8-19-2020 1:00 PM

Rates and patterns of indels in HIV-1 gp120 within and among hosts

John Lawrence Palmer, *The University of Western Ontario*

Supervisor: Poon, Art F.Y., *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Pathology and Laboratory Medicine

Recommended Citation

Palmer, John Lawrence, "Rates and patterns of indels in HIV-1 gp120 within and among hosts" (2020). *Electronic Thesis and Dissertation Repository*. 7308.
https://ir.lib.uwo.ca/etd/7308

Abstract

Insertions and deletions (indels) in the HIV-1 gp120 variable loops modulate sensitivity to neutralizing antibodies and are therefore implicated in HIV-1 immune escape. However, the rates and characteristics of variable loop indels have not been investigated within hosts. Here, I report a within-host phylogenetic analysis of gp120 variable loop indels, with mentions to my preceding study on these indels among hosts.

We processed longitudinally-sampled gp120 sequences collected from a public database (*n* = 11,265) and the Novitsky Lab (*n*=2,541). I generated time-scaled within-host phylogenies using BEAST, extracted indels by reconstructing ancestral sequences in Historian, and estimated variable loop indel rates by applying a Poisson-based model to indel counts and time data.

Variable loop indel rates appeared higher within hosts than among hosts in subtype C. Our findings improve understanding of indel evolution in HIV-1 gp120 and enable the evaluation of models describing indels, which I present as work in progress.

Keywords: HIV-1, gp120, env, variable loops, indels, indel rates, evolution, phylogenetics, immune escape

Lay Summary

The Human Immunodeficiency Virus (HIV) attaches to our immune cells using a protein on its surface called gp120. The nucleotide sequence that produces the gp120 protein undergoes numerous changes (mutations), with one type being the insertion or deletion of nucleotides (indels). Indels are most frequently found in five specific regions of the gp120 nucleotide sequence, or gene, that produce five sugar-covered loop structures (V1-V5) on the surface of this protein, referred to as "variable loops". Since gp120 is exposed on the surface of HIV, the human immune system commonly designs cells and proteins that target key patterns on the gp120 variable loops in order to detect and eliminate HIV. Importantly, indel mutations can change the shapes, lengths, and sugar positions of the gp120 variable loops, thereby altering the same targeted patterns on these loops until they become unrecognizable to the immune system. This process, known as "immune escape", occurs repeatedly during HIV infection, allowing the virus to remain unaffected by the immune response and continue to cause disease.

Rates of mutations provide an indication of how quickly an organism, or virus, can change itself to adapt to its environment, like in the process of immune escape in HIV, for example. In the gp120 variable loops, indels are involved in immune escape; however, indel rates have not been studied in these regions. This work provides the first estimates of indel rates in the gp120 variable loops of HIV by using computer software to estimate the amount of historical time between different virus samples. I first estimated variable loop indel rates at the population scale (among hosts) by analyzing one HIV nucleotide sequence per person worldwide. My main study then estimated indel rates by analyzing multiple sequences collected from a single individual (within hosts). I found that indel rates appear higher within hosts than among hosts. This work contributes to a better understanding of an important type of mutation within the gp120 variable loops that helps HIV adapt to our immune system. It does this by quantifying how frequently indels occur and their contributions to both variable loop changes and immune escape.

Co-Authorship Statement

John Palmer performed all the work included in this dissertation. Dr. Art Poon contributed to study design and data interpretation, and further provided funding support for this work. Chapter 2 of this dissertation contains contents reproduced from the published journal article 'Phylogenetic measures of indel rate variation among the HIV-1 group M subtypes' (Palmer J, Poon AFY. Virus Evol. 2019 Jul 21;5(2):vez022. doi: 10.1093/ve/vez022). I chose to include this publication in my dissertation because 1) my graduate research was based on the work performed in this study, and 2) it contains a substantial amount of work that was produced during my graduate training. Specifically, I produced Table 2.1, Figures 2.4 and 2.5, and Supplementary Figures A1 and A2 during the first 6 months of my masters program, while I produced Figures 2.1, 2.2, and 2.3 and Supplementary Table A1 during my Bachelor of Medical Sciences program. In Chapter 2, Dr. Poon also wrote small components of coding scripts and edited the final manuscript sent for publication.

Acknowlegements

I would first like to thank Dr. Art Poon in the Department of Pathology and Laboratory Medicine, as he has been an incredibly supportive mentor to me throughout my 4th year undergraduate thesis and my masters program. I am specifically grateful for his great mentoring, teaching, support in applications and presentations, and financial support. I attribute numerous skills to his teaching, some of which include computer scripting, statistical analysis, phylogenetic tree reconstruction, Bayesian inference, and statistical modelling. Above all, I am grateful for his continued patience and kindness.

I would also like to thank Dr. David Champredon in the Department of Pathology and Laboratory Medicine for teaching me about probability distributions, statistical modeling, and Bayesian inference. Dr. Champredon's generous teachings greatly improved my ability to use the RStan Bayesian inference framework, enabling me to produce the primary findings indel rates found in Chapter 3.

I would also like to thank Mr. Connor Chato in the Department of Pathology and Laboratory Medicine for his assistance in conceptualizing and implementing ideas pertaining to phylogenetic tree traversal and empirical modeling of insertions.

I would next like to thank Dr. Abayomi Olabode in the Department of Pathology and Laboratory Medicine for his continued support on many aspects of phylogenetic analysis and computer scripting.

I would like to thank Ms. Laura Muñoz Baena in the Department of Microbiology and Immunology for her very helpful feedback regarding BEAST analysis which helped me diagnose and address problems in my analysis.

I would finally like to thank my parents for their incredible support over the course of this program. You have always been my foundation and I dedicate this work to you.

Contents

Abstract **ii**

Lay Summary **iii**

Co-Authorship Statement **iv**

Acknowledgements **v**

List of Figures **x**

List of Tables **xi**

List of Appendices **xii**

1 Background and Context **1**

1.1 Human Immunodeficiency Virus 1

1.2 HIV-1 Genome and Structure 2

1.3 HIV Replicative Cycle 4

1.3.1 Binding and Entry 4

1.3.2 Reverse Transcription 4

1.3.3 Integration 5

1.3.4 Assembly of Virions 5

1.4 Susceptible Cells & Tropism 6

1.5 Mutation & Evolution 7

1.5.1 Nucleotide Substitutions and Evolutionary Models 8

1.5.2 Selection 9

1.5.3 Phylogenetic Inference 10

1.5.4 Indels 11

1.5.5 Recombination 13

1.5.6 Diversity & Evolution Within Hosts 13

1.5.7 Diversity & Evolution Among Hosts 15

1.6 HIV-1 Pathogenesis 16

1.6.1 Transmission 16

1.6.2 Immune Response to HIV-1 16

1.6.3 Immune Escape 17

1.6.4 Phases of Infection 18

1.7 gp120 19

1.7.1 Glycosylation 19
1.7.2 Variable Loops 20
1.7.3 Immune Escape 21
1.7.4 Variable Loop Indel Rates 21
1.8 Objectives 23
1.9 Hypothesis 23

Chapter 1 Bibliography **24**

2 Indel Evolution Among Hosts **41**
2.1 Background 41
2.1.1 Context 41
2.1.2 Pairwise Alignment 42
2.1.3 Phylogenetic Inference 43
2.1.4 Indel Rate Estimation 44
2.2 Introduction 46
2.3 Methods 48
2.3.1 Data processing 48
2.3.2 Phylogenetic analysis 49
2.3.3 Indel rate estimation 50
2.3.4 Analysis of indels 51
2.4 Results 52
2.5 Discussion 58
2.6 Acknowledgments 63
2.7 Data Availability 63

Chapter 2 Bibliography **64**

3 Indel Evolution Within Hosts **72**
3.1 Introduction 72
3.1.1 Sampling Differences 73
3.1.2 Evolutionary Processes 74
3.1.3 HIV-1 gp120 74
Immune Escape 75
Rates 75
3.1.4 Indels 76
3.2 Hypothesis 76
3.3 Methods 77
3.3.1 Sequence Retrieval and Processing 77
3.3.2 Sequence Slicing and Alignment 79
3.3.3 Patient Screening for Phylogenetic Analysis 79
Molecular Clock Signal 80
Hypermutation 80
Superinfection 81
3.3.4 Phylogenetic Tree Reconstruction 81

Parameters 82
Posterior Convergence 82
Coalescent Model & Model Selection 83
Tree Extraction 84
3.3.5 Ancestral State Reconstruction & Indel Extraction 85
3.3.6 Indel Rate Estimation 87
3.3.7 Indel Timings 88
3.3.8 Nucleotide Composition 89
3.3.9 Indel Lengths 89
3.3.10 Changes to Potential N-linked Glycosylation Sites 89
3.4 Results 91
3.4.1 Indel Rates 91
3.4.2 Indel Timings 91
3.4.3 Indel Lengths 93
3.4.4 Indel Nucleotide Compositions 94
3.4.5 Indel-Induced PNGS Changes 94
3.5 Discussion 97
3.5.1 Indel Rates 97
3.5.2 Indel Timings 100
3.5.3 Indel Lengths 101
3.5.4 Indel Nucleotide Compositions 102
3.5.5 Indel-Induced PNGS Changes 103
3.5.6 Limitations 103
3.5.7 Conclusions 104

