designed to elicit a highly variable response, mimicking the global HIV-1 diversity as well as the worldwide population's HLA class I allele frequencies. The algorithm was furthermore implemented to preserve the functionality of Gag, because of the immunologically beneficial properties of VLPs (A.3.2). For instance, the uptake of a single VLP feeds about 2000

Gag proteins to the cell[148]. This is a crucial advantage of VLPs over soluble antigen, because it helps to overcome MHC processing thresholds, where normally only 0.1% of all processed peptides escape destruction by cytosolic peptidases and associate with MHC class I molecules. This implies that only proteins present in a copy number higher than 1,000 will be recognized by the immune system[70].

E.1.1 The patient-derived input data exhibits a strong B clade bias

The set of Gag-specific CD8+ T cell epitopes to initiate the *Optimizer Algorithm* was downloaded from the LANL HIV Immunology database and subsequently curated manually (D.1.1.2). Since the entry of the first CTL epitope in 1987, the database grew almost exponentially and at the end of 2015 harbored over 8300 epitopes spanning all HIV-1 proteins, an increase of around 500 entries as compared to the year before[210,305]. Although multiple entries might describe the same epitope and reduce the number of unique epitopes, it nonetheless demonstrates the impressive number of patient-derived epitopes that can be employed for the teeGag design. Compared to other designer-antigens (B.1.1.3) that maximize the number of all naturally occurring 9-mers (mos) or calculate artificial sequences that minimize the genetic distance in a given set of viral sequences (anc, con, and cot), employing experimentally validated epitopes to improve immunological breadth has several advantages. During the multi-step epitope processing (A.2.1), for example, multiple factors, like the specificities of the proteasome or immunoproteasome, ERAAP/ERAP1, tapasin, TAP, and finally the HLA allele influence which epitopes are presented[306]. The complete procedure is extremely selective and only 1-2 of every 10,000 peptides generated by the proteasome will eventually bind to MHC class I molecules[307]. By using patient-derived epitopes that are known to be presented on the surface of infected cells, these limitations in the predictability of epitope presentation might be circumvented. These epitopes also represent peptides of all lengths, whereas the mosaic antigens only maximize coverage of PTEs with a single pre-defined length. Additionally, since most epitopes in the database were identified by screening for reactive T cells, it ensures that a TCR recognizing the pMHC-complex with sufficient avidity can be among a given patients' TCR repertoire. As mentioned before (A.2.2.1), any TCR undergoes a positive and a negative selection process that eliminates all nonfunctional and all self-reactive receptors, respectively. The patient-derived epitopes should therefore only contain sequences that are recognized by TCRs that prevailed in this selection process and should therefore be able to elicit a CD8+ T cell response in at least a fraction of people.

A disadvantage, however, of employing the epitope set from the LANL database is that most were identified by stimulating T cells from HIV-1+ patients with synthetic peptides. These peptides are usually between 11 and 15 amino acids in length and hence longer than the actual epitope, which is in 95% a 9-mer[91]. An epitope entry does therefore not necessarily map the optimal epitope. To compensate for this, features from subepitopes were allocated to their respective superepitopes during teeGag design (B.1.2), thus refining epitope mapping by exploiting the wealth of database entries. Thereby, a superepitope represents all epitopes that are fully embedded within it. Synthetic peptides also entail the problem that their sequence does not have to resemble the natural antigen, but might be a so called mimotope. Mimotopes are similar, but not identical sequences to the actual T cell epitope, which are still recognized by the

TCR. However, this effect might even be beneficial, since such epitopes would increase the depth of the immune response, allowing coverage of variable sequences by a single epitope.

Although, the LANL HIV database publishes the best compilation of HIV-1 CD8+ T cell epitopes that were identified in a broad range of studies from the last 30 years, it is still updated manually and therefore somewhat error-prone. This was evident from the 39 epitope entries that had to be deleted and the over 60 that had to be modified due to errors or inaccuracies, which occurred during the transfer to the database or that were already erroneous in the respective primary publication. Through intensive manual curation, a more accurate epitope data set was generated for this thesis (D.1.1.2). However, due to its size of 2688 epitope entries, it is still probable that there are individual inaccuracies in the metadata, whereas all major defects, like indel mutations or erroneous start-end annotations, were reliably removed. Additionally, although the LANL-published epitopes are a conglomerate from studies all over the world, the subtypes associated with the epitopes do not match the global clade frequencies. Subtype B isolates, for example, are only responsible for 11% of all global infections (Figure 18), but are largely overrepresented in the epitope list (30% of all registered epitope subtypes, Figure 26 D). This B clade bias is even more obvious for the filtered Gag sequence alignment that was used (D.1.1.1), where 38% are registered as subtype B sequences. This overrepresentation is probably due to the fact that subtype B is prevalent in Europe and Northern America (Figure 1), where most of the research has been performed.

The alignment of all Gag-specific CD8+ T cell epitopes to the HXB2-Gag reference sequence revealed 158 AAS at 128 variable positions (D.1.1.2). 372 positions were consequently conserved, meaning no AAS was detected. Since the epitopes are derived from all major HIV-1 subtypes, the number of variable positions was expected to be higher. In an alignment of 12,543 Gag sequences (including subtypes A1, B; C, D, F1, G, CRF01_AE, and CRF02_AG) against HXB2, Li et al.[188] detected 258 polymorphic positons and far over 600 AAS. The large difference between natural Gag polymorphism and the one observed in the epitope set might have several reasons: (1) the epitope set does not feature all subtypes equally and mainly B-clade-specific AAS are found. (2) Epitopes are determined by restimulation of patient's T cell repertoire with synthetic peptides. As these peptides are often based on consensus sequences or specific isolates (e.g. HXB2 for B clade sequences), not all polymorphic positions can be identified, either because the synthetic peptide does not restimulate the T cell reactive against the polymorphic epitope or it is a mimotope and can therefore not be differentiated. (3) Mutations might alter processing and presentation and certain polymorphic peptides are hence not presented. (4) Some TCRs, reactive against polymorphic epitopes do not pass through thymic selection. This shows that the quality of breadth-enhanced antigens, designed by the *Optimizer Algorithm*, is highly depending on the quality of the input data set. The steep annually increase of published epitopes (see above) therefore helps to increase the functionality of *Optimizer Algorithm* gradually.

E.1.2 Structure/sequence combination for best classification

To exclude integration of AAS in the teeGags that are harmful to VLP budding, a supervised, multidimensional FLD classifier was conceived (B.2.1). Four structural and one sequence-based feature were tested to discriminate between budding-retaining and budding-deleterious AAS. Since particle assembly and release is driven by the full-length Gag, a 3D-structure of the p55

Gag precursor protein (A.3.1) would be preferable for calculation of structure-based classifier features. Because such a structure was not available, domain-models were applied (B.2.1.1), which is likely an acceptable strategy, since the domains are mostly independent and behave like beads on a string[308].

The best feature combination for the FLD classifier (D.1.2.1), determined by 10-fold cross-validation (Extended Data Table 4), was $f_{3,\Delta DOPE\text{-}global}$ together with $f_{5,seq}$, which exhibited a prediction precision of 98%. Even for teeGag3, that with 40 AAS has the most combined mutations of all teeGags, this high precision would still entail a 45% probability of realization (i.e. 45% chance that none of the AAS is a false positive, see Figure 21). The calibration precision with 100% was even higher and would mean that the classification boundary is set in a way that it prohibits false positives (Figure 20). Any additional feature reduced the precision. This is especially evident for the combination of all five features that only exhibited a precision of 83%. For teeGag1, with only 28 AAS this low precision would result in a probability of realization of less than 1%. In other words, there would be an over 99% chance that at least one AAS was classified as false positive, which would possibly reduce or prohibit VLP budding and highlights the importance of choosing the classifier features carefully.

Since $f_{1,DOPE}$ on its own was not able to efficiently separate the training-set data (p=0.56, Figure 19), it can be excluded as suitable classifier feature in general. Interestingly, for the tested feature combinations, $f_{5,seq}$ was nearly always required to achieve a satisfactory prediction precision. This might be due to the fact that $f_{2,\Delta DOPE}$, $f_{3,\Delta DOPE\text{-}global}$, and $f_{4,\Delta DOPE\text{-}local}$ are all based on energy-changes between mutated and reference 3D-structures and might therefore be partially redundant (B.2.1.1). $f_{5,seq}$, by contrast, represents the sequence conservation on the respective AAS positon, addressing a completely different facet (B.2.1.2). Combining the sequence feature with any energy-function-based feature (except $f_{1,DOPE}$) improves classification by adducting two different aspects of an AAS introduced into the reference sequence (Extended Data Table 4).

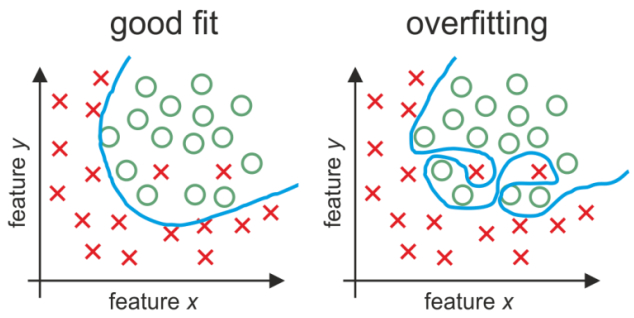


Figure 46. Schematic representation for a good fit and for overfitting of a classifier. The graphs depicts training-set entries of two different groups (red crosses or green circles) that are plotted according to their discriminatory features *x* and *y*. For a good fit (left) the training-set is used to calculate a discriminatory boundary (in blue) that separates both classes in the training-set, but is also able to classify new unknown data efficiently. Through overfitting of the classifier (right) the boundary is modeled too close to the training-set, which reduces the potential to group new data correctly.

However, this does not explain why $f_{2,\Delta DOPE}$ or $f_{4,\Delta DOPE\text{-}local}$, which both performed best on their own (p<0.0001 each, Figure 19), had negative effects on accuracy and precision, when added in a multidimensional FLD to $f_{3,\Delta DOPE\text{-}global}$ and $f_{5,seq}$. It seems that the addition of these features resulted in an overfitting of the classifier. Overfitting refers to a classification boundary that is modeled too close to the training-set data, by providing too many adjustable parameters. Thereby, random fluctuations in the training data are picked up and learned as concepts for the

classifier. These tightly fit classifier concepts might then not apply to new data (i.e. unknown AAS) and negatively impact the classifier's capacity to distinguish between the two groups[309] (Figure 46). As observed here, overfitting in FLD is described to show up in unstable results, especially during cross-validation[310] and can be circumvented by reduction to the essential features, here $f_{3,\Delta DOPE\text{-}global}$ and $f_{5,seq}$.

A limitation of determining functional conservation only through assessing single amino acid substitutions is that it cannot predict the effects of combining mutations, as was later on done during the generation of teeGags. Combination of two or more as positive predicted AAS might result in steric hindrance or other detrimental interactions, which could result in reduced or abolished formation of VLPs. Yet, the other way round is also conceivable: a negative effect of an as budding-deleterious predicted AAS might thereby get rescued through a second compensatory mutation. Addressing the consequences of such interactions *in silico* is difficult, since a suitable training-set to supervise an algorithm is not available and, in addition, the number of possible AAS-combinations rapidly outgrows the feasible number. For p24 with 60 naturally occurring AAS for example, the number of all combinations is 1.15×10^{18}. To compute the structure and energy profile of each combination is hence not realizable. But, since AAS that are located on different domains are spatially apart and the fact that all mutations are derived from natural, probably viable isolates, it seems reasonable to assume that combinations of budding-competent mutations should not alter the Gag functionality severely.

Interestingly, the percentage of as budding-deleterious predicted AAS was far higher for the permutation set (53%, D.1.2.3), than observed for the natural AAS identified in the epitope set (11%, D.1.2.2). This shows –as expected– that the natural CD8+ T cell epitopes are significantly depleted of budding-incompatible AAS ($p < 0.0001$ in Fisher's exact test). Of all Gag domains, by far the highest percentage of as budding-deleterious predicted AAS (83.6%) were observed for p24, the most conserved protein domain[188]. This supports the validity of the classifier, since highly conserved positions are most probably vital for functional Gag and mutations would more likely result in a defective protein. It is also in line with observations that p24 exhibits an enormous genetic fragility and tolerates only few substitutions[311]. Although all training-set data was based on AAS with known effects located in p17 (Table 4), the classification of unknown mutations in p24 or p2p7p1p6 still was very good, as the goal of avoiding false positives was reached. This might be due to the fact that $f_{5,seq}$ computes the position-specific conservation, irrespective of the protein. With an expanded set of validated AAS covering the complete Gag protein, future classifications might even get improved.

E.1.3 Flexible design of breadth-enhanced Gag antigens

Besides structural conservation of the Gag protein, the main focus for teeGag design was to combine as many immunologically-potent CD8+ T cells in as few antigens as possible. The quality-assessment of epitopes is very flexible and can include any number of epitope metadata characteristics (B.2.2). Only HLA- (B.2.2.1) and subtype score (B.2.2.2) were predefined, because they require an additional datasets containing frequencies of HLA allele groups and subtypes, respectively, for calculations, which makes it possible to adopt antigens to a specific target region. Both can also be excluded for score calculations, by assigning them a zero weight. A third predefined scoring parameter that is implemented in the *Optimizer Algorithm,* is that a list

of epitopes can be specified that get an additional scoring advantage. For the proof of concept teeGags this attribute was, however, not used.

The teeGags were designed as a globally applicable set of antigens. Through using the worldwide subtype frequencies (D.1.1.4), it was possible to reduce the effect of the B clade bias, which is evident in Figure 29 A: All depicted natural sequences that include high numbers of CD8+ T cell epitopes are derived from B clade isolates (Figure 29 A - left). Through assigning a lower subtype score to B clade epitopes than for example to those from clade C, the natural sequences are more balanced with regard to their antigen score (Figure 29 A - right). By choosing a more selective input epitope set or through different subtype attribute weighting, this B clade bias might be further reduced.

The teeGag sequences that were finally generated (D.1.4), already covered 87.8% of all epitopes and 91.0% of the maximally possible score within the first three iterations (teeGag1-3, Figure 27), making further antigens far less beneficial. This also meets the objective to restrict possible vaccine candidates to a maximum of three components[154]. For teeGag1, all but 6 amino acid positions were covered by at least one epitope. The non-covered positions were “filled” with the reference HXB2 Gag sequence. For each sequence generated in the following *Optimizer Algorithm* iterations, the fraction of the sequence that was covered got smaller, and more of the antigen was therefore completed with the reference. Filling the blanks with the respective amino acids from the reference sequence most likely maintains budding, but for future applications it might be advantageous to fill non-covered areas with compatible high-scoring epitopes. For the proof of concept, this was not important, since most high-scoring epitopes are B-clade-derived and are already integrated in HXB2 (Extended Data Table 5). It might however be unfavorable, for example, to use a subtype B sequence to fill non-covered positions, when designing a region-specific C clade antigen.

E.1.4 *In silico* validations highlight the reliability of the algorithm, but also the B clade bias in the input data

Through phylogenetic tree reconstruction, the teeGags and other antigen designs could be classified with regard to a preselected HIV-1 subtype reference set (Figure 28). Due to the B clade bias in the epitope data set and the fact that HXB2-Gag was applied as reference and to fill up non-covered areas, teeGag1-3 all cluster at the subtype B branch of the tree. As expected, conM and ancM, both designed to minimize the genetic distance of all HIV-1 subtypes, are located near the center of the tree. Surprisingly, this was not the case for cotM which was also classified as a B clade sequence. In addition, mosM that was designed to cover all naturally occurring Gag 9-mers was also not located at the center of the tree, but at the branch of C clade sequences. This might be due to the fact that the coverage calculation of the mosaic approach was based on a reference alignment that overrepresented C clades (64.4% subtype C)[199].

In further *in silico* analyses regarding antigen coverage (B.3.2) teeGag1 and teeGag1-3 were far superior to any monovalent or trivalent set of natural Gag sequences or designer antigens based on mos, anc, con, or cot approaches, respectively (Figure 29 and Extended Data Table 5). The best benchmarks to compare the teeGags are the mos sequences, since this is the only other approach that focuses on including as many CD8+ T cell epitopes as possible. For mosaic antigens, however, the number of naturally occurring 9-mers, as a surrogate for PTEs, was maximized[199], instead of patient-derived epitopes. Since the computational analysis programs

(i.e. antigen score, population-, and pathogen coverage) used the set of patient-derived epitopes to calculate the number and score of the integrated epitopes, mos is mostly clearly inferior as compared to the teeGags. Regarding the population (Figure 30 and Extended Data Figure 1), computed based on the global HLA allele frequencies (D.1.1.3), teeGags were far better than any of the other examined Gag sequences, including mos. This was expected, since teeGags were specifically optimized to mainly included epitopes presented on HLA alleles that had a high population frequency. In summary, the *Optimizer Algorithm* is able to design the best possible antigens, regarding antigen score and population coverage. Since the target populations' HLA allele frequencies and the attributes for antigen scores are customizable, the algorithm can design antigens for specific target populations that incorporate the best combinations of high-scoring epitopes.

Regarding pathogen coverage, teeGags were better than naturally occurring sequences, when using the global subtype frequencies for pathogen selection (Figure 31 and Extended Data Figure 2). However, for the monovalent sets, other antigen designs, especially mosM1.1, were slightly superior. This again is allegeable by the B clade bias in the teeGags. Using other distribution parameters for pathogen coverage, like B clades only, or subtype frequencies based on the occurrence in the epitope database, teeGag1 is again superior to all other antigens (Figure 32).

With the mosaic suite evaluation tools *Epicover* and *Posicover* (D.2.5), which calculate the covered 9-mers from a set of sequences by a specific antigen, no differences were observed for all monovalent antigen design approaches. This indicates that teeGag1 covers the PTEs similar well as mosM1.1, but is far superior in antigen- and population-coverage, which makes teeGag1 the preferable antigen in a monovalent vaccine regimen. Regarding trivalent antigen sets, the mosaic approach was best, which was expected, since mos3.1-3 was especially contrived for maximum coverage of M-group-derived Gag 9-mers. The disadvantage of considering all potential 9-mers as a surrogate for CD8+ T cell epitopes is that not all of these polypeptides will be presented on the surface of a given cell or prime a T cell response.

E.2 Experimental characterization of VLPs

E.2.1 Experimental classification of single AAS mutations

The quality of the FLD classifier was validated experimentally by determining the relative budding (RB) of all single AAS that were incorporated into teeGag1-3 and all as budding-deleterious predicted AAS (D.3.1). The mean RB of the latter was significantly reduced compared to the mean of all as budding-retaining predicted mutations (Figure 34), which demonstrates the effectiveness of the FLD classifier of the *Optimizer Algorithm*. The predefined objective to minimize the rate of false positives by maximizing the precision of the classifier was also met, since all as budding-retaining predicted mutations were indeed still compatible with budding, when included in the reference sequence. Of the budding-deleterious-predicted AAS, half were false negatives (9 of 18). This rate is higher than expected from the 10-fold cross-validation (Extended Data Table 4), but acceptable, as long as the main criterion of strictly no false positives is met. The experimental validation also highlights that the natural set of epitopes

is even more selective for mutations preserving structure and functions as estimated through *in silico* classification.

Although none of the positively-predicted AAS (Extended Data Table 6) did cause any statistically significant reduction of RB, some had special features that were part already addressed in the literature: mutations R15S, K28G, K30R, and K30M are all located in the membrane targeting domain[23] and substitute rather conserved basic amino acids, involved in binding to the inner leaflet of the plasma membrane (A.3.1). Interestingly, none of the mutations was classified as budding-deleterious or exhibited significantly decreased RB. However, K30M, which removes a basic residue and replaces it with a nonpolar residue, had only an RB of 0.65. K30R at the same position, which preserves the basic residue, had a distinctly higher relative budding of 1.27. Combined removal of the basic residues at position 30 and 32 (K30T and K32T) was described to result in reduced particle production[312]. The remaining as budding-retaining predicted AAS are mostly in good accordance with published data. For 34 of these mutations, information was found that stated that alterations at those sites do not alter particle release[212,311,313–324]. Conflicting data was observed for positions 268[315,319,325] and 362[321,322], where different alterations had variable effects on budding. However, none of the mutations was exactly the same as the one in the epitope set. Another 13 as budding-retaining predicted AAS were directly analyzed in various publications and none had negative effects on budding[311,313–316,322] (R91K[313], I138L[314], A146S[311], I147L[313,314], T190I[316], E203D[314], I223V[311], G248A[315], R286K[313,314], G357S[313,314], V362I[314], S373A[322], R384K[313]). This provides further validation for correct predictions and experimental assessment in this thesis. Nevertheless, some identical mutations were associated with reduced particle production. Moderate reduction was for example observed for T186M, I256V, and E260D[314]. For the latter two also a slightly reduced, but statistically not significant RB was observed in this thesis (0.85 and 0.77, respectively). Three further mutations -T190I[314], K335R[314], and V370A[322]- are described to have a stronger negative effect on budding. For T190I, the information is contradicting and in another publication, this mutation exhibits wildtype like particle release[316]. Similar for V370A, that showed reduced relative budding of 0.8 in this work, deletion of the same amino acid in another publication had no effect on budding[326]. Questionable mutations, with contradicting literature or non-significantly reduced RB, should be further characterized.

The budding-deleterious-predicted AAS (Extended Data Table 7) can be divided into two groups: experimentally validated true negatives and false negatives. The first false negative (S9R) is located in the membrane targeting domain[23], and although the relative budding was slightly reduced to 0.83, the effect was not significant and therefore was rejected as a correctly predicted budding-deleterious mutation. Negatively predicted mutations Q199E, N271H, and N271K all had wildtype-like or even enhanced particle production. In line with this, mutations at these positions were shown to have no effect on budding in previous works, but are accompanied by reduced infectivity of the virus[315]. Since the positions might be important for later steps in viral replication, and were probably identified as highly conserved positions by the sequence-based feature $f_{5,seq}$ (B.2.1.2) of the classifier, this could be the reason, why they were predicted as budding-deleterious. W316M was the only false negatively classified AAS located on a position that is described to negatively influence budding. Here, substitutions to alanine were described to prohibit particle production[315,327] and mutation to leucine or arginine resulted in replication-deficient viruses[311]. In this thesis, the AAS W316M was compatible Gag with functionality with an RB of 1.29.

True negative AAS were infrequent in the natural set of epitopes and seem to be derived from defective viruses that nevertheless could prime a T cell response or from a virus containing a Gag variant that harbored compensatory mutations to account for the fitness defect. The first true negative AAS was located in the NTD of p24. This T151L mutation showed moderate, but statistically significant reduced relative budding (RB=0.58). This result is contradicting other literature were the T151L was described to be tolerated for virus replication[313]. However, in that work, particle release was only addressed indirectly by assessing replication capacity in a viral mutation selection progress. Positions 267 to 273 were a hotspot with 7 as budding-deleterious predicted AAS. The high number of AAS was mainly due to one epitope (Record number: 54502; Epitope: KRWILGLNKIV), which had an isoleucine deletion compared to the HXB2 reference (KRWI**I**LGLNKIV), prohibiting a gap-free alignment. Of the 7 AAS, three were experimentally validated to be true-negatives and for G269L and L270N, budding was even completely abolished. Literature provides evidence that these positions are important, because an L268P mutation prohibits particle release[325] and mutations at position 296 and 270 result in non-viable viruses[311]. Next, R294I and Y301W were two validated budding-deleterious AAS, which are located within the highly conserved major homology region (MHR) that is necessary for assembly[138]. Although the mutations do not affect the MHR consensus sequence from orthoretroviruses[328], and no negative effects on budding are reported for both positions in the literature, it is conceivable that mutations within the MHR are detrimental. AAS Q311A was previously described to result in wt-like particle production[317], but had a moderately reduced RB (0.73) in this work. This is probably explainable by different assays used to quantify particles. Additionally, the experimental validation presented here is one of few accounting for transfection efficiency (through SEAP normalization), therefore probably able to detect more subtle differences during budding.

The last two truly budding-deleterious AAS, C392F and C395G, are located in p7 within the proximal zinc-finger-like domain, necessary to recruit viral genomic RNA to the particle assembly site[139]. In natural HIV-1 infections, interactions between NC and gRNA facilitate assembly, by promoting Gag multimerization[146]. Since VLP assembly remains efficient in the absence of genomic RNA it seems that host cell RNAs can fulfill this function[329]. It is therefore conceivable that mutations in the zinc-finger-like domains prohibit RNA binding and consequently reduce particle formation. This is especially true for C392F and C395G, since both mutate cysteines important for the zinc finger interactions (Figure 47). It was already shown elsewhere that mutations at these positions reduce particle production[330,331].

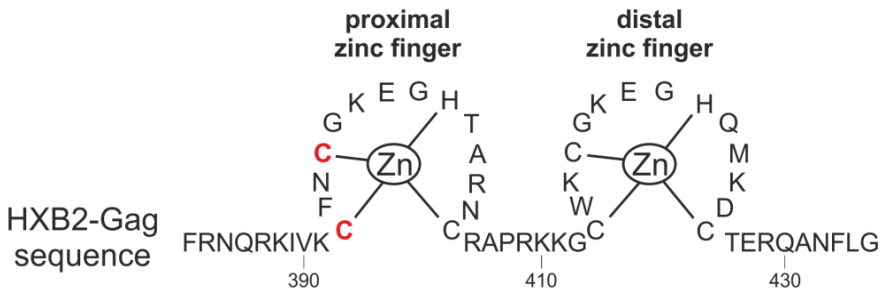


Figure 47. Schematic representation of the proximal and distal zinc finger, which that are located in p7 of Gag. Important interaction sites for the zinc fingers are indicated. The positions of the two validated mutations (C392F and C395G) with negative influence on budding are highlighted in red.

Interestingly, all AAS located in p6 were predicted as budding-retaining, which was experimentally validated. This might be due to the fact that p6 is largely unstructured[140] and, besides recruiting the ESCRT machinery through its late domains to facilitate particle release, seems to play a subordinate role during VLP production[148]. The AAS from the natural epitope set that were validated experimentally in this work could also be used to refine the training-set of the classifier. To further improve the quality of the classifier, the present training-set AAS might be also re-evaluated regarding RB, because the mutations were originally identified in virions rather than VLPs and with the reference sequence NL4-3 instead of HXB2[212].

E.2.2 Experimental characterization of teeGag particles

E.2.2.1 Biochemical characterization showed reference-like behavior for teeGag1 and teeGag3, but altered characteristics for teeGag2

After assessing the AAS individually, with regard to budding competence, their combination into the newly designed teeGags demanded further biochemical characterization (D.3.2). Since the sum of up to 40 AAS could probably alter antibody binding in the Gag-ELISA the RB was determined by adding a C-terminal 6x-His epitope tag to the Gag variants. This tag was used to assess the Gag budding densitometrically, without employing any Gag-specific antibodies (Figure 35). For teeGag1 and teeGag3, RB comparable to the reference HXB2-Gag was observed. In these cases, the assumption that single budding-competent AAS can be combined into a functionally intact Gag holds true. For teeGag2 on the other hand RB was moderately, but significantly reduced. This might be explained by a combinatory effect of any of five AAS (E260D, V370A, Q386S, T401I, and T427N) that were included in teeGag2 and displayed slightly reduced RB as separate mutations compared to the reference. Of course, other non-prominent AAS combinations could potentially have negative effects on budding as well. However, with 0.78, the RB of teeGag2 was high enough to efficiently produce VLPs in mammalian cells.

The special status of teeGag2 was confirmed during particle size analysis. Firstly, during the sucrose gradient ultracentrifugation (Figure 36) the main peak for teeGag1, teeGag3, and HXB2-Gag was observed at a density of 1.16 g/ml, which is in good accordance with literature on the characterization of HIV-1 Gag VLPs produced in mammalian cells [332,333]. teeGag2 on the other hand was identified at a slightly higher density of 1.18g/ml, which is nonetheless still in the range of typical VLPs[333]. As expected, no Gag particles were observed for the budding-deficient HXB2-Gag^{Myr-}, neither with the densitometrical measurement, nor after the gradient centrifugation

Electron microscopy imaging confirmed the uniformity of teeGag1, teeGag3, and HXB2-Gag which all had a median-diameter of about 100 nm. Again, teeGag2 was the odd one out with an increased median diameter of 145 nm (Figure 37 C). All constructs are nevertheless again in the normal range observed for HIV-1 Gag VLPs of 100-150 nm[148]. Whereas the increased median size of teeGag2 VLPs would not be too unusual, the infrequent, but distinct occurrence of aberrant, tubular particles is surprising (Figure 38). These tubular structures had a length of up to 2 µm and might be due to a set of four mutations within the p2 spacer (V370A, T371N, N372S, and S373A). It was proposed that the C-terminus of CA and a conserved N-terminal region of p2 together form an α-helix[322,334]. This CA-p2 boundary seems to play an active role during particle production and is thought to be essential for the induction of curvature. Alterations

in the p2 spacer of HIV-1 are therefore associated with particles that are very heterogeneous in size and shape, larger than normal, and often tubular[326,334,335]. To ensure that these teeGag2 AAS are responsible for the aberrant, tubular VLPs they should be analyzed in more detail. Especially mutation V370A could be of interest, since it is described to be associated with reduced budding[322] and alterations at this position are associated with replication-incompetent virions[313], although herein, it had a non-significantly decreased RB of 0.8.

