The Identification of Cytotoxic T Lymphocyte Escape in a Large, Longitudinal Subtype C HIV-1 Sequence Dataset

by

Ruth Mphahlele - MPHRUT003

Submitted to the University of Cape Town in fulfilment of the

requirements for the degree:

MSc in Medicine (by dissertation), specialization: **Bioinformatics**

Supervisors:

Professor Carolyn Williamson

Associate Professor Darren Martin

Co-supervisors:

Dr Melissa-Rose Abrahams

Dr Anna Yssel

Medical Virology and Computational Biology Divisions

Integrative Biomedical Sciences and Pathology Departments

Faculty of Health Science

University of Cape Town

Submission Date: 30 June 2023

Declaration

I, **Ruth Mphahlele**, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: …… Signed by candidate

Date: 18 September 2023 ……

Table of Contents

Acknowledgements

I would like to extend my gratitude to my supervisors, without whom this work would not be possible, Professor Carolyn Williamson, Professor Darren Martin, Dr Melissa-Rose Abrahams, Dr Anna Yssel, for their invaluable guidance, insightful feedback, and unwavering support throughout the entire process of this research. Their expertise has been instrumental in shaping this thesis.

All thanks the National Research Foundation (NRF) and CAPRISA for the funding support.

A huge thank you to the participants of the CAPRISA002 cohort, the base of this work is for you!

I would like to express my appreciation to my colleagues and friends who have provided encouragement, moral support, assistance, and valuable discussions during this research. Their insights and constructive feedback have helped shape my ideas and improve the quality of this thesis. All members of the HIV diversity group members. Special thanks to Gabriella Wilensky, Samuel Kariuki, Chivonne Moddley, Jennah Hurree, Nicole Chineka.

I extend my heartfelt thanks to my family and friends for their unwavering support, patience and understanding throughout this journey. Their encouragement and belief in me have been a constant source of motivation.

Lastly, I would like to express my deep gratitude to God for his mercy, which carried me through this research. All those mentioned above and more in the background, whose contribution collectively made this thesis possible.

Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral therapy
ANN	Artificial Neural Networks
BA	Binding Affinity
CAPRISA	Center for the AIDS programme of research in South Africa
CD4	Cluster of division
CRFs	Circulating Recombinant Forms
CTL	Cytotoxic T-cell Lymphocyte
EL	Eluted Ligands
ELF	Epitope Location Finder
FUBAR	Fast, Unconstrained Bayesian AppRoximation
HIV-1	Human Immunodeficiency Virus Type 1
HREC	Human Research Ethics Committee
LANL	Los Alamos National Lab
MHC	Major Histocompatibility Complex
Nef	negative regulating factor
NGS	Next-Generation Sequencing
PreP or PEP	Pre- or Post-exposure prophylaxis
Rev	RNA splicing-regulator
TCR	T Cell Receptor
Tat	transactivator protein
HyPhy	Hypothesis Testing using phylogenies
Vif	viral infectivity factor
Vpr	virus protein r
Vpu	virus protein unique
Wpi	Weeks post-infection
WT	Wildtype

List of Tables

List of Figures

Abstract

Human Immunodeficiency Virus (HIV) rapidly escapes cytotoxic T-cell lymphocyte (CTL) immune responses exerted by the host. Mutation patterns and HLA associated footprints linked to viral escape have been identified, making it possible to use viral sequence data, combined with the host HLA allele information, to predict escape. Next-Generation Sequencing (NGS) approaches enable the generation of large sequence datasets, and the detection of viral populations present at very low frequencies in an infected individual at any given time. These datasets allow for the study of changes in viral populations within a host over time and provide a means to understand the kinetics and pathway(s) of escape. While tools exist that allow the prediction of escape in sequence data with small sequence numbers per sampling timepoint, these tools often have limitations in analysing large NGS data sets.

In this project, we developed a workflow for identifying the kinetics of CTL escape in longitudinal HIV-1 next-generation datasets of *gag* sequences generated using an Illumina Miseq platform over the duration of drug-naïve infection. This acquired data set was generated from 15 women over a period of one to seven years and comprised of 4583 short read gag sequences (544 bp). We identified tools for identifying CTL escape in deep sequencing datasets and used pre-defined criteria to screen these tools. The outputs were validated using a test dataset from a previous study that identified escape. We selected the Epitope Matcher tool as having the most potential to identify CTL epitopes and escape mutations. To further support evidence of escape and identify additional putative escape mutations, we identified sites with high Shannon entropy (>=0.25) and sites evolving under positive selection using Hyphy-FUBAR. The sites were verified using the HLA association and CTL epitope variants and escape mutations lists, or data generated by Epitope Matcher.

Using the Epitope Matcher tool, we identified seven HLA-B restricted gag epitopes in six individuals of which putative escape was identified in seven epitopes, commonly occurring in the late chronic phase of infection. The most common epitope in the population was YL9 (found in 60% of the participants) (Gag HXB2 coordinates 296 to 304) restricted by HLA B*15:03, B*15:10 and B*42:01. Toggling of amino acids within

epitopes as a result of potential fitness cost associated with a specific change, was observed in five of seven epitopes. We further identified 35 high Shannon entropy sites, where nine of these sites were found within epitopes identified by Epitope Matcher. Additionally, nine of the high Shannon entropy sites were evolving under positive selection. With supporting evidence, we can predict that the mutation T310S (found in the AW11 epitope, restricted by allele B*58:01), is likely to be associated with escape. This study is important in that it provides a pipeline that will enable semi-automated analysis of NGS data. Using this approach, we have provided a better understanding of the kinetics and frequency of CTL escape over the course of HIV infection. Additionally, we have identified frequently targeted sites across the Gag p24 region and across individuals. This study is relevant to inform CTL-based vaccine prevention and treatment strategies.

Chapter 1: Literature review

1.1 Introduction

Human Immunodeficiency Virus Type 1 (HIV-1) was declared a global pandemic more than forty years ago. In 2021, nearly 40 million people globally were living with HIV-1 (1) (**Figure 1.1**), with more than half of these infected individuals residing in the southern African region. South Africa is the most affected country in the world with approximately 8.2 million people currently living with HIV-1 (2).

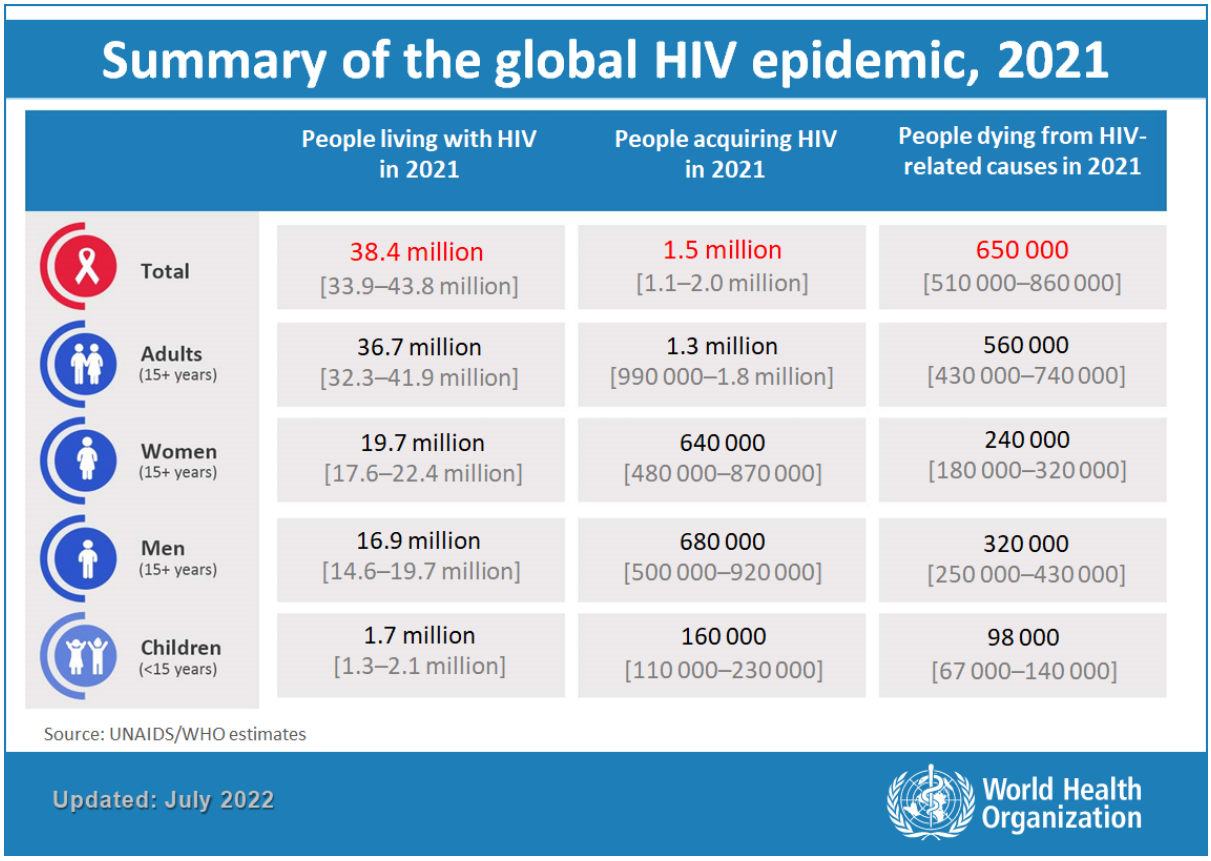


Figure 1.1: Estimated number of people living with HIV-1 in 2021. Image obtained from World Health Organization (3).

Antiretroviral therapy (ART) is an effective treatment for people living with HIV (PLWH) and has also been approved for use in pre- or post-exposure prophylaxis (PreP or PEP) to prevent people from getting infected (4). The development of long-acting ART as PreP, that does not require daily dosing, provides a significant step forward in prevention. The only long-acting PreP approved in South Africa is the dapivirine

vaginal ring which is given monthly (5). However, a demonstration project to evaluate cabotegravir administered intramuscularly to adolescent women every two months was recently approved (6, 7). Although these methods limit the risk of infection, they require adherence (8, 9), and there remains a need for other interventions, such as vaccines, that would provide lasting protection for a broad range of population groups. Despite multiple large vaccine trials, an effective preventative vaccine is yet to be discovered. To date, only one HIV-1 vaccine has provided some evidence of protection. This vaccine, tested in the RV144 clinical trial conducted in Thailand, comprised a combination of a canarypox viral vector carrying HIV-1 genes (*gag, pol*, and *env*) for stimulating cytotoxic T-lymphocyte (CTL) responses, with a viral envelope (Env) protein immunogen for stimulating antibody (Ab) responses (10, 11). Although this trial demonstrated 30% efficacy, the regimen failed to provide any demonstrable efficacy in a follow-up HVTN702 South African trial (12). While the immune responses that vaccines must elicit to protect against infection are not well defined, it is widely accepted that broadly neutralizing antibody responses will be needed (12). The importance of such responses was demonstrated in the recent antibody mediated protection (AMP) trial which evaluated the role of passive infusion of a broadly neutralizing antibody, VRC01, to prevent infection (13). AMP was the first study in humans to show that HIV-1 infection can be prevented by neutralizing antibodies (14).

CTL responses were shown to be important in protection against infection in non-human primate models (15, 16). However, the role of CTL responses in preventing infections in humans is less well defined with the halting of both the Imbokodo and Mosaico efficacy trials, which evaluated a rAd26 and protein boost vaccination regime (17). Other large vaccine trials that will evaluate the role of CTLs include the PrEPVacc trial, which is investigating a combination of rDNA, gp120 boost and MVA (18). CTL-eliciting vaccines may also be useful in cure strategies where CTL responses have been associated with delay to viral load rebound (19).

A major obstacle that an effective vaccine must overcome is the very high diversity of HIV-1. The global diversity of the virus has been largely driven by rapid evolution and host immune selection (20), including evasion from CTL responses resulting in loss of viral control (21). This has resulted in the emergence of different HIV-1 subtypes circulating in different human populations throughout the world. It is likely that vaccine

efficacy will be impacted by this diversity such that a vaccine that works for one subtype may not work for others. Furthermore, CTL epitope evolution is molded by host genetics: specifically, the major histocompatibility complex (MHC) alleles. MHC Class I proteins present virus epitopes on the surface of infected cells for recognition by CTLs, and certain escape mutations have become fixed in circulating viruses affecting immune control and disease progression in certain populations (22). Characterizing these immune escape variants provides insight into host immune responses associated with viral control, which informs vaccine strategies for prevention or cure. Next-generation sequencing methods provide a unique platform to investigate these escape mutations and their kinetics, as these mutations can arise at very low frequencies and fluctuate throughout infection (23). The aim of this review is to summarize current knowledge relevant to CTL escape, and to identify and highlight bioinformatics approaches used for the characterization of CTL immune escape.

1.2 Diversity of HIV-1

HIV-1 is characterized by extensive genetic diversity. HIV-1 strains are divided into four groups (M, N, O, and P), originating from four separate cross-species transmissions from chimpanzees and/or gorillas to humans (24). While HIV-1 groups O, N, and P are mainly restricted to Central Africa, group M has caused the HIV-1 pandemic (25-27). HIV-1 group M has been further classified into 10 distinct subtypes (A, B, C, D, F, G, H, J, K and L), sub-subtypes (A1 and A2 for subtype A, F1, and F2 for subtype F), and inter-subtype circulating recombinant forms (CRFs) (**Figure 1.2**). Subtypes and sub-subtypes arose from founder effects at different timepoints in the past, and inter-subtype recombinants can arise in individuals co-infected with two different subtypes. If these newly recombined strains have a significant degree of epidemic spread, they are called Circulating Recombinant Forms (CRFs) (28, 29). Geographically, subtype C is predominant in Southern Africa, Ethiopia, and India (30, 31). Subtype A is distributed across Eastern Europe, Central Asia, and extending from Western to Eastern Africa. Subtype B is predominant in the Americas, Western Europe, and East Asia. Subtype D is most prevalent in North Africa and the Middle East but is also common throughout East and Central Africa. Subtype G is commonly

found in Central Africa and West Africa (32). The genetic diversity of HIV-1 found among different populations is a major challenge for vaccine development.

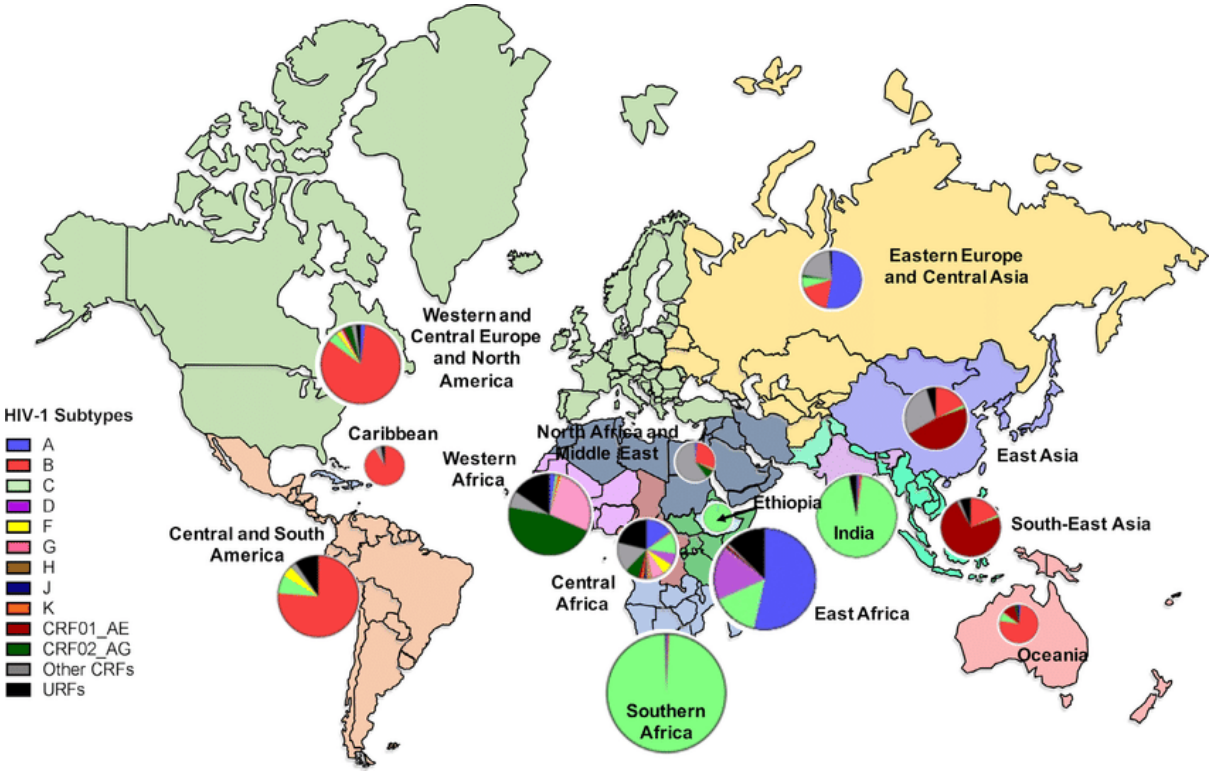


Figure 1.2: World map illustrating the geographical distribution of HIV-1 group M subtypes within each region. Pie graphs show the percentage of each subtype that circulates within a region and the size of each pie represents the total number of infections in that region. Each region is colour coded. This map was adapted from subtype prevalence data from Hemelaar et al., 2019 and infection prevalence data from UNAIDS Data 2019 (33).

Within an infected host, high genetic variability is a major facilitator of immune evasion. The virus population at a particular timepoint within an infected person consists of a complex mixture, or swarm, of genetically distinct variants, known as “quasispecies” (34, 35). The extensive genetic diversity is due to the high replication rate and the error-prone reverse transcriptase which lacks proofreading and results in high mutation rates per replication cycle. In addition to point mutations, HIV-1 evolves through recombination which occurs when there is strand switching between co-packaged viral RNA molecules during reverse transcription (36, 37).

1.3 The HIV-1 genome and proteins

The HIV-1 genome is approximately 9.8 kb long and encodes nine proteins including three major structural proteins, Gag, Pol, and Env, the regulatory proteins Tat and Rev, and the accessory proteins Vif, Vpu, Vpr and Nef (38) (**Figure 1.3**). The genome exists as two viral RNA molecules encapsulated in the virion, or as double-stranded DNA integrated into the human genome, referred to as the provirus. When in its DNA form the genome is flanked at both ends by LTR (long terminal repeat) sequences. The 5′ LTR region codes for the promoter of viral gene transcription. In the direction 5′ to 3′, the reading frame of the *gag* gene encodes the proteins of the outer core membrane, matrix protein (MA, p17), the capsid protein (CA, p24), the nucleocapsid (NC, p7), and the nucleic acid-stabilizing protein, p6. Following the *gag* reading frame, there is a *pol* reading frame which codes for the enzymes protease (PR, p12), reverse transcriptase (RT, p51) and Ribonuclease H (p15) or RT plus Ribonuclease H (together p66) and integrase (IN, p32). Adjacent to the *pol* gene, the *env* reading frame follows from which the two envelope glycoproteins, gp120 (surface protein, SU) and gp41 (transmembrane protein, TM), are derived. In addition to the structural proteins, the HIV-1 genome codes for several regulatory and accessory proteins: Tat (transactivator protein), Rev (RNA splicing-regulator), Nef (negative regulating factor), Vif (viral infectivity factor), Vpr (virus protein r), and Vpu (virus protein unique) (38, 39). **Table 1.1** details the functions of the different proteins found in the HIV-1 genome.

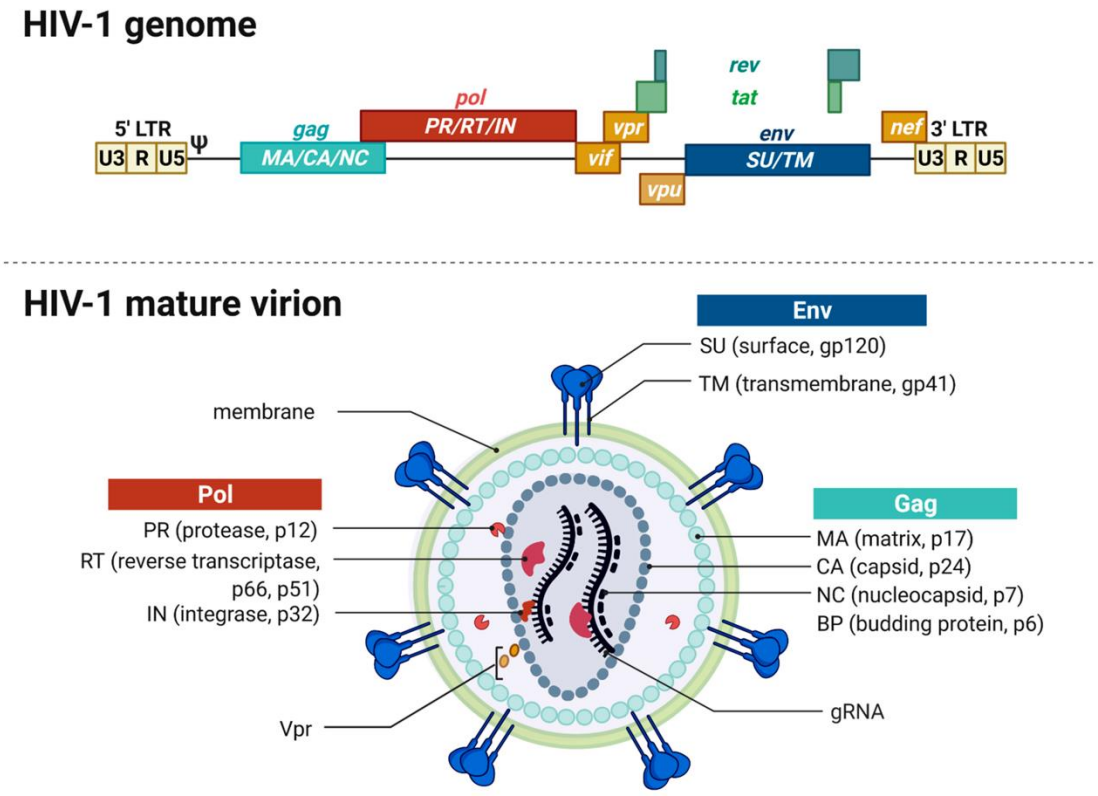


Figure 1.3: The HIV-1 genome organization, and virion structure. The illustration was obtained from Van Heuvel et al., 2022 (40).