Chapter 3 Bibliography 105

4 Insertion Modelling 115
4.1 Background 115
4.2 Methods 119
4.2.1 Insertion Data 119
4.2.2 Bayesian Framework 119
4.2.3 Fundamental Assumptions 120
4.3 Results 121
4.3.1 Single Parameter Models 121
4.3.2 Two State Model 123
Nucleotide Substitutions 125
Length Bias 127
Random Walk of Slip Events 127
4.4 Discussion 129

Chapter 4 Bibliography 130

5 Conclusions 132
5.1 Overview 132

5.2 Summary of Findings 133
5.3 Concluding Hypotheses 134
5.4 Significance & Contributions 135
5.5 Future Directions 136

Chapter 5 Bibliography **137**

A Chapter 2 Supplementary Material **139**
A.1 Supplementary Figures 139
A.2 Supplementary Tables 142

B Chapter 3 Supplementary Material **143**

Curriculum Vitae **147**

List of Figures

1.1 Phylogenetic Tree 12

2.1 Root-to-tip Regressions for Seven Group M Clades 53
2.2 Among-Host Indel Rates 55
2.3 Among-Host Indel Lengths 56
2.4 Among-Host Indel Nucleotide Compositions 57
2.5 Among-Host Indels Affecting Potential N-linked Glycosylation Sites 59

3.1 Within-Host Methods Flowchart 86
3.2 Within-Host Indel Rates 92
3.3 Within-Host Indel Timings 93
3.4 Within-Host Indel Lengths 95
3.5 Within-Host Indel Nucleotide Compositions 96
3.6 Within-Host Indel-Induced Changes to Potential N-linked Glycosylation Sites . 98

4.1 Distances to Next Matching Sequence 118
4.2 Data Format for Slip Model 120
4.3 Single Parameter Slip Models 121
4.4 Insertion Lengths 123
4.5 Two-State Model 124
4.6 Three-Parameter Model 126
4.7 Four-Parameter Model With Indel Alignment Algorithm 128

A1 Potential N-Linked Glycosylation Site Counts by gp120 Variable Loop 140
A2 Cherry Genetic Distances 141

B1 Within-Host Indel Nucleotide Compositions by Variable Loop 144
B2 Within-Host Indel Dinucleotide Compositions 145
B3 RStan Hierarchical Model Test 146

List of Tables

2.1 Molecular Clock Summaries for Seven Group M Clades 54

A1 Binomial Generalized Linear Model Output 142

List of Appendices

Appendix A Chapter 2 Supplementary Material 139
Appendix B Chapter 3 Supplementary Material 143

Chapter 1

Background and Context

1.1 Human Immunodeficiency Virus

The human immunodeficiency virus (HIV) is a highly diverse and rapidly-evolving virus, which is the causative agent of Acquired Immune Deficiency Syndrome (AIDS). HIV originates from the closely-related Simian Immunodeficiency Virus (SIV) found in various non-human primate species [1]. Specifically, reports suggest that there were multiple independent transmission events — the transfer of an infectious entity from one host to another — of SIV in these primate species to human hosts in West and Central Africa [1]. These transmission events generated multiple genetically-distinct forms of HIV that have been classified into two types (1 and 2) and further stratified into multiple groups per type [1]. HIV type 1 (HIV-1), which is the focus of this work, is more pathogenic — or more efficiently causes disease — and is suggested to originate specifically from the strain of SIV found in the chimpanzee species *Pan troglodytes troglodytes* [2, 3].

The HIV-1 genome, a sequence of roughly 9,200 nucleotides, contains 9 distinct genes that encode all the proteins this virus needs to function [4]. This genome is highly compact with minimal non-coding sequence amidst these functional genes, meaning that changes to its genetic sequence, or mutations, can often change the virus' biological characteristics, or phenotype [4, 5]. As a retrovirus, HIV is classified by its conversion of ribonucleic acid molecules (RNA) into deoxyribonucleic acids (DNA), a process that contradicts the standard "forward" flow of genetic information in the central dogma of biology: DNA is used to create RNA, which generates proteins [6]. Further explanation of this process and its implications will come later.

Of the four distinct groups within HIV-1 (M, N, O, and P), group M (main) is responsible for the current global pandemic, accounting for roughly 95% of HIV cases worldwide [7]. These HIV types and groups differ considerably in their genetic sequences and fittingly exhibit corresponding differences in their viral characteristics [8]. For example, HIV-1 group M has an approximate 100-fold higher fitness — an organism's capacity to successfully replicate in a specific environment — in human cells *in vitro* than either HIV-1 group O or HIV type 2, which has been suggested to contribute to its widespread prevalence [8]. As of 2018, the HIV-1 pandemic was infecting approximately 38 million people and has taken the lives of an estimated 32 million others since it began [9]. Despite improvements in therapeutic treatment and outbreak management, an estimated 1.7 million new HIV-1 infections and 770,000 HIV-related deaths were reported in 2018 alone, indicating potential for significant improvement regarding the prevention and treatment of this virus [9, 10].

1.2 HIV-1 Genome and Structure

HIV is a single-stranded positive sense RNA retrovirus, whose viral particle is comprised of a protein core surrounded by a lipid membrane derived from the human immune cells it infects [11]. Within the HIV-1 genome, there are three genes that encode for structural and enzymatic protein products: *gag*, *pol*, and *env* [11]. All three of these genes encode multiple protein products that are initially translated as polyprotein precursors and must be cleaved by a protease to be activated [11].

The *gag* gene encodes four proteins that are primarily responsible for forming the structures of the HIV particle. These structures include the matrix supporting the outer lipid membrane (matrix; p17), the protein shell of the internal viral core (capsid; p24), and the protein complexes encasing the copies of viral RNA (nucleocapsid; p9 and p6) [11, 12].

The *pol* gene codes for the reverse transcriptase (RT), integrase, and protease enzymes that are responsible catalyzing the reverse transcription of viral RNA to viral DNA, the integration of viral DNA into the host genome, and the cleavage of viral proteins, respectively [11, 12]. Thirdly, the *env* gene encodes two proteins that comprise the envelope spike protein complexes embedded in the lipid membrane of HIV-1.

These proteins, gp41 and gp120, are more specifically referred to as glycoproteins, as they are modified with the attachment of carbohydrate complexes composed of multiple sugar molecules called glycans [13]. Glycans are critical to the function, folding, and stability of the two *env* glycoproteins and will be discussed in more depth later [13, 14]. The gp41 anchor glycoprotein uses transmembrane protein structures to embed itself in the viral lipid membrane where it remains relatively hidden from the extracellular environment [15]. The gp120 glycoprotein, which is the focus of this work, attaches on top of gp41 using noncovalent interactions where it is fully exposed to the extracellular environment [16, 17]. Importantly, the sequence of gp120 is segmented into five conserved regions (C1-C5) that comprise the protein's interior, and five variable regions (V1-V5) that form disordered loop structures on its exterior [18]. Each protein spike consists of three copies of gp120 and three copies of gp41 bound together in a trimeric complex which, as a whole, allows HIV-1 to attach to and enter susceptible target cells [16, 17].

The remaining genes in the HIV genome code for non-structural regulatory and accessory proteins that perform various functions that support viral infection. These functions include, but are not limited to, enhancing the efficiency of viral transcription (*tat*), facilitating nuclear import (*vpr*) and export (*rev*) of viral components, suppressing intracellular (*vif*) and extracellular (*nef*) antiviral mechanisms, arresting the cell cycle (*vpr*), and degrading interfering proteins (*vpu*) [11, 19–22].

1.3 HIV Replicative Cycle

1.3.1 Binding and Entry

As a retrovirus, HIV undergoes a unique and complex replicative cycle dependent on the cellular machinery located within susceptible cells. The first stages of this process are to bind and enter a target cell, which are mediated by the envelope spike complexes on the HIV-1 lipid membrane. Target cell attachment of HIV-1 requires the binding of two receptors: the primary CD4 surface receptor and a coreceptor. Specifically, gp120 first attaches to the CD4 receptor via its binding site, which induces a conformational change in gp120 and causes exposure of the coreceptor binding site [17]. This enables subsequent binding of an essential coreceptor from the seven-transmembrane chemokine receptor family [17]. At this point, the protein interactions between gp120 and both surface receptors induce a series of conformational changes in the spike complex that enable viral entry [23]. More specifically, gp41 first adopts an intermediate conformation that allows it to insert its hydrophobic N-terminal fusion peptide into the target cell membrane [23]. From this stage, gp41 will bring the virus and cell membranes into close proximity and expand the initialized fusion pore by folding several of its helical structures together [23]. The contents within the viral protein core are then released into the cytoplasm of the susceptible immune cell as the capsid proteins dissociate. This leaves behind the reverse transcription complex, consisting of two copies of the HIV RNA genome held within nucleocapsid complexes, the RT enzyme and integrase enzymes encoded by the *pol* gene, and the Vpr protein [11].