Particle comparison by dynamic light scattering analysis resulted in generally larger particle diameters than those determined from analysis of the EM images (Figure 39). This is due to the fact that DLS does not measure the particle diameter directly, but rather the hydrodynamic diameter, which is the size of the particle in solution. The hydrodynamic diameter does therefore depend, besides the size of the particle, also on surface structures, shape, and the type of solvent. Nevertheless, teeGag2 again exhibited the largest mean diameter of all VLPs, correlating well with the EM analysis. The slightly enlarged size of teeGag3 compared to the reference was unexpected and could be addressed in further experiments, but since the direct analysis of electron microscopy images showed no increased diameter, this effect might be negligible.

Presence of microvesicles and exosomes are sometimes a problematic contamination in VLP preparations[333]. Since the control VLP preparations from cells transfected with the budding-deficient HXB2-Gag^{Myr-} or a control plasmid did not contain any pellet-able particles, interfering impurities of exosomes and microvesicles are unlikely.

E.2.2.2 Preliminary immunological characterization

As observed by Tanja Stief[267], HIV-1 Gag VLPs had to be pseudotyped with VSV-G for efficient uptake, processing, and presentation by mdDCs. With incorporated Env or without any viral surface protein, the VLPs were not presented efficiently enough to restimulate a Gag-specific CD8+ T cell clone. Since different DC subpopulations exhibit variable potential to cross-present antigens[80], it is possible that mdDCs are not the ideal cell type for cross-presentation studies, especially since other DC subpopulations are more likely to encounter antigens in the context of intramuscular vaccination. Nevertheless, it was previously shown that Gag-VLPs can activate mdDCs and induce *ex vivo* T-cell responses[166], but these VLPs were produced in insect cells using the baculovirus system. There, efficient uptake might be facilitated by baculovirus glycoprotein gp64, which gets incorporated into the VLPs and can mediate direct fusion, followed by classical MHC class I presentation[148]. It is therefore possible, that non-specific uptake of mammalian cell-derived naked or fusion-deficient viral particles by phagocytosis or micropinocytosis does, in *ex vivo* experiments, not result in detectable epitope presentation[336]. Of course, *in vivo* with many different APC populations, VLP formation of course might still be beneficial, as indicated for DNA- or viral-vector-delivered budding-competent Gag[176–178].

To analyze and compare the immunological potential of the novel teeGags to HXB2-Gag *ex vivo*, the VLPs were thus pseudotyped with VSV-G. The use of the widely expressed LDL receptor as major entry port of VSV and also of VSV-G-pseudotyped vectors confers a broad cell tropism[337]. This allowed efficient transduction of mdDCs and subsequent presentation of epitopes on the mdDCs-surface. However, since VSV-G induces membrane fusion, the antigen directly gained access to the cytoplasm for classical presentation on MHC class I, rather than the desired cross-presentation. For initial comparisons, this system was yet deemed sufficient. All teeGag- and HXB2-Gag-dervied, VSV-G-pseudotyped VLPs were immunogenic, able to be

efficiently presented by mdDCs as measured by restimulation of a Gag-specific CD8+ T cell clone (Figure 40 C). There were, however, subtle differences between the different Gag VLPs: teeGag3 VLPs were superior to all others regarding restimulation. By contrast, teeGag2 VLPs had a slightly reduced restimulatory capacity. Since the epitope sequence targeted by the readout CD8+ T cell clone was harmonized for all VLPs, these differences cannot be due to different TCR reactivity or MHC class I binding, but must be because of differences in uptake or processing. For teeGag2, the altered size and the aberrant, tubular particles might be responsible for the slightly reduced restimulation rate. Further experiments are needed to strengthen these initial observations and maybe *in vivo* experiments of naked, mammalian-cell-derived VLPs should be performed to determine their immunological potential for cross-presentation. Nevertheless, the key conclusion of these experiments is that all tested VLPs are similarly immunogenic, which further highlights functional conservation of the teeGags.

E.2.2.3 HIV-1 virions with teeGags

The problem for comparing teeGags to the HXB2-Gag reference included in molecular clones, is that mutations introduced in teeGag1-3 could have unknown influence on the ribosomal frameshift (A.1.3) or the C-terminally overlapping Pol. To circumvent this problem pNL4-3-AL was used as backbone to incorporate the different Gag variants, since the *gag* and *pol* reading frame were transcriptionally uncoupled in this molecular clone[251].

Analyzing particle release, only teeGag1 produced wildtype-like virions. teeGag2 and teeGag3, in contrast, did not produce any measurable virions (Figure 41 B-D). Since the proteins alone efficiently produce VLPs (D.3.2.1), the inability of teeGag2 and teeGag3 to produce virions is probably established before assembly and release. This is possibly due to the computational reverse translation of teeGag genes that was done based on a HIV-1-Gag-specifc-codon table (B.4.2) of teeGag genes. Yet, important influences on viral replication, like unknown RNA secondary structures and motifs that affect transcription and RNA export dynamics, as well as translation cannot be taken into account comprehensively during sequence design. The next step should be to check for intracellularly expressed Gag, to exclude effects of the teeGags on assembly and release. If no Gag is synthesized in the transfected cells, the problem probably occurs before particle production and should be analyzed through analysis of RNA levels in nucleus and cytoplasm.

Preserved virus replication of molecular clones including teeGags is however subordinate for their use as vaccine candidates and the observation that all human-codon-optimized teeGags on their own efficiently produce VLPs (Figure 35) is more important.

E.3 Next generation of antigens optimized for breadth

As outlined above, it was shown experimentally, that the novel teeGags are still able to form and release VLPs. Moreover, *in silico* analyses confirmed that teeGags include a broad range of potent CD8+ T cell epitopes. Nevertheless, there are some features that could improve the teeGags and also some that could be addressed in a next generation Gag antigens regarding the potentially elicited breadth.

The already existing teeGag could be improved by combining them with potent, bNAb-inducing Env antigens, incorporated into the VLP surface. As a vaccine, this approach would combine a

potent humoral immune response that is possibly able to protect from HIV-1 infection with a broad Gag-specific CD8+ T cell response that might facilitate control of breakthrough infections. teeGag1 was already used to generate VLPs presenting various Env variants, which exhibited a similar binding affinity hierarchy towards bNAbs as the respective soluble Env variants (Benjamin Zimmer, personal communication). Additionally, as previously addressed, CD8+ T cell priming can only be performed by APCs, mainly DCs, since they express the necessary co-receptors. It could therefore be advantageous to include molecules into the VLPs that target them to surface molecules of desired DC subsets. This could enhance CD8+ T cell responses, as has been shown for a broad range of DC molecules like DEC-205, DCIR2, mannose receptor, Dectin-1, and scavenger receptor CD36[338]. CD40L was already incorporated into Gag-VLPs to address the CD40 receptor on DCs and was able to activate DCs *in vitro*, as well as inducing immune responses against Gag in mice[339]. As discussed, mammalian-cell-derived VLPs might not be immunogenic enough to prime T cells and would therefore lead to peripheral tolerance and apoptosis of reactive T cells. For an efficient immune response, the use an adjuvant may be needed. This could be achieved by including immune-activating molecules into the VLP membrane, as was accidentally the case for baculovirus-derived VLPs (E.2.2.2). The immune-stimulatory molecule flagellin[340] and the GM-CSF and IL-4 fusion-protein GIFT4[341] were already anchored in VLP membranes, which led to enhanced immunogenicity and increased immune responses. For expression of VLPs from naked DNA plasmids or viral vectors encoding human-codon-optimized *gag* genes, the intrinsic immune-activating properties of both vector systems could already suffice to prime adequate responses. Motifs, like the PEST sequences[342], that enhance degradation could also added to Gag antigens to improve presentation. In this context, the FLD classifier could be applied to identify positions where insertions of such sequences can safely be made without compromising budding.

For a completely new generation of breadth-optimized, functionally-conserved Gag antigens there are some alterations that could be considered. For one, not all regions of Gag are equally associated with protective cellular immune responses[343]. Selective depletion of known CD8+ T cell epitopes in the non-protective regions, for example by mutating the epitope anchor positions that are necessary for HLA binding, could achieve a benefit, since cellular immune response might focus on protective regions. The FLD classifier could again ensure that such introduced mutations do not alter Gag functionality.

Another idea would be to focus CD8+ T cell responses on HLA-C- and HLA-E-restricted epitopes. Since both are not downregulated during HIV-1 infection, they could be a preferable target to control the virus. In NHPs it was shown that MHC-E restricted CD8+ T cell responses do not get primed during natural SIV infection, but vaccine-elicited MHC-E-restricted CD8+ T cells did recognize SIV-infected cells[162]. Moreover, because of limited MHC-E polymorphism, a vaccine targeted at MHC-E–restricted CD8+ T cell responses would elicit largely similar responses in all or most vaccinees, potentially providing efficient protection regardless of MHC-I genotype[162].

Additionally, since priming of CD8+ T cells normally requires additional help by CD4+ T cells that recognize related antigens on the same APC, ensuring efficient presentation of CD4 T cell epitopes could be beneficial for a cellular immunity-focused vaccine. The large set of CD4+ T cell epitopes from the LANL database (currently about 400 entries) could thus be combined with the CD8+ T cell epitopes from this work to generate a broad and balanced cellular response. In this scenario, preserved Gag functionality would also have the advantage that Gag can bud from

cells, when delivered via DNA or viral vectors. These exogenous particles can then be efficiently presented through the classical pathway on MHC class II molecules to mediate CD4+ T cell responses. Strong CD4+ T cell responses could additionally, through intrastructural help, improve humoral immune responses for VLPs that present Env on their surface. Gag-specific CD4+ T cell responses can thereby provide cognate help for Env-specific B cells, which would also facilitate induction of antibody responses[179].

E.4 Assessing the immunological breadth of antigens

E.4.1 LC-MS/MS peptide sequencing

To characterize the immunological breadth of antigens, a method to directly identify the HLA class I presented CD8+ T cell epitopes was conceived.

. For this, cells that express soluble variants of HLA class I alleles were transfected with the antigen of interest. Although maturation kinetics for sHLAs are slightly delayed, indicating minor difficulties to access the PLC, it does not significantly impact the presented immunopeptidome and the peptide yield is even increased compared to peptide purification from membrane-bound HLAs[110,111]. Additionally, the detergent used to solubilize the membrane-bound HLAs and the consequently arising cell debris hampers subsequent epitope identification. Therefore, and due to the fact that use of sHLAs allows examination of one specific allele and not the cells' complete haplotype, sHLAs were favored over membrane-bound HLAs.

Peptide-harboring sHLA complexes were isolated from the conditioned medium. It was shown in this thesis that it is possible to purify Gag-specific CD8+ T cell epitopes restricted to individual HLAs using this method, since only peptides from cells expressing sHLA-B*07:02 and Gag were able to restimulate an HLA B*07:02-restricted Gag-specific CTL clone (Figure 42 B). By quantifying the amount of isolated GL9 peptide it was verified that, in theory, enough peptides can be purified to allow identification by mass spectrometry (Figure 43 C). The maximum purification of 89 µg of sHLA protein should be enough for analyzing the immunopeptidome, since already a few micrograms of isolated sHLA molecules are sufficient for identification of hundreds to thousands of sHLA bound peptides[344].

Initial epitope sequencing of sB07-restricted epitopes by LC-MS/MS, however, led only to the identification of few peptide sequences (~90, Figure 44 E) per experiment. These sequences had a length distribution that is characteristic for CD8+ T cell epitopes (Figure 44 A) and most had a binding motif typical for HLA-B*07:02 (Figure 44 B). Although cells were transfected with Gag expressing plasmids, all identified epitopes were derived from endogenous proteins and none was Gag-specific. The biggest barrier for identifying more epitopes, and probably also Gag-specific epitopes, was a PEG contamination which superimposed the actual peptide signals during mass spectrometry (Figure 45). In an optimal setting, using membrane-bound HLAs for affinity purification, between 1000 and 3000 peptides can currently be identified from 10^9 cells[345], allowing identification of antigen-specific epitopes as well[125]. In the epitope sequencing experiments presented in this thesis, the number of sHLA producing cells was at least threefold higher, but only a tenth of the number of peptides was identified. Hence, as long as the sensitivity is not increased it is not possible to address the breadth of presented Gag epitopes.

The first step to improve the quality of the results would be to eliminate the PEG contamination. Besides keratin, PEG is the most common contamination in mass spectrometry experiments, and is present in many detergents like Triton X-100 or Tween and may also leach from some types of plastic-wares. Since trace amounts can already prohibit sensitive mass spectrometry analysis, only glassware and specific low-bind plastic-wares were used, wherever possible. Although, these harsh restrictions did reduce the contamination signal, it was not completely eliminated. The source of the contamination could, however, be narrowed down and is most probably caused by the 3 kDa cut-off filter, which might release the undesired substance when the eluate, which contains 10% acetic acid, is added. The best alternative would be to completely omit use of the cut-off filter and instead separate peptides from HLA-hc, β2m, and residual W6/32 antibody through RP-HPLC. This method has also been reported to increase the yield of peptides by a factor of ten[345] compared to purification with a cut-off filter, where a significant amount of peptides is lost.

Besides avoiding contamination, the signal would also be improved if more peptides could be isolated overall. Due to the strong link between protein abundance and HLA-presentation[120], elevated Gag production in the cells would have an additionally benefit. For most experiments performed, Flp-In™ T-Rex™ 293 cells that were adapted to grow in suspension and that stably expressed the sHLA heavy chains, were employed. In addition, these cell lines were transiently transfected with Gag-coding plasmids using PEI. The transfection rates were however rather low and often below 15%. Thus, most cells were not able to present Gag-derived epitopes, but only endogenous peptides. To circumvent this, HEK293F cells were transiently co-transfected with sHLA-hc and Gag on separate expression plasmids so that the expression of both proteins is coupled in most cells. However, due to the PEG contaminations, even with this adapted method, no Gag-specific epitopes were identified. The easiest way to screen antigen-specific epitopes presented by several HLA alleles would probably be to design a bidirectional sHLA vector library. Such mammalian expression vectors could in one direction encode a sHLA-hc allele and contain in the other direction a multiple cloning site to readily insert any antigen of interest. Thereby only one plasmid preparation would be necessary and it would be ensured that all transfected cells express both sHLA and the antigen.

Since it is described that the pool of cellular β2m in the ER can be limiting for extensive quantities of HLA heavy chains[346], it has to be tested, if additionally providing β2m would have a beneficial effect on the amount of sHLA complexes released to the cell supernatant. However, due to the fact that a substantial fraction of HLA molecules never binds to suitable peptides[347] and subsequently gets degraded through the ERAD system[78], the peptide concentration seems to be rate limiting during HLA loading[70]. To eliminate competition for β2m, peptides, and possible other factors, deleting of all membrane-bound HLA molecules from the HEK293F expression cell line, for example by CRISPR/Cas mediated excision, could be advantageous.

In APCs responsible for priming a CD8+ T cell response, the immunoproteasome alters degradation patterns compared to other cells. To resemble the processing in APCs, the immunoproteasome could be activated in HEK293F cells by treatment with IFNγ[348]. An additional benefit with this strategy would be that the immunoproteasome significantly elevates epitope abundance[71] and more sHLA complexes can presumably be loaded and secreted.

In conclusion, a method to identify sHLA-presented-peptides through LC-MS/MS analysis was adapted in this thesis. However, in order to determine the immunological breadth of the teeGags, the sensitivity of the assay has to be further improved.

E.4.2 Other ways to analyze immunological breadth

Besides the direct identification of presented epitopes through mass spectrometry sequencing, there is only a limited number of assays available to address the immunological breadth of antigens. The most meaningful approach would be to administer the teeGags in human vaccine trials and read-out the breadth of primed T cells using peptide sets matching the teeGag sequences. Most animal *in vivo* experiments, in contrast, are futile, since teeGags are optimized for humans and the incorporated potent CD8+ T cell epitopes would not be processed and presented on the MCH alleles of mice or NHPs. The only possibility might be to employ humanized mice that can present antigens on human HLAs. An alternative could be to design SIV CD8+ T cell epitope-enriched Gag antigens and test them in NHPs, but the number of known, directly available SIV epitopes is too low (only 14[349]) to allow a reasonable optimization process.

An *in vitro* method to get a first impression of the immunological breadth would be to use a broad set of different Gag-specific CD8+ T cell clones and identify presented epitopes on antigen-containing cells. Another more natural approach could be to isolate PBMCs from HIV-1-positive donors and restimulate those cells with teeGags- or the reference HXB2-Gag-VLPs. Since teeGags are designed to contain as many epitopes as possible, they should be able to restimulate more T cells than HXB2-Gag, provided the T cell response of a given HIV patient is sufficiently diverse. This postulated average increased magnitude could serve as a surrogate for breadth. It allows, however, only a rough estimation, since an immunodominant response against a single epitope could not be differentiated from a broad immune response with only low responses against the many different epitopes. To map the immunological breadth more closely, peptide pools spanning the teeGags or the reference could be used for individual restimulation of PBMCs from HIV-1+ donors.

The breadth could also be determined through *ex vivo* priming of naïve CD8+ T cells. Thereby, DCs from an HIV-1 negative donor would be loaded with Gag antigen and co-cultured with autologous lymphocytes[350]. The breadth would be assessed by measuring the number of primed CD8+ T cells, which recognize different epitopes. However, since only a small subset of all lymphocytes can be used, it is not possible to determine the maximum breadth in these *ex vivo* experiments, but on average the teeGags might prime more T cells than the reference.

Finally, rather easily adaptable tools to assess the immunological breadth are *in silico* processing and presentation predictions. Yet, the accuracy of such approaches is limited by several unclear steps during processing, like the influences of peptidases and how different chaperones influence the peptide pool. Therefore, such *in silico* predictions still perform relatively poor[70,351].

F Appendix

F.1 Abbreviations

1D	One-dimensional
3D	Three-dimensional
AAS	Amino acid substitutions
ADCC	Antibody-dependent cellular cytotoxicity
AIDS	Acquired immune deficiency syndrome
Amp^R	Ampicillin resistance gene
anc	Ancestral
APC	Allophycocyanin
APC	Antigen presenting cell
ART	Antiretroviral therapy
β2m	Beta-2-microglobulin
BFA	Brefeldin A
BLASTP	Protein Basic Local Alignment Search Tool
bNAb	broadly neutralizing antibody
BrdUTP	Bromo-deoxyuridine triphosphate
BrdU	Bromo-deoxyuridine
BRMP	Biological Response Modifiers Program
BSA	Bovine serum albumin
CA	Capsid protein
CCB	Crosslinking coupling buffer
CD	Cluster of differentiation
CDR3	Complementarity determining region 3
CIP	Alkaline Phosphatase, Calf Intestinal
con	Consensus
cot	Center-of-tree
cpz	Chimpanzee
CRF	Circulating recombinant form
CSV	Comma-separated values
CTB	Crosslinking termination buffer
CTL	Cytotoxic T lymphocyte
cv	Column volume
DC	Dendritic cell
DLS	Dynamic light scattering
DMEM	Dulbecco's Modified Eagle's Medium
DMP	Dimethyl pimelimidate dihydrochloride
DMSO	Dimethyl sulfoxide
DPBS	Dulbecco's Phosphate Buffered Saline
DRiP	defective ribosomal product
ds	Double-stranded
ELISA	Enzyme-linked immunosorbent assay
EM	Electron microscopy
ER	Endoplasmic reticulum
ERAP	ER aminopeptidases
FA	Formic acid
FDR	False discovery rate
FESEM	Field emission scanning electron microscopy
FBS	Fetal Bovine Serum
FC	Flow Cytometry
FITC	Fluorescein isothiocyanate
FLD	Fisher's Linear Discriminant Analysis
FRT	Flippase recognition target
fwd	Forward
Gag	Group-specific antigen
GMP	Good manufacturing practices
GOI	Gene of interest
gor	Gorilla
hc	Heavy chain
HESN	HIV-exposed seronegative
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HPLC	High performance liquid chromatography
hpt	Hours post transfection
HRP	Horseradish peroxidase
$Hygro^R$	Hygromycin resistance gene
IFN-γ	Interferon gamma
Ig	Immunoglobulin
IN	Integrase
INSTI	Integrase strand transfer inhibitor
ITAM	Immunoreceptor tyrosine-based activation motifs
Kan^R	Kanamycin resistance gene
LB	lysogeny broth
LC	Liquid chromatography
LCL	Lymphoblastoid cell line
LTNP	Long-term nonprogressors
LTR	Long terminal repeat

MA	Matrix protein
MACS	Magnetic-activated cell sorting
mdDCs	Monocyte-derived dendritic cells
merGag	Multiple epitope refined Gag antigen
MHC	Major histocompatibility complex
MLA	Mixed leukocyte reaction
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MSA	Multiple sequence alignment
MWIS	Maximum weight independent set
NC	Nucleocapsid protein
Neo^R	Neomycin resistance gene
NHP	Non-human primates
NNRTI	Non-nucleoside reverse transcriptase inhibitor
NRTI	nucleoside reverse transcriptase inhibitor
PBMCs	Peripheral blood mononuclear cells
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PE	Phycoerythrin
PEI	Polyethylenimine
PIC	Pre-integration complex
PLC	Peptide-loading complex
POD	Horseradish peroxidase
PTE	Potential T cell epitopes
PR	Protease
RB	Relative budding
rev	Reverse
RP-HPLC	Reversed phase High performance liquid chromatography
RRE	Rev response element
RT	Reverse transcriptase
SD	Standard deviation
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEAP	Secreted alkaline phosphatase
SEM	Standard error of the mean
SHIV	Simian-human immunodeficiency virus
sHLA	Soluble human leukocyte antigen
SIV	Simian immunodeficiency virus
ss	Single-stranded
sus	Suspension
TAP	Transporter associated with antigen presentation
TBS	Tris-buffered saline
TCR	T cell receptor
teeGag	T cell epitope-enriched Gag antigen
TEM	Transmission electron microscopy
TFA	Trifluoroacetic acid
T_{FH}	Follicular helper CD4+ T cells
TMB	3,3',5,5'-Tetramethylbenzidin
URF	Unique recombinant form
VLP	Virus-like particle
VSV	Vesicular stomatitis virus
VSV-G	Vesicular stomatitis virus glycoprotein
WB	Western Blotting

F.2 Extended Data

F.2.1 Extended Data Tables

Extended Data Table 1. Changelog for the list of epitope entries that were identified as erroneous or undesired (D.1.1.2). Each entry can be assigned to the LANL database through its unique record number (“RecNo”). In 154 cases it was possible to correct the entry. For these lines marked as “Modified”, the detailed change log is stated. However 99 epitope entries had to be deleted (Action “Deleted”) with the reason for removing the respective database entry being listed in the last column.

RecNo	Epitope	Action	Changelog	Reason
56549	EKIRLRPGGKKYKL	Modified	Epitope: Changed to "EKIRLRPGGKKKYKL"	LANL database error
52872	KIRLRPGGKK	Modified	Subtype: "multiple" deleted	
12	KIRLRPGGKKKYKLKHIVWASRELE	Deleted		Epitope length > 20aa
14	KIRLRPGGKKKYKLKHIVWASRELE	Deleted		Epitope length > 20aa
13	KIRLRPGGKKKYKLKHIVWASRELE	Deleted		Epitope length > 20aa
52864	IRLRPGGKKK	Modified	Subtype: "multiple" deleted	
58852	RLRPGGKKK	Modified	HLA: "A02:01" --> "A03:01"	
52892	RLRPGGKKK	Modified	Subtype: "multiple" deleted	
53186	RPRPGGKKK	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
58435	RLRPGGKKK	Modified	Subtype: "A1" changed to "A"	
57939	RLRPGGKKK	Modified	Subtype: "A1" changed to "A"	
58974	LYNTVATL	Modified	Start: 20 --> 78; End: 28--> 85	LANL database error
55913	RLRPGGKKKKY	Deleted		Conservation score = 0; No access to reference paper
55912	RLRPGGKKKK	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
55704	RLRPGGKKKK	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
58436	RLRPGGKKKY	Modified	Subtype: "A1" changed to "A"	
30	CLRPGGKKKYKLKHIV	Deleted		Conservation score = 0; N-terminal cysteine added to aid synthesis
34	LRPGGKKKYKLKHIV	Modified	HLA: "not B8" deleted	
56297	RPGGKKKKYK	Modified	Epitope: Changed to "RPGGKKKYK"	LANL database error
59061	GGKKKYKMK	Deleted		Subtype: "HIV-2 "
58437	GGKKKYKLK	Modified	Subtype: "A1" changed to "A"	
58433	KYKLKHIVW	Modified	Subtype: "A1" changed to "A"	
54592	HYMLNHIVW	Modified	Epitope: Changed to "HYMLKHIVW"	Erroneous sequence in reference paper
55475	HYMLKHLVW	Modified	Subtype: "A1" changed to "A"	
57	QTGSEELRSLYNTVATLYCVHQRIE	Deleted		Epitope length > 20aa
56959	EELRSLYNT	Modified	Subtype: "A1" changed to "A"	
58430	ELRSLYNTV	Modified	Subtype: "A1" changed to "A"; HLA: "A*0801" --> "B*0801"	Erroneous HLA specification in reference paper; HLA A*08 does not exist
52873	ELRSLYNTVA	Modified	Subtype: "multiple" deleted	
53400	RSLYNTATLY	Deleted		Conservation score = 0; Sequence in reference paper maybe erroneous
1279	RLSYNTVATLY	Modified	Epitope: Changed to "RSLYNTVATLY"	Erroneous sequence in reference paper
58431	RSLYNTVATLY	Modified	Subtype: "A1" changed to "A"	
58427	SLYNTVATL	Modified	Subtype: "A1" changed to "A"	
55467	SLYNTVATL	Modified	Subtype: "A1" changed to "A"	
55756	SLFNTVATL	Deleted		Immunogen: "computer prediction" without experimental validation
52777	SLYNTAVTL	Modified	Epitope: Changed to "SLYNTVATL"	LANL database error
52875	SLYNTVATLY	Modified	Subtype: "multiple" deleted	
55638	SLYNIVATLWCVH	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
58973	TSTLQEQIGW	Modified	Protein: "p17" --> "p24"; Start: 78 --> 108; End: 88 --> 117	LANL database error
55634	YNIVATLWCVHQ	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
55772	ATLYCVHEKIEVRDTKEALDK	Deleted		Epitope length > 20aa
111	CRIDVKDTKEALEKIE	Deleted		Conservation score = 0; N-terminal cysteine added to aid synthesis
112	VHQRIEIKDTKEALDKIEEEQNKSKKKA	Deleted		Epitope length > 20aa
58438	VQNLQGQMV	Modified	Subtype: "A1" changed to "A"	
58439	QGQMVHQSL	Modified	Subtype: "A1" changed to "A"	
1174	VQHAISPRTLNAWV	Modified	Epitope: Changed to "VHQAISPRTLNAWV"	LANL database error
134	VHQAISPRTLNAWVKVVEEKAF	Deleted		Epitope length > 20aa
58440	HQSLSPRTL	Modified	Subtype: "A1" changed to "A"	
1537	QAISPRTLNAV	Modified	Epitope: Changed to "QAISPRTLNAW"	LANL database error
57775	AISPRTNLAW	Modified	Epitope: Changed to "AISPRTLNAW"	LANL database error
58441	SLSPRTLNA	Modified	Subtype: "A1" changed to "A"	
58812	ISPRTLNAW	Modified	Start: 14 --> 15	LANL database error
56956	AISPRTLNAW	Modified	Subtype: "A1" changed to "A"	
57597	ISPRTNLAW	Modified	Epitope: Changed to "ISPRTLNAW"	LANL database error
57258	ISPRTLNAW	Modified	Subtype: "A1" changed to "A"	
56591	TPQDLNTML	Modified	Start: 15 --> 48; End: 23 --> 56	LANL database error