Table 1.1: HIV-1 proteins and their functions (41).

Gene	Size	Protein	Function
gag		Pr55Gag	precursor of the inner structural proteins
	p24	capsid protein (CA)	formation of conical capsid
	p17	matrix protein (MA)	myristoylated protein, forming the inner membrane layer
	p7	nucleoprotein (NC)	formation of the nucleoprotein/RNA complex
	p6		involved in virus particle release
pol		Pr160GagPol	precursor of the viral enzymes
	p10	protease (PR)	proteolytic cleavage of Gag (Pr55) and Gag-Pol (Pr160GagPol) precursor protein; release of structural proteins and viral enzymes
	p51	reverse transcriptase (RT)	transcription of HIV-1 RNA to cDNA
	p15 (66)	RNase H	degradation of viral RNA in the viral RNA/DNA replication complex
	p32	integrase (IN)	integration of proviral DNA into the host genome

env	gp160	PrGp160	precursor of the envelope proteins SU and TM, cleavage by cellular protease
	gp120	surface glycoprotein (SU)	attachment of virus to the target cell
	gp41	transmembrane protein (TM)	anchorage of gp120, fusion of viral and cell membrane
tat	p14	transactivator protein	activator of transcription of viral genes
rev	p19	RNA splicing regulator	regulates the export of non-spliced and partially spliced viral mRNA
nef	p27	negative regulating factor	myristilated protein, influence on HIV-1 replication, enhancement of infectivity of viral particles, downregulation of CD4 on target cells and HLA cells on target
vif	p23	viral infectivity protein	critical for infectious virus production in vivo
vpr	p15	virus protein r	component of virus particles, interaction with p6, facilitates virus infectivity, effect on the cell cycle
vpu	p16	virus protein unique	efficient virus particle release, control of CD4 degradation, modulates intracellular trafficking

1.4 The natural history of HIV-1 in an individual

HIV-1 infections can be categorized into three phases. The first is the acute or primary infection phase and is characterized by rapidly rising viral loads which then decline to a steady state, lasting three to six months (42). At this phase, HIV-1-specific CD8+ T-cell lymphocyte (CTL) responses appear and expand to ~10% of all circulating CD8+ T cells as viremia peaks. This early T cell response is probably responsible for much of the reduction in virus load (43, 44). The viral load steady state attained following acute infection, is a predictor of disease progression (43).

The second phase is the chronic or asymptomatic phase. During this phase, the viral load remains relatively stable due to the equilibrium that is reached between virus production and clearance under the action of CTL and neutralizing antibody responses (45). The ongoing emergence of CTL escape variants implies the virus is continuously under selective pressures to evade at least partially effective CTL responses. The CTL

response evolves over time, often with different patterns of immunodominance during the chronic phase compared with the acute infection phase, which may in part reflect the selection of virus escape mutants and consequent new CTL responses (46).

The final phase, AIDS (acquired immunodeficiency syndrome) is characterized by a decline in CD4+ T cells and a sharp increase in viral loads (47). Understanding the natural history of infection is a significant step for researchers in designing cure and vaccine strategies. In the CAPRISA (Centre for the AIDS Programme of Research in South Africa) 002 cohort investigated in this thesis, predictors of disease progression in subtype C infected South African women were classified according to CD4 counts and viral load estimates within a particular time range (48). This cohort was set-up in 2004 and recruited women within 3 months of infection and followed them over time. Women accessed treatment according to prevailing treatment guidelines of South Africa. In this study individuals were classified as rapidly progressing to disease when they had CD4 counts less than 350 cells/μL between six months to two years following infection. Individuals with viral loads consistently below 2000 copies/mL or CD4 counts greater than 500 for more than 8 years were classified as slow progressors. An intermediate progressor was classified as anyone who did not fit in the slow or rapid categories or had CD4 counts greater than 350 cells/μL beyond two years (49, 50).

1.5 Cytotoxic T Lymphocyte (CTL) escape

HIV-1 rapidly escapes immune responses exerted by an infected host (37). After cell entry, reverse transcription, and integration of proviral DNA into the infected cell genome, viral proteins are produced and processed into peptide epitopes by the cellular machinery. Epitopes, which typically range from 9 to 14 amino acids in length, are loaded onto HLA class I molecules for presentation at the cell surface. CTLs are CD8-positive T cells that are specialized in the direct killing of infected cells (particularly those infected with viruses), cancerous cells, or cells that are damaged in other ways (51). Recognition of the epitope–HLA complex by a T cell receptor (TCR) expressed by a CTL results in CTL-mediated elimination of the infected/cancerous/damaged cell (22). Immediately following the initiation of an HIV-1 infection, the virus begins to adaptively evolve and, in so doing, viral genomes begin accumulating mutations that enable escape from immune responses (52, 53). CTL

escape takes place when virus-infected cells are no longer recognized by CTLs. The mutations mediating this escape, called escape mutations, can fall into three categories based on different mechanisms of evasion. Firstly, escape mutations upstream of the epitope can impact intracellular epitope processing. Secondly, escape mutations can directly impact epitope-HLA binding, such as those that occur at sites within epitopes, called anchor residues, that are crucial for the strength and specificity of epitope binding to HLA molecules (usually at sites 2 or 9 of the epitope). Thirdly, some escape mutations reduce or abrogate recognition of epitope-HLA complexes by T cell receptors TCRs (54).

Genes encoding HLA class I are among the most polymorphic in the human genome with everyone expressing up to six different class I alleles (two at each of the A, B, and C clusters) out of a pool of thousands of known human HLA allelic variants (55). This extensive host genetic diversity serves as a mechanism whereby the human immune system, on the scale of both the individual and the population, is equipped to recognize a vast array of epitopes from a broad range of pathogens. Some HLA alleles, (for example, HLA-B*57, HLA-B*27, and HLA-B*51) are more likely than others (for example B*35) to mediate successful control of infection and slower disease progression (54, 56). Also, the correlation between a host's HLA alleles and their ability to control HIV-1 is mediated by the ability of the alleles to target specific viral epitopes. This is supported by studies suggesting that the targeting of epitopes in the HIV-1 Gag protein is correlated with control while the targeting of epitopes in Env and Nef is correlated with disease progression (57, 58). After viral entry, Gag has been shown to have a high abundance relative to other viral proteins, and it contains protective epitopes. For these reasons it has been hypothesized that, the targeting of Gag leads to better disease control (59).

On a population level, the value of characterizing CTL escape patterns or targeted epitopes would be to identify correlates of protective immunity. Specifically, HLA-associated polymorphisms in the genomes of circulating viral variants can serve as markers of viral sites evolving under HLA-allele-mediated immune pressures that are sufficiently strong *in vivo* to promote the selection of virus escape mutations (60). Protective HLA alleles exert strong selection pressures on functionally constrained

sites and preferentially select escape mutations at HLA anchor residues more than non-protective HLA alleles. For example, B*27 restricted KK10 epitope escape proceeds via selection of the R264K mutation, which results in a significant defect in viral replication (61). Information on the accumulation of such escape mutations can be incorporated into HIV-1 vaccine strategies (62-64).

Immunodominant epitopes are known to provoke strong immune responses, because these epitopes are most easily recognized by the immune system and have the most influence on the specificity of the induced CTL responses. A study by Liu et al., 2013 shows that CTL escape occurs more rapidly in high entropy epitopes and that, although CTL responses to conserved epitopes arise later, they are more effective in viral control. Additionally, the study observed shifts in T cell immunodominance hierarchies early in HIV-1 infection (65). The dominant T cell response was mostly focused on the *Nef*, *Env*, and *Gag* proteins while *Pol* was targeted infrequently. Examples of immunodominant epitopes include KK10 (*Gag*131-140), GL9 (*Nef*94-102), RM9 (*Nef*71-79), and TL9 (*Gag*180-188) (66, 67). It is advantageous to identify epitopes containing mutations that reduce viral fitness as these make good vaccine targets (68).

1.6 The identification of CTL escape epitopes and mutations

The prediction of CTL escape using HIV-1 sequences is not straightforward. It requires (i) knowing how to analyze sequences for changes (polymorphisms) associated with CTL pressure, (ii) distinguishing these changes from those occurring due to random mutation processes or other selective factors and (iii) whether over time the mutation becomes dominant and fixed in the population. CTL escape can initially be detected within the first weeks of infection and is identifiable as: (i) changes in known HLA-targeted epitope sequences where single amino acid polymorphisms arise and gradually increase in frequency; (ii) different amino acid residues within a 9-11 epitopic region toggling back-and-forth over time with/without one of the alternatives eventually becoming fixed and (iii) clusters of amino acid polymorphisms occurring within 9 to 11 amino acid stretches (69-71). These polymorphic amino acid sites will frequently be detectably evolving under positive selection such that standard positive selection analyses can be used to support the identification of predicted escape mutations.

Many studies have attempted to identify evidence of escape using participant HLA information and these studies have applied a variety of tools including:

- ELF (Epitope Location Finder): a tool that is used to search a submitted protein sequence for both known epitopes from an immunology database, and epitopes predicted based on consensus binding motifs (72). ELF does this by initially screening the given epitope against the LANL database (an annotated, searchable inventory of more than 17,000 entries of HIV-1 cytotoxic and helper T-cell epitopes and antibody binding sites, integrated with sequence variability data from the LANL HIV-1 Sequence Database) which consists mostly of epitopes identified in HIV-1 subtype B studies. ELF Identifies anchor residues within a query sequence, highlights differences between the predicted query epitopes and the database epitopes, and estimates binding affinity scores between predicted epitopes and HLA molecules. https://www.hiv.lanl.gov/content/sequence/ELF/epitope_analyzer.html
- MotifScan: a web-based tool for finding HLA anchor residues in proteins or peptides (73) that works similarly to ELF. https://www.hiv.lanl.gov/content/immunology/motif_scan/motif_scan
- NetMHCpan: a tool that predicts the binding of any potential epitope peptide irrespective of sequence to a wide range of known MHC molecules using artificial neural networks (ANNs) (74, 75). This tool is not restricted by HIV-1 subtypes (i.e., it has not been exclusively trained using the known binding affinities of HIV-1 derived peptides) and it considers a broader range of HLA alleles than other tools. It is more likely to yield epitope binding predictions than ELF. If an identified HLA type or genotype is not found in the database of known MHC molecules, the closest relative of that HLA genotype will be used for epitope prediction. https://services.healthtech.dtu.dk/service.php?NetMHCpan-4.1
- Phylogenetic dependency networks: measure the strength of selection exerted by an HLA allele on a given polymorphism (76). https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000225
- CTL epitope databases: repositories of curated data relevant to immune reactions and specific pathogens. These are used as a reference for validating

found/predicted epitopes and/or escape mutations (77). https://www.hiv.lanl.gov/content/immunology/index.html

- Manual identification of polymorphisms in known epitopes in amino acid sequence alignments.

As outlined previously (see section 1.5), CTL escape mutations can be of three main types, two of which are expected to cause amino acid polymorphisms either within epitopes (such as at an anchor residue that impacts binding to HLA molecules or changes in residues recognized by TCR) or in sequences flanking epitopes (such as at a site which inhibits processing of epitopes for presentation on HLA molecules). The impacts of escape mutations within epitopes on HLA binding can be either verified experimentally using IFN-γ ELISPOT assays, or they can be computationally inferred based on the predicted impacts of potential escape mutations on HLA binding (78, 79).

Table 5.1 (appendix) provides a list of studies, mostly from southern African populations, that screened for CTL escape mutations in sequence data, with a focus on the tools and methods used to identify/predict CTL escape mutations. This table provides a summary of the sequencing methods, analysis tools/methods and cohorts/datasets that were used for identifying targeted epitopes for CTL escape previously. From this selection of studies, only five used the known tools in combination while the rest mostly used different models as well as manual identification. Despite the variety of tools that are available there remains no straightforward, robust, flexible, and widely applicable analysis pipeline for the identification of CTL escape mutations in deep sequencing data. Furthermore, deep sequencing provides the advantage of enabling a more detailed insight into CTL escape kinetics.

1.7 Next-generation sequencing as a tool for characterizing CTL escape

Deep sequencing is a high-throughput method that allows the sequencing of billions of nucleotides in a single run with the aim of sampling the viral genome extensively. The critical difference between Sanger sequencing and deep sequencing approaches is the sequencing volume. While the Sanger method only sequences a single DNA fragment at a time, deep sequencing is massively parallel, sequencing millions of fragments simultaneously per run. This process translates into sequencing hundreds

to thousands of genes at one time. Deep sequencing allows the detection of low-frequency variants with high sensitivity. It has a faster turnaround time for high sample volumes, provides comprehensive genomic coverage, lower limits of detection, higher capacity with sample multiplexing and the capacity to sequence hundreds to thousands of genes or gene regions simultaneously (80-82).

1.8 Rationale

There are a limited number of studies that have investigated HIV-1 CTL escape dynamics in deep sequencing data. Because of this there is also a lack of pipelines, software, and more automated workflows to handle the large volumes of data that are needed to reliably detect CTL escape mutations. The importance of this study is to yield a better understanding of immune responses, prevalence, and the kinetics of CTL immune escape with the use of next-generation sequencing datasets and a combination of predictive tools.

1.9 Research Aim and Objectives

The aim of the study was to develop a workflow for identifying the kinetics of CTL escape in longitudinal HIV-1 next-generation datasets of *gag* sequences generated using an Illumina Miseq platform.

The objectives of this study were:

1. To explore existing tools used for identifying CTL escape and their applicability to deep sequencing datasets.
2. To develop a workflow to identify CTL escape patterns in deep sequencing data.
3. To describe CTL escape dynamics in a longitudinal deep sequencing dataset from 15 women from the CAPRISA 002 cohort.

Chapter 2: Methods and Materials

2.1 Study population

This study was approved by the Human Research Ethics Committee (HREC), Faculty of Health Sciences at the University of Cape Town (UCT), South Africa (HREC 797/2020). This study is a sub-study of an NIH-funded R01 study (HREC 588/2015) and utilized Illumina Miseq generated sequences of a partial p24 region of Gag (HXB2 nucleotide position 1255 to 1798; amino acid position 156 to 337). These data were generated for 15 women living with HIV, who were part of the Centre for the AIDS Program of Research in South Africa (CAPRISA) 002 acute infection cohort. This cohort recruited recently HIV-infected women from KwaZulu-Natal and followed them from infection to the time of ART initiation (according to the WHO / South African guidelines at the time) and for up to five years thereafter (83). Sequences were generated by Lynn Tyers and Deelan Doolabh (Division of Medical Virology, UCT) from samples taken at 6-month intervals from infection until ART initiation.

2.2 NGS sequence alignment processing

The sequencing method used to generate and process the sequences has been published in Abrahams et al., 2019 (84). The HXB2 *gag* nucleic acid reference sequence (position 1255 to 1798) was added to the longitudinal Illumina Miseq generated p24 Gag sequences using an in-house script written by Dr Anna Yssel, Division of Medical Virology, UCT (Appendix 5.2). The sequences were codon-aligned using the MACSE tool (v2) (85), and the AliView tool (version 1.1) (86) was used for manual inspection and editing of misaligned regions. The nucleotide sequences were then translated to amino acid sequences. Following these steps, the sequences were de-gapped using the web tool, GapStreeze (https://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html) (87) with default settings and a gap tolerance of 50%. Sequences of epitopes known to bind to a participant's HLA were extracted from the aligned sequences using the AliView tool for further analysis. Epitopes with missing amino acids were removed.

2.3 Epitope analysis

EpitopeMatcher was run under Rstudio using the shiny package, run_EpitopeMatcher_app() (https://github.com/philliplab/EpitopeMatcher.git) (88). The user interface required (i) a patient HLA genotype csv file which lists all the participant's HLA alleles, (ii) a list of experimentally-verified HIV CTL/CD8+ epitopes also in a csv file (https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html) which is regularly updated with experimentally verified epitopes, and (iii) a fasta format amino acid sequence file of the query alignment (89). An amino acid and nucleotide sequence alignment from which identical sequences within a timepoint were collapsed into a single sequence, was used for this analysis. A summary of input and output files are recorded in **Table 2.1**. As output, the tool provides a match and mismatch score between the query and reference sequence and the location of mismatches within the epitope, providing a csv file of all CTL epitopes reported as restricted by the participant's HLA alleles. Once CTL epitopes were identified, these were trimmed from the p24 Gag alignment (prior to de-duplication) and used to generate logograms using the web tool, AnalyzeAlign (https://www.hiv.lanl.gov/content/sequence/ANALYZEALIGN/analyze_align.html) (90). Default program settings were used throughout, and sequences were grouped by timepoint or by phase of infection, where information per phase was provided in a text file. NetMHCpan (version 4.1) was used for the validation experiment. As input, we used a Sanger sequence test dataset where immune selection was previously mapped to full-genome sequences (69). Settings were left on default, and the peptide length was set between 8 and 14-mer for this analysis. A summary of input and output files for NetMHCpan can be found in **Table 2.1**).

2.4 Positive selection analyses

The Hypothesis Testing using phylogenies - Fast, Unconstrained Bayesian AppRoximation (HyPhy–FUBAR), a Bayesian statistic orientated approach to inferring nonsynonymous (dN) and synonymous (dS) substitution rates on per-site basis for a given coding alignment and corresponding phylogeny (91), was used to identify codon sites that were potentially evolving under positive selection. The tool was run via the command line and as input the tool accepts HXB2 codon aligned sequences in fasta

format (without stop codons) and a phylogenetic tree describing the evolutionary relationships of the sequences in .nwk file format. The phylogenetic tree was generated using Hyphy. FUBAR outputs a Json file containing information of the codon sites evolving under both positive (probability[alpha<beta]) and negative (probability [alpha>beta]) selection. The Json file was converted to a csv file and the sites were mapped to Gag HXB2 positions corresponding to amino acid positions in sequence files (HXB2 GAG coordinates 156 to 336). The Entropy (Shannon Entropy-one) tool, (https://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_one.html) which measures variation in sequence alignments was used to calculate the entropy score as well as the amino acid frequency at each site in all the 15 sequence alignments. As input the tool accepts fasta format query amino acid alignments and outputs entropy plots indicating the entropy score at each site of the amino acid alignment. Additionally, the tool outputs tables with entropy scores and amino acid frequency at each position (92) (**Table 2.1**).

Table 2.1: Input and output options for Epitope Matcher, NetMHCpan, Hyphy-FUBAR and Entropy tools.

Tool	Input and settings	Output
Epitope macher	● Csv format participant HLA genotype file. ● Csv format LANL CTL epitope database. ● Fasta file query amino acid sequences.	● Csv format list of HLA-restricted epitopes and identified mismatches to the HXB2 reference sequence
NetMHCpan	● Fasta format file of query amino acid sequences. Options on tool: ● Peptide length (desired k-mers) ● HLA alleles (participant HLA) ● Additional configuration: ○ Threshold for strong binder (%rank): 0.5 ○ Threshold for weak binder (%rank): 2 ○ Include BA predictions: tick ○ Sort by prediction score: tick ○ Save prediction on XLS file: tick	● Web server output ● Csv format output of potential restricted epitopes and binding scores (downloadable).
HyPhy FUBAR	● Fasta format query coding sequence file ● .nwk format tree file corresponding to query coding sequence file. Options on Hyphy tool: ● 1: Natural selection ● 4:FUBAR ● Path to coding sequence and tree files	● JSON format probabilities of sites evolving under natural selection (positive and negative selection)
Entropy (Shannon Entropy-one)	● Fasta format query amino acid sequence file.	● .png format entropy plot ● Table format entropy scores ● Amino acid frequency per site.

Chapter 3: Results

3.1 Review of tools for identifying CTL epitopes and putative escape mutations in deep sequencing datasets

In order to explore the ability of various available software tools and published methods to identify CTL epitopes and putative escape mutations in next-generation sequence datasets, a list of tools was identified through online searches and scientific journal articles. Four tools were identified: (i) Epitope Location Finder (72); (ii) MotifScan (73); (iii) NetMHCpan (75); and (iv) Epitope Matcher (88). Another two reported methods identified involved the use of phylogenetic models (76) and manual mapping of CTL epitopes using online CTL epitope databases. The abilities of these different approaches to identify CTL epitopes and putative escape mutations were evaluated based on the following criteria:

1. Function: Can the tool/method detect potential HLA-epitope interactions and possible escape mutations in our subtype C dataset?
2. Processivity: Is the tool/method capable of analyzing NGS datasets?
3. Source database: Does the tool/method screen the given sequence against a database of known CTL epitopes?
4. Accessibility: Is the tool/method current and operational?
5. Reliability: Has the tool/method been published and verified for accuracy?
6. Output: Does the tool/method output HLA-associated CTL epitopes and polymorphisms?
7. Design: What kind of algorithm was utilized in designing the tool?

