HIV evolution during ART failures revealed by using long-read sequencing and bioinformatics tools

Shiyi Wang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Bruce Torbett, Chair

Deborah Fuller

Pejman Mohammadi

Program Authorized to Offer Degree:

Department of Laboratory Medicine & Pathology

University of Washington

Abstract

HIV evolution during ART failures revealed by using long-read sequencing and bioinformatics tools

Shiyi Wang

Chair of the Supervisory Committee:

Bruce Torbett

Department of Pediatrics

HIV resistance often leads to antiretroviral therapy (ART) failures, involving two crucial mutation categories: drug-resistance mutations (DRMs) and compensatory mutations. DRMs reside in HIV enzyme active sites (protease, reverse transcriptase, and integrase), hindering drug binding, while compensatory mutations restore enzyme stability and function, compensating for DRMs. With the increase of drug potency, more compensatory mutations are involved in compensating for one DRM, forming complex mutational patterns. However, the interplay between DRMs and compensatory mutations remains elusive.

In this thesis work, I combined a long-read sequencing approach and bioinformatic tools to unveil the complex mutational patterns driving HIV resistance development. Long-read

sequencing yielded 4.5kb *gag-pol* sequences from individual HIV genomes within clinical serum samples, preserving co-varying mutations critical for pattern identification. Mutational patterns were inferred based on pairwise correlations detected in the sequencing data and quantified using a custom bioinformatic tool. I utilized Hamming-distance-based phylogenetic analysis (HDBPA) and paired post-ART HIVs with their pre-ART most recent common ancestors (MRCAs) based on sequence similarity. In this way, I divided mutations in mutational patterns into different categories (mutations inherited from pre-ART MRCA, and mutations acquired during ART) and revealed the order of mutation development. I demonstrated the utility of this approach by studying the HIV evolution in two PWHs facing ART failures. The findings revealed different mutational patterns selected and enriched during ART and inferred evolutionary pathways taken by HIVs during resistance development.

Alongside substitution mutation involved in HIV evolution, I participated in a collaborative study, aiming to measure linkage disequilibrium between recombination events and SNVs. The findings revealed novel correlations between $p6^{Gag}$ insertions and Gag cleavage site mutations in drug-resistant HIV genomes.

Taken together, my work deciphered mutational patterns and recombination events driving HIV evolution during ART using long-read sequencing and custom bioinformatics tools. The findings of this study indicated interactions both within HIV proteins and among proteins, which could guide anti-viral drug design. The methods introduced could be used for identifying complex mutational patterns required for resistance development and revealing the order of mutation development in HIV as well as other fast-evolving viruses and bacteria.

TABLE OF CONTENTS

Chapter 1. Introduction.....1

1.1 Drug Resistance Mutations (DRMs) and Compensatory Mutations Co-evolve Within the HIV Genome During Antiretroviral Therapy (ART).....1

1.2 Long-Read Sequencing Strategy Provides the Sequences of Individual HIVs from Genetically Diverse HIV-1 Populations.....6

1.3 Novel Associations Between Recombination Events and SNVs in Viral Genomes Revealed Using Co-Variation Mapper (CoVaMa).....10

Chapter 2. Co-varying mutations within HIV *gag-pol* region revealed by long-read sequencing.....13

2.1 Abstract.....13

2.2 Results.....15

2.2.1 MrHAMER Detects Linked Mutations with High Sensitivity and Preserved Original Proportions.....15

2.2.2 The Enrichment of a Pre-ART, Drug-Resistant HIV *gag-pol* Mutational Pattern During ART Failures in one PWH Revealed by Long-read Sequencing.....18

2.3 Discussion.....24

2.4 Methods.....27

Chapter 3. Complex mutational patterns that drive HIV resistance development revealed using bioinformatics analysis tools.....32

3.1 Abstract....32

3.2 Results....34

3.2.1 HDBPA Reveals the Evolution of Individual HIVs Druing ART Failures....34

3.2.2 Identification of Enriched Mutational Patterns and Uncovering the Order of Mutation Development in Two PWHs During ART Failures....37

3.3 Discussion....47

3.4 Methods....51

Chapter 4. Co-variation of viral recombination with single nucleotide variants (SNVs) during viral evolution revealed by an improved co-variation mapper (CoVaMa)....53

4.1 Abstract....53

4.2 Results....55

4.2.1 Overview of the CoVaMa Pipeline (v0.7)....55

4.2.2 Associations Between Recombination Events and SNVs in Defective Flock House Virus (FHV) RNAs Revealed by CoVaMa....58

4.2.3 Associations Between Insertion in $p6^{Gag}$ and Mutations in Gag Cleavage Sites Revealed by CoVaMa....66

4.3 Discussion....72

4.4 Methods....76

Chapter 5. Estimation of the optimal sample size needed for correlated mutation identification using an individual-virus-based simulator....81

5.1 Abstract....81

5.2 Results....83

5.2.1 The Forward Simulator Effectively Simulates the Acquisition of Correlated Mutations in the HIV-1 Protease During Resistance Development....83

5.2.2 The Simulator Mimics Viral Rebound Resulting from the Selection of DRMs and Recaptures the Linkage Distribution in Drug-Resistant HIV Protease Sequences....92

5.2.3 Sample Size Needed for Identifying Correlated DRMs and Compensatory Mutations Estimated Using Synthetic Samples....94

5.3 Discussion....101

5.4 Methods....105

Chapter 6. Conclusion....112

List of Figures

Figure 1.1. Correlated mutations are detected across the HIV Gag-protease in protease inhibitors-resistant viral populations.....3

Figure 1.2. Compensatory mutations compensate for DRMs by stabilizing the protease structure or improving protease-substrate binding..... 5

Figure 1.3. Workflow of MrHAMER sequencing methodology..... 9

Figure 2.1. MrHAMER detects linked mutations in the HIV *gag-pol* with high sensitivity and preserves the original proportion of different mutation pairs in complex synthetic libraries.....16

Figure 2.2. Diversity reduction is detected in the HIV population during ART failures.....19

Figure 2.3. Sequencing of longitudinal serum samples from one PWH experiencing ART failures reveals the enrichment of a pre-ART, drug-resistant *gag-pol* mutational pattern.....22

Figure 3.1. HDBPA reveals the evolution of individual HIVs and the linkage disequilibrium-based analysis reveals mutational patterns from sequencing data..... 35

Figure 3.2. Individual HIVs in one PWH follow two distinct evolutionary pathways during the ART failure..... 39

Figure 3.3. Three major evolutionary pathways taken by individual HIVs during therapy failures in one PWH.....44

Figure 4.1. Overview of the CoVaMa pipeline.....56

Figure 4.2. Detection of recombination events in FHV RNA2 using CoVaMa..... 60

Figure 4.3. CoVaMa reveals associations between recombination events and SNVs in FHV RNA2..... 61

Figure 4.4. A226G's potential influence on the secondary structure of FHV RNA2..... 65

Figure 4.5. CoVaMa reveals associations between insertions in $p6^{Gag}$ and mutations in Gag cleavage sites..... 68

Figure 4.6. CoVaMa reveals associations between insertions and mutation in the HIV matrix protein..... 71

Figure 5.1. The simulator recaptures the fitness increase along with the acquisition of correlated mutations under drug pressure..... 88

Figure 5.2. Overview of the simulator and the simulation pipeline..... 91

Figure 5.3. The simulator recaptures the linkage distribution in drug-resistant protease sequences. 93

Figure 5.4. Overview of the downstream analysis pipeline..... 95

Figure 5.5. Using the unique reads with a read count ≥ 2 in each sample improves the precision of correlated mutation identification..... 97

Figure 5.6. Precision-recall trade-off across varies sample sizes..... 100

List of Tables

Table 2.1. The 25 SNVs that are enriched during ART failures..... 21

Table 3.1. Two predominant mutational patterns are identified in the virological-failure sample.... 41

Table 3.2. Genetic patterns formed in each evolutionary pathway two weeks after starting the second therapy.....45

Table 4.1. A226G and G575A are negatively associated with each other in FHV RNA2..... 62

Table 4.2. Frequency of A226G in full-length FHV RNA2 and D-RNA2 over passage..... 63

Table 4.3. Frequency of G575A in full-length FHV RNA2 and D-RNA2 over passage..... 63

Table 4.4. Insertions detected in longitudinal serum samples from one PWH..... 67

Table 4.5. Command lines used to analyze two viral datasets in this study.....79

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Dr. Bruce E. Torbett, who is always supportive of my projects. His expert guidance and warm encouragement have been invaluable to me.

I would like to express my gratitude to my co-advisor, Dr. Pejman Mohammadi, who has shaped the bioinformatics aspect of my thesis. His prompt responses to my questions have been tremendously helpful.

I would like to express my gratitude to my committee members, Dr. Lisa M. Frenkel, Dr. Deborah H. Fuller, and Dr. Jesse D. Bloom. Their essential comments and suggestions at every stage of my thesis project have greatly contributed to its development.

I would like to express my gratitude to Dr. Andrew L. Routh at UTMB, Dr. David Smith at UCSD, Dr. James Mullins at UW, Dr. Christian M. Gallardo at SCRI, and Dr. Daniel Montiel-Garcia at Scripps for their invaluable contributions to my thesis project.

I would like to express my gratitude to my wonderful lab mates, Jade Wolff, Dr. Shentian Zhuang, Tien Le, Dr. Eirini Vamva, Dr. Chet R. Ojha, Dr. Mia Faerch, Dr. Taiwei Li, and Dr. Craig Schindewolf. Collaborating with them has enriched my research experience in countless ways.

I would like to thank Megan V. Barker and Dr. William Mahoney at the M3D program for their support during my transfer and continuous assistance throughout my graduate journey.

I would like to express my gratitude to my husband, Dr. Yisong Deng, for his unwavering emotional support. I am also grateful to my family and his family for standing by me during this lengthy journey. Additionally, I owe appreciation to my faithful cream-colored hamster, who was a constant, quiet companion for the past two years since my move to Seattle.

Finally, I would like to extend my thanks to all those who have supported me over the past five years. Your encouragement has been invaluable in helping me reach this milestone.

Chapter 1. INTRODUCTION

1.1 Drug Resistance Mutations (DRMs) and Compensatory Mutations Co-evolve Within the HIV Genome During Antiretroviral Therapy (ART)

HIV-1 remains a global health concern, with approximately 39 million people living with HIV-1 worldwide (UNAIDS, 2023). Although antiretroviral therapy (ART) that targets HIV-1 enzymes and other viral proteins have considerably improved over the past few decades, some individuals still fail ART, resulting in loss of viral suppression[1]. This could be due to poor treatment adherence when not clinically managed appropriately or infection with inhibitor-resistant HIVs. Like other RNA viruses, HIV-1 exhibits error-prone reverse transcription and a short replication time[2]. Therefore, when not fully ART suppressed, HIV-1 replicates fast and forms a highly heterogeneous viral population in people living with HIV (PWHs)[3]. This genetically diverse viral population provides the opportunity for viral selection and, consequently, the development of ART resistance.

During ART, drug resistance mutations (DRMs) located in primary sites (usually enzyme active sites) are selected due to the resulting changes to viral protein structures that limit inhibitor binding[1,4,5]. While providing the virus with inhibitor resistance, DRMs can come at the expense of viral fitness, given that mutational changes that impede drug binding often result in changes in the enzyme active site, leading to loss of function[6]. To counteract these deleterious effects and improve viral fitness, compensatory mutations arise elsewhere within the same viral protein or in different viral proteins[7,8]. These mutations, not directed at the inhibitor binding site *per se*, lead to protein or viral alterations that mitigate the impact of DRMs[7–9]. For instance, compensatory

mutations could increase the stability of the viral enzymes (protease, integrase, and reverse transcriptase) or alter distal enzyme substrates, such as protease cleavage sites in the Gag polyprotein[10–13]. Other mitigating mutations that alter envelope function and enhance viral spread also aid in compensating for DRMs[9]. Driven by continuous selection and viral fitness improvement, DRMs and compensatory mutations tend to coevolve in the HIV-1 genome during ART.

Strong correlations between mutations in the HIV-1 genome have been observed in drug-resistant HIV-1 populations[14]. Flynn et al. (2015) sequenced serum samples collected from 93 PWHs who failed ART consisting of protease inhibitors[14]. They estimated the bivariate joint probabilities from the observed single-site frequencies detected by deep sequencing to identify correlated residues across the Gag and protease regions. Their results revealed strongly correlated mutations located in the Gag polyprotein as well as between the Gag polyprotein and protease (Figure 1.1).

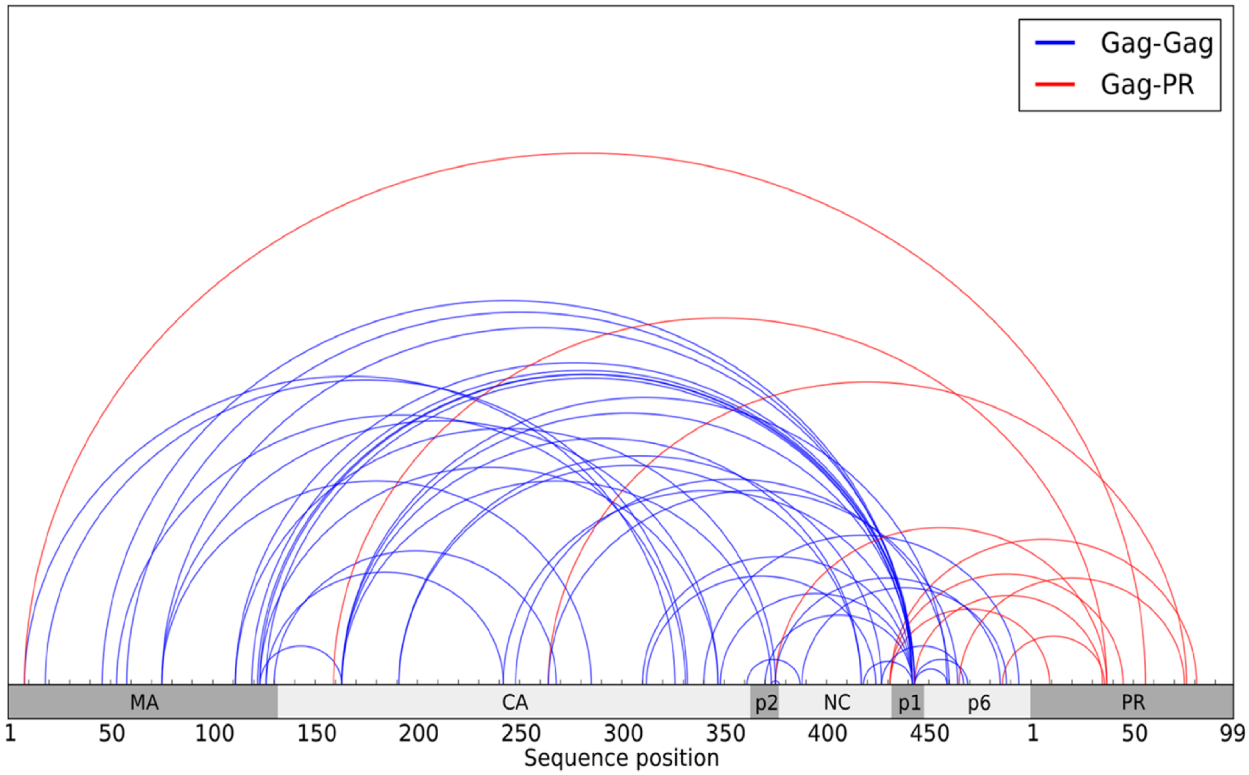

Figure 1.1. Correlated mutations are detected across the HIV Gag-protease in protease inhibitors-resistant viral populations.

Serum samples collected from 93 PWHs who failed protease inhibitor-based ART were sequenced. The sequencing dataset was analyzed using mutual information-based covariation analysis to infer correlated mutation within the HIV-1 genome. The 50 most significant correlations between mutations that are within the Gag polyprotein (blue) and across Gag and protease (red) are shown in this plot. This figure is acquired from Flynn et al. (2015). PWH: person living with HIV. ART: antiretroviral therapy. PR: protease. MA: matrix protein. CA: capsid protein. NC: nucleocapsid protein.

Additionally, several studies have been conducted investigating the biological and structural basis for the co-evolution between compensatory mutations and DRMs. Chang et al. (2011) investigated the evolutionary cost of DRMs in HIV-1 protease in terms of protein stability

(Figure 1.2A)[8]. They measured the melting temperatures of the wild-type HIV-1 protease and mutant HIV-1 proteases containing DRMs. Compared to the wild-type HIV-1 protease, mutant HIV-1 proteases with DRMs displayed reduced melting temperatures, indicating a destabilized protein structure. Thereafter, the researchers measured the melting temperatures of mutant HIV-1 proteases exhibiting both DRMs and compensatory mutations. The results indicated that the acquisition of compensatory mutations restored the stability of the mutant proteases to the level of, or even beyond, the wild-type baseline.

HIV-1 protease cleaves the Gag polyprotein at different cleavage sites (CSs), e.g., the P1/p6 cleavage site (Figure 1.2C). This cleavage step plays an essential role during the maturation of HIV-1 particles. Kolli et al. (2014) investigated the structural basis of co-evolution of the P1/p6 CS with the protease DRMs, D30N/N88D (Figure 1.2D)[11]. Their work showed that a P1/p6 CS containing compensatory mutations (L449F and S451N) acts as a better substrate for both wild-type and mutant HIV-1 protease (D30N/N88D) than the wild-type CS.

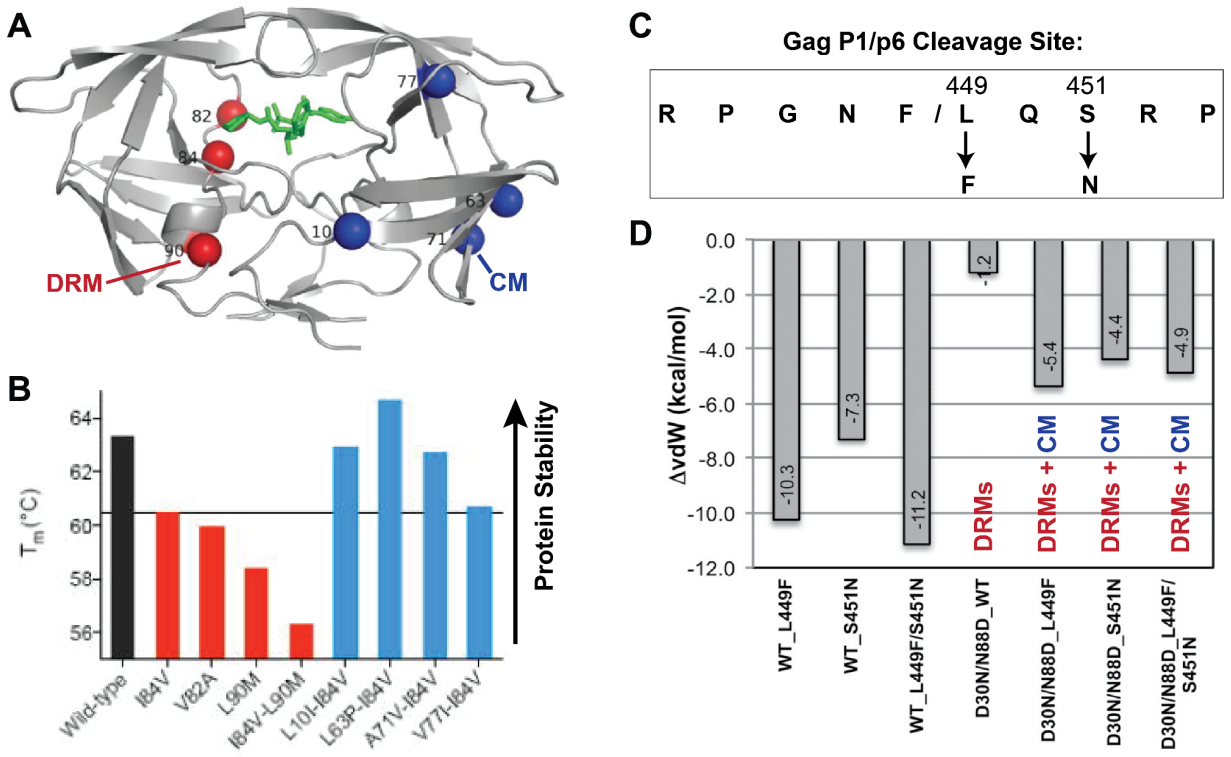

Figure 1.2. Compensatory mutations compensate for DRMs by stabilizing the protease structure or improving protease-substrate binding.

(A) This diagram illustrates the structure of an HIV-1 protease dimer, highlighting major DRMs in red and compensatory mutations (CMs) in blue. This figure is acquired from Chang and Torbett (2011). (B) Melting temperatures (T_m) and protein stability comparison among the HIV-1 wild-type protease (black), proteases with DRMs (red), and proteases with DRMs and compensatory mutations (blue). The melting temperatures were determined by differential scanning calorimetry. This figure is acquired from Chang and Torbett (2011). (C) The amino acid sequence of the wild-type HIV-1 Gag P1/p6 CS. The two compensatory mutations, L449F and S451N, are annotated. (D) Overall van der Waals (vdW) interactions between mutant P1/p6 and mutant protease in comparison to the wild-type complex (PR_{WT}-$P1/p6_{WT}$). More negative total vdW contact energy suggests more favorable binding. This figure is acquired from Kolli et al. (2014). DRM: drug resistance mutation. CM: compensatory mutation. CS: cleavage site.

1.2 Long-Read Sequencing Strategy Provides the Sequences of Individual HIVs from Genetically Diverse HIV-1 Populations

Next generation sequencing (NGS) has been used to identify correlated mutations selected in the HIV-1 genome during ART. However, like Illumina sequencing, NGS relies on the short-sequencing technique (250 - 800 bp)[15]. Considering the ~4.5kb HIV *gag-pol* genome, a short read length hinders the direct identification of distal-linked mutations in the HIV genome. Haplotype assembly tools were developed to construct haplotype sequences based on short reads[16]. However, the high similarity among HIV genomes brings extra complexity to haplotype construction. Moreover, the haplotypes in the HIV swarm have largely varied abundances, and minority haplotypes challenge the sensitivity of haplotype construction.

To acquire the sequence of individual viral genomes, the Single-Genome Amplification (SGA) technique was developed. In this method, viral genomes were serially diluted to one copy per reaction before amplification and sequencing[17]. In this way, SGA avoided template switching and chimera formation, which were seen in the bulk PCR, thereby preserving the integrity of the original viral genomes. As a result, SGA has been used to generate sequences of individual HIVs in different studies[18,19]. However, SGA is time-consuming and labor-intensive. Hence, it has been limited to a relatively low sampling depth. Therefore, to capture HIV genomes in genetically diverse HIV populations, researchers started developing more efficient sequencing methodologies based on long-read sequencing strategy.

Long-read sequencing methodologies, such as PacBio and Oxford Nanopore Technologies (ONT), provide sequences of long, single DNAs or RNAs, which can cover the HIV *gag-pol*

region[20]. However, they have a high sequencing error rate, which hinders the detection of minor mutations in the viral population[21]. To improve the sequencing accuracy, PacBio introduced circular consensus sequencing (CCS), which generates sequential repeats of the target genome[22]. These sequential repeats are then used to construct a consensus sequence of the target genome. In this way, sequencing accuracy is increased to 99.8%[22]. Similarly, Wilson et al. (2019) utilized the Rolling-Circle Amplification (RCA) during ONT sequencing library preparation. They generated concatemeric repeats of ultrashort DNA sequences (<100bp) and successfully reduced the ONT sequencing error rate to < 10%[23].

Alternative to the CCS, Karst et al. (2021) combined Unique Molecular Identifiers (UMIs) and long-read sequencing techniques (PacBio and ONT)[24]. Unique UMIs were added to both ends of target genomes during PCR amplification. After sequencing, reads were binned based on their terminal UMI-pairs. For each UMI-pair bin, a consensus sequence was generated and polished. In this way, they acquired the sequences of long target genomes with thousands of base pairs with an error rate of < 0.01%. Additionally, chimera genomes would have UMIs that were detected in other more common UMI-pairs. This made chimera genomes easy to identify and remove. By implementing this UMI-pair filtering, the researchers reduced the chimera rate in the sequencing output to less than 0.02%.

To acquire individual HIV-1 *gag-pol* genomes, in 2021, our lab developed the Multi-read Hairpin Mediated Error-correction Reaction (MrHAMER) sequencing methodology based on the ONT sequencing technology and established downstream bioinformatic pipelines tailored for HIV studies[25]. In this methodology, HIV-1 RNAs were first reverse-transcribed and PCR amplified to generate double-stranded cDNAs. To decrease the template-switching rate during PCR

amplification, we performed two rounds of emulsion PCR (20 cycles each). The emulsion PCRs avoided the occurrence of more than one template in each droplet (reaction) and thus reduced the template-switching rate to 0.45%. After this, we ligated hairpins to both ends of double-stranded cDNAs. Using primers targeted on the hairpin sequence, we utilized RCA to produce concatemers consisting of several sequential repeats of the target viral genome. Long concatemers were then made double-stranded, size-selected, and sequenced using an in-house MinION sequencing device. The downstream polishing process used the sequential repeats of each target viral genome to reconstruct the consensus sequence. This step eliminated the random errors generated during the PCR amplification and sequencing process. By doing this, the error rate of the ONT sequencing was reduced from over 10% to ~0.133%. Additionally, we validated that MrHAMER could detect an input RNA amount as low as 5,000 viral genomic copies (2,500 virions or 2,500 viruses/ml of sera), which provided relevance for use with clinical samples (Figure 1.3).

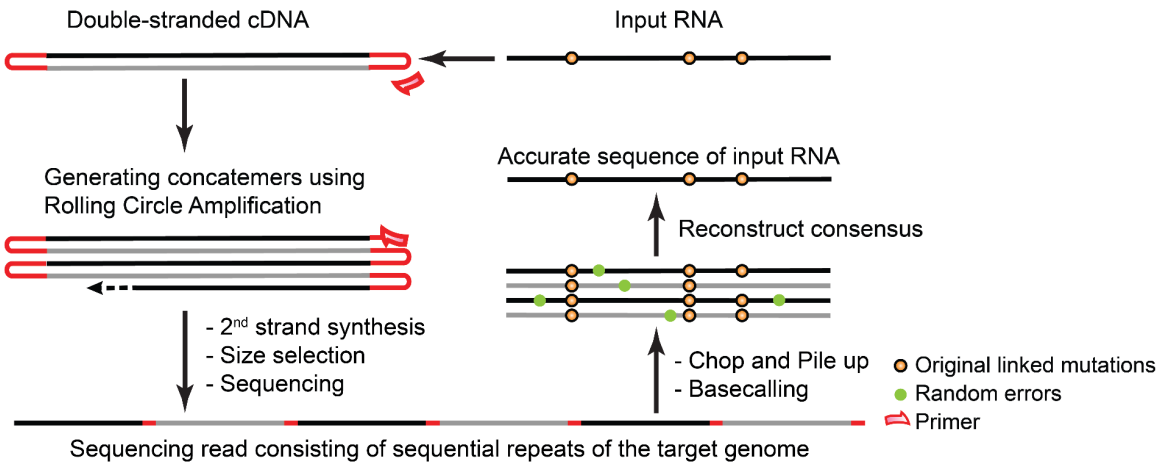

Figure 1.3. Workflow of MrHAMER sequencing methodology.

The MrHAMER sequencing pipeline involves several key steps to capture the target RNA sequence accurately. Initially, the input target RNA is reverse-transcribed into cDNA. Hairpins are added to both ends of the double-stranded cDNA. Subsequently, primers located within the hairpin regions are used to start the Rolling Circle Amplification (RCA), generating concatemers consisting of sequential repeats of the target sequence. The concatemers are then made double-stranded, size-selected, and sequenced using Oxford Nanopore sequencing technology. After sequencing and basecalling, sequential repeats of the target sequence are extracted from the sequencing read and used to reconstruct the consensus sequence. In this figure, mutations in the target RNA sequence are marked by orange dots, while green dots represent random errors that can occur during PCR amplification and sequencing.

In Chapter 2, I validated MrHAMER's ability to identify linked mutations with a frequency as low as 1%. Additionally, I confirmed that MrHAMER could preserve the original proportion of different haplotypes in the heterogeneous viral swarm. This allows the early detection of rare founder species that might escape ART pressure, which could assist in selecting the optimal

ART. Furthermore, the abundance of different viral haplotypes in the sequencing outputs could be used to estimate their replication capacities under drug pressure.

1.3 Novel Associations Between Recombination Events and SNVs in Viral Genomes Revealed Using Co-Variation Mapper (CoVaMa)

Recombination that occurs within viral genomes, particularly RNA viruses, is another powerful driving force behind viral evolution and adaptation[26], besides the acquisition of Single-Nucleotide Variants (SNVs). Recombination results from template-switching events during the replication of viral RNAs. In this process, the viral polymerase accidentally disassociates from its original template and then re-associates to another region in the same or a different template[26]. This could result in various recombination events in the viral genome, including insertions, deletions, and duplications, and can dramatically impact viral fitness, viral intra-host diversity, and the development of resistance to antivirals[27–30].

Due to the compactness of viral genomes, viral adaptation events are more likely to cooperate with each other[14,31,32]. To identify correlated SNVs in viral genomes, several methods have been published[33], such as Co-Variation Mapper (CoVaMa, v0.1) introduced by Routh et al. (2015)[34]. CoVaMa measures the linkage disequilibrium (LD) between SNVs within viral NGS datasets[34]. CoVaMa identifies all loci with SNVs from input NGS reads and interrogates all possible pairwise associations between loci. LD values are calculated for each pairwise association: an LD value of 0 indicates no association, and an LD value that is close to +/- 0.25 indicates a strong association. In the same study, Routh et al. demonstrated the utility of this approach by

identifying associated mutations in Flock House virus (FHV) passaged *in vitro* and inhibitor-resistant HIVs collected from PWHs.