RecNo	Epitope	Action	Changelog	Reason
56974	ISPRTLNAW	Modified	Subtype: "A1" changed to "A"	
58442	LSPRTLNAW	Modified	Subtype: "A1" changed to "A"	
58443	SPRTLNAWV	Modified	Subtype: "A1" changed to "A"	
58746	EKIRLRPGGKKKYKL	Modified	Protein: "p24" --> "p17"; End: 28 --> 31	LANL database error
55357	TLNAWVKLV	Deleted		Subtype: "HIV-2 "
53231	NAWVKIEEK	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
157	NAWVKVVEEKAFSPEVIPMFSA	Deleted		Epitope length > 20aa
58444	EEKAFSPEV	Modified	Subtype: "A1" changed to "A"	
58840	KAFSPEVIPMF	Modified	Start: 29 --> 30; End: 41 --> 40	LANL database error
52366	EKAFSPEVPMFTALSEGAT	Modified	Epitope: Changed to "EKAFSPEVIPMFTALSEGAT"	LANL database error
58285	KAFSPEVI	Modified	Subtype: "A1" changed to "A"	
58451	KAFSPEVIP	Modified	Subtype: "A1" changed to "A"	
57523	KAFSPEVIPMF	Modified	HLA: "A*0201" --> "B*5701"	LANL database error
57259	KAFSPEVIPMF	Modified	Subtype: "A1" changed to "A"	
58423	KAFSPEVIPMF	Modified	Subtype: "A1" changed to "A"	
56377	KAFSPEVIPMF	Modified	Subtype: "ACD" --> "B"	LANL database error
58445	FSPEVIPMF	Modified	Subtype: "A1" changed to "A"	
56557	SPEVIPMFASLSEGA	Modified	Epitope: Changed to "SPEVIPMFSALSEGA"	LANL database error
58694	VIPMSFAL	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
58446	VIPMFSALS	Modified	Subtype: "A1" changed to "A"	
1456	IPMFSALSEGATPDQL	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
58447	PMFSALSEG	Modified	Subtype: "A1" changed to "A"	
52367	MFTALSEGTPQDLNTMLNT	Modified	Epitope: Changed to "MFTALSEGATPQDLNTMLNT"	LANL database error
58448	FSALSEGAT	Modified	Subtype: "A1" changed to "A"	
55224	QALSEGCTPYDINQML	Deleted		Subtype: "HIV-2 "
179	SALSEGATPQDLNTMLNTVGGH	Deleted		Epitope length > 20aa
58449	LSEGATPQD	Modified	Subtype: "A1" changed to "A"	
58452	GATPQDLNM	Modified	Subtype: "A1" changed to "A"	
181	CTPYDINQMLNC	Deleted		Subtype: "HIV-2 "
58425	TPQDLNTNL	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
55683	TPYDINQML	Deleted		Subtype: "HIV-2 "
53711	TPYDINQML	Deleted		Subtype: "HIV-2 "
182	TPQDLNQML	Modified	Epitope: Changed to "TPQDLNTML"	LANL database error
55688	TPYDINQML	Deleted		Subtype: "HIV-2 "
187	TPYDINQML	Deleted		Subtype: "HIV-2 "
188	TPYDINQML	Deleted		Subtype: "HIV-2 "
59027	TPYDINQML	Deleted		Subtype: "HIV-2 "
55522	TPYDINQML	Deleted		Subtype: "HIV-2 "
55165	TPQDLNTML ?	Deleted		"?" = Amino acid sequence not defined by reference
55483	TPQDLNTML	Modified	Subtype: "A1" changed to "A"	
58426	TPQDLNTML	Modified	Subtype: "A1" changed to "A"	
58453	TPQDLNTML	Modified	Subtype: "A1" changed to "A"	
52903	TPQDLNTMLN	Modified	Subtype: "multiple" deleted	
52877	TPQDLNMMLN	Modified	Subtype: "multiple" deleted	
53705	TPQDLNQMLNTV	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
52904	DLNTMLNTVG	Modified	Subtype: "multiple" deleted	
190	DLNTMLNTVGGHQAAMQMLKETINEEAAEWDR	Deleted		Epitope length > 20aa
58454	LNMMLNIVG	Modified	Subtype: "A1" changed to "A"	
58455	MMLNIVGGH	Modified	Subtype: "A1" changed to "A"	
58456	LNIVGGHQA	Modified	Subtype: "A1" changed to "A"	
58338	GGHQAAMQMLKDTINEEEA	Modified	Epitope: Changed to "GGHQAAMQMLKDTINEEA"	LANL database error
203	GHQAAMQMKETINEEAAEW	Modified	Epitope: Changed to "GHQAAMQMLKETINEEAAEW"	LANL database error
204	GHQAAMQMLKETINEEAAEWDR	Deleted		Epitope length > 20aa
58457	HQAAMQMLK	Modified	Subtype: "A1" changed to "A"	
208	LKETINEEAAEWDRVPV	Modified	Epitope: Changed to "LKETINEEAAEWDRVHPV"	LANL database error
212	EIINEEAAEW	Deleted		Subtype: "HIV-2 "
58796	ETINEEAAAEW	Modified	Epitope: Changed to "ETINEEAAEW"	LANL database error
213	ETINEEAAEWDRVHPVVHAGP	Modified	Epitope: Changed to "ETINEEAAEWDRVHPVHAGP"	Erroneous sequence in reference paper
58458	EEAAEWDRL	Modified	Subtype: "A1" changed to "A"	
58459	EAAEWDRLH	Modified	Subtype: "A1" changed to "A"	
58460	AAEWDRLHP	Modified	Subtype: "A1" changed to "A"	
58461	RLHPVHAGP	Modified	Subtype: "A1" changed to "A"	
57432	HPVHAGPVA	Modified	Subtype: "A1" changed to "A"	
55477	HPVHAGPVA	Modified	Subtype: "A1" changed to "A"	
56524	HVPHAGPIA	Modified	Epitope: Changed to "HPVHAGPIA"	LANL database error
58462	VHAGPIPPG	Modified	Subtype: "A1" changed to "A"	
58463	PGQMREPRG	Modified	Subtype: "A1" changed to "A"	
58464	MREPRGSDI	Modified	Subtype: "A1" changed to "A"	
58465	SDIAGTTST	Modified	Subtype: "A1" changed to "A"	
55226	SDIAGTTSTVDEQIQWMY	Deleted		Subtype: "HIV-2 "
58466	IAGTTSTLQ	Modified	Subtype: "A1" changed to "A"	
58467	GTTSTLQEQ	Modified	Subtype: "A1" changed to "A"	
58842	TSTLQEQIGW	Modified	Start: 107 --> 108; End: 118 --> 117	LANL database error
57351	TSTLQEQIGW	Modified	HLA: "A*02" --> "B*57"	

RecNo	Epitope	Action	Changelog	Reason
56715	TSTLQEQIGW	Modified	HLA: "A*02" --> "B*57"	
57524	TSTLQEQIGW	Modified	HLA: "A*0201" --> "B*5701"	LANL database error
58779	TSTLQEQIGW	Modified	HLA: "B*2705" --> "B*5701"	
57952	TSTLQEQIGW	Modified	Subtype: "A1" changed to "A"	
57261	TSTPQEQIGW	Modified	Subtype: "A1" changed to "A"	
52759	TSTLQEQIGN	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
1335	TSTLQRQIGW	Modified	Epitope: Changed to "TSTLQEQIGW"	LANL database error
221	TSTVEEQQIW	Deleted		Subtype: "HIV-2 "
59066	TSTVEEQIQW	Deleted		Subtype: "HIV-2 "
55525	TSTVEEQIQW	Deleted		Subtype: "HIV-2 "
59025	TSTVDEQIQW	Deleted		Subtype: "HIV-2 "
53712	TSTVEEQIQW	Deleted		Subtype: "HIV-2 "
2010	TSTLQEQIGWF	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
55707	TSTLQEQIGWF	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
57981	TSTLQEQIGWF	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
219	TSTLQEQIGWF	Modified	Epitope: Changed to "TSTLQEQIGW"; End: 118 --> 117	LANL database error
57976	TSTLQEQIGWF	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
58471	KRWIILGLN		Start: 112 --> 131; End: 120 --> 139; Subtype: "A1" changed to "A"	
58468	TLQEQIGWM	Modified	Subtype: "A1" changed to "A"	
58469	QEQIGWMTG	Modified	Subtype: "A1" changed to "A"	
55225	MYRQQNPVPVGNIYRRWI	Deleted		Subtype: "HIV-2 "
53072	MTSNPPIPV	Modified	Subtype: "M" --> "C"	
58807	MTNNPPIPV	Modified	Subtype: "multiple" deleted	
233	NPPIPGEIKRWIILGNIK	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
57705	NPPIPVGEIYKRWIILGLNKIV	Deleted		Epitope length > 20aa
235	NPPIPVGEIYKRWIILGLNKIV	Deleted		Epitope length > 20aa
229	NPPIPVGEIYKRWIILGLNKIVRMYSPTSILD	Deleted		Epitope length > 20aa
227	NPPIPVGEIYKRWIILGLNKIVRMYSPTSILD	Deleted		Epitope length > 20aa
55663	NPVPVGNIY	Deleted		Subtype: "HIV-2 "
56346	NPVPVGNIY	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
1326	PPIPVGDIH	Modified	Epitope: Changed to "PPIPVGDIY"	LANL database error
245	NPVPVGNIY	Deleted		Subtype: "HIV-2 "
59013	NPVPVGNIY	Deleted		Subtype: "HIV-2 "
55227	PVGNIYRRWIQIGLQKCV	Deleted		Subtype: "HIV-2 "
52476	VGEIYKRWIIGLNK	Modified	Epitope: Changed to "VGEIYKRWIILGLNK"	LANL database error
55676	GDIYWKRWI	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
57520	EIYKRWII	Modified	HLA: "A*0201" --> "B*0801"	LANL database error
53743	EIYRKWII	Modified	Epitope: Changed to "EIYKRWII"	LANL database error
55159	EIYKRWII ?	Deleted		"?" = Amino acid sequence not defined by reference
52752	IYKLWIILGLNKIVRMYSPT	Modified	Epitope: Changed to "IYKRWIILGLNKIVRMYSPT"	LANL database error
58599	SLYNTVATL	Deleted		Erroneous start-end annotations
262	RRWIQLGLQK	Deleted		Subtype: "HIV-2 "
57522	KRWIILGLNK	Modified	HLA: "A*0201" --> "B*2705"	LANL database error
281	KRWIIMGNK	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
282	KRWIILGNK	Modified	Epitope: Changed to "KRWIILGLNK"	LANL database error
1802	RRWIQLGLQK	Deleted		Conservation score = 0
55658	KIRWIIGLNK	Modified	Epitope: Changed to "KRWIILGLNK"	LANL database error
57187	KRWIIILGLNK	Modified	Epitope: Changed to "KRWIILGLNK"	LANL database error
263	RRWIQLGLQK	Deleted		Subtype: "HIV-2 "
56729	RRWIQLGLQK	Deleted		Conservation score = 0
264	KRWIILGGLNK	Modified	Epitope: Changed to "KRWIILGLNK"	LANL database error
57945	KRWIILGLNK	Modified	Subtype: "A1" changed to "A"	
291	KRWIILGLNKIVMRY	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
292	KRWIILGLNKIVMRY	Modified	Epitope: Changed to "KRWIILGLNKIVRMY"	LANL database error
1183	KRWILGLNKIVRMY	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
294	KRWIILGLNKIVRMYC	Deleted		C-terminal cysteine added for chemical coupling
295	KRWIILGLNKIVRMYSPTSI	Modified	HLA: "B62?" deleted	
297	KRWIILGLNKIVRMYSPTSILD	Deleted		Epitope length > 20aa
1180	KWILGLNKIVRMY	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
1181	KWILGLNKIVRMY	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
58473	LGLNKIVRM	Modified	Start: 133 --> 136; End: 141 --> 144; Subtype: "A1" changed to "A"	LANL database error
58472	WIILGLNKI	Modified	Subtype: "A1" changed to "A"	
53764	GLNKVRMY	Modified	Epitope: Changed to "GLNKIVRMY"	Erroneous sequence in reference paper
58474	RMYSPVIL	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
55641	STLDIRQGPKEPFID	Modified	Epitope: Changed to "SILDIRQGPKEPFID"	LANL database error
58475	ILDIRQGPK	Modified	Subtype: "A1" changed to "A"	

RecNo	Epitope	Action	Changelog	Reason
55365	IRQGPKEPFRDYVDRFFKTLRAEQA	Deleted		Epitope length > 20aa
58476	PKEPFRDYV	Modified	Subtype: "A1" changed to "A"	
308	PKEPFRDYVDRFYKTLRAEQAS	Deleted		Epitope length > 20aa
58480	DYVDRFFKT	Modified	Subtype: "A1" changed to "A"	
55690	SYVDRFYKSL	Deleted		Subtype: "HIV-2 "
1275	YVDRFFKRL	Modified	Epitope: Changed to "YVDRFFKTL"	LANL database error
56596	QATQDVKNW	Modified	Start: 164 --> 176; End: 172 --> 184	LANL database error
55166	YVDRFFKTL ?	Deleted		"?" = Amino acid sequence not defined by reference
55223	YVDRFYKSLRAEQTDPAV	Deleted		Subtype: "HIV-2 "
54538	VDRFYKLTRAEQAS	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
54537	DRFYKLTRA	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
322	DRFWKTLRA	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
329	DRFYKLTRA	Modified	Epitope: Changed to "DRFYKTLRA"	LANL database error
59026	DRFYKSLRA	Deleted		Subtype: "HIV-2 "
58477	DRFFKTLRA	Modified	Subtype: "A1" changed to "A"	
52757	DRFYKTRAE	Deleted		Conservation score = 0
58478	FKTLRAEQA	Modified	Subtype: "A1" changed to "A"	
52481	YKTLRAEQASQDVKNWN	Modified	Epitope: Changed to "YKTLRAEQASQDVKNWM"	LANL database error
55228	LRAEQTDPAVKNWMTQTL	Deleted		Subtype: "HIV-2 "
58841	QASQEVKNW	Modified	Start: 175 --> 176; End: 185 --> 184	LANL database error
1699	QASGEVKNW	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
58434	QASQEVKNW	Modified	Subtype: "A1" changed to "A"	
1554	QASKEVKNWV	Deleted		Conservation score = 0
55689	QTDPAVKNWM	Deleted		Subtype: "HIV-2 "
58479	ETLLVQNAN	Modified	Subtype: "A1" changed to "A"	
58481	LVQNANPDC	Modified	Subtype: "A1" changed to "A"	
52221	VQNANPDCKTILKAL	Deleted		Immunogen: "computer prediction" without experimental validation
53160	NANPDSKTI	Modified	Epitope: Changed to "NANPDCKTI"	LANL database error
348	NANPDCKTI?	Deleted		"?" = Amino acid sequence not defined by reference
58482	NPDCKSILR	Modified	Subtype: "A1" changed to "A"	
58432	DCKTILKAL	Modified	Subtype: "A1" changed to "A"; HLA: "A*0801" --> "B*0801"	Erroneous HLA specification in reference paper; HLA A*08 does not exist
58483	DCKSILRAL	Modified	Subtype: "A1" changed to "A"	
58681	RALGPGATM	Modified	Epitope: Changed to "RALGPGATL"	LANL database error
55479	RALGPGATL	Modified	Subtype: "A1" changed to "A"	
58484	RALGPGATL	Modified	Subtype: "A1" changed to "A"	
58485	GATLEEMMT	Modified	Subtype: "A1" changed to "A"	
58486	LEEMMTACQ	Modified	Subtype: "A1" changed to "A"	
359	LEEMMTACQGVGGPGHKARVL	Deleted		Epitope length > 20aa
58487	EEMMTACQG	Modified	Subtype: "A1" changed to "A"	
58488	MMTACQGVG	Modified	Subtype: "A1" changed to "A"	
58489	MTACQGVGG	Modified	Subtype: "A1" changed to "A"	
57517	ACQGVGGPGHK	Modified	HLA: "A*0201" --> "A*1101"	LANL database error
58490	GGPGHKARV	Modified	Subtype: "A1" changed to "A"	
56599	GPGHKARVL	Modified	HLA: "B5" --> "B*0705"	
58978	RQANFLGKI	Modified	Protein: "p24" --> "p2p7p1p6"; Start: 223 --> 66; End: 231 --> 74	LANL database error
56597	AEQATQDVKNW	Deleted		Erroneous start-end annotations
57518	GPGHKARVL	Modified	HLA: "A*0201" --> "B*0702"	LANL database error
52884	GPGHKARVLA	Modified	Subtype: "multiple" deleted	
367	GHKARVLAEATLSQVN	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
53071	VLAEAMSQV	Modified	Subtype: "M" deleted	
52894	LARNCRAPRK	Modified	Subtype: "multiple" deleted	
56975	KAPRKKGCW	Modified	Subtype: "A1" changed to "A"	
58429	FLGKIWPSYK	Modified	Subtype: "A1" changed to "A"	
55466	FLGKIWPSHK	Modified	Subtype: "A1" changed to "A"	
55273	PPSGKGGNY	Deleted		Subtype: "HIV-2 "
53094	GNFLQSRPTAPPF	Deleted		Conservation score = 0; Sequence in reference paper inconclusive
52222	GNFLQSRPEPTAPPF	Deleted		Immunogen: "computer prediction" without experimental validation
56465	QNRPEPRPEPTAPPAENFRES	Deleted		Conservation score = 0; Sequence in reference paper inconclusive

Extended Data Table 2. Final set of unique epitopes (D.1.1.2). The epitope raw data from LANL immunology database was screened for multiple records of the same epitope. Such duplicates were merged into one consensus entry. The remaining 691 unique entries are listed below. Each epitope is specified by the [a]HIV-1 Gag protein it is located in, [b]start- and [c]end position in the respective protein, [d]epitope sequence, [e]associated subtypes, [f]HLA molecules that present it, [g]affiliation with LTNP, [h]expected immune response score, and [i]conservation score.

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p17	1	10	MGARASVLSG	CRF01_AE				0.29
p17	5	13	ASVLSGGEL	B				0.05
p17	5	15	ASILRGGKLDK	C				0.09
p17	5	18	ASVLSGGELDRWEK	B				0.03
p17	5	19	ASVLSGGELDRWEKI	B				0.03
p17	6	15	SVLSGGQLDR	B	A*11			0.01
p17	6	19	SVLSGGELDRWEKI	B				0.03
p17	9	23	SGGELDRWEKIRLRP	B	B*40; B*44			0.04
p17	11	19	GELDRWEKI	B	B*40		0.09	0.04
p17	11	19	GQLDRWEKI	B			0.01	0.01
p17	11	22	GKLDSWEKIRLR	A; CRF01_AE; CRF02_AG				0.02
p17	11	22	GKLDAWEKIRLR	CRF01_AE				0.20
p17	11	30	GELDRWEKIRLRPGGKKKYK	B	B*15			0.01
p17	12	21	ELDRWEKIRL	B; C	B*15			0.04
p17	13	27	LDRWEKIRLRPGGKK	B	A*03; B*40			0.08
p17	16	30	WEKIRLRPGGKKKYK	B				0.05
p17	17	31	EKIRLRPGGKKKYKL	B	B*07; B*27		0.00	0.04
p17	17	31	EKIRLRPGGKKHYML	C				0.10
p17	17	34	EKIRLRPGGKKHYMLKHL	B; C	Cw*17		0.00	0.06
p17	17	34	EKIRLRPGGKKKYKLKHI	B				0.02
p17	17	34	EKIRLRPGGKKKYRLKHL	B				0.09
p17	18	26	KIRLRPGGK	B; A; CRF01_AE	A*03; A*11; B*07; B*27		0.04	0.60
p17	18	27	KIRLRPGGKK	B; C	A*03; A*11; B*07; B*27; B*35	1	0.05	0.60
p17	18	31	KIRLRPGGKKKYKL	B	A*03; B*15			0.04
p17	18	32	KIRLRPGGKKKYKLK	B	B*57			0.04
p17	19	27	IRLRPGGKK	B	B*27		0.01	0.74
p17	19	28	IRLRPGGKKK	B	A*03; A*11; B*15			0.30
p17	20	28	RLRPGGKKK	B; A; C; CRF02_AG	A*03; A*11; A*30; B*07; B*15; B*42		0.05	0.30
p17	20	28	RLRQGGKKK	B	A*03			0.00
p17	20	29	RLRPGGKKKY	B; A; C; D	A*03; A*30; B*15; B*42		0.04	0.30
p17	20	29	RLRPGGKKHY	C	A*30; B*42			0.20
p17	20	30	RLRPGGKKKYK	B				0.05
p17	20	30	RLRPGGKKHYM	C				0.17
p17	20	31	RLRPGGKKRYRL	A; CRF01_AE; CRF02_AG				0.01
p17	21	29	LRPGGKKKY	B				0.32
p17	21	35	LRPGGKKKYKLKHIV	B	A*03; A*11; B*07; B*08			0.03
p17	21	35	LRPGGKKKYRLKHLV	A; D				0.11
p17	21	40	LRPGGKKKYRLKHLVWASRE	A	Cw*04			0.11
p17	22	29	RPGGKKHY	A; C; D	B*07; B*42			0.24
p17	22	30	RPGGKKHYM	C	B*07; B*42			0.21
p17	22	30	RPGGKKKYK	B	B*07			0.06
p17	22	30	RPGGKKRYM	C				0.05
p17	22	31	RPGGKKKYML	A; C; D	B*07; B*42			0.01
p17	22	31	RPGGKKKYKL	B; D	B*51; Cw*04			0.04
p17	22	31	RPGGKKRYKL	B	B*07			0.01
p17	23	34	PGGKKRYRLKHL	A; CRF02_AG				0.01
p17	24	31	GGKKKYKL	B; A; D; CRF02_AG	B*08		0.02	0.05
p17	24	31	GGKKKYRL	B	B*08			0.16
p17	24	32	GGKKKYKLK	B; A; F	B*08			0.05
p17	24	35	GGKKKYKLKHIV	B	B*08			0.03
p17	24	39	GGKKKYKLKHIVWASR	B				0.03
p17	25	39	GKKKYKLKHIVWASR	B	A*24; B*07			0.03
p17	25	39	GKKHYMLKHLVWASR	C				0.11
p17	25	41	GKKHYMLKHLVWASREL	B; C	Cw*03		0.07	0.11
p17	25	42	GKKQYKLKHIVWASRELE	B				0.01
p17	26	40	KKHYMLKHIVWASRE	C				0.05
p17	27	35	KRYMIKHLV	C				0.01
p17	28	36	KYKLKHIVW	B; A; C; F; CRF01_AE	A*23; A*24; A*26		0.19	0.04
p17	28	36	HYMLKHLVW	B; A; C	A*23; A*24			0.12
p17	28	36	HYMLKHIVW	A; C; D	A*23			0.06
p17	28	36	KYMLKHIVW	C	A*24			0.00
p17	28	36	KYRLKHIVW	B	A*24			0.05
p17	28	36	KYRLKHLVW	B; A	Cw*04		0.13	0.13
p17	28	38	HYMLKHLVWAS	C				0.11
p17	28	43	KYKLKHIVWASRELER	B				0.03
p17	29	37	YKLKHIVWA	B				0.08
p17	31	45	MKHLVWASRELERFA	C; CRF01_AE			0.11	0.09
p17	32	46	KHIVWASRELERFAV	B				0.07
p17	33	41	HLVWASREL	B; C	Cw*06; Cw*08			0.60
p17	33	50	HIVWASRELERFAVNPSL	B				0.00
p17	34	43	IVWASRELER	B	A*11			0.27
p17	34	44	LVWASRELERF	B; C	A*30; B*57		0.01	0.52
p17	36	44	WASRELERF	B; C; CRF01_AE	B*35		0.04	0.84
p17	37	51	ASRELERFAVNPGLL	B				0.10
p17	37	51	ASRELERFALNPGLL	C				0.53
p17	41	55	LERFALNPGLLETAE	CRF01_AE			0.06	0.13
p17	41	58	LERFAVNPSLLETSEGCR	B				0.00
p17	42	50	ERFAVNPGL	B	B*27			0.10
p17	42	51	ERFAVNPGLL	B	B*27			0.10
p17	42	56	ERFALNPGLLETSEG	C				0.26
p17	42	58	ERFAVNPGLLETSEGCR	B				0.04
p17	42	58	ERFALNPGLLETSEGCK	C				0.21
p17	43	51	RFAVNPGLL	B; C	B*15			0.10
p17	44	52	FAVNPGLLE	B				0.10
p17	47	55	NPGLLETSE	B				0.35
p17	49	57	GLLESSEGC	B	A*02			0.01
p17	49	64	GLLETSEGCKQIMKQL	C	A*68			0.06
p17	58	72	KQIMKQLQPALQTGT	C				0.06
p17	59	67	QILEQLQPA	B	A*02			0.02
p17	59	68	QILEQLQPAL	B	A*02			0.02
p17	63	72	QLQPSLQTGS	B	A*02			0.07
p17	63	79	QLQPSLQTGSEELRSLY	B				0.01
p17	70	86	TGTEELRSLYNTVATLY	B; C	Cw*14		0.04	0.12
p17	70	86	TGSEELRSLYNTVATLY	B				0.01
p17	71	79	GSEELKSLY	B	A*01			0.08
p17	71	79	GSEELRSLY	B	A*01		0.06	0.03
p17	71	79	GTEELRSLY	A	A*01			0.16
p17	71	85	GSEELRSLYNTVATL	B				0.01
p17	71	90	GSEELRSLYNTVATLYCVHQ	B			0.39	0.01
p17	73	81	EELRSLYNT	A				0.18
p17	73	82	EELRSLYNTV	C	B*40			0.15
p17	73	87	EELRSLYNTVATLYC	B	A*02			0.14
p17	74	82	ELRSLYNTV	B; A; F; CRF01_AE	B*08			0.15
p17	74	82	ELKSLYNTV	B	B*08			0.14
p17	74	82	ELKSLFNTI	B	B*08			0.05
p17	74	83	ELRSLYNTVA	B	A*02			0.15
p17	74	88	ELKSLYNTVATLYCV	C				0.10
p17	76	86	RSLYNTVATLY	B; A; C; F	A*02; A*30; B*15; B*57; B*58		0.03	0.15
p17	76	86	RSLYNTVAVLY	B; F	A*30			0.01
p17	76	86	RSLFNTVATLY	B				0.13
p17	76	90	KSLYNTVVTLWCVHQ	B; C; CRF01_AE			0.02	0.00
p17	77	85	SLYNTVATL	B; A; C; D; F; G; K; CRF02_AG	A*02; A*68		0.20	0.32
p17	77	85	SLFNTVATL	B; A; C	A*02		0.09	0.28
p17	77	85	SLYNTVVTL	B	A*02			0.01