1.3.2 Reverse Transcription

The next stage in the replicative cycle is reverse transcription of the viral RNA genome into viral DNA. Briefly, this process involves generating a copy of negative-sense viral DNA from RNA, degrading the original RNA sequence, and synthesizing positive-sense DNA from the negative-sense strand [11]. When complete, this produces the pre-integration complex [11]. It is important to mention here that the RT enzyme does not have 3' to 5' exonuclease activity to proofread incorporated base pairs like other eukaryotic polymerases. This means that the reverse transcription process is highly error prone, reportedly introducing between 1.4 and 3.4 $\times 10^{-5}$ misincorporated nucleotides per site each time it replicates [24, 25]. In fact, other sources estimate this rate reaching as high as 4 $\times 10^{-3}$ mutations/nucleotide/replication *in vivo* [26]. For reference, these retrovirus mutation rates are between a thousand- and a million-fold higher than the mutation rates of eukaryotic organisms, which exhibit between 10^{-8} to 10^{-11} mutations/site/replication [27].

1.3.3 Integration

At this point, the pre-integration complex, consisting of reverse transcribed viral DNA, integrase enzymes, matrix protein, viral protein R (Vpr), and nucleocapsid proteins, is transported across the nuclear membrane [28]. Viral protein R and the phosphorylated matrix proteins both play critical roles in this process [28]. Once in the nucleus, integrase cleaves host DNA and subsequently inserts the viral DNA into the gap, now being referred to as proviral DNA [28].

1.3.4 Assembly of Virions

The proviral DNA is then transcribed to viral mRNA using the host RNA polymerase II enzyme, similar to other human genes in the genome [11]. which is amplified greatly after translation of the Tat protein [11]. Next, viral RNA is shuttled out of the nucleus with the help of the Rev protein, followed by viral protein translation and cleavage in the cytoplasm [11]. Finally, all viral proteins, including two packaged copies of the viral RNA transcripts, are transported to the cell membrane where aggregate together and bud off into a new HIV viral particle [11].

1.4 Susceptible Cells & Tropism

HIV infects cell populations that express the CD4 surface receptor protein, which include dendritic cells (DCs), macrophages, and importantly, helper T cells [29]. There are also other less common cell types that are susceptible to HIV based on their expression of CD4, such as microglia in the brain [30]. While HIV has generally been found in all of these cell types, a given strain of HIV cannot bind to all of these cell types with the same efficiency and typically exhibits preference for one population. The different tendencies of HIV-1 to infect specific cell types are modulated by the virus's two possible tropisms, or different abilities to infect specific target cells. The viral tropisms of HIV-1 are associated with the virus' use of two different coreceptors belonging to the seven-transmembrane chemokine protein family: the CC-chemokine receptor 5 (CCR5), and the CXC-chemokine receptor 4 (CXCR4) [31, 32]. The R5 (or M-tropic) HIV strain primarily uses the CCR5 coreceptor for entry, allowing it to infect macrophages and memory T cell populations [31, 32]. The second tropism, namely the X4 (or T-tropic) HIV strain, predominantly utilizes the CXCR4 coreceptor for viral entry and can infect both naive and memory T cell populations, being predominantly found in the latter [31, 32]. Moreover, there is the possibility for HIV to acquire dual-tropism, termed R5X4, in which case HIV can use both coreceptors. It is important to know that, while these trends describe the majority of tropism behaviour, they certainly should not be considered concrete biological rules as these trends contain at least a few exceptions[31, 33].

The R5-tropic strain of HIV is the tropism associated with establishment of new infection transmission, as supported by its presence in the essentially every transmitted-founder (T/F) variant — the HIV strain that establishes infection in a new host [31, 32]. Research suggests this strong filter for R5 tropism is caused by predominant coreceptor expression patterns in the mucosal layers where HIV transmission often takes place [34–36]. While R5-tropic HIV-1 is predominant throughout early infection, X4-tropic variants arise in approximately 50% of patients after roughly 5 years [34, 37]. Importantly, the X4 tropism of HIV is associated with faster depletion of T cell populations and progression to AIDS, due to the wider range of T cells that susceptible to these viruses [32].

1.5 Mutation & Evolution

Evolution refers to changes in the frequencies of genetic variants (or alleles) within biological populations over successive generations [38]. Mutations (genetic changes) are responsible for creating genetic diversity within populations that is required in order for evolution to occur. There are multiple different types of mutations which include, but are not limited to, nucleotide substitutions, insertions and deletions (indels), or even recombination of entire genomes [38]. Recall that a genome encodes the functional proteins that constitute an organism and thus, mutations can, but not always, have direct effects on an organism's phenotype [38]. The interactions between an organism's phenotype and its surrounding environment affects its ability to survive and reproduce in a process known as selection, which can further shape evolution and will be discussed more later.

Viruses occupy a unique position on the tree of life, which has often been debated due to their lack of autonomous metabolic function and dependence on hijacking cellular machinery to replicate [39]. However, viruses still undergo many of the same core biological processes found in all biological organisms, including a reproductive cycle involving the inheritance of genetic material that can affect phenotype [39, 40]. In this regard, evolution occurs within viruses in the same way it occurs within humans, animals, or any other organisms, and can be measured as such [41].

HIV-1 belongs to the family of RNA retroviruses, which exhibit some of the highest evolutionary rates of any biological entity [42]. The evolution of HIV-1 is so fast than it is detectable not only between different individuals, but also within a single person over the course of their HIV-1 infection [42, 43]. This remarkable evolution, which is central to this work, is facilitated by several factors. For one, HIV-1 has a very fast generation time — the time from a virus leaving a cell until it produces a new virus in a new cell — which has been estimated at approximately 2.6 days [44]. Moreover, there are high counts, approximately 10^7, of infected CD4$^+$ cells within the body during infection, each of which is capable of producing roughly 10^3 new virions [45]. As mentioned previously, the highly error-prone reverse transcription process introduces between 1.4 and 3.4 $\times 10^{-5}$ mutations/nucleotide/replication which, when accounting for the 9.2 kb genome length of HIV-1, equates to roughly one new mutation per replication [24–26, 46]. Together, the genetic instability of HIV-1 along with the remarkable size and growth of its population enables the efficient generation of enormous diversity.

1.5.1 Nucleotide Substitutions and Evolutionary Models

Nucleotide substitutions are central to the study of evolution, as they are the specific mutation on which commonly-used models of evolution are based [47]. In other words, most models of evolution (also referred to as substitution models) depend on nucleotide substitutions in order to estimate evolutionary relatedness [47, 48]. Specifically, these models operate under the assumption that substitutions are a stochastic process and account for the probabilities of each nucleotide changing to any other nucleotide in a genetic sequence over time [47, 48]. The number of parameters in these models can vary considerably, though at least one of these parameters is typically an expected substitution rate. An example of one substitution model (Felsenstein ‘81) is represented below.

$$Q = \begin{pmatrix} * & \mu\pi_C & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & * & \mu\pi_G & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & * & \mu\pi_T \\ \mu\pi_A & \mu\pi_C & \mu\pi_G & * \end{pmatrix}$$

The above rate matrix Q contains a single rate parameter μ that controls how frequently all substitutions occur, which is multiplied by three different parameters describing the frequencies of all four nucleotides (π_A, π_C, π_G; not four parameters because $\pi_A + \pi_C + \pi_G + \pi_T = 1$). Diagonal values (*) represent a quantity such that each row must sum to 0.

Importantly, evolutionary models can be used to reconstruct the evolutionary history of a population in a procedure known as phylogenetic inference, which is central to this work which will be further discussed in a coming section. There are many different substitution models, each utilizing a unique set of parameters and making different assumptions. Briefly, some of the most prominent substitutions models, which are named by their authors, include the Jukes-Cantor, Kimura, Felsenstein, Hasegawa–Kishono–Yano, and Tamura-Nei models [49–53]. For instance, the Jukes-Cantor model, being among the earliest models, uses a single rate parameter, assumes equal substitution rates between nucleotides, and equal starting distributions of all nucleotides [52]. Additional complexity was then added to the subsequent advancements of these models, two of which — the Felsenstein and Tamura-Nei models — are used in this work. The Felsenstein 81 (F81) model adds parameterization of nucleotide frequencies allowing their starting distributions to differ, while still maintaining a single substitution rate parameter [49]. The Tamura-Nei 93 (TN93) not only parameterizes nucleotide frequencies, but also incorporates three different rate parameters: one describing the rate of nucleotide transversions ($A/G \leftrightarrow C/T$) and one for each type of transition ($A \leftrightarrow G$, $C \leftrightarrow T$) [51]

Substitution rates are a fundamental parameter in the field of viral evolution as they reflect a virus' ability to adapt to an individual host or entire population (*i.e.* humans) [54]. By rapidly acquiring genetic change, HIV-1 populations are remarkably resilient and versatile when infecting humans, as they can quickly respond to different environmental pressures such as those induced by the host-specific immune responses.

1.5.2 Selection

Selection, a concept originally proposed by Charles Darwin, refers to the interactions between an organism's phenotype and the surrounding environment that affect the organism's survivability and reproductive success, and is a concept central to the study of evolution [55]. Positive selection describes the process by which mutations conferring a fitness advantage increase in frequency within a population and is indicative of evolution that promotes adaptation to the environment (adaptive evolution) [56]. On the contrary, negative or purifying selection refers to the removal of mutations with relatively lower fitness from the next generation [57, 58]. Populations can also be subject to neutral selection in which case organism-environment interactions have minimal effects on allele frequencies, meaning that fluctuations in variants are primarily modulated by stochastic processes [59]. Evidently, selection is an important consideration when studying viral mutations, as these forces can significantly modulate their prevalence within populations. For instance, selection can permit mutations to reach fixation: the increase of a genetic variant's prevalence to 100% in the population [59]. Selection is applicable to all biological entities which contain a genome coding for its phenotypic traits and inherit genetic information, which importantly includes viruses like HIV-1 [40].