However, unlike the correlation between SNVs, studying the correlation between recombination events with each other or with SNVs has been long hampered by their inherent genetic complexity and a lack of bioinformatic tools. Characterizing these correlations may be necessary for understanding why certain recombination events might be selected and their role in viral evolution. Thus, I aimed to identify the associations between recombination events, i.e., insertions or deletions (InDels) and duplications, as well as between recombination events and SNVs.

Collaborating with the Routh Lab, we expanded the previously reported CoVaMa pipeline (v0.1) to measure linkage disequilibrium between recombination events and SNVs within both short-read and long-read sequencing datasets[35]. I revised the new version of CoVaMa (v0.7) and demonstrated its applicability by reanalyzing two different viral datasets: an FHV dataset that was collected during *in vitro* culture and sequenced using long-read sequencing[36] and an HIV dataset that was collected from PWHs who failed ART and sequenced using NGS[14,37].

From the FHV dataset, I identified multiple SNVs that were either correlated or anti-correlated with large deletion events in FHV Defective-RNAs (D-RNAs). I hypothesized that these mutations were adaptive mutations that either allowed increased replication of D-RNAs or allowed full-length genomes to escape replication attenuation by D-RNAs. From the HIV dataset, I identified insertions in the PTAP region of $p6^{Gag}$, which correlated with mutations found in Gag cleavage sites. The location of these linked SNVs and InDels proximal to Gag cleavage sites suggested a role for these adaptations in supporting drug resistance development.

Overall, CoVaMa (v0.7) provided a powerful tool to characterize the molecular details of viral adaptation and the impact of RNA recombination upon virus evolution. The CoVaMa (v0.7) script is publicly available at https://sourceforge.net/projects/covama/.

Chapter 2. Co-varying mutations within HIV gag-pol region revealed by long-read sequencing

2.1 Abstract

During antiretroviral therapy (ART), drug resistance mutations (DRMs) are selected in the active sites of HIV enzymes due to their ability to alter the enzyme's structure and hinder inhibitor binding. However, while providing HIV with inhibitor resistance, DRMs can come at the cost of viral fitness, given that mutational changes impeding drug binding often destabilize enzyme structure and impair enzymatic function. To counteract this deleterious effect and improve viral fitness, compensatory mutations are selected elsewhere within the same HIV enzyme or in a different HIV protein. These compensatory mutations, not directed at the inhibitor mutations *per se*, could lead to protein or viral alterations that mitigate the impact of DRMs. Driven by continuous resistance selection and viral fitness improvement, DRMs and compensatory mutations coevolve within the HIV genome during ART. However, our understanding of correlated DRMs and compensatory mutations as well as their contribution to HIV fitness remains incomplete. This is due to the lack of appropriate long-read sequencing methodologies and matched downstream analyzing pipelines.

Back in 2021, the Torbett lab developed a Nanopore-based, long-read sequencing pipeline, named Multi-read Hairpin Mediated Error-correction Reaction (MrHAMER). It could generate accurate sequences of the *gag-pol* region in individual HIVs in serum samples. In my study, I assessed MrHAMER's sensitivity and reliability in identifying linked mutations in HIV genomes. To achieve this, I utilized MrHAMER to sequence synthetic RNA libraries consisting

of mixed HIV genomes with different mutation pairs inserted. MrHAMER demonstrated its sensitivity in detecting rare mutation pairs with a frequency as low as 1% while maintaining a favorable signal-to-noise ratio. In addition, MrHAMER preserved the original proportion of different mutation pairs in high-complexity viral populations.

With MrHAMER sequencing methodology validated, I sequenced and analyzed longitudinal serum samples collected from one person living with HIV (PWH, Patient Identification: 3JQ) who failed ART. From the sequencing data, I identified 25 *gag-pol* mutations that were enriched during ART failures. These enriched mutations contained a canonical reverse transcriptase (RT) DRM, RT M184I. Further analysis indicated that, rather than being individually selected in different HIV genomes, these 25 mutations were collectively enriched in the same HIV genome. And this 25-aa mutational pattern was enriched to 53.4% in the virological-failure sample. Additionally, I identified a rare HIV genome in the drug-naive sample that carried this 25-aa mutational pattern.

To summarize, by sequencing longitudinal serum samples collected from one PWH who failed ART, I revealed the selection of a pre-existing HIV with a drug-resistant *gag-pol* mutational pattern during ART failures.

2.2 RESULTS

2.2.1 MrHAMER Detects Linked Mutations with High Sensitivity and Preserved Original Proportions

MrAHMER sequencing methodology was designed to acquire accurate *gag-pol* sequences of individual HIVs present in the viral swarm[25]. In this study, I assessed MrHAMER's ability in identifying linked mutations in the HIV genome, both in terms of sensitivity and reliability.

To achieve this, I designed two synthetic RNA libraries for testing[25]. The synthetic RNA libraries consisted of the "wild-type" HIV RNA and "mutant" HIV RNAs. To generate the "wild-type" HIV RNA, a non-infectious HIV strain, pSG3.1-ΔEnv-D25A[38], was used as the template ("wild-type"). The "wild-type" RNAs were *in vitro* reverse transcribed from the *gag-pol* region of the template. To generate "mutant" HIV RNAs, different "mutant" constructs were generated by adding unique mutation pairs to the *gag-pol* region of the template. All built-in mutations were selected from DRMs listed in the HIV drug resistance database (HIVDB)[39], to mimic HIV sequences that were observed in the clinic. The "mutant" RNAs were *in vitro* reverse transcribed from the *gag-pol* region of each "mutant" construct.

To evaluate the sensitivity of MrHAMER, I prepared the first library by mixing the "wild-type" RNA and one "mutant" RNA in a ratio of 99:1 (Figure 2.1A). The "mutant" RNA sequence had a pair of built-in SNVs, A1597C and A4106G. This created a pair of mutations with a frequency of 1% in the mixed genomes. Sequencing of this library using MrHAMER generated 458 *gag-pol* reads. Close to the input ratio, 0.66% of the sequencing reads had the built-in mutation pair. To determine whether the built-in mutation pair could be identified from the background

noise, I calculated the linkage disequilibrium in the sequencing reads using CoVaMa[34]. CoVaMa detected the built-in mutation pair with an outstanding R-squared value of around 0.8 (Figure 2.1B). To summarize, MrHAMER demonstrated its sensitivity in detecting rare mutation pairs in viral populations, even at frequencies as low as 1%, while maintaining a favorable signal-to-noise ratio.

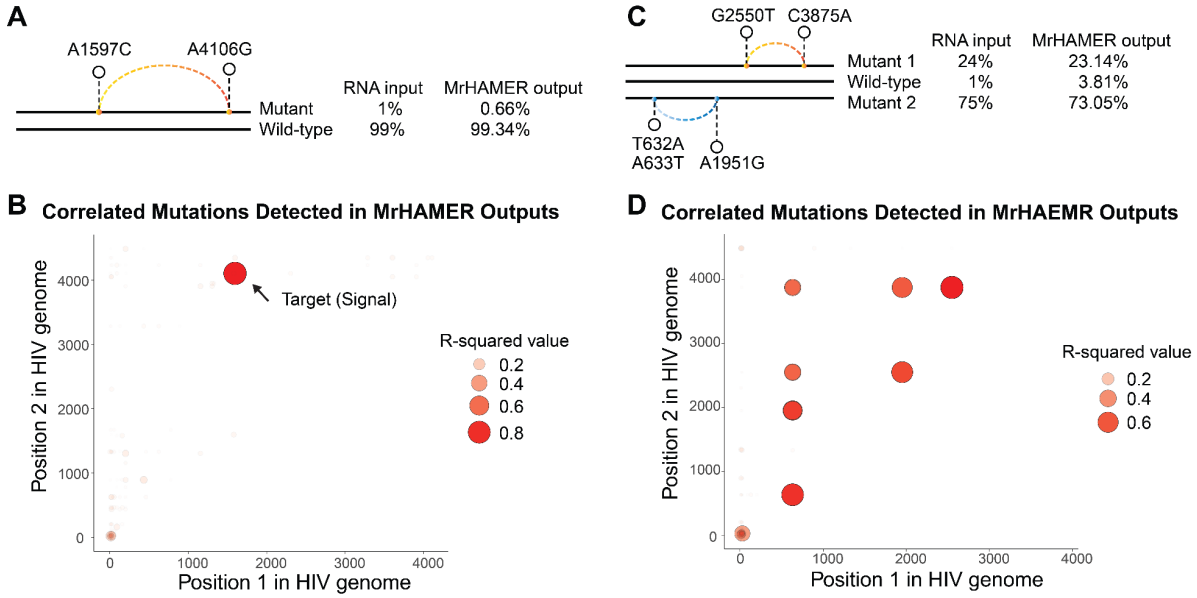

Figure 2.1. MrHAMER detects linked mutations in the HIV *gag-pol* with high sensitivity and preserves the original proportion of different mutation pairs in complex synthetic libraries. (A) The first synthetic library is generated by mixing the "wild-type" RNA and one "mutant" RNA in a ratio of 99:1. The "mutant" RNA is reverse transcribed from the HIV *gag-pol* region with two SNVs, A1597C and A4106G. After sequencing, the built-in mutation pair is detected in 0.66% of the reads. (B) Additionally, CoVaMa detects the built-in mutation pair with an outstanding R-squared value of around 0.8. (C) In the second synthetic library, Mutant 1 RNA, "wild-type" RNA, and Mutant 2 RNA are mixed in a ratio of 24:1:75. Each mutant sequence contains its specific built-in mutation pairs (Mutant 1: G2550T and C3875A; Mutant 2: T632A/A633G and A1951G). After sequencing, the three sequences are detected in proportions closely mirroring the original

mixture. (D) Additionally, CoVaMa detects all mutation pairs with R-squared values that are significantly higher than the background noise level.

To evaluate whether MrHAMER could maintain the original proportion of different mutation pairs in a viral population, I prepared the second library using the "wild-type" RNA and two distinct "mutant" RNAs, named Mutant 1 and Mutant 2 (Figure 2.1C). The Mutant 1 sequence had a pair of SNVs, G2550T and C3875A. And the Mutant 2 sequence had T632A/A633G and A1951G. The Mutant 1, "wild-type", and Mutant 2 RNAs were mixed in a ratio of 24:1:75 in the 200,000-RNA library. This created mutation pairs with different abundances in the mixed genomes. In the sequencing output of this library, all mutation pairs were detected. In addition, the three mutation pairs were detected in a ratio close to the original proportion. The "wild-type" RNA was slightly over-represented (3.81% vs. 1%). Moreover, CoVaMa detected all mutation pairs with R-squared values that were significantly higher than the background noise (Figure 2.1D). This result was validated using one more technical repeat. To summarize, I confirmed MrHAMER's ability in preserving the original proportion of different mutation pairs in high-complexity viral populations. This indicated that MrHAMER could be used to identify linked mutations that are selected in the HIV *gag-pol* region during ART failures.

2.2.2 The Enrichment of a Pre-ART, Drug-Resistant HIV gag-pol Mutational Pattern During ART Failures in one PWH Revealed by Long-read Sequencing

Using the validated MrHAMER sequencing methodology, I sequenced and analyzed longitudinal serum samples collected from one PWH (Patient Identification: 3JQ) who failed ART, aiming to identify linked mutations that were associated with ART failures. This PWH failed two rounds of ART over 6 years. The first round of ART consisted of one Protease Inhibitor (PI), one non-nucleoside reverse transcriptase inhibitor (NNRTI), and two nucleoside reverse transcriptase inhibitors (NRTIs). The second round of ART contained NRTIs only. I sequenced one drug-naive sample collected before the therapy initiation and one virological-failure sample collected after ART failures (Figure 2.2A). After sequencing, I acquired 1002 individual HIV *gag-pol* genomes from the drug-naive sample and 1270 individual HIV *gag-pol* genomes from the virological-failure sample.

To reveal the viral sequence diversity in the HIV population during ART, I identified viral haplotypes in the drug-naive sample and the virological-failure sample using CliqueSNV[40]. CliqueSNV revealed 38 haplotypes in the drug-naive sample, none of which had an abundance of over 6%. On the contrary, in the virological-failure sample, only 12 haplotypes were detected and the predominant one was enriched to 35% (Figure 2.2B). The reduction in the sequence diversity and the selection of the predominant haplotype in the viral population was consistent with the strong selection sweep mediated by anti-viral inhibitors.

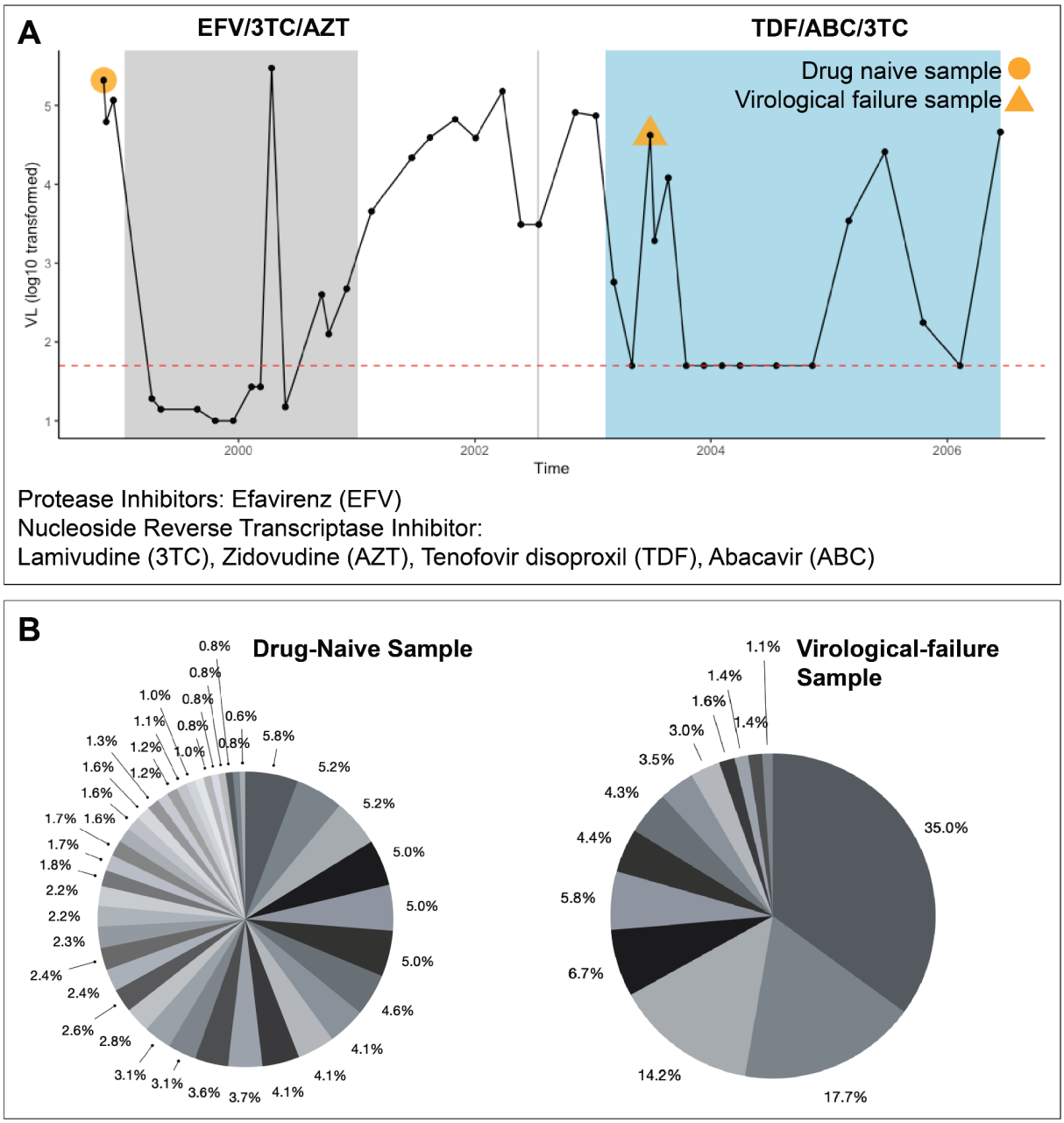

Figure 2.2. Diversity reduction is detected in the HIV population during ART failures. (A) Changes in HIV viral load (copies/ml) during ART failures in one PWH. Gray and blue shading highlights two therapy periods with annotated inhibitors. The drug-naive

sample is marked using the orange circle, and the virological-failure sample is marked using the orange triangle. The red dashed line represents a viral load of 50 copies/ml. (B) The 38 haplotypes detected in the drug-naive sample (left) and the 12 haplotypes detected in the virological-failure (right) sample are shown with sector diagrams based on haplotype abundance. ART: antiretroviral therapy.

To uncover the genetic changes (mutations) within the HIV genomes during ART failures, I compared the HIV genomes from the virological-failure sample to those from the drug-naive sample. I created a drug-naive consensus sequence using HIV genomes from the drug-naive sample and aligned the virological-failure sample's HIV genomes to this consensus sequence. The alignment revealed 25 positions across the HIV *gag-pol* region where mutations were selected and enriched during ART failures (Table 2.1). One of these was a DRM in the reverse transcriptase, RT M184I[41] (Figure 2.3A). In contrast, these 25 mutations were either rarely detected or absent in the drug-naive sample.

To determine the mutation combinations at these 25 positions within individual HIVs, I developed a *Python* script named Linked Mutation Extractor (LiME)[25]. LiME identified different nucleotide combinations at a series of positions in the viral genome and grouped viral genomes based on their unique nucleotide combinations. Using LiME, I quantified the nucleotide combinations at these 25 positions within the virological-failure sample. Rather than being individually selected in different genomes, these 25 mutations were collectively enriched in the same genome, accounting for 53.4% of the virological-failure sample’s viral genomes (Figure 2.3B). Additionally, I quantified the nucleotide combinations at these 25 positions in the drug-naive sample. A founder species with a frequency of ~0.1% was detected in the drug-naive

sample, carrying the 25-aa mutational pattern that later became enriched during ART (Figure 2.3B).

Table 2.1. The 25 SNVs that are enriched during ART failures.

Nucleotide	Type	Mutation	Protein	Domain
G816C	Non-synonymous	R9S	Gag	p17
A867G	Synonymous	-	Gag	p17
A1395G	Synonymous	-	Gag	p24
A1497T	Synonymous	-	Gag	p24
A1514C	Non-synonymous	N242T	Gag	p24
C1572T	Synonymous	-	Gag	p24
A1909G	Non-synonymous	I374A	Gag	p2
C1990A	Non-synonymous	L401I	Gag	p7
T2064C	Synonymous	-	Gag	p7
A2286G	Non-synonymous	T12A	protease	-
C2319T	Synonymous	-	protease	-
C2357A	Non-synonymous	D35E	protease	-
A2375G	Synonymous	-	protease	-
A2436G	Non-synonymous	I62V	protease	-
G2522A	Synonymous	-	protease	-
G2913A	Non-synonymous	E122K	RT	-
T2953C	Non-synonymous	I135T	RT	-
G3101A	Non-synonymous	M184I	RT	-
T3257G	Synonymous	-	RT	-
C3439A	Non-synonymous	A297E	RT	-
G3672A	Non-synonymous	V375I	RT	-
A3935C	Synonymous	-	RNAseH	-
T4349C	Synonymous	-	IN	-
A4379G	Non-synonymous	I50M	IN	-
G4405A	Non-synonymous	G59E	IN	-

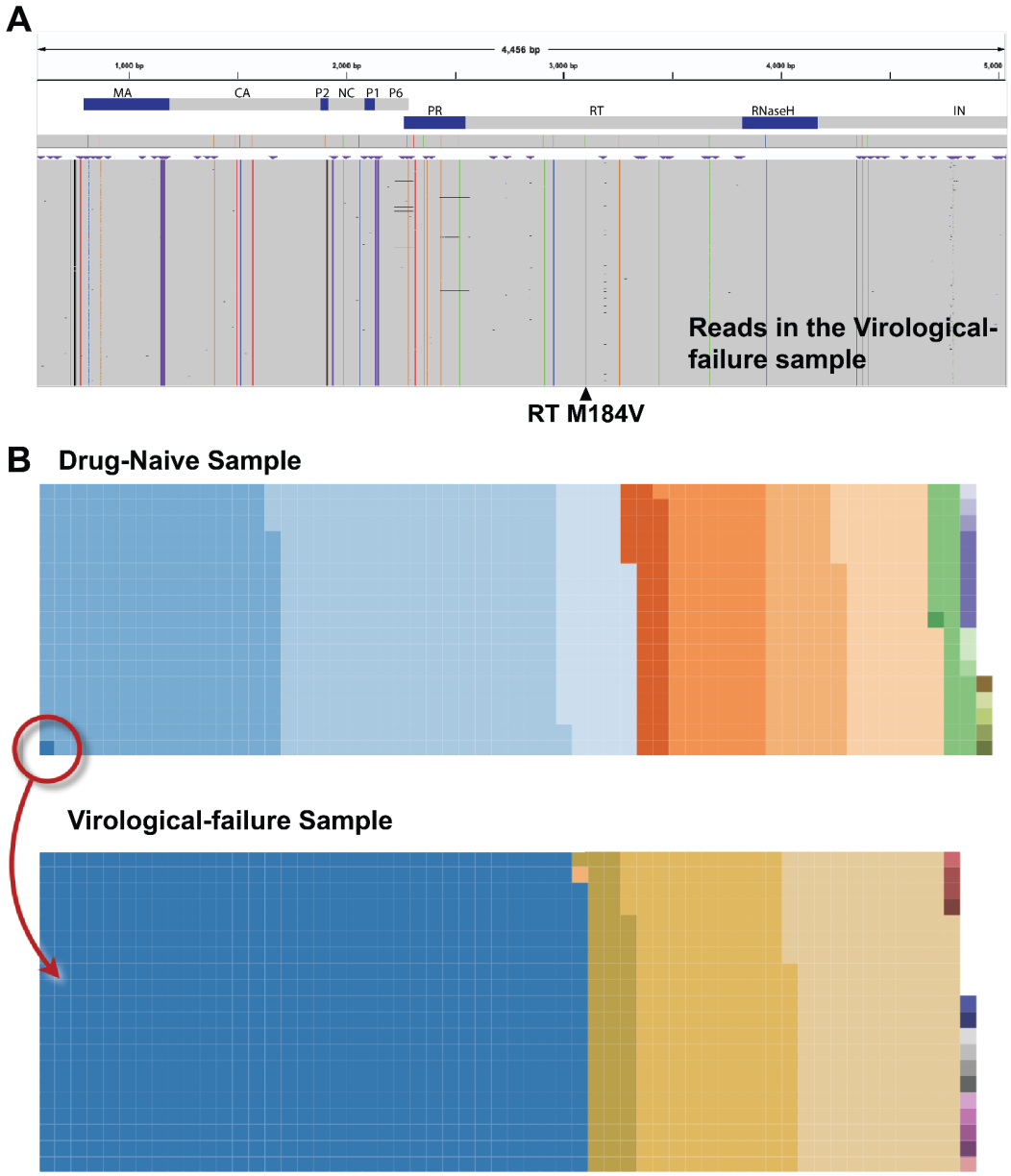

Figure 2.3. Sequencing of longitudinal serum samples from one PWH experiencing ART failures reveals the enrichment of a pre-ART, drug-resistant *gag-pol* mutational pattern. (A) HIV genomes in the virological-failure sample are aligned to the drug-naive consensus sequence and visualized using Integrative Genomics Viewer (IGV)[42]. Alignment reveals a total of 25 positions across the HIV *gag-pol* region where mutations are selected and enriched in the virological-failure sample. These 25 positions are

highlighted by vertical lines in different colors. Additionally, the DRM RT M184V is marked using a black arrowhead. RT: reverse transcriptase. (B) The nucleotide combinations at the 25 *gag-pol* positions in individual HIV genomes from the drug-naive sample (top) and the virological failure sample (bottom) are shown using waffle plots. In the waffle plots, each square represents one HIV genome. HIV genomes with the same nucleotide combination are colored in the same color, and HIV genomes with distinct nucleotide combinations are colored in different colors. The red arrow indicates a rare nucleotide combination from the drug-naive sample (top) that is enriched in HIV genomes in the virological failure sample (bottom). DRM: drug resistance mutation. PWH: person living with HIV. ART: antiretroviral therapy.

2.3 Discussion

Correlated DRMs and compensatory mutations selected in the HIV *gag-pol* region during ART provide HIV with enhanced resistance and replication fitness[10,12,13,43]. These correlated mutations, distributed either within the same viral protein (protease, reverse transcriptase) or between interacting proteins (Gag polyprotein and protease)[44,45], affect protein-protein interactions, play essential roles in the viral life cycle, and have been targeted for anti-viral inhibitor development. Understanding the correlated mutations that arose within the HIV genome would provide insights into the evolutionary strategies of HIV during ART failures. However, previous studies on correlated mutations were constrained by the short read length of deep-sequencing techniques, often relying on bioinformatic inference or statistical analysis[14,34].

In this study, I assessed the ability of MrHAMER, a long-read sequencing methodology, in identifying linked mutations within the HIV *gag-pol* region. MrHAMER exhibited high sensitivity, detecting even rare mutation pairs with frequencies as low as 1%. This capability is crucial for identifying infrequent HIV variants in the viral population, which could impact the outcomes of subsequent ART. Furthermore, I confirmed that MrHAMER accurately maintains the original proportions of different mutation pairs in complex viral populations. Consequently, by analyzing the abundance of various HIV genomes in the sequencing outputs, we can infer their relative fitness within the viral population.

With MrHAMER validated, I applied it to sequence longitudinal serum samples from one PWH experiencing ART failures, aiming to identify linked mutations associated with ART failures. MrHAMER provided accurate sequences of the ~4.5kb *gag-pol* region in individual HIVs. From the sequencing data, I detected 25 mutations that were enriched across the HIV *gag-pol* region

during ART. Analyzing individual sequencing reads confirmed that these 25 mutations were collectively enriched in the same HIV genome.

One known DRM, RT M184V, was included in these 25 mutations. Previous studies have shown that, while providing HIV with high resistance against NRTIs, RT M184V would significantly reduce the HIV replication capacity[46]. Compensatory mutations often arise in the HIV genomes, compensating for the loss in the viral replication capacity. In this study, I observed the co-occurrence of RT M184V and 24 other mutations across the HIV *gag-pol* region. These 24 mutations were rarely or not detectable pre-ART but became enriched together during ART failures. This suggests they likely compensate for the reduced fitness caused by RT M184V. The effects of these mutations need to be validated using *in vitro* fitness assays. Alternatively, using more samples from different PWHs treated with the same drug class could help distinguish passenger mutations from significant findings (more details will be provided in Chapter 5).

In the drug-naive viral population of this PWH, I identified a rare founder species with an abundance of ~0.1% that had all the above 25 mutations. This indicated that this founder species was presumably selected during ART and likely contributed to ART failures. Notably, this founder species had not been detected in the genotypic drug resistance testing prior to ART initiation due to its low abundance[47]. This highlighted MrHAMER's ability to identify infrequent HIV strains, which could later contribute to enhanced resistance and viral fitness, ultimately resulting in suboptimal treatment outcomes.

For this analysis, I utilized one drug-naive sample and one virological-failure sample to illustrate the evolution of HIV during ART failures. Having more intermediate samples will provide us with more details of the HIV evolution during ART. More analyses and methods are provided in

the next Chapter, which focus on tracing the evolution of individual HIV genomes during ART failures.

To identify co-occurring mutations in individual HIV genomes, I provided LiME, a *Python* script that was designed to identify and quantify mutational patterns in long-read sequencing datasets[25]. In this study, LiME identified 25 co-occurring mutations across the HIV Gag polyprotein, protease, reverse transcriptase, and integrase. These co-occurring mutations might be involved in protein-protein interactions and compensation mechanisms. Moreover, LiME could be used to reveal sequence diversity in viral samples. Consistent with the CliqueSNV outputs, LiME detected more *gag-pol* mutational patterns in the drug-naive sample and fewer mutational patterns after ART failure, with predominant mutational patterns detected. LiME's optimization facilitates the simultaneous identification of nucleotide combinations across tens of positions in thousands of reads within seconds. Its efficiency surpasses that of haplotype construction and enables users to target specific positions or regions of interest. In summary, this simple and convenient tool can be used in diverse sequencing datasets and can compensate for variant calling tools and linkage disequilibrium calculation tools in studying the polymorphism in viral and other genomes.

In this study, I validated MrHAMER's sensitivity and reliability in identifying linked mutations within the HIV genome. I showed its application by identifying co-occurring mutations across the HIV *gag-pol* region that were selected during ART failures in one PWH. Furthermore, the ability of MrHAMER to analyze higher-order linked mutations would benefit the identification of complex evolutionary pathways[48–50]. And this method can be adapted for identifying correlated mutations in other viruses or bacteria[32,51].

2.4 Methods

Generation of plasmids for in vitro transcription of HIV RNAs.

To generate a plasmid for *in vitro* transcription of HIV RNAs, the HIV insert from the pSG3.1 strain was PCR amplified with Q5 HotStart Master Mix (NEB M0494S) in two fragments, with an overlap in the protease locus to add a D25A mutation and both an EcoRI/T7 promoter and PolyA/BamHI sites at the 5′ and 3′ ends of the insert. A pUC19 backbone was PCR amplified with overlaps to the T7 promoter site at the 5′ of insert and PolyA tail at the 3′ ends. PCR-amplified insert and vector fragments were assembled with the NEBuilder HiFi DNA Assembly kit (NEB E2621S) and plated onto LB-Amp. Single colonies were grown, mini-prepped, and sequenced to verify the plasmid identity and orientation of all fragments. For nomenclature purposes, this sequence is referred to as a 'wild-type' strain throughout. Additional mutants were designed based on this SG3.1 'wild-type' background, containing one of the following mutation pairs: T632A/A633G with A1951G, G2550T with C3875A, A1597C with A4106G. These mutations were added via PCR amplification with primers that generated overlaps for subsequent NEBuilder HiFi assembly. All modified plasmids containing mutation pairs were grown from single colonies and sequenced to verify sequence identity.