A summary of the comparison of these tools/methods is provided in **Table 3.1**.

Two of four tools did not meet the selection criteria:

(i) Epitope Location Finder (ELF) (https://www.hiv.lanl.gov/content/sequence/ELF/epitope_analyzer.html) is limited as the input it accepts is protein sequences of less than 50 amino acids long. (ii) MotifScan (https://www.hiv.lanl.gov/content/immunology/motif_scan/_motif_scan) searches for HLA anchor residue motifs, but does not automatically compare this output with an experimentally verified epitope list, thus requiring manual evaluation. Due to the vastness of deep sequencing

datasets, manually screening databases is time consuming. Phylogenetic methods, as previously reported by Carlson et al., 2012 have been used to identify HLA associated polymorphisms in HIV sequences and can be used to support findings, however the dataset in this thesis is not large enough to apply this method.

Two methods met the criteria and were explored further in this thesis: (i) Epitope Matcher, which uses a participant's HLA to find how well that HLA characterises experimentally verified epitopes, and (ii) NetMHCpan, which predicts binding of epitopes to any known HLA types, showed potential for identifying CTL epitopes and putative escape mutations using deep sequencing datasets. These tools are described in greater detail below.

Table 3.1: List of tools to predict CTL epitopes and escape mutations.

Tool	Epitope Location Finder	Motif Scan	NetMHCpan	Epitope Matcher
Function	Uses HLA anchor motifs to predict epitopes.	Searches any single protein for all known HLA anchor residue motifs, allows viewing of motif libraries.	Predicts peptides binding to an MHC molecule of known sequence using artificial neural networks.	An R package that finds how well the epitopes in a participant's virus will be recognized by the patient's HLA.
Input	Protein sequence < 50 aa.	Protein sequence.	At most 5000 sequences; not 6<AA>20000. Select HLA/MHC to be predicted.	Sequence fasta file Participant HLA type LANL CTL epitope database.

Output	Potential targeted epitopes ordered by HLA.	A table of anchor residues recognized by a given HLA type. A motif score p-value measuring the binding affinity for each region sequence and motif target site.	Table of predicted epitopes based on anchor residue binding affinity.	List of all epitopes restricted by participant's HLA alleles with mismatches indicated, in a csv format file.
Reference HIV CTL epitope source	Experimentally verified epitopes (LANL)	Marsh2000, SYFPEITHI database	Binding affinity and eluted ligands peptides	Experimentally verified HIV CTL epitopes-LANL
HLA allele required?	Yes	Yes	Yes	Yes

3.1.1 Epitope Matcher

Epitope Matcher is a tool that predicts how well the epitopes in a participant's viral population will be recognized by user specified HLAs based on amino acid sequence similarity to a reference. The algorithm calculates a match and mismatch score when comparing the epitopes from the query sequence(s) and reference sequences of experimentally verified HLA-targeted epitopes. An epitope is seen as not found, if the query epitope and the reference do not contain the same number of bases (88).

The tool is provided as an R package that can either be run with a browser-based GUI or from the command-line. The input requires the HLA genotype of the participant, the CTL escape epitope list, which is a list of curated and verified CTL epitope information (https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html) and the HIV sequence fasta file. As output, it generates a csv file with all searches from the query alignment, which are of all the epitopes that are picked up in the individual fasta file.

For a given epitope associated with the individual's HLA genotype, it provides a hamming distance between the query sequences and reference for that epitope location. It therefore allows one to identify sites that are changing over time when examining longitudinal data.

3.1.2 NetMHCpan

NetMHCpan is an online tool that predicts binding of any potential epitope peptide, irrespective of sequence length, to a wide range of known MHC molecules using artificial neural networks (ANNs) (74, 75). This tool has not been exclusively trained using the known binding affinities of HIV derived peptides, and therefore considers a broader range of HLA alleles than other tools. It comprises a large database of known peptide-HLA-Class I interactions and is trained on a combination of more than 850,000 quantitative Binding Affinity (BA) and Mass-Spectrometry Eluted Ligands (EL) peptides). If the HLA type or genotype is found to be unknown, the closest relative of that HLA genotype will be used for epitope prediction. The tool considers the protein sequence and the HLA information and generates quantitative predictions of the affinity of any peptide-HLA-I interactions. NetMHCpan returns as default the likelihood of a peptide being a natural ligand of the selected MHC(s) and provides information on the binding affinity. The peptide is identified as a strong binder if the % rank is below the specified threshold for strong binders (which usually is 0.5 %), while a weak binder will be between 0.5 and 2 %. Therefore, when examining longitudinal data, it is possible to identify changes in HLA-targeted epitopes that result in a change in MHC binding affinity and therefore add support to predicting CTL escape. Thus, results generated from Epitope Matcher and NetMHCPan could potentially complement each other.

3.2 Validation of selected tools

To validate whether the two tools selected could predict targeted epitopes and identify putative escape mutations, we analyzed their ability to identify previously reported epitope targets and immune selection from a published study. This study generated a longitudinal full-genome HIV-1 dataset from five CAPRISA 002 participants sampled in acute and early HIV-1 infection using Sanger sequencing (69). Tables with detailed information (demographic and sequencing information for five CAPRISA002 participants, putative CTL escape epitopes and polymorphisms) are found in the Appendix **Tables 5.3.1 and 5.3.2**. This dataset consisted of 112 near full-length viral genomes generated from, on average, three timepoints, including, on average nine sequences at screening/enrolment (2-5 weeks post-infection), six at three months (11-

13 weeks post-infection), and nine at six months post-infection (22-29 weeks post-infection). Additional sequences (half-genome, SGA and clonal) were generated from various timepoints ranging from 2 to 117 weeks post-infection. The study identified 21 epitopes with evidence of CTL escape in 8 regions of the genome (Vif, Tat, Rev, Nef, Pol, Vpr, Gp41, Gag) (**Table 5.3.1**). For CTL epitope prediction, the study used the tools Epitope Location Finder and NetMHCpan 2.2 as well as manual identification through CTL escape databases. Furthermore, predictions of 13 of the 21 epitopes were supported by IFN-γ ELISPOT screening (Liu et al., 2013) (65). We investigated whether Epitope Matcher and NetMHCpan would detect these epitopes and the sites undergoing change.

Epitope Matcher identified 81% of the epitopes previously reported for the test dataset with evidence of CTL escape (**Figure 3.1A**). Epitope Matcher further identified 93% of the sites undergoing change (**Figure 3.1B**). NetMHCpan predicted only 10% of epitopes previously reported with evidence of CTL escape. In both approaches, the epitopes with long amino acid length (i.e. >15 amino acids) were not analyzed by the tools. There were therefore size restrictions that were identified in these tools, as each tool was provided with a full alignment to identify or predict possible epitopes, and Epitope Matcher only identified known epitopes from the B list that were between 8 to 14 amino acids long. In addition, NetMHCpan only provides outputs for epitopes in the range of 8 to 14 amino acids as per tool settings. Two epitopes with lengths 18 amino acids each were therefore missed. Two epitopes that were predicted by NetMHCpan, described in Abrahams et al., 2013 (69), were not detected by Epitope Matcher. Epitope Matcher further identified a total 46 epitopes which were not reported by Abrahams et al., 2013 (69). This may be due to the regular updating of the epitope database on LANL.

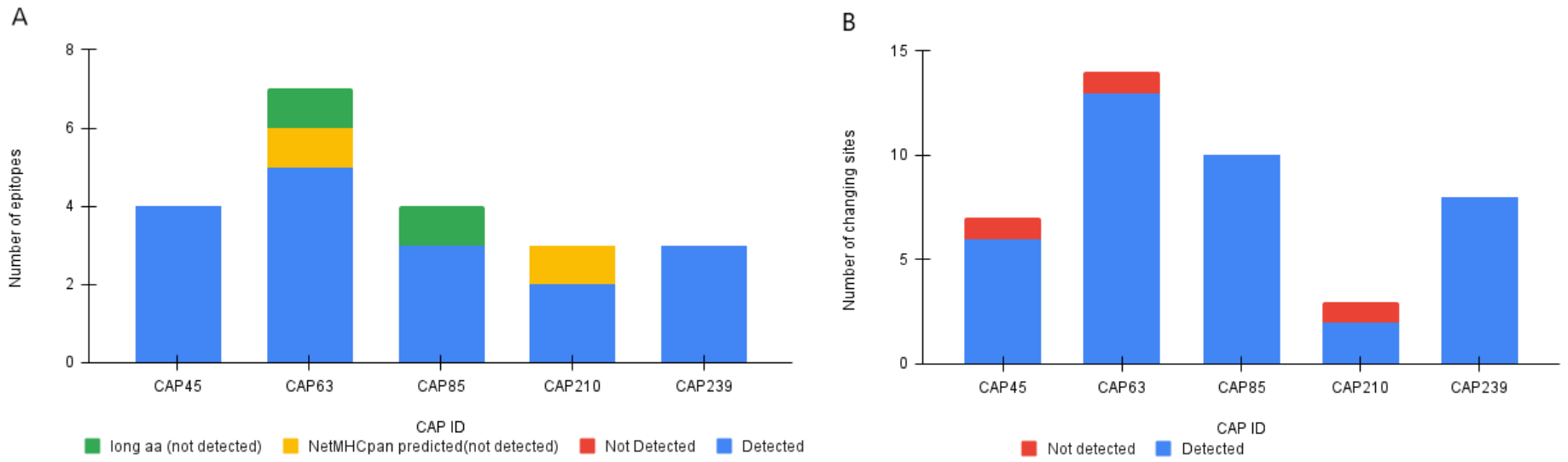


Figure 3.1 Analysis of the ability of the Epitope Matcher tool to detect previously identified epitopes and/or mutations in a near full-length genome dataset. **A**, Number of epitopes detected by Epitope Matcher. Colour codes: Blue represents detected epitopes, red represents epitopes not detected, yellow represents epitopes that were not detected by Epitope Matcher because they were predicted by NetMHCpan in the original study, green represents epitopes not detected because of the size of epitope (>14 amino acids long). **B**, Amino acid sites undergoing change as identified by Abrahams et al., 2013 compared to those identified by Epitope Matcher. Changing sites detected by Epitope Matcher (Blue) and changing sites not detected by Epitope Matcher (Red) are indicated.

3.3 Workflow optimization

Based on the validation test explored in section 3.2, Epitope Matcher identified HLA-restricted, experimentally verified CTL epitopes and polymorphisms within these epitopes. In contrast, NetMHCpan did not perform well, in terms of independently identifying all epitopes/polymorphisms described in Abrahams et al., 2013. For 90% of the investigated epitopes that were known to bind a participant's HLA, NetMHCpan incorrectly predicted that the epitopes were non-binders, even prior to the introduction of mutations associated with escape (data not shown). NetMHCpan was therefore not utilized for the analysis of CTL epitopes and escape mutations in this study.

A workflow for identifying CTL epitopes and mutations putatively associated with escape and describing the kinetics of escape in next-generation sequencing datasets that incorporated the Epitope Matcher tool was therefore developed and is detailed in **Figure 3.2**. The workflow begins with applying quality control and processing of sequence alignments followed by running the sequences through tools to identify CTL epitopes and escape mutations. For specific tools, sequences were trimmed to specified lengths of epitopes. AnalyzeAlign was used for generating logograms and frequency tables showing percentage frequencies of mutations per timepoint and per infection phase, over time. Entropy-one was used to identify high entropy (>=0.25) sites. Hyphy-FUBAR was used to identify sites evolving under positive selection. GapStreeze was used to remove gaps from the sequences with 50% tolerance.

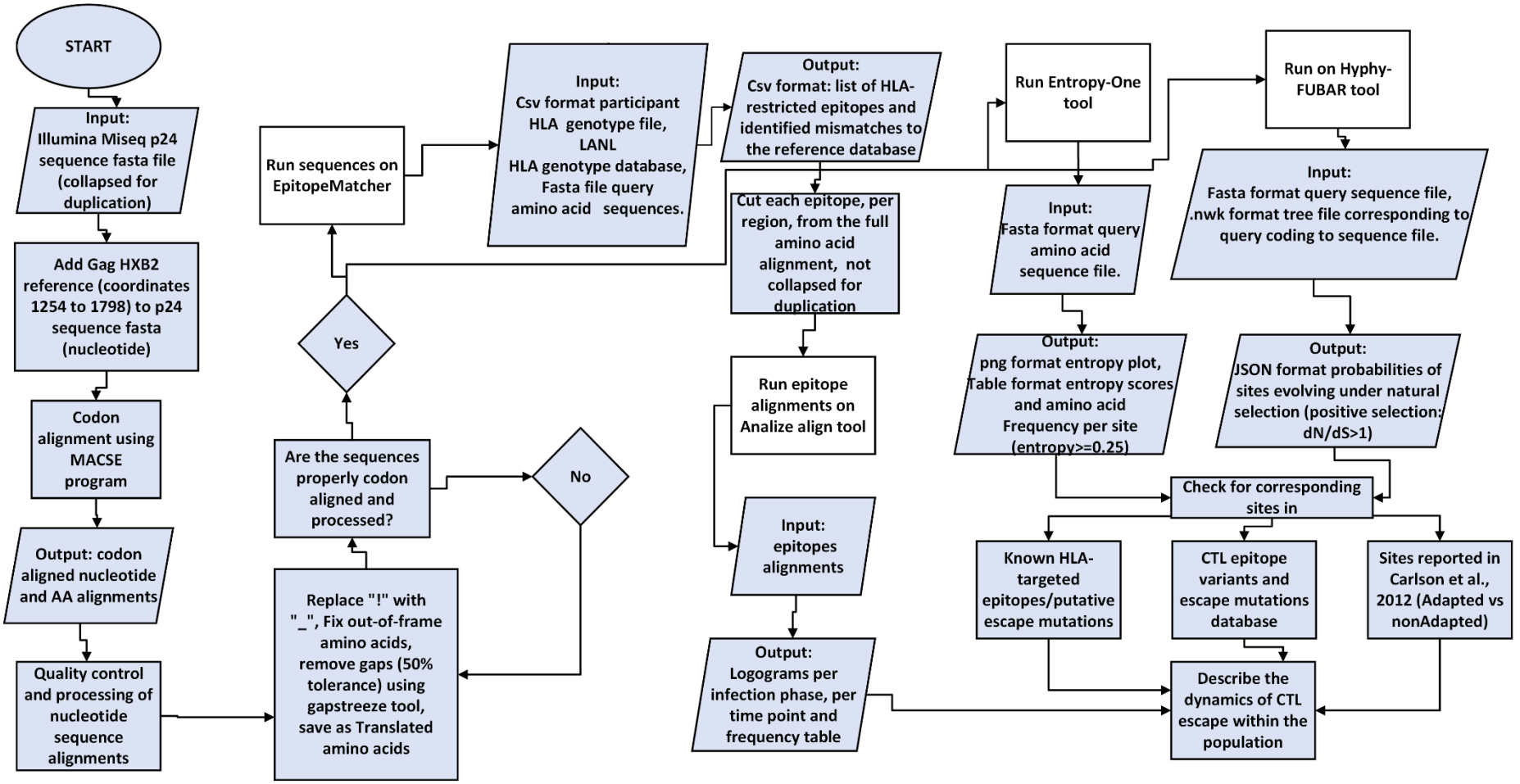


Figure 3.2 Workflow for identifying putative CTL escape in deep sequencing datasets.

3.4 CTL escape dynamics in a longitudinal deep sequencing data set from 15 CAPRISA 002 women

We evaluated immune escape in 15 participants from the CAPRISA 002 acute infection cohort (**Table 3.2**) using the designed workflow described (**Figure 3.2**). Disease progression was classified according to CD4 count and viral load within a particular time range prior to the initiation of antiretroviral therapy, as described by Mlisana et al., 2014. Rapid disease progression was classified as that individual having CD4 counts less than 350 cell/µL between six months to two years after infection. Slow progression was classified as that individual having viral loads consistently below 2000 copies/mL (or CD4 count greater than 500 cell/µL for more than eight years). Individuals experiencing intermediate disease progression were classified as those that did not fit in the rapid or slow progressor categories, or who had CD4 counts greater than 350 but below 500 cell/µL between two to eight years following infection (49, 50). Of the 15 women, seven were defined as rapid progressors, seven as intermediate progressors and one as a slow progressor. Thirteen individuals harboured protective HLA-B alleles (B*44:03, B*15:03, B*58:01, B*81:01, B*58:02, B*57:03, B*08:01), which included the slow progressor (**Table 3.2**)

The sequence data was provided by M.R Abrahams (Medical Virology, University of Cape Town). Sequences were generated using the Illumina MiSeq approach, and comprised sequence reads from the *gag p24* genome region that were 544 bp in length (HXB2 genome coordinates 1255 to 1798 nucleotide positions). The median number of sequences generated per participant was 281 (range: 41 to 566) and per timepoint was 126 (range: 3 to 429). The median number of timepoints sequenced was seven (range: 2 to 11) ranging from enrolment to 380 months post-infection. We only analysed epitopes restricted by the MHC class I HLA-B allele, as these are most frequently associated with impact on HIV disease outcome (93).

Putative escape was classified as changes found in amino acid sites of wildtype epitopes which persisted or increased to a frequency of over 5% over the sampling period (**Table 3.3**). In total, 73 epitopes were identified for the 15 women as targeted by their HLA-B allele, and putative escape mutations and/or introduction of polymorphisms over the infection period were identified in seven (9.5%) of these

epitopes (from six women), with a total of nine mutations. Eight out of nine of the mutations are found in epitopes that are known to be restricted by HLA alleles that are associated with HIV control: including B*15:03 (AEQA**T**QEVKNW, AEQ**A**TQ**D**VKNW, VKV**I**EEKAF), B*44:03 (AEQATQ**E**VK**N**W), B*08:01 (GDIYK**R**WI), and B15:10 (YVDRFFK**A**L). Sixty-six of the 73 epitopes had low (<5%) frequency mutations and were not investigated further for escape.

Table 3.2: CAPRISA 002 participant demographic data

CAPID	HLA profile	Duration of follow-up (weeks)	No. sampling timepoints	Sampling timepoints (weeks post-infection)	Disease progression
188	B*15:03, B*15:16	247	7	4, 30, 69, 132, 158, 185, 237	Intermediate
206	B*07:02, B*44:03	274	6	8, 26, 81, 133, 159, 254	Rapid
217	B*15:03, B*58:01	360	11	10, 31, 60, 138, 165, 190, 218, 242, 282, 321, 358	Intermediate
222	B*53:01, B*81:01	318	2	122, 148	Slow
244	B*44:03, B*58:02	380	11	9, 33, 56, 150, 191, 230, 255, 282, 307, 332, 359	Rapid
256	B*15:03, B*58:02	327	5	6, 106, 159, 206, 429	Rapid
257	B*42:02, B*44:03	249	7	7, 54, 80, 107, 161, 191, 240	Intermediate
280	B*15:03, B*15:10	300	9	13, 36, 63, 117, 156, 209, 222, 256, 296	Rapid
286	B*15:10, B*49:01	255	9	16, 36, 66, 93, 118, 147, 162, 175, 225	Rapid
287	B*15:10, B*42:01	260	11	9, 28, 61, 99, 112, 137, 152, 192, 216, 229, 225	Intermediate
288	B*07:02, B*15:01	210	8	6, 24, 54, 82, 107, 134, 174, 210	Intermediate
316	B*08:01, B*15:01	215	4	12, 36, 53, 90	Intermediate
336	B*08:01, B*58:02	142	4	22, 51, 73, 126	Rapid
337	B*15:10, B*57:03	166	5	9, 59, 87, 111, 163	Rapid
372	B*15:10, B*45:01	182	8	3, 27, 53, 106, 132, 161, 172, 182	Intermediate

*green illustrate HLAs that have been associated with control of viral replication (56, 93, 94).

*red illustrates HLA's that have been associated with rapid disease progression (56, 95, 96).

Table 3.3: Known HLA-associated CTL epitopes in all the participants identified by Epitope Matcher.