Interestingly, both positive and purifying selection shape the evolution of HIV-1 within hosts and have been detected in numerous regions throughout the HIV-1 genome, with some conflicting findings in particular genes [57, 60, 61]. Although these trends are more complex when analyzed at a smaller resolution, the *gag* and *gp41* genes generally tend to undergo more purifying selection, while *nef* and *pol* are often subject to more positive selection [57, 62, 63]. Importantly, HIV-1 *gp120* undergoes both positive and purifying selection. The effects of purifying selection are most prominent in the conserved regions of gp120, meaning that mutations here are frequently detrimental and get filtered out of the population [42, 58]. Areas experiencing positive selection, on the other hand, are primarily found in this protein's variable regions (V1, V2, V3) and also show strong colocalization with known antibody and CTL binding sites. It is widely accepted that mutations at these sites confer a notable advantage and undergo corresponding positive selection because they enable gp120 to avoid recognition by the host immune response [57, 60, 64, 65].

1.5.3 Phylogenetic Inference

The study of evolution frequently relies on phylogenetic inference: the inference of a population's evolutionary relationships [66]. Central to phylogenetic inference procedures is the reconstruction of a phylogenetic tree, which is a model that describes the evolutionary relatedness between different biological entities [67]. Phylogenetic trees contain their own parameters — the order of branching (topology) and the lengths of each branch — and depend on a substitution model to calculate their likelihood, *i.e.* the probability of observing the data given the model parameters [67]. Generally, the process of tree reconstruction first involves fitting a model of evolution (discussed previously) to the population's genetic sequence data, and subsequently solving for the parameter values of this model that best fit the data [66]. This procedure, which is central to this thesis work, can be accomplished using multiple different statistical inference methods. For example, a common approach is estimation by maximum likelihood (ML), which involves determination of the parameter values that maximize the model's associated likelihood function. Application of phylogenetic inference to viruses can be quite fascinating and remarkable because, as previously mentioned, the virus evolutionary rates far surpass those of other organisms [68]. Thus, we are able to observe evolution in short time intervals that, in another organism, may take thousands or millions of years to occur (Figure 1.1) [68].

1.5.4 Indels

Insertions and deletions (indels) are a unique mutational mechanism given that they add or remove genetic information and alter the overall length of sequences, unlike nucleotide substitutions that only change sequence composition [69]. While nucleotide substitutions form the basis of commonly-used evolutionary models, there is inherent difficulty and uncertainty associated with reconstructing indels and utilizing their information in phylogenetic inference due to the fact that indels cannot be directly observed like substitutions [70]. In other words, there is no direct indication that sequence has been inserted or deleted and thus the locations of indels can only be inferred [70]. As a result, the process of reconstructing indels in a sequence alignment is somewhat arbitrary and can change with sequence alignment parameters, leading to the exclusion of indels from many common phylogenetic methods [70–72]. Newer phylo-

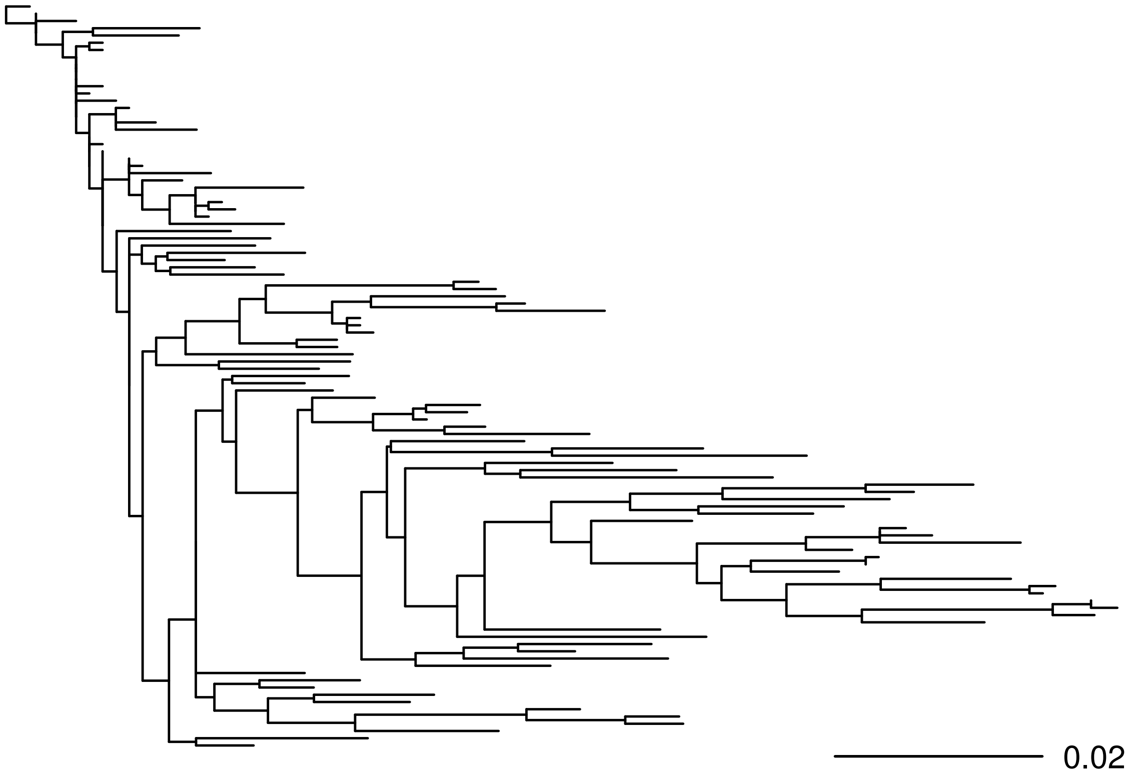

Figure 1.1: A phylogenetic tree generated using maximum likelihood methods that describes an HIV-1 infection within a single host. The bar in the bottom right corner provides a scale of tree branch length units: expected substitutions per site.

genetics methods, while not yet widespread, have begun testing and using the incorporation of indels and have seen increases in accuracy for doing so [69–71] Despite their absence from phylogenetic inference methods, indels still play a significant role in virus evolution and have been shown to confer unique phenotypic traits in multiple viruses, not just HIV-1. For example, specific insertions in the *gag* and *pol* genes of HIV-1 have been shown to confer enhanced infectivity and drug resistance to the virus, respectively [73–75].

1.5.5 Recombination

Recombination refers to the incorporation of genetic information from two or more different sources into the same genetic sequence [76]. In HIV-1, recombination occurs very frequently and plays a significant role in shaping the evolution of this virus [42]. The estimated recombination rate of HIV-1 is roughly 1.4×10^{-5} events per site per cycle, which is comparable to the estimated rate of substitutions [42]. HIV-1 is predisposed to undergo recombination, based on its structure and mechanisms of replication [11]. Foremost, recombination is facilitated by the diploid nature of HIV-1, meaning that it contains two copies of its genome per viral particle [76]. If a susceptible cell becomes infected with two different HIV-1 particles, it is possible for two genetically-distinct HIV-1 genomes to get incorporated into a single HIV-1 particle [76]. Recombination may then occur when this HIV-1 particle enters a new cell and undergoes reverse transcription [76]. Additionally, reverse transcription of HIV-1 involves the dissociation and transfer of RT between template strands, meaning that affinity between the RT enzyme and the nucleotide sequence is relatively low [11]. It is plausible that RT could be transferred incorrectly to a second proximal RNA molecule during this step, thereby creating a recombinant RNA genome containing portions of both templates [11].

1.5.6 Diversity & Evolution Within Hosts

Recall that HIV-1 evolves so rapidly that substantial diversity and evolution is found not only among different hosts, but within the HIV-1 population circulating in a single infected host too

[77]. This provides two different scales at which HIV-1 diversity and evolution can be studied.

As previously discussed, evolution on a within-host scale is driven by the virus' large population sizes, efficient replication, and rapid accumulation of mutations via substitutions, indels, and recombination [24, 25, 42, 44]. Interestingly, these factors facilitate a highly heterogeneous HIV-1 population structure within hosts, suggested to be comprised of numerous smaller subpopulations that each originate from a different virus variant and thus have a distinct genetic profile [78–80]. This population structure, described as a metapopulation or quasispecies, further enables HIV-1 to rapidly adapt to the host environment [81] and is believed to be partially facilitated by its localization to multiple different bodily compartments, which include the blood, lymph nodes, genital tract, and brain [78, 80, 82].

This complex population structure is further shaped by evolutionary forces, which include, but are not limited to, positive and negative (purifying) selection. For one, HIV-1 primarily undergoes adaptive evolution within hosts to evade the cell- and antibody-mediate host immune responses, which involves the positive selection of advantageous mutations that confer immune escape [42, 60, 61, 65]. This adaptive evolution is so prevalent that HIV-1 is further suggested to be “short-sighted” by some, in that increases in viral fitness are almost always favoured even to the extent of sacrificing transmission fitness [83]. These advantageous mutations get successively fixed in the population over time, causing the gradual and consistent divergence away from the founding variant and toward a well-adapted one, such as those capable of evading the immune system [54, 59]. Adaptive evolution accounts for the highly directional nature of within host evolution, as demonstrated by the asymmetrical shape of the phylogenetic tree in Figure 1.1. Importantly, while positive selection predominates, there is also purifying selection acting within hosts which filters out detrimental mutations that compromise virus functionality [58]. Therefore, both positive and purifying selection exert modulating effects on genetic variant frequencies and importantly, influence which mutations will reach fixation in the population [59].