In vitro transcription of HIV RNAs.

HIV plasmid was treated with T5 exonuclease (NEB M0363S) to digest any fragmented vector, and DNA was cleaned with Monarch PCR & DNA Cleanup Kit (NEB T1030S). The resulting supercoiled plasmid was linearized at the 3′ ends of the PolyA tail using BamHI-HF (NEB R3136S) and checked for reaction completion by running on an agarose gel. The linearized

plasmid was DNA-cleaned and eluted in nuclease-free water. Standard RNA Synthesis was carried out with the HiScribe T7 High Yield RNA Synthesis kit (NEB E2040S) for 1.5 hours according to the manufacturer's instructions, using 500–1000 ng of linearized plasmid as input, followed by DNase I digestion as instructed. RNA was purified using RNA Clean & Concentrator–5 kits (Zymo Research R1013) and eluted in nuclease-free water. RNA samples were serially diluted in order to arrive at the desired number of input RNA molecules. When using a complex mixture of samples, RNA species were first mixed at a high concentration at the right proportion, followed by serial dilutions to the appropriate RNA molecule number.

Sequencing library preparation.

For the sequencing library preparation process, including reverse transcription, emulsion PCRs, MrHAMER template preparation, and generating concatemers from the MrHAMER templates, please review the previous publication[25].

Nanopore sequencing.

Sequencing libraries were prepared using the Ligation Sequencing Kit (SQK-LSK109). All samples were sequenced with MinION R9.4.1 flowcells, basecalled with Guppy basecaller 3.6.0. Quality Control was performed on reads using the NanoPlot package.

Bioinformatics.

For the downstream bioinformatics analyzing process, including basecalling, splitting long concatemers, and reconstructing genomes, please review the previous publication[25].

Detection of correlated mutations using CoVaMa.

MrHAMER generated FASTA files containing reconstructed genomes at high accuracy (using at least eight repetitive units per genome). For downstream analysis, these reconstructed genomes were mapped to the HIV reference using *minimap2* (default mode). The CoVaMa package consists of two *Python* scripts[34]. The first script is *CoVaMa_Make_Matrices.py*. I ran the first script using the *minimap2* outputs with the following options *–Mode2 Nucs –SAM1 –PileUp_Fraction 0.005 NT*. Since our effective limit of detection of linked mutations was between 0.5 and 1%, the *PileUp_Fraction* was set to 0.005 so that only contingency tables with mutants present at a frequency >0.5% were considered. The first script generated large matrices consisting of populated contingency tables for every pairwise association between SNVs. For the next step, matrices were analyzed using *CoVaMa_Analyze_Matrices.py* with the following options *–Min_Coverage 5 -OutArray -Weighted NT*. Setting the *Min_Coverage* to 5 discarded contingency tables populated with fewer than 5 aligned reads. Results were output in TEXT format, with each row indicating a pair of correlated mutations, their linkage disequilibrium (LD) value, weighted LD value, R-squared value, and the entire contingency table. For result visualization, the top 150 pairs of correlated SNVs with the highest LD values were plotted via *R*.

Viral RNA extraction from samples collected from PWHs.

Frozen plasma aliquots from PWHs experiencing ART failures were obtained from the UC San Diego Primary Infection Resource Consortium (PIRC). From every sample, 1 ml plasma was transferred to 1.5 ml DNA LoBind tubes (Eppendorf), and centrifuged for 75 min at 18 000 x g and 4 °C to concentrate the virus. After centrifugation, 860 µl of supernatant was carefully aspirated from the top of the tube and frozen at −80 °C. The remaining 140 µl of the

concentrated virus was processed according to the QIAamp Viral RNA Mini kit (QIAGEN 52904) protocol, with the only deviation from protocol being viral RNA elution in 30 µl of nuclease-free water. 11 µl of this eluate was used as input for subsequent RT and serial emulsion PCRs as described in the previous publication[25].

Generation of the sample-specific consensus sequence for use as a reference for single molecule reconstruction.

Given high HIV intra-host diversity (both at the SNV and In-Del level), a sample-specific reference was generated for each sample prior to single molecule reconstruction. FASTQ reads generated after demultiplexing with qfilesplitterV3.1.py (each containing at least ten repeating units) are concatenated, mapped to pSG3.1 reference, followed by *racon* and *medaka* sequence correction to generate a sample-specific consensus assembly. This sample-specific consensus assembly is used as a reference for subsequent single molecule error-correction using the protocolV3.3.py script.

Generation of a PWH-specific naive reference using standardized HXB2 coordinates.

For downstream analysis of error-corrected *gag-pol* reads, the drug-naive consensus assembly for the serum sample was mapped to the HXB2 reference sequence (accession number K03455) using MAFFT v7.471 with the following options *–ep 20 –keeplength –addfragments*. This yielded a drug-naive reference sequence that preserved HXB2 ORFs and positional coordinates while taking into account the prevailing genetic background of the drug-naive error-corrected *gag-pol* reads. Error-corrected sequencing reads and haplotype sequences could then be mapped to this newly generated reference to determine patterns of SNVs and insertion/deletion events.

Generation of viral haplotype clusters from high-accuracy gag-pol reads.

High-accuracy *gag-pol* reads from drug-naive and virological-failure (VF) samples were separately mapped to their respective sample-specific consensus sequence using *minimap2*. The resulting drug-naive and VF SAM files were then used as input for CliqueSNV v1.5.4 (https://github.com/vtsyvina/CliqueSNV) with the following options *-m snv-pacbio -rn fdf extended4 -tf 0.001*. The resulting outputs delineated clusters (haplotypes) of *gag-pol* reads with sharing mutation patterns and included the FASTA sequence and percent enrichment of each haplotype cluster.

Waffle plot analysis (LiME + Waffle plots).

To better understand the possible evolutionary pathways taken by individual HIV genomes during ART, Linked-Mutation Extractor (LiME), a custom *Python* script, was developed to acquire the higher-order linkage information from sequencing reads. LiME was designed to classify long sequences into several groups based on their unique combination of nucleotides at a series of positions (Genetic Patterns). LiME took a BAM file containing the aligned target sequences and a list of positions. In order to accelerate the analyzing process, the list of positions given was first chucked into several smaller subgroups with five or less than five positions. Genetic pattern identification and read classification were done independently in each subgroup, after which all genetic patterns belonging to the same read were collected and combined, forming an entire pattern covering all the given positions. LiME generated a waffle plot using the *PyWaffle Python* package and stored the analyzing results in a CSV file, each line containing a genetic pattern observed, the read count with this pattern, and ID of those reads.

Chapter 3. COMPLEX MUTATIONAL PATTERNS THAT DRIVE HIV RESISTANCE DEVELOPMENT REVEALED USING BIOINFORMATICS ANALYSIS TOOLS

3.1 ABSTRACT

Correlated drug resistance mutations (DRMs) and compensatory mutations selected in the HIV *gag-pol* region provide HIV with enhanced resistance and replication fitness during Antiretroviral Therapy (ART). Among these correlated mutations, DRMs grant HIV with inhibitor resistance but at the cost of enzyme stability and enzymatic function. This fitness reduction is mitigated by the selection of compensatory mutations, either within the same HIV enzyme or in a different HIV protein. To better understand the evolutionary strategies taken by HIVs during drug-mediated selection, it is essential to investigate the sequential order in which these correlated mutations arise in individual HIVs. Furthermore, identifying mutational patterns selected in the HIV genome under drug pressure is critical.

In the previous Chapter, I demonstrated the application of MrHAMER, a novel long-read sequencing methodology. MrHAMER provided accurate *gag-pol* sequences from individual HIVs present in serum samples from people living with HIV (PWHs) experiencing ART failures. Despite this advancement, we still lack bioinformatic tools to reveal evolutionary pathways taken by individual HIVs during ART failures and identify mutational patterns associated with resistance development.

In response to this need, I utilized the Hamming-distance-based phylogenetic analysis (HDBPA). This method was designed to handle longitudinal samples sequenced using long-read sequencing

technologies, e.g., MrHAMER. To show this method's application for analyzing different HIV regions that have evolved under distinct drug classes, I described the analyses of ART failures in two PWHs using the HDBPA. One PWH received a Protease Inhibitor (PI)-based therapy; the other PWH received Nucleoside Reverse Transcriptase Inhibitors (NRTIs) therapies.

From the PWH who failed a PI-based therapy, I analyzed one pre-treatment serum sample and one post-treatment serum sample using the HDBPA. The findings revealed two predominant subpopulations selected during this ART failure. Additionally, the results indicated that these two subpopulations originated from distinct Most Recent Common Ancestors (MRCAs) and acquired different sets of mutations under drug pressure, indicating their different evolutionary strategies in response to the treatment.

The other PWH failed two successive dual-NRTI therapies. Analyzing longitudinal serum samples collected during therapy failures revealed a clear pattern. It showed that the resistance development in individual HIVs in this viral swarm mainly followed one of three distinct evolutionary pathways. Each pathway was characterized by a unique combination of DRMs and compensatory mutations in the HIV reverse transcriptase.

Besides HIV, HDBPA could be applied to study the evolution of other viruses and bacteria. This analysis pipeline/methodology is being formatted into a method manuscript, and the methodology will be submitted to GitHub for public use.

3.2 RESULTS

3.2.1 HDBPA Reveals the Evolution of Individual HIVs Druing ART Failures

Phylogenetic trees have been widely used to correlate sequences from different time points and reveal genetic changes over time[52]. Phylogenetic tree construction usually relies on consensus sequences. It faces challenges while dealing with large numbers of genomes generated by long-read sequencing methodologies like MrHAMER, which generated thousands of individual HIV *gag-pol* genomes from each serum sample[25]. Therefore, traditional tree construction methods are incompatible with the sequencing depth of MrHAMER and other long-read sequencing methodologies. Moreover, conventional phylogenetic tree construction ignores the abundance of individual viral genomes in the viral swarm and lacks a direct representation of mutational changes between Most Recent Common Ancestors (MRCAs) and their descendants.

To overcome these challenges, I utilized Hamming distance to reveal the relationship between individual HIV genomes. I used HDBPA, a method tailored for analyzing HIV genomes collected over time and sequenced using MrHAMER[53] (Figure 3.1B). Considering aligned nucleotide sequences or amino acid sequences as strings of characters, the Hamming distance between two sequences is the number of positions where their characters do not match. From an evolutionary perspective, it represents the minimum number of mutations required for one sequence to evolve into another sequence[53].

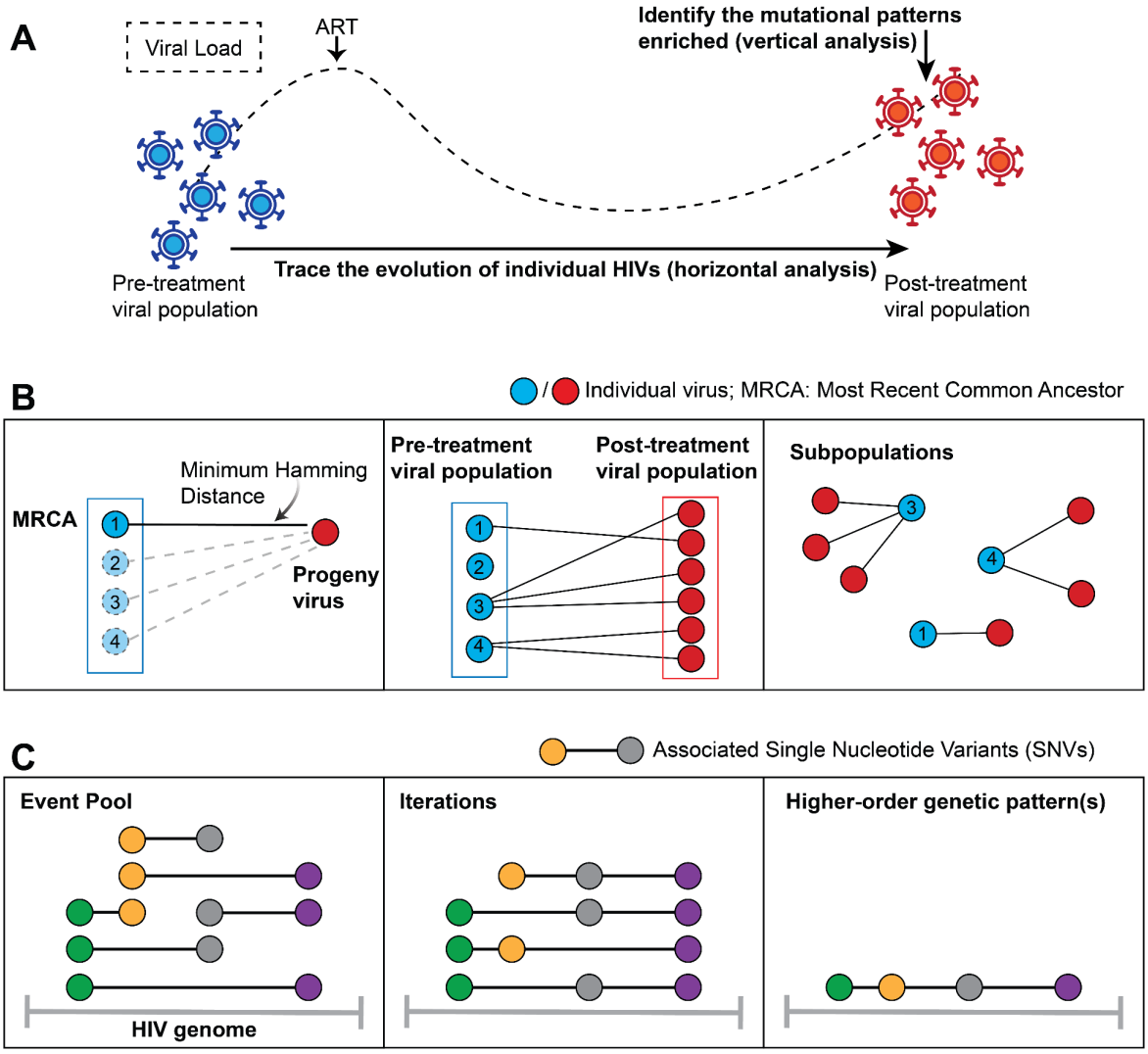

Figure 3.1. HDBPA reveals the evolution of individual HIVs and the linkage disequilibrium-based analysis reveals mutational patterns from sequencing data. (A) Longitudinal serum samples are subjected to long-read sequencing, after which two distinct bioinformatic analyses are conducted. The horizontal analysis tracks the evolutionary trajectories of individual HIVs during ART using the Hamming-distance-based phylogenetic analysis (HDBPA), while the vertical analysis identifies mutational patterns enriched in the viral population at specific time points, such as after ART failure. (B) HDBPA correlates each HIV genome to its Most

Recent Common Ancestor (MRCA) from a prior time point. For instance, each post-treatment HIV genome is linked to its corresponding MCRA in the pre-treatment viral population. The post-treatment HIV genomes derived from the same MRCA are grouped into one subpopulation. (C) Constructing Higher-Order Mutational Patterns. Pairwise associated SNVs are identified using CoVaMa. These SNVs become starting events in an event pool. Through iterative fusion, pairs of events are combined if all their SNVs are pairwise associated. This process continues until no new fused events can be formed, generating all possible higher-order mutational patterns. ART: antiretroviral therapy. SNV: Single-Nucleotide Variants. CoVaMa: Co-Variation Mapper.

HDBPA required at least two HIV samples sequenced at different time points (Figure 3.1A, Figure 3.1B). After sequencing, individual HIV genomes generated were aligned to the HIV HXB2 reference sequence. This ensured that all HIV genomes would have the standardized HXB2 coordinates. After this, HDBPA correlated each aligned HIV genome in the latter time point to its MRCA detected at a prior time point based on the minimum Hamming distance[53]. In this way, HIV genomes detected at different sampling time points could be correlated based on their sequence similarity. Additionally, HIV genomes derived from the same MRCA were clustered into one subpopulation. The abundance of each subpopulation, relative to the total viral genomes, indicated its relative fitness under drug pressure during the observation period. In this way, HDBPA revealed the predominant subpopulations selected.

Unlike the HDBPA, which provided a time-based (horizontal) analysis of the viral evolution, a vertical analysis targeted a single sample to uncover enriched mutational patterns at that time point (Figure 3.1A). Mutational patterns selected during ART provide HIV with enhanced viral

fitness under drug pressure. Higher-order mutational patterns contain several, i.e., ≥ 2, mutations, which could be located across different Open-Reading Frames (ORFs) in the HIV genome. Hence, mutational pattern identification exceeds the capabilities of CoVaMa and other pairwise linkage disequilibrium (LD) calculation methods[14,34]. To address this, additional modifications were implemented after acquiring pairwise association information using CoVaMa (Figure 3.1C). First, significant pairwise associated SNVs (with LD values ≥ 3 sigma) generated by CoVaMa were collected as starting points. Then, every two events were fused if all SNVs involved were significantly pairwise associated. For instance, association among SNV A, B, and C could only be established if significant association was detected between A and B, between A and C, and between B and C. These newly formed fused events were added to the event pool. This step would be repeated until no new events could be formed. In the end, this generated all possible higher-order mutational patterns. The abundance of each higher-order genetic pattern formed could be calculated using the Linked-Mutation Extractor as described in Chapter 2. This pipeline was accomplished using a custom *Python* script.

3.2.2 Identification of Enriched Mutational Patterns and Uncovering the Order of Mutation Development in Two PWHs During ART Failures

I analyzed the HIV evolution during ART failures in two PWHs using the HDBPA and showed its applicability in investigating the evolution of individual HIVs under drug pressure. The first PWH failed a PI-based therapy. Thus, analysis of this PWH considered the HIV *gag-pol* region but with an emphasis on the HIV *protease* region. The other PWH failed dual-NRTI therapies,

and the analysis focused on the HIV RT region. Together, these two analyses showed that the HDBPA could be used to study various HIV regions that have evolved under different drug pressures.

The first PWH (PID: 3IB) failed a PI-based therapy containing the PI Nelfinavir (NFV) and NRTIs. From this PWH, two longitudinal serum samples were collected and sequenced using MrHAMER: one was collected before the therapy initiation (the drug-naive sample); the other was collected after the therapy failure (the virological-failure sample) (Figure 3.2A). MrHAMER acquired 1832 individual HIV *gag-pol* genomes from the drug-naive sample and 632 individual HIV *gag-pol* genomes from the virological-failure sample. HDBPA revealed 27 subpopulations that were formed under drug pressure with varying subpopulation sizes. Additionally, two subpopulations substantially expanded and were detected as the predominant subpopulations in the virological-failure sample (Figure 3.2B). These two subpopulations were named after their MRCAs, Subpopulation 608 and Subpopulation 540.

I analyzed the drug-naive MRCAs of these two predominant subpopulations, looking for unique pre-existing mutations. The MRCA of Subpopulation 540 carried transmitted DRMs in the protease (PR): PR D30N, PR V77I, and PR N88D. Previous research has reported that these DRMs could provide HIV with high NFV resistance[54,55]. Thus, MRCA 540 and its progenies had high PI resistance. On the contrary, the MRCA of Subpopulation 608 did not contain any protease DRMs and thus was susceptible to the PI used. Instead, it contained several mutations in CSs of the Gag polyprotein. The cleavage of Gag polyprotein at CSs by HIV protease is an essential step in the viral life cycle[56]. Several CS mutations were identified in the MRCA of Subpopulation 608, including A374T and T375A in the P2/NC and S451N in the P1/p6. Among

them, S451N mutation has been reported to improve the cleavage efficiency of protease[11]. Moreover, I compared the MRCA of Subpopulation 608 to other drug-susceptible genomes in the drug-naive sample. None of the other drug-susceptible genomes contained CS mutations or were selected during ART.

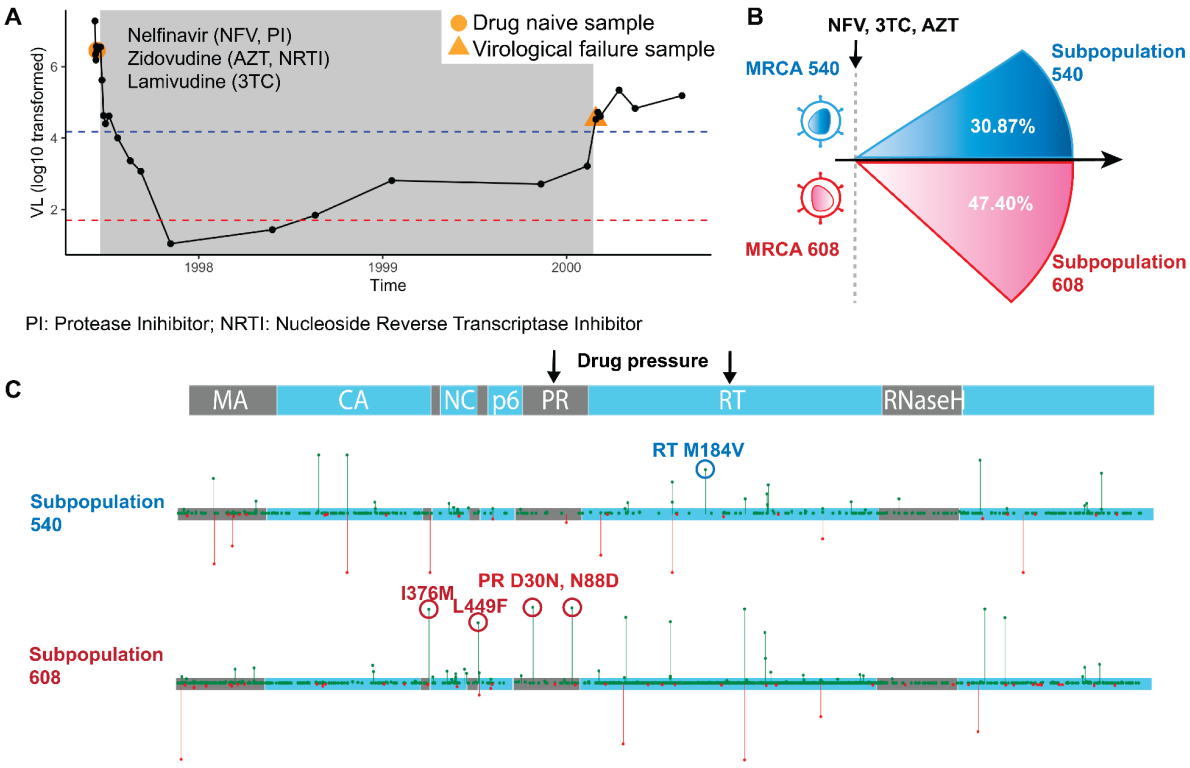

Figure 3.2. Individual HIVs in one PWH follow two distinct evolutionary pathways during the ART failure.

(A) Changes in HIV viral load (copies/ml) during ART failures in one PWH. Gray shading highlights the therapy period with inhibitors annotated. The drug-naive sample is marked using the orange circle, and the virological-failure sample is marked using the orange triangle. The red dashed line represents a viral load of 50 copies/ml. (B) Two major subpopulations emerge during ART failure. They are named after their MRCAs and colored in blue and red. MRCA: Most Recent Common Ancestors. (C) Genetic changes in individual HV genomes within the two predominant subpopulations during

ART. Proteins encoded in the HIV *gag-pol* are annotated on the top. In each subpopulation, the green vertical lines above the HIV genome represent the mutations acquired; the red vertical lines below the HIV genome represent the mutations lost. The height of a vertical line corresponds to the times that a genetic change is detected in each subpopulation. Known DRMs in protease and reverse transcriptase and compensatory mutations in Gag cleavage sites are labeled. ART: antiretroviral therapy. PWH: person living with HIV.

To identify the genetic changes (mutations) occurring in the two predominant subpopulations during ART, I quantified the mutational changes in individual HIV genomes within each subpopulation. The quantification result indicated that distinct genetic changes occurred in HIV genomes in these two subpopulations during ART failure (Figure 3.2C). The HIV genomes in Subpopulation 540 inherited protease DRMs (D30N, V77I, and N88D) from their MRCA. Hence, no further genetic changes were detected in their *protease* region. In this subpopulation, most genetic changes occurred in the capsid (CA), reverse transcriptase (RT), and integrase (IN). The mutations developed in the RT included a canonical DRM, RT M184V, which provided high resistance to the NRTIs used[46]. On the contrary, most HIV genomes in Subpopulation 608 acquired PR D30N and PR N88D under drug pressure. Interestingly, HIV genomes in Subpopulation 608 further developed more CS mutations in the Gag P1/p6 and P2/NC, including I376M in the P2/NC and L449F in P1/p6. According to previous studies, L449F improves the cleavage efficiency of protease with PR D30N and PR N88D[57,58].

Two predominant higher-order mutational patterns, Pattern A and B, were identified in the virological-failure viral population (Table 3.1). Pattern A contained 12 mutations across the HIV *gag-pol* region and was enriched in Subpopulation 608 but not in Subpopulation 540. On the

contrary, Pattern B contained 12 mutations and was enriched in Subpopulation 540 but not in Subpopulation 608. Compared to Pattern A, Pattern B contained more DRM mutations, including PR V77I and RT M184V. On the contrary, Pattern A contained several unique Gag CS mutations, including A374T, T375A, and I376M in the P2/NC, and L449F and S541N in the P1/p6. L449F and S541N have been reported to increase the cleavage efficiency of both wild-type protease and mutant protease with PR D30N and PR N88D[11]. Although no canonical RT DRM was detected in Pattern A, several other RT mutations developed together under RT inhibitor pressure. In summary, these two subpopulations, originating from distinct MRCAs, followed divergent evolutionary trajectories, selecting distinct mutational patterns that enhanced drug resistance and viral fitness during ART.

Table 3.1. Two predominant mutational patterns are identified in the virological-failure sample.

Region in HIV	**Pattern A**	**Pattern B**
MA		K28Q, S54P, V82I
P2	A374T*, T375A*, I376M*	
p6	L449F*, S451N*	S451I*
PR	D30N◇, N88D◇	D30N◇, V77I◇, N88D◇
RT	T69N, I135M, K275Q, R277K	I135R, M184V◇
IN	K71R	V31I, E35Q, K211S

* Cleavage site (CS) mutations; ◇ Drug resistance mutations (DRMs)
MA: matrix protein; PR: protease; RT: reverse transcriptase; IN: integrase

The other PWH (PID:1008) failed two consecutive dual-NRTI therapies (Figure 3.3A). The first therapy consisted of Stavudine (D4T) and Didanosine (DDI), while the second therapy comprised Stavudine (D4T) and Lamivudine (3TC). To trace the evolution of HIV over time, I sequenced one drug-naive sample and three virological-failure samples: one was collected after the first therapy's failure, one was collected ~2 weeks after the second therapy started, and a 3rd sample was collected after the failure of the second therapy (Figure 3.3A). Using the HDBPA, I identified the evolutionary tracks of individual HIVs during ART failures. Notably, these evolutionary tracks could be grouped into three major evolutionary pathways. All three pathways started with a pre-existing DRM, RT M41L, and developed a Thymidine Analog Mutation (TAM) T215Y during the first therapy[59]. The combination of RT M41L and RT T215Y increased the viral resistance against D4T and DDI to an intermediate/high level[39]. Other than RT M41L and RT T215Y, these three pathways were characterized by unique combinations of RT DRMs (Figure 3.3B):

1) Pathway A: During the first therapy, HIV genomes in this pathway developed the accessory NRTI-selected mutation, RT V75T[60]. About two weeks into the second therapy, these viral genomes developed another TAM, RT L210W. The combined mutations M41L-V75T-L210W-T215Y resulted in a total resistance mutation score of 160 against D4T, according to HIVDB. Compared to other HIV genomes within the same viral population, those with this specific mutation combination exhibited the highest resistance

to D4T. Consistent with this, two weeks after the second therapy started, the viral genome in this pathway made up around 45% of the viral population. However, all HIV genomes in this pathway failed to develop RT M184V and remained susceptible to the 3TC, the other drug used in the second therapy[46]. Consequently, the HIV genomes within this pathway could not survive through continuous 3TC pressure, and their proportion decreased to a mere 0.5% after the second therapy's failure.

2) Pathway B: During the first therapy, viral genomes in this pathway developed RT V75T as well. About two weeks into the second therapy, these genomes additionally acquired the drug-resistant mutation M184V, which imparts high resistance to 3TC. The combined mutations M41L-V75T-M184V-T215Y gave the virus high resistance to D4T and 3TC. Consequently, the HIV genomes following this pathway were selected during the second therapy, constituting 44% of the viral population after the second therapy's failure.

3) Pathway C: During the first therapy, viral genomes in this pathway developed V75S and L210W. Around two weeks after the second therapy started, these viral genomes developed RT M184V. The combined mutations M41L-V75S-M184V-L210W-T215Y gave the virus high resistance to D4T and 3TC. Consequently, HIV genomes in this pathway were selected during the second therapy, comprising 51% of the viral population after the second therapy's failure.

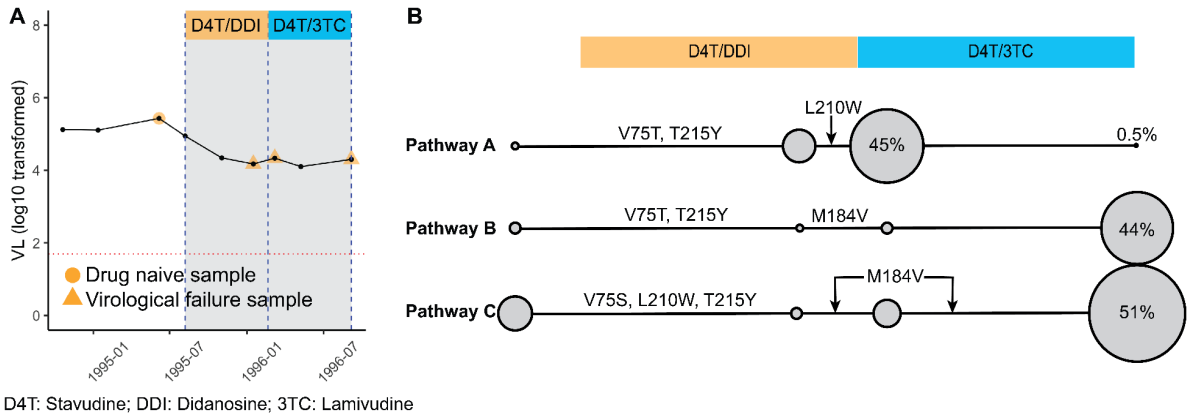

Figure 3.3. Three major evolutionary pathways taken by individual HIVs during therapy failures in one PWH.