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p17	77	85	SLYNTIATL	B; CRF01_AE	A*02		0.22	0.03
p17	77	85	SLYNTVAAL	B	A*02			0.00
p17	77	85	SLYNTVAVL	B	A*02			0.03
p17	77	86	SLYNTVATLY	B	A*02			0.31
p17	77	86	SLFNTVATLY	C				0.27
p17	77	91	SLYNTVATLYCVHQR	B; A; D	A*02; A*30; B*40; B*44			0.05
p17	77	94	SLYNTVATLYCVHQRIEV	B				0.01
p17	78	85	LYNTVATL	B; D	A*24; Cw*14			0.32
p17	78	86	LYNTVATLY	B; C	A*29; B*44			0.31
p17	78	86	LFNTVATLY	B; C	A*29			0.27
p17	80	88	NTVATLYCV	B	A*02			0.59
p17	82	90	VATLYCVHQ	B				0.17
p17	82	91	IATLWCVHQR	CRF01_AE	A*11		0.09	0.04
p17	82	92	VATLYCVHQRI	B	A*11			0.09
p17	83	91	AVLYCVHQR	B	A*11			0.07
p17	83	91	ATLYCVHQR	B	A*11			0.11
p17	83	91	ATLYCVHQK	B; C				0.05
p17	83	94	ATLWCVHQRIDI	CRF01_AE; CRF02_AG				0.03
p17	84	91	TLYCVHQR	B	A*11			0.11
p17	84	91	TLYCVHQK	B	A*11			0.05
p17	84	92	TLYCVHQRI	B; F	A*11		0.05	0.11
p17	84	92	TLYCVHQKI	B	A*11			0.05
p17	84	92	VLYCVHQRI	B	A*02			0.06
p17	85	92	LYCVHQKI	D	A*24			0.06
p17	86	96	YCVHAGIEVRD	C				0.03
p17	86	101	YCVHQRIEIKDTKEAL	B				0.02
p17	87	95	CVHQRIEIK	B	A*11			0.05
p17	89	98	HQRIEIKDTK	B	A*11			0.04
p17	91	101	RIDVKDTKEAL	B				0.07
p17	91	105	RIDVKDTKEALEKIE	B				0.01
p17	92	101	IEIKDTKEAL	B; F	B*40		0.07	0.06
p17	92	101	IDIKDTKEAL	B	B*40			0.05
p17	93	101	EVKDTKEAL	B	B*08			0.14
p17	93	101	EIKDTKEAL	B; CRF01_AE	B*08; B*40			0.06
p17	93	101	DVKDTKEAL	B				0.11
p17	97	111	TKEALEKIEEEQNKS	BC				0.02
p17	103	112	KIEEEQNKSK	B	A*11			0.08
p17	114	122	KTQQAAADK	B; F	B*57			0.00
p17	119	127	AADTGNSSQ	B				0.03
p17-p24	119	3	AADTGNSSQVSQNYPIV	B				0.02
p17	121	132	DTGHSNQVSQNY	B	A*33			0.00
p17	123	132	GNSSQVSQNY	B				0.05
p17	124	132	NSSKVSQNY	B	B*35		0.03	0.04
p17	124	132	NSSQVSQNY	B	B*35		0.02	0.07
p17-p24	124	1	NSSQVSQNYP	B				0.07
p17-p24	125	3	GKKVSQNYPIV	C				0.00
p17-p24	126	11	GKVSQNYPIVQNLQGQMV	C	B*13			0.21
p17-p24	126	11	SQVSQNYPIVQNLQGQMV	B				0.06
p17-p24	127	3	QVSQNYPIV	B; D	A*68			0.17
p17-p24	129	11	SQNYPIVQNLQGQMV	C				0.49
p17-p24	131	6	NYPIVQNL	B	A*24			0.52
p24	3	11	VQNLQGQMV	B; A; C	B*13		0.02	0.53
p24	7	15	QGQMVHQSL	A				0.05
p24	8	15	GQMVHQAI	B	B*48			0.34
p24	8	16	GQMVHQAIS	B				0.33
p24	8	17	GQMVHQAISP	B; A; C; D	B*57; B*58			0.33
p24	8	18	GQMVHQAISPR	B; C	A*74			0.33
p24	8	20	GQMVHQAISPRTL	B	Cw*03			0.33
p24	8	21	GQMVHQAISPRTLN	B; CRF01_AE	A*03; Cw*03			0.33
p24	8	27	GQMVHQAISPRTLNAWVKVV	B	B*14			0.08
p24	9	18	QMVHQAISPR	B	A*03			0.33
p24	9	23	QMVHQAISPRTLNAW	B	B*58			0.33
p24	9	23	QMVHQSLSPRTLNAW	A; D				0.04
p24	10	18	MVHQAISPR	B	A*03; A*33			0.33
p24	10	23	MVHQSMSPRTLNAW	A; CRF02_AG				0.01
p24	11	20	VHQAISPRTL	C	B*15			0.34
p24	11	24	VHQAISPRTLNAWV	B				0.34

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p24	11	25	VHQAISPRTLNAWVK	B				0.34
p24	12	20	HQAISPRTL	B	B*15			0.35
p24	12	20	HQSLSPRTL	A				0.07
p24	12	20	HQPISPRTL	B			0.03	0.09
p24	13	20	QAISPRTL	B	Cw*03; Cw*07			0.35
p24	13	23	QAISPRTLNAW	B	A*25; B*57		0.00	0.35
p24	14	22	SLSPRTLNA	A				0.07
p24	14	23	AISPRTLNAW	B; A; C; CRF02_AG	B*15; B*57; B*58		0.03	0.35
p24	14	24	AISPRTLNAWV	B	B*57			0.35
p24	15	23	ISPRTLNAW	B; A; C; D	B*15; B*57; B*58; Cw*06	1	0.03	0.48
p24	15	23	LSPRTLNAW	B; A; C	B*57; B*58			0.34
p24	15	24	ISPRTLNAWV	C	B*57			0.48
p24	16	24	SPRTLNAWV	B; A; C; D; CRF02_AG	B*07; B*81		0.03	0.94
p24	16	26	SPRTLNAWVKV	A; D	B*07			0.92
p24	17	28	PRTLNAWVKVVE	B				0.24
p24	17	31	PRTLNAWVKVVEEKA	B	A*02			0.16
p24	17	32	PRTLNAWVKVIEEKAF	B; C	Cw*02		0.02	0.59
p24	18	26	RLLNAWVKV	B; A; CRF02_AG	A*02		0.11	0.00
p24	18	26	RTLNAWVKV	B	A*02		0.03	0.96
p24	19	27	TLNAWVKVV	B	A*02			0.25
p24	19	27	TLNAWVKVI	B; A; C; D	A*02			0.71
p24	21	40	NAWVKVVEEKAFSPEVIPMF	B	B*57			0.16
p24	22	36	AWVKVVEEKGFNPEV	CRF01_AE				0.06
p24	22	36	AWVKVIEEKAFSPEV	C				0.54
p24	23	40	WVKVIEEKAFSPEVIPMF	B; C	Cw*16		0.01	0.54
p24	24	32	VKVIEEKAF	B; C	B*15			0.61
p24	24	32	VKVVEEKAF	B; C	B*15			0.17
p24	24	41	VKVIEEKAFSPEVIPMFT	C				0.39
p24	25	39	KVVEEKAFSPEVIPM	B; BC	B*44			0.16
p24	26	40	VIEEKAFSPEVIPMF	A; CRF01_AE; CRF02_AG				0.54
p24	27	36	IEEKAFSPEV	C				0.56
p24	27	37	IEEKAFSPEVI	C	B*45			0.55
p24	28	36	EEKAFSPEV	B; A; C; D	B*44; B*45		0.02	0.73
p24	28	47	EEKAFSPEVIPMFSALSEGA	B	B*27			0.23
p24	29	36	EKAFSPEV	C				0.73
p24	29	43	EKAFSPEVIPMFSAL	B; BC				0.25
p24	29	43	EKGFNPEVIPMFSAL	B; CRF01_AE			0.03	0.06
p24	29	48	EKAFSPEVIPMFTALSEGAT	C				0.44
p24	30	37	KAFSPEVI	B; A	B*57			0.74
p24	30	38	RAFSPEVIP	B; A; C	B*57			0.01
p24	30	38	KAFSPEVIP	A				0.74
p24	30	40	KAFSPEVIPMF	B; A; C; D; G; CRF02_AG	B*15; B*57; B*58	1	0.02	0.73
p24	30	40	KGFNPEVIPMF	B	B*57			0.10
p24	31	41	AFSPEVIPMFT	C				0.47
p24	31	44	AFSPEVIPMFSALS	B				0.25
p24	31	47	AFSPEVIPMFTALSEGA	B; C	Cw*14		0.04	0.45
p24	31	47	AFSPEVIPMFSALSEGA	B				0.24
p24	31	50	AFSPEVIPMFSALSEGATPQ	B				0.22
p24	32	40	FSPEVIPMF	B; A; C	B*15; B*57; B*58			0.81
p24	33	40	SPEVIPMF	C				0.81
p24	33	47	SPEVIPMFSALSEGA	B	A*02			0.27
p24	35	43	EVIPMFSAL	B; A; C; D; CRF01_AE	A*26; Cw*03		0.05	0.37
p24	35	43	EVIPMFTAL	C	A*26			0.53
p24	35	49	EVIPMFSALSEGATP	B				0.34
p24	36	43	VIPMFSAL	B; D	Cw*01; Cw*02			0.37
p24	36	44	VIPMFSALS	A				0.36
p24	37	46	IPMFSALSEG	B	B*07		0.06	0.36
p24	37	51	IPMFSALSEGATPQD	B	A*02; B*44			0.33
p24	37	52	IPMFSALSEGATPQDL	B	B*44			0.32
p24	38	46	PMFSALSEG	A				0.36
p24	38	48	PMFTALSEGAT	C				0.56
p24	38	55	PMFTALSEGATPQDLNTM	B; C	Cw*08		0.01	0.45
p24	38	55	PMFSALSEGATPQDLNTM	B; A; CRF01_AE				0.13
p24	39	53	MFSALSEGATPHDLN	A; D				0.00

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p24	39	58	MFTALSEGATPQDLNTMLNT	C				0.44
p24	40	48	FSALSEGAT	A				0.35
p24	41	55	SALSEGATPQDLNTM	B	B*44			0.14
p24	41	56	SALSEGATPQDLNTML	B				0.14
p24	41	60	SALSEGATPQDLNMMLNIVG	A	B*81			0.17
p24	41	60	SALSEGATPQDLNTMLNTVG	B; CRF01_AE				0.12
p24	42	52	ALSEGATPQDL	C				0.85
p24	42	55	ALSEGATPQDLNMM	A; CRF01_AE; CRF02_AG				0.21
p24	43	51	LSEGATPQD	A				0.86
p24	43	52	LSEGATPQDL	B; A	B*42; B*44			0.85
p24	44	52	SEGATPQDL	B; CRF02_AG	B*40; B*44		0.04	0.86
p24	45	56	EGATPQDLNMML	A; CRF01_AE; CRF02_AG				0.21
p24	45	59	EGATPQDLNTMLNTV	B	B*07; B*57			0.59
p24	45	60	EGATPQDLNTMLNTVG	B				0.59
p24	46	54	GATPQDLNM	A				0.22
p24	46	59	GATPQDLNMMLNIV	A; CRF01_AE; CRF02_AG				0.21
p24	46	59	GATPQDLNTMLNTV	B				0.59
p24	46	62	GATPQDLNTMLNTVGGH	B; C				0.59
p24	47	55	ATPQDLNTM	B	B*07			0.63
p24	47	56	ATPQDLNTML	B	B*07; B*58			0.62
p24	47	56	ATPQDLNMML	A	B*53			0.22
p24	47	58	ATPQDLNTMLNT	B; C; D	B*58			0.61
p24	48	55	TPQDLNTM	B; C	B*07; B*81			0.63
p24	48	56	TPQDLNTML	B; A; C; D	B*07; B*39; B*42; B*53; B*81; Cw*08		0.02	0.62
p24	48	56	TPQDLNTMLNT	B; C	B*14			0.61
p24	48	56	TPTDLNTML	B	B*42		0.01	0.01
p24	48	56	TPQDLNMML	B; A; CRF02_AG	B*42; B*53		0.02	0.22
p24	48	57	TPQDLNTMLN	B; C; D	B*07; B*14			0.62
p24	48	57	TPQDLNMMLN	B	B*07			0.22
p24	48	62	TPQDLNMMLNIVGGH	A; D				0.21
p24	49	57	PQDLNTMLN	B	B*14; Cw*07; Cw*08			0.62
p24	49	59	PQDLNTMLNTV	B	B*14			0.60
p24	49	63	PQDLNTMLNTVGGHQ	B; BC	A*02			0.59
p24	51	59	DLNTMLNTV	B	B*14; Cw*08			0.65
p24	51	59	DLNMMLNIV	B; A	B*14			0.23
p24	51	60	DLNTMLNTVG	B	A*02; B*14			0.64
p24	51	70	DLNTMLNTVGGHQAAMQMLK	B				0.62
p24	52	60	LNMMLNIVG	A				0.23
p24	53	66	NTMLNTVGGHQAAM	C				0.64
p24	53	70	NTMLNTVGGHQAAMQMLK	C				0.62
p24	54	62	MMLNIVGGH	A				0.23
p24	56	64	LNIVGGHQA	A				0.25
p24	57	71	NTVGGHQAAMQMLKE	B	A*02			0.14
p24	59	72	VGGHQAAMQMLKET	B; C; CRF01_AE			0.02	0.19
p24	60	70	GGHQAAMQMLK	C				0.95
p24	60	78	GGHQAAMQMLKDTINEEA	C				0.69
p24	61	69	GHQAAMQML	B; C	B*15; B*38; B*39		0.02	0.96
p24	61	71	GHQAAMQMLKE	B	A*02		0.16	0.20
p24	61	71	GHQAAMQMLKD	C; D	A*02			0.75
p24	61	71	GHQAAMEMLKD	B; A	A*02			0.00
p24	61	75	GHQAAMQMLKETINE	B	A*02; B*15			0.18
p24	61	78	GHQAAMQMLKETINEEAA	B				0.18
p24	61	80	GHQAAMQMLKETINEEAAEW	B				0.17
p24	62	70	HQAAMQMLK	B; A	A*11; B*52			0.96
p24	62	75	HQAAMQMLKETINE	CRF01_AE				0.18
p24	64	78	AAMQMLKDTINEEAA	B; C				0.69
p24	64	80	AAMQMLKETINEEAAEW	B				0.17
p24	65	73	AMQMLKETI	B; CRF02_AG	A*02			0.20
p24	65	73	AMQMLKDTI	BC				0.75
p24	65	79	AMQMLKETINEEAAE	B	B*40			0.17
p24	66	79	MQMLKDTINEEAAE	A; CRF01_AE				0.67
p24	69	83	LKETINEEAAEWDRV	B	A*25			0.04
p24	69	86	LKDTINEEAAEWDRLHPV	C	A*68			0.41
p24	69	86	LKETINEEAAEWDRVHPV	B				0.04

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p24	69	86	LKETINEEAAEWDRLHPV	B				0.11
p24	70	78	KETINEEAA	B	B*40		0.09	0.19
p24	70	83	KDTINEEAAEWDRL	A; CRF01_AE				0.47
p24	71	80	ETINEEAAEW	B; A; D	A*25; B*53; B*58		0.00	0.18
p24	71	80	DTINEEAAEW	B; A; C	B*53; B*58		0.01	0.69
p24	71	90	ETINEEAAEWDRVHPVHAGP	B				0.03
p24	72	80	TINEEAAEW	B	B*57			0.87
p24	73	87	INEEAAEWDRVHPVH	B	A*02			0.12
p24	73	87	INEEAAEWDRLHPVH	BC				0.43
p24	75	83	EEAAEWDRL	B; A; F	B*40			0.61
p24	75	83	EEAAEWDRV	B	B*40			0.15
p24	76	84	EAAEWDRLH	A				0.64
p24	77	85	AAEWDRLHP	A				0.65
p24	77	86	AAEWDRLHPV	B				0.55
p24	77	87	AAEWDRLHPVH	C				0.46
p24	77	91	AAEWDRVHPVHAGPI	B	A*02; B*40; B*44			0.08
p24	77	91	AAEWDRLHPVHAGPI	BC				0.25
p24	77	92	AAEWDRLHPVHAGPIA	B; C				0.19
p24	78	86	AEWDRLHPV	B; C; F	A*02; B*40			0.55
p24	78	86	AEWDRVHPV	B	B*40		0.13	0.15
p24	79	86	EWDRVHPV	B	A*02			0.15
p24	81	95	DRVHPVHAGPIAPGQ	B	B*07			0.03
p24	81	95	DRLHPVHAGPIAPGQ	B				0.19
p24	81	100	DRLHPVHAGPAAPGQMREPR	B				0.01
p24	82	90	RLHPVHAGP	A				0.48
p24	82	92	RLHPVHAGPIA	C				0.20
p24	83	91	LHPVHAGPI	B				0.27
p24	83	92	VHPVHAGPIA	B	B*55			0.03
p24	84	91	HPVHAGPI	D	B*35			0.42
p24	84	91	HPVHAGPV	B	B*35			0.21
p24	84	92	HPVHAGPIA	B; C; D; F	B*07; B*35; B*39		0.06	0.26
p24	84	92	HPVHAGPVA	B; A; C; D	B*07			0.20
p24	84	100	HPVHAGPIAPGQMREPR	B	A*02			0.22
p24	86	94	VHAGPIPPG	A				0.16
p24	87	101	HAGPIAPGQMREPRG	B	A*02			0.23
p24	89	96	GPIAPGQM	C; D	B*35			0.25
p24	91	105	IAPGQMREPRGSDIA	B; C	B*13			0.24
p24	91	110	IAPGQMREPRGSDIAGTTST	B				0.20
p24	93	101	PGQMREPRG	A				0.68
p24	93	107	PGQMREPRGSDIAGT	B				0.65
p24	94	104	GQMREPRGSDI	B; C	B*13		0.03	0.68
p24	94	105	GQMREPRGSDIA	C				0.68
p24	96	104	MREPRGSDI	A				0.68
p24	101	120	GSDIAGTTSTLQEQIGWMTN	B				0.07
p24	102	110	SDIAGTTST	A				0.79
p24	102	118	SDIAGTTSTLQEQIAWM	C				0.30
p24	104	112	IAGTTSTLQ	A				0.67
p24	105	119	AGTTSTLQEQIGWMT	B	A*02			0.20
p24	106	114	GTTSTLQEQ	A				0.67
p24	106	120	GTTSTLQEQIAWMTS	C				0.16
p24	107	116	TTSTLQEQIA	C	A*68			0.33
p24	108	117	TSTLQEQIGW	B; A; C; CRF01_AE	B*07; B*15; B*57; B*58	1	0.03	0.21
p24	108	117	TSTLQEQIAW	B; C	B*57; B*58	1	0.03	0.33
p24	108	117	TSTPQEQIGW	A	B*57			0.05
p24	108	117	TSTLQEQVGW	B	B*57			0.01
p24	108	117	TSTLQEQVAW	C				0.02
p24	109	117	STLQEQIGW	B; D; CRF02_AG	B*57; B*58		0.02	0.21
p24	109	117	STLQEQIAW	B				0.33
p24	109	118	STLQEQIGWM	B	A*02			0.21
p24	109	123	STLQEQIGWMTNNPP	B	A*02			0.07
p24	109	124	STLQEQIGWMTNNPPI	B				0.07
p24	109	126	STLQEQIGWMTNNPPIPV	B				0.07
p24	110	118	TLQEQIGWM	B; A; CRF01_AE	A*02			0.21
p24	110	118	NLQEQIGWM	B	A*02			0.03
p24	110	119	NLQEQIGWMT	B	A*02			0.03
p24	112	120	QEQIGWMTG	A				0.03
p24	114	128	QIGWMTNNPPIPVGE	B; A; CRF02_AG				0.05

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p24	116	130	AWMTSNPPVPVGDIY	C				0.05
p24	117	126	WMTNNPPIPV	B	A*02			0.16
p24	117	131	WMTNNPPIPVGEIYK	B	A*02			0.08
p24	117	134	WMTNNPPIPVGEIYKRWI	B				0.08
p24	118	126	MTNNPPIPV	B	A*02; A*03; B*07			0.16
p24	118	126	MTSNPPIPV	C	A*02			0.37
p24	121	129	NPPIPVGEI	B	B*07		0.02	0.29
p24	121	130	NPPIPVGDIY	B; C	B*35		0.03	0.37
p24	121	130	NPPIPVGEIY	B	B*35		0.01	0.29
p24	121	135	NPPIPVGEIYKRWII	B	B*08; B*44			0.24
p24	121	140	NPPIPVGEIYKRWIILGLNK	B				0.21
p24	122	130	PPIPVGEIY	B; A; C; D	B*07; B*35; B*53		0.05	0.29
p24	122	130	PPIPVGDIY	B; A; C; CRF02_AG	B*35		0.05	0.38
p24	122	130	PPVPVGDIY	C	B*35			0.14
p24	122	136	PPIPVGDIYKRWIIL	C				0.34
p24	123	131	PIPVGDIYK	B; C				0.37
p24	124	138	IPVGEIYKRWIILGL	B; A; C	B*08			0.23
p24	125	135	PVGDIYKRWII	C				0.52
p24	125	139	PVGEIYKRWIILGLN	B	A*24; B*40			0.32
p24	125	142	PVGEIYKRWIILGLNKIV	B				0.32
p24	126	140	VGEIYKRWIILGLNK	B				0.32
p24	127	135	GEIYKRWII	B; A; C; D; CRF02_AG	B*08		0.05	0.35
p24	127	135	GDIYKRWII	B	B*08			0.52
p24	127	136	GEIYKRWIIL	B	A*24; B*08		0.03	0.33
p24	127	143	GEIYKRWIILGLNKIVR	B	B*27			0.32
p24	127	143	GEIYKRWIILGLNKIVRMY	B	B*27			0.32
p24	128	135	EIYKRWII	B	B*08			0.35
p24	128	135	DIYKRWII	B; C	B*08			0.53
p24	128	136	EIYKRWIIL	B; D	A*02; A*24; B*08; Cw*07		0.08	0.33
p24	128	142	DIYKRWIILGLNKIV	C				0.49
p24	129	136	IYKRWIIL	B	A*24; B*08			0.83
p24	129	137	IYKRWIILG	B	A*24			0.82
p24	129	138	IYKRWIILGL	B; D	A*24; B*27			0.82
p24	129	140	IYKRWIILGLNK	B; A; C; CRF01_AE	A*24			0.81
p24	129	143	IYKRWIILGLNKIVR	B	A*24; B*27			0.81
p24	129	148	IYKRWIILGLNKIVRMYSPT	B	B*15; B*27			0.09
p24	130	148	YKRWIILGLNKIVRMYSPT	B	B*27			0.09
p24	131	139	KRWIILGLN	A				0.83
p24	131	140	KRWIILGLNK	B; A; C; D; CRF01_AE; CRF02_AG	B*27	1	0.04	0.82
p24	131	140	KRWIIMGLNK	B	B*27		0.03	0.04
p24	131	140	KRWIIMGLHK	B	B*27			0.00
p24	131	140	KRWILLGLNK	B				0.00
p24	131	142	KRWILGLNKIV	B	B*27			0.00
p24	131	142	KRWIILGLNKIV	B	B*27			0.82
p24	131	145	KRWIILGLNKIVRMY	B	B*27			0.81
p24	131	150	KRWIILGLNKIVRMYSPTSI	B				0.08
p24	132	140	RWIILGLNK	B	B*27			0.84
p24	133	141	WIILGLNKI	A				0.84
p24	133	147	WIILGLNKIVRMYSP	B	A*03; A*11; B*15; B*27			0.83
p24	133	150	WIILGLNKIVRMYSPVSI	B; C	Cw*17; Cw*18		0.01	0.70
p24	134	141	IILGLNKI	B	A*02; A*03			0.85
p24	134	142	IILGLNKIV	B	A*02			0.84
p24	134	143	IILGLNKIVR	B; A; C	A*03; A*11; A*33		0.01	0.84
p24	135	142	ILGLNKIV	B; C; CRF01_AE	A*02; A*03; B*27			0.85
p24	135	143	ILGLNKIVR	B	A*03; A*11			0.84
p24	135	145	ILGLNKIVRMY	B	B*07; B*27			0.84
p24	136	144	LGLNKIVRM	A				0.91
p24	136	145	LGLNKIVRMY	B	B*15			0.91
p24	136	146	LGLNKIVRMYS	B; C	B*15			0.90
p24	136	150	LGLNKIVRMYSPTSI	B				0.09
p24	136	153	LGLNKIVRMYSPTSILDI	B				0.09
p24	137	145	GLNKIVRMY	B; A; CRF01_AE	B*15; B*27		0.07	0.97
p24	137	151	GLNKIVRMYSPTSIL	B	A*02; B*15			0.10
p24	139	153	NKIVRMYSPVSILDI	A; CRF01_AE;			0.25	0.80

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
				CRF02_AG				
p24	141	155	IVRMYSPTSILDIRQ	B	A*02; A*24			0.06
p24	142	150	VRMYSPVSI	B; C; F	Cw*18			0.83
p24	143	150	RMYSPTSI	B; F	B*52		0.01	0.11
p24	143	151	RMYSPTSIL	B; CRF01_AE; CRF02_AG	A*02			0.11
p24	143	151	RMYSPVSIL	B; A				0.82
p24	144	151	MYSPTSIL	B	A*24			0.11
p24	144	158	MYSPVSILDIRQGPK	CRF01_AE			0.05	0.38
p24	145	153	YSPVSILDI	CRF01_AE	Cw*01		0.12	0.81
p24	145	153	YSPTSILDI	B	Cw*12			0.11
p24	149	158	SILDIRQGPK	B	A*11			0.46
p24	149	163	SILDIKQGPKEPFRD	A				0.46
p24	149	163	SILDIRQGPKEPFID	CRF02_AG				0.00
p24	149	165	SILDIKQGPKEPFRDYV	C				0.46
p24	150	158	ILDIRQGPK	A				0.47
p24	151	170	LDIRQGPKEPFRDYVDRFYK	B				0.14
p24	152	162	DIRQGPKEPFR	B	B*27			0.46
p24	154	168	RQGPKEPFRDYVDRF	B; A; CRF02_AG				0.45
p24	156	164	GPKEPFRDY	B	B*07		0.11	0.95
p24	156	173	GPKEPFRDYVDRFFKTLR	B; C	Cw*03; Cw*18		0.12	0.60
p24	156	173	GPKEPFRDYVDRFYKTLR	B				0.18
p24	157	165	PKEPFRDYV	A				0.95
p24	159	168	EPFRDYVDRF	B; C; D	A*02			0.94
p24	159	169	EPFRDYVDRFF	A; C; D	B*81			0.74
p24	159	178	EPFRDYVDRFFKTLRAEQAT	C	B*44		0.04	0.52
p24	160	168	PFRDYVDRF	D				0.94
p24	160	169	PFRDYVDRFF	C				0.75
p24	161	169	FRDYVDRFY	B	A*01			0.20
p24	161	169	FRDYVDRFF	B; C	Cw*18			0.77
p24	161	170	FRDYVDRFYK	B; D	B*18; B*27			0.19
p24	161	170	FRDYVDRFFK	B				0.75
p24	161	172	FRDYVDRFFKAL	A	A*03			0.04
p24	161	174	FRDYVDRFYKTLRA	B				0.18
p24	161	175	FRDYVDRFYKTLRAE	B; A; D; BC	B*44			0.18
p24	161	180	FRDYVDRFYKTLRAEQASQD	B	B*15			0.06
p24	161	180	FRDYVDRFYKTLRAEQASQE	B				0.06
p24	162	172	RDYVDRFFKTL	B; A	A*24		0.16	0.62
p24	162	172	RDYVDRFYKTL	B; D	A*24; A*26; B*15; B*18; B*44		0.11	0.19
p24	163	171	DYVDRFYKT	B	A*24			0.19
p24	163	171	DYVDRFFKT	A				0.63
p24	163	172	DYVDRFFKTL	B	A*24			0.63
p24	163	172	DYVDRFYKTL	B	A*24			0.19
p24	163	173	DYVDRFYKTLR	B	A*33			0.19
p24	164	172	YVDRFYKTL	B	A*02; B*15			0.19
p24	164	172	YVDRFFKTL	B; A; C; D	A*26; B*15; Cw*03		0.28	0.63
p24	164	172	YVDRFFKAL	A; CRF01_AE	B*15; Cw*03			0.04
p24	164	178	YVDRFYKTLRAEQAT	CRF01_AE			0.17	0.06
p24	164	181	YVDRFFKTLRAEQATQDV	C				0.30
p24	164	181	YVDRFYKTLRAEQASQEV	B				0.06
p24	165	178	VDRFYKTLRAEQAS	B				0.12
p24	165	179	VDRFYKTLRAEQASQ	B	B*57			0.12
p24	166	174	DRFYKTLRA	B; D	B*14; B*27	1	0.03	0.19
p24	166	174	DRFFKTLRA	B; A; C	B*14			0.63
p24	166	174	DRFWKTLRA	B	B*14			0.00
p24	166	175	DRFYKTLRAE	B	B*14			0.19
p24	166	176	DRFYKTLRAEQ	B; A	B*14			0.19
p24	168	176	FYKTLRAEQ	B	Cw*04			0.19
p24	169	177	FKTLRAEQA	A				0.58
p24	169	183	YKTLRAEQASQEVKN	B	B*58			0.06
p24	169	185	YKTLRAEQASQDVKNWM	B				0.05
p24	169	188	YKTLRAEQASQEVKNWMTET	B	B*57			0.06
p24	169	188	FKTLRAEQATQDVKNWMTDT	C				0.23
p24	171	180	TLRAEQATQD	C	Cw*03			0.32
p24	172	189	LRAEQASQEVKNWMTETL	B				0.07
p24	172	189	LRAEQATQDVKNWMTDTL	C				0.27
p24	173	181	RAEQASQEV	B	B*51; Cw*08			0.08