CAPID	Duration of follow-up (weeks)	# Of timepoints sequenced	# of longitudinal seqs	Known epitopes identified by Epitope Matcher*	Epitope name	Associated HLA allele/s	Gag HxB2 position
188	247	7	351	VKVIEEKAF	VF9	B*15:03	156-164
				VIPMFTAL	VL8	B*15:16	168-175
				YVDRFFKTL	YL9	B*15:03	296-304
				AEQATQEVKNW	AW11	B*15:03	306-316
206	274	6	499	TPQDLNTML	TL9	B*07:02	180-181
				TVGGHQAAM	TM9	B*07:02	190-198
				HPVPAGPPA	HA9	B*07:02	216-224
				GPPAPGQLR	GR9	B*07:02	221-229
				VRMYSPVSI	VI9	B*07:02	274-282
				YVDRFFKTL	YL9	B*07:02	296-304
				LSEGATPQDL	LL10	B*44:03	175-184
				AEQATQEVKNW	AW11	B*44:03	306-316
217	360	11	566	VKVIEEKAF	VF9	B*15:03	156-164
				KAFSPEVIPMF	KF11	B*58:01	161-172
				NPQDLNTML	NL9	B*58:01	180-188
				DTINEEAAEW	DW10	B*58:01	203-212

				YVDRFFKTL	YL9	B*15:03	296-304
				AEQA**T**QEVKNW	AW11	B*15:03	306-316
222	318	2	49	SDGATPSDL	SL9	B*53:01	176-184
				TPSDLNSML	TL9	B*53:01	180-188
				GHQAAMQML	GL9	B*53:01	193-201
				KDTINEEAA	KA9	B*53:01	202-210
				AEWDRLHPV	AV9	B*53:01	210-218
				HPVHAGP**V**A	HA9	B*53:01	216-224
				PPIPVGDIY	PY9	B*53:01	254-262
				DIYKRWII	DI8	B*53:01	260-267
				GLNKIVRMY	GY9	B*53:01	269-277
				RAEQATQDV	RV9	B*53:01	305-313
				QATQDVKNW	QW9	B*81:01	308-316
244	380	11	555	ATPQDLNTMLNT	AT12	B*58:02	179-190
				AEQATQ**E**VK**N**W	AW11	B*44:03	306-316
256	327	5	120	VKV**I**EEKAF	VF9	B*15:03	156-164
				ATPQDLNTMLNT	AT12	B*58:02	179-190
				YVDRFFKTL	YL9	B*15:03	296-304
				AEQ**A**TQ**D**VKNW	AW11	B*15:03	306-316

257	249	7	281	LSEGATPQDL	LL10	B*44:03	175-184
				TPQDLNTML	TL9	B*42:02	180-188
				FRDYVDRFF	FF9	B*42:02	293-301
				AEQATQEVKNW	AW11	B*44:03	306-316
280	300	9	516	VKVIEEKAF	VF9	B*15:03	156-164
				GHQAAMQML	GL9	B*15:10	193-201
				YVDRFFKTL	YL9	B*15:03	296-304
				AEQATQEVKNW	AW11	B*15:03	306-316
286	255	9	222	GHQAAMLML	GL9	B*15:10	193-201
				YVDRFFKTL	YL9	B*15:10	296-304
287	260	11	364	EEKAFSPEV	EV9	B*42:01	160-168
				TPQDLNTML	TL9	B*42:01	180-188
				GHQAAMQML	GL9	B*15:10, B*42:01	193-201
				EIYKRWII	EI8	B*42:01	260-267
				YVDRFFKTL	YL9	B*15:10	296-304
288	210	8	192	TPQDLNTML	TL9	B*07:02	180-188
				TVGGHQAAM	TM9	B*07:02	190-198
				HPVHAGPIA	HA9	B*07:02	216-224
				GPIAPGQMR	GR9	B*07:02	221-224

				VRMYSPVSI	VI9	B*07:02	274-282
				YVDRFFKTL	YL9	B*07:02	296-304
				VKVIEEKAF	VF9	B*15:10	156-164
				GLNKIVRMY	GY9	B*15:10	269-277
316	215	4	196	VKVIEEKAF	VF9	B*15:01	156-164
				DIYKRWIIL	DL9	B*08:01	260-268
				GLNKIVRMY	GY9	B*15:01	269-277
				NPDCKTIL	NL8	B*08:01	327-334
336	142	4	41	ATPQDLNTMLNT	AT12	B*58:02	179-190
				GDIYK**R**WI	GI8	B*08:01	259-266
				NPDCKTIL	NL8	B*08:01	327-334
337	166	5	188	VKVVEEKNF	VF9	B*15:01	156-164
				GLNKIVRMY	GY9	B*15:01	269-277
				VEEKNFSPEVI	VI11	B*57:03	159-169
				KNFSPEVIPMF	KF11	B*57:03	161-172
				QATQDVKNW	QW9	B*57:03	308-316
372	182	8	458	GHQAAMQML	GL9	B*15:01	193-201
				YVDRFFK**A**L	YL9	B*15:10	296-304
				EEKAFSPEV	EV9	B*45:01	160-168

*Amino acids in red and **bolded** indicate a change that was present at > 5% frequency.

*Underlined amino acids are sites with dN/dS > 1 (sites evolving under positive selection).

3.4.1 Kinetics of CTL escape in known epitopes

To characterize CTL escape, we plotted the kinetics and timing of putative escape identified in the six out of fifteen women, as illustrated in **Figures 3.3-3.9**, in the acute/early (0 to 26 weeks), early chronic (27 to 52 weeks) and late chronic (> 52 weeks) phases of infection. The wildtype (WT) virus was taken as the majority/consensus variant from the earliest sampled timepoint for this analyses.

To investigate if escape impacted the viral load trajectories of these women, we compared the occurrence of escape mutations to the participant's viral load over the duration when the escape mutation increased in frequency. To determine if there was an effect on viral load setpoint, we measured the median viral load one year before and after the occurrence of a mutation (**Figure 3.3-3.9 D**).

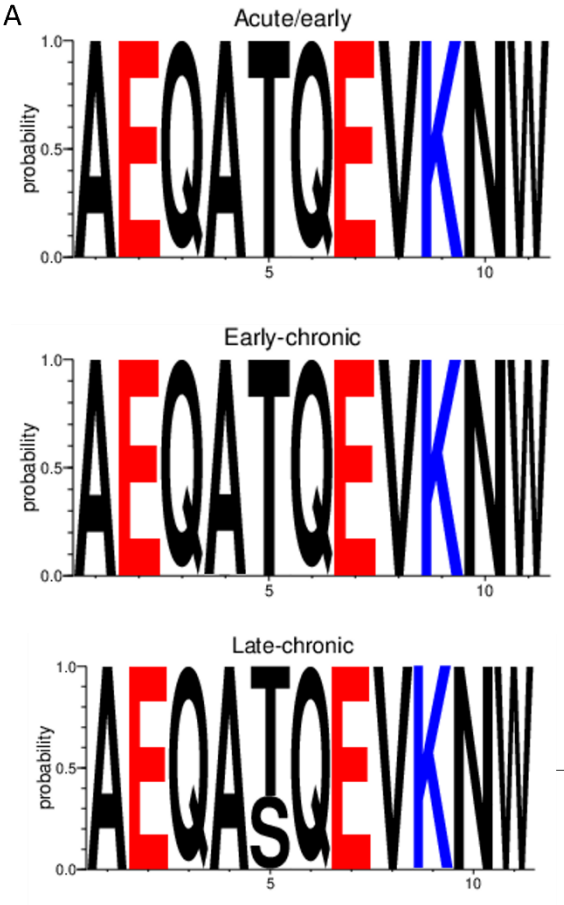
Toggling between the escape form and reversion of amino acids within an epitope has been reported to be associated with sites where escape may have a deleterious effect on virus fitness (97). Toggling was observed in five epitopes: AW11 (CAP217, CAP244, CAP256), VF9 (CAP256), and YL9 (CAP372). All of the mutations potentially associated with escape were observed in the late chronic phase of infection.

Kinetics of putative escape are described for each of the six women below.

Participant CAP217

CAP217 was an intermediate progressor, her CD4 counts remained above 350 and below 500 cell/µL during infection. Escape was identified in the AW11 epitope (Gag HXB2 coordinates 306 to 316) restricted by the HLA B*15:03, an allele associated with control of viral replication (98). We observed changes at multiple amino acid positions during the chronic phase of infection. These changes include low frequency mutations at positions 4, 8, and 9 of the epitope (A309T (0.65%), K314E (0.08%), K314R (0.08%), N315D (0.08%)) (**Figure 3.3A, B**). The T310S mutation was initially detected at 190 wpi and became dominant by 350 wpi. The increase in frequency of the T310S mutation was slow, taking nearly 52 weeks to reach 50% frequency and a further 116 weeks to reach 80% frequency (**Figure 3.3C**). We observed a spike in viral loads at 203 wpi of 269 000 copies/mL, 13 weeks after the T310S mutation was first detected. Thereafter, at 218 wpi, the viral loads decreased to 53800 copies/mL suggesting there was a temporary loss of control of viral replication associated with the occurrence and rise in frequency of this mutation (**Figure 3.3C, D**). The T310S mutation has previously been reported as an escape mutation as per the CTL epitope variants and escape mutations database (https://www.hiv.lanl.gov /content/immunology/variants/ctl_variant.html) and reported as adapted, that is, it is enriched in the presence of B*58:01, by Carlson et al., 2012.

A



B

Variant	Count	Pct.	No. of mutations
AEQATQEVKNW			
----------	16	100	0

Total sequences = 16
Number of variants = 1

Variant	Count	Pct.	No. of mutations
AEQATQEVKNW			
----------	153	99.35	0
---T-------	1	0.65	1

Total sequences = 154
Number of variants = 2

Variant	Count	Pct.	No. of mutations
AEQATQEVKNW			
----------	753	63.76	0
----S------	424	35.9	1
----S---E--	1	0.08	2
----S----D-	1	0.08	2
--------R--	1	0.08	1
--------E--	1	0.08	1

Total sequences = 1181
Number of variants = 6

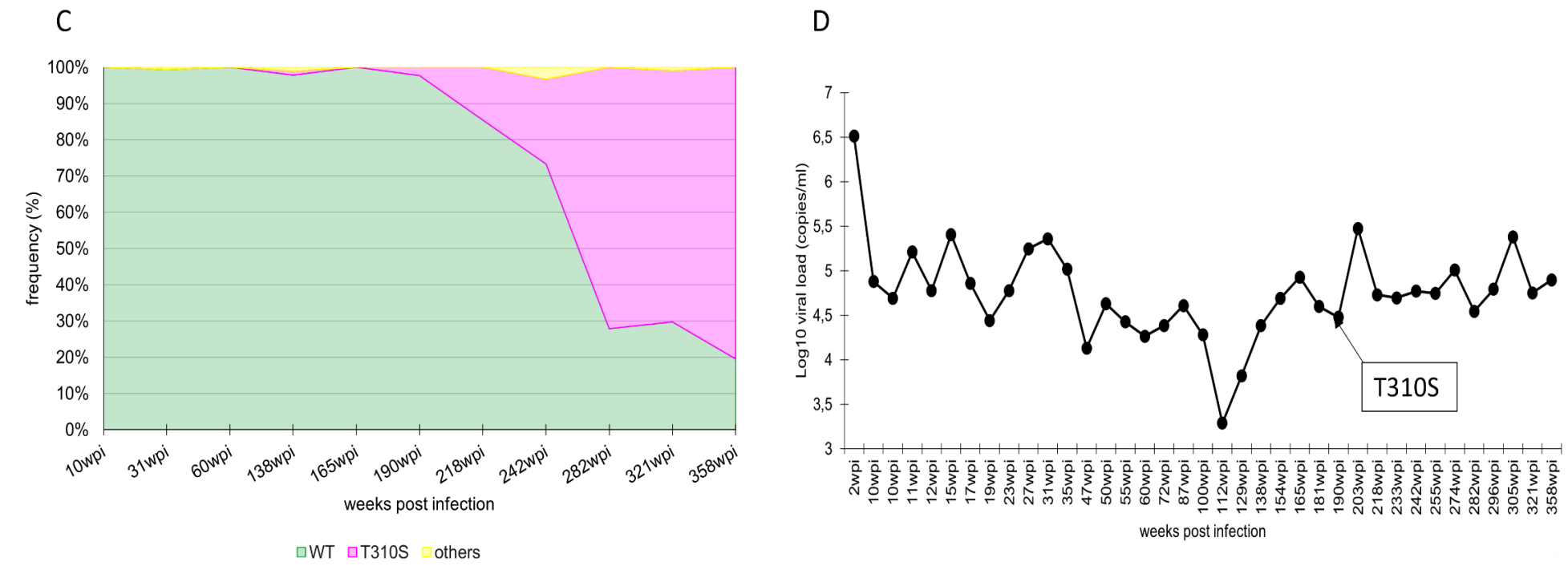


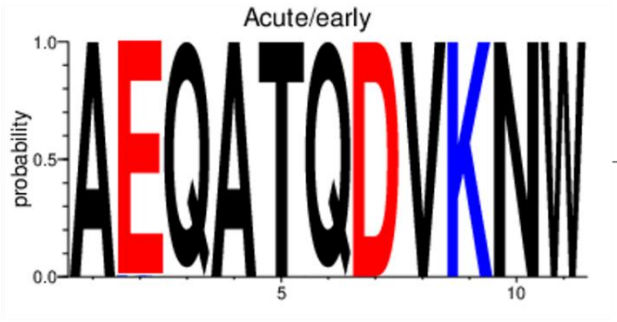
Figure 3.3: CAP217. The kinetics of escape in the HLA-B*15:03 restricted AW11 epitope (306 to 316). **A**, Amino acid logogram at Acute/Early (0 to 26 weeks), Early-chronic (27 to 52 weeks), and Late-chronic infection (> 52 weeks). Colour codes for amino acid charges are as follows: Black is hydrophobic, red is negatively charged, and blue is positively charged. **B**, Frequency table containing amino acid alignment of epitope variants over time showing the location of changes (variant), number of each variant (count), percentage of each variant (pct), and number of mutations. **C**, A stacked plot showing the change in frequency of AW11 epitope variants over six years. WT is the wildtype sequence. **D**, Viral load plot showing the viral load trajectory over the course of infection

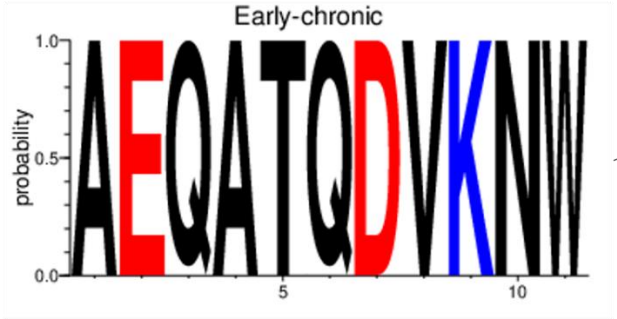
Participant CAP244

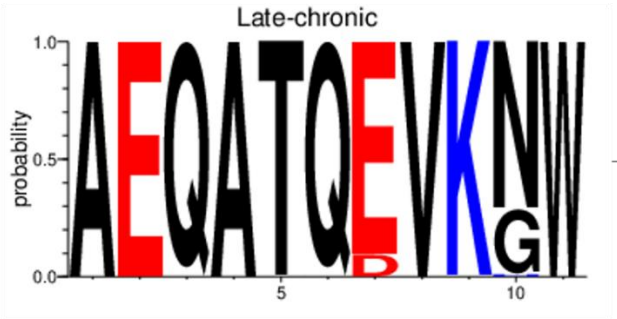
CAP244 has a protective HLA allele (HLA-B*44:03), and an allele associated with rapid disease progression (HLA-B*58:02) (99, 100). Despite having a protective allele, she was a rapid progressor with her CD4 counts reaching 350 cell/µL within six months of infection and remaining below 350 cell/µL for the two-year duration of follow-up. A putative escape mutation was identified in the AW11 epitope (Gag HXB2 coordinates 306 to 316), which is restricted by the HLA-B*44:03 allele. Position 312 encoded aspartic acid (D) in 100 % of sequences sampled in both the acute/early and early chronic phases of infection, and the epitope remained highly conserved (**Figure 3.4 A and B**). However, in the late chronic phase of infection, diversity was observed, with low frequency mutations detected at positions 3, 5, 7 and 10 of the epitope (Q308H (0.14%), T310S (0.14%), T310A (0.14%), K314E (0.14%), N315H (1.29%)). Furthermore, the AW11 wildtype epitope was present at 9.44% frequency in this phase, while the putative escape variant D312E and N315G emerged and reached a peak frequency of 61.37% and 27.61%, respectively.

The D312E mutation in AW11 increased to become the dominant amino acid at this site with 60.34% frequency at 150 wpi and reached 70% by 359 wpi. The N315G mutation fluctuated between 150 wpi at 18.97%, 191 wpi at 40%, 282 wpi at 76.47% and reached 29.36% frequency by 359 wpi (the last sampled timepoint) (**Figure 3.4C**). Viral loads in CAP244 remained constant before and after the appearance of the D312E and N315G mutations with a difference of 1480 copies/mL, suggesting that there was no effect on viral replication associated with the appearance of these mutations (**Figure 3.4D**). The N315G mutation has been reported as nonadapted/depleted in the presence of B*44:03, while the D312E mutation is enriched and reported as a known escape mutation as per the CTL epitope variants and escape mutations database (100, 101).

A







B

Variant	Count	Pct.	No. of mutations
AEQATQEVKNW			
------D----	124	99.2	1
-K----D----	1	0.8	2

Total sequences = 125
Number of variants = 2

Variant	Count	Pct.	No. of mutations
AEQATQEVKNW			
------D----	52	100	1

Total sequences = 52
Number of variants = 1

Variant	Count	Pct.	No. of mutations
AEQATQEVKNW			
----------	428	61.23	0
---------G-	192	27.47	1
------D----	66	9.44	1
---------H-	9	1.29	1
--H--------	1	0.14	1
----S-D----	1	0.14	2
----A----G-	1	0.14	2
--------E--	1	0.14	1

Total sequences = 699
Number of variants = 8

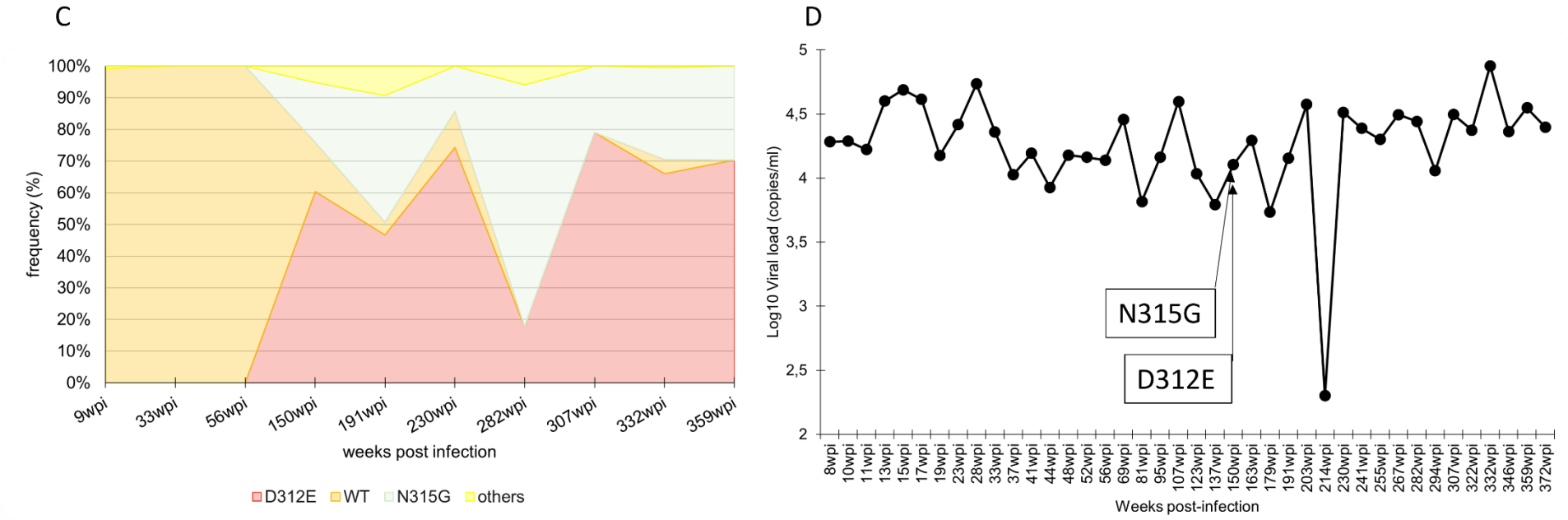


Figure 3.4: CAP244. The kinetics of escape in the HLA-B*44:03 restricted AW11 epitope (HXB2 Gag coordinates 306 to 316). **A,** Amino acid logogram at Acute/Early (0 to 26 weeks), Early-chronic (27 to 52 weeks), and Late-chronic Infection (> 52 weeks). Colour codes for amino acid charges are as follows: Black is hydrophobic, red is negatively charged, and blue is positively charged. **B**, Frequency table containing amino acid alignment of epitope variants over time showing the location of changes (variant), number of each variant (count), percentage of each variant (pct), and number of mutations. The master sequence is taken as the consensus of the full longitudinal alignment, which does not always correspond to the transmitted virus sequence. **C**, A stacked plot showing the change in frequency of AW11 epitope variants over seven years. **D**, Viral load plot showing the viral load trajectory during the 7 years of infection.

Participant CAP222

CAP222 was a slow progressor with viral loads below 500 copies/mL at most timepoints. As a result of these low viral loads, deep sequencing data were available from only two timepoints, both in late chronic infection (122 and 148 wpi). However, 13 *gag p24* Sanger sequences were publicly available from this participant (LANL). These were obtained from eleven different timepoints (6, 13, 26, 52, 56, 70, 83, 95, 108, 122, and 174 wpi), with two sequences obtained from 6- and 13-weeks post-infection each, and one at each of the remaining timepoints (**Figure 3.5.1**). We were therefore able to identify the transmitted virus Gag sequence for this individual using this supplementary sequence data. We identified one epitope undergoing change in this individual, HA9 (Gag HXB2 coordinates 216 to 224), restricted by HLA B*53:01, an allele associated with an increased risk of progression in HIV-1 infection (102). A change from amino acid A to V was observed at position 223 of this epitope. At the earliest sampled timepoints generated by Sanger sequencing, only the A was present at this site. Whereas, in the deep sequenced dataset, there were both amino acids A and V at site 223 of the HA9 epitope. At 122 wpi, 223A was at 18.9% frequency and 223V was at 79.4%, whilst at 148 wpi, 223A was at 5% frequency and 223V was at 95%. At 122 wpi, at 1.35% frequency the H219Q mutation was also identified in the HA9 epitope (**Figure 3.5.2**). In the chronic phase, there was a change from 80% to 95% frequency of the variant A223V (**Figure 3.5.2**).