(A) Changes in HIV viral load (copies/ml) during therapy failures in one PWH. Gray shading highlights the therapy periods with inhibitors annotated on the top. The start and stop of each therapy are marked using blue vertical dashed lines. The drug-naive sample is marked using the orange circle, and the virological-failure samples are marked using orange triangles. The red dashed line represents a viral load of 50 copies/ml. (B) Three major evolutionary pathways followed by HIV genomes during consecutive therapy failures. The RT DRMs involved in each pathway are annotated. The number of HIV genomes following each pathway is quantified at four sampling time points. This quantitative data is visualized using circles, the sizes of which reflect the relative abundance of HIV genomes following each pathway in relation to the total count of HIV genomes. PWH: person living with HIV; RT: reverse transcriptase.

Besides different combinations of RT DRMs mentioned above, HIV genomes in these three pathways also developed distinct mutations located on positions 53, 293, 297, 334, 335, and 417 in the reverse transcriptase. A significant negative association was detected between RT V75S and RT E53D during the first therapy failure. Most HIV genomes with pre-existing RT E53D

developed RT V75T during the first NRTI therapy and followed Pathway A. On the contrary, RT V75S were acquired mainly in HIV genomes in Pathway C, which did not have the pre-existing RT E53D. After switching to the second NRTI therapy, HIV genomes in the three evolutionary pathways kept the mutations developed in the previous therapy and further acquired different mutations. Sampling 2 weeks after the second therapy started provided a snapshot of the mutational pattern in each pathway (Figure 3.5B):

1) Pathway A: M41L-E53D-V75T-L210W-T215Y-E297S-Q334N-V417I, named Pattern (A).
2) Pathway B: M41L-V75T-M184V-T215Y, named Pattern (B).
3) Pathway C: M41L-V75S-M184V-L210W-T215Y-I293V-E297L-G335D, named Pattern (C).

Analyzing results indicated that HIV genomes with Pattern A failed to develop RT M184V and got eliminated under 3TC pressure. On the contrary, HIV genomes with Pattern B and Pattern C were selected and expanded under 3TC pressure. To summarize, I identified three major evolutionary pathways utilized by HIVs in this PWH during consecutive therapy failures. The HIV genomes following each pathway had a unique combination of RT DRMs and compensatory mutations.

Table 3.2. Genetic patterns formed in each evolutionary pathway two weeks after starting the second therapy.

RT positions	**53**	**75**	**184**	**210**	**293**	**297**	**334**	**335**	**417**
Pattern (A)	D	T◇		W◇		S	N		I

Pattern (B)		T◇	V◇			L			
Pattern (C)		S◇	V◇	W◇	V	L		D	
◇ Drug resistance mutations (DRMs). RT: reverse transcriptase.									

3.3 Discussion

In this study, I utilized the MrHAMER sequencing methodology to acquire individual HIV *gag-pol* genomes from longitudinal serum samples of PWHs[25]. Subsequently, these individual HIV genomes were linked across different time points using the HDBPA. This method allowed for the visualization of the evolutionary trajectories of individual HIVs during ART failure, thus revealing the specific genetic changes (mutations) that occurred during ART and the sequential order in which they arose. Additionally, this method parsed mutations detected in HIV genomes into two groups: 1) mutations inherited from MRCAs; 2) mutations developed *de novo* under drug pressure. The mutations in the former group likely supported viral fitness in the absence of drug pressure. In contrast, those in the latter group were likely associated with enhanced fitness under drug pressure.

Furthermore, by illustrating the evolution of individual HIVs, this approach identified groups of individual HIVs originating from the same MRCA and revealed predominant subpopulations formed during ART. The predominant subpopulations selected during ART likely contributed the most to ART failures. And the shared genetic changes within these predominant subpopulations likely provided HIV with enhanced fitness during ART.

To show the ability of HDBPA to analyze the evolution of various HIV regions that have evolved under different drug classes, I presented the analyses of ART failures in two PWHs.

From the first PWH (PID: 3IB) who failed a PI-based therapy, HDBPA revealed two predominant HIV subpopulations that arose from distinct MRCAs: one contained protease (PR) DRMs presumably coming from the naive infecting HIV population, and the other was

PI-susceptible, but circumvented inhibitor activity presumably due to compensatory mutations in Gag CSs that enhanced PR cleavage and viral replication capacity. During ART, these PI-susceptible HIV genomes developed PR DRMs *de novo* or through recombination with other drug-resistant genomes. Furthermore, these HIV genomes developed more mutations in the Gag CSs under drug pressure, besides those that were inherited from its MRCA. The mutations developed in P1/p6 have been reported to improve the PR cleavage efficiency, thus increasing viral replication capacity[11,57,58]. The mutations in P2/NC might be selected accordingly, maintaining the sequential order of cleavage in the Gag polyprotein[61]. In summary, I observed different evolutionary strategies utilized in the two predominant subpopulations: one evolved to higher resistance with moderate replication capacity, and the other evolved to moderate resistance with higher replication capacity. Moreover, this indicated that pre-existing compensatory mutations in the drug-naive HIV genomes could also contribute to resistance development and affect ART outcomes.

From the other PWH (PID: 1008) who failed two consecutive dual-NRTI therapies, HDBPA revealed three major evolutionary pathways individual HIVs took during therapy failures. These pathways were characterized by unique combinations of RT DRMs and RT compensatory mutations on a series of RT positions (53, 75, 184, 210, 293, 297, 334, 335, and 417). Notably, HIV genomes in Pathway A acquired the combined mutations M41L-E53D-V75T-L210W-T215Y-E297S-Q334N-V417I, which provided them with the highest D4T resistance in the viral population. However, HIV genomes with this mutation set failed to acquire RT M184V during the second therapy and were eliminated by 3TC, the other drug used in the second therapy, pressure. Conversely, HIV genomes in the other two pathways acquired

distinct combined mutations, enabling the successful development of RT M184V and driving enrichment during the second therapy. This observation indicated that different combinations of RT mutations might have either promoting or disruptive effects on acquiring RT M184V. Therefore, compensatory mutations emerging in HIV genomes earlier, such as those from prior therapies, might significantly shape their evolutionary outcomes (regression or selection) in subsequent therapies. Moreover, like the previous analysis, this analysis again revealed several viral subpopulations involved during ART failure(s). Different mutation combinations in these viral subpopulations presumably provided them with different flexibility when facing new drug pressure and increased the difficulty in suppressing viral replication.

This study focused on non-synonymous mutations, which lead to structural and functional changes in HIV proteins. Therefore, the Hamming distance was calculated based on non-synonymous mutations in the HIV genome. However, it is important to note that the Hamming distance calculation criteria can vary depending on the specific objectives of each study. Criteria options include nucleotide variants, amino acid mutations, or specifically non-synonymous amino acid mutations. Moreover, during Hamming distance calculation, specific mutation groups or HIV regions can be emphasized by assigning distinct weights to different components. For example, if we want to consider the entire HIV *gag-pol* region but also emphasize the similarity between *protease* sequences over other HIV regions, a higher score for protease mutations would be appropriate. Similarly, a higher score could be given to non-synonymous mutations over synonymous mutations to emphasize non-synonymous mutations in the HIV genome.

In this study, I focused on substitution mutations' contribution to the evolution of HIV genomes. However, recombination is another essential driving force of viral evolution. In theory, HIV genomes recombined between two or more MRCAs would have an unusually high Hamming distance when compared to all HIV genomes detected at a previous time point. Based on this, recombined HIV genomes could be identified and investigated in future updates of HDBPA.

To summarize, I validated the HDBPA's ability to analyze the evolution of various HIV regions that have evolved under different drug classes. Moreover, this method can be extended to various sequencing datasets for the study of evolutionary processes in other viruses and bacteria. Notably, there are no constraints on sequence length and depth. However, longer target region sequencing and higher sequencing depth in each longitudinal sample could improve the matching accuracy between progeny genomes and their MRCAs.

3.4 METHODS

Viral RNA extraction from samples collected from PWHs.

Frozen plasma aliquots from PWHs experiencing ART failures were obtained from the UC San Diego Primary Infection Resource Consortium (PIRC). From every sample, 1 ml plasma was transferred to 1.5 ml DNA LoBind tubes (Eppendorf), and centrifuged for 75 min at 18 000 x g and 4 °C to concentrate the virus. After centrifugation, 860 µl of supernatant was carefully aspirated from the top of the tube and frozen at −80 °C. The remaining 140 µl of the concentrated virus was processed according to the QIAamp Viral RNA Mini kit (QIAGEN 52904) protocol, with the only deviation from protocol being viral RNA elution in 30 µl of nuclease-free water. 11 µl of this eluate was used as input for subsequent RT and serial emulsion PCRs as described in the previous publication[25].

Sequencing library preparation.

For the sequencing library preparation process, including reverse transcription, emulsion PCRs, MrHAMER template preparation, and generating concatemers from the MrHAMER templates, please review the previous publication[25].

Nanopore sequencing.

Sequencing libraries were prepared using the Ligation Sequencing Kit (SQK-LSK109). All samples were sequenced with MinION R9.4.1 flowcells, basecalled with Guppy basecaller 3.6.0. Quality Control was performed on reads using the NanoPlot package.

Bioinformatics.

For the downstream bioinformatics analyzing process, including basecalling, splitting long concatemers, and reconstructing genomes, please review the previous publication[25].

Insertion of individual gag-pol reads into the HXB2 reference sequence.

Individual HIV genomes should have the same coordinates, e.g., HXB2 coordinates, for the Hamming distance calculation. Hence, after sequencing and error-correction, individual *gag-pol* reads in each sample were mapped to the HXB2 reference sequence (accession number K03455) using *minimap2* (default mode). Aligned reads were stored in a SAM file. Based on CIGAR strings in the SAM file, the insertions detected every *gag-pol* read ('I' and 'S') were removed, and the deletions ('D') were filled using the corresponding sequence in the HXB2 reference. In this way, all *gag-pol* reads were uniform in length as HXB2 reference sequence, i.e., 297 nt, but kept their unique mutations. This process was completed using a *Python* script.

Calculating the Hamming distance and targeting MRCAs for individual HIV genomes.

Individual HIV genomes from two longitudinal samples of the same PWH were sequenced and aligned to the HIV HXB2 reference genome. Non-synonymous mutations were identified in each aligned HIV genome, making a mutation set. For each mutation set detected at the later time point, a comparison was made to those mutation sets from the prior time point. The Hamming distance between these sets was then calculated. This process enabled the identification of the Most Recent Common Ancestor (MRCA) for each *gag-pol* read detected at the later time point. The MRCA was determined based on the lowest Hamming distance between the mutation sets. This analysis was performed using a custom *Python* script.

Chapter 4. Co-variation of viral recombination with single nucleotide variants (SNVs) during viral evolution revealed by an improved co-variation mapper (CoVaMa)

4.1 Abstract

In the previous Chapters, I identified correlated mutations within the HIV genome, which were encoded by co-varying Single-Nucleotide Variants (SNVs) that arose during antiretroviral therapy (ART) failures. Besides the acquisition of SNVs, the recombination process is another strong driving force of viral evolution. It results from the template-switching events during the replication of viral RNAs. In this process, the viral polymerase accidentally disassociates from its original temple viral genome and then re-associates to another region in the same template or a different template. These results in recombination events in the viral genome, including insertions, deletions, and duplications. Recombination events can have a dramatic impact on viral fitness, viral intra-host diversity, and the development of resistance against antiviral drugs.

The co-occurrence of SNVs in viral genomes during evolution has been well-described. However, unlike the correlation between SNVs, studying the correlation between recombination events with each other or with SNVs was hampered by their inherent genetic complexity and a lack of bioinformatic tools. In this collaborative research with the Routh lab, we expanded the previously reported Co-Variation Mapper pipeline (CoVaMa, v0.1) to measure linkage

disequilibrium between recombination events and SNVs within both short-read and long-read sequencing datasets.

I revised the new version of CoVaMa (v0.7) and demonstrated its applications by analyzing two viral datasets acquired in previous studies. One viral dataset was acquired from Flock House virus (FHV) serially passaged *in vitro* and sequenced using long-read sequencing technology. The analysis of this dataset revealed SNVs that were either correlated or anti-correlated with large deletions in FHV Defective-RNAs (D-RNAs). The other viral dataset was acquired from longitudinal serum samples collected from a person living with HIV (PWH) who failed ART. In the HIV genome, I found correlations between insertions in the $p6^{Gag}$ and mutations in Gag cleavage sites.

To summarize, this collaborative study provided a tool to probe both NGS datasets and long-read datasets for evidence of the correlation between intra-host variants. This study confirmed previous findings and provided insights into novel associations between SNVs and specific recombination events within the viral genome.

4.2 Results

4.2.1 Overview of the CoVaMa Pipeline (v0.7)

CoVaMa (v0.7) pipeline utilized aligned reads in SAM files as input. It identified recombination events (insertions, deletions, and duplications) and SNVs by comparing aligned reads to the reference genome. To identify associations between SNVs, recombination events, and SNVs with recombination events, CoVaMa built three types of contingency tables accordingly: 4x4, 2x2, and 4x2 (Figure 4.1A). For instance, to explore the relationship between an SNV on locus *i* and a recombination event, a 4x2 contingency table would be built. In the contingency table, rows represented nucleotide possibilities on locus *i* (A, T, G, or C), and the columns represented the presence or absence of the recombination event. CoVaMa checked each input read for the nucleotide on locus *i* and the presence of that recombination event to populate the contingency table. All eight possible combinations (‘A-presence’, ‘T-presence’, ‘G-presence’, ‘C-presence’, ‘A-absence, ‘T-absence, ‘G-absence, ‘C-absence) were quantified and used to populate the contingency table. The exact process was used to populate contingency tables between SNVs or between recombination events.

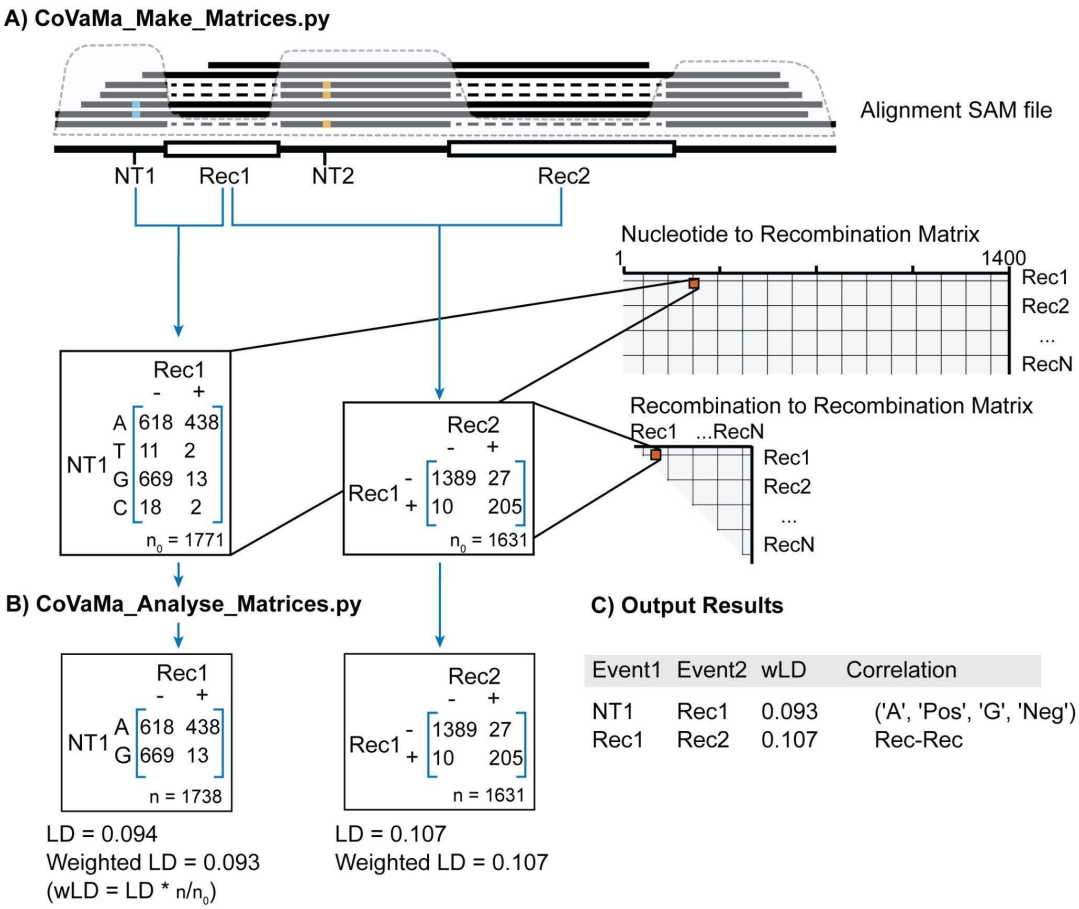

Figure 4.1. Overview of the CoVaMa pipeline.

The CoVaMa pipeline contains two scripts. (A) The CoVaMa_Make_Matrices.py extracts information from each aligned read and generates large matrices containing 4x4 SNV-vs-SNV contingency tables, 4x2 SNV-vs-Recombination contingency tables, and 2x2 Recombination-vs-Recombination contingency tables. In each contingency table, the columns correspond to either A, T, G, and Cs for nucleotides, or the presence and the absence of recombination events. Rec, Recombination; SNV, Single-Nucleotide Variant. (B) CoVaMa_Analyse_Matrices.py analyzes contingency tables from each matrix for evidence of linkage disequilibrium. From each 4x4, 4x2, and 2x2 contingency table, every possible 2x2 table populated by sufficient reads is extracted to calculate the LD value. The LD and R-squared values are normalized by the proportion of reads

populating the 2x2 contingency table, generating the weighted LD and R-squared values. LD value: linkage disequilibrium value. (C) An example of the CoVaMa output.

Once contingency tables had been generated, they would be used to calculate the LD values between every two adaptation events (recombination events and SNVs) (Figure 4.1B). From the 4x4, 4x2, and 2x2 contingency table, every possible 2x2 sub-table was extracted, and only those populated with a sufficient number of reads were subjected to LD calculation. LD calculation was performed using the standard formula: $LD = p_{AB}p_{ab} - p_{Ab}p_{aB}$. 'A' and 'a' were placeholders for two possible options at the first coordinate, which could be either two nucleotide possibilities at one locus, or the presence or absence of a recombination event. Similarly, 'B' and 'b' represented the corresponding possibilities at the second coordinate. The LD value ranged from 0 to +/- 0.25, where 0 indicated no association and values close to +/- 0.25 indicated a strong association. Considering that a single 4x4 or 4x2 table could yield multiple potential LD reports due to diversity within these tables, LD values were normalized by the proportion of reads populating the 2x2 contingency table from the entire 4x4 or 4x2 contingency table. This normalization yielded weighted LD values (wLD).

The output of CoVaMa was reported using a TEXT file (Figure 4.1C), which provided detailed information on associated adaptation events that were detected, as well as the LD value for each association.

4.2.2 Associations Between Recombination Events and SNVs in Defective Flock House Virus (FHV) RNAs Revealed by CoVaMa

Flock House virus (FHV) is an insect-specific small bipartite RNA virus. Its genome consists of two segments, RNA1 (3.1 kb) and RNA2 (1.4kb), encoding for the viral polymerase and viral capsid protein, respectively[62]. During evolution, Defective-RNAs (D-RNAs) arise spontaneously through nonhomologous RNA recombination in FHV RNA[63]. Thus, FHV is an ideal model for studying viral recombination and evolution[34,64]. In a previous study[36], Jaworski *et al.* (2017) serially passaged FHV in S2 Drosophila cells in culture, aiming to characterize the emergence, selection, and adaptation of D-RNAs *in vitro*. After each passage, the researchers extracted encapsidated RNA from purified virions and generated full-length cDNA copies of the FHV genomic segments. Purified cDNA was sequenced using Nanopore sequencing to identify the emergence of insertions, deletions, and other RNA recombination events. Interestingly, the sequencing results revealed that, after several passages, multiple deletions were most commonly found in individual reads. On the contrary, reads containing only single-deletions were seldom seen. In the same study, Jaworski *et al.* postulated that the co-occurrence of multiple deletion events within individual reads indicated a selective advantage for ‘mature’ D-RNAs over the ‘immature’ D-RNAs. These paired deletions were consistent with major defective RNA2 species previously characterized[36,65–67].

The researchers also observed the emergence of multiple SNVs in the FHV RNA2 over passage, including A226G and G575A. However, the function of these two mutations still needed to be determined. A226G leads to a synonymous mutation. On the contrary, G575A results in the Alanine to Threonine substitution at amino acid position 185 of the capsid protein. Looking at

the FHV particle (T=3 icosahedral), this A185T mutation is located at the five-fold and quasi-three-fold symmetry axes of the virus particle, indicating possible interference with the viral assembly[68].

To investigate the relationship between A226G, G575A, and recombination events constituting D-RNA2 species, I analyzed the FHV dataset using the CoVaMa (v0.7). Few recombination events were detected in the first two passages, consistent with the findings of Jaworski *et al.* (2017) (Figure 4.2A). Diverse recombination events emerged throughout passaging in FHV RNA2, with their frequency increasing (Figure 4.2A, Figure 4.2B). This increase in diversity and abundance reflected Defective RNA species' selective or replicative advantage. Additionally, I observed fluctuation in the abundance of recombination events over time (Figure 4.2B). This could be due to the competition between D-RNA encapsulation and the requirement of generating enough complete capsid proteins to form mature capsids[69]. Interestingly, the majority of recombination events observed located in two regions with only slight variance in the exact coordinates of the recombination junction: 1) deletion events that excised nucleotides between nt 240 and nt 530, termed 'Group 1' (such as 248^512 and 250^513); 2) deletion events that excised nucleotides between nt 720 and nt 1240, termed 'Group 2' (such as 736^1219) (Figure 4.2B, Figure 4.2C). By aligning these recombination events to the FHV RNA2 reference genome, I confirmed that none of the recombination events removed the packing motif or the replication cis-acting motif in FHV RNA2[66,70].

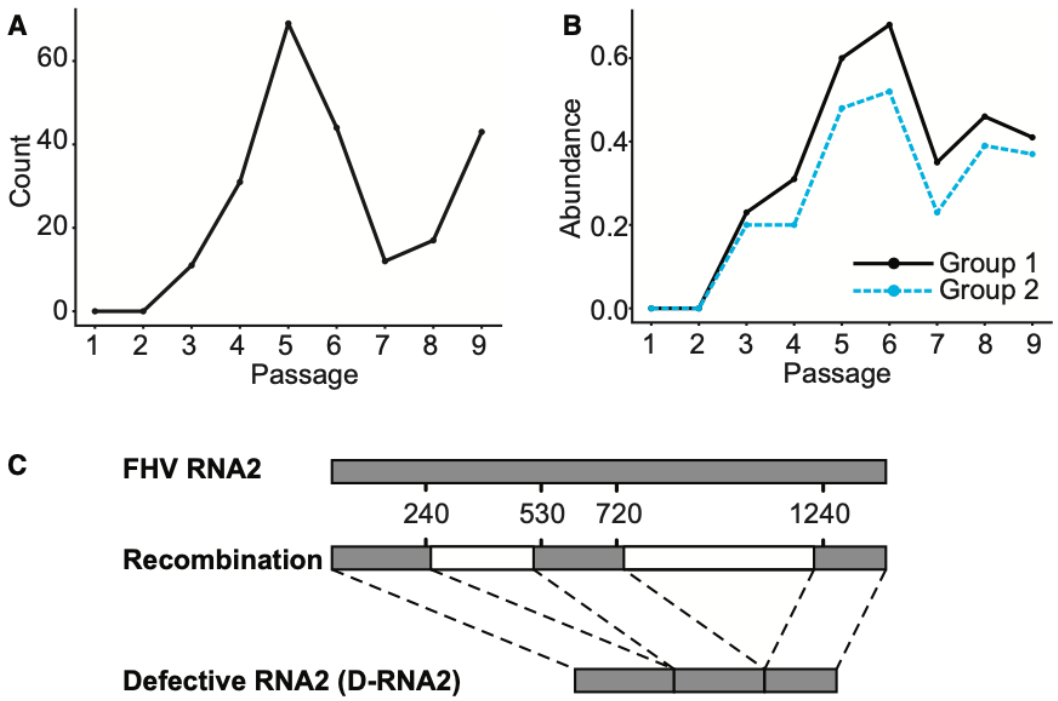

Figure 4.2. Detection of recombination events in FHV RNA2 using CoVaMa. (A) Count of unique recombination events detected in FHV RNA2 across different passages. (B) Recombination events tend to cluster into two distinct groups, denoted "Group 1" and "Group 2," located in different regions of FHV RNA2. The abundance of each group is plotted over passages. (C) Mapping of the two recombination-enriched regions to the FHV RNA2 reference genome. These recombination events collectively form FHV Defective RNA2. FHV: Flock House virus.

CoVaMa detected strong associations between recombination events, e.g., 248^512 and 736^1219, 250^513 and 736^1219 (Figure 4.3A). Additionally, it detected a strong association between recombination events and SNVs, e.g., between G575A or A226G and recombination events (Figure 4.3B). The non-synonymous mutation G575A positively correlated with recombination events 250^513, 248^512, and 736^1219, suggesting a co-dependent evolution (Figure 4.3B). In contrast to the G575A mutation, the A226G mutation negatively correlated

with recombination events 250^513, 248^512, and 736^1219 (Figure 4.3B). Consistent with these observations, CoVaMa also reported that A226G and G575A negatively correlated with one another (Figure 4.3B, Table 4.1).

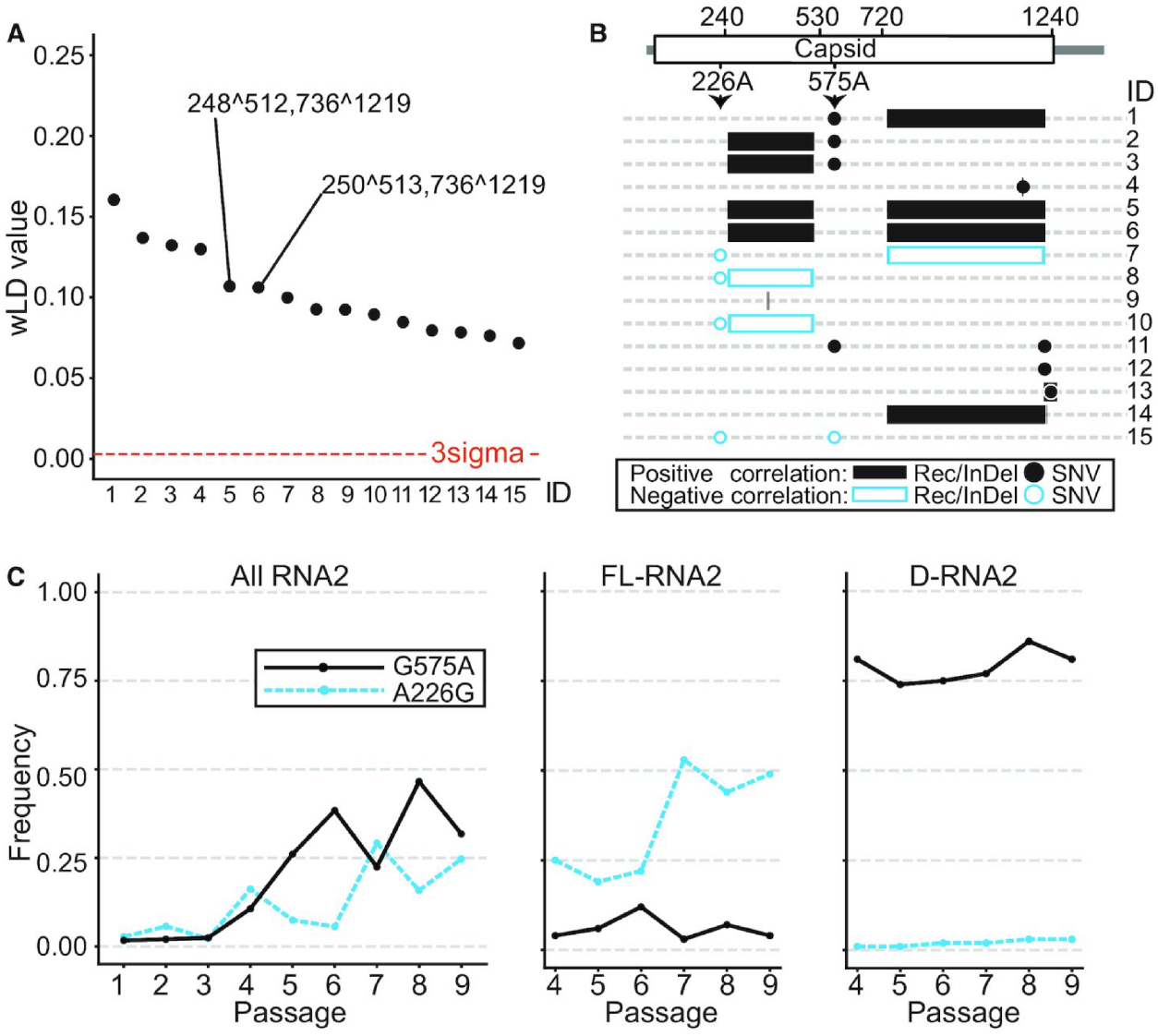

Figure 4.3. CoVaMa reveals associations between recombination events and SNVs in FHV RNA2.