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p24	173	181	RAEQATQEV	A				0.35
p24	173	182	RAEQASQEVK	B	A*11			0.08
p24	173	183	RAEQATQDVKN	C				0.37
p24	173	187	RAEQASQEVKNWMTE	B	B*44			0.07
p24	173	187	RAEQATQDVKNWMTD	A				0.27
p24	173	187	RAEQATQEVKNWMTE	A; CRF01_AE				0.12
p24	174	184	AEQASQDVKNW	B; C; D	B*44; B*57; Cw*04		0.07	0.08
p24	174	184	AEQASQEVKNW	B	B*44		0.04	0.07
p24	174	184	AEQASADVKNW	B	B*44			0.00
p24	174	184	AEQATQDVKNW	B; C	B*44			0.37
p24	174	185	AEQASQEVKNWM	B	Cw*05			0.07
p24	174	186	AEQASQEVKNWMT	B	B*44			0.07
p24	175	186	EQASQEVKNWMT	B	B*44			0.07
p24	176	184	QATQDVKNW	C	B*53; B*57; B*58		0.03	0.37
p24	176	184	QASQEVKNW	B; A; D	B*53; B*57; B*58; Cw*04	1	0.04	0.07
p24	176	184	QATQEVKNW	B; A; CRF02_AG	B*53		0.02	0.24
p24	176	184	QATQEVKNM	A	B*53		0.03	0.00
p24	176	184	QATQEVKGW	B; CRF02_AG				0.09
p24	177	185	ASQEVKNWM	B	B*53			0.07
p24	177	185	ATQEVKNWM	B	B*53		0.02	0.24
p24	179	193	QEVKNWMTETLLVQN	CRF01_AE			0.05	0.19
p24	180	189	EVKNWMTETL	B	B*53			0.21
p24	180	190	EVKNWMTETLL	B				0.20
p24	181	189	VKNWMTETL	B	B*48		0.02	0.38
p24	181	190	VKNWMTETLL	B	B*08			0.37
p24	181	191	VKNWMTETLLV	B				0.36
p24	185	202	MTDTLLVQNANPDCKTIL	C	B*08			0.37
p24	187	195	ETLLVQNAN	A				0.41
p24	190	198	LVQNANPDC	A				0.87
p24	190	199	LVQNSNPDCK	B	A*11			0.03
p24	190	204	LVQNANPDCKTILRA	C				0.40
p24	191	199	VQNSNPDCK	B	A*11		0.13	0.03
p24	191	199	VQNANPDCK	B	A*03			0.86
p24	191	205	VQNANPDCKTILKAL	B	B*08; B*51			0.17
p24	191	210	VQNANPDCKTILKALGPAAT	B				0.08
p24	193	201	NANPDCKTI	B; A	B*08; B*51		0.08	0.65
p24	193	202	NSNPDCKTIL	B	B*51			0.02
p24	193	205	NANPDCKTILRAL	C	B*39			0.42
p24	193	207	NANPDCKTILKALGP	B	B*07			0.14
p24	193	209	NANPDCKTILRALGPGA	B; C	Cw*05; Cw*08; Cw*12		0.05	0.33
p24	194	202	ANPDCKTIL	B; C; CRF01_AE	B*07			0.65
p24	195	202	NPDCKTIL	B; C	B*08; B*35		0.03	0.67
p24	195	203	NPDCKSILR	A				0.12
p24	195	205	NPDCKTILRAL	C	B*39			0.42
p24	196	204	PDCKTILKA	B				0.20
p24	197	205	DCKTILKAL	B; A	B*08		0.01	0.20
p24	197	205	DCKTILRAL	C	B*08			0.43
p24	197	205	DCKSILRAL	A				0.11
p24	197	211	DCKTILKALGPAATL	B	B*57			0.10
p24	197	212	DCKTILKALGPAATLE	B				0.10
p24	199	213	KSILKALGTGATLEE	CRF01_AE			0.18	0.05
p24	199	213	KSILRGLGAGATLEE	A				0.00
p24	199	218	KTILRALGPGATLEEMMTAC	C				0.22
p24	200	214	TILRALGPGASLEEM	C				0.09
p24	200	217	TILRALGPGASLEEMMTA	B; C	Cw*03; Cw*07		0.03	0.09
p24	201	215	ILRALGPGATLEEMM	CRF02_AG				0.29
p24	203	211	KALGPAATL	B	B*15; Cw*03			0.11
p24	203	211	RALGPGATL	B; A; C; D	Cw*08			0.30
p24	204	214	ALGPGASLEEM	C				0.12
p24	205	219	LGPAATLEEMMTACQ	B	A*02			0.13
p24	206	214	GPGATLEEM	A; C; D				0.40
p24	208	216	GATLEEMMT	A				0.61
p24	208	226	AATLEEMMTACQGVGGPSH	B				0.04
p24	209	217	ATLEEMMTA	B; A; D; CRF01_AE	A*02		0.08	0.74
p24	209	223	ATLEEMMTACQGVGG	B	A*02			0.73
p24	211	219	LEEMMTACQ	A				0.94
p24	211	230	LEEMMTACQGVGGPGHKARV	B	B*07			0.47

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p24	212	220	EEMMTACQG	A				0.96
p24	212	222	EEMMTACQGVG	C				0.95
p24	213	221	EMMTACQGV	B	A*02		0.03	0.96
p24	213	227	EMMTACQGVGGPGHK	B	A*11			0.50
p24	214	222	MMTACQGVG	A				0.96
p24	215	223	MTACQGVGG	A				0.96
p24	216	230	TACQGVGGPSHKARV	C				0.41
p24	217	227	ACQGVGGPGHK	B; CRF01_AE	A*11		0.14	0.51
p24	217	227	ACQGVGGPSHK	B			0.03	0.44
p24	217	231	ACQGVGGPGHKARVL	B	A*11			0.49
p24	217	231	ACQGVGGPSHKARIL	A; CRF01_AE				0.02
p24-p2p7p1p6	217	1	ACQGVGGPGHKARVLA	B				0.48
p24-p2p7p1p6	217	3	ACQGVGGPGHKARVLAEA	B				0.48
p24	218	227	CQGVGGPGHK	B	A*11		0.14	0.51
p24	219	231	QGVGGPGHKARVL	B; C				0.49
p24	220	227	GVGGPGHK	B	A*11			0.52
p24	220	227	GVGGPSHK	B			0.05	0.45
p24	220	229	GVGGPGHKAR	B; F	A*11			0.51
p24	221	231	VGGPGHKARVL	B; C; CRF01_AE	B*07			0.49
p24-p2p7p1p6	221	4	VGGPGHKARVLAEAM	B	A*02; B*07			0.49
p24-p2p7p1p6	221	4	VGGPSHKARILAEAM	A; CRF01_AE				0.02
p24-p2p7p1p6	221	5	VGGPGHKARVLAEAMS	B				0.48
p24	222	230	GGPGHKARV	A				0.50
p24	223	231	GPGHKARVL	B; C; D; F	B*07; B*35		0.06	0.50
p24	223	231	GPSHKARVL	B; A; C; D; CRF01_AE	B*07; B*42; B*81		0.05	0.42
p24-p2p7p1p6	223	1	GPSHKARVLA	C	B*07			0.42
p24-p2p7p1p6	223	1	GPGHKARVLA	B	B*07			0.49
p24-p2p7p1p6	225	8	GHKARVLAEAMSQVT	B	A*02; B*07			0.06
p24-p2p7p1p6	225	11	SHKARVLAEAMSQANSA	B; C	Cw*08		0.00	0.00
p24-p2p7p1p6	226	10	HKARVLAEAMSQTNS	C				0.04
p24-p2p7p1p6	229	7	RVLAEAMSQV	B	A*02			0.28
p24-p2p7p1p6	229	12	RVLAEAMSQVTNSAT	B	A*02			0.02
p24-p2p7p1p6	230	7	VLAEAMSQV	B; A; C; D	A*02		0.09	0.28
p24-p2p7p1p6	230	7	VLAEAMSQA	B	A*02			0.45
p24-p2p7p1p6	230	8	VLAEAMSQVT	B	A*02			0.07
p2p7p1p6	1	10	AEAMSQVTNS	B	B*40; B*45			0.04
p2p7p1p6	1	10	AEAMSQANS	C	B*45			0.09
p2p7p1p6	2	16	EAMSQVTNSATIMMQ	B	A*02			0.01
p2p7p1p6	5	13	SQVTNSATI	B	A*02			0.02
p2p7p1p6	6	14	QVTNSATIM	B				0.02
p2p7p1p6	7	15	VTNSATIMM	B				0.02
p2p7p1p6	8	17	TNSANIMMQR	B				0.01
p2p7p1p6	9	17	NSATIMMQR	B				0.01
p2p7p1p6	14	22	MMQRGNFRN	B				0.04
p2p7p1p6	14	28	MMQRGNFRNQRKIVK	B				0.01
p2p7p1p6	16	24	QRGNFRNQR	B				0.04
p2p7p1p6	18	32	GNFRNQRKIVKCFNC	B				0.01
p2p7p1p6	18	37	SNFKGNKRMVKCFNCGKEGH	C	A*02		0.14	0.00
p2p7p1p6	19	33	NFRGPKRIKCFNCG	A				0.01
p2p7p1p6	21	29	RNQRKTVKC	B				0.02
p2p7p1p6	22	30	GPKRIVKCF	C	B*42			0.06
p2p7p1p6	23	33	SKRIVKCFNCG	C				0.06
p2p7p1p6	25	34	KTVKCFNCGR	B	A*11			0.00
p2p7p1p6	32	40	CGKEGHIAR	B	A*31			0.32

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p2p7p1p6	34	48	KEGHIAKNCRAPRKK	B				0.14
p2p7p1p6	37	52	HIARNCRAPRKKGCWK	C				0.28
p2p7p1p6	38	46	IARNCRAPR	C	A*30			0.31
p2p7p1p6	38	47	LARNCRAPRK	B	A*03; A*31			0.29
p2p7p1p6	38	48	IAKNCRAPRKK	C				0.17
p2p7p1p6	38	52	TARNCRAPRKKGCWK	B	A*03			0.00
p2p7p1p6	38	52	IAKNCRAPRKKGCWK	B				0.16
p2p7p1p6	42	50	CRAPRKKGC	B	B*14		0.01	0.77
p2p7p1p6	42	51	CRAPRKKGCW	B	B*57			0.77
p2p7p1p6	43	51	KAPRKKGCW	A				0.06
p2p7p1p6	43	51	RAPRKKGCW	C				0.77
p2p7p1p6	43	52	KAPRKKGCWK	B	A*11			0.06
p2p7p1p6	52	61	KCGKEGHQMK	B	A*11			0.79
p2p7p1p6	55	70	KEGHQMKDCTERQANF	B	A*02		0.05	0.64
p2p7p1p6	58	69	HQMKDCNERQAN	B; G	A*02			0.02
p2p7p1p6	60	74	MKDCTERQANFLGKI	B; CRF01_AE			0.05	0.65
p2p7p1p6	62	72	DCTERQANFLG	B; C				0.79
p2p7p1p6	63	71	CTERQANFL	B	B*40			0.83
p2p7p1p6	64	71	TERQANFL	B	B*40		0.09	0.84
p2p7p1p6	65	75	ERQANFLGKIW	C				0.77
p2p7p1p6	66	74	RQANFLGKI	B; C	B*13; B*48		0.02	0.79
p2p7p1p6	66	80	RQANFLGKIWPSYKG	B	A*02			0.01
p2p7p1p6	66	80	RQANFLGKIWPSHKG	B				0.41
p2p7p1p6	66	81	RQANFLGKIWPSHKGR	B; C	Cw*02		0.00	0.40
p2p7p1p6	70	77	FLGKIWPS	B; A; CRF02_AG	A*02		0.17	0.80
p2p7p1p6	70	79	FLGKIWPSHK	B; A; C; CRF01_AE	A*02		0.09	0.43
p2p7p1p6	70	79	FLGKIWPSYK	B; A; C; D	A*02		0.15	0.02
p2p7p1p6	70	84	FLGKIWPSYKGRPGN	B	A*02			0.01
p2p7p1p6	70	84	FLGKIWPSHKGRPGN	B; C				0.41
p2p7p1p6	72	89	GKIWPSHKGRPGNFLQSR	B; C				0.20
p2p7p1p6	73	81	KIWPSYKGR	B	A*31			0.02
p2p7p1p6	74	88	IWPSHKGRPGNFLQS	B				0.22
p2p7p1p6	82	96	PGNFLQSRPEPTAPP	B	A*02			0.28
p2p7p1p6	83	97	GNFLQSRPEPTAPPF	B	A*02			0.00
p2p7p1p6	85	94	FLQSRPEPTA	B	A*02		0.34	0.30
p2p7p1p6	89	97	RPEPTAPPA	C; CRF01_AE	B*07		0.04	0.53

Prot[a]	S[b]	E[c]	Epitope[d]	Subtype[e]	HLA[f]	L[g]	R[h]	C[i]
p2p7p1p6	91	100	EPTAPPEESF	D	B*35; B*58			0.07
p2p7p1p6	91	100	EPTAPPAESF	C	B*07		0.02	0.50
p2p7p1p6	91	102	EPTAPPAESFRF	C	B*58		0.01	0.30
p2p7p1p6	93	112	TAPPEESFRFGEETTTPSQK	B	Cw*08			0.02
p2p7p1p6	93	112	TAPPAESFRFEETTPAPK	C				0.12
p2p7p1p6	94	102	APPAESFRF	C	B*07		0.04	0.36
p2p7p1p6	97	107	AESFRFEET	CRF01_AE				0.27
p2p7p1p6	98	112	ESFRFGEETTTPSQK	B				0.02
p2p7p1p6	103	120	GEETTTPSQKQEPIDKEL	B				0.01
p2p7p1p6	108	116	TPSQKQEPI	B; D	B*35; B*53			0.02
p2p7p1p6	110	118	SQKQEQIDK	B	A*11			0.00
p2p7p1p6	111	127	QKQGTIDKELYPLASLK	B				0.00
p2p7p1p6	111	127	QKQEPIDKELYPLASLK	B				0.01
p2p7p1p6	113	121	QEPIDKELY	D	B*44			0.03
p2p7p1p6	114	123	EPKDREPL	C	B*08			0.11
p2p7p1p6	114	123	EPIDKELYPL	B; D	B*35; B*42; B*53		0.01	0.02
p2p7p1p6	114	124	EPIDKELYPLA	B				0.02
p2p7p1p6	118	126	KELYPLASL	B	A*02; B*40			0.07
p2p7p1p6	118	126	KELYPLTSL	B	B*40		0.06	0.02
p2p7p1p6	118	135	KELYPLASLRSLFGNDPS	B				0.01
p2p7p1p6	118	135	KELYPLASLKSLFGNDPS	B				0.02
p2p7p1p6	118	137	KEMYPLASLRSLFGNDPSSQ	B				0.00
p2p7p1p6	120	129	LYPLASLRSL	D	A*24			0.02
p2p7p1p6	121	129	YPLTSLRSL	B	B*35		0.03	0.02
p2p7p1p6	121	129	YPLASLRSL	B; D	B*53			0.03
p2p7p1p6	121	130	YPLASLRSLF	B; D	B*07; B*35		0.03	0.03
p2p7p1p6	121	130	YPLTSLRSLF	B	B*07			0.02
p2p7p1p6	122	132	PLTSLKSLFGS	C				0.23
p2p7p1p6	123	130	LASLRSLF	D	B*58			0.04

Extended Data Table 3. Global class I HLA allele group frequencies of loci A, B, and C (D.1.1.3). The frequencies were derived from the overall average HLA allele frequencies across 497 population samples as published by Solberg et al.[291]. The published high four-digit allele resolution was condensed to a low two digit resolution, by summing up all specific HLA protein frequencies to the respective allele group.

HLA-A allele	Frequency [%]	HLA-B allele	Frequency [%]	HLA-C allele	Frequency [%]
A*01	5.0	B*07	7.8	Cw*01	8.6
A*02	25.3	B*08	3.1	Cw*02	3.1
A*03	4.5	B*13	4.1	Cw*03	17.4
A*11	12.6	B*14	1.8	Cw*04	13.8
A*23	2.4	B*15	10.5	Cw*05	2.6
A*24	20.1	B*18	2.5	Cw*06	6.2
A*25	0.5	B*27	2.5	Cw*07	21.9
A*26	4.1	B*35	9.7	Cw*08	7.4
A*29	2.1	B*37	2.5	Cw*12	5.3
A*30	4.2	B*38	1.7	Cw*14	4.0
A*31	4.2	B*39	3.9	Cw*15	4.3
A*32	1.5	B*40	11.9	Cw*16	2.8
A*33	5.2	B*41	0.6	Cw*17	1.9
A*34	2.2	B*42	1.1	Cw*18	0.6
A*36	0.4	B*44	6.9		
A*43	0.0	B*45	1.0		
A*66	0.7	B*46	2.4		
A*68	3.8	B*47	0.2		
A*69	0.1	B*48	2.1		
A*74	0.8	B*49	0.9		
A*80	0.1	B*50	1.0		
A*92	0.0	B*51	5.9		
		B*52	2.3		
		B*53	1.7		
		B*54	1.4		
		B*55	1.7		
		B*56	2.0		
		B*57	1.6		
		B*58	3.7		
		B*59	0.2		
		B*67	0.2		
		B*73	0.1		
		B*78	0.2		
		B*81	0.4		
		B*82	0.1		
		B*95	0.0		

Extended Data Table 4. Ranking of all classification feature combinations (D.1.2.1). All possible feature permutations were examined in a 10-fold cross-validation (100 repeats) of the training-set. The listed feature combinations were sorted first by precision and then by accuracy.

Features					Accuracy [%]	Specificity [%]	Sensitivity [%]	Precision, PPV [%]	NPV [%]
1	2	3	4	5					
		x		x	86 ± 3	99 ± 3	75 ± 5	**98 ± 3**	78 ± 3
x		x		x	84 ± 3	99 ± 3	71 ± 5	**98 ± 3**	75 ± 4
			x		80 ± 2	87 ± 0	75 ± 4	**86 ± 1**	75 ± 3
			x	x	80 ± 3	87 ± 1	74 ± 5	**86 ± 1**	75 ± 4
	x		x		78 ± 1	87 ± 0	71 ± 2	**86 ± 0**	73 ± 2
	x				78 ± 1	87 ± 0	71 ± 1	**86 ± 0**	72 ± 1
		x	x		78 ± 2	87 ± 0	71 ± 4	**86 ± 1**	72 ± 3
	x			x	78 ± 2	87 ± 1	70 ± 3	**86 ± 1**	72 ± 2
	x		x	x	77 ± 2	87 ± 0	69 ± 5	**85 ± 1**	71 ± 3
		x	x	x	77 ± 3	87 ± 0	69 ± 6	**85 ± 1**	71 ± 4
	x	x			75 ± 1	87 ± 2	65 ± 2	**85 ± 2**	69 ± 1
	x	x		x	75 ± 2	87 ± 2	64 ± 3	**85 ± 2**	68 ± 2
	x	x	x		75 ± 1	87 ± 1	65 ± 2	**85 ± 1**	69 ± 1
x			x	x	75 ± 3	87 ± 1	64 ± 5	**84 ± 1**	68 ± 3
x			x		74 ± 3	87 ± 0	63 ± 5	**84 ± 1**	68 ± 3
	x	x	x	x	74 ± 2	87 ± 1	63 ± 4	**84 ± 1**	68 ± 2
x	x			x	73 ± 2	87 ± 0	61 ± 3	**84 ± 1**	66 ± 2
x	x				73 ± 2	87 ± 0	60 ± 3	**84 ± 1**	66 ± 2
x	x		x	x	73 ± 2	87 ± 0	60 ± 4	**84 ± 1**	66 ± 2
x	x		x		72 ± 2	87 ± 1	59 ± 4	**84 ± 1**	65 ± 2
x		x	x	x	72 ± 3	87 ± 0	60 ± 5	**84 ± 1**	66 ± 3
x		x	x		72 ± 3	87 ± 1	59 ± 5	**83 ± 1**	65 ± 3
x	x	x			71 ± 2	87 ± 1	57 ± 3	**83 ± 2**	64 ± 2
x	x	x		x	71 ± 2	87 ± 0	57 ± 4	**83 ± 1**	64 ± 2
x	x	x	x	x	70 ± 2	87 ± 1	56 ± 4	**83 ± 2**	64 ± 2
x		x			79 ± 2	81 ± 5	76 ± 1	**82 ± 4**	75 ± 2
x	x	x	x		70 ± 2	87 ± 1	55 ± 4	**82 ± 1**	63 ± 2
		x			79 ± 3	77 ± 6	80 ± 3	**80 ± 4**	77 ± 3
				x	72 ± 2	75 ± 3	69 ± 2	**76 ± 2**	68 ± 2
x				x	67 ± 3	73 ± 3	62 ± 6	**72 ± 3**	63 ± 4
x					58 ± 3	67 ± 3	51 ± 4	**64 ± 3**	55 ± 2

Extended Data Table 5. Number of epitopes and antigen scores for an extended antigen panel (see D.2.2).

	Gag antigen	Number of sequences	Number of epitopes	Antigen score
---Monovalent---	teeGag1	1	402	590
	teeGag2	1	266	393
	teeGag3	1	222	332
	conM	1	283	424
	conA1	1	238	350
	conA2	1	226	336
	conB	1	401	542
	conC	1	271	472
	conD	1	288	456
	conF1	1	247	412
	conG	1	231	353
	conH	1	217	330
	conK	1	223	350
	con01_AE	1	199	307
	con02_AG	1	196	295
	con03_AB	1	220	359
	con04_CPX	1	195	310
	con06_CPX	1	251	397
	con07_BC	1	235	369
	con08_BC	1	249	415
	con10_CD	1	241	360
	con11_CPX	1	222	302
	con12_BF	1	242	357
	con14_BG	1	219	338
	ancM	1	293	426
	ancA1	1	250	374
	ancB	1	352	465
	ancC	1	269	430
	mosM1.1	1	277	466
	mosB1.1	1	400	542
	mosC1.1	1	275	478
	cotM	1	311	450
	cotB	1	393	537
	cotC	1	275	480
	HXB2 (HVTN 505)	1	380	522
	CAM1 (HVTN	1	381	524
	LAI IIIb (RV144)	1	382	524
---Trivalent---	teeGag1-3	3	588	949
	conA1+B+C	3	551	867
	ancA1+B+C	3	493	752
	mosM3.1-3	3	524	808
	mosB3.1-3	3	484	791
	mosC3.1-3	3	341	537
	cotM+B+C	3	504	828

Extended Data Table 6. List of all budding-retaining AAS incorporated in teeGag1-3. [a]Each budding-retaining AAS was assigned a unique code based on their HXB2-Gag reference position. If one site harbored more than one mutation they were distinguished by adding a character. [b]The mutation description consists of the original HXB2-Gag amino acid, the location numbering relative to HXB2-Gag, and the new amino acid introduced through the AAS. [c]Relative budding, determined as described in D.3.1 .For generation of Gag sequences with a single AAS, the [d]forward and [e]reverse mutagenesis primers used are given. For * marked list entries the AAS mutated Gag sequences were provided by GeneArt.

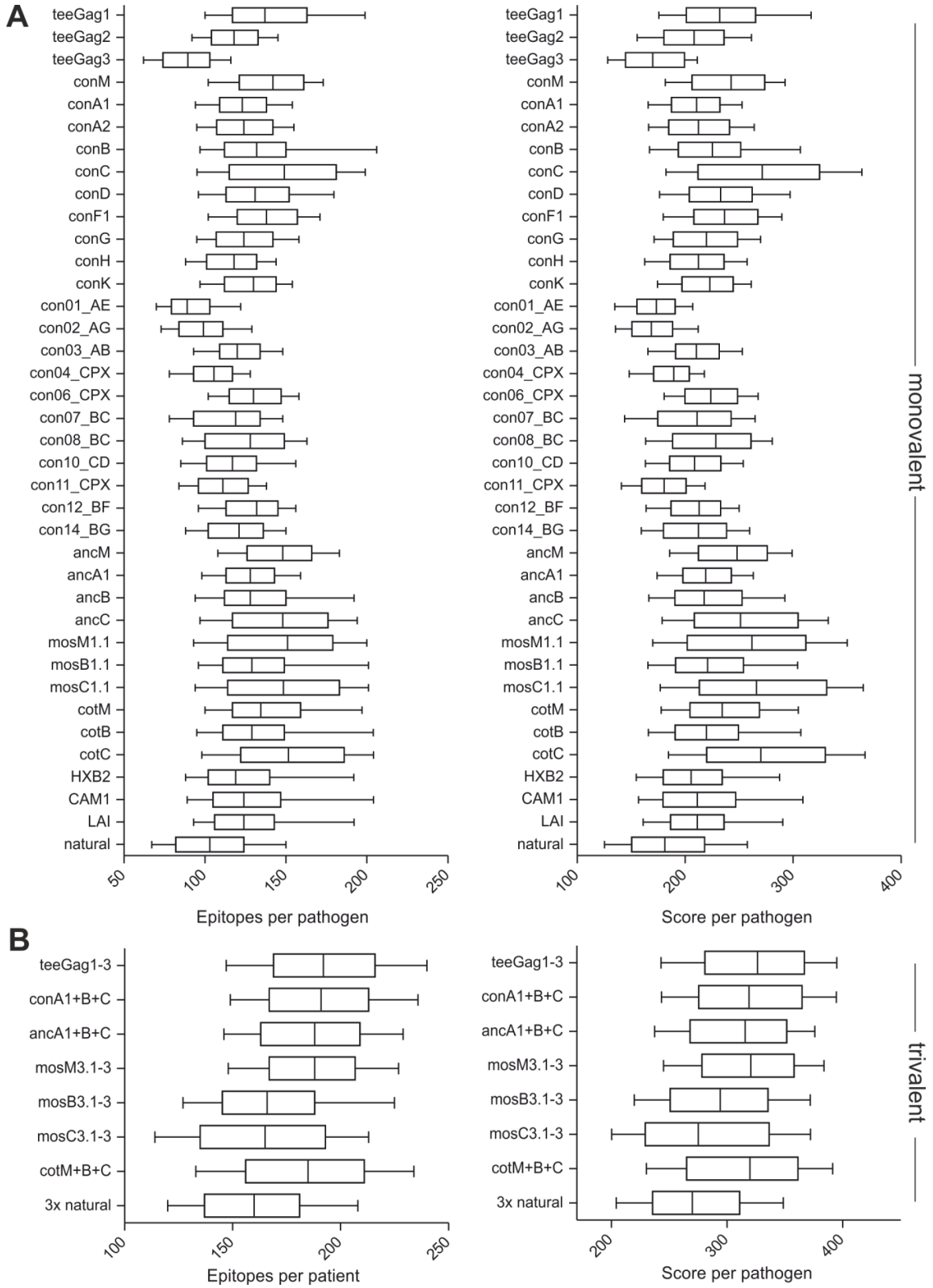
Code[a]	Mutation[b]	RB[c]	fwd mutagenesis primer[d] (5'→3')	rev-mutagenesis primer[e] (5'→3')
12a	E12Q	1.20	*	
12b	E12K	1.42	CTGGCGGCAAGCTGGACAGATGGGAGAAG	CCAGCTTGCCGCCAGACAGCAC
15	R15S	1.32	GCTGGACAGCTGGGAGAAGATCCGGCTG	CTTCTCCCAGCTGTCCAGCTCGCCG
28	K28H	1.22	GCAAGAAGCACTACAAGCTGAAGCACATCGTG	CTTCAGCTTGTAGTGCTTCTTGCCGCCAGGTCTC
30a	K30R	1.27	GAAGTACAGACTGAAGCACATCGTGTGGG	CGATGTGCTTCAGTCTGTACTTCTTCTTGCCGCCAG
30b	K30M	0.65	GAAGTACATGCTGAAGCACATCGTGTGGG	CGATGTGCTTCAGCATGTACTTCTTCTTGCCGCCAG
34	I34L	1.85	*	
46	V46L	1.84	TTCGCCCTGAACCCCGGCCTCC	CGGGGTTCAGGGCGAATCTTTCCAGCTCTC
53	T53S	0.92	*	
58	R58K	1.21	GGGCTGCAAGCAGATCCTGGGCCAGC	GGATCTGCTTGCAGCCCTCGCTGG
72	S72T	0.98	GACAGGCACCGAGGAACTGCGGAGC	GTTCCTCGGTGCCTGTCTGCAGGC
76	R76K	1.12	CGAGGAACTGAAGAGCCTGTACAACACCGTG	GTACAGGCTCTTCAGTTCCTCGCTGCCTG
91	R91K	1.52	GTGCACCAGAAGATCGAGATCAAGGACACCAAAG	GATCTCGATCTTCTGGTGCACGCAGTACAG
93	E93D	1.87	*	
94	I94V	1.53	*	
102	D102E	1.65	*	
115	A115T	1.71	*	
122	T122K	0.85	*	
124	H124N	1.14	*	
125	S125G	1.58	GGCCACGGCAACCAGGTGTCCCAG	CTGGTTGCCGTGGCCGGTGTCGG
126a	N126S	1.20	*	
126b	N126G	1.11	CACAGCGGCCAGGTGTCCCAGAACTACC	CACCTGGCCGCTGTGGCCGGTGTC
126c	N126K	1.56	CCACAGCAAGCAGGTGTCCCAGAACTACCC	GACACCTGCTTGCTGTGGCCGGTG
127	Q127K	1.14	CCACAGCAACAAGGTGTCCCAGAACTACCC	GGACACCTTGTTGCTGTGGCCGGTG
138	I138L	1.21	*	
146	A146S	1.61	GCACCAGAGCATCAGCCCCCGGAC	GGCTGATGCTCTGGTGCACCATCTGGC
147	I147L	0.88	CAGGCCCTGAGCCCCCGGACCC	GGGGCTCAGGGCCTGGTGCACCATC
159	V159I	1.06	*	
173	S173T	1.29	*	
186	T186M	1.07	*	
190	T190I	1.36	*	
203	E203D	1.01	*	
215	V215L	1.07	*	
223	I223V	1.22	GACCTGTGGCCCCTGGCCAGATG	CAGGGGCCACAGGTCCGGCGTGCA
247	I247V	0.86	GGAACAGGTGGGCTGGATGACCAACAACC	CCAGCCCACCTGTTCCTGCAGGGTGC
248	G248A	1.05	GAACAGATCGCCTGGATGACCAACAACCCC	GTCATCCAGGCGATCTGTTCCTGCAGGG
252	N252S	1.02	*	
256	I256V	0.85	CCCCGTGCCCGTGGGCGAGATC	CACGGGCACGGGGGGGTTGTTGGTCA
260	E260D	0.77	*	
268	L268M	1.18	GTGGATCATCATGGGCCTGAACAAGATCGTG	CAGGCCCATGATGATCCACCGCTTGTAGATC
280	T280V	1.32	*	
286	R286K	0.98	CTGGACATCAAGCAGGGACCCAAAGAGC	GTCCCTGCTTGATGTCCAGGATGCTGGTG
301	Y301F	0.94	*	
310	S310T	1.00	*	
312	E312D	1.31	CAGCCAGGACGTCAAGAACTGGATGACCGAG	GTTCTTGACGTCCTGGCTGGCCTGC
319	E319D	1.04	*	
326	A326S	0.97	*	
335	K335R	1.46	*	
340	A340G	1.33	GCCCTGGCGCCACCCTGGAAGAGA	GTGGCGCCAGGGCCCAGGGCC
342	T342S	1.49	GCCGCCAGCCTGGAAGAGATGATGACCG	CCAGGCTGGCGGCAGGGCC
357	G357S	1.74	*	
362	V362I	1.83	GCCAGAATCCTGGCCGAGGCCATG	CGGCCAGGATTCTGGCTTTGTGTCCAGGTC
370	V370A	1.05	*	
371	T371N	1.01	CAGGTCAACAACAGCGCCACCATCATG	GGCGCTGTTGTTGACCTGGCTCATGGC
372	N372S	1.31	GGTCACAAGCAGCGCCACCATCATG	GGCGCTGCTTGTGACCTGGCTCATGG