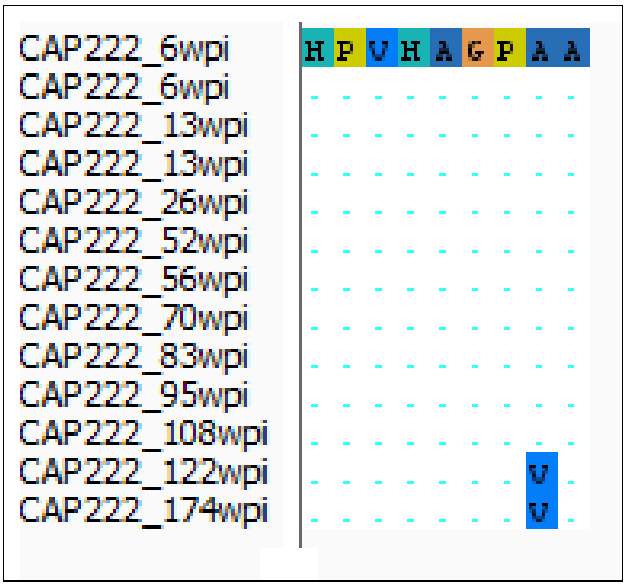


Figure 3.5.1: Identity plot of the HA9 epitope restricted by HLA-B*53:01 from 13 Gag p24 Sanger sequences. Timepoints: 6,13,26, 52, 56, 70, 83, 95, 108, 122, 174 wpi.

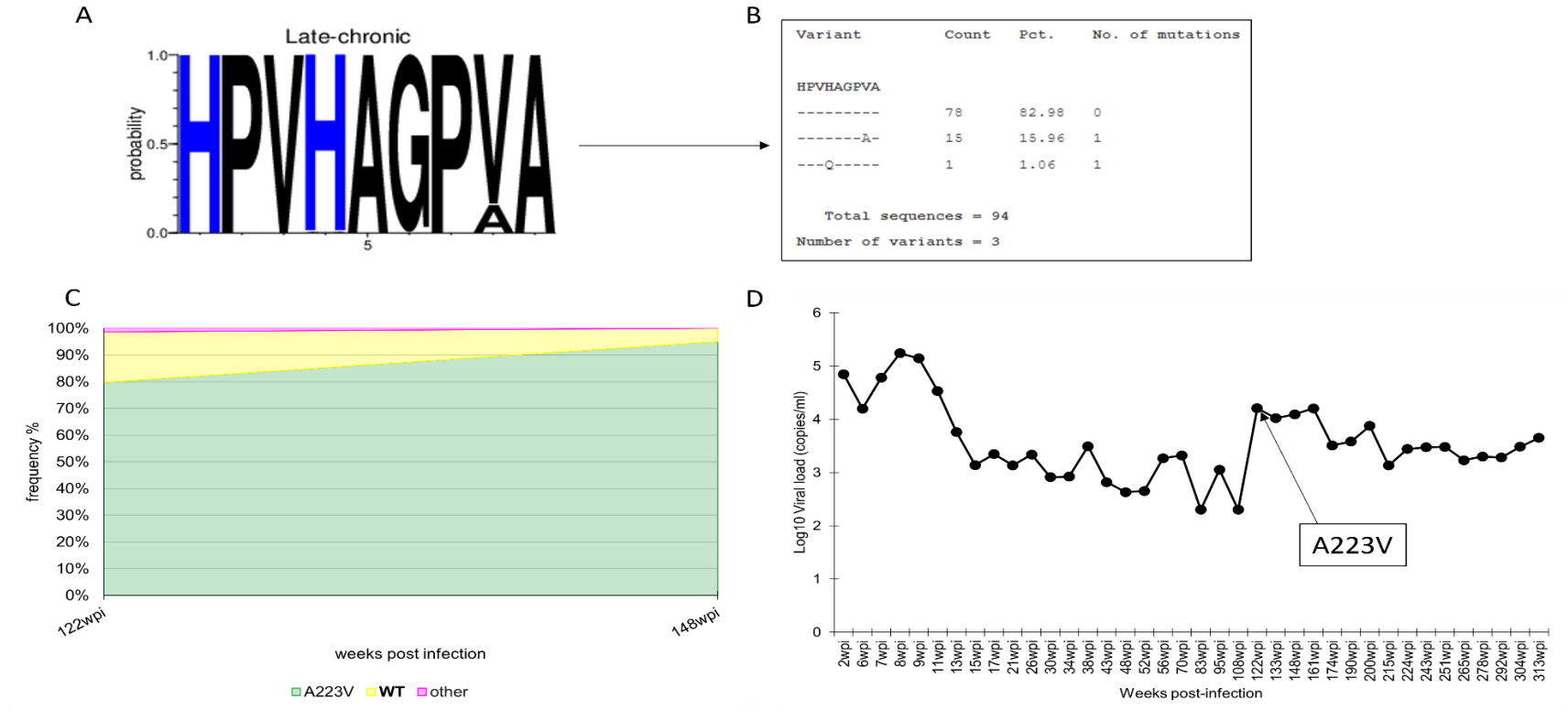


Figure 3.5.2: **CAP222**. The kinetics of escape in the HLA-B*53:01 restricted HA9 epitope (Gag HXB2 coordinates 216 to 224). **A**, Amino acid logogram of sequences at Late-chronic Infection (> 52 weeks). Colour codes for amino acid charges are as follows: Black is hydrophobic, and blue is positively charged. **B**, Frequency table containing amino acid alignment of epitope variants over time showing the location of changes (variant), number of each variant (count), percentage of each variant (pct), and number of mutations. **C,** A stacked plot showing the change in frequency of HA9 epitope variants between 122 and 148 weeks post-infection. **D**, Viral load plot showing the viral load trajectory over the course of infection.

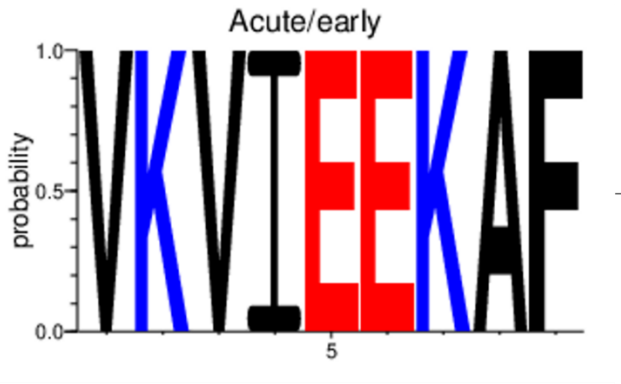
Participant CAP256

CAP256 was a rapid progressor, her CD4 counts over the two years of follow-up remained below 350 cell/µL. This individual was superinfected at 13 wpi with a second strain of HIV (103). CAP256 had five timepoints sampled, one in the acute/early phase (6 wpi) and four in the late chronic phase (at 106, 159, 206, and 429 wpi). Putative escape mutations were identified in the VF9 (Gag HXB2 coordinates 156 to 164), and AW11 (Gag HXB2 coordinates 306 to 316) epitopes.

The VF9 (156 to 164) epitope, is restricted by HLA-B*15:03, an allele associated with control of viral replication. In the late chronic phase, the I159V mutant was observed at a frequency of 67.26%. Low frequency mutations were also observed at this time at positions 4, 6 and 9 of the epitope (E162K (0.88%), I159V&A163T (0.88%)) (**Figure 3.6A, B**). While the only variants detected at 6 wpi had an I at position 159 of Gag, over a period of two years, the frequency of sequences with I at this position decreased until they dropped below the detection limits of our sequencing protocol at 106 wpi. Thereafter, the frequency of sequences with an I at position 159 increased, reaching a frequency of 70% by 206 wpi.

Between 106 wpi and 159 wpi the I159V mutation went from 0% to 85% frequency and thereafter it fluctuated between 85% at 159 wpi, 27% at 206 wpi, and 64% at 429 wpi (**Figure 3.6C**). There was no significant increase or decrease in viral loads associated with the appearance of the I159V mutation. Thus, no loss of control of replication was associated with this escape mutation (**Figure 3.6D**).

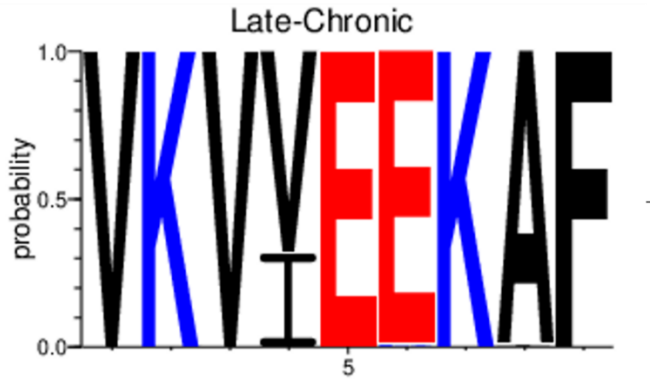
A



B

Variant	Count	Pct.	No. of mutations
VKVIEEKAF			
---------	766	100	0

Total sequences = 766
Number of variants = 1



Variant	Count	Pct.	No. of mutations
VKVIEEKAF			
---V-----	76	67.26	1
---------	35	30.97	0
-----K---	1	0.88	1
---V---T-	1	0.88	2

Total sequences = 113
Number of variants = 4

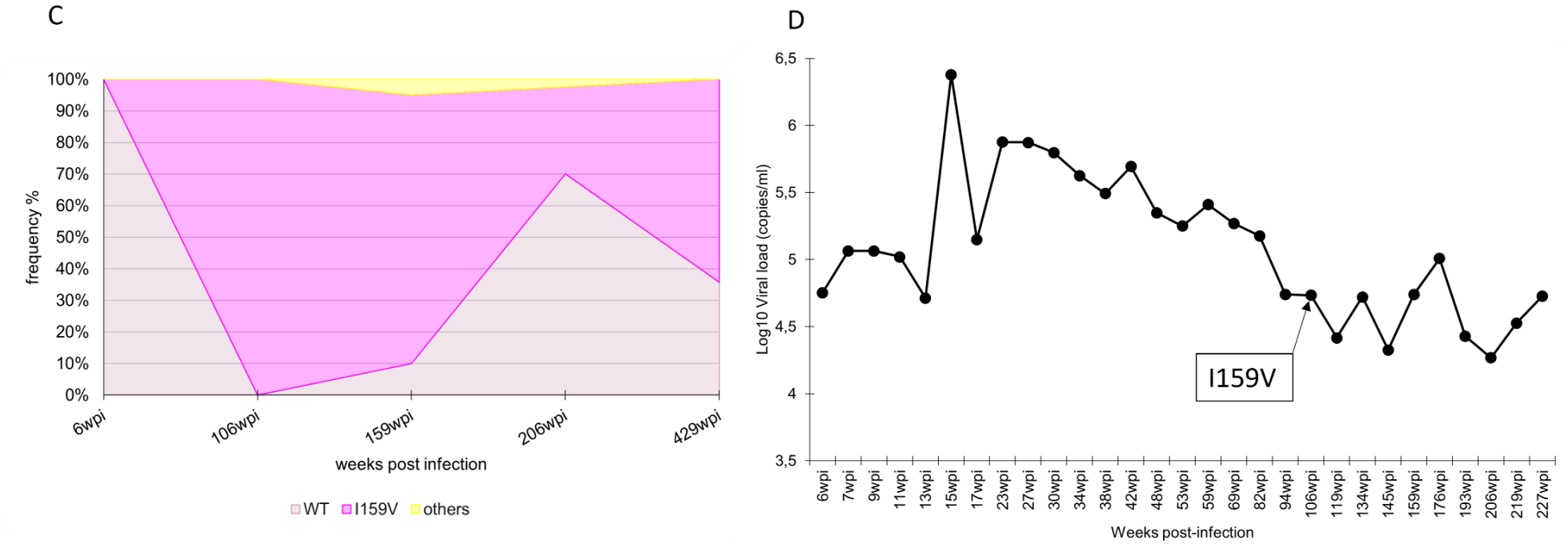
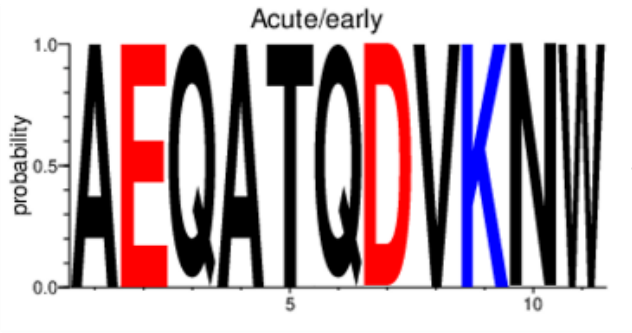


Figure 3.6: CAP256. The kinetics of escape in the HLA-B*15:03 restricted VF9 epitope (HXB2 Gag coordinates 156 to 164). **A**, Amino acid logogram at Acute/early (0 to 26 weeks) and Late-chronic Infection (> 52 weeks). Colour codes for amino acid charges are as follows: Black is hydrophobic, red is negatively charged, and blue is positively charged. **B**, Frequency table containing amino acid alignment of epitope variants over time showing the location of changes (variant), number of each variant (count), percentage of each variant (pct), and number of mutations. **C**, A stacked plot showing the change in frequency of VF9 epitope variants over eight years. **D**, Viral load plot showing the viral load trajectory over eight years of infection.

Also in this participant, 99.40% of sequences carried the WT AW11 epitope (Gag HXB2 coordinates 306 to 316) – which is restricted by HLA-B*15:03 - at 6 wpi (during the acute/early phase of infection). Low frequency mutations at this timepoint were observed at positions 3, 7, 8 and 10 of the epitope (Q308H (0.13%), Q311R (0.13%), D312N (0.13%), and N315K (0.13%)). By the late chronic phase, A309S and D312E mutations were present in 100% of the sequences sampled at 106 wpi and remained at 100% frequency until 429 wpi (**Figure 3.7 A, B, C**). We found a small increase in viral loads before and after the emergence of mutants A309S and D312E, suggesting no loss of control of replication was associated with these mutations (**Figure 3.7D**).

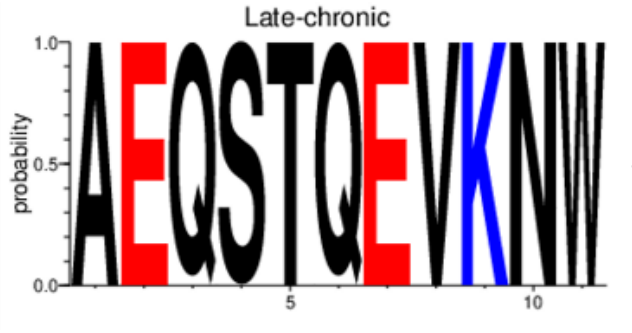
A



B

Variant	Count	Pct.	No. of mutations
AEQATQDVKNW			
----------	760	99.48	0
------N----	1	0.13	1
-----R-----	1	0.13	1
---------K-	1	0.13	1
--H--------	1	0.13	1

Total sequences = 764
Number of variants = 5



Variant	Count	Pct.	No. of mutations
AEQATQDVKNW			
---S--E----	112	100	2

Total sequences = 112
Number of variants = 1

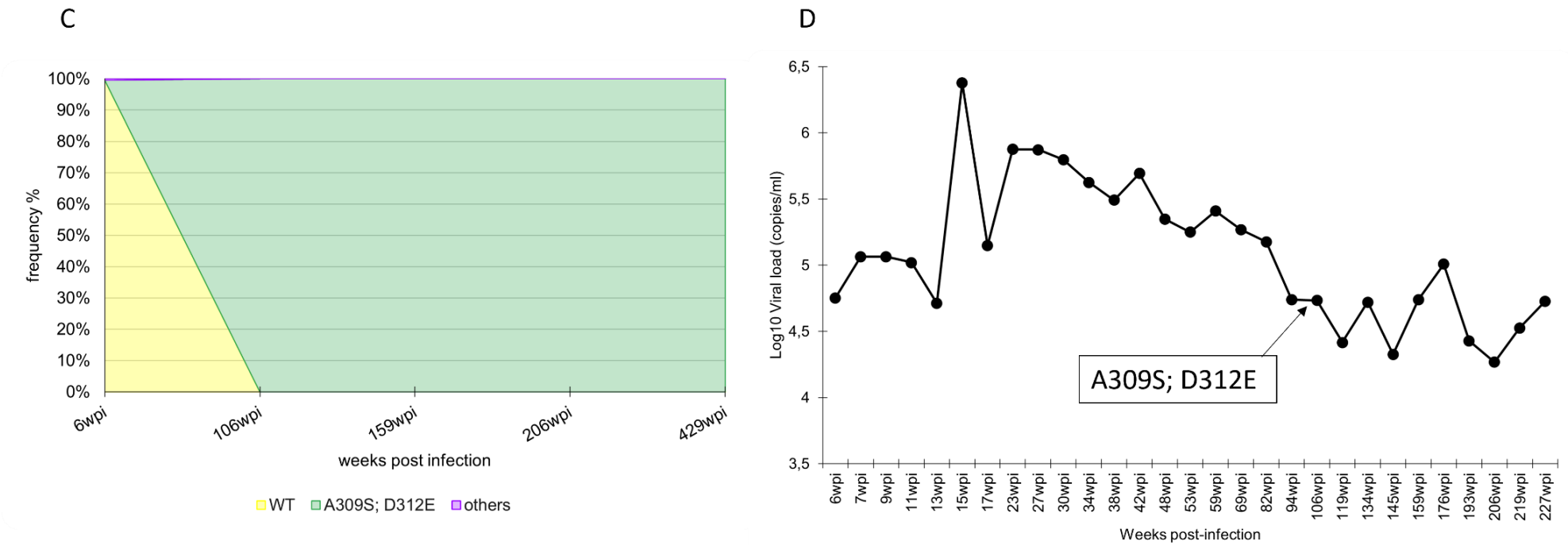
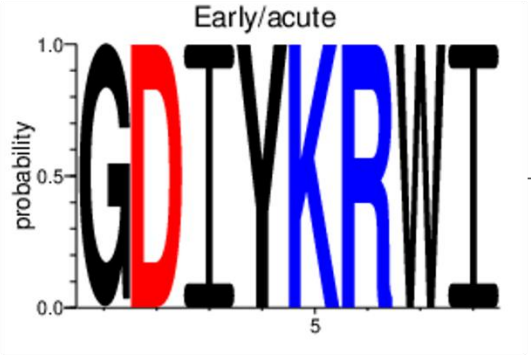


Figure 3.7: CAP256. The kinetics of escape in the HLA-B*15:03 restricted AW11 epitope (306 to 316). **A**, Amino acid logogram at Acute/early (0 to 26 weeks) and Late-chronic Infection (> 52 weeks). Colour codes for amino acid charges are as follows: Black is hydrophobic, red is negatively charged, and blue is positively charged. **B**, Frequency table containing amino acid alignment of epitope variants over time showing the location of changes (variant), number of each variant (count), percentage of each variant (pct), and number of mutations. **C**, a Stacked plot showing the change in frequency of AW11 epitope variants over eight years. **D**, a viral load plot showing the viral load trajectory over the eight years of disease progression**.**

Participant CAP336

CAP336 was a rapid progressor, her CD4 counts over two years of follow-up remained below 350 cell/µL. The GI8 wildtype epitope (Gag HXB2 coordinates 259 to 266), which is restricted by HLA-B*08:01, was present at a frequency >= 97.62% at all three phases of infection, except for at 126 wpi when the R264K mutation was observed at a frequency of 6% (**Figure 3.8A, B, C**) (98). There was no significant increase in viral load (**Figure 3.8D**) associated specifically with the occurrence of the R264K mutation in the GI8 epitope. This mutation was previously reported as a known escape mutation as per the CTL epitope variants and escape mutations database (101).