The top 15 associations with the highest weighted LD (wLD) values revealed by CoVaMa in Passage 9 are labeled with IDs from 1 to 15. (A) The wLD values of the 15

associations are plotted from high to low. The associations between major recombination events are labeled in the plot. The three-sigma threshold is shown by a red dashed line. (B) This schematic diagram shows the distribution of the 15 most significant associations on FHV RNA2. The FHV RNA genome and the capsid ORF are shown on the top. Positive associations are colored in black. And negative associations are colored in blue. Recombination events (deletions) are plotted using blocks, and SNVs are plotted using dots. (C) The abundance of mutation A226G and G575A in all RNA2 reads (left), in the full-length RNA2 (FL-RNA2, middle), and in the D-RNA2 (right) in each passage. ORF: open reading frame.

Table 4.1. A226G and G575A are negatively associated with each other in FHV RNA2.

Passage	Position 1	Position 2	wLD value	wR-squared value	correlations
6	226	5775	0.013454	0.014234	('A', 'A', 'G', 'G')
9	226	575	0.071722	0.125011	('A', 'A', 'G', 'G')

To confirm these results, I separated reads mapping to either the full-length RNA2 or D-RNA2 into two groups. The separation was accomplished based on whether one read had the recombination event 736^1219, one of the major recombination events constituting D-RNA2. In the full-length RNA2 group and D-RNA2 group, I quantified the frequency of A226G and G575A. The quantification results revealed that A226G was enriched in full-length RNA2 from 25% in Passage 4 to 50% in Passage 9, while the abundance of A226G in D-RNA2 remained low at about 2% in all passages (Figure 4.3C). The ratio of the abundance of this mutation in full-length RNA2 and D-RNA2 was 18.8 ± 6.0 (SD) over passage (Table 4.2). This SNV was therefore anti-correlated with the emergence of D-RNAs. The enrichment of A226G in the full-length RNA2 suggested a possibly beneficial effect on the replication or packaging of the

mutant full-length RNA2. In contrast, G575A was enriched in the D-RNA2 and relatively depleted in the full-length genomic RNA (Figure 4.3C). The ratio of the abundance of this mutation in D-RNA2 and full-length RNA2 was 16.2 ± 7.1 (SD) over passage (Table 4.3). This SNV was, therefore, positively correlated with the emergence of D-RNAs, suggesting a D-RNA-specific adaptation.

Table 4.2. Frequency of A226G in full-length FHV RNA2 and D-RNA2 over passage.

	226A			**226G**		
Passage	All RNA2	In full-length RNA2	In D-RNAs	All RNA2	In full-length RNA2	In D-RNAs
1	0.96	NA	NA	0.03	NA	NA
2	0.93	NA	NA	0.06	NA	NA
3	0.97	NA	NA	0.02	NA	NA
4	0.83	0.73	0.99	0.16	0.25	0.01
5	0.92	0.79	0.97	0.07	0.19	0.01
6	0.94	0.75	0.96	0.06	0.22	0.02
7	0.69	0.45	0.97	0.29	0.53	0.02
8	0.83	0.54	0.05	0.16	0.44	0.03
9	0.74	0.49	0.96	0.25	0.49	0.03

NOTE: The frequency of SNV in “wild-type” RNA2 and defective RNA2 was calculated from the contingency table formed between SNV and the major recombination event 736^1219, which was not available in Passage 1, 2, and 3 due to the low frequency of the SNV or the recombination event 736^1219. NA: Not available. D-RNA2: defective-RNA2.

Table 4.3. Frequency of G575A in full-length FHV RNA2 and D-RNA2 over passage.

	575A			**575G**		
Passage	All RNA2	In full-length RNA2	In D-RNAs	All RNA2	In full-length RNA2	In D-RNAs
1	0.02	NA	NA	0.97	NA	NA
2	0.02	NA	NA	0.97	NA	NA
3	0.02	NA	NA	0.97	NA	NA
4	0.11	0.04	0.81	0.89	0.95	0.17
5	0.26	0.06	0.74	0.73	0.93	0.24
6	0.38	0.12	0.75	0.61	0.86	0.23
7	0.22	0.03	0.77	0.77	0.95	0.17
8	0.47	0.07	0.86	0.52	0.90	0.11
9	0.32	0.04	0.81	0.68	0.95	0.16

NOTE: The frequency of SNV in "wild-type" RNA2 and defective RNA2 was calculated from the contingency table formed between SNV and the major recombination event 736^1219, which was not available in Passage 1, 2, and 3 due to the low frequency of the SNV or the recombination event 736^1219. NA: Not available. D-RNA2: defective-RNA2.

A226G in FHV RNA2 leads to a synonymous substitution. Thus I investigated its potential influence on the secondary structure of FHV RNA2. I used Vienna RNA Website[71] to generate predicted RNA structures of both D-RNA2 and full-length RNA2, with and without A226G. A226G showed no significant influence on the adjacent RNA2 packaging motif in full-length RNA2 (Figure 4.4A) and D-RNA2 (Figure 4.4B). However, the A226G mutation was predicted to destabilize a region of D-RNA2 secondary structure that formed a long-range interaction with residues 580-600, which in turn was predicted to alter additional long-range interactions even further downstream within the 50 nts of the 3' terminus (Figure 4.4B). Interestingly, this region contains the cis-acting motif essential for RNA2 replication[72–74], indicating a possible deleterious effect of A226G in the replication of D-RNA2. Moreover, the predicted structure of full-length RNA2 without the A226G mutation placed position 736 and position 1219 close to each other, between which one of the major recombination events in D-RNA2 was formed (Figure 4.4C). However, when there was an A226G mutation in full-length RNA2, this distance was significantly increased, which might preclude the emergence of recombination events (Figure 4.4D).

To validate the reported correlation between the non-synonymous substitution G575A with the recombination events found in D-RNA2, our collaborator engineered a pMT-FHVRNA2 expression vector with the G575A SNV, using a reverse genetics approach that has been previously used extensively[64]. However, this mutant virus could not be rescued. This observation

suggested that this point mutation was a lethal mutation, rationalizing why it was not observed in the full-length RNA2 but only in the D-RNA2, which was not constrained by a need to express FHV capsid.

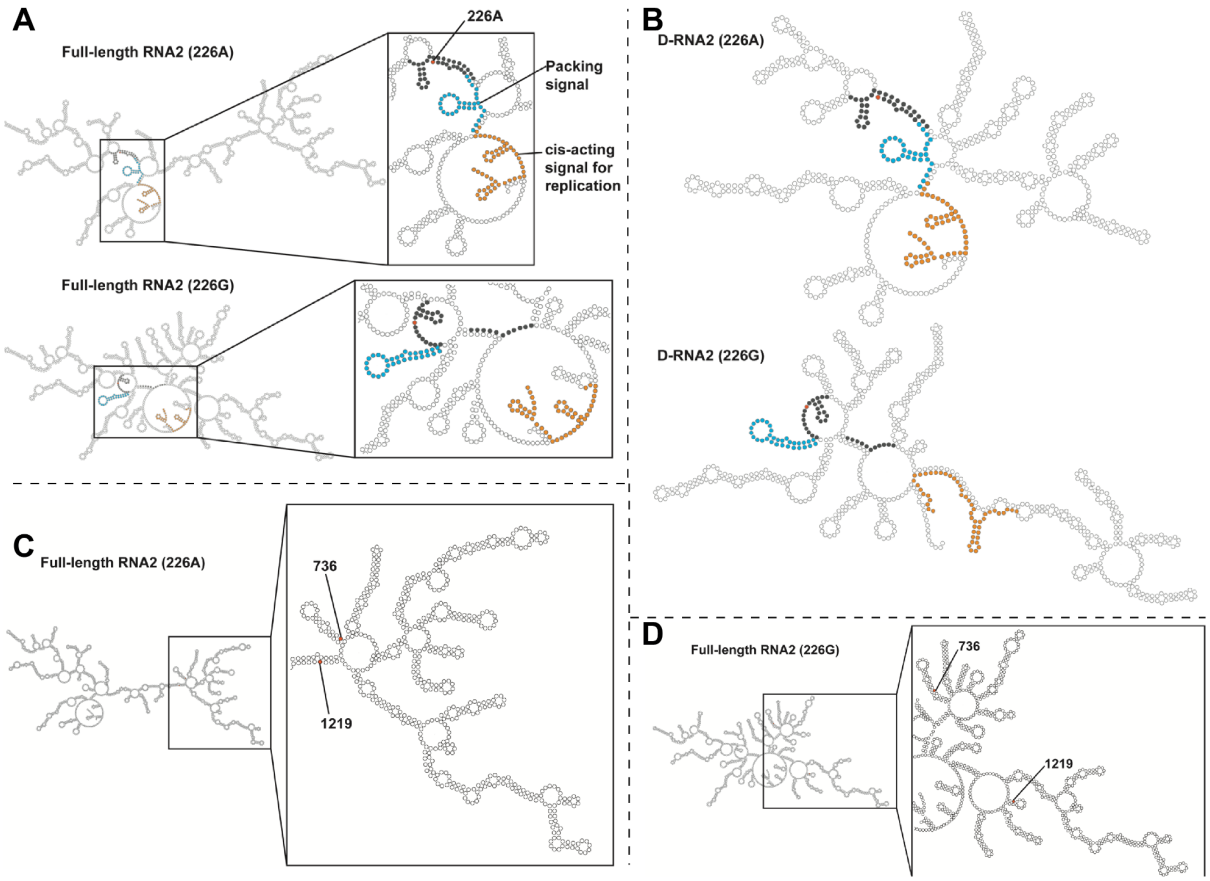

Figure 4.4. A226G's potential influence on the secondary structure of FHV RNA2. (A) Predicted secondary structure of full-length FHV RNA2 with and without A226G. The secondary structures are predicted using the Vienna RNA Websuite and viewed in Forna (Gruber et al., 2008). The nucleotide at nt 226 is highlighted in red and labeled. The packing signal region at the 5' terminus is colored in blue, and the cis-acting signal at the 3' terminus is colored in orange. The predicted long-range interaction between nt 590-600 and nt 218-241 is colored in gray. (B) Predicted D-RNA2 structures with and without A226G. (C) Recombination sites in the predicted secondary structure of

full-length FHV RNA2 without A226G. The secondary structures are predicted using the Vienna RNA Websuite and viewed in Forna. Recombination sites at nucleotides 736 and 1219 are highlighted in red. (D) Recombination sites in the predicted secondary structure of full-length FHV RNA2 with A226G.

4.2.3 Associations Between Insertion in $p6^{Gag}$ and Mutations in Gag Cleavage Sites Revealed by CoVaMa

In 2015, Routh *et al.* reported correlated mutations within HIV genomes collected from a large cohort of PWHsas part of the US Military HIV Natural History Longitudinal study[34]. This study used the previous version of CoVaMa (v0.1) and revealed correlated mutations between HIV Gag and protease, as well as within protease. However, this study was limited to assessing only SNVs and did not consider recombination events. Recombination events, like insertions, deletions, and duplications, have been frequently observed in the HIV genome, indicating their critical contribution to viral evolution[75].

To investigate the association between recombination events and SNVs in the HIV genome, I analyzed the HIV evolution in one PWH experiencing ART failures using the new version of CoVaMa (v0.7). Five longitudinal serum samples have been collected from this PWH and sequenced using NGS in previous studies[14,37]. Additionally, our collaborator detected several recombination events using ViReMa (Viral-Recombination-Mapper) from the sequencing outputs[76]. An insertion in the P(T/S)AP region of the $p6^{Gag}$ was detected in all five samples, which added ‘RPEPS’ or ‘RLEPS’ downstream of the P1/p6 cleavage site (Table 4.1, Figure 4.4A). Similar insertions in the $p6^{Gag}$ have been observed in other studies[25,77,78]. Another consistent 6-nt insertion occurred close to the proteolysis sites between the matrix and capsid

proteins. This insertion encoded an extra ‘AA’ between amino acid 120 and amino acid 121 (Figure 4.4A). This extra ‘AA’ was observed in the HIV-1 group M subtype B isolate ARV2/SF2[79].

Table 4.4. Insertions detected in longitudinal serum samples from one PWH.

Events	Region	Amino acid change	Samples
2157_AGACCAGAGCCATCA_2158, 2157_AGACTAGAGCCATCA_2158	$p6^{Gag}$	RPEPS, RLEPS	1 - 5
1148_GGCAGC_1149, 1151_CAGCTG_1151	MA	AA, AA	1 - 5
1174_CCAGCAGCCAAGTCA_1175, 1174_CCAGCAGCCAGGTCA_1175	MA/CA cleavage site	TSSQV, TSSQV	2, 4

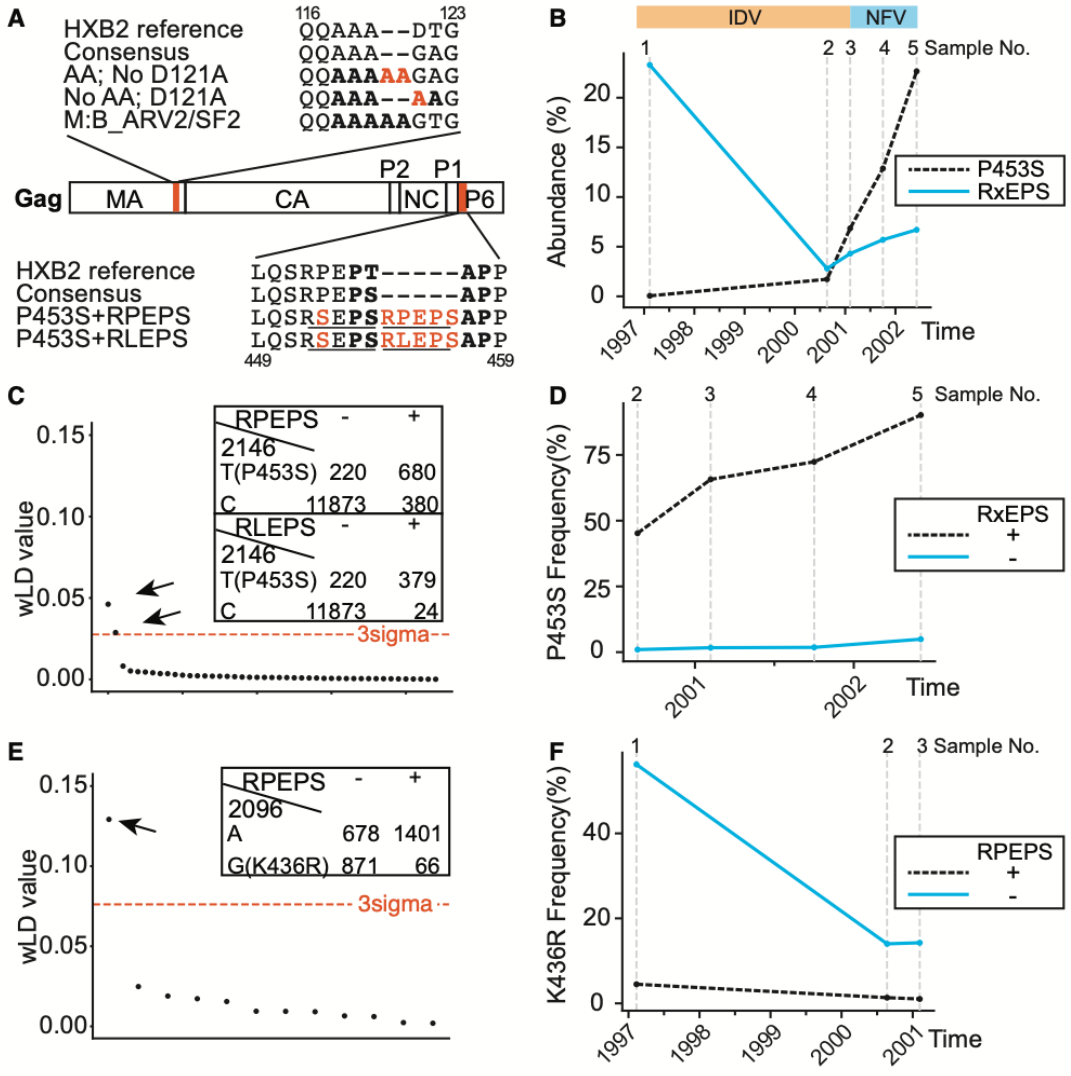

Figure 4.5. CoVaMa reveals associations between insertions in p6^{Gag} and mutations in Gag cleavage sites.

(A) Mapping of associated insertions and mutations in the HIV Genome to the HIV Gag region. The Gag open reading frames (ORFs) are labeled: MA, CA, P2, NC, P1, and p6^{Gag}. At the top are shown the negative association between 'AA' insertion and D121A mutation: the HXB2 reference sequence, the sample's consensus sequence, two representative sequences with 'AA' insertion or D121A mutation, and the sequence for

HIV-1 group M subtype B isolate ARV2/SF2. The associated events are highlighted in red. The five-alanine motif is highlighted in bold. At the bottom are shown the positive association between 'RxEPS' insertion (x for P/L) and P453S mutation: the HXB2 reference sequence, the sample's consensus sequence, and two representative sequences with the associated events. The associated events are highlighted in red. The PTAP region is highlighted in bold, and the "duplication" region is marked underlined. (B) The abundance of P453S mutation and 'RxEPS' insertion in the viral population over time. Protease inhibitors used are annotated: Indinavir (IDV) or Nelfinavir (NFV). (C) Significant positive associations between 'RxEPS' insertions and P453S are indicated by arrows among all associations involving 'RxEPS' in Sample 4, with their contingency tables at the top right. (D) The abundance of P453S in reads with and without 'RxEPS' insertions. (E) A significant negative association between 'RPEPS' insertion and K436R is indicated by an arrow among all associations involving 'RPEPS' in Sample 1, with its contingency table at the top right. (F) The abundance of K436R in reads with and without 'RPEPS' insertions.

CoVaMa revealed significant associations between insertion events and specific SNVs within the HIV genome. For instance, the P453S mutation in the Gag P1/p6 cleavage site became more prevalent after sampling point 2, coinciding with the enrichment of the 'RxEPS' insertion in the $p6^{Gag}$ region (Figure 5.4B). CoVaMa revealed a positive correlation between C2146T and 'RxEPS' insertion (Figure 5.4C), with a maintained elevated LD between them over time. When quantifying the abundance of P453S in reads with and without 'RxEPS' insertion, a clear pattern emerged: P453S was notably more common in reads with the insertion (Figure 5.4D). The ratio of P453S abundance in reads with and without 'RLEPS' was 27.74 ± 17.66 (SD), and the ratio of P453S abundance in reads with and without 'RPEPS' was 36.43 ± 16.67 (SD). In the last sample, while only 5% of the reads without 'RxEPS' insertion carried the P453S substitution (x for P/L),

this number increased to 99% and 69% in reads with ‘RLEPS’ and ‘RPEPS’, respectively. This finding indicated the possible cooperativity between P453S in the P1/p6 cleavage site and the ‘RxEPS’ insertion in the PTAP region. In contrast to the P453S mutation, the K436R mutation in the Gag NC/P1 cleavage site was negatively correlated with the ‘RPEPS’ insertion (Figure 5.4E). The ratio of the K436R abundance in reads without and with ‘RPEPS’ was 12.27 ± 1.71 (SD) in the first three samples before its abundance decreased to lower than 1% (Figure 5.4F). Consistent with the above findings, a combination of K436R substitution and P453S substitution was less favored in the HIV genome.

In addition to the associations between ‘RxEPS’ in $p6^{Gag}$ and mutations in the Gag cleavage sites, the ‘AA’ insertion in the MA negatively correlated with the D121A substitution (Figure 4.6A). This negative association resulted in a five-alanine motif in the C terminus of MA, which has been seen in the isolate ARV2/SF2. Additionally, the ‘TSSQV’ insertion in the MA/CA cleavage site negatively correlated with D121A (Figure 4.6B) and positively correlated with the ‘AA’ insertion (wLD = 0.021805 in sample 4).

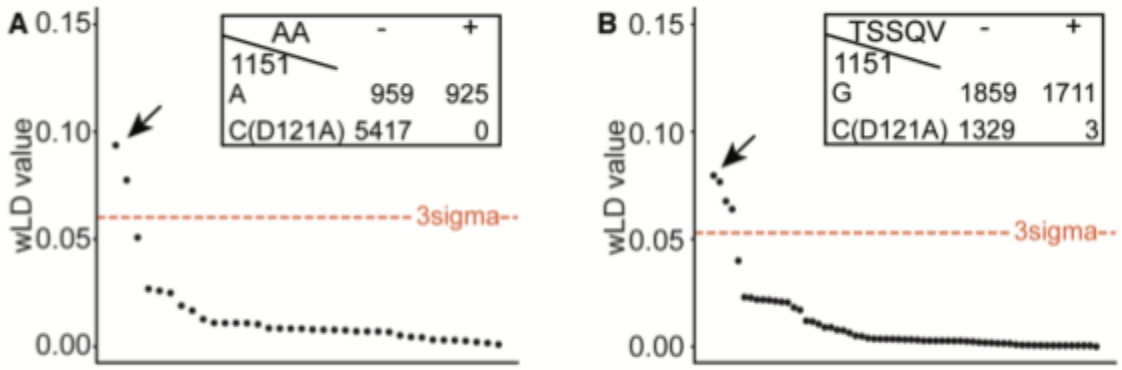

Figure 4.6. CoVaMa reveals associations between insertions and mutation in the HIV matrix protein.

(A) A significant negative association between 'AA' insertion and D121A is indicated by an arrow among associations involving 'AA' in Sample 3, with its contingency table shown at the top right. (B) A significant negative association between 'TSSQV' insertion and D121A is indicated by an arrow among associations involving 'TSSQV' in Sample 4, with its contingency table shown at the top right. The three-sigma threshold is shown by the red dashed line.

4.3 Discussion

Viral evolution and adaptation occur through acquiring novel SNVs and recombination events such as small structural variants or larger insertions and deletions[80–82]. These adaptation events seldom occur in isolation. Instead, multiple adaptation events work together to confer the virus with novel biological properties. The correlation between SNVs during viral evolution has been well-described for a range of viral systems[14,32] using various bioinformatic tools (reviewed by Posada-Cespedes *et al.*, 2017[33]). However, computational tools that determine whether SNVs are correlated with recombination events or whether multiple recombination events are correlated with one another have to date been lacking. To address this gap, we extended the previously reported CoVaMa (v0.1) pipeline. Using the new version of CoVaMa (v0.7), I analyzed two viral sequencing datasets and identified novel associations between recombination events and specific SNVs in the viral genome.

The first viral dataset was collected from FHV passaged *in vitro*. Consistent with previous studies, CoVaMa revealed a strong positive association between deletions that constituted D-RNA species[36]. In this study, I also found SNVs that were either positively or negatively associated with deletions in FHV D-RNA2. These SNVs were previously too distantly spaced from each other or from deletions to be correlated in the NGS data, illustrating the value of performing long-read sequencing for this type of analysis.

Our data demonstrated that the FHV D-RNAs acquired adaptation events that were not found in the full-length FHV RNA. These adaptation events are possibly driven by the formation of new secondary structures, which could confer replicative or packaging advantages to the D-RNAs. Additionally, these events may be favored due to the reduced genetic constraints on D-RNA2

mutations, as D-RNAs are not responsible for functional viral expression. Notably, the SNV G575A observed in this study resulted in an A185T substitution at the symmetry axis of the FHV particle. Such a substitution at this interface might hinder efficient virus assembly, potentially explaining the enrichment of G575A in the D-RNAs rather than in full-length RNAs.

Adaptation events in the full-length FHV genome that were rarely found in the D-RNAs were as well observed. Our findings indicated that these adaptations only started to enrich after the emergence of D-RNAs (such as the synonymous A226G SNV found here), suggesting that the full-length 'helper' virus may be adapting or escaping from the 'interfering' properties of the D-RNAs. Although the mechanism of escape was not clear from this data, such a phenomenon was originally postulated by DePolo *et al*., (1987)[83] who demonstrated that vesicular stomatitis virus (VSV) isolated from late viral passages was not subject to attenuation from defective interfering viral particles that arose in earlier viral passages. Further experimental characterization of these SNVs would reveal the precise molecular mechanisms driving their selection and competition.

Previous studies have reported that amino acid substitutions in HIV Gag compensate for the compromised catalytic functions of protease with DRMs, contributing to viral fitness under drug pressure[8,12,14,84]. Additionally, insertions in Gag could increase HIV infectivity and drug resistance. Tamiya *et al.* (2004) reported that 'SRPE' duplication in $p6^{Gag}$ in multi-PI resistant HIV subtype G could increase the cleavage efficiency of protease with DRMs[77]. Moreover, full or partial PTAP duplications have been reported to be selected during ART[85,86]. Martins *et al.* (2016) showed that, under PI pressure, full PTAP duplication in the Gag p6 significantly

increased the cleavage efficiency of protease with DRMs, thus increasing drug resistance and infectivity[87].

In this study, I analyzed the NGS dataset from longitudinal serum samples collected from one PWH experiencing ART failures using CoVaMa (v0.7). The analyzing results indicated strong associations between Gag cleavage site (CS) mutations and an 'R(P/L)EPS' insertion in the P(T/S)AP region of the $p6^{Gag}$. The P453S mutation in the Gag P1/p6 CS was rarely detected in HIV genomes without 'R(P/L)EPS' insertion, whereas it was enriched to over 90% in HIV isolates with 'R(P/L)EPS' insertion over time under drug pressure. 'RPEPS' belongs to the proline-rich motif 'RPEP(S/T)APP' in the N terminus of $p6^{Gag}$, which plays essential roles in packaging processed Pol proteins during late assembly[43,88]. Replacing the P453 and P455 showed reduced viral replication in primary monocytes[88], which might account for the consistently low frequency of P453S in the viral strains analyzed. However, the duplication of 'R(P/L)EPS' restored the prolines or added more prolines in this motif, which released the restriction and allowed the P453S substitution, resulting in the strong positive correlation detected between 'R(P/L)EPS' and P453S.

The final output of CoVaMa comprises a large table detailing the linkage disequilibrium found in each contingency table measured. As per the original report, CoVaMa also reports the LD values comprising the threshold values for three- and five-sigma. This value and the LD table can be used to provide a ranked list of genetic co-variations ranging from the most to least co-varying pairs of genetic adaptations. However, CoVaMa does not imply or provide a statistical framework with which to determine false discovery rates or degrees of significance with multiple hypothesis testing. CoVaMa can be applied to study diverse data types, including Illumina and

Nanopore reads, which each have their own inherent error rates and profiles. These sequencing platforms are also applied in different manners to yield volumes of data with different numbers of technical and/or biological replicates that will vary according to the investigator's specifications. Furthermore, the underlying templates are naturally highly diverse, ranging from small multiple-partite viruses (such as FHV) to long single-stranded RNA viruses, each also having its own error profile and different profiles of viral adaptations. As a result, different statistical frameworks that reflect the parameters used to detect and report covariance in the original sample must be deployed in each scenario to provide appropriate and robust estimates of statistical significance. These are not inherently provided by CoVaMa but can be developed based on the table of LD values reported in the CoVaMa output.

Overall, CoVaMa provides a simple and intuitive tool that probes both NGS and Nanopore datasets for evidence of the correlation between intra-host variants. Importantly, we here expanded this approach to detect and report the co-occurrence of SNVs with recombination events. While I focused here on viral intra-host diversity, the same approach and pipeline could equally be applied for NGS analysis of other organisms where diversity or correlation of sequence variants is anticipated, such as in bacterial or other complex mixtures of populations.

4.4 METHODS

Confirm the presence or absence of various recombination events in aligned sequencing reads.

To quantify the abundance of each recombination event in the viral population, in this algorithm, we utilize specific criteria to confirm the presence or absence of different recombination events (insertions, deletions, and duplications) in aligned sequencing reads. These criteria are described below:

1) Deletions/Splicing: To confirm the absence of a deletion event, a read must map inside the putative deletion site with at least a minimum number of nucleotides (--Rec_Exclusion X; default is ten nucleotides). These mapped nucleotides can map anywhere within the deletion and/or overlap with the recombination junction. With a large number of nucleotides required, the exclusion of the recombination event has high confidence. If too few nucleotides are required, ambiguity may arise due to sequence similarity between deleted nucleotides and the nucleotides upstream of the recombination event. For instance, if only one nucleotide is required to map inside the recombination event, a negative mapping will be scored erroneously for one out of every four recombination events as the first deleted nucleotide has a one in four chance of also being the same as the nucleotide upstream of the recombination event. Care must also be taken in repetitive regions in case the nucleotides preceding the recombination 5' site are close or identical to those preceding the recombination 3' site. In cases like these, a large 'X' value is required.
2) Micro-Deletions: If a deletion event is very small (i.e., it is a micro-deletion smaller than the --MicroInDel_Length Y parameter specified in the command line), then it is not

possible for --InDel_Exclusion X nts (default is 5) to map inside the recombination event. In this case, the aligned read should map over the possible micro-deletion event and have X number of mapped nucleotides on each side of the event to confirm the absence of the putative deletion.

3) Insertions: A similar strategy is required to confirm the absence of a putative insertion event. Insertion events can either correspond to inserted nucleotides, insertions of fragments of host or other viral genes, or small duplications. To confirm the absence of an insertion event, an aligned read should have an appropriate number of mapped nucleotides on each side of the insert site. The number of mapped nucleotides on each side is controlled by the --Rec_Exclusion parameter (default is 10) in the command line for large insertions or --Indel_Exclusion parameter (default is 5) for micro-insertions smaller than the --MicroInDel_Length parameter (default is 5). An inappropriate setting alters the number of events detected and might influence the results.

Long-read data from serial passaged FHV analyzed using CoVaMa.