373	S373A	1.36	GTCACAAACGCCGCCACCATCATGATGC	GTGGCGGCGTTTGTGACCTGGCTCATGG
375	T375N	0.80	*	
381	G381S	1.07	GCAGCGGAGCAACTTCCGGAACCAGC	GGAAGTTGCTCCGCTGCATCATGATGG
384	R384K	1.36	GGGCAACTTCAAGAACCAGCGGAAGATCGTG	GCTGGTTCTTGAAGTTGCCCCGCTGC
385	N385G	0.99	*	
386a	Q386N	0.65	CGGAACAACCGGAAGATCGTGAAGTGCTTC	CGATCTTCCGGTTGTTCCGGAAGTTGCCCC
386b	Q386S	1.22	*	
387	R387K	1.21	*	
388	K388R	0.94	*	
389a	I389M	1.36	CAGCGGAAGATGGTGAAGTGCTTCAACTGCGG	CACTTCACCATCTTCCGCTGGTTCCGG
389b	I389T	0.85	GCGGAAGACCGTGAAGTGCTTCAACTGCG	CACTTCACGGTCTTCCGCTGGTTCCGG
397	K397R	1.36	CTGCGGCAGAGAGGGCCACACCGC	CCCTCTCTGCCGCAGTTGAAGCAC
401	T401I	0.82	*	
403	R403K	1.02	*	
427	T427N	0.85	*	
460a	E460A	1.19	CTCCCGCCGAAAGCTTCAGATCCGGCG	GAAGCTTTCGGCGGGAGGGGCGGTG
460b	E460F	1.17	*	
465	S465F	1.28	GCTTCAGATTCGGCGTGGAAACCACC	CCACGCCGAATCTGAAGCTTTCCTCGGGAG
466	G466E	1.00	CTTCAGATCCGAGGTGGAAACCACCACCCCC	GGTTTCCACCTCGGATCTGAAGCTTTCCTCG
467	V467E	1.06	*	
468	E468T	0.83	CGGCGTGACAACCACCACCCCCCC	GGTGGTTGTCACGCCGGATCTGAAGC
470	T470P	1.36	GGAAACCCCCACCCCCCCACAGAAG	GGGGGTGGGGGTTTCCACGCCGGATC
471	T471A	1.13	CCACCGCCCCCCCACAGAAGCAG	GGGGGGGCGGTGGTTTCCACGCCG
473a	P473K	0.72	CCACCCCCAAACAGAAGCAGGAACCCATCG	GCTTCTGTTTGGGGGTGGTGGTTTCCAC
473b	P473S	0.78	*	
478	P478Q	0.89	GAAGCAGGAACAGATCGACAAAGAGCTGTACCCC	CTTTGTCGATCTGTTCCTGCTTCTGTGGGG
487	T487A	1.00	*	
490	R490K	1.59	CAGCCTGAAGAGCCTGTTCGGCAACG	GAACAGGCTCTTCAGGCTGGTCAGGGG
495	N495S	1.02	CGGCAGCGACCCCAGCAGCC	GGGGTCGCTGCCGAACAGGCTTCTCAG

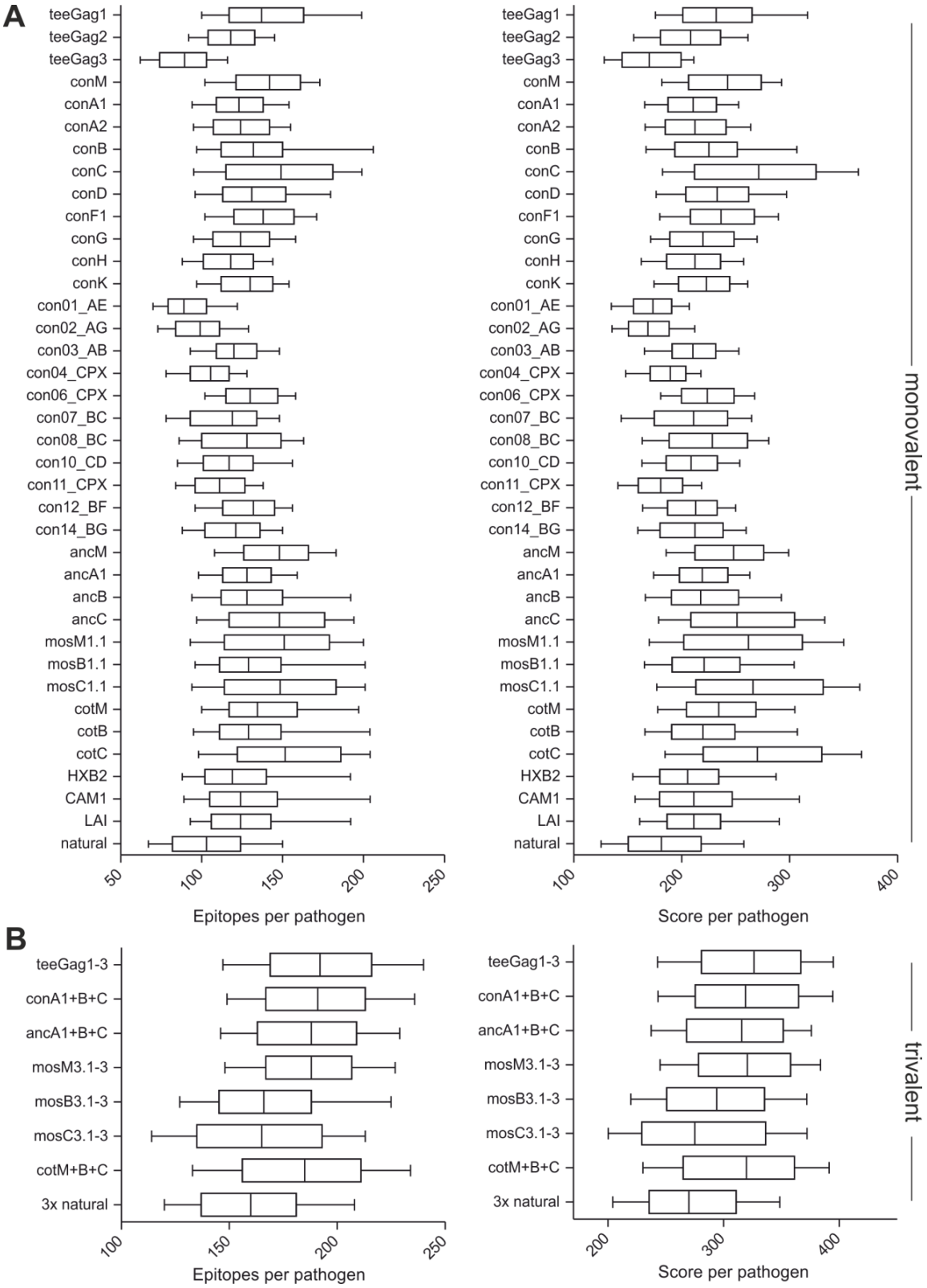
Extended Data Table 7. List of all budding-deleterious AAS identified in the epitope set. [a]Each as budding-deleterious predicted AAS was assigned a unique code based on their HXB2-Gag reference position followed by the letter "N" for negative. If one site harbored more than one mutation, they were distinguished by adding a character. [b]The mutation description consists of the original HXB2-Gag amino acid, the location numbering relative to HXB2-Gag, and the new amino acid introduced by the ASS. The AAS are classified according to the outcome of the [c]experimental validation as false or true negatives. [d]Relative budding, determined as described in D.3.1. For generation of Gag sequences with a single AAS the [e]forward and [f]reverse mutagenesis primers used are given.

Code[a]	Mutation [b]	Exp. Valid.[c]	RB[d]	fwd mutagenesis primer[e] (5'→3')	rev-mutagenesis primer[f] (5'→3')
9N	S9R	false neg.	0.83	GTGCTGAGAGGCGGCGAGCTG	GCCGCCTCTCAGCACAGAGGCTCTGG
62Na	G62K	false neg.	1.25	GATCCTGAAGCAGCTGCAGCCCAGC	GCAGCTGCTTCAGGATCTGTCTGCAGCC
62Nb	G62E	false neg.	1.26	GATCCTGGAGCAGCTGCAGCCCAGC	GCAGCTGCTCCAGGATCTGTCTGCAGC
151N	T151L	true neg.	0.58	CCCCGGCTGCTGAACGCCTGGGTCA	GTTCAGCAGCCGGGGGCTGATG
199N	Q199E	false neg.	1.17	CCGCCATGGAGATGCTGAAAGAGACAATCAA	GCATCTCCATGGCGGCCTGG
267N	I267L	false neg.	1.09	GTGGATCCTGCTGGGCCTGAACAAGATCG	GCCCAGCAGGATCCACCGCTTGTAGATCTC
268N	L268G	true neg.	0.56	ATCATCGGCGGCCTGAACAAGATCGTG	AGGCCGCCGATGATCCACCGCTTGTAG
269N	G269L	true neg.	0.00	CATCCTGCTGCTGAACAAGATCGTGCGG	TGTTCAGCAGCAGGATGATCCACCGC
270N	L270N	true neg.	0.00	CCTGGGCAACAACAAGATCGTGCGGATG	TCTTGTTGTTGCCCAGGATGATCCAC
271Na	N271H	false neg.	1.00	GGGCCTGCACAAGATCGTGCGGATG	GATCTTGTGCAGGCCCAGGATGATC
271Nb	N271K	false neg.	2.15	GGCCTGAAGAAGATCGTGCGGATGTACA	CACGATCTTCTTCAGGCCCAGGATGAT
273N	I273V	false neg.	0.74	GAACAAGGTGGTGCGGATGTACAGCCC	CCGCACCACCTTGTTCAGGCCCAGG
294N	R294I	true neg.	0.31	AGCCCTTCATCGACTACGTGGACCGGTTC	CGTAGTCGATGAAGGGCTCTTTGGGTC
301N	Y301W	true neg.	0.57	CCGGTTCTGGAAGACCCTGCGGGC	GGGTCTTCCAGAACCGGTCCACGTAGT
311N	Q311A	true neg.	0.73	CCAGCGCCGAAGTCAAGAACTGGATGACC	GACTTCGGCGCTGGCCTGCTCGG
316N	W316M	false neg.	1.29	GTCAAGAACATGATGACCGAGACACTGCTG	TCGGTCATCATGTTCTTGACTTCCTGGCTG
392N	C392F	true neg.	0.31	TCGTGAAGTTCTTCAACTGCGGCAAA	GCAGTTGAAGAACTTCACGATCTTCCGCT
395N	C395G	true neg.	0.46	CTTCAACGGCGGCAAAGAGGGC	TGCCGCCGTTGAAGCACTTCACGATCTT

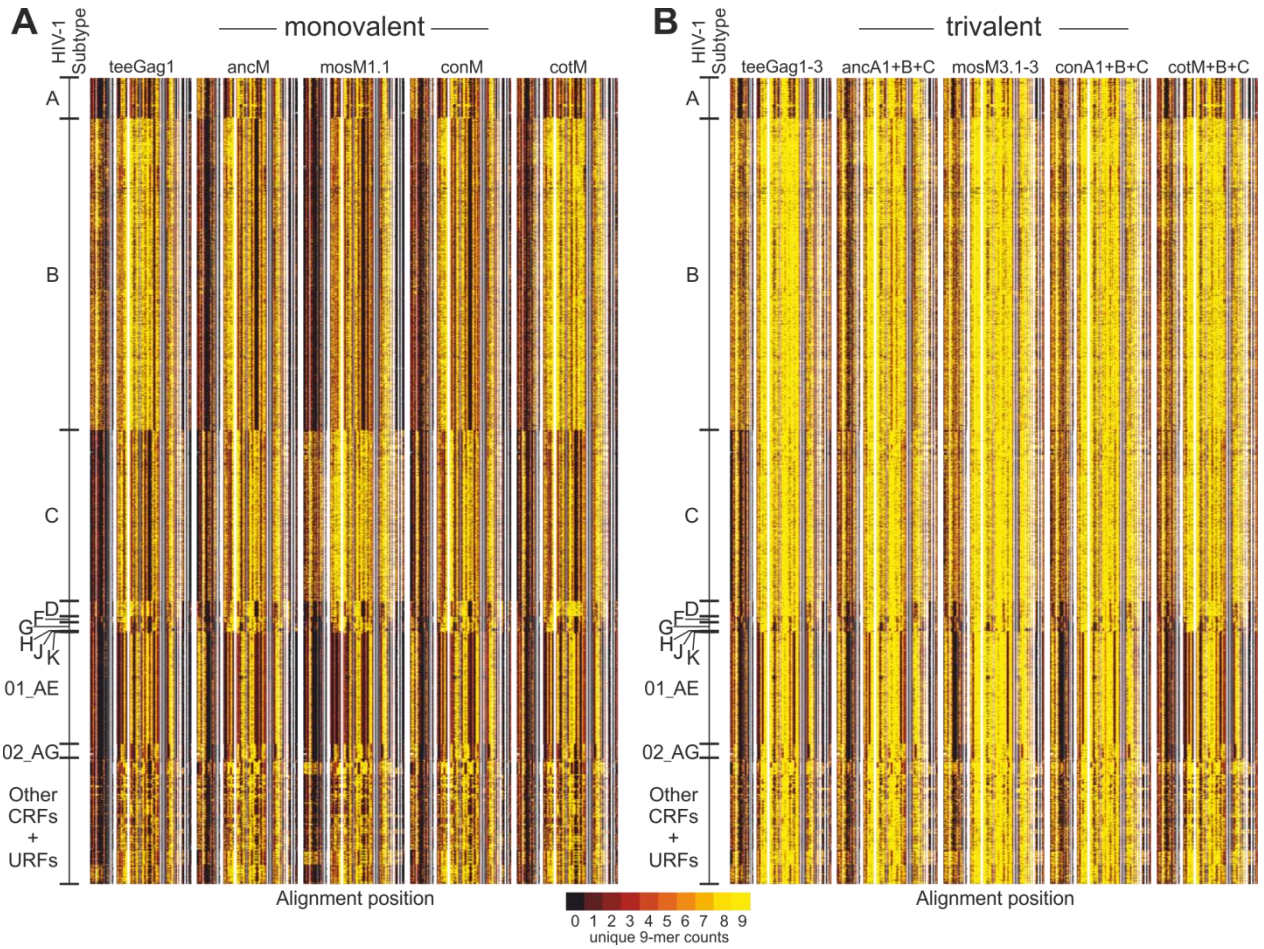
F.2.2 Extended Data Figures



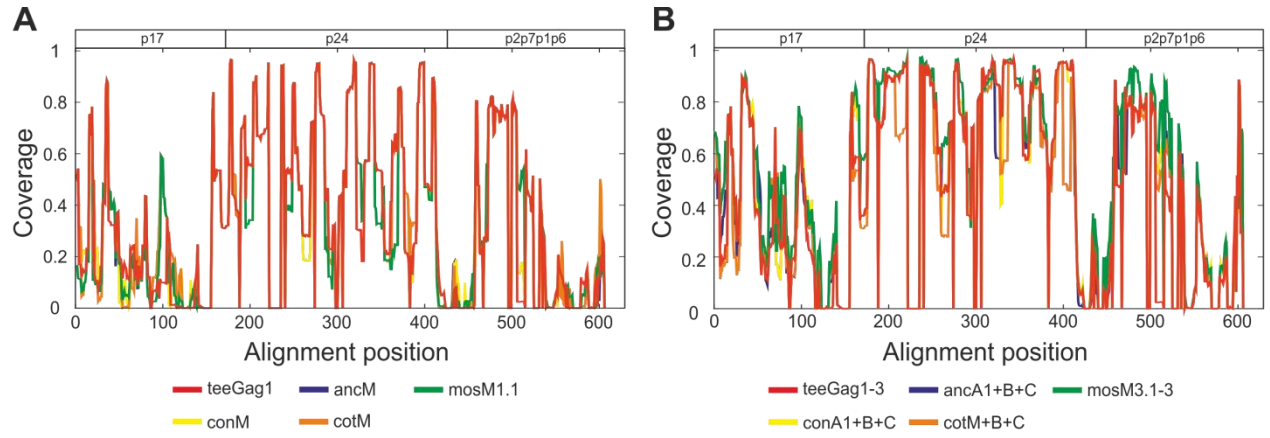
Extended Data Figure 1. Population coverage for an extended antigen panel (D.2.3). Number of epitopes (left panel) and score (right panel) per pathogen of (A) mono- or (B) trivalent antigen sets are displayed as boxes that represent the median with 50 % quartiles, whiskers represent the 10 and 90 % percentiles.



Extended Data Figure 2. Pathogen coverage for an extended antigen panel (D.2.4). Number of epitopes (left panel) and score (right panel) per pathogen of (A) mono- or (B) trivalent antigen sets are displayed as boxes that represent the median with 50 % quartiles, whiskers represent the 10 and 90 % percentiles.



Extended Data Figure 3. 9-mer coverage by mono- or trivalent antigen sets mapped on the curated filtered web alignment (D.2.5). (A) Mono- or (B) trivalent antigen sets analyzed are indicated above each column. Each row represents one sequence of the Gag alignment. Each column of the row indicates one amino acid of the respective sequence, color-coded by the number of matched unique 9-mers. The alignment sequences were grouped according to their subtype, as indicated at the left edge of both graphs.



Extended Data Figure 4. Antigen-matched 9-mers in the alignment sequence (D.2.5). Filtered web alignment position-specific 9-mer coverage-rates of (A) mono- or (B) trivalent antigen sets are plotted on the y-axis. Positions on the x-axis are according to the natural alignment position, partitioned into p17, p24, and p2p7p1p6 as indicated above the graphs.

F.2.3 Extended Data Sequences

F.2.3.1 HXB2-Gag and teeGag1-3 protein sequences

Variant positions in teeGag1-3 compared to HXB2-Gag are highlighted orange.

>HXB2-Gag reference sequence

```
MGARASVLSG GELDRWEKIR LRPGGKKKYK LKHIVWASRE LERFAVNPGL
LETSEGCRQI LGQLQPSLQT GSEELRSLYN TVATLYCVHQ RIEIKDTKEA
LDKIEEEQNK SKKKAQQAAA DTGHSNQVSQ NYPIVQNIQG QMVHQAISPR
TLNAWVKVVE EKAFSPEVIP MFSALSEGAT PQDLNTMLNT VGGHQAAMQM
LKETINEEAA EWDRVHPVHA GPIAPGQMRE PRGSDIAGTT STLQEQIGWM
TNNPPIPVGE IYKRWIILGL NKIVRMYSPT SILDIRQGPK EPFRDYVDRF
YKTLRAEQAS QEVKNWMTET LLVQNANPDC KTILKALGPA ATLEEMMTAC
QGVGGPGHKA RVLAEAMSQV TNSATIMMQR GNFRNQRKIV KCFNCGKEGH
TARNCRAPRK KGCWKCGKEG HQMKDCTERQ ANFLGKIWPS YKGRPGNFLQ
SRPEPTAPPE ESFRSGVETT TPPQKQEPID KELYPLTSLR SLFGNDPSSQ
```

>teeGag1 - 28 variant positions

```
MGARASVLSG GELDRWEKIR LRPGGKKKYK LKHIVWASRE LERFAVNPGL
LETSEGCRQI LGQLQPSLQT GSEELRSLYN TVATLYCVHQ RIEIKDTKEA
LEKIEEEQNK SKKKTQQAAA DKGNSSQVSQ NYPIVQNLQG QMVHQAISPR
TLNAWVKVIE EKAFSPEVIP MFSALSEGAT PQDLNTMLNT VGGHQAAMQM
LKETINEEAA EWDRLHPVHA GPIAPGQMRE PRGSDIAGTT STLQEQIGWM
TNNPPIPVGE IYKRWIILGL NKIVRMYSPT SILDIRQGPK EPFRDYVDRF
YKTLRAEQAS QEVKNWMTET LLVQNANPDC KTILRALGPG ATLEEMMTAC
QGVGGPGHKA RVLAEAMSQV TNSATIMMQR SNFKGNKRMV KCFNCGKEGH
IAKNCRAPRK KGCWKCGKEG HQMKDCTERQ ANFLGKIWPS YKGRPGNFLQ
SRPEPTAPPA ESFRFEETTP APKQKQEPID KELYPLASLR SLFGNDPSSQ
```

>teeGag2 - 39 variant positions

```
MGARASVLSG GQLDRWEKIR LRPGGKKKYR LKHLVWASRE LERFALNPGL
LETSEGCKQI LGQLQPSLQT GTEELRSLYN TVATLYCVHQ RIEVKDTKEA
LDKIEEEQNK SKKKAQQAAA DTGHSGKVSQ NYPIVQNLQG QMVHQAISPR
TLNAWVKVVE EKAFSPEVIP MFTALSEGAT PQDLNMMLNI VGGHQAAMQM
LKDTINEEAA EWDRVHPVHA GPIAPGQMRE PRGSDIAGTT STLQEQIAWM
TSNPPIPVGD IYKRWIILGL NKIVRMYSPV SILDIKQGPK EPFRDYVDRF
FKTLRAEQAT QDVKNWMTDT LLVQNANPDC KTILKALGPA ATLEEMMTAC
QGVGGPSHKA RVLAEAMSQA NSAATIMMQR GNFRNSKRIV KCFNCGKEGH
IARNCRAPRK KGCWKCGKEG HQMKDCNERQ ANFLGKIWPS YKGRPGNFLQ
SRPEPTAPPF ESFRFGEETT TPSQKQEPID KELYPLASLK SLFGNDPSSQ
```

>teeGag3 - 40 variant positions

```
MGARASVLSG GKLDSWEKIR LRPGGKKHYM LKHLVWASRE LERFAVNPGL
LESSEGCRQI LGQLQPSLQT GSEELKSLYN TVATLYCVHQ KIDIKDTKEA
LDKIEEEQNK SKKKAQQAAA DTGHGKKVSQ NYPIVQNIQG QMVHQSLSPR
TLNAWVKVIE EKAFSPEVIP MFTALSEGAT PQDLNTMLNT VGGHQAAMQM
LKETINEEAA EWDRVHPVHA GPVAPGQMRE PRGSDIAGTT STLQEQVAWM
TSNPPVPVGD IYKRWIIMGL NKIVRMYSPV SILDIRQGPK EPFRDYVDRF
YKTLRAEQAS QDVKNWMTET LLVQNSNPDC KTILRALGPG ASLEEMMTAC
QGVGGPSHKA RILAEAMSQV TNSANIMMQR GNFRNQRKTV KCFNCGREGH
TARNCRAPRK KGCWKCGKEG HQMKDCTERQ ANFLGKIWPS YKGRPGNFLQ
SRPEPTAPPE ESFRFGEETT TPSQKQEQID KELYPLTSLK SLFGSDPSSQ
```