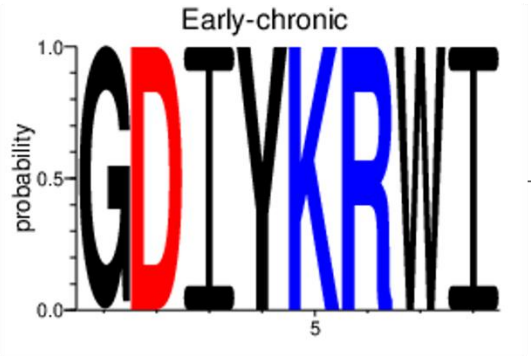
A



B

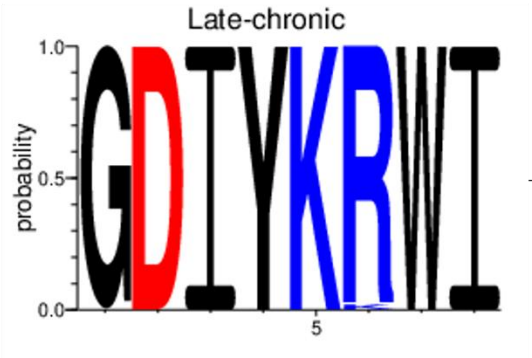
Variant	Count	Pct.	No. of mutations
GDIYKRWI			
--------	20	100	0

Total sequences = 20
Number of variants = 1



Variant	Count	Pct.	No. of mutations
GDIYKRWI			
--------	13	100	0

Total sequences = 13
Number of variants = 1



Variant	Count	Pct.	No. of mutations
GDIYKRWI			
--------	41	97.62	0
-----K--	1	2.38	1

Total sequences = 42
Number of variants = 2

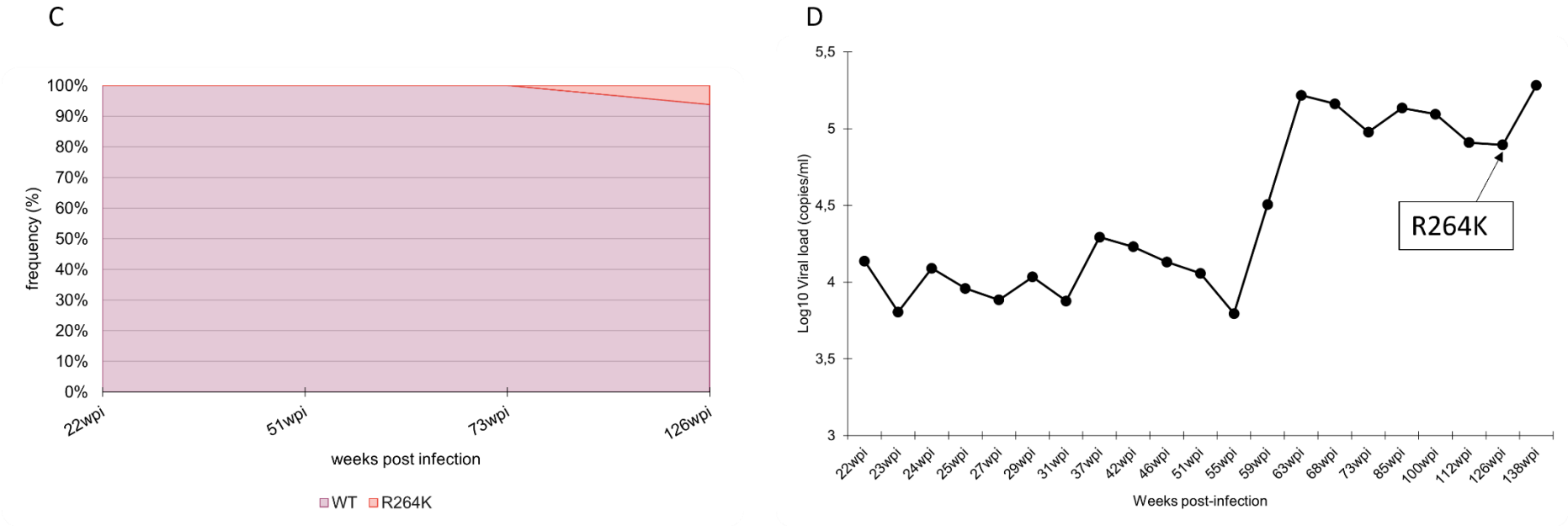
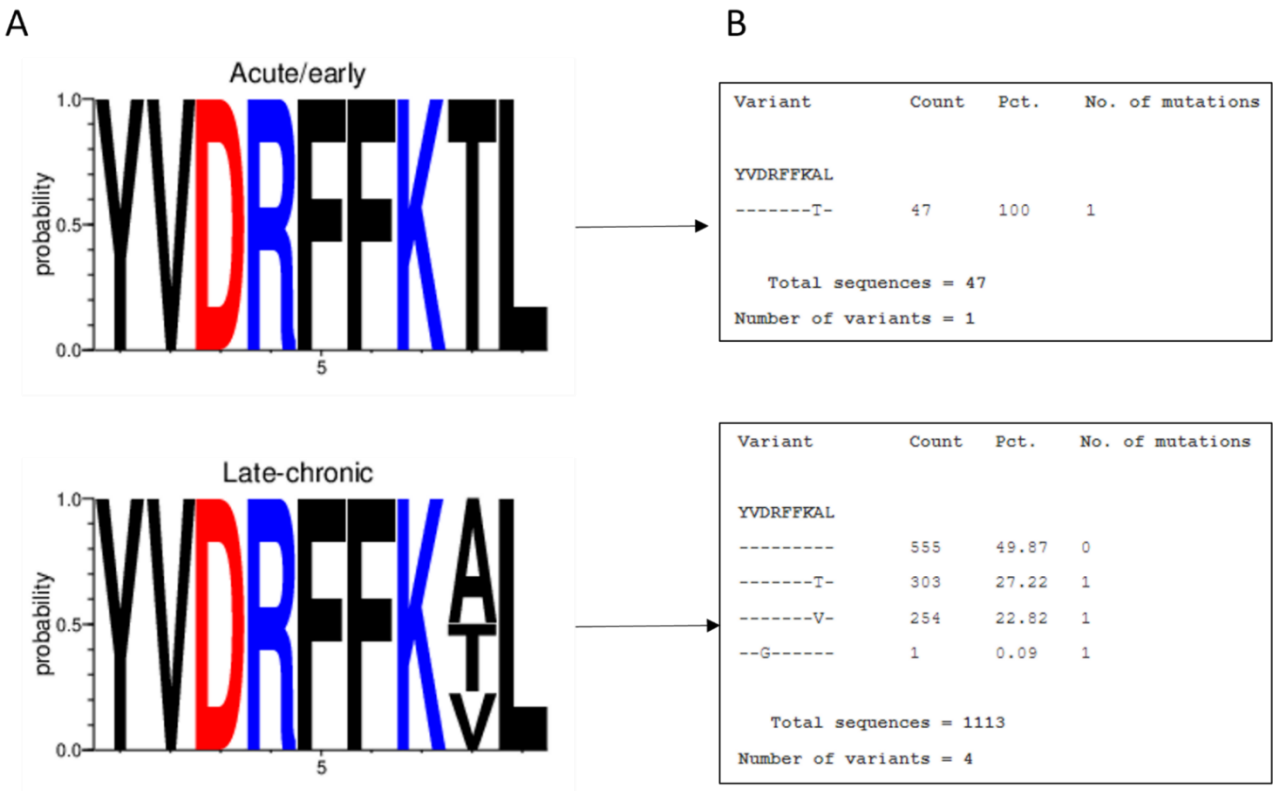


Figure 3.8: CAP336. The kinetics of escape in the HLA-B*08:01 restricted GI8 epitope (259 to 266). **A**, Amino acid logogram at Acute/early (0 to 26 weeks), Early-chronic (27 to 52 weeks), and Late-chronic Infection (> 52 weeks). Colour codes for amino acid charges are as follows: Black is hydrophobic, red is negatively charged, and blue is positively charged. **B**, Frequency table containing amino acid alignment of epitope variants over time showing the location of changes (variant), number of each variant (count), percentage of each variant (pct), and number of mutations. **C**, a Stacked plot showing the change in frequency of GI8 epitope variants over two years. **D**, A viral load plot showing the viral load trajectory over two years.

Participant CAP372

CAP372 was an intermediate progressor, maintaining her viral loads over the three years of follow-up and taking three years for CD4 counts to decline to 350 cell/µL. She has the HLA-B*15:01 allele, which is associated with better control of HIV (104). The wild-type YL9 (Gag HXB2 coordinates 296 to 304) epitope, which is restricted by HLA B*15:10 carried by this individual, was at 100 % frequency at 3 wpi, the earliest sampled timepoint. The T303A mutant was detectable at a frequency of 14% at 106 wpi, rising to a peak of 86% at 161 wpi. At 106 wpi, the T303V mutation was observed at 36% frequency, fluctuating between 10% frequency at 161 wpi and 22 % at 182 wpi. Low frequency mutation D298G was observed at position 3 of the epitope at 0.09% at 161 wpi (**Figure 3.9 A, B, C**). The T303V/A mutations appeared at 106 wpi where viral loads increased by 47344 copies/mL (**Figure 3.9D**). The T303V mutation was reported by Carlson et al., 2012 to be depleted in the presence of C*03 (this participant is restricted by C*03:04).



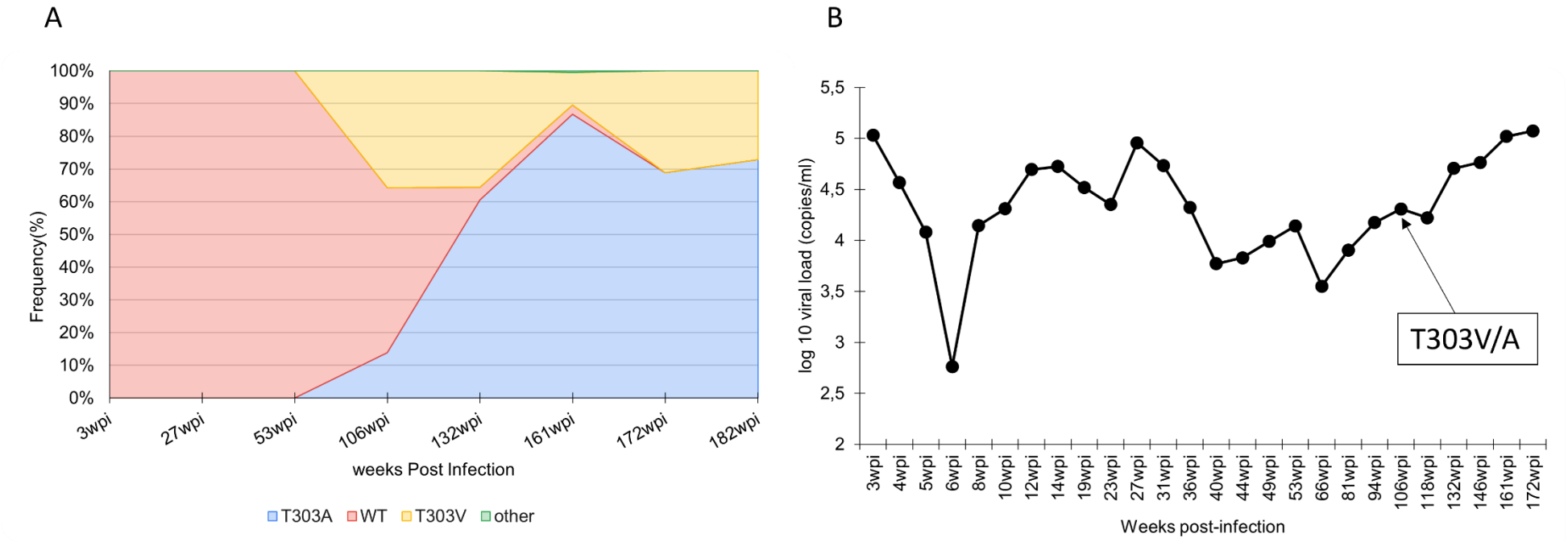


Figure 3.9: CAP372. The kinetics of escape in the HLA-B*15:10 restricted YL9 epitope (296 to 304). **A**, Amino acid logogram at acute/early (0 - 26 weeks), and late chronic Infection (> 52 weeks). Colour codes for amino acid charges. Black is hydrophobic, red is negatively charged; blue is positively charged. **B**, Frequency table containing amino acid alignment of epitope variants over time showing the location of changes (variant), number of each variant (count), percentage of each variant (pct), and number of mutations. The master sequence is taken as the consensus of the full longitudinal alignment, which does not always correspond to the transmitted virus sequence **C**, a Stacked plot showing the change in frequency of YL9 epitope variants over three years. **D**, Viral load plot showing the viral load trajectory over three years.

3.5 Identification of sites under positive selection and high entropy sites

As Epitope Matcher only identified known epitopes reported in the CTL epitope database (https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html), escape in lesser known or novel CTL epitopes would have been missed. Positive selection analysis, identifies sites, like immune escape mutations, that confer an advantage to the virus. Here we used positive selection analysis to support predicted CTL escape, but also to identify new sites undergoing immune selection that may have been missed by Epitope Matcher. The Hyphy-FUBAR analysis tool was used to identify dN/dS ratios of >1 in p24 gag sequences for all participants (91). We also used Entropy tool (https://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_one.html), to identify sites undergoing change over the course of infection that increased in frequency or diversity. High entropy was taken as a value greater or equal to 0.25 (s-unit) over the full sampled infection period (105) (**Table 3.4**). These sites under positive section and/or high entropy site, were inspected for either corresponding to epitopes within the LANL database CTL epitope variants and escape mutations list (101), and/or to HLA adapted sites identified by Carlson et al., 2012 (**Table 3.4**).

A total of 35 amino acid sites in Gag across the 15 participants were identified as high entropy over the infection period (**Table 3.4**), of which nine (out of 35) mutations/sites were within known epitopes: T310S(AW11), V223A(HA9), D312E(AW11), N315G(AW11), V159I(VF9), A306S(AW11), D309E(AW11), E316D(AW11), and T303V/A(YL9). Furthermore, there were six high entropy sites that corresponded to HLA associated sites reported by Carlson et al., 2012 as either “adapted” or “nonadapted” forms, where adapted mutations were defined as those that are significantly enriched in the presence of the HLA allele in question (referred to as either escape mutants or resistant forms) while nonadapted mutations are those that are significantly depleted in the presence of the HLA allele in question (susceptible, wild-type and/or reversion forms) (100). This included the high entropy site where mutation G226N was identified in both participants CAP244 and CAP256, which has previously been reported as nonadapted, that is amino acid ‘N’ was depleted in the presence of C*04. CAP244 further harboured the D312E mutation where, the ‘E’ amino acid was reported as adapted in the presence of B*44. In addition, participant CAP372

harboured the K335R mutation at a high entropy site, where the amino acid ‘R’ was reported to be adapted in the presence of C*03:04. Additionally, CAP372 harboured the T303V/A mutation at a high entropy site, which is found in the known epitope YL9 and the ‘A’ amino acid at this site was reported as non-adapted with the presence of the C*03 allele. CAP217 harboured the T310S mutation where the amino acid ‘S’ was reported as adapted in the presence of B*58:01.

In terms of the positive selection analyses, positive selection was identified in 10 participants, with one site evolving under positive selection in each participant (**Table 3.4**). Only one (position 310 seen in CAP217) of the 10 sites fell within epitopes identified by Epitope Matcher and another one site (position 335 seen in CAP372) was previously reported by Carlson et al., 2012 as a HLA adapted site. In total, nine sites that were detectably evolving under positive selection were also high entropy sites. The remaining 23 high entropy sites, along with eight sites evolving under positive selection (all of which overlapped), were not found either in the CTL epitope variants and escape mutation list (https://www.hiv.lanl.gov/content/immunology/variants/ctl_variant.html) or HLA-adapted sites (100).

The AW11 epitope (Gag HXB2 coordinates 306 to 316) undergoing change in CAP217 and restricted by alleles B*15:03 and B*58:01 harboured a T310S mutation. As described in 3.4.1 (**Figure 3.3**), this mutation had reached a frequency of 80% by 358 wpi in this individual (the end of the pre-ART infection phase). In these additional analyses, position 310 was further identified as a high entropy site and to be evolving under positive selection. T310S has also been proven to be an escape mutation by Carlson et al., 2012 (100), and described as adapted to HLA allele B*58:01, which means, the amino acid ‘S’ is significantly enriched in the presence of B*58:01. This data provided supporting evidence that the T310S mutation in the AW11 CTL epitope was likely to be associated with escape.

Table 3.4: High Shannon entropy sites and sites evolving under positive selection.

CAPID	HLA profile	Mutation* (dN/dS>1 ratio)	Entropy score	Mutation in known epitopes? Y/N	Mutation associated with escape/adaptation to HLA (source/reference) +
188	A*02:02, A*74:01 B*15:03, B*15:16 C*02:10, C*14:02	P225A/S* (0.995)	0.333	No	None
206	A*03:01, A*32:01 B*07:02, B*44:03 C*02:10, C*07:02	E207D	0.693	No	None
		D230E	0.327	No	None
217	A*02:02, A*29:01 B*15:03, B*58:01 C*02:10, C*06:02	T310S* (0.996)	0.697	Yes, AW11 (HXB2 306 to 316)	EpitopeMatcher (B*58:01), S Adapted by B*58:01, Carlson2012 (100)
222	A*30:01, A*33:03 B*53:01, B*81:01 C*04:01, C*04:01	V223A	0.650	Yes, HA9 (HXB2 216 to 224)	EpitopeMatcher (B53:01)
244	A*23:01, A*30:04 B*44:03, B*58:02 C*04:01, C*06:02	H219Q* (0.998)	0.382	No	None
		G226N	0.452	No	N NonAdapted by C*04
		D312E	0.485	Yes, AW11 (HXB2 306 to 316)	EpitopeMatcher (B*44:03), E Adapted by B*44, Geels2015 (103)
		N315G	0.642	Yes, AW11 (HXB2 306 to 316)	EpitopeMatcher (B44:03)

256	A*29:01, A*66:01 B*15:03, B*58:02 C*04:01, C*06:02	V159I	0.692	Yes, VF9 (HXB2 156 to 164)	EpitopeMatcher (B*15:03), Frahm2006 (98)
		G226N	0.637	No	N NonAdapted by C*04
		I228M	0.631	No	None
		T215L	0.675	No	None
		N249S	0.637	No	None
		R283K	0.631	No	None
		A306S	0.637	Yes, AW11 (HXB2 306 to 316)	EpitopeMatcher (B*15:03)
		D309E	0.669	Yes, AW11 (HXB2 306 to 316)	EpitopeMatcher (B*15:03)
		E316D	0.637	Yes, AW11 (HXB2 306 to 316)	EpitopeMatcher (B*15:03)
257	A*23:01, A*29:02 B*42:02, B*44:03 C*17:01, C*17:01	A223V/I* (0.967)	0.972	No	None
		E245D	0.650	No	None
		D255E	0.542	No	None
280	A*29:02, A*74:01 B*15:03, B*15:10 C*02:10, C*08:04	V215T	0.527	No	None
		K286R	0.634	No	None
		A332T	0.445	No	None
286	A*03:01, A*29:02 B*15:10, B*49:01 C*16:01, C*16:01	S242T* (0.923)	0.436	No	None

287	A*24:02, A*43:01 B*15:10, B*42:01 C*04:01, C*17:01	V256I	0.276	No	None
288	A*24:02, A*43:01 B*07:02, B*15:01 C*04:01, C*07:02	E230D* (0.904)	0.579	No	None
336	A*23:01, A*68:01 B*08:01, B*58:02 C*06:02, C*07:01	I223N	0.669	No	None
		A224P	0.494	No	None
		N242T	0.582	No	None
		S281N* (0.980)	0.669	No	None
337	A*23:01, A*30:02 B*15:10, B*57:03 C*16:01, C*18:01	N253T* (0.952)	0.638	No	None
372	A*03:01, A*34:02 B*15:10, B*45:01 C*03:04, C*06:02	T303V/A	0.975	Yes, YL9 (HXB2 296 to 304)	EpitopeMatcher (B*15:10), A Adapted by C*03, Honeyborne2010 (106), Holzermer2015 (107), Leitman2017 (108), VanTeijlingen2014 (109)
		T332S	0.705	No	None
		R335K* (0.962)	0.617	No	K Adapted by C*03:04

*High Shannon entropy sites (ratio of sites evolving under positive selection, dN/dS >1, in brackets)

+Adapted- amino acid enriched by HLA, non-adapted- amino acid depleted by HLA (100)

3.6 Population level analysis of CTL escape epitopes and mutations

The most abundant HLA-B allele in our participants was B*15:03 (found in 4 out of 15 women), a protective allele, associated with control of viral replication. Of the 73 known epitopes identified by Epitope Matcher in this study, the most commonly HLA-B targeted epitope identified was YL9, found in 9 out of 15 participants (60%) and restricted by the B*15:03, B*15:10 and B*42:01 alleles. The AW11 (Gag HXB2 coordinates 306 to 316) epitope was the second most targeted epitope and was found in 7 out of the 15 participants (47%). This epitope acquired changes at multiple amino acid positions with high frequencies, specifically at positions 312, 314 and 315. Individuals with alleles B*15:03, B*44:02 or B*44:03 commonly target the AW11 epitope. Other commonly targeted epitopes included TL9, VF9, and GL9 identified in five of the participants each, and GY9, identified in four of the participants. Other remaining epitopes, identified by Epitope Matcher, in this population were less common or unique to individuals and were restricted by a diverse set of HLA-B alleles, while other common HLA-B alleles amongst participants, in no order of commonality, included B*07:02, B*53:01, B*15:01 and B*15:10.

On average, Epitope Matcher identified five epitopes per participant, with a range of two to eleven. At the last point of follow-up, 62.5% of the putative escape mutations were found in at least 50% of the sampled sequences. Overall in the seven participants where escape was quantified, most escape (8/9) mutations were observed in the late chronic phase of infection (> 52 weeks). On average, putative escape mutations took 52 weeks to reach 50% and/or 113 weeks to reach 80% and mutations arose at any position within an epitope. Viral loads were not obviously impacted by the presence or absence of the potential escape mutations.

In addition, 9/9 CTL escape mutations identified by Epitope Matcher were supported by entropy analyses and 1/9 supported by positive selection analysis. While HLA-C associated CTL escape mutations were identified by these additional analyses, there appear to be no HLA-B associated epitopes/escape mutations missed by the Epitope Matcher analyses.

Chapter 4: Discussion

HIV rapidly escapes immune responses exerted by the host including escape from CTLs. These escape mutations arise over the course of infection and may be more frequent in certain regions of the viral genome than others. Deep sequencing approaches enable the study of changes in viral quasispecies and provide a means to understand the kinetics and pathway(s) of escape. The aim of the study was to evaluate HIV-1 CTL escape kinetics in longitudinal NGS datasets of *gag* generated using the Illumina MiSeq platform. In this thesis we have identified tools for identifying CTL escape in deep sequencing datasets and developed a workflow to facilitate this. Lastly, we applied these tools to describe CTL escape dynamics in a longitudinal deep sequencing dataset from 15 women in the CAPRISA 002 cohort.