Jaworski *et al.* (2017) previously described a serial passaging experiment of FHV followed by the analysis of viral genetic changes by the parallel use of short-read Illumina and long-read Nanopore sequencing[36]. Briefly, pMT vectors containing the cDNA of each FHV genomic RNA were transfected into S2 cells in culture, and expression of genomic RNA was induced with copper sulphate. After 3 days, 10 ml of supernatant was retained. 1 ml of supernatant was passaged directly onto fresh S2 cells in culture, and FHV virions were purified from the remaining supernatant using sucrose cushions and sucrose gradients as described. This process was repeated for a total of nine passages. Encapsidated RNA was extracted from each stock of

purified virions and reverse transcribed using sequence-specific primers cognate to the 3' ends of FHV RNA1 and FHV RNA2. Full-length cDNAs were then PCR amplified in 19 cycles with NEB Phusion using primers targeting the 5' and 3' end of each genomic segment. Final PCR amplicons were used as input for Oxford Nanopore Amplicon sequencing protocol by ligating on native barcodes and then the ONT adaptor. Pooled libraries were sequenced on an ONT MinION MkIB with R9 (2017) flowcells. Data was demultiplexed and basecalled using the Metrichor software and *poretools*[89]. Datasets are publicly available in NCBI SRA under the accession code SRP094723. Basecalled data were mapped to the FHV Genomic RNAs (FHV RNA1: NC_004146, 3107 nts) and (FHV RNA2: NC_00414, 1400 nts) using *BBMap*. SAM files of the aligned data were passed to the CoVaMa (Ver 0.7) for analysis, using the command lines shown in Table 4.6.

Short-read data from longitudinal serum samples collected from one PWH analyzed using CoVaMa.

In a previous study in 2015, Chang *et al.* used an RT-PCR approach to amplify the *gag-pol* regions of HIV from 93 clinical specimens for NGS[14,37]. Briefly, two overlapping cDNA amplicons were RT-PCR amplified using two pairs of primers targeting *gag-pol* (F1: 737-754, R1: 1759-1736; F2: 1548-1569, R2: 2608-2589). Full-length cDNA amplicons were sheared to approximately 175 bp and submitted for paired-end sequencing (2x150bp) on an Illumina HiSeq. I selected one PWH from this previous study, from which five longitudinal samples were selected and sequenced. Paired-end reads were merged using *BBMerge* and mapped to the HIV genome (HXB2: K03455.1, 9719bp) using *ViReMa* (v0.21) with default settings plus the following parameters (*--X 3 --MicroInDel 5 –BackSplice_limit 25*)[76]. As the cDNA amplicon

strategy for Illumina sequencing was not directional, only reads mapping to the positive sense viral genome were used for further analysis. These SAM files were passed to CoVaMa (Ver 0.7) for analysis, using the command lines shown in Table 4.6.

Statistical analysis in CoVaMa outcomes.

Linkage disequilibrium was measured for every pairwise interaction of each nucleotide position with each detected recombination event within the viral genome using CoVaMa. To distinguish the significant associations from the background noise, the three-sigma rule for comparing a single data point to a very large distribution of other data points was applied. CoVaMa automatically generated the standard deviation and mean of all LD values, based on which the 3sigma could be calculated. Associations passing the three-sigma in each CoVaMa outcome were considered significant.

Table 4.5. Command lines used to analyze two viral datasets in this study.

Command lines used to analyze Nanopore sequencing reads acquired from each passage of FHV:
1. Making matrices containing contingency tables for each association using CoVaMa_Make_Matrices script: Output_Tag defined the name for the output pickle file [Output_Tag].Total_Matrices.py.pi. Data_directory defined the folder that contained the FHV reference sequence in FASTA file format and the aligned sequences in SAM file format. Contingency tables populated by over 100 reads and the mutant frequency higher than 0.05 were generated and passed to the output pickle file. *python2 CoVaMa_Make_Matrices.py * *[Output_Tag] Data_directory/FHV_Genome_corrected.txt * *–Mode2 All –SAM1 Data_directory/FHV_mapping.sam –PileUp_Fraction 0.05 * *–Min_Fusion_Coverage 10 NT*

2. Calculating linkage disequilibrium values for each contingency table using the CoVaM_ Analyse_Matrices script:

   [Output Tag].Total_Matrices.py.pi generated by CoVaMa_Make_Matrices script was used as the input. Linkage disequilibrium information was stored in the output file in TXT file format. For the contingency table to be analyzed, a minimum coverage of over ten pairs of associated nucleotides and recombination events was required. The LD and R-squared values for each association were normalized.

   *python2 CoVaMa_Analyse_Matrices.py [Output_Tag].Total Matrices.py.pi *
   CoVaMa output.txt –Min_Coverage 10 –Min_Fusion_Coverage 10 -OutArray -Weighted NT

Command lines used to analyze NGS sequencing reads acquired from each longitudinal HIV sample:

1. Making matrices containing contingency table for each association using the CoVaMa_Make_Matrices script:

   Output_Tag defined the name for the output pickle file [Output Tag].Total_Matrices.py.pi. Data_directory defined the folder that contained the HIV reference sequence for each sample in FASTA file format and the aligned sequences in SAM file format. Contingency tables populated by over 100 reads and mutant frequency higher than 0.01 were generated and passed to the output pickle file.

   a) For the 15-nt insertion event at nt 2158, 15 nucleotides were used to exclude negative recombination events. The analysis region started at nt 1958 and ended at nt 2358.
   b) For the 6-nt insertion event at nt 1149, six nucleotides were used to exclude negative recombination events. The analysis region started at nt 949 and ended at nt 1349.

   *python2 CoVaMa_Make_Matrices.py [Output_Tag] *
   *Data_directory/HIV_reference_seq.txt *
   *–Mode2 Recs –SAM1 Data_directory/HIV_mapping.sam *
   –PileUp_Fraction 0.01–Rec_Exclusion 15 –NtStart1958 –NtFinish 2358 NT

2. Calculating linkage disequilibrium values for each contingency table using the CoVaMa_Analyse_Matrices script:

   [Output_Tag].Total Matrices.py.pi generated by CoVaMa_Make_Matrices script was used as the input. Linkage disequilibrium information was stored in the output file in TXT file format. For the contingency table to be analyzed, a minimum coverage of 100 pairs of associated nucleotides and recombination events was required. The LD values and R-squared values for each association were normalized.

   *python2 CoVaMa_Analyse_Matrices.py [Output_Tag].Total Matrices.py.pi *
   *CoVaMa_output.txt –Min_Coverage 100 –Min_Fusion_Coverage 100 *
   -OutArray -Weighted NT

Chapter 5. Estimation of the optimal sample size needed for correlated mutation identification using an individual-virus-based simulator

5.1 Abstract

Drug resistance mutations (DRMs) and their associated compensatory mutations co-evolved within the HIV genome during antiretroviral therapy (ART) in people living with HIV (PWHs). DRMs are located in the enzyme active sites, impeding the inhibitor binding at the cost of enzyme stability and function. Conversely, compensatory mutations are selected outside the enzyme's active site, restoring the enzyme's stability and function. This interplay between correlated DRMs and compensatory mutations provides HIV with resistance and enhanced replication fitness. Chapters 2 and 3 investigated the identification of correlated mutations selected in the HIV *gag-pol* region during ART failures using long-read sequencing methodologies and downstream analysis tools.

Passenger mutations in the HIV genome are inherited from the founder viruses or acquired alongside DRMs. These mutations have no alteration on the HIV fitness under ART. However, when selecting HIV with DRMs, passenger mutations arising in the same HIV genome can also be selected and enriched, posing a challenge to correlated mutation identification.

Correlated DRMs and compensatory mutations tend to emerge consistently across PWHs treated with the same drug pressure. On the contrary, passenger mutations occur randomly and are unlikely to be shared across PWHs. Therefore, by utilizing samples from diverse PWHs treated with the same drug class, we can effectively eliminate passenger mutations, enhancing the accuracy of correlated mutation identification.

However, the methodology to determine the sample size needed for a statistically robust analysis has yet to be studied. Hence, this study provided a method to explore the relationship between sample size and the accuracy of correlated mutation identification. This method involves the generation of thousands of synthetic virological-failure HIV samples using a forward simulator and the identification of correlated mutations shared across samples.

The findings of this study revealed the precision and recall that could be achieved with a specific sample size. This, in turn, provided a practical approach for estimating the required sample size for future studies aiming to identify novel correlated mutations in the HIV genome.

5.2 Results

5.2.1 The Forward Simulator Effectively Simulates the Acquisition of Correlated Mutations in the HIV-1 Protease During Resistance Development

Forward simulators have been widely applied to mimicking natural mutation patterns, generating synthetic sequencing datasets for downstream analysis, and predicting the outcomes of evolutionary models[90–92]. In this study, I constructed an individual virus-based forward simulator to simulate the acquisition of correlated DRMs and compensatory mutations in the HIV genome under drug pressure.

The HIV protease is one of the three key enzymes crucial for the virus's fitness during ART. Additionally, correlated DRMs and compensatory mutations have been reported in the HIV protease resulting from ART[44]. Therefore, in my initial studies, I solely considered the contribution of HIV protease to HIV fitness under ART and focused on simulating the evolution of the HIV protease gene under protease inhibitor pressure. Overall, the forward simulator should be able to:

1) Use a number of *protease* sequences as starting materials.

2) Simulate the evolution of each *protease* sequence under drug pressure.

3) Generate synthetic virological-failure viral populations.

Additionally, this study focused on non-synonymous mutations, which influenced HIV protein structure and function, and ignored synonymous mutations. Hence, if not specified, all "mutations" mentioned below refer to non-synonymous mutations only.

Previous research has shown that HIV protease DRMs are usually selected in the protease's active site due to their changes in the enzyme structure that impede inhibitor binding[4]. In contrast, compensatory mutations are located distal to the protease's active site[93,94]. Based on this, I generated synthetic DRMs and synthetic compensatory mutations in the HIV protease. The synthetic DRMs were confined to the loci where canonical DRMs have been detected[39], while synthetic compensatory mutations were positioned distal to these canonical DRMs (Methods).

To mimic the observed correlation between protease DRMs and compensatory mutations[13,95], I started with pairing synthetic DRMs and synthetic compensatory mutations in a one-to-one ("1-to-1") relationship. This "1-to-1" relationship reflected the scenario in which a specific compensatory mutation can mostly counteract the fitness decline resulting from acquiring one DRM[13,95]. Using this strategy, I constructed a mutation dataset containing linked synthetic DRMs and synthetic compensatory mutations.

Two equations were established to quantify the fitness of HIV with different mutation combinations in protease and provide synthetic mutation pairs with a fitness advantage over passenger mutations under drug pressure. These equations together quantified the average number of progenies (*R)* that an HIV with protease sequence *g* could produce based on four fitness-related factors: the total number of mutations in protease (mutational burden), drug pressure, DRMs, and compensatory mutations.

$$R = R_0 * 2/(1 + e^{Ef(p)-x}) \quad (5.1)$$

$$x = DR_{DRM}(g, p) - RC_{DRM}(g) + RC_{CM}(g) - RC_{per-mutation} * MB(g) \quad (5.2)$$

In Equation 5.1, R_0 represents the average number of progenies a "wild type" HIV could generate in the untreated condition. For an HIV with protease sequence g, its mutations' combined effect (x) counteracts drug pressure's effect, *Ef(p)*, which together determines its fitness (R). Equation 5.1 assures that the resulting R of a "wild-type" HIV under no drug pressure equals R_0. Equation 5.2 quantifies the combined effects of all mutations in protease sequence g on its fitness:

1) DRM provides drug resistance and thus improves fitness: $+ DR_{DRM}(g, p)$.

2) DRM destabilizes protein structure and thus decreases fitness: $- RC_{DRM}(g)$.

3) Compensatory mutation improves fitness by compensating for DRM: $+ RC_{CM}(g)$.

4) Mutational burden decreases fitness. This reduction is calculated by multiplying the total protease mutation number ($MB(g)$) by the fitness loss of getting a single mutation ($- RC_{per-mutation}$).

In *in vitro* experiments, it was observed that the drug-resistant HIV isolates grown under drug pressure replicated less efficiently than the 'wild-type' HIV isolates grown without drug pressure[96]. Based on this observation, I designed an extreme scenario where the drug-resistant HIV genome had only one DRM and one compensatory mutation in protease, without any additional mutations. By assuming that one compensatory mutation could fully counteract the fitness reduction caused by one DRM (the "1-to-1" relationship), the presence of the compensatory mutation in this specific drug-resistant HIV could restore its fitness in the

drug-treated condition to the level of the 'wild-type' HIV in an untreated condition. Consequently, a synthetic drug-resistant HIV under drug pressure, harboring only one DRM and one compensatory mutation in protease but no other mutations, would have an *R* that is equal to (or infinitely close to) R_0.

Additionally, I assessed the maximum level of protease mutational burden (number of total protease mutations) that drug-sensitive HIV genomes and drug-resistant HIV genomes could sustain while maintaining efficient replication. To gather this data, I acquired HIV drug-naive *protease* sequences[18,19,97–105], along with treatment-experienced HIV *protease* sequences collected during non-suppressive ART[106–110], from the HIVDB SGS program. About 95% of drug-naive proteases had ten mutations or fewer (Figure 5.1A). This indicated that drug-naive HIVs with up to ten protease mutations maintained normal replication, while a higher mutation count could potentially destabilize protease structure and impair enzymatic function. This empirical data informed the simulator equations, indicating that synthetic drug-naive HIV with ten or fewer protease mutations should have an *R* greater than 1 in untreated conditions. Similarly, within treatment-experienced *protease* sequences, 95% had 12 protease mutations or fewer, including both active and non-active site mutations in protease (Figure 5.1A). This indicated that drug-resistant HIVs with up to 12 protease mutations still had normal replication, while an increased mutation load would presumably destabilize protease structure and impair enzymatic function. This empirical data informed the simulator equations, indicating that synthetic drug-resistant HIVs with 12 or fewer protease mutations should have an *R* above 1 in the drug-treated condition.

The assigned values to variables in the equations ensure that the simulation results are consistent with empirical mutational burden data and the extreme scenario mentioned earlier. A synthetic drug-resistant HIV with only one pair of correlated DRM and compensatory mutations in protease had an R equal to R_0. Additionally, in the drug-untreated condition, synthetic drug-sensitive HIV *protease* sequences ($DRM^-CM^-_{untreated}$, "-": absent) with ≤ 10 mutations resulted in an R value above 1, indicating efficient replication. Similarly, under drug-treated conditions, synthetic drug-resistant HIV *protease* sequences ($DRM^+CM^+_{treated}$, "+": present) with ≤ 12 mutations yielded an R value above 1 (Figure 5.1B).

Furthermore, the values assigned recaptured the fitness increase during the acquisition of correlated mutations under drug pressure (Figure 5.1C): As can be seen, under drug pressure, the synthetic drug-sensitive *protease* sequence ($DRM^-CM^-_{treated}$) yielded the lowest fitness; the addition of one DRM ($DRM^+CM^-_{treated}$) improved fitness, which was further enhanced by acquiring one compensatory mutation ($DRM^+CM^+_{treated}$). Additionally, these equations captured the fitness decrease with increasing mutational burden in the protease (Figure 5.1C).

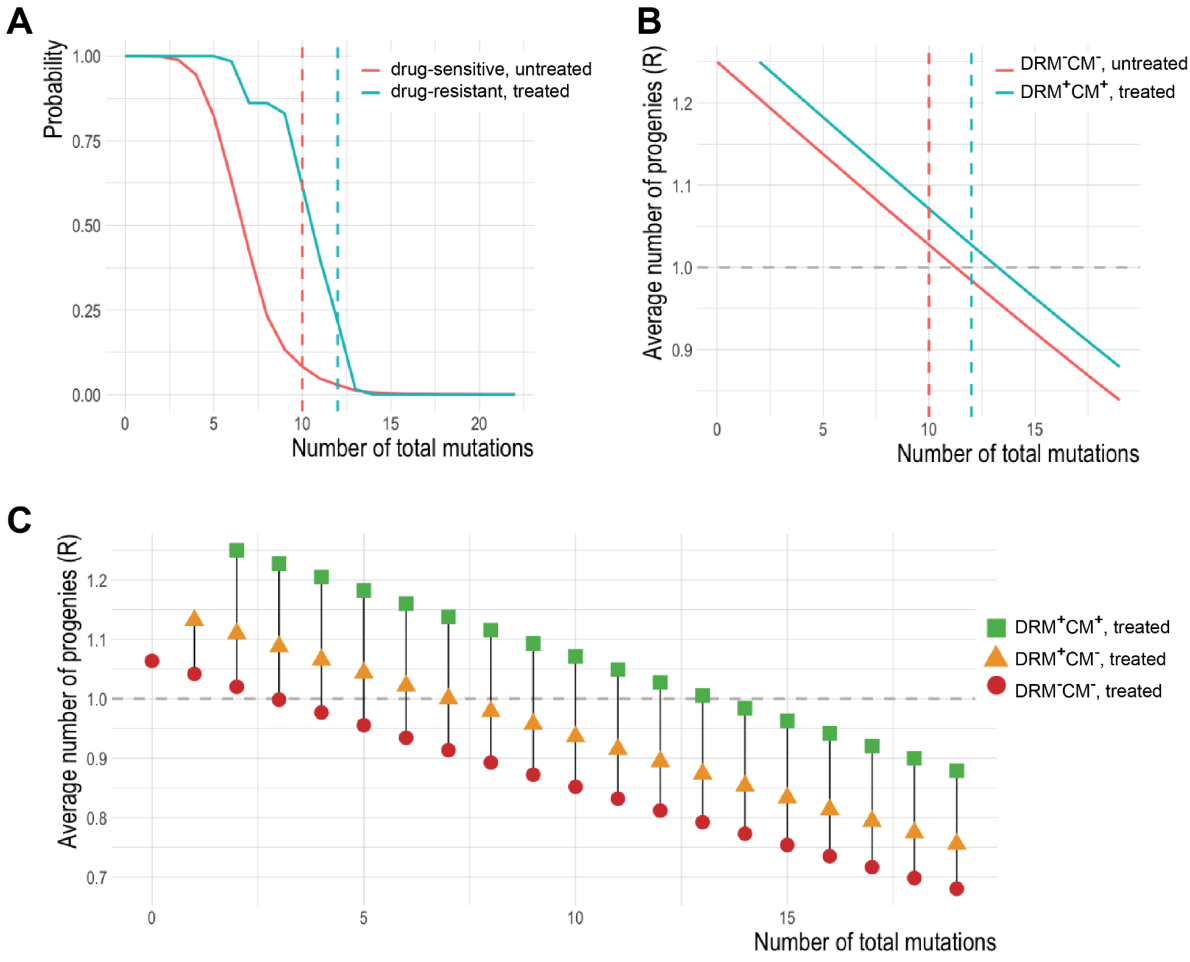

Figure 5.1. The simulator recaptures the fitness increase along with the acquisition of correlated mutations under drug pressure.

(A) Using 1757 drug-naive HIV *protease* sequences collected from 322 drug-naive plasma samples acquired from PWHs, I quantified the number of non-synonymous mutations in drug-sensitive HIV proteases in the untreated condition. Similarly, using 431 drug-resistant HIV *protease* sequences collected from 74 virological-failure plasma samples acquired from PWHs, I quantified the number of non-synonymous mutations in drug-resistant HIV proteases in the drug-treated condition. Using this empirical data, I plotted the probability of observing a certain number of mutations in drug-sensitive HIV proteases in the untreated condition (red) and drug-resistant HIV proteases in the treated

condition (green). Two vertical dashed lines at 10 or 12 mutations represent 95% thresholds — 95% of cases would fail on the left of this dashed line. The vertical dashed lines are colored in red for drug-sensitive HIV proteases or green for drug-resistant HIV proteases. (B) Based on the empirical data, I assigned appropriate values to variables within the two equations. After this, using these two equations, I calculated the fitness of HIVs that possessed either group of synthetic protease sequences and had different levels of protease mutational burden: untreated drug-sensitive ($DRM^{-}CM^{-}_{untreated}$) and treated drug-resistant ($DRM^{+}CM^{+}_{treated}$). The vertical dashed lines at 10 and 12 mutations were adapted from Figure A. The gray horizontal dashed line highlights the scenario wherein one viral sequence generates, on average, one progeny viral sequence. (C) Using these equations, I calculated the fitness of HIVs that had three distinct groups of synthetic *protease* sequences and different levels of protease mutational burden under drug pressure: $DRM^{-}CM^{-}_{treated}$, $DRM^{+}CM^{-}_{treated}$, and $DRM^{+}CM^{+}_{treated}$. Those having the same levels of protease mutational burden (number of total mutations) in these three groups were grouped and compared with each other.

With variables within the two equations properly assigned, the fitness of HIVs with different synthetic *protease* sequences was quantified, considering the presence or absence of correlated mutations, drug pressure, and the extent of protease mutational burden. As previously stated, only the protease's contribution to HIV fitness was considered, and the other regions in the HIV were ignored while constructing the simulator.

The simulator initiated with individual HIV *protease* sequences stored in a FASTA file and simulated their evolution over generations. In each generation, individual *protease* sequences would mutate, recombine, and replicate (Figure 5.2A). At the end of each generation, a new sequence pool was formed and passed to the next generation.

Two different treatment outcomes were observed in the clinic: viral rebound due to the selection of drug resistance or viral suppressio[111,112]. According to these observations, the simulator monitored the size of the simulated viral population over generations and stopped in either of the following two conditions. In the first condition, the size of the simulated viral population should exceed a given threshold, e.g., 30,000 HIV genomes. This was considered a simulated viral rebound, and the simulator stopped and collected the simulated viral population. In the second condition, the size of the simulated viral population should be continuously lower than a given threshold, e.g., 100 HIV genomes, for at least three generations[113]. This was considered a simulated viral suppression, and the simulator would stop as well (Figure 5.2B). The simulator is now available for public use on GitHub.

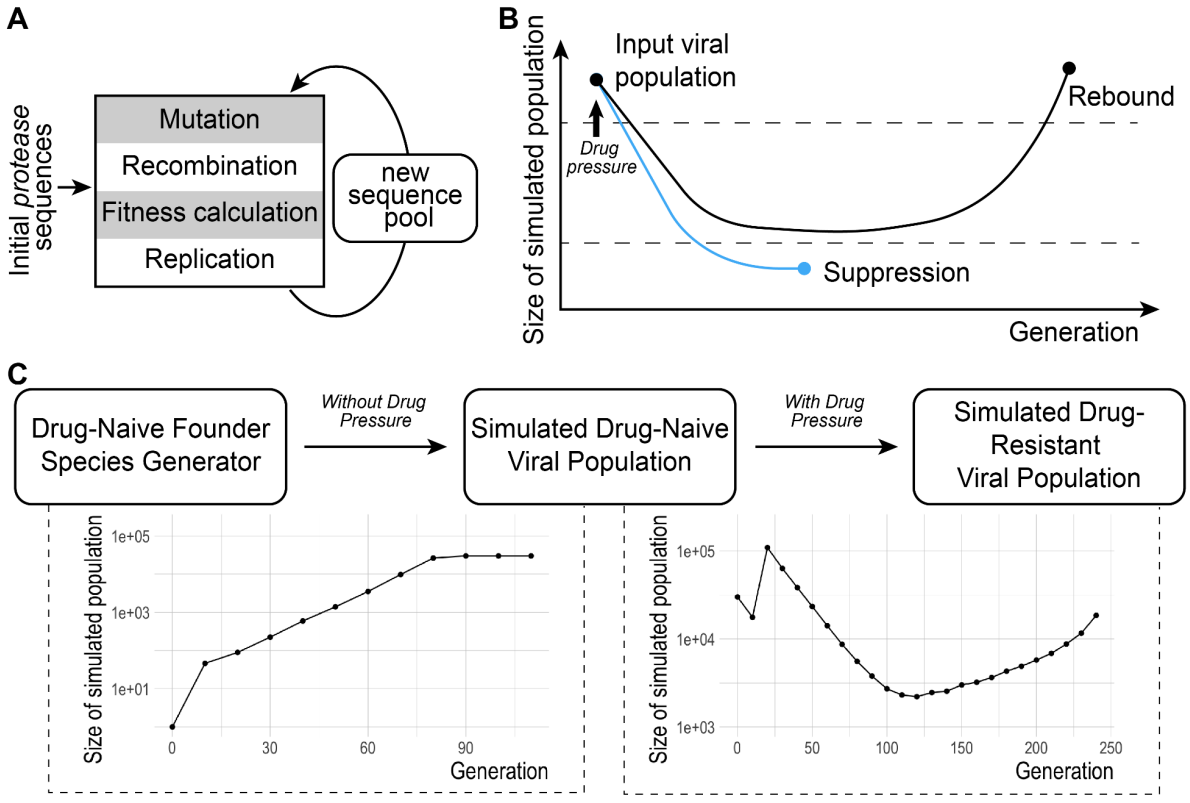

Figure 5.2. Overview of the simulator and the simulation pipeline.

(A) The simulator simulates viral evolution over generations. In each generation, individual viral sequences go through four steps sequentially: mutation, recombination, fitness calculation, and replication. After each generation, a new simulated viral population is generated and passed to the next generation. (B) The simulator monitors the size of the simulated viral population over generations and stops in two conditions: simulated viral rebound (black) and simulated viral suppression (blue). The x-axis represents the generations, and the y-axis represents the size of the simulated population. The two gray horizontal dashed lines represent the thresholds defining simulated viral rebound and simulated viral suppression, respectively. (C) The simulation pipeline is constructed with three stages. First, staring *protease* sequences are generated using an HIV-1 drug-naive *protease* sequence generator. Thereafter, the first simulation starts from each synthetic starting sequence with drug pressure turned off, generating one synthetic drug-naive viral population. Lastly, the second simulation starts from each synthetic drug-naive viral population with the drug pressure turned on, generating a synthetic virological-failure viral population. Typical changes in simulated population size during the first and second simulations are shown below the workflow.

With the simulator constructed, I used it as a module and built the entire pipeline to generate synthetic virological-failure viral populations (Figure 5.2C). To achieve this, first, I generated starting *protease* sequences using an HIV drug-naive *protease* sequence generator (Methods). Thereafter, I used each synthetic starting sequence to start the first simulation with no drug pressure. This generated synthetic drug-naive viral populations. In the end, each synthetic drug-naive viral population was used to start the third simulation with drug pressure, which generated the synthetic virological-failure populations. These synthetic virological-failure viral populations would be used to explore the sample size needed for correlated mutation identification.

5.2.2 The Simulator Mimics Viral Rebound Resulting from the Selection of DRMs and Recaptures the Linkage Distribution in Drug-Resistant HIV Protease Sequences

To assess the performance of the simulation pipeline, I utilized it to generate synthetic drug-naive viral populations and synthetic virological-failure viral populations. I quantified the number of non-synonymous mutations in these synthetic HIV *protease* sequences. As anticipated from empirical data, synthetic drug-experienced HIV *protease* sequences showed a better tolerance for higher mutational burden than synthetic drug-naive HIV *protease* sequences (Figure 5.3A).

In addition, I selected three representative synthetic drug-naive viral populations and three synthetic virological-failure viral populations to quantify the abundance of built-in DRMs and compensatory mutations in these synthetic viral populations. Synthetic drug-naive viral populations showed no detected DRM (0% ± 0%, SD) and a low frequency of compensatory mutations (0.47% ± 0.37%, SD). On the contrary, DRMs and compensatory mutations were highly enriched in synthetic virological-failure viral populations (96% ± 3.6% for DRMs and 85.1% ± 13.7% for compensatory mutations, SD). This indicated that the simulated viral rebound was a result of the selection of DRMs and compensatory mutations under drug pressure.

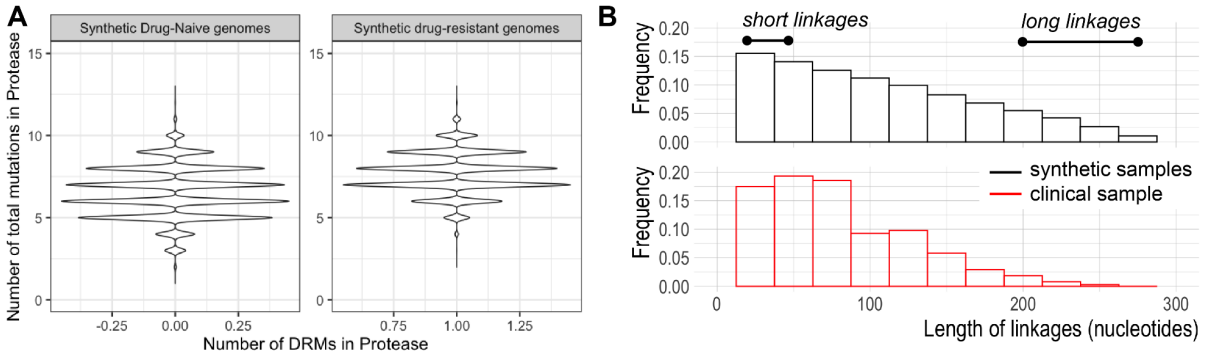

Figure 5.3. The simulator recaptures the linkage distribution in drug-resistant *protease* sequences.

Using the simulator pipeline and five pairs of synthetic linked mutations, I generated 1690 synthetic drug-naive populations and 1011 synthetic virological-failure populations. (A) From each population, 100 sequences are randomly selected and aligned to the HXB2 *protease* sequence. The total number of non-synonymous mutations encoded is quantified in every synthetic drug-naive *protease* sequence (with 0 DRM) and every synthetic drug-resistant *protease* sequence (with one DRM). The quantification results are plotted using violin plots. (B) To compare the linkages detected in synthetic virological-failure populations to those observed in clinical virological-failure populations, I sampled each of the 1011 synthetic virological-failure populations with a sampling depth of 1000 *protease* sequences and added a sequencing mismatch rate of 0.02% to the *protease* sequences selected. Using CoVaMa[34], I identified the linkage disequilibrium in 1011 synthetic virological-failure samples and one clinical virological-failure sample. The size of linkages detected in each sample was quantified. The distribution of the size of linkages in synthetic samples (black) and the clinical sample (red) is plotted using a histogram with relative frequencies.