F.3 References

1. Gallo, R. C. *et al.* Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science* **220,** 865–867 (1983).
2. Barré-Sinoussi, F. *et al.* Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220,** 868–871 (1983).
3. Centers for Disease Control (CDC). Pneumocystis pneumonia--Los Angeles. *MMWR Morb. Mortal. Wkly. Rep.* **30,** 250–252 (1981).
4. Cohen, M. S., Hellmann, N., Levy, J. A., DeCock, K. & Lange, J. The spread, treatment, and prevention of HIV-1: evolution of a global pandemic. *J. Clin. Invest.* **118,** 1244–1254 (2008).
5. UNAIDS. *Global AIDS Update -UNAIDS report*. (2016).
6. UNAIDS. *Global AIDS Response Progress Reporting (GARPR)*. (2016).
7. UNAIDS. *UNAIDS - Fact sheet 2016*. (2016).
8. Hemelaar, J., Gouws, E., Ghys, P. D., Osmanov, S. & WHO-UNAIDS Network for HIV Isolation and Characterisation. Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS Lond. Engl.* **25,** 679–689 (2011).
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,** 607–614 (2000).
10. Korber, B. *et al.* Timing the ancestor of the HIV-1 pandemic strains. *Science* **288,** 1789–1796 (2000).
11. Sharp, P. M. & Hahn, B. H. The evolution of HIV-1 and the origin of AIDS. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **365,** 2487–2494 (2010).
12. Ndung'u, T. & Weiss, R. A. On HIV diversity. *AIDS Lond. Engl.* **26,** 1255–1260 (2012).
13. Faivre-Moskalenko, C. *et al.* RNA control of HIV-1 particle size polydispersity. *PloS One* **9,** e83874 (2014).
14. Liu, J., Bartesaghi, A., Borgnia, M. J., Sapiro, G. & Subramaniam, S. Molecular architecture of native HIV-1 gp120 trimers. *Nature* **455,** 109–113 (2008).
15. Zhu, P. *et al.* Distribution and three-dimensional structure of AIDS virus envelope spikes. *Nature* **441,** 847–852 (2006).
16. Briggs, J. A. G. *et al.* The mechanism of HIV-1 core assembly: insights from three-dimensional reconstructions of authentic virions. *Struct. Lond. Engl. 1993* **14,** 15–20 (2006).
17. Modrow, S., Falke, D., Truyen, U. & Schätzl, H. *Molekulare Virologie*. (Spektrum Akademischer Verlag, 2010).
18. 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).
19. Rizzuto, C. D. *et al.* A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. *Science* **280,** 1949–1953 (1998).
20. Huang, H., Chopra, R., Verdine, G. L. & Harrison, S. C. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* **282,** 1669–1675 (1998).
21. Engelman, A. & Cherepanov, P. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat. Rev. Microbiol.* **10,** 279–290 (2012).
22. Bryant, M. & Ratner, L. Myristoylation-dependent replication and assembly of human immunodeficiency virus 1. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 523–527 (1990).
23. Zhou, W., Parent, L. J., Wills, J. W. & Resh, M. D. Identification of a membrane-binding domain within the amino-terminal region of human immunodeficiency virus type 1 Gag protein which interacts with acidic phospholipids. *J. Virol.* **68,** 2556–2569 (1994).
24. Levin, J. G., Guo, J., Rouzina, I. & Musier-Forsyth, K. Nucleic acid chaperone activity of HIV-1 nucleocapsid protein: critical role in reverse transcription and molecular mechanism. *Prog. Nucleic Acid Res. Mol. Biol.* **80,** 217–286 (2005).
25. Checkley, M. A., Luttge, B. G. & Freed, E. O. HIV-1 envelope glycoprotein biosynthesis, trafficking, and incorporation. *J. Mol. Biol.* **410,** 582–608 (2011).
26. Maartens, G., Celum, C. & Lewin, S. R. HIV infection: epidemiology, pathogenesis, treatment, and prevention. *Lancet Lond. Engl.* **384,** 258–271 (2014).
27. Quinn, T. C. *et al.* Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. *N. Engl. J. Med.* **342,** 921–929 (2000).
28. Crawford, N. D. & Vlahov, D. Progress in HIV reduction and prevention among injection and noninjection drug users. *J. Acquir. Immune Defic. Syndr. 1999* **55 Suppl 2,** S84-87 (2010).
29. Keele, B. F. *et al.* Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 7552–7557 (2008).
30. Cohen, M. S., Shaw, G. M., McMichael, A. J. & Haynes, B. F. Acute HIV-1 Infection. *N. Engl. J. Med.* **364,** 1943–1954 (2011).
31. Walker, L. M. *et al.* Broad neutralization coverage of HIV by multiple highly potent antibodies. *Nature* **477,** 466–470 (2011).
32. Bacchetti, P. *et al.* Survival patterns of the first 500 patients with AIDS in San Francisco. *J. Infect. Dis.* **157,** 1044–1047 (1988).
33. WHO. *Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis for HIV*. (2015).
34. Walker, B. D. & Hirsch, M. S. Antiretroviral therapy in early HIV infection. *N. Engl. J. Med.* **368,** 279–281 (2013).
35. Rodger, A. J. *et al.* Mortality in well controlled HIV in the continuous antiretroviral therapy arms of the SMART and ESPRIT trials compared with the general population. *AIDS Lond. Engl.* **27,** 973–979 (2013).
36. Chun, T. W. *et al.* Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature* **387,** 183–188 (1997).
37. Finzi, D. *et al.* Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat. Med.* **5,** 512–517 (1999).
38. Banga, R. *et al.* PD-1(+) and follicular helper T cells are responsible for persistent HIV-1 transcription in treated aviremic individuals. *Nat. Med.* **22,** 754–761 (2016).
39. Siliciano, J. D. *et al.* Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat. Med.* **9,** 727–728 (2003).
40. Check Hayden, E. How to beat HIV. *Nature* **523,** 146–148 (2015).
41. Flynn, N. M. *et al.* Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *J. Infect. Dis.* **191,** 654–665 (2005).
42. 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).
43. Buchbinder, S. P. *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. *Lancet Lond. Engl.* **372,** 1881–1893 (2008).
44. Gray, G. E. *et al.* 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.* **11,** 507–515 (2011).
45. Hammer, S. M. *et al.* Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. *N. Engl. J. Med.* **369,** 2083–2092 (2013).
46. Rerks-Ngarm, S. *et al.* Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *N. Engl. J. Med.* **361,** 2209–2220 (2009).
47. Haynes, B. F. *et al.* Immune-correlates analysis of an HIV-1 vaccine efficacy trial. *N. Engl. J. Med.* **366,** 1275–1286 (2012).
48. Corey, L. *et al.* Immune correlates of vaccine protection against HIV-1 acquisition. *Sci. Transl. Med.* **7,** 310rv7 (2015).
49. Euler, Z. *et al.* Cross-reactive neutralizing humoral immunity does not protect from HIV type 1 disease progression. *J. Infect. Dis.* **201,** 1045–1053 (2010).
50. Klein, F. *et al.* HIV therapy by a combination of broadly neutralizing antibodies in humanized mice. *Nature* **492,** 118–122 (2012).
51. Caskey, M. *et al.* Viraemia suppressed in HIV-1-infected humans by broadly neutralizing antibody 3BNC117. *Nature* **522,** 487–491 (2015).
52. Parren, P. W. *et al.* Antibody protects macaques against vaginal challenge with a pathogenic R5 simian/human immunodeficiency virus at serum levels giving complete neutralization in vitro. *J. Virol.* **75,** 8340–8347 (2001).
53. Gautam, R. *et al.* A single injection of anti-HIV-1 antibodies protects against repeated SHIV challenges. *Nature* **533,** 105–109 (2016).
54. Liao, H.-X. *et al.* Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496,** 469–476 (2013).
55. Klein, F. *et al.* Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. *Cell* **153,** 126–138 (2013).
56. Sliepen, K. & Sanders, R. W. HIV-1 envelope glycoprotein immunogens to induce broadly neutralizing antibodies. *Expert Rev. Vaccines* **15,** 349–365 (2016).
57. Borrow, P., Lewicki, H., Hahn, B. H., Shaw, G. M. & Oldstone, M. B. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J. Virol.* **68,** 6103–6110 (1994).
58. Migueles, S. A. & Connors, M. Success and failure of the cellular immune response against HIV-1. *Nat. Immunol.* **16,** 563–570 (2015).
59. Rowland-Jones, S. *et al.* HIV-specific cytotoxic T-cells in HIV-exposed but uninfected Gambian women. *Nat. Med.* **1,** 59–64 (1995).
60. Ruiz-Riol, M. *et al.* Alternative effector-function profiling identifies broad HIV-specific T-cell responses in highly HIV-exposed individuals who remain uninfected. *J. Infect. Dis.* **211,** 936–946 (2015).
61. Walker, B. & McMichael, A. The T-cell response to HIV. *Cold Spring Harb. Perspect. Med.* **2,** (2012).
62. Picker, L. J., Hansen, S. G. & Lifson, J. D. New paradigms for HIV/AIDS vaccine development. *Annu. Rev. Med.* **63,** 95–111 (2012).
63. Kiepiela, P. *et al.* CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nat. Med.* **13,** 46–53 (2007).
64. Sacha, J. B. *et al.* Gag-specific CD8+ T lymphocytes recognize infected cells before AIDS-virus integration and viral protein expression. *J. Immunol. Baltim. Md 1950* **178,** 2746–2754 (2007).
65. Dengjel, J. *et al.* Autophagy promotes MHC class II presentation of peptides from intracellular source proteins. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 7922–7927 (2005).
66. Yewdell, J. W., Antón, L. C. & Bennink, J. R. Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? *J. Immunol. Baltim. Md 1950* **157,** 1823–1826 (1996).
67. Rock, K. L. *et al.* Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. *Cell* **78,** 761–771 (1994).
68. Khan, S. *et al.* Cutting edge: neosynthesis is required for the presentation of a T cell epitope from a long-lived viral protein. *J. Immunol. Baltim. Md 1950* **167,** 4801–4804 (2001).
69. Toes, R. E. *et al.* Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. *J. Exp. Med.* **194,** 1–12 (2001).
70. Neefjes, J., Jongsma, M. L. M., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11,** 823–836 (2011).
71. Kincaid, E. Z. *et al.* Mice completely lacking immunoproteasomes show major changes in antigen presentation. *Nat. Immunol.* **13,** 129–135 (2012).
72. van Endert, P. M. *et al.* A sequential model for peptide binding and transport by the transporters associated with antigen processing. *Immunity* **1,** 491–500 (1994).
73. Kelly, A. *et al.* Assembly and function of the two ABC transporter proteins encoded in the human major histocompatibility complex. *Nature* **355,** 641–644 (1992).
74. Schmitt, L. & Tampé, R. Structure and mechanism of ABC transporters. *Curr. Opin. Struct. Biol.* **12,** 754–760 (2002).
75. Serwold, T., Gaw, S. & Shastri, N. ER aminopeptidases generate a unique pool of peptides for MHC class I molecules. *Nat. Immunol.* **2,** 644–651 (2001).
76. York, I. A. *et al.* The ER aminopeptidase ERAP1 enhances or limits antigen presentation by trimming epitopes to 8-9 residues. *Nat. Immunol.* **3,** 1177–1184 (2002).
77. Saveanu, L. *et al.* Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat. Immunol.* **6,** 689–697 (2005).
78. Hughes, E. A., Hammond, C. & Cresswell, P. Misfolded major histocompatibility complex class I heavy chains are translocated into the cytoplasm and degraded by the proteasome. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 1896–1901 (1997).
79. Jung, S. *et al.* In vivo depletion of CD11c+ dendritic cells abrogates priming of CD8+ T cells by exogenous cell-associated antigens. *Immunity* **17,** 211–220 (2002).
80. Joffre, O. P., Segura, E., Savina, A. & Amigorena, S. Cross-presentation by dendritic cells. *Nat. Rev. Immunol.* **12,** 557–569 (2012).
81. Ackerman, A. L., Giodini, A. & Cresswell, P. A role for the endoplasmic reticulum protein retrotranslocation machinery during crosspresentation by dendritic cells. *Immunity* **25,** 607–617 (2006).
82. Burgdorf, S., Schölz, C., Kautz, A., Tampé, R. & Kurts, C. Spatial and mechanistic separation of cross-presentation and endogenous antigen presentation. *Nat. Immunol.* **9,** 558–566 (2008).
83. Lizée, G. *et al.* Control of dendritic cell cross-presentation by the major histocompatibility complex class I cytoplasmic domain. *Nat. Immunol.* **4,** 1065–1073 (2003).
84. Basha, G. *et al.* A CD74-dependent MHC class I endolysosomal cross-presentation pathway. *Nat. Immunol.* **13,** 237–245 (2012).
85. Delamarre, L., Pack, M., Chang, H., Mellman, I. & Trombetta, E. S. Differential lysosomal proteolysis in antigen-presenting cells determines antigen fate. *Science* **307,** 1630–1634 (2005).
86. Robinson, J. *et al.* The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* **43,** D423-431 (2015).

87. Falk, K., Rötzschke, O., Stevanović, S., Jung, G. & Rammensee, H. G. Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* **351,** 290–296 (1991).
88. Matsumura, M., Fremont, D. H., Peterson, P. A. & Wilson, I. A. Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science* **257,** 927–934 (1992).
89. Kondo, A. *et al.* Prominent roles of secondary anchor residues in peptide binding to HLA-A24 human class I molecules. *J. Immunol. Baltim. Md 1950* **155,** 4307–4312 (1995).
90. Ruppert, J. *et al.* Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules. *Cell* **74,** 929–937 (1993).
91. Granados, D. P., Laumont, C. M., Thibault, P. & Perreault, C. The nature of self for T cells-a systems-level perspective. *Curr. Opin. Immunol.* **34,** 1–8 (2015).
92. Hassan, C. *et al.* The human leukocyte antigen-presented ligandome of B lymphocytes. *Mol. Cell. Proteomics MCP* **12,** 1829–1843 (2013).
93. Kiepiela, P. *et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432,** 769–775 (2004).
94. Goulder, P. J. *et al.* Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. *Nat. Med.* **3,** 212–217 (1997).
95. Schneidewind, A. *et al.* Escape from the dominant HLA-B27-restricted cytotoxic T-lymphocyte response in Gag is associated with a dramatic reduction in human immunodeficiency virus type 1 replication. *J. Virol.* **81,** 12382–12393 (2007).
96. Migueles, S. A. *et al.* HLA B*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 2709–2714 (2000).
97. Carrington, M. *et al.* HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science* **283,** 1748–1752 (1999).
98. Roeth, J. F., Williams, M., Kasper, M. R., Filzen, T. M. & Collins, K. L. HIV-1 Nef disrupts MHC-I trafficking by recruiting AP-1 to the MHC-I cytoplasmic tail. *J. Cell Biol.* **167,** 903–913 (2004).
99. Cohen, G. B. *et al.* The selective downregulation of class I major histocompatibility complex proteins by HIV-1 protects HIV-infected cells from NK cells. *Immunity* **10,** 661–671 (1999).
100. Rötzschke, O., Falk, K., Wallny, H. J., Faath, S. & Rammensee, H. G. Characterization of naturally occurring minor histocompatibility peptides including H-4 and H-Y. *Science* **249,** 283–287 (1990).
101. Rötzschke, O. *et al.* Isolation and analysis of naturally processed viral peptides as recognized by cytotoxic T cells. *Nature* **348,** 252–254 (1990).
102. Storkus, W. J., Zeh, H. J., 3rd, Salter, R. D. & Lotze, M. T. Identification of T-cell epitopes: rapid isolation of class I-presented peptides from viable cells by mild acid elution. *J. Immunother. Emphas. Tumor Immunol. Off. J. Soc. Biol. Ther.* **14,** 94–103 (1993).
103. Barnstable, C. J. *et al.* Production of monoclonal antibodies to group A erythrocytes, HLA and other human cell surface antigens-new tools for genetic analysis. *Cell* **14,** 9–20 (1978).
104. Brodsky, F. M. & Parham, P. Monomorphic anti-HLA-A,B,C monoclonal antibodies detecting molecular subunits and combinatorial determinants. *J. Immunol. Baltim. Md 1950* **128,** 129–135 (1982).
105. Prilliman, K. *et al.* Large-scale production of class I bound peptides: assigning a signature to HLA-B*1501. *Immunogenetics* **45,** 379–385 (1997).
106. Paul, P. *et al.* Identification of HLA-G7 as a new splice variant of the HLA-G mRNA and expression of soluble HLA-G5, -G6, and -G7 transcripts in human transfected cells. *Hum. Immunol.* **61,** 1138–1149 (2000).
107. van Rood, J. J., van Leeuwen, A. & van Santen, M. C. Anti HL-A2 inhibitor in normal human serum. *Nature* **226,** 366–367 (1970).
108. Charlton, R. K. & Zmijewski, C. M. Soluble HL-A7 antigen: localization in the beta-lipoprotein fraction of human serum. *Science* **170,** 636–637 (1970).
109. Adamashvili, I., Kelley, R. E., Pressly, T. & McDonald, J. C. Soluble HLA: patterns of expression in normal subjects, autoimmune diseases, and transplant recipients. *Rheumatol. Int.* **25,** 491–500 (2005).
110. Scull, K. E. *et al.* Secreted HLA recapitulates the immunopeptidome and allows in-depth coverage of HLA A*02:01 ligands. *Mol. Immunol.* **51,** 136–142 (2012).
111. Lazarus, D. *et al.* Efficient peptide recovery from secreted recombinant MHC-I molecules expressed via mRNA transfection. *Immunol. Lett.* **165,** 32–38 (2015).
112. Hickman, H. D. *et al.* C-terminal epitope tagging facilitates comparative ligand mapping from MHC class I positive cells. *Hum. Immunol.* **61,** 1339–1346 (2000).
113. Hunt, D. F. *et al.* Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* **255,** 1261–1263 (1992).
114. Granados, D. P. *et al.* Impact of genomic polymorphisms on the repertoire of human MHC class I-associated peptides. *Nat. Commun.* **5,** 3600 (2014).
115. Crotzer, V. L. *et al.* Immunodominance among EBV-derived epitopes restricted by HLA-B27 does not correlate with epitope abundance in EBV-transformed B-lymphoblastoid cell lines. *J. Immunol. Baltim. Md 1950* **164,** 6120–6129 (2000).
116. Croft, N. P. *et al.* Kinetics of antigen expression and epitope presentation during virus infection. *PLoS Pathog.* **9,** e1003129 (2013).
117. Keskin, D. B. *et al.* Physical detection of influenza A epitopes identifies a stealth subset on human lung epithelium evading natural CD8 immunity. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 2151–2156 (2015).
118. Testa, J. S. *et al.* MHC class I-presented T cell epitopes identified by immunoproteomics analysis are targets for a cross reactive influenza-specific T cell response. *PloS One* **7,** e48484 (2012).
119. Yaciuk, J. C. *et al.* Direct interrogation of viral peptides presented by the class I HLA of HIV-infected T cells. *J. Virol.* **88,** 12992–13004 (2014).
120. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics MCP* **14,** 658–673 (2015).
121. Hickman, H. D. *et al.* Cutting edge: class I presentation of host peptides following HIV infection. *J. Immunol. Baltim. Md 1950* **171,** 22–26 (2003).
122. Wahl, A., Schafer, F., Bardet, W. & Hildebrand, W. H. HLA class I molecules reflect an altered host proteome after influenza virus infection. *Hum. Immunol.* **71,** 14–22 (2010).
123. Berlin, C. *et al.* Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy. *Leukemia* **29,** 647–659 (2015).
124. Weidanz, J. A. *et al.* Development and implementation of a direct detection, quantitation and validation system for class I MHC self-peptide epitopes. *J. Immunol. Methods* **318,** 47–58 (2007).
125. Ternette, N. *et al.* Early Kinetics of the HLA Class I-Associated Peptidome of MVA.HIVconsv-Infected Cells. *J. Virol.* **89,** 5760–5771 (2015).

126. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334,** 395–402 (1988).
127. Starr, T. K., Jameson, S. C. & Hogquist, K. A. Positive and Negative Selection of T Cells. *Annu. Rev. Immunol.* **21,** 139–176 (2003).
128. James, J. R. & Vale, R. D. Biophysical mechanism of T-cell receptor triggering in a reconstituted system. *Nature* **487,** 64–69 (2012).
129. Murphy, K. M. *Janeway's Immunobiology*. (Taylor & Francis Group, 2011).
130. Borowski, A. B. *et al.* Memory CD8+ T cells require CD28 costimulation. *J. Immunol. Baltim. Md 1950* **179,** 6494–6503 (2007).
131. Sallusto, F., Geginat, J. & Lanzavecchia, A. Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annu. Rev. Immunol.* **22,** 745–763 (2004).
132. Geginat, J., Lanzavecchia, A. & Sallusto, F. Proliferation and differentiation potential of human CD8+ memory T-cell subsets in response to antigen or homeostatic cytokines. *Blood* **101,** 4260–4266 (2003).
133. Lanzavecchia, A. & Sallusto, F. Dynamics of T lymphocyte responses: intermediates, effectors, and memory cells. *Science* **290,** 92–97 (2000).
134. Gheysen, D. *et al.* Assembly and release of HIV-1 precursor Pr55gag virus-like particles from recombinant baculovirus-infected insect cells. *Cell* **59,** 103–112 (1989).
135. Massiah, M. A. *et al.* Three-dimensional structure of the human immunodeficiency virus type 1 matrix protein. *J. Mol. Biol.* **244,** 198–223 (1994).
136. Gitti, R. K. *et al.* Structure of the amino-terminal core domain of the HIV-1 capsid protein. *Science* **273,** 231–235 (1996).
137. Gamble, T. R. *et al.* Structure of the carboxyl-terminal dimerization domain of the HIV-1 capsid protein. *Science* **278,** 849–853 (1997).
138. Mammano, F., Ohagen, A., Höglund, S. & Göttlinger, H. G. Role of the major homology region of human immunodeficiency virus type 1 in virion morphogenesis. *J. Virol.* **68,** 4927–4936 (1994).
139. Summers, M. F. *et al.* Nucleocapsid zinc fingers detected in retroviruses: EXAFS studies of intact viruses and the solution-state structure of the nucleocapsid protein from HIV-1. *Protein Sci. Publ. Protein Soc.* **1,** 563–574 (1992).
140. Fossen, T. *et al.* Solution structure of the human immunodeficiency virus type 1 p6 protein. *J. Biol. Chem.* **280,** 42515–42527 (2005).
141. Tang, C. *et al.* Entropic switch regulates myristate exposure in the HIV-1 matrix protein. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 517–522 (2004).
142. 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,** 11364–11369 (2006).
143. Shkriabai, N. *et al.* Interactions of HIV-1 Gag with assembly cofactors. *Biochemistry (Mosc.)* **45,** 4077–4083 (2006).
144. Ono, A. & Freed, E. O. Plasma membrane rafts play a critical role in HIV-1 assembly and release. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 13925–13930 (2001).
145. Briggs, J. a. G. *et al.* Structure and assembly of immature HIV. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 11090–11095 (2009).
146. Freed, E. O. HIV-1 assembly, release and maturation. *Nat. Rev. Microbiol.* **13,** 484–496 (2015).
147. Briggs, J. A. G. *et al.* The stoichiometry of Gag protein in HIV-1. *Nat. Struct. Mol. Biol.* **11,** 672–675 (2004).
148. Deml, L., Speth, C., Dierich, M. P., Wolf, H. & Wagner, R. Recombinant HIV-1 Pr55gag virus-like particles: potent stimulators of innate and acquired immune responses. *Mol. Immunol.* **42,** 259–277 (2005).
149. Votteler, J. & Sundquist, W. I. Virus budding and the ESCRT pathway. *Cell Host Microbe* **14,** 232–241 (2013).
150. Hanson, P. I., Roth, R., Lin, Y. & Heuser, J. E. Plasma membrane deformation by circular arrays of ESCRT-III protein filaments. *J. Cell Biol.* **180,** 389–402 (2008).
151. Poropatich, K. & Sullivan, D. J. Human immunodeficiency virus type 1 long-term non-progressors: the viral, genetic and immunological basis for disease non-progression. *J. Gen. Virol.* **92,** 247–268 (2011).
152. Autran, B., Descours, B., Avettand-Fenoel, V. & Rouzioux, C. Elite controllers as a model of functional cure. *Curr. Opin. HIV AIDS* **6,** 181–187 (2011).
153. Prendergast, A. *et al.* Gag-specific CD4+ T-cell responses are associated with virological control of paediatric HIV-1 infection. *AIDS Lond. Engl.* **25,** 1329–1331 (2011).
154. McMichael, A. J. & Koff, W. C. Vaccines that stimulate T cell immunity to HIV-1: the next step. *Nat. Immunol.* **15,** 319–322 (2014).
155. Schell, J. B. *et al.* Significant protection against high-dose simian immunodeficiency virus challenge conferred by a new prime-boost vaccine regimen. *J. Virol.* **85,** 5764–5772 (2011).
156. Schell, J. B. *et al.* Antigenic requirement for Gag in a vaccine that protects against high-dose mucosal challenge with simian immunodeficiency virus. *Virology* **476,** 405–412 (2015).
157. Stephenson, K. E., Li, H., Walker, B. D., Michael, N. L. & Barouch, D. H. Gag-specific cellular immunity determines in vitro viral inhibition and in vivo virologic control following simian immunodeficiency virus challenges of vaccinated rhesus monkeys. *J. Virol.* **86,** 9583–9589 (2012).
158. Hansen, S. G. *et al.* Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine. *Nature* **473,** 523–527 (2011).
159. Hansen, S. G. *et al.* Effector memory T cell responses are associated with protection of rhesus monkeys from mucosal simian immunodeficiency virus challenge. *Nat. Med.* **15,** 293–299 (2009).
160. Lilja, A. E. & Shenk, T. Efficient replication of rhesus cytomegalovirus variants in multiple rhesus and human cell types. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 19950–19955 (2008).
161. Hansen, S. G. *et al.* Cytomegalovirus vectors violate CD8+ T cell epitope recognition paradigms. *Science* **340,** 1237874 (2013).
162. Hansen, S. G. *et al.* Broadly targeted $CD8^+$ T cell responses restricted by major histocompatibility complex E. *Science* **351,** 714–720 (2016).
163. Ludwig, C. & Wagner, R. Virus-like particles-universal molecular toolboxes. *Curr. Opin. Biotechnol.* **18,** 537–545 (2007).
164. Ruedl, C., Storni, T., Lechner, F., Bächi, T. & Bachmann, M. F. Cross-presentation of virus-like particles by skin-derived CD8(-) dendritic cells: a dispensable role for TAP. *Eur. J. Immunol.* **32,** 818–825 (2002).
165. Grgacic, E. V. L. & Anderson, D. A. Virus-like particles: passport to immune recognition. *Methods San Diego Calif* **40,** 60–65 (2006).
166. Buonaguro, L. *et al.* Baculovirus-derived human immunodeficiency virus type 1 virus-like particles activate dendritic cells and induce ex vivo T-cell responses. *J. Virol.* **80,** 9134–9143 (2006).
167. Tsunetsugu-Yokota, Y. *et al.* Yeast-derived human immunodeficiency virus type 1 p55(gag) virus-like particles activate dendritic cells (DCs) and induce perforin expression

in Gag-specific CD8(+) T cells by cross-presentation of DCs. *J. Virol.* **77,** 10250–10259 (2003).

168. Buonaguro, L. *et al.* Induction of neutralizing antibodies and cytotoxic T lymphocytes in Balb/c mice immunized with virus-like particles presenting a gp120 molecule from a HIV-1 isolate of clade A. *Antiviral Res.* **54,** 189–201 (2002).

169. Buonaguro, L. *et al.* Induction of systemic and mucosal cross-clade neutralizing antibodies in BALB/c mice immunized with human immunodeficiency virus type 1 clade A virus-like particles administered by different routes of inoculation. *J. Virol.* **79,** 7059–7067 (2005).

170. Paliard, X. *et al.* Priming of strong, broad, and long-lived HIV type 1 p55gag-specific CD8+ cytotoxic T cells after administration of a virus-like particle vaccine in rhesus macaques. *AIDS Res. Hum. Retroviruses* **16,** 273–282 (2000).

171. Sailaja, G., Skountzou, I., Quan, F.-S., Compans, R. W. & Kang, S.-M. Human immunodeficiency virus-like particles activate multiple types of immune cells. *Virology* **362,** 331–341 (2007).

172. Su, L. *et al.* Characterization of a virtually full-length human immunodeficiency virus type 1 genome of a prevalent intersubtype (C/B') recombinant strain in China. *J. Virol.* **74,** 11367–11376 (2000).

173. Wild, J. *et al.* Preclinical evaluation of the immunogenicity of C-type HIV-1-based DNA and NYVAC vaccines in the Balb/C mouse model. *Viral Immunol.* **22,** 309–319 (2009).

174. Mooij, P. *et al.* Comparison of human and rhesus macaque T-cell responses elicited by boosting with NYVAC encoding human immunodeficiency virus type 1 clade C immunogens. *J. Virol.* **83,** 5881–5889 (2009).

175. Harari, A. *et al.* An HIV-1 clade C DNA prime, NYVAC boost vaccine regimen induces reliable, polyfunctional, and long-lasting T cell responses. *J. Exp. Med.* **205,** 63–77 (2008).

176. Böckl, K. *et al.* Altering an artificial Gagpolnef polyprotein and mode of ENV co-administration affects the immunogenicity of a clade C HIV DNA vaccine. *PloS One* **7,** e34723 (2012).

177. 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,** 970–988 (2015).

178. Asbach, B. *et al.* Potential To Streamline Heterologous DNA Prime and NYVAC/Protein Boost HIV Vaccine Regimens in Rhesus Macaques by Employing Improved Antigens. *J. Virol.* **90,** 4133–4149 (2016).

179. Nabi, G. *et al.* GagPol-specific $CD4^+$ T-cells increase the antibody response to Env by intrastructural help. *Retrovirology* **10,** 117 (2013).

180. Crooks, E. T. *et al.* Vaccine-Elicited Tier 2 HIV-1 Neutralizing Antibodies Bind to Quaternary Epitopes Involving Glycan-Deficient Patches Proximal to the CD4 Binding Site. *PLoS Pathog.* **11,** e1004932 (2015).

181. Kuate, S. *et al.* Immunogenicity and efficacy of immunodeficiency virus-like particles pseudotyped with the G protein of vesicular stomatitis virus. *Virology* **351,** 133–144 (2006).

182. Skountzou, I. *et al.* Incorporation of glycosylphosphatidylinositol-anchored granulocyte-macrophage colony-stimulating factor or CD40 ligand enhances immunogenicity of chimeric simian immunodeficiency virus-like particles. *J. Virol.* **81,** 1083–1094 (2007).

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

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

185. Phillips, R. E. *et al.* Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. *Nature* **354,** 453–459 (1991).

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

187. Li, G. *et al.* An integrated map of HIV genome-wide variation from a population perspective. *Retrovirology* **12,** 18 (2015).

188. Li, G. *et al.* Functional conservation of HIV-1 Gag: implications for rational drug design. *Retrovirology* **10,** 126 (2013).

189. McElrath, M. J. *et al.* HIV-1 vaccine-induced immunity in the test-of-concept Step Study: a case-cohort analysis. *Lancet Lond. Engl.* **372,** 1894–1905 (2008).