We defined a list of criteria for screening HIV sequences for CTL escape. These criteria included the functionality of the tool, it's processivity, accessibility, reliability, source of database and output provided. Four tools were assessed with only two meeting the criteria. Only one of these tools, Epitope Matcher, was ultimately selected based on a validation test using a test dataset from a previous study that identified epitopes with evidence of CTL escape (69). This tool identified epitopes targeted by the participants HLA and identified mismatches to a reference of experimentally verified epitopes. Other tools/methods were excluded due to limitations on sequence length, lack of use of a CTL epitope database as a reference/training set for predicting CTL epitopes and escape mutations, or a requirement that CTL epitopes be manually mapped.

In a p24 *gag* NGS dataset of 4583 total sequences from 15 HIV-1 subtype C infected women from the CAPRISA 002 cohort, KwaZulu Natal, we identified 73 HLA-B targeted CTL epitopes using the Epitope Matcher tool. Of the 73 epitopes, seven acquired putative escape mutations that increased in frequency over time (frequency of > 5% at a timepoint). In total, there were nine mutations identified within the seven epitopes. Five of nine of these mutations corresponded to known, reported CTL escape mutations in the CTL epitope variants and escape mutations database (101) and/or by Carlson et al., 2012 (100). All nine mutations corresponded to high entropy

sites, however only one of these mutations (T310S) was identified as a site evolving under positive selection. An additional 23 amino acid sites within Gag (not in epitopes detected by Epitope Matcher) were identified as high entropy and/or evolving under positive selection, of which two were associated with HLA-C alleles expressed by the corresponding participants. These 23 sites may be potential novel HLA-targeted sites or escape from other immune responses or mutations impacting viral fitness. The trajectory of frequency of the putative escape mutations was different in each epitope, where in some epitopes (55%), the frequency either constantly or decreased overtime while in others (44%), the frequency fluctuated between frequencies (low and/or high).

Historically, the identification of CTL epitopes and escape mutations typically required the use of different tools/resources in combination, at times including manual inspection of multiple sequence alignments. In this thesis we developed a streamlined pipeline to perform all steps required for predicting escape. Using this approach, we could take a multiple sequence alignment from an individual and identify mutations associated with CTL pressure. We assessed tools reported in the literature and/or available online that were previously used in some way in the prediction of CTL escape.

CTL escape has been detected within the first weeks of HIV infection using standard Sanger sequencing methods (69, 110). Furthermore, clusters of amino acid polymorphisms may occur within epitopes (9 to 14 amino acid stretches). Such clusters of polymorphisms have been reported to include those toggling back-and-forth between two residues or between different amino acid sites over time, in some cases eventually becoming either fixed or dominant within the population, potentially to optimise the cost of virus fitness (70). Additionally, the rate of escape differs between epitopes and HLAs, with some epitopes escaping rapidly in acute infection while others take long to escape, typically in the chronic phase of infection (63, 65, 111, 112). The use of deep sequencing to generate datasets provides an advantage for identifying the kinetics of CTL escape due to its sensitivity providing accuracy and an improved way of detecting low frequency mutations (113). In this study, putative escape was evaluated longitudinally up to 380 weeks post-infection in seven (out of 15) CAPRISA 002 women. Over this time, most of the escape mutations (8/9) were observed in the late chronic phase of infection (> 52 weeks). Escape in these

mutations was slow taking 52 weeks to reach 50% frequency and/or 113 weeks to reach 80% frequency, such as the T310S (AW11) and T303A (YL9) mutations. The mutations arose at different positions within the 9 to 14-mer epitope stretches. Previously, studies have suggested that escape mutations preferably occur at HLA-binding sites (anchor residues) of CTL epitopes (79). In our results, all seven epitopes did not have mutations at HLA-binding sites. Sampling was done every six months, thus there is a probability that the actual time a mutation first arose may have been missed due to the big sampling window and this may have affected the ability to completely characterize the kinetics of escape.

Protective HLA alleles have previously been reported to better control HIV-1. The correlation between a host's HLA alleles and its ability to control HIV-1 is mediated by the ability of the alleles to target specific viral epitopes (114). In this study, 5/9 of mutations were found in epitopes restricted by participants with protective alleles, however, we did not observe any obvious evidence of the impact of protective alleles on the outcome of clinical progression overtime. Additionally, only 3/9 of the putative escape mutations were associated with temporary increases in viral replication.

A study by Liu et at., 2013 previously reported that CTL escape occurs more rapidly in high entropy epitopes. These high entropy sites in epitopes are more variable and thus can escape faster compared to conserved (low entropy) sites which mutate slowly due to functional reasons (65). In this study, there were nine out of 35 high entropy sites that were found in epitopes known to bind to the participant's HLA (as identified by Epitope Matcher). From these epitopes, the YL9 and HA9 epitopes, which have previously been reported as immunodominant epitopes, were identified. Within YL9 and HA9 the T303V/A and V223A mutations were selected for, respectively (115, 116). Our current findings support the immunodominance of the YL9 and HA9 epitopes.

For future work, Epitope Matcher could be improved by (I) incorporating a Graphical User Interface (GUI) utilizing the R Shiny app or a similar platform, this would facilitate user interaction through visual icons and controls. (II) Automatically searching for and using the latest versions of LANL CTL epitope databases, (III) broadening the focus of the tool, in addition to searching for CTL epitopes, to also include other motifs such as N-linked glycosylation sites associated with neutralizing antibody targeting in

Envelope. Additional work can be done to automate the entire workflow of identifying CTL epitopes and escape mutations. The potentially novel sites with high entropy that were not identified or found in known epitopes would need to be validated using IFN-γ ELISPOT assays. To broaden our understanding in characterizing the kinetics of CTL escape in HIV, the analysis performed in this study can be applied to the entire genome. Furthermore, the Gag p24 region analyzed here was amplified using an Illumina MiSeq method that generates non-overlapping forward and reverse sequence reads due to an amplicon size greater than the limits of the technology (>600bp). Therefore, the region is missing amino acids where the reads should overlap. This region spans the well-known TW10 epitope targeted by individuals with the B*58:01 allele, therefore escape in this immunodominant site may have been missed resulting in under-estimation of putative escape mutation frequency.

The characterization of CTL escape kinetics in next-generation sequencing datasets has been described in a limited number of studies (**Table 5.1** Appendix). A challenge remains the lack of automated workflows or tools needed to confidently detect CTL epitopes and putative escape mutations. Given that NGS is now commonly used to study HIV evolution, these tools will need to handle the volumes of data provided by sequencing. In this study, we show that Epitope Matcher is a promising tool when used in combination with supporting resources into a pipeline to identify and characterize dynamics of CTL epitopes and putative escape mutations.

Chapter 5: Appendices

Table 5.1: Summary table for studies that screened for CTL escape mutations.

Paper	**Participants #**	**Cohort description (period/freq of sequencing)**	**Method of identifying CTL escape from sequence data**	**Gene regions sequenced**	**Proven or predicted?**
Carlson et al.,2012	2126 treatment naïve	Durban (n=1218); Bloemfontein (261); Kimberly (n=31); Gaborone, Botswana (n=514); SA subjects in UK clinics (n=102)	HLA-associated polymorphisms	gag, pol, nef	Proven. IFN-γ ELISPOT
Chopera et al., 2017	78 treatment naïve	Durban	HLA associated polymorphisms	gag	Predicted
Radebe et al.,2015	18 treatment naïve	Blood samples collected at weeks 2,4,6,8,12,18,26 and 52 post-infection.	Pearson's and Spearman's correlation test (sanger sequence)	gag	Proven; Cultured IFNγ ELISPOT assay
Carlson et al., 2008	1089	HOMER cohort: British Columbia, Canada(n=567); Clade B gag seqs. Durban cohort SA (n=522); Clade C p17/p24 gag seqs	Phylogenetic Dependency networks	gag	Predicted

Brumme et al., 2008	98 untreated HIV subtype B	Berlin, Germany (n=38); Massachusetts General Hospital, Boston (n =25); Aaron Diamond AIDS Research Center, New York, NY (n= 24); National Centre in HIV Epidemiology and Clinical Research, University of New South Wales, Sydney, Australia (n=11). In general, patients were monitored at baseline, 1 month, and every 2 or 3 months thereafter,	HLA associated polymorphisms	gag pol nef	Proven IFNγ ELISPOT
Goulder et al., 2015	22 treatment naïve	Durban, SA; Blood collection: at enrollment, 2 weeks, and 2,3 and 6 months post infection and then every 6 months thereafter.	SNAP and Highlighter Tool; Manually; https://www.hiv.lanl.gov/content/immunology/variants/ctl_variant.html	gag	Predicted
Tumiotto 2017 et al.,	47 art treated (successfully) for at least 6 months	Bordeaux Uni Hospital ,France(n=35) Primary infection cohort, Montreal, Canada(n=6) IMPACTA, Lima, Peru (n=6)	HLA-associated polymorphisms; Tutu Genetics	gag pol	Predicted
Rousseau et al., 2008	375	-sampled prior to ART;	Phylogenetic analysis	gag, env, nef, pol	Proven IFNγ ELISPOT
Boutwell et al., 2013	(gag ctl escape mutations)	pNL4-3 HIV-1 subtype B	correction for multiple comparisons	gag p17, p24 pol	proven
Abrahams et al., 2013	5	CAPRISA 002 Acute infection cohort, Durban SA	-SGA -ELF -NetHMCpan -Positive selection analyses	Full genome	Proven; IFNγ ELISPOT assays

Kløvepris et al., 2013	C clade Cohort n=1010 (143 Expressed HLA-A *62:02)	Durban and UK cohort	HLA associated polymorphisms	gag pol vpr	proven; Elispot assays
Goonetilleke et al., 2009	3	CHAVI 001 cohort.	SGA; positive selection analyses	Full genome	Proven; IFNγ ELISPOT assays.
Treurnicht et al., 2010	14 treatment naive	CAPRISA002 cohort; been infected for a median of 39 Days (range 22 to 62 days) at enrolment	CTL epitope database (https://www.hiv.lanl.gov/content/immunology/variants/ctl_variant.html); NetMHCpan tool; ELF Motif Scan	env	Predicted
Wood et al., 2009	102 subtype B	81 - single virion infection 21 - multiply infected	Evolution model to identify diversifying selection in experimentally confirmed CTL epitopes. (Los Alamos HIV Immunology Database, ELF)	env	Proven; Ex-vivo IFNγ ELISPOT

5.2A Python code written by Dr Anna Yssel for adding HXB2 reference sequence (position 1255 to 1798) to p24 sequences:

```
#!/bin/bash
#set -euo pipefail
TASK="add reference"
WD=$(pwd)
NOW=$(date +"%Y%m%d%H%M")
mkdir -p ${TASK}_runinfo/{log,tmp}
declare -a array=( $(ls *.fasta) )
arraylength=${#array[@]}
for ((i=1; i<${arraylength}+1; i++));
do
echo  ${array[$i-1]}
x=${array[$i-1]}
y=${x%.fasta}_HXB2.fasta
cat 1255to1798.fasta ${x} > ${y}
done
```

5.2B **MACSE program**

MACSE was accessed via the virtual machine “srvubugal004.uct.ac.za” and the following code was run to codon align the participant’s virus sequence with the HxB2 reference:

java -jar -Xmx600m /opt/software/macse/macse_v1.2.jar -prog alignSequences -seq *HXB2.fasta file path -out_NT output file path/macse_out_NT.fasta -out_AA output file path/macse_out_AA.fasta

5.2C **Epitope matcher**

Steps for installation and running of the epitope matcher tool can be found on Github, link: https://github.com/philliplab/EpitopeMatcher.git

5.2D **NetMHCpan**

Additional configuration for the netMHCpan tool can be found on the website:

https://services.healthtech.dtu.dk/service.php?NetMHCpan-4.1

5.3

Table 5.3.1 Demographic and sequencing information for five CAPRISA002 participants (69).

Participant ID	Gender	Age	Risk factor	Subtype	Disease Progression	Fiebig Stage[1] at first sequenced timepoint	BEAST mean # days since MRCA (95% CI)[2]	HLA Type	Sample date	Weeks post infection	Whole (half) genome sequences	Sub-genomic clone/SGA sequences
CAP45	Female	41	Heterosexual	C	Slow	I/II	18 (4–35)	A*23:01, 29:02 B*15:10, 45:01 Cw*06:02, 16:01	20-Apr-05 11-May-05 07-June-05 28-June-05 27-Jul-05 04-Apr-06 04-Jul-06	2 5 9 12 16 52 65	3 6 3 1 3	16 1 8 7
CAP63	Female	32	Heterosexual	C	Rapid	III	30 (10–53)	A*02:01, 23:01 B*45:01 Cw*04:01, 16:01	06-Jan-05 09-Mar-05 13-Jul-05 07-Sep-05	2 11 29 37	11 7 10 5	19 26
CAP85	Female	24	Heterosexual	C	Intermediate	V	53 (20–101)	A*30:02 B*08:01, 45:01 Cw*07:01, 16:01	22-Jun-05 18-Aug-13 07-Dec-05 16-Feb-06 10-May-06 07_Jun-06	5 13 29 39 51 55	8 9 7 2	21 11 1 9 9

Participant ID	Gender	Age	Risk factor	Subtype	Disease Progression	Fiebig Stage[1] at first sequenced timepoint	BEAST mean # days since MRCA (95% CI) [2]	HLA Type	Sample date	Weeks post infection	Whole (half) genome sequences	Sub-genomic clone/SGA sequences
									07-Dec-06	81		1
									06-Jun-07	107		7
CAP210	Female	43	Heterosexual	C	Rapid	I/II	11 (3–25)	A*68:02	03-May-05	2	9	21
								B*15:10	13-Jun-05	12	7	8
								Cw*03:04	21-Sep-05	22	11	3
									19-Oct-05	26		3
									23-Nov-05	31		
CAP239	Female	44	Heterosexual	C	Intermediate	V*	34 (11–58) *	A*01:23, 29:02	19-Jul-05	2	2 (3)	21
									10-Aug-05	5	8	75
								B*42:01, 58:01	17-Aug-05	6	2	19
									21-Sep-05	11	9	9
								Cw*06:02, 17:01	07-Dec-05	22	0 (4)	9
									14-Jun-06	49	3	
									11-Jan-07	79		
									04-Oct-07	117		

[1]Fiebig stages (Fiebig et al., 2003) are (I/II) HIV RNA positive but antibody negative; (III) ELISA positive but non-reactive Western blot; (V) reactive Western blot without p31 band.

Table 5.3.2: Putative Cytotoxic T lymphocyte escape epitopes and polymorphisms (Abrahams et al., 2013) (69)

Participant ID	ORF	Epitope/genome region sequence*	HXB2 psn.	Participant HLA association/s**	Reference	Time of first AA change (range)(wks)	High entropy epitope/peptide (LANL subtype C database)	Shuffling/Toggling of AA mutations
CAP45	Vif	**D**WHLGHGVSI-	78–87	B*15:10	LANL database[#]	12–65	No	No
	Rev	I**H**SISERI**L**	52–60	B*15:10	LANL database	5–12	Yes	Yes
	Tat	**N**CYCK**H**CSY	24–32	A*29:02	LANL database	5–12	Yes	Yes
	Nef	**EE**VGFPVRPQV	64–74	B*45:01	Matthews et al., 2008	5–9	No	Yes
CAP63	Pol	ALT**E**IC**EE**M	188–196	A*02:01	LANL database	5–11	Yes	Yes
	Pol	QLTE**A**V**H**KI	522–530	Predicted A*02:01		11–29	No	Yes

Participant ID	ORF	Epitope/genome region sequence*	HXB2 psn.	Participant HLA association/s**	Reference	Time of first AA change (range)(wks)	High entropy epitope/peptide (LANL subtype C database)	Shuffling/Toggling of AA mutations
	Vpr	AL**IRI**LQQL	59–67	A*02:01	LANL database	5–11	Yes	Yes
	Gp41	SWSNKSE**E**DIW**G**NMTWMQ	102–119	A*23:01/Cw*04:01	LANL database	11–29	Yes	Yes
	Gp41	L**LD**SI**A**I**TV**	303–311	A*02:01	LANL database	2–4	Yes	Yes
	Nef	ALT**SS**NT**AA**	42–50	A*02:01	LANL database	5–11	Yes	Yes
	Nef	**EE**VGFPVRPQV	64–74	B*45:01	Matthews et al., 2008	0–2	No	Yes
CAP85	Pol	K**A**GYVTD**R**GRQKV**VS**LTE	609–626	B*08:01	Matthews et al., 2008	0–5	Yes	Yes

Participant ID	ORF	Epitope/genome region sequence*	HXB2 psn.	Participant HLA association/s**	Reference	Time of first AA change (range)(wks)	High entropy epitope/peptide (LANL subtype C database)	Shuffling/Toggling of AA mutations
	Gp41	**RY**L**GS**L**V**Q**Y**	283–291	A*30:02	LANL database	0–5	Yes	Yes
	Nef	**KE**VGFP**V**RPQV	64–74	B*45:01	Matthews et al., 2008	0–5	No	Yes
	Nef	**Y**FPDWQNY	120–127	A*30:02	LANL database	13–29	No	No
CAP210	Gag	VHQ**A**ISPRTL	143–152	B*15:10	Matthews et al., 2008	12–16	No	No
	Vif	**D**WHLGHG**V**SI	78–87	B*15:10	LANL database	12–16	No	Yes
	Gp41	**EA**TDR**I**L**EL**	313–321	Predicted A*68:02		2–5	Yes	Yes

Participant ID	ORF	Epitope/genome region sequence*	HXB2 psn.	Participant HLA association/s**	Reference	Time of first AA change (range)(wks)	High entropy epitope/peptide (LANL subtype C database)	Shuffling/Toggling of AA mutations
CAP239	Gag	TS**TL**QEQV**A**W	240–249	B*58:01	Matthews et al., 2008	0–2	No	Yes
	Pol	IVLP**EKE**SW	399–407	B*58:01	Matthews et al., 2008	2–5	Yes	Yes
	Nef	K**A**AV**V**DLSFF	82–90	B*58:01	Matthews et al., 2008	11–22	Yes	No

*Bold amino acids indicate sites undergoing mutation; underlined amino acids indicate sites evolving under positive selection
**Predicted epitopes obtained using Net MHCPan2.0 (www.cbs.dtu.dk/services/NetMHCPan)
#Los Alamos National Laboratory (LANL) database (www.hiv.lanl.gov) HIV Molecular Immunology 2008 compendium used

References

1. UNAIDS. Global HIV & AIDS statistics 2022 [Available from: https://www.unaids.org/en/resources/fact-sheet.
2. UNAIDS. Country Factsheets - South africa 2022 [Available from: https://www.unaids.org/en/regionscountries/countries/southafrica.
3. Organization WH. HIV Data and Statistics. 2022.
4. Siddiqui GF, Siddiqui SA, Verma P, Jaiswal R, Adhaulia A. Pre- and post-sexual exposure prophylaxis of HIV: An update. Indian J Sex Transm Dis AIDS. 2019;40(2):184-5.
5. Johnson&Johnson. South Africa Health Products Regulatory Authority Approves Dapivirine Ring Developed by the International Partnership for Microbicides 2022 [Available from: https://www.jnj.com/south-africa-health-products-regulatory-authority-approves-dapivirine-ring-developed-by-the-international-partnership-for-microbicides.
6. WHO. WHO recommends long-acting cabotegravir for HIV prevention 2022 [Available from: https://www.who.int/news/item/28-07-2022-who-recommends-long-acting-cabotegravir-for-hiv-prevention.
7. Overton ET, Richmond G, Rizzardini G, Thalme A, Girard P-M, Wong A, et al. Long-acting Cabotegravir and Rilpivirine Dosed Every 2 Months in Adults With Human Immunodeficiency Virus 1 Type 1 (HIV-1) Infection: 152-Week Results From ATLAS-2M, a Randomized, Open-label, Phase 3b, Noninferiority Study. Clinical Infectious Diseases. 2023.
8. Center for Disease Control and Prevention. Effectiveness of Prevention Strategies to Reduce the Risk of Acquiring or Transmitting HIV 2022 [Available from: https://www.cdc.gov/hiv/risk/estimates/preventionstrategies.html#print.
9. Centers for Disease Control and Prevention. Effectiveness of Prevention Strategies to Reduce the Risk of Acquiring or Transmitting HIV 2022 [Available from: https://www.cdc.gov/hiv/risk/estimates/preventionstrategies.html#print.
10. de Souza MS, Ratto-Kim S, Chuenarom W, Schuetz A, Chantakulkij S, Nuntapinit B, et al. The Thai phase III trial (RV144) vaccine regimen induces T cell responses that preferentially target epitopes within the V2 region of HIV-1 envelope. J Immunol. 2012;188(10):5166-76.