Furthermore, to validate whether the synthetic drug-experienced *protease* sequences recaptured the size distribution of linkages observed in clinical samples, I compared the synthetic

virological-failure samples to a clinical virological-failure sample sequenced in a previous study[25]. To mimic the sampling and sequencing of clinical samples, from each synthetic virological-failure viral population, I randomly selected 1000 *protease* sequences and introduced a sequencing mismatch rate of 0.02% to these sequences (Methods)[25]. This process yielded 1011 synthetic virological-failure samples. Thereafter, I calculated the linkage disequilibrium in each synthetic virological-failure sample, as well as in the clinical virological-failure sample, using the LD calculator CoVaMa[34]. CoVaMa provided a list of linked SNVs detected in each sample. Quantification results showed a similar trend to the clinical sample (Figure 5.3B), where shorter linkages (less than 100 nucleotides in length) were more frequent compared to longer ones in the synthetic virological-failure samples. Additionally, the abundance of linkages decreased quickly with increasing linkage length (Figure 5.3B).

The above results indicated that our simulator successfully mimicked the selection of DRMs and compensatory mutations in HIV protease under drug pressure and recaptured the linkage distribution in drug-experienced HIV *protease* sequences.

5.2.3 Sample Size Needed for Identifying Correlated DRMs and Compensatory Mutations Estimated Using Synthetic Samples

To identify correlated mutations from different samples and eliminate passenger mutations, I developed an analyzing approach (Figure 5.4). This approach involved two filtering steps to identify candidate mutation pairs that were frequently observed among samples and were enriched in those samples. After these filterings, the approach constructed a 2x2 contingency

table for every candidate mutation pair, populated the contingency table using sequences from all input samples, and then calculated the pairwise association using Fisher's exact test.

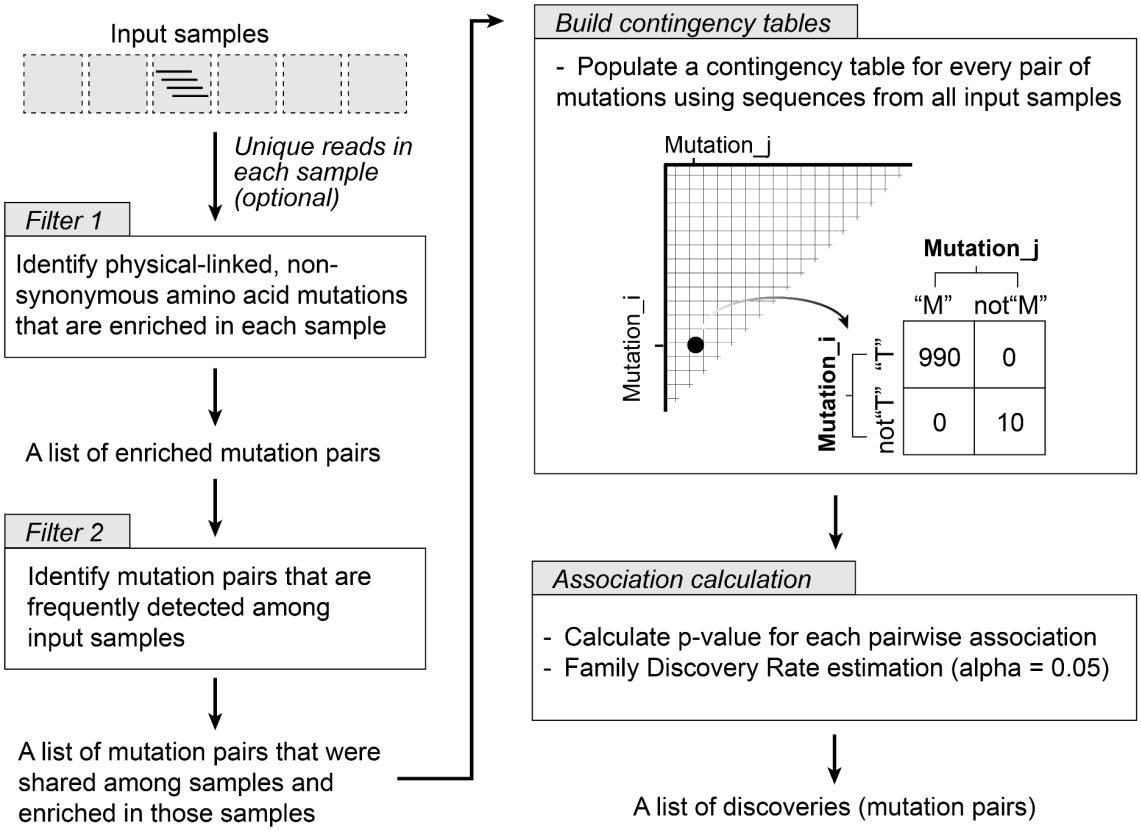

Figure 5.4. Overview of the downstream analysis pipeline.

In each input sample, all physical-linked amino acid mutations that are enriched to a certain frequency, e.g., 50%, are identified (Filter 1). This generated a list of mutation pairs that are enriched in at least one input sample. Then, all the enriched mutation pairs are organized in descending order according to the count of input samples in which they are enriched. The top few mutation pairs that are most frequently detected among input samples are selected (Filter 2). After these two filters, a list of mutation pairs that are both shared among samples and enriched within those samples is determined. For every mutation pair, a contingency table is constructed and populated using sequences in all input samples. Fisher's exact test is performed to calculate the pairwise association based

on each contingency table. In the end, a Family Discovery Rate estimation[114] is performed using the *Python* library *statsmodels*[115], which generates a list of the final discoveries (mutation pairs).

To assess the relationship between the sample size and the analyzing accuracy, I conducted tests using varying numbers of synthetic samples (ranging from 2 to 200) while keeping all other settings constant among tests. Synthetic samples were generated by the simulation pipeline, utilizing five pairs of built-in linked DRMs and compensatory mutations, with a sequencing depth of 1000 genomes and a sequencing mismatch rate of 0.02%. In all tests, Filter 1 selected mutation pairs that were enriched to over 50% in at least one sample, while Filter 2 selected the top 75 mutation pairs that were most frequently detected among samples and enriched in those samples. Following each test, the identified mutation pairs were compared to the predefined five built-in mutation pairs, generating precision and recall values (as detailed in the Methods section). The test outcomes revealed a positive correlation between sample size and resulting precision and recall values. As the sample size increased, recall improved from around 30% to 100% (Figure 5.5A), and precision increased from approximately 5% to around 12.5% (Figure 5.5B).

In an alternative approach, rather than using all sequences in each sample as described above, I conducted additional tests employing only unique reads with a read count ≥ 2 in each sample. Statistically, using unique reads would decrease the abundance of random passenger mutations in the population without significantly impacting the frequency of fitness-related mutations. The new tests were also conducted using varying numbers of synthetic samples (ranging from 2 to 200), and the precision and recall were calculated following each test. Compared to using all the

sequences in each sample, using unique sequences in each sample showed no significant effect on the recall (Figure 5.5A, Figure 5.5C). However, it increased the precision of the analysis (Figure 5.5B, Figure 5.5D). Hence, for subsequent analyses, I continued using unique reads with a read count ≥2 in each sample for analysis.

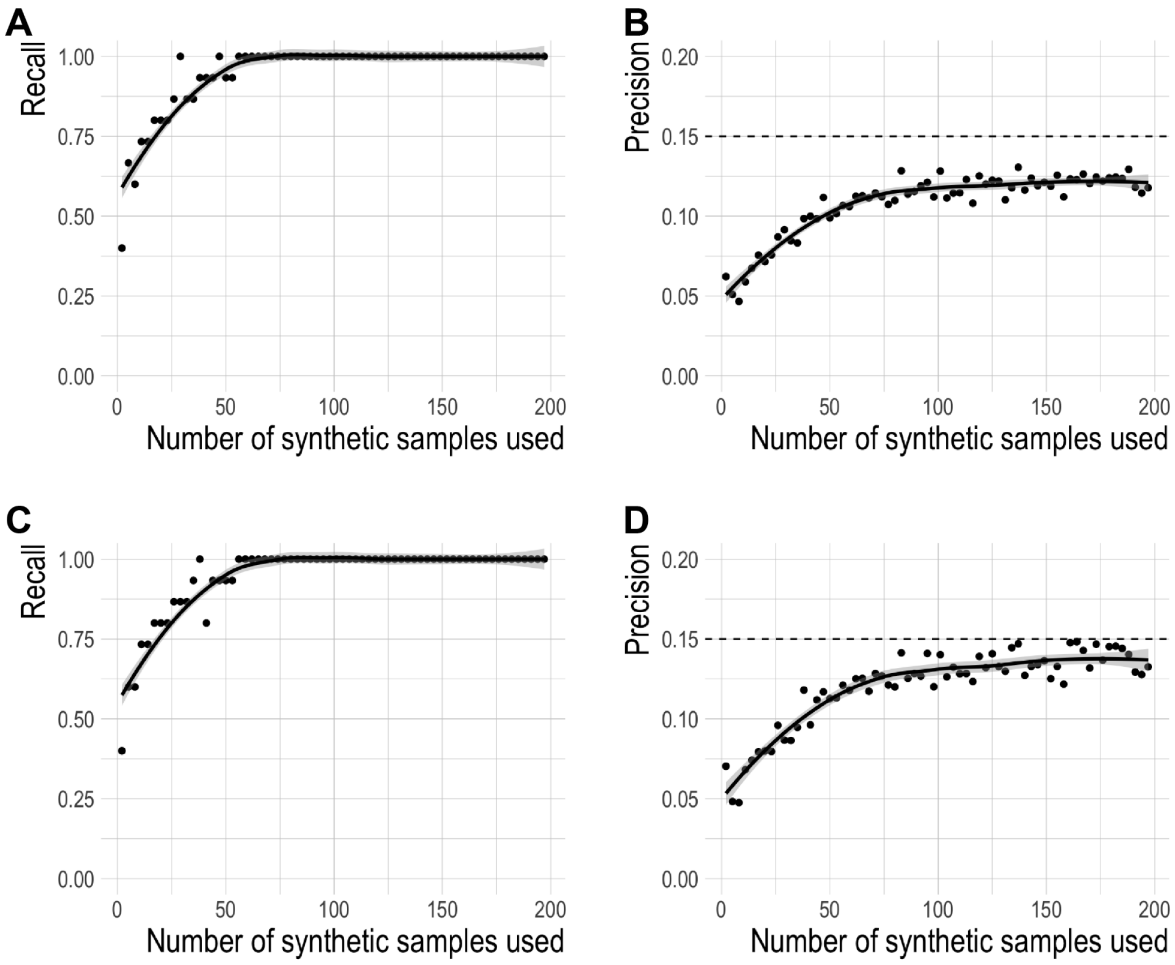

Figure 5.5. Using the unique reads with a read count ≥ 2 in each sample improves the precision of correlated mutation identification.

Using the simulation pipeline and five pairs of built-in linked mutations, I generated over 1000 synthetic virological-failure samples. (A) First, I used all the sequences in each

sample for analysis. In Filter 1, mutation pairs that were enriched in 50% in each sample were selected. In Filter 2, the top 75 mutation pairs that were most frequently detected among samples were selected. I conducted tests with sample size ranging from 2 to 200. Each test was repeated three times independently, and the average precision and recall were calculated. The recall changing with the sample size is plotted using a scatter plot (*geom_point* in *ggplot2)*. Additionally, a smooth line is generated using the local regression method LOESS (*geom_smooth*). Moreover, the precision changing with the number of samples used is plotted in (B). Secondly, I kept the two filters unchanged but only used the sequences with a read count ≥ 2 in each sample. Using sample numbers ranging from 2 to 200, I performed a series of analyses and calculated the precision and recall of each analysis. The recall change with the number of samples used is plotted in (C), as well as the precision change with the number of samples used (D), using scatter plots and smooth lines.

To explore the optimal precision-recall balance that could be achieved with a specific sample size, I fine-tuned Filter 2 in the analysis pipeline. A more stringent Filter 2 would theoretically eliminate more false positives (mutation pairs involving passenger mutations) and increase precision. However, it would also increase the probability of discarding true positives (correlated DRMs and compensatory mutations), reducing recall.

Using over 1000 synthetic samples generated with five pairs of built-in linked DRMs and compensatory mutations, I conducted tests with varying sample sizes (10, 20, 30, 80, and 200 samples) and explored Filter 2 adjustment ranging from 10 to 200 (Figure 5.6A). Precision and recall were calculated following each test, as described above. As expected, the test outcomes showed a trade-off between precision and recall at a specific sample size. Additionally, using more samples resulted in an overall increase in both precision and recall. For instance, with only

ten samples, the upper precision limit reached roughly 12.5%, accompanied by a low recall of 25%. However, by increasing the sample size to 200, precision increased to over 20% while maintaining a recall of 93%.

The number of correlated DRMs and compensatory mutations (evolutionary pathways) that HIV could use to evade drug pressure also influences this analysis. To investigate the impact of pathway numbers on analyzing accuracy, I generated over 1000 synthetic virological-failure samples using ten pairs of built-in linked DRMs and compensatory mutations.

Using synthetic samples generated with these ten pairs of linked mutations, I conducted tests with varying sample sizes (30, 80, 150, 200, and 350 samples) and a Filter 2 ranging between 10 and 400 (Figure 5.6B). Much like the observations from the five-pathway dataset, I observed the trade-off between precision and recall when a specific number of samples was used, as well as the positive correlation between sample size and overall precision/recall. For ten pathways, using 200 samples yielded a precision of ~12% and a recall of ~60%. Compared to having five pathways, it was observed that having a total of ten pathways resulted in a reduction in precision. This finding indicated the increased difficulty in identifying an increasing number of pathways.

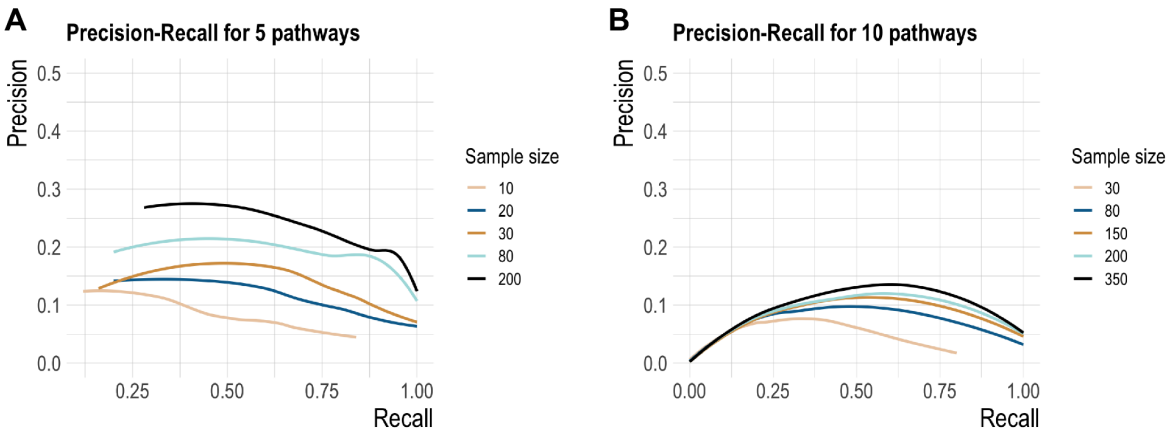

Figure 5.6. Precision-recall trade-off across varies sample sizes.

(A) Using 1011 synthetic virological-failure samples generated with five pairs of built-in linked DRMs and compensatory mutations, I conducted tests with varying sample sizes (10, 20, 30, 80, and 200 samples) and a Filter 2 ranging from 10 to 200. Each test was repeated five times independently, and the average precision and recall were calculated. With the test outputs, Precision-Recall plots are generated using *geom_smooth* in *ggplot2* and a local regression method LOESS. (B) Using 1057 virological-failure samples generated with ten pairs of built-in linked DRMs and compensatory mutations, I conducted tests with varying sample sizes (30, 80, 150, 200, and 350 samples) and a Filter 2 ranging from 10 to 400. These tests were repeated three times independently. The average precision and recall were calculated for each test. With the test outputs, Precision-Recall plots are generated using *geom_smooth* in *ggplot2* and a local regression method LOESS.

5.3 Discussion

During ART, correlated DRMs and compensatory mutations are selected within the HIV genome, contributing to enhanced resistance and viral fitness. Understanding these correlated mutations could provide insight into the evolutionary trajectories individual HIV took during resistance development and ART failures. However, passenger mutations that arose in the same HIV genome with resistance mutations could be simultaneously selected and enriched, thus interfering with the correlated mutation identification.

In this study, I aimed to improve the accuracy of correlated resistance mutation identification by using samples from different PWHs treated with the same drug class (independent samples). Additionally, I explored the sample size needed for a statistically robust analysis. To accomplish these goals, I utilized an individual virus-based forward simulator to generate synthetic virological-failure samples. Additionally, I provided an analytical approach to identify linked mutations that were shared among samples and enriched in those samples.

My findings showed a positive correlation between the sample size and the precision/recall of identifying correlated DRMs and compensatory mutations. Furthermore, the findings of this study provided a practical method for estimating the sample size required to identify correlated mutations within the HIV genome at a specified level of precision. For example, assuming we were to analyze resistance development against one antiviral drug class, HIV might acquire resistance via one of five possible pairs of correlated mutations. If we would like to test no more than 30 candidate mutation pairs and capture at least four out of the five pairs of correlated mutations, the simulation findings suggested that using around 30 samples from different PWHs would be necessary to achieve the desired results.

I compared the analyzing results of various synthetic samples generated with different numbers of built-in correlated DRMs and compensatory mutations, where each pair of correlated mutations represented a distinct evolutionary pathway. This study revealed a negative relationship between the number of pathways and the precision and recall of the correlated mutation identification. When considering five pathways, utilizing around 30 samples would enable the identification of four out of five pairs of correlated mutations by testing fewer than 30 candidate mutation pairs. However, if the number of pathways increased from five to ten, in the best scenario, we could achieve a precision of ~13.6% and a recall of ~60%. Moreover, to achieve that result, we would need to collect around 350 independent samples and test around 45 candidate mutation pairs. This indicated that more samples would be required to reveal more complex ART-evading mechanisms.

This study focused on the "1-to-1" relationship between compensatory mutations and DRMs, wherein a specific compensatory mutation can largely counteract the fitness decline resulting from acquiring one DRM. Examples of this "1-to-1" relationship have been observed in the sequencing and functional assessment of clinical HIV samples; for instance, the decrease in protease (PR) enzymatic function caused by the acquisition of DRM protease PR D30N could be restored by acquiring PR N88D in protease[13,95]. However, the relationship between complementary mutations and DRMs can be more complex. First, the replication fitness loss caused by one PR DRM might be compensated by diverse compensatory mutations operating through various mechanisms. For instance, PR D30N can be compensated by PR M36I, PR A71V, or PR N88D, which could affect protease monomer/dimer stability or influence protease structure[116]. Additionally, mutations in Gag substrate, e.g., Gag P453L, can also complement PR

D30N and restore fitness, presumably by improving protease-Gag binding[117]. To consider this complex series of complementary mutations, we would need to build a new synthetic mutation dataset, wherein one PR DRM could cooperate with any of several compensatory mutations. Secondly, with the increase in the potency of antiviral drugs, more than one compensatory mutation may be involved in compensating for one DRM[14]. Therefore, a more comprehensive, "1-to-n" relationship is worth investigating. Thirdly, DRMs with varying levels of resistance can cooperate to provide the virus with drug resistance. An example is Thymidine Analog Mutations (TAMs) in the reverse transcriptase, which provide resistance to RT inhibitors. This observation suggests another "n-to-n" relationship between DRMs and compensatory mutations in HIV enzymes.

With the individual virus-based forward simulator constructed, different types of relationships between DRMs and complementary mutations (such as "1-to-1", "1-to-n", "n-to-n") can be developed and incorporated as plug-in motifs for distinct research objectives in subsequent stages. Furthermore, besides investigating the acquisition of correlated mutations in the HIV *protease* region, this approach can be used for other HIV proteins (reverse transcriptase, integrase) and extended for correlated mutations across HIV proteins, e.g., between Gag and protease.

In this study, only virological-failure samples were considered for correlated mutation identification. However, the accuracy of this analysis can further be improved by incorporating drug-naive samples as well. First, by using a consensus sequence constructed based on drug-naive sequences instead of the HIV-1 HXB2 sequence used in this study, we could remove

the mismatch between clinical sequences and the HXB2 sequence. Secondly, physical linkages commonly detected in drug-naive samples can be used to infer false positives in the discoveries.

To summarize, this study provides a better understanding of the relationship between the number of samples used, the number of mutational drug-resistant pathways, and the accuracy of identifying correlated resistance mutations. The use of the simulator-based method allows for estimating the required sample size for specific studies in the future.

5.4 METHODS

Design synthetic linked DRMs and compensatory mutations.

From HIVDB, I collected canonical DRMs detected in HIV-1 protease[39]. These canonical DRMs are located on 12 loci in the protease: 30, 32, 46, 47, 48, 50, 54, 76, 82, 84, 88, and 90. A total of 71 synthetic DRMs ("DRM pool") on these 12 loci were generated using a custom *Python* script. Considering that compensatory mutations are located distal from DRMs, I selected the loci that were three amino acids away from all of the above 12 DRM loci. This generated 54 loci for synthetic compensatory mutation generation, based on which a total of 309 synthetic compensatory mutations were generated ("CM pool").

To generate five pairs of linked mutations, five synthetic DRMs were selected from the "DRM pool" without replacement, and five synthetic compensatory mutations were selected from the "CM pool" without replacement. The selected synthetic DRMs and compensatory mutations were paired in a 1-to-1 relationship. In this way, five unique pairs of linked synthetic DRMs and compensatory mutations were generated. The exact process was used for generating ten pairs of linked mutations.

Collect drug-naive and drug-experienced HIV-1 protease sequences *from HIVDB.*

To quantify the total number of non-synonymous mutations encoded in drug-naive HIV *protease* sequences and drug-resistant HIV *protease* sequences, I collected data from the HIV-1 Single Genome Sequence (SGS) Database in HIVDB[39]. I used the website filter to find samples that were infected by HIV-1 subtype B. A total of 322 drug-naive plasma samples were collected from 142 drug-naive individuals enrolled in 11 studies. From these 322 drug-naive plasma

samples, I acquired 1757 drug-naive HIV *protease* sequences. A total of 74 virological-failure plasma samples were collected from 27 individuals with unsuppressed viral replication enrolled in five studies. From these 74 virological-failure plasma samples, I acquired 430 HIV drug-resistant HIV *protease* sequences.

Assign values to variables within two equations designed for fitness calculation.

Considering that 'wild-type' HIV could replicate well in the untreated condition, R_0 should be higher than 1. In this study, a constant value of 1.25 was given to the R_0. Additionally, 0.3 was assigned to $Ef(p)$, quantifying the effect of drug pressure on viral fitness. In this study, I also assumed that the extra reduction in fitness caused by getting one DRM is 10-fold that caused by getting one mutation.

I quantified the number of mutations in drug-naive *protease* sequences collected from HIVDB. Empirical data informed that drug-sensitive HIV *protease* sequences in the untreated condition ($DRM^-CM^-_{untreated}$) with up to 10 mutations maintained normal replication. Based on this, I solved the equation $R = 1.25 * 2/(1 + e^{RC_{per-mutation}*10}) \geq 1$ and acquired the upper limit of $RC_{per-mutation}$ of 0.0405. Hence, I assigned 0.036 to $RC_{per-mutation}$. This assured that the $DRM^-CM^-_{untreated}$ would have an R higher than 1 if it had ten mutations or fewer, an R around 1 if it had 11 mutations, and an R below 1 if it had 12 mutations or more.

All variables involved in the fitness calculation of drug-sensitive HIV ($DRM^-CM^-_{untreated}$ and $DRM^-CM^-_{treated}$) had values assigned. Thus, we could calculate the fitness (R) for HIV with various drug-sensitive *protease* sequences with varying levels of mutational burden in the treated

or untreated condition. Additionally, the $RC_{DRM}(g)$ was defined as 10-fold of $RC_{per-mutation}$ to be 0.36.

In the extreme scenario, an assumption was made that a drug-treated HIV with only one DRM and one corresponding compensatory mutation in the protease would have an *R* equal R_0. Based on this, I solved the equation:

$Ef(p) = DR_{DRM}(g, p) - 0.36 + RC_{CM}(g) + 2 * 0.036, Ef(p) = 0.3$. This resulted in $DR_{DRM}(g, p) + RC_{CM}(g) = 0.732$ (Equation 1).

In this study, I assumed that the fitness increase from $DRM^{-}CM^{-}_{treated}$ to $DRM^{+}CM^{+}_{treated}$ was attributed to the DRM accounting for 60%, while the corresponding compensatory mutation contributed the remaining 40%. With this assumption, the intermediate stage, $DRM^{+}CM^{-}_{treated}$, would have an $R \geq 1$ with seven mutations or fewer. Based on this, I solved the equation $R = 1.25 * 2/(1 + e^{0.3-x}) \geq 1, x = DR_{DRM}(g, p) - 0.36 - 0.036 * 7$ and acquired the upper limit of $DR_{DRM}(g, p)$ of 0.507. Hence, I assigned 0.507 to $DR_{DRM}(g, p)$. Based on Equation 1 above, 0.225 was assigned to $RC_{CM}(g)$.

Generation of synthetic drug-naive protease sequences using an HIV-1 drug-naive protease sequence generator.

To generate synthetic drug-naive HIV *protease* sequences, I acquired the nucleotide distribution on each locus of the 297-nt HIV *protease* sequence. To fulfill this, I collected HIV *protease* sequences from a total of 83 drug-naive plasma/serum samples collected from different PWHs.

All PWHs were infected with HIV-1 subtype B. Among these 83 samples, nine of them were provided by the UC San Diego Primary Infection Resource Consortium (PIRC) (David Smith, M.D.) and the UW Viroverse study (James Mullins, Ph.D.). These nine samples were sequenced in our lab using MrHAMER[25]. The other 74 samples were acquired from 74 PWHs enrolled in 12 studies in the HIVDB HIV-1 Single Genome Sequence (SGS) program[39]. From each sample, five HIV *protease* sequences were selected. This generated a total of 415 drug-naive HIV *protease* sequences. Based on these 415 drug-naive HIV *protease* sequences, I acquired the empirical distribution of nucleotides on each locus of the HIV *protease* sequence. A sequence generator was constructed using this empirical data, generating synthetic drug-naive HIV *protease* sequences. The generator was written in *Python*. Using this sequence generator, I generated a total of 2000 synthetic drug-naive HIV *protease* sequences and used them to initiate the simulation pipeline in this study.

Generation of synthetic drug-naive viral populations.

Using each drug-naive HIV *protease* sequence generated by the sequencing generator, I started the first round of simulation with drug pressure turned off. During the simulation, once the size of the simulated population reached ≥ 30,000, the simulator would run 30 more generations with a constant population size of 30,000. This mimicked the set point in the HIV-1 infection[118]. After completing these ten generations, the simulator collected the synthetic drug-naive population and stopped.

From a total of 2000 synthetic drug-naive HIV *protease* sequences, 2000 simulations were performed independently. Considering that the simulated population could die out and fail to

establish a drug-naive population, each simulation was allowed to be repeated 100 times until one drug-naive population was established.

Generation of synthetic virological-failure viral populations.

Using each synthetic drug-naive population, I performed the second round of simulation with drug pressure turned on. To increase the probability of the viral rebound and save computing power and time, at Generation 15, the simulated viral population was scaled up to 150,000 genomes, mimicking a short treatment break. The size of the simulated population was monitored during the simulation. And once the population size continued to increase for 3 generations and exceeded 30,000 genomes (the threshold for a simulated viral rebound in this study), the simulator collected the synthetic virological-failure population and stopped.

Considering that the simulated population could be suppressed and failed to reach a simulated viral rebound, each simulation was allowed to repeat 20 times until a simulated viral rebound was achieved.

Data preparation.

From each synthetic virological-failure viral population, 1000 sequences were randomly selected. A sequencing error of 0.02% was added to the selected sequences randomly using the ARGS module of the *mutation-simulator* (ver2.0.3)[91]. This process resulted in synthetic virological-failure samples, or called synthetic virological-failure samples. The execution of this procedure, along with bulk processing, was facilitated using custom *Python* scripts and the *Snakemake* workflow management system[119].

Linkage Disequilibrium calculation.

Individual sequences in each sample were aligned to the HXB2 reference sequence using *minimap2* with the default mode[120]. Then, the linkage disequilibrium calculation was performed using Co-Variation Mapper[34]. For this calculation, *PileUp_Fraction* was set to 0.005, and *Min_Coverage* was set to 2 in command lines.

Downstream analysis pipeline.

This analysis pipeline was designed to identify linked mutations shared among samples and enriched in those samples. It used the sequencing outputs stored in FASTA files as the input. The pipeline output was a list of correlated mutations detected (discoveries) and their p-values.

The pipeline first checked each sample to identify mutation pairs enriched to a specific frequency (50% was used in this study) (Filter 1). After checking all the input samples, the pipeline combined all enriched mutation pairs detected. Duplicate items were removed. Then, for each mutation pair, the pipeline counted the number of input samples that had this mutation pair enriched. The top few mutation pairs that were most frequently detected among input samples would be selected (Filter 2). In this way, the mutation pairs that were shared among samples and were enriched in those samples were identified (candidate mutation pairs).

For each candidate mutation pair, a 2x2 contingency table was constructed. For instance, assume that we had a pair of linked mutations, T on position 10 and M on position 20. In the 2x2 table generated, rows would present T or all other amino acids on position 10, and columns would present M or all other amino acids on position 20. The pipeline identified and quantified different mutation combinations on positions 10 and 20 in reads in all input samples, clustering them into four categories: T-M, T-not M, not T-M, not T-not M. Read numbers in these four categories

were used to populate the contingency table. With the contingency tables populated, the pipeline calculated the p-value for each pairwise association using Fisher's exact test. This calculator was accomplished using the *fisher_exact* function within the *scipy*, with the *alternative* set to "*greater*"[121]. In the end, a False Discovery Rate (FDR) approach with a significance level of 0.05 was performed using the *statsmodel* library (*sm.stats.multitest.fdrcorrection*)[115], generating discoveries (mutation pairs). This analysis pipeline was accomplished using a custom script written in *Python*.