1.5.7 Diversity & Evolution Among Hosts

HIV-1 exhibits considerable genetic differences between different human hosts, which are even more pronounced between human populations in different geographical regions. After initial transmission to humans in Central Africa, HIV-1 group M spread globally and formed further distinct clades: populations containing all the descendants that originate from a single ancestor [3]. Based on these distinct clades, HIV-1 group M was stratified into nine major subtypes (A-D, F-J) which differ in their amino acid sequences by 10% to 35% depending on the examined gene [3]. Currently, the group M subtypes are localized to different geographical regions due to historical founder effects: the reduction and change of genetic diversity caused by few individuals from a large population entering a novel environment [3]. For example, HIV-1 subtype B is predominantly found in North America, while subtype C is found heavily in Africa and Eastern Asia [3]. The clades of group M are an important factor to consider when studying HIV-1 given they are suggested to differ in their mother-to-child transmission efficiency [84], tropism phenotype [85], genital mucosa injury [86], Nef-mediated suppression of the CTL response [87], and rate of disease progression [88].

The evolutionary processes that shape among host diversity are primarily associated with transmission. The transmission of HIV-1 to a new host is associated with strong population bottleneck effects, which refers to the drastic changes to genetic diversity caused by a reduction in population size [78]. In fact, it has been found that approximately 80% of new HIV subtype B and C infections originate from only a single T/F virus, despite there often being a large and diverse HIV population within the transmitting host [89–91]. The genotypes of the one or few founding viruses will therefore, dictate those of the whole population after sufficient viral expansion, allowing the new viral population to become considerably different from its original. Additionally, transmitted HIV-1 variants have been found to be more similar to the virus that originally established its infection than the current HIV-1 population. [92]. This apparent reversion of genetic diversity has been described in literature as the “store and retrieve” hypothesis [92]. This proposed trend in HIV-1 evolution accounts for the lower evolutionary rates observed among hosts than within hosts and the apparent loss of genetic changes in new founding viruses relative to their populations of origin [93, 94]. Mechanistically, this may be associated with the strong transmission bottleneck of HIV-1 that appears to select for a particular phenotype, *i.e.* the R5 tropism phenotype present in 100% of T/F viruses [32, 95].

1.6 HIV-1 Pathogenesis

Pathogenesis refers to the biological mechanisms responsible for inducing disease [96]. HIV-1 induces disease in humans by infecting and destroying CD4$^+$ immune cell populations, the most important of which are those of T helper cells [97]. The pathogenesis of HIV-1 involves multiple different stages of infection, each with unique and important concepts [97].

1.6.1 Transmission

Transmission of HIV-1 to a new host involves exposure of the virus either to bodily mucosal layers or the subcutaneous tissue [98]. The most common routes of HIV transmission include homo- and heterosexual contact, mother-to-child exposure, and subcutaneous injection [98]. Each route of transmission has a different level of associated risk based on the permissiveness of the tissue and the common medium by which virus is transferred [98].

1.6.2 Immune Response to HIV-1

Shortly after the establishment of HIV-1 infection, a substantial cell-mediated immune response is mounted against HIV-1, involving activation of numerous CD8$^+$ cytotoxic T lymphocytes (CTLs) [99]. CTLs play a critical role in fighting viral infections due to their functional focus on eliminating intracellular pathogens. To detect infected cells, CTLs rely on the presentation of viral peptides on human leukocyte antigen (HLA) class 1 molecules, and upon doing so, can then lyse these cells using stored perforin and granzyme proteins [99]. The CTL response can target a wide range of viral peptides spanning many of the HIV-1 functional proteins, including those encoded by *gag*, *pol*, and *env* [100]. The efficiency of the CTL response is further dependent on different HLA class 1 molecules, which display substantial variation particularly between ethnicities [101]. People can therefore present with considerable differences in response to HIV-1 infection based on their type of HLA class 1 molecule [101]. While CTL-induced cell death is the primary antiviral response, CTLs also release cytokines that suppress HIV-1 replication and entry in nearby cells [99].

A few weeks after the start of infection and the initiation of the CTL response, the humoral immune system — the portion of the immune system responsible for antibody production — begins mounting a response against HIV-1, involving the production of both neutralizing and non-neutralizing antibodies [102]. Briefly, neutralizing antibodies suppress infection by binding to and incapacitating free-floating virus within circulation, while non-neutralizing ones activate other components of the antiviral immune response — including the protein complement system, phagocytes, and natural killer cells — to destroy virus and infected cells in an indirect manner [102]. To bind to the virus, neutralizing antibodies can specifically target the gp120, gp41, matrix (p17), and capsid (p24) proteins of HIV-1, though the majority of responses are elicited against gp120 due to its extracellular position [102, 103]. The majority of subjects infected with HIV-1 elicit a neutralizing antibody response against gp120, which contributes to suppression of HIV-1 viral load [103].

1.6.3 Immune Escape

The cell- and antibody-mediated immune response both contribute to the initial suppression of circulating HIV-1 populations, or viral load [60]. In the majority of cases however, neither response is able to effectively clear HIV-1 infection from the body, resulting in the progression to AIDS for which there is no cure. The inability to clear HIV-1 is primarily due to the virus population's high heterogeneity and rapid evolution, which enables the generation of variants that evade this response in a process referred to as immune escape [60, 104]. Recall that both the CTL and humoral immune responses bind to specific viral protein structures in order to detect and neutralize virally-infected cells or the virus itself [60]. Of the numerous HIV-1 variants in circulation, those viruses containing the targeted protein variant will be less likely to propagate into further generations. On the other hand however, virus variants that acquire mutations at the targeted protein site which prevent recognition by the immune responses will undergo positive selection, as they receive a strong evolutionary advantage and experience less competition for susceptible cells [99, 105]. Under positive selection, these immune escape variants will increase in frequency, gradually making more of the HIV-1 population resistant to the immune response [106]. It is important to note, however, that this process is a constant battle between virus and immune response as opposed to a single pathway described above.

1.6.4 Phases of Infection

There are three major phases that describe the course of HIV-1 infection: a short acute phase, a long asymptomatic phase, and the final immune compromised phase (AIDS) [107]. During the acute phase of infection which typically lasts between 2 and 10 weeks, HIV-1 rapidly replicates within susceptible CD4$^+$ cells, resulting in a large viral load and importantly, a sharp decline of helper T cells [107]. The affected individual will present with symptoms of an acute infection, which commonly include fever, headache and pharyngitis [107]. Later in this phase, an HIV-specific CTL immune response reaches full capacity, which depletes the viral population considerably and permits the re-population of CD4$^+$ helper T cells [107]. At the end of the acute phase, HIV-1 is typically under sufficient control at low to moderate levels and will establish the set point viral load: the stabilized level of circulating virus that persists for the next asymptomatic phase of infection [108, 109]. The set point viral load can differ dramatically between infected hosts and importantly predicts the rate of disease progression over the remaining course of infection [108, 109].

Next is the asymptomatic phase of HIV-1 infection which typically lasts between 7 and 10 years but demonstrates large variation in duration between patients [107, 110]. During this period, the patient presents without symptoms of infection as the virus is suppressed to relatively moderate or low levels in the body [107]. However, the virus continues to infect susceptible cells and evolve in this suppressed state [111]. Importantly, it is during this stage that the HIV-1 population predominantly undergoes adaptive evolution in response to host immune responses, thereby giving rise to immune escape variants under positive selection [59, 111]. The continual generation of new escape variants gradually increases the viral load during the asymptomatic phase until this resistant population becomes large enough to fully overwhelm the immune system [59, 111].

The final stage, AIDS, occurs when the virus begins to overwhelm the immune response, typically due to high prevalence of escape mutants in the viral population and the sufficient depletion of helper T lymphocytes [112]. Although all CD4$^+$ cell types play an important role in immune responses, AIDS is specifically defined by the sufficient depletion of CD4$^+$ helper T lymphocytes specifically [112, 113]. Helper T cells are essential to the activation and specificity of both the antibody- and cell-mediated responses of the adaptive immune system [114]. In AIDS, helper T cell populations are depleted, compromising the entire adaptive immune system and leading to an inability to combat even low virulence pathogens [112–114].

1.7 gp120

My research focuses specifically on the gp120 surface glycoprotein of HIV-1. Recall that this protein mediates initial attachment of HIV-1 to susceptible immune cells by binding the CD4 receptor and a second chemokine coreceptor [115]. The *env* gene of HIV-1 that codes for gp120 and gp41 exhibits higher rates of evolution than other genes such as *gag* and *pol* [43].