11. Karasavvas N, Billings E, Rao M, Williams C, Zolla-Pazner S, Bailer RT, et al. The Thai Phase III HIV Type 1 Vaccine Trial (RV144) Regimen Induces Antibodies That Target Conserved Regions Within the V2 Loop of gp120. AIDS Research and Human Retroviruses. 2012;28(11):1444-57.
12. Laher F, Bekker L-G, Garrett N, Lazarus EM, Gray GE. Review of preventative HIV vaccine clinical trials in South Africa. Archives of Virology. 2020;165(11):2439-52.
13. Walsh SR, Seaman MS. Broadly Neutralizing Antibodies for HIV-1 Prevention. Front Immunol. 2021;12:712122.
14. Edupuganti S, Mgodi N, Karuna ST, Andrew P, Rudnicki E, Kochar N, et al. Feasibility and Successful Enrollment in a Proof-of-Concept HIV Prevention Trial of VRC01, a Broadly Neutralizing HIV-1 Monoclonal Antibody. J Acquir Immune Defic Syndr. 2021;87(1):671-9.
15. Alter G, Barouch D. Immune Correlate-Guided HIV Vaccine Design. Cell Host & Microbe. 2018;24(1):25-33.
16. Barouch DH, Stephenson KE, Borducchi EN, Smith K, Stanley K, McNally AG, et al. Protective efficacy of a global HIV-1 mosaic vaccine against heterologous SHIV challenges in rhesus monkeys. Cell. 2013;155(3):531-9.
17. Barouch DH, Tomaka FL, Wegmann F, Stieh DJ, Alter G, Robb ML, et al. Evaluation of a mosaic HIV-1 vaccine in a multicentre, randomised, double-blind, placebo-controlled, phase 1/2a clinical trial (APPROACH) and in rhesus monkeys (NHP 13-19). The Lancet. 2018;392(10143):232-43.
18. Clincal Trials. A Combination Efficacy Study in Africa of Two DNA-MVA-Env Protein or DNA-Env Protein HIV-1 Vaccine Regimens With PrEP (PrEPVacc) 2022 [Available from: https://clinicaltrials.gov/ct2/show/NCT04066881.
19. Fan J, Liang H, Ji X, Wang S, Xue J, Li D, et al. CTL-mediated immunotherapy can suppress SHIV rebound in ART-free macaques. Nature Communications. 2019;10.
20. Cuevas JM, Geller R, Garijo R, López-Aldeguer J, Sanjuán R. Extremely High Mutation Rate of HIV-1 In Vivo. PLOS Biology. 2015;13(9):e1002251.
21. Bangham CRM. CTL quality and the control of human retroviral infections. European Journal of Immunology. 2009;39(7):1700-12.
22. Crux NB, Elahi S. Human Leukocyte Antigen (HLA) and Immune Regulation: How Do Classical and Non-Classical HLA Alleles Modulate Immune Response to

Human Immunodeficiency Virus and Hepatitis C Virus Infections? Front Immunol. 2017;8:832.
23. Quer J, Colomer-Castell S, Campos C, Andrés C, Piñana M, Cortese MF, et al. Next-Generation Sequencing for Confronting Virus Pandemics. Viruses. 2022;14(3):600.
24. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH. The origins of acquired immune deficiency syndrome viruses: where and when? Philos Trans R Soc Lond B Biol Sci. 2001;356(1410):867-76.
25. Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, Liu W, et al. SIV infection in wild gorillas. Nature. 2006;444(7116):164.
26. Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH. The origins of acquired immune deficiency syndrome viruses: where and when? Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2001;356(1410):867-76.
27. Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, et al. A new human immunodeficiency virus derived from gorillas. Nature Medicine. 2009;15(8):871-2.
28. Tang R, Yu Z, Ma Y, Wu Y, Phoebe Chen Y-P, Wong L, et al. Genetic source completeness of HIV-1 circulating recombinant forms (CRFs) predicted by multi-label learning. Bioinformatics. 2021;37(6):750-8.
29. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, et al. HIV-1 nomenclature proposal. Science. 2000;288(5463):55-6.
30. Wilkinson E, Engelbrecht S, De Oliveira T. History and origin of the HIV-1 subtype C epidemic in South Africa and the greater southern African region. Scientific Reports. 2015;5(1):16897.
31. Gartner MJ, Roche M, Churchill MJ, Gorry PR, Flynn JK. Understanding the mechanisms driving the spread of subtype C HIV-1. EBioMedicine. 2020;53:102682.
32. Bbosa N, Kaleebu P, Ssemwanga D. HIV subtype diversity worldwide. Current Opinion in HIV and AIDS. 2019;14(3):153-60.
33. Hemelaar J, Elangovan R, Yun J, Dickson-Tetteh L, Fleminger I, Kirtley S, et al. Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis. The Lancet Infectious Diseases. 2019;19(2):143-55.

34. McMichael AJ, Rowland-Jones SL. Cellular immune responses to HIV. Nature. 2001;410(6831):980-7.
35. Santoro MM, Perno CF. HIV-1 Genetic Variability and Clinical Implications. ISRN Microbiology. 2013;2013:1-20.
36. Fisher AG, Ensoli B, Looney D, Rose A, Gallo RC, Saag MS, et al. Biologically diverse molecular variants within a single HIV-1 isolate. Nature. 1988;334(6181):444-7.
37. Guha D, Ayyavoo V. Innate Immune Evasion Strategies by Human Immunodeficiency Virus Type 1. ISRN AIDS. 2013;2013:1-10.
38. Berkhout B. Structure and function of the human immunodeficiency virus leader RNA. Prog Nucleic Acid Res Mol Biol. 1996;54:1-34.
39. Human Immunodeficiency Virus (HIV). Transfus Med Hemother. 2016;43(3):203-22.
40. Van Heuvel Y, Schatz S, Rosengarten JF, Stitz J. Infectious RNA: Human Immunodeficiency Virus (HIV) Biology, Therapeutic Intervention, and the Quest for a Vaccine. Toxins. 2022;14(2):138.
41. Frankel AD, Young JAT. HIV-1: Fifteen Proteins and an RNA. Annual Review of Biochemistry. 1998;67(1):1-25.
42. Cohen MS, Shaw GM, McMichael AJ, Haynes BF. Acute HIV-1 Infection. New England Journal of Medicine. 2011;364(20):1943-54.
43. Henrard DR, Phillips JF, Muenz LR, Blattner WA, Wiesner D, Eyster ME, et al. Natural History of HIV-1 Cell-Free Viremia. JAMA. 1995;274(7):554-8.
44. Demers KR, Makedonas G, Buggert M, Eller MA, Ratcliffe SJ, Goonetilleke N, et al. Temporal Dynamics of CD8+ T Cell Effector Responses during Primary HIV Infection. PLOS Pathogens. 2016;12(8):e1005805.
45. Moir S, Chun T-W, Fauci AS. Pathogenic Mechanisms of HIV Disease. Annual Review of Pathology: Mechanisms of Disease. 2011;6(1):223-48.
46. Ford ES, Puronen CE, Sereti I. Immunopathogenesis of asymptomatic chronic HIV Infection: the calm before the storm. Current Opinion in HIV and AIDS. 2009;4(3):206-14.
47. Okoye AA, Picker LJ. CD4 T-cell depletion in HIV infection: mechanisms of immunological failure. Immunological Reviews. 2013;254(1):54-64.

48. CAPRISA. CAPRISA002: Viral set point and clinical disease progression: The role of immunological, genetic and viral factors over the course of disease and during antiretroviral therapy.
49. Mlisana K, Werner L, Garrett NJ, McKinnon LR, van Loggerenberg F, Passmore JA, et al. Rapid disease progression in HIV-1 subtype C-infected South African women. Clin Infect Dis. 2014;59(9):1322-31.
50. Pantaleo G, Fauci AS. Immunopathogenesis of HIV infection. Annual Review of Microbiology. 1996;50(1):825-54.
51. Sigal LJ. Activation of CD8 T Lymphocytes during Viral Infections. Encyclopedia of Immunobiology. 2016:286-90.
52. Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nature Communications. 2016;7(1):11660.
53. McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF. The immune response during acute HIV-1 infection: clues for vaccine development. Nature Reviews Immunology. 2010;10(1):11-23.
54. Phillips RE, Rowland-Jones S, Nixon DF, Gotch FM, Edwards JP, Ogunlesi AO, et al. Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature. 1991;354(6353):453-9.
55. Gouder PJR, Walker BD. HIV and HLA Class I: An Evolving Relationship. Immunity. 2012;37(3):426-40.
56. Philip, Bruce. HIV and HLA Class I: An Evolving Relationship. Immunity. 2012;37(3):426-40.
57. Sacha JB, Chung C, Rakasz EG, Spencer SP, Jonas AK, Bean AT, et al. Gag-specific CD8+ T lymphocytes recognize infected cells before AIDS-virus integration and viral protein expression. J Immunol. 2007;178(5):2746-54.
58. Brumme ZL, John M, Carlson JM, Brumme CJ, Chan D, Brockman MA, et al. HLA-Associated Immune Escape Pathways in HIV-1 Subtype B Gag, Pol and Nef Proteins. PLoS ONE. 2009;4(8):e6687.
59. Jia M, Hong K, Chen J, Ruan Y, Wang Z, Su B, et al. Preferential CTL targeting of Gag is associated with relative viral control in long-term surviving HIV-1 infected former plasma donors from China. Cell Research. 2012;22(5):903-14.

60. Carlson JM, Brumme CJ, Martin E, Listgarten J, Brockman MA, Le AQ, et al. Correlates of Protective Cellular Immunity Revealed by Analysis of Population-Level Immune Escape Pathways in HIV-1. Journal of Virology. 2012;86(24):13202-16.
61. Schneidewind A, Brockman MA, Yang R, Adam RI, Li B, Le Gall S, 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. Journal of Virology. 2007;81(22):12382-93.
62. Chopera DR, Wright JK, Brockman MA, Brumme ZL. Immune-mediated attenuation of HIV-1. Future Virology. 2011;6(8):917-28.
63. Wang C, Liu D, Zuo T, Hora B, Cai F, Ding H, et al. Accumulated mutations by 6 months of infection collectively render transmitted/founder HIV-1 significantly less fit. Journal of Infection. 2020;80(2):210-8.
64. Le AQ, Shahid A, Brumme ZL. HIV-1 Mutational Escape from Host Immunity. In: Hope TJ, Richman DD, Stevenson M, editors. Encyclopedia of AIDS. New York, NY: Springer New York; 2018. p. 863-78.
65. Liu MKP, Hawkins N, Ritchie AJ, Ganusov VV, Whale V, Brackenridge S, et al. Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. Journal of Clinical Investigation. 2012.
66. Kløverpris HN, Cole DK, Fuller A, Carlson J, Beck K, Schauenburg AJ, et al. A molecular switch in immunodominant HIV-1-specific CD8 T-cell epitopes shapes differential HLA-restricted escape. Retrovirology. 2015;12(1):20.
67. Zhang H, Cao S, Gao Y, Sun X, Jiang F, Zhao B, et al. HIV-1–Specific Immunodominant T-Cell Responses Drive the Dynamics of HIV-1 Recombination Following Superinfection. Frontiers in immunology. 2022;12.
68. Sahay B, Nguyen CQ, Yamamoto JK. Conserved HIV Epitopes for an Effective HIV Vaccine. J Clin Cell Immunol. 2017;8(4).
69. Abrahams MR, Treurnicht FK, Ngandu NK, Goodier SA, Marais JC, Bredell H, et al. Rapid, complex adaptation of transmitted HIV-1 full-length genomes in subtype C-infected individuals with differing disease progression. Aids. 2013;27(4):507-18.
70. Chopera DR, Woodman Z, Mlisana K, Mlotshwa M, Martin DP, Seoighe C, et al. Transmission of HIV-1 CTL Escape Variants Provides HLA-Mismatched Recipients with a Survival Advantage. PLoS Pathogens. 2008;4(3):e1000033.

71. Fryer HR, Frater J, Duda A, Palmer D, Phillips RE, McLean AR. Cytotoxic T-Lymphocyte Escape Mutations Identified by HLA Association Favor Those Which Escape and Revert Rapidly. Journal of Virology. 2012;86(16):8568-80.
72. Epitope Location Finder 2018 [Available from: https://www.hiv.lanl.gov/content/sequence/ELF/explanation.html.
73. HIV HLA Anchor Residue Motifs (Motif Scan) [Available from: https://www.hiv.lanl.gov/content/immunology/motif_scan/motif_scan.
74. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. Nucleic Acids Res. 2020;48(W1):W449-w54.
75. NetMHCpan - 4.1. Pan-specific binding of peptides to MHC class I proteins of known sequence [Available from: https://services.healthtech.dtu.dk/services/NetMHCpan-4.1/.
76. Carlson JM, Brumme ZL, Rousseau CM, Brumme CJ, Matthews P, Kadie C, et al. Phylogenetic Dependency Networks: Inferring Patterns of CTL Escape and Codon Covariation in HIV-1 Gag. PLoS Computational Biology. 2008;4(11):e1000225.
77. HIV Molecular Immunology Database [Available from: https://www.hiv.lanl.gov/content/immunology/index.html.
78. Yang F, Patton K, Kasprzyk T, Long B, Gupta S, Zoog SJ, et al. Validation of an IFN-gamma ELISpot assay to measure cellular immune responses against viral antigens in non-human primates. Gene Therapy. 2022;29(1-2):41-54.
79. Bronke C, Almeida C-AM, McKinnon E, Roberts SG, Keane NM, Chopra A, et al. HIV escape mutations occur preferentially at HLA-binding sites of CD8 T-cell epitopes. AIDS. 2013;27(6):899-905.
80. Tumiotto C, Riviere L, Bellecave P, Recordon-Pinson P, Vilain-Parce A, Guidicelli G-L, et al. Sanger and Next-Generation Sequencing data for characterization of CTL epitopes in archived HIV-1 proviral DNA. PLOS ONE. 2017;12(9):e0185211.
81. Chabria SB, Gupta S, Kozal MJ. Deep Sequencing of HIV: Clinical and Research Applications. Annual Review of Genomics and Human Genetics. 2014;15(1):295-325.

82. Goldman D, Domschke K. Making sense of deep sequencing. Int J Neuropsychopharmacol. 2014;17(10):1717-25.
83. Van Loggerenberg F, Mlisana K, Williamson C, Auld SC, Morris L, Gray CM, et al. Establishing a Cohort at High Risk of HIV Infection in South Africa: Challenges and Experiences of the CAPRISA 002 Acute Infection Study. PLoS ONE. 2008;3(4):e1954.
84. Abrahams M-R, Joseph SB, Garrett N, Tyers L, Moeser M, Archin N, et al. The replication-competent HIV-1 latent reservoir is primarily established near the time of therapy initiation. Science Translational Medicine. 2019;11(513):eaaw5589.
85. Ranwez V, Chantret N, Delsuc F. Aligning Protein-Coding Nucleotide Sequences with MACSE. Springer US; 2021. p. 51-70.
86. Larsson A. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics. 2014;30(22):3276-8.
87. Gap Strip/Squeeze v2.1.0 [Available from: https://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html.
88. Labuschagne P. Epitope Matcher.
89. CTL/CD8+ Epitope Summary [Internet]. Available from: https://www.hiv.lanl.gov/content/immunology/tables/ctl_summary.html.
90. AnalyzeAlign [Available from: https://www.hiv.lanl.gov/content/sequence/ANALYZEALIGN/analyze_align.html.
91. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol. 2013;30(5):1196-205.
92. Shannon Entropy-One [Available from: https://www.hiv.lanl.gov/content/sequence/ENTROPY/entropy_one.html.
93. Kiepiela P, Leslie AJ, Honeyborne I, Ramduth D, Thobakgale C, Chetty S, et al. Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature. 2004;432(7018):769-75.
94. Prentice Heather A, Porter Travis R, Price Matthew A, Cormier E, He D, Farmer Paul K, et al. HLA-B*57 versus HLA-B*81 in HIV-1 Infection: Slow and Steady Wins the Race? Journal of Virology. 2013;87(7):4043-51.
95. Blackwell JM, Jamieson SE, Burgner D. HLA and Infectious Diseases. Clinical Microbiology Reviews. 2009;22(2):370-85.
96. Huang J, Goedert JJ, Sundberg EJ, Cung TDH, Burke PS, Martin MP, et al. HLA-B*35-Px–mediated acceleration of HIV-1 infection by increased inhibitory

immunoregulatory impulses. Journal of Experimental Medicine. 2009;206(13):2959-66.
97. Delport W, Scheffler K, Seoighe C. Frequent Toggling between Alternative Amino Acids Is Driven by Selection in HIV-1. PLoS Pathogens. 2008;4(12):e1000242.
98. Frahm N, Kiepiela P, Adams S, Linde CH, Hewitt HS, Sango K, et al. Control of human immunodeficiency virus replication by cytotoxic T lymphocytes targeting subdominant epitopes. Nature Immunology. 2006;7(2):173-8.
99. Zhang X, Huang X, Xia W, Li W, Zhang T, Wu H, et al. HLA-B*44 Is Associated with a Lower Viral Set Point and Slow CD4 Decline in a Cohort of Chinese Homosexual Men Acutely Infected with HIV-1. Clinical and Vaccine Immunology. 2013;20(7):1048-54.
100. Carlson JM, Listgarten J, Pfeifer N, Tan V, Kadie C, Walker BD, et al. Widespread Impact of HLA Restriction on Immune Control and Escape Pathways of HIV-1. Journal of Virology. 2012;86(9):5230-43.
101. CTL/CD8+ Epitope Variants and Escape Mutations [Internet]. Available from: https://www.hiv.lanl.gov/content/immunology/variants/ctl_variant.html.
102. Thomas M, Hopkins C, Duffy E, Lee D, Loulergue P, Ripamonti D, et al. Association of the HLA-B*53:01 Allele With Drug Reaction With Eosinophilia and Systemic Symptoms (DRESS) Syndrome During Treatment of HIV Infection With Raltegravir. Clin Infect Dis. 2017;64(9):1198-203.
103. Geels MJ, Dubey SA, Anderson K, Baan E, Bakker M, Pollakis G, et al. Broad Cross-Clade T-Cell Responses to Gag in Individuals Infected with Human Immunodeficiency Virus Type 1 Non-B Clades (A to G): Importance of HLA Anchor Residue Conservation. Journal of Virology. 2005;79(17):11247-58.
104. Chopera DR, Mlotshwa M, Woodman Z, Mlisana K, De Assis Rosa D, Martin DP, et al. Virological and Immunological Factors Associated with HIV-1 Differential Disease Progression in HLA-B*58:01-Positive Individuals. Journal of Virology. 2011;85(14):7070-80.
105. Bansal A, Gough E, Sabbaj S, Ritter D, Yusim K, Sfakianos G, et al. CD8 T-cell responses in early HIV-1 infection are skewed towards high entropy peptides. AIDS. 2005;19(3).
106. Honeyborne I, CodoñEr FM, Leslie A, Tudor-Williams G, Luzzi G, Ndung'U T, et al. HLA-Cw*03-Restricted CD8 T-Cell Responses Targeting the HIV-1 Gag Major

Homology Region Drive Virus Immune Escape and Fitness Constraints Compensated for by Intracodon Variation. Journal of Virology. 2010;84(21):11279-88.

107. Hölzemer A, Thobakgale CF, Jimenez Cruz CA, Garcia-Beltran WF, Carlson JM, Van Teijlingen NH, et al. Selection of an HLA-C*03:04-Restricted HIV-1 p24 Gag Sequence Variant Is Associated with Viral Escape from KIR2DL3+ Natural Killer Cells: Data from an Observational Cohort in South Africa. PLOS Medicine. 2015;12(11):e1001900.

108. Leitman EM, Thobakgale CF, Adland E, Ansari MA, Raghwani J, Prendergast AJ, et al. Role of HIV-specific CD8+ T cells in pediatric HIV cure strategies after widespread early viral escape. Journal of Experimental Medicine. 2017;214(11):3239-61.

109. van Teijlingen NH, Hölzemer A, Körner C, García-Beltrán WF, Schafer JL, Fadda L, et al. Sequence variations in HIV-1 p24 Gag-derived epitopes can alter binding of KIR2DL2 to HLA-C*03:04 and modulate primary natural killer cell function. Aids. 2014;28(10):1399-408.

110. Schweighardt B, Wrin T, Meiklejohn DA, Spotts G, Petropoulos CJ, Nixon DF, et al. Immune Escape Mutations Detected Within HIV-1 Epitopes Associated With Viral Control During Treatment Interruption. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2010;53(1):36-46.

111. Roberts HE, Hurst J, Robinson N, Brown H, Flanagan P, Vass L, et al. Structured Observations Reveal Slow HIV-1 CTL Escape. PLOS Genetics. 2015;11(2):e1004914.

112. Ntale RS, Chopera DR, Ngandu NK, Assis De Rosa D, Zembe L, Gamieldien H, et al. Temporal Association of HLA-B*81:01- and HLA-B*39:10-Mediated HIV-1 p24 Sequence Evolution with Disease Progression. Journal of Virology. 2012;86(22):12013-24.

113. Casadellà M, Paredes R. Deep sequencing for HIV-1 clinical management. Virus Res. 2017;239:69-81.

114. Chakraborty S, Rahman T, Chakravorty R. Characterization of the Protective HIV-1 CTL Epitopes and the Corresponding HLA Class I Alleles: A Step towards Designing CTL Based HIV-1 Vaccine. Advances in Virology. 2014;2014:1-17.

115. Currier JR, Visawapoka U, Tovanabutra S, Mason CJ, Birx DL, McCutchan FE, et al. BMC Immunology. 2006;7(1):8.

116. Honeyborne I, Leslie A, Crawford H, Rousseau C, Mullins J, Walker B, et al. P09-18. Cw*0303/0304 HIV specific CTL response toward GagYL9 select for HIV escape variants with low fitness that is compensated by intra-codon variation. Retrovirology. 2009;6(S3):P131.