Calculation of precision and recall.

Precision was calculated as the number of true positives divided by the number of discoveries[122]. The recall was calculated as the number of true positives divided by the size of the truth set[122]. For instance, if the simulations were performed using five pairs of built-in linked mutations, the size of the truth set was five. After analyzing, assume that a total of 20 discoveries were identified, including three out of five pairs of built-in linked mutations. Based on this, the number of true positives would be five. The precision of this analysis would be 0.15 (3/20), and the recall would be 0.6 (3/5).

Chapter 6. CONCLUSION

Co-varying mutations within the HIV genome confer drug resistance while preserving high replication fitness under drug pressure[8,11]. Identifying these co-varying mutations not only enhances our understanding of HIV protein structure and function but also uncovers potential novel protein-protein interactions, which would benefit antiviral drug design.

Prior studies utilized next-generation sequencing (NGS) techniques to sequence viral populations and developed several methods to extract mutational covariation from NGS data[14,34]. These methods either directly measure the frequency of mutations that co-occurred within the same NGS read or employ mathematical models to infer associations between mutations. However, these methods have limitations, such as constraints related to read length or their inability to handle more than two mutations simultaneously.

To address these challenges and identify co-varying *gag-pol* mutations driving resistance development, I provided a new approach combining long-read sequencing methodology MrHAMER[25] and a series of customized bioinformatics tools. MrHAMER provided accurate *gag-pol* sequences of individual HIV genomes, maintaining genetic linkages in individual viral genomes, even capturing distal linkages that were previously elusive. Pairwise correlations across the HIV *gag-pol* region were extracted from the long-read sequencing dataset using Co-Variation Mapper[34]. Building on these pairwise correlations, I inferred and quantified higher-order mutational patterns using customized *Python* scripts. The predominant mutational patterns detected after ART failures likely drive the viral resistance development and could be experimentally validated using *in-vitro* fitness assays. Additionally, I used the

Hamming-distance-based phylogenetic analysis (HDBPA) to trace the evolution of individual HIV genomes, revealing step-by-step construction of higher-order mutational patterns during ART-driven selection and the evolutionary strategies of different HIV genomes.

I demonstrated the efficacy of this approach by analyzing viral evolution in two distinct PWHs experiencing ART failures. The findings identified different mutational patterns associated with ART failure within each viral swarm, suggesting that resistance development may result from the selection of distinct subpopulations following different evolutionary pathways. Additionally, the findings underscored the impact of pretreatment genotypes on the viral evolutionary pathway for selecting new DRMs under drug pressure.

In future work, the *Python* scripts used in this approach would be integrated into one pipeline and submitted to GitHub for public use. The recombination process during viral evolution, which has not been considered in the current version of HDBPA yet, could be added to provide a more concise prediction of MRCA. Additionally, this approach could be adapted for identifying correlated mutations in other viruses or bacteria.

Experimental validation of the mutational patterns identified in this study is essential to distinguish resistance-related mutations from passenger mutations, which have no alteration on viral fitness but are inherited from founder species or acquired early along with DRMs. In contrast to resistance-related mutations, which tend to be shared among PWHs treated by the same drug class, passenger mutations arise more randomly. Hence, using samples collected from different PWHs could help distinguish resistance-related mutations from passenger mutations, reducing the number of experiments needed.

In the last part of my thesis work, I estimated the optimal sample size needed for identifying resistance-related mutations using synthetic virological-failure samples, providing a basis for future studies. The synthetic virological-failure samples were generated using a forward simulator, which simulated the acquisition of correlated DRMs and compensatory mutations in HIV protease under drug pressure. The findings revealed the relationship between the sample size, the number of mutational drug-resistant pathways, and the accuracy of identifying correlated mutations, and estimated the optimal sample size for future studies.

The forward simulator used in this study could be improved in many ways for wider application. First, more complex relationships between correlated DRMs and compensatory mutations could be explored, including “one-to-n” and “n-to-n”. Secondly, expanding this approach to other HIV proteins besides protease or inter-protein correlated mutations, e.g., between Gag and protease, could yield valuable insights. Additionally, investigating the influence of sequencing quality (sequencing error rate and sequencing depth) on correlated mutation identification would provide more accurate sample size estimates, particularly when different sequencing platforms are employed.

To summarize, unlike previous population-level HIV evolution studies, my study provided a new approach, which uncovered *gag-pol* mutational changes within individual HIV genomes and subpopulations during ART failures, identifying the higher-order mutational patterns associated with HIV resistance development and unveiling the order of mutation development. These findings not only advanced our understanding of resistance mechanisms but also unveiled novel correlations among mutations across the vast landscape of the HIV *gag-pol* region.

Bibliography

1. Weber, I. T., Wang, Y.-F. & Harrison, R. W. HIV Protease: Historical Perspective and Current Research. *Viruses* **13**, 839 (2021).
2. Simon, V. & Ho, D. D. HIV-1 dynamics in vivo: implications for therapy. *Nat. Rev. Microbiol.* **1**, 181–190 (2003).
3. Barouch, D. H. Challenges in the development of an HIV-1 vaccine. *Nature* **455**, 613–619 (2008).
4. Weber, I. & Agniswamy, J. HIV-1 Protease: Structural Perspectives on Drug Resistance. *Viruses* **1**, 1110–1136 (2009).
5. Wensing, A. M. *et al.* 2019 Update of the Drug Resistance Mutations in HIV-1. *Top. Antivir. Med.* **27**, 111–121 (2019).
6. Kuroda, M. J., el-Farrash, M. A., Choudhury, S. & Harada, S. Impaired infectivity of HIV-1 after a single point mutation in the POL gene to escape the effect of a protease inhibitor in vitro. *Virology* **210**, 212–216 (1995).
7. Doyon, L. *et al.* Second locus involved in human immunodeficiency virus type 1 resistance to protease inhibitors. *J. Virol.* **70**, 3763–3769 (1996).
8. Chang, M. W. & Torbett, B. E. Accessory mutations maintain stability in drug-resistant HIV-1 protease. *J. Mol. Biol.* **410**, 756–760 (2011).
9. Tedbury, P. R., Mercredi, P. Y., Gaines, C. R., Summers, M. F. & Freed, E. O. Elucidating the Mechanism by which Compensatory Mutations Rescue an HIV-1 Matrix Mutant Defective for Gag Membrane Targeting and Envelope Glycoprotein Incorporation. *J. Mol. Biol.* **427**, 1413–1427 (2015).
10. Kolli, M., Lastere, S. & Schiffer, C. A. Co-evolution of nelfinavir-resistant HIV-1

protease and the p1–p6 substrate. *Virology* **347**, 405–409 (2006).

11. Kolli, M., Özen, A., Kurt-Yilmaz, N. & Schiffer, C. A. HIV-1 Protease-Substrate Coevolution in Nelfinavir Resistance. *J. Virol.* **88**, 7145–7154 (2014).

12. Kolli, M., Stawiski, E., Chappey, C. & Schiffer, C. A. Human Immunodeficiency Virus Type 1 Protease-Correlated Cleavage Site Mutations Enhance Inhibitor Resistance. *J. Virol.* **83**, 11027–11042 (2009).

13. Mitsuya, Y. *et al.* N88D facilitates the co-occurrence of D30N and L90M and the development of multidrug resistance in HIV type 1 protease following nelfinavir treatment failure. *AIDS Res. Hum. Retroviruses* **22**, 1300–1305 (2006).

14. Flynn, W. F. *et al.* Deep Sequencing of Protease Inhibitor Resistant HIV Patient Isolates Reveals Patterns of Correlated Mutations in Gag and Protease. *PLOS Comput. Biol.* **11**, e1004249 (2015).

15. Hu, T., Chitnis, N., Monos, D. & Dinh, A. Next-generation sequencing technologies: An overview. *Hum. Immunol.* **82**, 801–811 (2021).

16. Garg, S. Computational methods for chromosome-scale haplotype reconstruction. *Genome Biol.* **22**, 101 (2021).

17. Butler, D. M., Pacold, M. E., Jordan, P. S., Richman, D. D. & Smith, D. M. The Efficiency of Single Genome Amplification and Sequencing is Improved by Quantitation and Use of a Bioinformatics Tool. *J. Virol. Methods* **162**, 280–283 (2009).

18. Maldarelli, F. *et al.* HIV Populations Are Large and Accumulate High Genetic Diversity in a Nonlinear Fashion. *J. Virol.* **87**, 10313–10323 (2013).

19. Kearney, M. *et al.* Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. *J. Virol.* **83**, 2715–2727 (2009).

20. Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. & Thermes, C. The Third Revolution in Sequencing Technology. *Trends Genet.* **34**, 666–681 (2018).

21. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

22. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

23. Wilson, B. D., Eisenstein, M. & Soh, H. T. High-Fidelity Nanopore Sequencing of Ultra-Short DNA Targets. *Anal. Chem.* **91**, 6783–6789 (2019).

24. Karst, S. M. *et al.* High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods* **18**, 165–169 (2021).

25. Gallardo, C. M. *et al.* MrHAMER yields highly accurate single molecule viral sequences enabling analysis of intra-host evolution. *Nucleic Acids Res.* **49**, e70–e70 (2021).

26. Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine? *Nat. Rev. Microbiol.* **9**, 617–626 (2011).

27. Domingo, E., Sheldon, J. & Perales, C. Viral Quasispecies Evolution. *Microbiol. Mol. Biol. Rev.* **76**, 159–216 (2012).

28. Vignuzzi, M. & López, C. B. Defective viral genomes are key drivers of the virus–host interaction. *Nat. Microbiol.* **4**, 1075–1087 (2019).

29. Charpentier, C., Nora, T., Tenaillon, O., Clavel, F. & Hance, A. J. Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. *J. Virol.* **80**, 2472–2482 (2006).

30. Nora, T. *et al.* Contribution of Recombination to the Evolution of Human Immunodeficiency Viruses Expressing Resistance to Antiretroviral Treatment. *J. Virol.* **81**,

7620–7628 (2007).

31. Flynn, W. F., Haldane, A., Torbett, B. E. & Levy, R. M. Inference of Epistatic Effects Leading to Entrenchment and Drug Resistance in HIV-1 Protease. *Mol. Biol. Evol.* **34**, 1291–1306 (2017).

32. Aurora, R., Donlin, M. J., Cannon, N. A. & Tavis, J. E. Genome-wide hepatitis C virus amino acid covariance networks can predict response to antiviral therapy in humans. *J. Clin. Invest.* **119**, 225–236 (2009).

33. Posada-Cespedes, S., Seifert, D. & Beerenwinkel, N. Recent advances in inferring viral diversity from high-throughput sequencing data. *Virus Res.* **239**, 17–32 (2017).

34. Routh, A., Chang, M. W., Okulicz, J. F., Johnson, J. E. & Torbett, B. E. CoVaMa: Co-Variation Mapper for disequilibrium analysis of mutant loci in viral populations using next-generation sequence data. *Methods* **91**, 40–47 (2015).

35. Wang, S. *et al.* Covariation of viral recombination with single nucleotide variants during virus evolution revealed by CoVaMa. *Nucleic Acids Res.* **50**, e41–e41 (2022).

36. Jaworski, E. & Routh, A. Parallel ClickSeq and Nanopore sequencing elucidates the rapid evolution of defective-interfering RNAs in Flock House virus. *PLOS Pathog.* **13**, e1006365 (2017).

37. Chang, M. W. *et al.* Rapid deep sequencing of patient-derived HIV with ion semiconductor technology. *J. Virol. Methods* **189**, 232–234 (2013).

38. Ghosh, S. K. *et al.* A Molecular Clone of HIV-1 Tropic and Cytopathic for Human and Chimpanzee Lymphocytes. *Virology* **194**, 858–864 (1993).

39. Vondrasek, J. & Wlodawer, A. HIVdb: A Database of the Structures of Human Immunodeficiency Virus Protease. *Proteins Struct. Funct. Bioinforma.* **49**, 429–431 (2002).

40. Knyazev, S. *et al.* Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction. *Nucleic Acids Res.* **49**, e102–e102 (2021).

41. Boucher, C. A. *et al.* High-level resistance to (-) enantiomeric 2'-deoxy-3'-thiacytidine in vitro is due to one amino acid substitution in the catalytic site of human immunodeficiency virus type 1 reverse transcriptase. *Antimicrob. Agents Chemother.* **37**, 2231–2234 (1993).

42. Robinson, J. T. *et al.* Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

43. Maguire, M. F. *et al.* Changes in Human Immunodeficiency Virus Type 1 Gag at Positions L449 and P453 Are Linked to I50V Protease Mutants In Vivo and Cause Reduction of Sensitivity to Amprenavir and Improved Viral Fitness In Vitro. *J. Virol.* **76**, 7398–7406 (2002).

44. Rhee, S.-Y., Liu, T. F., Holmes, S. P. & Shafer, R. W. HIV-1 Subtype B Protease and Reverse Transcriptase Amino Acid Covariation. *PLOS Comput. Biol.* **3**, e87 (2007).

45. Wu, T. D. *et al.* Mutation Patterns and Structural Correlates in Human Immunodeficiency Virus Type 1 Protease following Different Protease Inhibitor Treatments. *J. Virol.* **77**, 4836–4847 (2003).

46. Gallant, J. The M184V mutation: What it does, how to prevent it, and what to do with it when it's there. *AIDS Read.* **16**, 556–9 (2006).

47. Mayer, K. H., Hanna, G. J. & D'Aquila, R. T. Clinical Use of Genotypic and Phenotypic Drug Resistance Testing to Monitor Antiretroviral Chemotherapy. *Clin. Infect. Dis.* **32**, 774–782 (2001).

48. Beerenwinkel, N. *et al.* Estimating HIV Evolutionary Pathways and the Genetic Barrier to Drug Resistance. *J. Infect. Dis.* **191**, 1953–1960 (2005).

49. Svicher, V. *et al.* Involvement of Novel Human Immunodeficiency Virus Type 1 Reverse

Transcriptase Mutations in the Regulation of Resistance to Nucleoside Inhibitors. *J. Virol.* **80**, 7186–7198 (2006).

50. Mammano, F., Trouplin, V., Zennou, V. & Clavel, F. Retracing the Evolutionary Pathways of Human Immunodeficiency Virus Type 1 Resistance to Protease Inhibitors: Virus Fitness in the Absence and in the Presence of Drug. *J. Virol.* **74**, 8524–8531 (2000).

51. Cresswell-Clay, E. & Periwal, V. Genome-wide covariation in SARS-CoV-2. *Math. Biosci.* **341**, 108678 (2021).

52. Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10–19 (2019).

53. Pinheiro, H. P., de Souza Pinheiro, A. & Sen, P. K. Comparison of genomic sequences using the Hamming distance. *J. Stat. Plan. Inference* **130**, 325–339 (2005).

54. Gupta, A. *et al.* Structural studies on molecular mechanisms of Nelfinavir resistance caused by non-active site mutation V77I in HIV-1 protease. *BMC Bioinformatics* **16**, S10 (2015).

55. Bandaranayake, R. M. *et al.* The Effect of Clade-Specific Sequence Polymorphisms on HIV-1 Protease Activity and Inhibitor Resistance Pathways. *J. Virol.* **84**, 9995–10003 (2010).

56. Sundquist, W. I. & Krausslich, H.-G. HIV-1 Assembly, Budding, and Maturation. *Cold Spring Harb. Perspect. Med.* **2**, a006924–a006924 (2012).

57. Verheyen, J. *et al.* Compensatory mutations at the HIV cleavage sites p7/p1 and p1/p6 in therapy-naïve and therapy-experienced patients. *Antivir. Ther.* **11**, 879–87 (2006).

58. Bally, F., Martinez, R., Peters, S., Sudre, P. & Telenti, A. Polymorphism of HIV Type 1 Gag p7/p1 and p1/p6 Cleavage Sites: Clinical Significance and Implications for Resistance to Protease Inhibitors. *AIDS Res. Hum. Retroviruses* **16**, 1209–1213 (2000).

59. Betancor, G. *et al.* Clinical, virological and biochemical evidence supporting the association of HIV-1 reverse transcriptase polymorphism R284K and thymidine analogue resistance mutations M41L, L210W and T215Y in patients failing tenofovir/emtricitabine therapy. *Retrovirology* **9**, 68 (2012).

60. Lacey, S. F. & Larder, B. A. Novel mutation (V75T) in human immunodeficiency virus type 1 reverse transcriptase confers resistance to 2',3'-didehydro-2',3'-dideoxythymidine in cell culture. *Antimicrob. Agents Chemother.* **38**, 1428–1432 (1994).

61. Bell, N. M. & Lever, A. M. L. HIV Gag polyprotein: processing and early viral particle assembly. *Trends Microbiol.* **21**, 136–144 (2013).

62. Venter, P. A. & Schneemann, A. Recent insights into the biology and biomedical applications of Flock House virus. *Cell. Mol. Life Sci.* **65**, 2675–2687 (2008).

63. Li, Y. & Ball, L. A. Nonhomologous RNA recombination during negative-strand synthesis of flock house virus RNA. *J. Virol.* **67**, 3854–3860 (1993).

64. Zhou, Y. & Routh, A. Mapping RNA–capsid interactions and RNA secondary structure within virus particles using next-generation sequencing. *Nucleic Acids Res.* **48**, e12–e12 (2020).

65. Dasgupta, R., Cheng, L.-L., Bartholomay, L. C. & Christensen, B. M. Flock house virus replicates and expresses green fluorescent protein in mosquitoes. *J. Gen. Virol.* **84**, 1789–1797 (2003).

66. Zhong, W., Dasgupta, R. & Rueckert, R. Evidence that the packaging signal for nodaviral RNA2 is a bulged stem-loop. *Proc. Natl. Acad. Sci.* **89**, 11146–11150 (1992).

67. Jovel, J. & Schneemann, A. Molecular characterization of Drosophila cells persistently infected with Flock House virus. *Virology* **419**, 43–53 (2011).

68. Fisher, A. J. & Johnson, J. E. Ordered duplex RNA controls capsid architecture in an icosahedral animal virus. *Nature* **361**, 176–179 (1993).

69. Rezelj, V. V., Levi, L. I. & Vignuzzi, M. The defective component of viral populations. *Curr. Opin. Virol.* **33**, 74–80 (2018).

70. Ball, L. A. & Li, Y. cis-acting requirements for the replication of flock house virus RNA 2. *J. Virol.* **67**, 3544–3551 (1993).

71. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. The Vienna RNA Websuite. *Nucleic Acids Res.* **36**, W70–W74 (2008).

72. Kaesberg, P. *et al.* Structural homology among four nodaviruses as deduced by sequencing and X-ray crystallography. *J. Mol. Biol.* **214**, 423–435 (1990).

73. Albariño, C. G., Eckerle, L. D. & Ball, L. A. The cis-acting replication signal at the 3′ end of Flock House virus RNA2 is RNA3-dependent. *Virology* **311**, 181–191 (2003).

74. Rosskopf, J. J. *et al.* A 3′ terminal stem–loop structure in Nodamura virus RNA2 forms an essential cis-acting signal for RNA replication. *Virus Res.* **150**, 12–21 (2010).

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

76. Routh, A. & Johnson, J. E. Discovery of functional genomic motifs in viruses with ViReMa–a Virus Recombination Mapper–for analysis of next-generation sequencing data. *Nucleic Acids Res.* **42**, e11–e11 (2014).

77. Tamiya, S., Mardy, S., Kavlick, M. F., Yoshimura, K. & Mistuya, H. Amino Acid Insertions near Gag Cleavage Sites Restore the Otherwise Compromised Replication of Human Immunodeficiency Virus Type 1 Variants Resistant to Protease Inhibitors. *J. Virol.* **78**, 12030–12040 (2004).

78. Marlowe, N. *et al.* Analysis of Insertions and Deletions in the gag p6 Region of Diverse HIV Type 1 Strains. *AIDS Res. Hum. Retroviruses* **20**, 1119–1125 (2004).

79. Sanchez-Pescador, R. *et al.* Nucleotide sequence and expression of an AIDS-associated retrovirus (ARV-2). *Science* **227**, 484–492 (1985).

80. Johnson, B. A. *et al.* Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis. *Nature* **591**, 293–299 (2021).

81. Gribble, J. *et al.* The coronavirus proofreading exoribonuclease mediates extensive viral recombination. *PLOS Pathog.* **17**, e1009226 (2021).

82. Langsjoen, R. M. *et al.* Differential Alphavirus Defective RNA Diversity between Intracellular and Extracellular Compartments Is Driven by Subgenomic Recombination Events. *mBio* **11**, e00731-20 (2020).

83. DePolo, N. J., Giachetti, C. & Holland, J. J. Continuing coevolution of virus and defective interfering particles and of viral genome sequences during undiluted passages: virus mutants exhibiting nearly complete resistance to formerly dominant defective interfering particles. *J. Virol.* **61**, 454–464 (1987).

84. Dam, E. *et al.* Gag Mutations Strongly Contribute to HIV-1 Resistance to Protease Inhibitors in Highly Drug-Experienced Patients besides Compensating for Fitness Loss. *PLoS Pathog.* **5**, e1000345 (2009).

85. Martins, A. N., Arruda, M. B., Pires, A. F., Tanuri, A. & Brindeiro, R. M. Accumulation of P(T/S)AP Late Domain Duplications in HIV Type 1 Subtypes B, C, and F Derived from Individuals Failing ARV Therapy and ARV Drug-Naive Patients. *AIDS Res. Hum. Retroviruses* **27**, 687–692 (2010).

86. Peters, S. *et al.* Resistance to Nucleoside Analog Reverse Transcriptase Inhibitors

Mediated by Human Immunodeficiency Virus Type 1 p6 Protein. *J. Virol.* **75**, 9644–9653 (2001).

87. Martins, A. N. *et al.* Elucidation of the Molecular Mechanism Driving Duplication of the HIV-1 PTAP Late Domain. *J. Virol.* **90**, 768–779 (2016).

88. Dettenhofer, M. & Yu, X.-F. Proline Residues in Human Immunodeficiency Virus Type 1 p6Gag Exert a Cell Type-Dependent Effect on Viral Replication and Virion Incorporation of Pol Proteins. *J. Virol.* **73**, 4696–4704 (1999).

89. Loman, N. J. & Quinlan, A. R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30**, 3399–3401 (2014).

90. Peng, B. *et al.* Genetic data simulators and their applications: an overview. *Genet. Epidemiol.* **39**, 2–10 (2015).

91. Kühl, M. A., Stich, B. & Ries, D. C. Mutation-Simulator: fine-grained simulation of random mutations in any genome. *Bioinformatics* **37**, 568–569 (2021).

92. Jariani, A. *et al.* SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination. *Virus Evol.* **5**, (2019).

93. Weikl, T. R. & Hemmateenejad, B. Accessory mutations balance the marginal stability of the HIV-1 protease in drug resistance. *Proteins Struct. Funct. Bioinforma.* **88**, 476–484 (2020).

94. Henes, M. *et al.* Picomolar to Micromolar: Elucidating the Role of Distal Mutations in HIV-1 Protease in Conferring Drug Resistance. *ACS Chem. Biol.* **14**, 2441–2452 (2019).

95. Sugiura, W. *et al.* Interference between D30N and L90M in Selection and Development of Protease Inhibitor-Resistant Human Immunodeficiency Virus Type 1. *Antimicrob. Agents Chemother.* **46**, 708–715 (2002).

96. Resch, W., Ziermann, R., Parkin, N., Gamarnik, A. & Swanstrom, R. Nelfinavir-Resistant, Amprenavir-Hypersusceptible Strains of Human Immunodeficiency Virus Type 1 Carrying an N88S Mutation in Protease Have Reduced Infectivity, Reduced Replication Capacity, and Reduced Fitness and Process the Gag Polyprotein Precursor Aberrantly. *J. Virol.* **76**, 8659–8666 (2002).

97. Salazar-Gonzalez, J. F. *et al.* Genetic identity, biological phenotype, and evolutionary pathways of transmitted/founder viruses in acute and early HIV-1 infection. *J. Exp. Med.* **206**, 1273–1289 (2009).

98. Mens, H. *et al.* HIV-1 Continues To Replicate and Evolve in Patients with Natural Control of HIV Infection. *J. Virol.* **84**, 12971–12981 (2010).

99. Herbeck, J. T. *et al.* Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. *J. Virol.* **85**, 7523–7534 (2011).

100. Manak, M. *et al.* Pilot Studies for Development of an HIV Subtype Panel for Surveillance of Global Diversity. *AIDS Res. Hum. Retroviruses* **28**, 594–606 (2012).

101. Koning, F. A. *et al.* Dynamics of HIV Type 1 Recombination Following Superinfection. *AIDS Res. Hum. Retroviruses* **29**, 963–970 (2013).

102. Iyer, S. S. *et al.* Resistance to type 1 interferons is a major determinant of HIV-1 transmission fitness. *Proc. Natl. Acad. Sci.* **114**, E590–E599 (2017).

103. deCamp, A. C. *et al.* Sieve analysis of breakthrough HIV-1 sequences in HVTN 505 identifies vaccine pressure targeting the CD4 binding site of Env-gp120. *PloS One* **12**, e0185959 (2017).

104. Song, H. *et al.* Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nat. Commun.* **9**, 1928 (2018).

105. Rolland, M. *et al.* Molecular dating and viral load growth rates suggested that the eclipse phase lasted about a week in HIV-1 infected adults in East Africa and Thailand. *PLoS Pathog.* **16**, e1008179 (2020).

106. Imamichi, H. *et al.* Human immunodeficiency virus type 1 quasi species that rebound after discontinuation of highly active antiretroviral therapy are similar to the viral quasi species present before initiation of therapy. *J. Infect. Dis.* **183**, 36–50 (2001).

107. Winckelmann, A. *et al.* Romidepsin-induced HIV-1 viremia during effective antiretroviral therapy contains identical viral sequences with few deleterious mutations. *AIDS Lond. Engl.* **31**, 771–779 (2017).

108. Winckelmann, A. *et al.* Genetic characterization of the HIV-1 reservoir after Vacc-4x and romidepsin therapy in HIV-1-infected individuals. *AIDS Lond. Engl.* **32**, 1793–1802 (2018).

109. Colby, D. J. *et al.* Rapid HIV RNA rebound after antiretroviral treatment interruption in persons durably suppressed in Fiebig I acute HIV infection. *Nat. Med.* **24**, 923–926 (2018).

110. St Bernard, L., Abolade, J., Mohri, H., Markowitz, M. & Evering, T. H. Drug Resistance Mutation Frequency of Single-Genome Amplification-Derived HIV-1 Polymerase Genomes in the Cerebrospinal Fluid and Plasma of HIV-1-Infected Individuals under Nonsuppressive Therapy. *J. Virol.* **94**, e01824-19 (2020).

111. Palmer, A. *et al.* Viral suppression and viral rebound among young adults living with HIV in Canada. *Medicine (Baltimore)* **97**, e10562 (2018).

112. Min, S., Gillani, F. S., Aung, S., Garland, J. M. & Beckwith, C. G. Evaluating HIV Viral Rebound Among Persons on Suppressive Antiretroviral Treatment in the Era of "Undetectable Equals Untransmittable (U = U)". *Open Forum Infect. Dis.* **7**, ofaa529 (2020).

113. DIEPSTRA, K. *et al.* What we talk about when we talk about durable viral suppression.

AIDS Lond. Engl. **34**, 1683–1686 (2020).

114. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

115. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. *Proc. 9th Python Sci. Conf.* **2010**, (2010).

116. Clemente, J. C., Hemrajani, R., Blum, L. E., Goodenow, M. M. & Dunn, B. M. Secondary Mutations M36I and A71V in the Human Immunodeficiency Virus Type 1 Protease Can Provide an Advantage for the Emergence of the Primary Mutation D30N. *Biochemistry* **42**, 15029–15035 (2003).

117. Kožíšek, M. *et al.* Molecular Analysis of the HIV-1 Resistance Development: Enzymatic Activities, Crystal Structures, and Thermodynamics of Nelfinavir-resistant HIV Protease Mutants. *J. Mol. Biol.* **374**, 1005–1016 (2007).

118. Goulder, P. J., Lewin, S. R. & Leitman, E. M. Paediatric HIV infection: the potential for cure. *Nat. Rev. Immunol.* **16**, 259–271 (2016).

119. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

120. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

121. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

122. Raghavan, V., Bollmann, P. & Jung, G. S. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.* **7**, 205–229 (1989).

VITA

Shiyi Wang completed her Ph.D. in Molecular Medicine and Mechanisms of Disease from the University of Washington. Her research interests primarily lie in the correlated mutations that arose in the HIV genome during antiretroviral therapy (ART), which provided HIV with enhanced resistance and fitness, leading to ART failures. During her doctoral study, Shiyi utilized long-read sequencing methodology to acquire individual HIV genomes from longitudinal serum samples from people living with HIV (PWHs) and developed bioinformatic analyzing tools to reveal the evolutionary trajectories of individual HIVs during ART in PWHs and identify mutational patterns associated with ART failures. Furthermore, she explored the relationship between sample size and correlated mutation identification accuracy using an individual virus-based forward simulator. In Oct 2023, she successfully defended her thesis titled "Investigating the evolution of individual HIVs during ART failures using long-read sequencing" under the supervision of Bruce E. Torbett, Ph.D., MSPH.

Prior to pursuing her Ph.D., Shiyi Wang earned her Bachelor's degree from Zhejiang University, China, with third-class scholarships for outstanding merits for 2015 and 2016. During her undergraduate studies, Shiyi Wang worked as a visiting scholar in Dr. Haoxin Xu's Lab at the University of Michigan, investigating synthetic small molecules' effects in modulating lysosomal Two-Pore Channel Subtype 2 (TPC2).

As a Ph.D. graduate, Shiyi Wang looks forward to continuing her research journey and making meaningful contributions to the investigation of viral evolution using sequencing and bioinformatics methods.