1.7.1 Glycosylation

The gp120 glycoprotein is one of the most heavily glycosylated proteins known in nature, as approximately 50% of its molecular weight is associated with its numerous glycan sugar complexes [116]. The abundant glycosylation on gp120 is facilitated by this protein's numerous potential N-linked glycosylation sites (PNGS): amino acid sequences that encode the post-translational attachment of a glycan to asparagine (N) residues. This dense, protective layer of glycan sugars is referred to as the “glycan shield” and protects the susceptible conserved regions of this protein often targeted by neutralizing antibodies [117]. Interestingly, the gp120 glycans themselves can be the target of neutralizing antibodies generated by the host immune response as well [118]. The broadly neutralizing antibodies 2G12 and PGT128, for example, have been shown to bind to specific glycans on gp120 [118, 119]. Therefore, N-linked glycans on gp120 are susceptible to selective pressures due to their interactions with the immune response [118]. Importantly, the glycan shield can respond to these pressures because it is highly dynamic in nature. The frequent mutations in gp120 can create or remove PNGSs in its protein sequence and thus alter the protein’s overall glycan sugar patterns [117]. In fact, increases in gp120 glycan counts over the course of infection, particularly in the variable loops, has been reported by both Derdeyn et al. [120] and Sagar et al. [121], supporting the hypothesis that gp120 “raises” its glycan shield to evade immune responses.

1.7.2 Variable Loops

Another important feature of the gp120 gene is its distinct genetic segmentation into five conserved (C1-C5) and five variable regions (V1-V5), which has been well documented since being reported in 1986 by Starcich et al. [122]. In the physical protein structure, the conserved regions of this protein are primarily hidden in its interior, while the variable regions form five disordered loop structures on the protein’s exterior anchored by disuphide bonds at their base. The variable loops are located on the exterior on gp120 and therefore, experience notable selective pressures from neutralizing antibody responses [117]. This selection facilitates the rapid accumulation of sequence mutations in the variable loops, which include both substitutions and indels [18]. The substitution rates in these regions, for example, are even higher than that of *env* as a whole [18, 123]. Mutations can generate substantial structural variability in these loops by not only changing the compositions and lengths of the loops, but also by changing variable loop glycosylation patterns [18]. The high variability observed in the variable loops makes them a difficult target and thus offers protection for some of the more conserved regions of gp120, including the CD4 binding site for example [18].

1.7.3 Immune Escape

While gp120 is targeted by neutralizing antibody responses, these responses remain unable to effectively bind gp120 due to its dense glycan shield and rapidly mutating variable loops [18, 117, 117, 121]. The elusive nature of gp120 is driven by mutations in the variable loops that frequently generate immune escape variants by rapidly changing the lengths, compositions, and glycan content of these loop structures [18, 117]. As discussed previously, gp120 escape variants experience a significant evolutionary advantage and increase in prevalence due to the considerable positive selection applied by the neutralizing antibody response [18, 121, 124]. Escape from these antibodies allows HIV-1 to remain functional within circulation and continue to infect new immune cells [118].

The gp120 variable loops have been shown to accumulate an abundance of indel events [18]. Indels specifically can induce many of the aforementioned changes to the variable loops (*i.e.* lengths, glycosylation patterns, amino acid compositions) that have been reported to generate gp120 immune escape variants [18, 105, 117, 125]. These findings implicate variable loop indels in the onset of gp120 immune escape and provide the core rationale for investigating these particular mutations in this work [18].

1.7.4 Variable Loop Indel Rates

Rates of sequence evolution are important parameters to the study of HIV-1 for a few reasons. For instance, substitution rates are generally suggested to influence a biological entity's capacity to adapt to dynamic environmental pressures and in the case of viral infections, are suggested to correlate with the rate of disease progression [26, 54]. Rates of sequence evolution also allow mutations to be used as a source of phylogenetic information in inference methods when incorporated into a suitable model, *i.e.* like the use of nucleotide substitutions in current phylogenetic methods [49, 51].

While nucleotide substitution rates have been extensively studied in HIV-1 [43, 123, 126], there are no studies, to our knowledge, that have investigated and estimated rates of indel evolution in HIV-1 occurring in human hosts. It is important to note that Mansky and Temin [24] estimated the rate of indel accumulation in the HIV-1 genome using *in vitro* methods; however, this study reports an indel mutation rate, not a rate of indel evolution that we are investigating in this work. A mutation rate measures how frequently cellular machinery generates mutations during a single viral replication cycle within a cell [43]. On the other hand, evolutionary rates measure the relative accumulation of mutations among population lineages as they diverge over time and importantly, are driven by forces such as natural selection and genetic drift [43]. In other words, mutation rates are describing the creation of raw genetic diversity, while evolutionary rates are describing how mutations are shaped within a population as it interacts with its environment.

Without rates of indel evolution or an effective model describing these processes, indels remain an unused source of phylogenetic information in the study of HIV-1. Estimation of HIV-1 indel evolution rates would therefore, mark an important first step towards their effective incorporation into phylogenetic inference methods. Moreover, the contributions of indels towards HIV-1 sequence evolution and adaptation in specific genetic regions (*i.e.* the gp120 variable loops) also remain unknown without estimates of their rates. No studies have attempted to estimate indel rates specifically within the gp120 gene or the variable regions where indels are biologically significant in their contributions toward immune escape. It is therefore clear that indel rates, particularly in HIV-1 gp120, are an important topic needing to be addressed in modern HIV-1 research through the use of patient-derived genetic sequence data.

1.8 Objectives

In this work, I first address this knowledge gap by providing the first estimates of gp120 variable loop indel rates on both a within-host and among-host scale using phylogenetic methods and compare these estimates to each other. I further examine lengths and compositions of recovered indels, and importantly, how frequently indels induce changes to PNGSs that form the glycan shield of gp120. Finally, I investigate whether the characteristics of observed variable loop indels can be effectively described using an empirical model, with aims of understanding the parameters governing indel generation.

1.9 Hypothesis

I foremost hypothesize that indel rates can be determined in the variable loops of gp120 using dated phylogenetic analyses on both among-host and within-host scales. On an among-host scale, I predict that indel rates will differ significantly among the group M subtypes, showing concordance with subtype differences in phenotype. I further expect to observe significantly higher indel rates within hosts than among hosts, due to the strong purifying selection acting on the level of transmission events in HIV-1. I also predict finding more indels that induce frameshift mutations within hosts than among hosts, on the same basis that there is less purifying selection acting within hosts than among hosts. Finally, I postulate that the characteristics of gp120 indels can be effectively described using an empirical model of replication slippage.

Bibliography

[1] Joris Hemelaar. The origin and diversity of the HIV-1 pandemic. *Trends in molecular medicine*, 18(3):182–192, 2012.

[2] Feng Gao, Elizabeth Bailes, David L Robertson, Yalu Chen, Cynthia M Rodenburg, Scott F Michael, Larry B Cummins, Larry O Arthur, Martine Peeters, George M Shaw, et al. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature*, 397 (6718):436–441, 1999.

[3] Denis M Tebit and Eric J Arts. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *The Lancet infectious diseases*, 11 (1):45–56, 2011.

[4] Joseph M Watts, Kristen K Dang, Robert J Gorelick, Christopher W Leonard, Julian W Bess Jr, Ronald Swanstrom, Christina L Burch, and Kevin M Weeks. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 460(7256):711–716, 2009.

[5] Rafael Correa and Ma Angeles Muñoz-Fernández. Viral phenotype affects the thymic production of new T cells in HIV-1-infected children. *Aids*, 15(15):1959–1963, 2001.

[6] A Telesnitsky and SP Goff. Reverse transcriptase and the generation of retroviral DNA. 1997.

[7] Paul M Sharp and Beatrice H Hahn. Prehistory of HIV-1. *Nature*, 455(7213):605–606, 2008.

[8] Kevin K Ariën, Awet Abraha, Miguel E Quinones-Mateu, Luc Kestens, Guido Vanham, and Eric J Arts. The replicative fitness of primary human immunodeficiency virus type 1 (HIV-1) group M, HIV-1 group O, and HIV-2 isolates. *Journal of virology*, 79(14): 8979–8990, 2005.

[9] Global HIV & AIDS statistics — 2019 fact sheet, url=https://www.unaids.org/en/resources/fact-sheet, journal=United Nations AIDS, publisher=United Nations AIDS, year=2019, month=.

[10] HIV/AIDS, url=https://www.who.int/health-topics/hiv-aids/, journal=World Health Organization, publisher=World Health Organization, year=2020, month=.

[11] Eric O Freed. HIV-1 replication. *Somatic cell and molecular genetics*, 26(1-6):13–33, 2001.

[12] Steven R King. HIV: virology and mechanisms of disease. *Annals of emergency medicine*, 24(3):443–449, 1994.

[13] Y Li, Lizhong Luo, N Rasool, and C Yong Kang. Glycosylation is necessary for the correct folding of human immunodeficiency virus gp120 in CD4 binding. *Journal of virology*, 67(1):584–588, 1993.

[14] Chi-Huey Wong. Protein glycosylation: new challenges and opportunities. *The Journal of organic chemistry*, 70(11):4219–4225, 2005.

[15] Beatriz Apellániz, Edurne Rujas, Soraya Serrano, Koldo Morante, Kouhei Tsumoto, Jose MM Caaveiro, M Ángeles Jiménez, and José L Nieva. The atomic structure of the HIV-1 gp41 transmembrane domain and its connection to the immunogenic membrane-

proximal external region. *Journal of Biological Chemistry*, 290(21):12999–13015, 2015.

[16] Stephen A Gallo, Catherine M Finnegan, Mathias Viard, Yossef Raviv, Antony Dimitrov, Satinder S Rawat, Anu Puri, Stewart Durell, and Robert Blumenthal. The HIV Env-mediated fusion reaction. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1614(1):36–50, 2003.