190. Dommaraju, K. *et al.* CD8 and CD4 epitope predictions in RV144: no strong evidence of a T-cell driven sieve effect in HIV-1 breakthrough sequences from trial participants. *PloS One* **9,** e111334 (2014).

191. Williamson, A.-L. & Rybicki, E. P. Justification for the inclusion of Gag in HIV vaccine candidates. *Expert Rev. Vaccines* **15,** 585–598 (2016).

192. Kesturu, G. S. *et al.* Minimization of genetic distances by the consensus, ancestral, and center-of-tree (COT) sequences for HIV-1 variants within an infected individual and the design of reagents to test immune reactivity. *Virology* **348,** 437–448 (2006).

193. Létourneau, S. *et al.* Design and pre-clinical evaluation of a universal HIV-1 vaccine. *PloS One* **2,** e984 (2007).

194. Rolland, M., Nickle, D. C. & Mullins, J. I. HIV-1 group M conserved elements vaccine. *PLoS Pathog.* **3,** e157 (2007).

195. Liu, M. K. P. *et al.* Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *J. Clin. Invest.* **123,** 380–393 (2013).

196. Kulkarni, V. *et al.* HIV-1 p24(gag) derived conserved element DNA vaccine increases the breadth of immune response in mice. *PloS One* **8,** e60245 (2013).

197. Kulkarni, V. *et al.* Altered response hierarchy and increased T-cell breadth upon HIV-1 conserved element DNA vaccination in macaques. *PloS One* **9,** e86254 (2014).

198. Koopman, G. *et al.* DNA/long peptide vaccination against conserved regions of SIV induces partial protection against SIVmac251 challenge. *AIDS Lond. Engl.* **27,** 2841–2851 (2013).

199. Fischer, W. *et al.* Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nat. Med.* **13,** 100–106 (2007).

200. Barouch, D. H. *et al.* Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. *Nat. Med.* **16,** 319–323 (2010).

201. Barouch, D. H. *et al.* Vaccine protection against acquisition of neutralization-resistant SIV challenges in rhesus monkeys. *Nature* **482,** 89–93 (2012).

202. Foley, B. *et al. HIV Sequence Compendium 2013*. (Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2013).

203. Ratner, L. *et al.* Complete nucleotide sequence of the AIDS virus, HTLV-III. *Nature* **313,** 277–284 (1985).

204. Korber, B., Foley, B., Kuiken, C., Pillai, S. & Sodroski, J. Numbering Positions in HIV Relative to HXB2CG. (2014). Available at: http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html. (Accessed: 28th April 2016)

205. Nickle, D. C. *et al.* Consensus and ancestral state HIV vaccines. *Science* **299,** 1515-1518-1518 (2003).

206. Gaschen, B. *et al.* Diversity considerations in HIV-1 vaccine selection. *Science* **296,** 2354–2360 (2002).
207. Nickle, D. C. *et al.* Coping with viral diversity in HIV vaccine design. *PLoS Comput. Biol.* **3,** e75 (2007).
208. Rolland, M. *et al.* Reconstruction and function of ancestral center-of-tree human immunodeficiency virus type 1 proteins. *J. Virol.* **81,** 8507–8514 (2007).
209. Gray, G. E. *et al.* Recombinant adenovirus type 5 HIV gag/pol/nef vaccine in South Africa: unblinded, long-term follow-up of the phase 2b HVTN 503/Phambili study. *Lancet Infect. Dis.* **14,** 388–396 (2014).
210. Yusim, K. *et al. HIV Molecular Immunology 2014*. (Los Alamos National Laboratory, Theoretical Biology and Biophysics, 2015).
211. Pfeifer, M. Development of an Algorithm for the in silico Design of T Cell Immunogens as exemplified by the HIV-1 Group specific antigen. (University of Regensburg, 2010).
212. Freed, E. O., Orenstein, J. M., Buckler-White, A. J. & Martin, M. A. Single amino acid changes in the human immunodeficiency virus type 1 matrix protein block virus particle production. *J. Virol.* **68,** 5311–5320 (1994).
213. Lesk, A. M. & Chothia, C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136,** 225–270 (1980).
214. Illergård, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* **77,** 499–508 (2009).
215. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **47,** 5.6.1-32 (2014).
216. Martí-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).
217. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).
218. Fiser, A., Do, R. K. & Sali, A. Modeling of loops in protein structures. *Protein Sci. Publ. Protein Soc.* **9,** 1753–1773 (2000).
219. Shen, M.-Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci. Publ. Protein Soc.* **15,** 2507–2524 (2006).
220. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
221. Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36,** W5-9 (2008).
222. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).
223. Sunyaev, S., Ramensky, V. & Bork, P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet. TIG* **16,** 198–200 (2000).
224. Miller, M. P. & Kumar, S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* **10,** 2319–2328 (2001).
225. Gribskov, M., McLachlan, A. D. & Eisenberg, D. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 4355–4358 (1987).
226. Claverie, J. M. Some useful statistical properties of position-weight matrices. *Comput. Chem.* **18,** 287–294 (1994).
227. Tatusov, R. L., Altschul, S. F. & Koonin, E. V. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 12091–12095 (1994).
228. Dayhoff, M. O. & Schwartz, R. M. A model of evolutionary change in proteins. *Atlas Protein Seq. Struct.* (1978).
229. Henikoff, J. G. & Henikoff, S. Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci. CABIOS* **12,** 135–143 (1996).
230. Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **7,** 179–188 (1936).
231. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. in *International Joint Conference on Artificial Intelligence* 1137–1143 (1995).
232. Hopcroft, J. & Tarjan, R. Algorithm 447: Efficient Algorithms for Graph Manipulation. *Commun ACM* **16,** 372–378 (1973).
233. Garey, M. R. & Johnson, D. S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. (W H Freeman & Co, 1979).
234. Du, D.-Z. & Pardalos, P. M. *Handbook of Combinatorial Optimization*. (Springer, 1999).
235. Holland, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. (Mit University Press Group Ltd, 1992).
236. Leitner, T., Korber, B., Daniels, M., Calef, C. & Foley, B. HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences, 2005. *HIV Seq. Compend. 2005* 41–48 (2005).
237. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinforma. Oxf. Engl.* **23,** 2947–2948 (2007).
238. Paradis, E. *Analysis of Phylogenetics and Evolution with R*. (Springer New York, 2006).
239. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinforma. Oxf. Engl.* **20,** 289–290 (2004).
240. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2,** 13–34 (1985).
241. Gouy, M. & Gautier, C. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10,** 7055–7074 (1982).
242. Hernan, R. A. *et al.* Human hemoglobin expression in Escherichia coli: importance of optimal codon usage. *Biochemistry (Mosc.)* **31,** 8619–8628 (1992).
243. Graf, M. *et al.* Concerted action of multiple cis-acting sequences is required for Rev dependence of late human immunodeficiency virus type 1 gene expression. *J. Virol.* **74,** 10822–10826 (2000).
244. Fath, S. *et al.* Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression. *PloS One* **6,** e17596 (2011).
245. Kypr, J. & Mrázek, J. Unusual codon usage of HIV. *Nature* **327,** 20 (1987).
246. van der Kuyl, A. C. & Berkhout, B. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. *Retrovirology* **9,** 92 (2012).
247. Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460,** 711–716 (2009).
248. Page, K. A., Landau, N. R. & Littman, D. R. Construction and use of a human immunodeficiency virus vector for analysis of virus infectivity. *J. Virol.* **64,** 5270–5276 (1990).
249. Winslow, D. L. *et al.* Construction of infectious molecular clones of HIV-1 containing defined mutations in the protease gene. *Biochem. Biophys. Res. Commun.* **205,** 1651–1657 (1994).

250. Saksela, K., Cheng, G. & Baltimore, D. Proline-rich (PxxP) motifs in HIV-1 Nef bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of Nef+ viruses but not for down-regulation of CD4. *EMBO J.* **14,** 484–491 (1995).

251. Leiherer, A., Ludwig, C. & Wagner, R. Uncoupling Human Immunodeficiency Virus Type 1 gag and pol Reading Frames: Role of the Transframe Protein p6* in Viral Replication. *J. Virol.* **83,** 7210–7220 (2009).

252. O'Gorman, S., Fox, D. T. & Wahl, G. M. Recombinase-mediated gene activation and site-specific integration in mammalian cells. *Science* **251,** 1351–1355 (1991).

253. Glaser, R. W. & Hausdorf, G. Binding kinetics of an antibody against HIV p24 core protein measured with real-time biomolecular interaction analysis suggest a slow conformational change in antigen p24. *J. Immunol. Methods* **189,** 1–14 (1996).

254. Küttner, G. *et al.* Immunoglobulin V regions and epitope mapping of a murine monoclonal antibody against p24 core protein of HIV-1. *Mol. Immunol.* **29,** 561–564 (1992).

255. *HIV Immunology and HIV SIV Vaccine Databases 2003*. (DIANE Publishing).

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

257. Grant, S. G., Jessee, J., Bloom, F. R. & Hanahan, D. Differential plasmid rescue from transgenic mouse DNAs into Escherichia coli methylation-restriction mutants. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 4645–4649 (1990).

258. Green, R. & Rogers, E. J. Chemical Transformation of E. coli. *Methods Enzymol.* **529,** 329–336 (2013).

259. Green, M. R. & Sambrook, J. *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor Laboratory Press,U.S., 2012).

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

261. Wolf, H. *et al.* Production, mapping and biological characterisation of monoclonal antibodies to the core protein (p24) of human immunodeficiency virus type 1. *AIDS Forsch.* **1,** (1990).

262. Brunner, T. Charakterisierung eines induzierbaren eukaryotischen Expressionssystems zur simultanen Produktion HI-viraler Envelope-Proteine. (University of Regensburg, 2013).

263. Graham, F. L., Smiley, J., Russell, W. C. & Nairn, R. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J. Gen. Virol.* **36,** 59–74 (1977).

264. Rio, D. C., Clark, S. G. & Tjian, R. A mammalian host-vector system that regulates expression and amplification of transfected genes by temperature induction. *Science* **227,** 23–28 (1985).

265. Sauer, B. Site-specific recombination: developments and applications. *Curr. Opin. Biotechnol.* **5,** 521–527 (1994).

266. Thurner, B. *et al.* Generation of large numbers of fully mature and stable dendritic cells from leukapheresis products for clinical application. *J. Immunol. Methods* **223,** 1–15 (1999).

267. Stief, T. Capacity of recombinant HI-VLPs for T-cell stimulation. (University of Regensburg, 2015).

268. Ukkonen, P., Korpela, J., Suni, J. & Hedman, K. Inactivation of human immunodeficiency virus in serum specimens as a safety measure for diagnostic immunoassays. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* **7,** 518–523 (1988).

269. Horowitz, B. *et al.* WHO Expert Committee on Biological Standardization. *World Health Organ. Tech. Rep. Ser.* **924,** 1–232, backcover (2004).

270. Ekstrand, D. H., Awad, R. J., Källander, C. F. & Gronowitz, J. S. A sensitive assay for the quantification of reverse transcriptase activity based on the use of carrier-bound template and non-radioactive-product detection, with special reference to human-immunodeficiency-virus isolation. *Biotechnol. Appl. Biochem.* **23 (Pt 2),** 95–105 (1996).

271. Steindl, F., Armbruster, C., Pierer, K., Purtscher, M. & Katinger, H. W. A simple and robust method for the complete dissociation of HIV-1 p24 and other antigens from immune complexes in serum and plasma samples. *J. Immunol. Methods* **217,** 143–151 (1998).

272. Ameismeier, M. Quantitative Analyse der Partikelfreisetzung von Varianten des HIV-1 Gag Proteins. (University of Regensburg, 2013).

273. Watzlowik, M. Analyse der Freisetzung Virus-ähnlicher Partikel von Varianten des HIV-1 Gag Proteins. (University of Regensburg, 2015).

274. Berger, J., Hauber, J., Hauber, R., Geiger, R. & Cullen, B. R. Secreted placental alkaline phosphatase: a powerful new quantitative indicator of gene expression in eukaryotic cells. *Gene* **66,** 1–10 (1988).

275. Berger, J., Howard, A. D., Gerber, L., Cullen, B. R. & Udenfriend, S. Expression of active, membrane-bound human placental alkaline phosphatase by transfected simian cells. *Proc. Natl. Acad. Sci. U. S. A.* **84,** 4885–4889 (1987).

276. McComb, R. B. & Bowers, G. N. Study of optimum buffer conditions for measuring alkaline phosphatase activity in human serum. *Clin. Chem.* **18,** 97–104 (1972).

277. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72,** 248–254 (1976).

278. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227,** 680–685 (1970).

279. De Ruiter, G. A., Smid, P., Schols, H. A., Van Boom, J. H. & Rombouts, F. M. Detection of fungal carbohydrate antigens by high-performance immunoaffinity chromatography using a protein A column with covalently linked immunoglobulin G. *J. Chromatogr.* **584,** 69–75 (1992).

280. Purcell, A. W. *et al.* Quantitative and qualitative influences of tapasin on the class I peptide repertoire. *J. Immunol. Baltim. Md 1950* **166,** 1016–1027 (2001).

281. Hauck, S. M. *et al.* Deciphering membrane-associated molecular processes in target tissue of autoimmune uveitis by label-free quantitative mass spectrometry. *Mol. Cell. Proteomics MCP* **9,** 2292–2305 (2010).

282. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).

283. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10,** 1794–1805 (2011).

284. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749-755 (2014).

285. Miller, L. Quantifying western blots without expensive commercial quantification software. *Stuff* (2007).

286. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinforma. Oxf. Engl.* **32,** 511–517 (2016).

287. Nielsen, M. *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci. Publ. Protein Soc.* **12,** 1007–1017 (2003).

288. Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18,** 6097–6100 (1990).

289. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).

290. Dunn, O. J. Multiple Comparisons among Means. *J. Am. Stat. Assoc.* **56,** 52–64 (1961).

291. Solberg, O. D. *et al.* Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum. Immunol.* **69,** 443–464 (2008).

292. Wang, Z. *et al.* Complete human immunodeficiency virus-1 specific T lymphocyte response to Chinese human immunodeficiency virus-1 B/C chronic infectors. *Biomed. Environ. Sci. BES* **22,** 522–528 (2009).

293. Thurmond, J. *et al.* Web-based design and evaluation of T-cell vaccine candidates. *Bioinforma. Oxf. Engl.* **24,** 1639–1640 (2008).

294. Trolle, T. *et al.* The Length Distribution of Class I-Restricted T Cell Epitopes Is Determined by Both Peptide Supply and MHC Allele-Specific Binding Preference. *J. Immunol. Baltim. Md 1950* **196,** 1480–1487 (2016).

295. Bai, Y., Ni, M., Cooper, B., Wei, Y. & Fury, W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. *BMC Genomics* **15,** 325 (2014).

296. Vyas, J. M., Van der Veen, A. G. & Ploegh, H. L. The known unknowns of antigen processing and presentation. *Nat. Rev. Immunol.* **8,** 607–618 (2008).

297. de Jong, A. Contribution of mass spectrometry to contemporary immunology. *Mass Spectrom. Rev.* **17,** 311–335 (1998).

298. Rammensee, H. G., Friede, T. & Stevanoviíc, S. MHC ligands and peptide motifs: first listing. *Immunogenetics* **41,** 178–228 (1995).

299. Lund, O. *et al.* Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* **55,** 797–810 (2004).

300. Vita, R. *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* **43,** D405-412 (2015).

301. Tekaia, F., Yeramian, E. & Dujon, B. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* **297,** 51–60 (2002).

302. Tekaia, F. & Yeramian, E. Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. *BMC Genomics* **7,** 307 (2006).

303. Li, F. *et al.* Mapping HIV-1 vaccine induced T-cell responses: bias towards less-conserved regions and potential impact on vaccine efficacy in the Step study. *PloS One* **6,** e20479 (2011).

304. Stephenson, K. E. & Barouch, D. H. A global approach to HIV-1 vaccine development. *Immunol. Rev.* **254,** 295–304 (2013).

305. Yusim, K. *et al. HIV Molecular Immunology 2015*. (Los Alamos National Laboratory, Theoretical Biology and Biophysics, 2016).

306. Akram, A. & Inman, R. D. Immunodominance: a pivotal principle in host response to viral infections. *Clin. Immunol. Orlando Fla* **143,** 99–115 (2012).

307. Yewdell, J. W., Reits, E. & Neefjes, J. Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nat. Rev. Immunol.* **3,** 952–961 (2003).

308. Lingappa, J. R., Reed, J. C., Tanaka, M., Chutiraka, K. & Robinson, B. A. How HIV-1 Gag assembles in cells: Putting together pieces of the puzzle. *Virus Res.* **193,** 89–107 (2014).

309. Kantz, H. & Schreiber, T. *Nonlinear Time Series Analysis*. (Cambridge University Press, 2003).

310. Luo, D., Ding, C. & Huang, H. Linear Discriminant Analysis: New Formulations and Overfit Analysis. in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* 417–422 (AAAI Press, 2011).

311. Rihn, S. J. *et al.* Extreme genetic fragility of the HIV-1 capsid. *PLoS Pathog.* **9,** e1003461 (2013).

312. Freed, E. O., Englund, G. & Martin, M. A. Role of the basic domain of human immunodeficiency virus type 1 matrix in macrophage infection. *J. Virol.* **69,** 3949–3954 (1995).

313. Al-Mawsawi, L. Q. *et al.* High-throughput profiling of point mutations across the HIV-1 genome. *Retrovirology* **11,** 124 (2014).

314. Manocheewa, S., Swain, J. V., Lanxon-Cookson, E., Rolland, M. & Mullins, J. I. Fitness costs of mutations at the HIV-1 capsid hexamerization interface. *PloS One* **8,** e66065 (2013).

315. von Schwedler, U. K., Stray, K. M., Garrus, J. E. & Sundquist, W. I. Functional surfaces of the human immunodeficiency virus type 1 capsid protein. *J. Virol.* **77,** 5439–5450 (2003).

316. Wright, J. K. *et al.* Impact of HLA-B*81-associated mutations in HIV-1 Gag on viral replication capacity. *J. Virol.* **86,** 3193–3199 (2012).

317. Yufenyuy, E. L. & Aiken, C. The NTD-CTD intersubunit interface plays a critical role in assembly and stabilization of the HIV-1 capsid. *Retrovirology* **10,** 29 (2013).

318. Wagner, R., Deml, L. & Wolf, H. Polyvalent, recombinant HIV-1 virus-like particles: novel HIV-1 vaccine strategies. *Antibiot. Chemother.* **46,** 48–61 (1994).

319. Dorfman, T., Bukovsky, A., Ohagen, A., Höglund, S. & Göttlinger, H. G. Functional domains of the capsid protein of human immunodeficiency virus type 1. *J. Virol.* **68,** 8180–8187 (1994).

320. Abdurahman, S., Höglund, S., Höglund, A. & Vahlne, A. Mutation in the loop C-terminal to the cyclophilin A binding site of HIV-1 capsid protein disrupts proper virus assembly and infectivity. *Retrovirology* **4,** 19 (2007).

321. Melamed, D. *et al.* The conserved carboxy terminus of the capsid domain of human immunodeficiency virus type 1 gag protein is important for virion assembly and release. *J. Virol.* **78,** 9675–9688 (2004).

322. Liang, C. *et al.* Characterization of a putative alpha-helix across the capsid-SP1 boundary that is critical for the multimerization of human immunodeficiency virus type 1 gag. *J. Virol.* **76,** 11729–11737 (2002).

323. Liang, C., Hu, J., Whitney, J. B., Kleiman, L. & Wainberg, M. A. A structurally disordered region at the C terminus of capsid plays essential roles in multimerization and membrane binding of the gag protein of human immunodeficiency virus type 1. *J. Virol.* **77,** 1772–1783 (2003).

324. Reicin, A. S. *et al.* Linker insertion mutations in the human immunodeficiency virus type 1 gag gene: effects on virion particle assembly, release, and infectivity. *J. Virol.* **69,** 642–650 (1995).

325. Hong, S. S. & Boulanger, P. Assembly-defective point mutants of the human immunodeficiency virus type 1 Gag precursor phenotypically expressed in recombinant baculovirus-infected cells. *J. Virol.* **67,** 2787–2798 (1993).

326. Datta, S. A. K. *et al.* On the role of the SP1 domain in HIV-1 particle assembly: a molecular switch? *J. Virol.* **85,** 4111–4121 (2011).

327. Robinson, B. A., Reed, J. C., Geary, C. D., Swain, J. V. & Lingappa, J. R. A temporospatial map that defines specific steps at which critical surfaces in the Gag MA and CA domains act during immature HIV-1 capsid assembly in cells. *J. Virol.* **88,** 5718–5741 (2014).

328. Tanaka, M. *et al.* Mutations of Conserved Residues in the Major Homology Region Arrest Assembling HIV-1 Gag as a Membrane-Targeted Intermediate Containing Genomic RNA and Cellular Proteins. *J. Virol.* **90,** 1944–1963 (2016).

329. Eckwahl, M. J., Telesnitsky, A. & Wolin, S. L. Host RNA Packaging by Retroviruses: A Newly Synthesized Story. *mBio* **7,** e02025-2015 (2016).

330. Dorfman, T., Luban, J., Goff, S. P., Haseltine, W. A. & Göttlinger, H. G. Mapping of functionally important residues of a cysteine-histidine box in the human immunodeficiency virus type 1 nucleocapsid protein. *J. Virol.* **67,** 6159–6169 (1993).

331. Chatel-Chaix, L., Boulay, K., Mouland, A. J. & Desgroseillers, L. The host protein Staufen1 interacts with the Pr55Gag zinc fingers and regulates HIV-1 assembly via its N-terminus. *Retrovirology* **5,** 41 (2008).

332. Cruz, P. E. *et al.* Characterization and downstream processing of HIV-1 core and virus-like-particles produced in serum free medium. *Enzyme Microb. Technol.* **26,** 61–70 (2000).

333. Steppert, P. *et al.* Purification of HIV-1 gag virus-like particles and separation of other extracellular particles. *J. Chromatogr. A* **1455,** 93–101 (2016).

334. Accola, M. A., Höglund, S. & Göttlinger, H. G. A Putative α-Helical Structure Which Overlaps the Capsid-p2 Boundary in the Human Immunodeficiency Virus Type 1 Gag Precursor Is Crucial for Viral Particle Assembly. *J. Virol.* **72,** 2072–2078 (1998).

335. Gross, I. *et al.* A conformational switch controlling HIV-1 morphogenesis. *EMBO J.* **19,** 103–113 (2000).

336. Buseyne, F. *et al.* MHC-I-restricted presentation of HIV-1 virion antigens without viral replication. *Nat. Med.* **7,** 344–349 (2001).

337. Finkelshtein, D., Werman, A., Novick, D., Barak, S. & Rubinstein, M. LDL receptor and its family members serve as the cellular receptors for vesicular stomatitis virus. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 7306–7311 (2013).

338. Shortman, K., Lahoud, M. H. & Caminschi, I. Improving vaccines by targeting antigens to dendritic cells. *Exp. Mol. Med.* **41,** 61–66 (2009).

339. Franco, D., Liu, W., Gardiner, D. F., Hahn, B. H. & Ho, D. D. CD40L-containing virus-like particle as a candidate HIV-1 vaccine targeting dendritic cells. *J. Acquir. Immune Defic. Syndr. 1999* **56,** 393–400 (2011).

340. Vassilieva, E. V. *et al.* Enhanced mucosal immune responses to HIV virus-like particles containing a membrane-anchored adjuvant. *mBio* **2,** e00328-310 (2011).

341. Feng, H. *et al.* Incorporation of a GPI-anchored engineered cytokine as a molecular adjuvant enhances the immunogenicity of HIV VLPs. *Sci. Rep.* **5,** 11856 (2015).

342. Rogers, S., Wells, R. & Rechsteiner, M. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science* **234,** 364–368 (1986).

343. Mothe, B. *et al.* Definition of the viral targets of protective HIV-1-specific T cell responses. *J. Transl. Med.* **9,** 208 (2011).

344. Bassani-Sternberg, M. *et al.* Soluble plasma HLA peptidome as a potential source for cancer biomarkers. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 18769–18776 (2010).

345. Dudek, N. L., Croft, N. P., Schittenhelm, R. B., Ramarathinam, S. H. & Purcell, A. W. A Systems Approach to Understand Antigen Presentation and the Immune Response. *Methods Mol. Biol. Clifton NJ* **1394,** 189–209 (2016).

346. Myers, N. B., Wormstall, E. & Hansen, T. H. Differences among various class I molecules in competition for beta2m in vivo. *Immunogenetics* **43,** 384–387 (1996).

347. Neefjes, J. J. & Ploegh, H. L. Allele and locus-specific differences in cell surface expression and the association of HLA class I heavy chain with beta 2-microglobulin: differential effects of inhibition of glycosylation on class I subunit association. *Eur. J. Immunol.* **18,** 801–810 (1988).

348. Aki, M. *et al.* Interferon-gamma induces different subunit organizations and functional diversity of proteasomes. *J. Biochem. (Tokyo)* **115,** 257–269 (1994).

349. Loffredo, J. T., Valentine, L. E. & Watkins, D. I. *Beyond Mamu-A*01+ Indian Rhesus Macaques: Continued Discovery of New MHC Class I Molecules that Bind Epitopes from the Simian AIDS Viruses*. (Los Alamos National Laboratory, Theoretical Biology and Biophysics, 2007).

350. Niu, L. *et al.* Preclinical evaluation of HIV-1 therapeutic ex vivo dendritic cell vaccines expressing consensus Gag antigens and conserved Gag epitopes. *Vaccine* **29,** 2110–2119 (2011).

351. Hoppes, R., Ekkebus, R., Schumacher, T. N. M. & Ovaa, H. Technologies for MHC class I immunoproteomics. *J. Proteomics* **73,** 1945–1953 (2010).

F.4 Danksagung

Ohne die Unterstützung zahlreicher Personen wäre es mir unmöglich gewesen diese Arbeit anzufertigen. Einige dieser Personen möchte ich hier extra hervorheben.

An erste Stelle möchte ich mich bei Prof. Dr. Ralf Wagner herzlich bedanken, der es mir ermöglicht hat diese Dissertation in seiner Arbeitsgruppe anzufertigen. Sowohl, die exzellente fachliche Begleitung, als auch die äußerst angenehme Arbeitsatmosphäre und die Möglichkeiten zur persönlichen Weiterentwicklung die mir geboten wurden, weiß ich sehr zu schätzen.

Weiter möchte ich mich bei Dr. Benedikt Asbach dafür bedanken, dass er immer ein offenes Ohr für mich hatte und für die zahlreichen Gespräche, Ideen und Verbesserungsvorschläge, die diese Arbeit so bereichert haben.

Herzlichen Dank auch an Prof. Dr. Rainer Merkl, dass er diese Arbeit als Mentor mitbetreut hat und immer mit Rat und Tat zur Seite stand. Es war immer wieder gewinnbringend die Ergebnisse nochmal von einem anderen Blickwinkel betrachten zu lassen.

Großer Dank auch an allen aktuellen und ehemaligen Kolleginnen und Kollegen der AG Wagner, die mir unvergessliche Jahre bereitet haben und meine Zeit in der AG und Regensburg enorm bereichert haben. Speziell erwähnen will ich hierbei Richard B. Kiener. Danke, für all die fachlichen Gespräche und deine Anmerkungen zu dieser Arbeit, aber auch für die gemeinsame Zeit nach Feierabend. Bedanken möchte ich mich auch bei Thomas Schuster, für die schöne gemeinsame Zeit innerhalb und außerhalb der Arbeit (jetzt hab ich hoffentlich wieder mehr Zeit fürs Kino) und bei Anja W. Schütz, dass sie mir als einzige über die ganze Zeit die Treue im 80er Labor gehalten hat, man hätte sich wohl kaum eine bessere Arbeitskollegin wünschen können. Um diese Arbeit nicht noch weiter zu verlängern seinen hier nur noch Basti, Benny, David, Julia, Kri/ystina, Meli, Miri, Tanja, Tobi und Vroni hervorgehoben. Extra erwähnen will ich nur noch Cino, für seine unermüdliche Unterstützung, auch in den spätesten Abendstunden.

Zuletzt möchte ich mich noch bei meiner Familie, meiner besten Freundin Claudi und besonders bei meinen Eltern, Franz und Marille, bedanken. Danke, dass ihr es mir ermöglicht habt ohne Einschränkungen oder Vorgaben meinen eigenen Weg zu gehen und, dass ihr mich immer auf jede erdenkliche Art und Weise unterstützt habt.