[17] Joseph G Sodroski. HIV-1 entry inhibitors in the side pocket. *Cell*, 99(3):243–246, 1999.

[18] Natasha Wood, Tanmoy Bhattacharya, Brandon F Keele, Elena Giorgi, Michael Liu, Brian Gaschen, Marcus Daniels, Guido Ferrari, Barton F Haynes, Andrew McMichael, et al. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS pathogens*, 5(5):e1000414, 2009.

[19] Ann M Sheehy, Nathan C Gaddis, Jonathan D Choi, and Michael H Malim. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–650, 2002.

[20] Wei Chun Goh, Mark E Rogel, C Matthew Kinsey, Scott F Michael, Patricia N Fultz, Martin A Nowak, Beatrice H Hahn, and Michael Emerman. HIV-1 Vpr increases viral expression by manipulation of the cell cycle: a mechanism for selection of Vpr in vivo. *Nature medicine*, 4(1):65–71, 1998.

[21] Stephan Bour and Klaus Strebel. The HIV-1 Vpu protein: a multifunctional enhancer of viral particle release. *Microbes and Infection*, 5(11):1029–1039, 2003.

[22] Kathleen L Collins, Benjamin K Chen, Spyros A Kalams, Bruce D Walker, and David Baltimore. HIV-1 Nef protein protects infected primary cells against killing by cytotoxic T lymphocytes. *Nature*, 391(6665):397–401, 1998.

[23] Craig B Wilen, John C Tilton, and Robert W Doms. HIV: cell binding and entry. *Cold Spring Harbor perspectives in medicine*, 2(8):a006866, 2012.

[24] Louis M Mansky and Howard M Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, 69(8):5087–5094, 1995.

[25] Michael E Abram, Andrea L Ferris, Wei Shao, W Gregory Alvord, and Stephen H Hughes. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *Journal of virology*, 84(19):9864–9878, 2010.

[26] José M Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely high mutation rate of HIV-1 in vivo. *PLoS Biol*, 13(9):e1002251, 2015.

[27] Bodo Linz, Helen M Windsor, John J McGraw, Lori M Hansen, John P Gajewski, Lynn P Tomsho, Caylie M Hake, Jay V Solnick, Stephan C Schuster, and Barry J Marshall. A mutation burst during the acute phase of helicobacter pylori infection in humans and rhesus macaques. *Nature communications*, 5(1):1–8, 2014.

[28] Serguei Popov, Michael Rexach, Gabriele Zybarth, Norbert Reiling, May-Ann Lee, Lee Ratner, Cynthia M Lane, Mary Shannon Moore, Günter Blobel, and Michael Bukrinsky. Viral protein R regulates nuclear import of the HIV-1 pre-integration complex. *The EMBO journal*, 17(4):909–917, 1998.

[29] Milena S Espíndola, Luana S Soares, Leonardo J Galvao-Lima, Fabiana A Zambuzi, Maira C Cacemiro, Verônica S Brauer, and Fabiani G Frantz. HIV infection: focus on the innate immune cells. *Immunologic research*, 64(5):1118–1132, 2016.

[30] K Kure, KM Weidenheim, WD Lyman, and Dennis W Dickson. Morphology and distribution of HIV-1 gp41-positive microglia in subacute AIDS encephalitis. *Acta neuropathologica*, 80(4):393–400, 1990.

[31] Paul R Clapham and Áine McKnight. HIV-1 receptors and cell tropism. *British medical bulletin*, 58(1):43–59, 2001.

[32] Jean-Charles Grivel, Robin J Shattock, and Leonid B Margolis. Selective transmission of R5 HIV-1 variants: where is the gatekeeper? *Journal of translational medicine*, 9(1): 1–17, 2011.

[33] Maureen M Goodenow and Ronald G Collman. HIV-1 coreceptor preference is distinct from target cell tropism: a dual-parameter nomenclature to define viral phenotypes. *Journal of leukocyte biology*, 80(5):965–972, 2006.

[34] John P Moore, Scott G Kitchen, Pavel Pugach, and Jerome A Zack. The CCR5 and CXCR4 coreceptors—central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection. *AIDS research and human retroviruses*, 20(1):111–126, 2004.

[35] Z-Q Zhang, T Schuler, M Zupancic, S Wietgrefe, KA Staskus, KA Reimann, TA Reinhart, M Rogan, Winston Cavert, Chris J Miller, et al. Sexual transmission and propagation of SIV and HIV in resting and activated CD4+ T cells. *Science*, 286(5443): 1353–1357, 1999.

[36] Paul U Cameron, Peter S Freudenthal, Jeanne M Barker, Stuart Gezelter, Kayo Inaba, and Ralph M Steinman. Dendritic cells exposed to human immunodeficiency virus type-1 transmit a vigorous cytopathic infection to CD4+ T cells. *Science*, 257(5068): 383–387, 1992.

[37] Donald E Mosier. How HIV changes its tropism: evolution and adaptation? *Current Opinion in HIV and AIDS*, 4(2):125, 2009.

[38] Muhammad Aqeel Ashraf and Maliha Sarfraz. Biology and evolution of life science. *Saudi journal of biological sciences*, 23(1):S1, 2016.

[39] Marc HV Van Regenmortel and Brian WJ Mahy. Emerging issues in virus taxonomy. *Emerging infectious diseases*, 10(1):8, 2004.

[40] U Kutschera. Evolution. Reference module in life sciences. Article 06399, 2017.

[41] Esteban Domingo, Cristina Escarmís, Noemi Sevilla, Andres Moya, Santiago F Elena, Josep Quer, Isabel S Novella, and John J Holland. Basic concepts in RNA virus evolution. *The FASEB Journal*, 10(8):859–864, 1996.

[42] Richard A Neher and Thomas Leitner. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol*, 6(1):e1000660, 2010.

[43] Samuel Alizon and Christophe Fraser. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology*, 10(1):49, 2013.

[44] Alan S Perelson, Avidan U Neumann, Martin Markowitz, John M Leonard, and David D Ho. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586, 1996.

[45] Tae-Wook Chun, Lucy Carruth, Diana Finzi, Xuefei Shen, Joseph A DiGiuseppe, Harry Taylor, Monika Hermankova, Karen Chadwick, Joseph Margolick, Thomas C Quinn, et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature*, 387(6629):183–188, 1997.

[46] Martin A Nowak, Robert M May, and Roy M Anderson. The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease. *Aids*, 4(11): 1095–1103, 1990.

[47] Supratim Choudhuri. *Bioinformatics for beginners: genes, genomes, molecular evolution, databases and analytical tools*. Elsevier, 2014.

[48] Korbinian Strimmer, Arndt von Haeseler, Anne-Mieke Salemi, et al. Nucleotide substitution models. In *The Phylogenetics Handbook A Practical Approach to DNA and Protein Phylogeny*, pages 72–100. Cambridge University Press, 2003.

[49] Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

[50] Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120, 1980.

[51] Koichiro Tamura and Masatoshi Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512–526, 1993.

[52] Thomas H Jukes, Charles R Cantor, et al. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.

[53] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2):160–174, 1985.

[54] Ha Youn Lee, Alan S Perelson, Su-Chan Park, and Thomas Leitner. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. *PLoS Comput Biol*, 4(12):e1000240, 2008.

[55] Charles Darwin and William F Bynum. *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. AL Burt New York, 2009.

[56] Pardis C Sabeti, Stephen F Schaffner, Ben Fry, Jason Lohmueller, Patrick Varilly, Oleg

Shamovsky, Alejandro Palma, TS Mikkelsen, D Altshuler, and ES Lander. Positive natural selection in the human lineage. *science*, 312(5780):1614–1620, 2006.

[57] Stacy A Seibert, Carina Y Howell, Marianne K Hughes, and Austin L Hughes. Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1). *Molecular biology and evolution*, 12(5):803–813, 1995.

[58] CTT Edwards, EC Holmes, OG Pybus, DJ Wilson, RP Viscidi, EJ Abrams, RE Phillips, and AJ Drummond. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics*, 174(3):1441–1453, 2006.

[59] Andrew Rambaut, David Posada, Keith A Crandall, and Edward C Holmes. The causes and consequences of HIV evolution. *Nature Reviews Genetics*, 5(1):52–61, 2004.

[60] David A Price, Philip JR Goulder, Paul Klenerman, Andrew K Sewell, Philippa J Easterbrook, Maxine Troop, Charles RM Bangham, and Rodney E Phillips. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proceedings of the National Academy of Sciences*, 94(5):1890–1895, 1997.

[61] Lamei Chen, Alla Perlina, and Christopher J Lee. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. *Journal of virology*, 78(7):3722–3732, 2004.

[62] Calvin Pan, Joseph Kim, Lamei Chen, Qi Wang, and Christopher Lee. The HIV positive selection mutation database. *Nucleic acids research*, 35(suppl_1):D371–D375, 2007.

[63] M de A Paolo, Esper G Kallas, Robson F de Souza, and Edward C Holmes. Genealogical evidence for positive selection in the nef gene of HIV-1. *Genetics*, 153(3):1077–1089, 1999.

[64] Susanna L Lamers, John W Sleasman, Jin-Xiong She, Kimberly A Barrie, Steven M Pomeroy, Douglas J Barrett, and Maureen M Goodenow. Independent variation and positive selection in env V1 and V2 domains within maternal-infant strains of human immunodeficiency virus type 1 in vivo. *Journal of virology*, 67(7):3951–3960, 1993.

[65] Rasmus Nielsen and Ziheng Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–936, 1998.

[66] Stephen Jay Gould. *Ontogeny and phylogeny*. Harvard University Press, 1977.

[67] G Weyenberg and R Yoshida. Phylogenetic Tree Distances. 2016.

[68] David A Steinhauer and JJ Holland. Rapid evolution of RNA viruses. *Annual Reviews in Microbiology*, 41(1):409–431, 1987.

[69] Bhakti Dwivedi and Sudhindra R Gadagkar. Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evolutionary Biology*, 9(1):211, 2009.

[70] Haim Ashkenazy, Ofir Cohen, Tal Pupko, and Dorothee Huchon. Indel reliability in indel-based phylogenetic inference. *Genome biology and evolution*, 6(12):3199–3209, 2014.

[71] Benjamin D Redelings and Marc A Suchard. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC evolutionary biology*, 7(1):40, 2007.

[72] Joan Pons and Alfried P Vogler. Size, frequency, and phylogenetic signal of multiple-residue indels in sequence alignment of introns. *Cladistics*, 22(2):144–156, 2006.

[73] Abdelrahim Rakik, Mounir Ait-Khaled, Philip Griffin, Deborah A Thomas, Margaret Tisdale, et al. A Novel Genotype Encoding a Single Amino Acid Insertion and Five Other Substitutions Between Residues 64 and 74 of the HIV-1 Reverse Transcriptase

Confers High-Level Cross-Resistance to Nucleoside Reverse Transcriptase Inhibitors. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 22(2):139–145, 1999.

[74] J Jacques de Jong, Jaap Goudsmit, Vladimir V Lukashov, Milly E Hillebrand, Elly Baan, Raymond Huismans, Sven A Danner, H Jacobus, Frank de Wolf, and Suzanne Jurriaans. Insertion of two amino acids combined with changes in reverse transcriptase containing tyrosine-215 of HIV-1 resistant to multiple nucleoside analogs. *Aids*, 13(1):75–80, 1999.

[75] Shambhu G. Aralaguppe, Dane Winner, Kamalendra Singh, Stefan G. Sarafianos, Miguel E. Quiñones-Mateu, Anders Sönnerborg, and Ujjwal Neogi. Increased replication capacity following evolution of PYxE insertion in Gag-p6 is associated with enhanced virulence in HIV-1 subtype C from East Africa. *Journal of Medical Virology*, 89 (1):106–111, January 2017. ISSN 1096-9071. doi: 10.1002/jmv.24610.

[76] Rafael Nájera, Elena Delgado, Lucía Pérez-Alvarez, and Michael M Thomson. Genetic recombination and its role in the development of the HIV-1 pandemic. *Aids*, 16:S3–S16, 2002.

[77] Philippe Lemey, Andrew Rambaut, and Oliver G Pybus. HIV evolutionary dynamics within and among hosts. *AIDs Rev*, 8(3):125–140, 2006.

[78] Simon DW Frost, Marie-Jeanne Dumaurier, Simon Wain-Hobson, and Andrew J Leigh Brown. Genetic drift and within-host metapopulation dynamics of HIV-1 infection. *Proceedings of the National Academy of Sciences*, 98(12):6975–6980, 2001.

[79] Andreas Meyerhans, Rémi Cheynier, Jan Albert, Martina Seth, Shirley Kwok, John Sninsky, Linda Morfeldt-Månson, Birgitta Asjö, and Simon Wain-Hobson. Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. *Cell*, 58(5):901–910, 1989.

[80] Li Yin, Li Liu, Yijun Sun, Wei Hou, Amanda C Lowe, Brent P Gardner, Marco Salemi, Wilton B Williams, William G Farmerie, John W Sleasman, et al. High-resolution

deep sequencing reveals biodiversity, population structure, and persistence of HIV-1 quasispecies within host ecosystems. *Retrovirology*, 9(1):1–9, 2012.

[81] Antonio V Bordería, Ramon Lorenzo-Redondo, Maria Pernas, Concepción Casado, Tamara Alvaro, Esteban Domingo, and Cecilio Lopez-Galindez. Initial fitness recovery of HIV-1 is associated with quasispecies heterogeneity and can occur without modifications in the consensus sequence. *PloS one*, 5(4):e10319, 2010.

[82] Jason T Blackard. HIV compartmentalization: a review on a clinically important phenomenon. *Current HIV research*, 10(2):133–142, 2012.

[83] Kristof Theys, Pieter Libin, Andrea-Clemencia Pineda-Pena, Ann Nowe, Anne-Mieke Vandamme, and Ana B Abecasis. The impact of HIV-1 within-host evolution on transmission dynamics. *Current opinion in virology*, 28:92–101, 2018.

[84] Boris Renjifo, Peter Gilbert, Beth Chaplin, Gernard Msamanga, Davis Mwakagile, Wafaie Fawzi, Max Essex, Tanzanian Vitamin, HIV Study Group, et al. Preferential in-utero transmission of HIV-1 subtype C as compared to HIV-1 subtype A or D. *Aids*, 18(12):1629–1636, 2004.

[85] Charlotte Tscherning, Annette Alaeus, Robert Fredriksson, Åsa Björndal, HongKui Deng, Dan R Littman, Eva Maria Fenyö, and Jan Albert. Differences in chemokine coreceptor usage between genetic subtypes of HIV-1. *Virology*, 241(2):181–188, 1998.

[86] Grace C John-Stewart, Ruth W Nduati, Christine M Rousseau, Dorothy A Mbori-Ngacha, Barbra A Richardson, Stephanie Rainwater, Dana D Panteleeff, and Julie Overbaugh. Subtype C is associated with increased vaginal shedding of HIV-1. *The Journal of infectious diseases*, 192(3):492–496, 2005.

[87] Jaclyn K Mann, Helen Byakwaga, Xiaomei T Kuang, Anh Q Le, Chanson J Brumme, Philip Mwimanzi, Saleha Omarjee, Eric Martin, Guinevere Q Lee, Bemuluyigza Baraki,

et al. Ability of HIV-1 Nef to downregulate CD4 and HLA class I differs among viral subtypes. *Retrovirology*, 10(1):100, 2013.

[88] Noah Kiwanuka, Oliver Laeyendecker, Merlin Robb, Godfrey Kigozi, Miguel Arroyo, Francine McCutchan, Leigh Anne Eller, Michael Eller, Fred Makumbi, Deborah Birx, et al. Effect of human immunodeficiency virus Type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection. *The Journal of infectious diseases*, 197(5):707–13, 2008.

[89] M-R Abrahams, Jeffrey A Anderson, EE Giorgi, Cathal Seoighe, K Mlisana, L-H Ping, GS Athreya, Florette K Treurnicht, Brandon F Keele, N Wood, et al. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *Journal of virology*, 83(8):3556–3567, 2009.

[90] Brandon F Keele, Elena E Giorgi, Jesus F Salazar-Gonzalez, Julie M Decker, Kimmy T Pham, Maria G Salazar, Chuanxi Sun, Truman Grayson, Shuyi Wang, Hui Li, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proceedings of the National Academy of Sciences*, 105(21): 7552–7557, 2008.

[91] Richard E Haaland, Paulina A Hawkins, Jesus Salazar-Gonzalez, Amber Johnson, Amanda Tichacek, Etienne Karita, Olivier Manigart, Joseph Mulenga, Brandon F Keele, George M Shaw, et al. Inflammatory genital infections mitigate a severe genetic bottleneck in heterosexual transmission of subtype A and C HIV-1. *PLoS pathogens*, 5(1): e1000274, 2009.

[92] Andrew D Redd, Aleisha N Collinson-Streng, Nikolaos Chatziandreou, Caroline E Mullis, Oliver Laeyendecker, Craig Martens, Stacy Ricklefs, Noah Kiwanuka, Phyu Hninn Nyein, Tom Lutalo, et al. Previously transmitted HIV-1 strains are pref-

erentially selected during subsequent sexual transmissions. *The Journal of infectious diseases*, 206(9):1433–1442, 2012.

[93] Bram Vrancken, Andrew Rambaut, Marc A Suchard, Alexei Drummond, Guy Baele, Inge Derdelinckx, Eric Van Wijngaerden, Anne-Mieke Vandamme, Kristel Van Laethem, and Philippe Lemey. The genealogical population dynamics of hiv-1 in a large transmission chain: bridging within and among host evolutionary rates. *PLoS Comput Biol*, 10(4):e1003505, 2014.

[94] Katrina A Lythgoe and Christophe Fraser. New insights into the evolutionary rate of hiv-1 at the within-host and epidemiological levels. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741):3367–3375, 2012.

[95] Sarah B Joseph, Ronald Swanstrom, Angela DM Kashuba, and Myron S Cohen. Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. *Nature reviews Microbiology*, 13(7):414–425, 2015.

[96] Mukesh Meena, Prashant Swapnil, Andleeb Zehra, Mohd Aamir, Manish Kumar Dubey, Chandra Bali Patel, and RS Upadhyay. Virulence factors and their associated genes in microbes. In *New and Future Developments in Microbial Biotechnology and Bioengineering*, pages 181–208. Elsevier, 2019.

[97] Andrew J McMichael, Persephone Borrow, Georgia D Tomaras, Nilu Goonetilleke, and Barton F Haynes. The immune response during acute HIV-1 infection: clues for vaccine development. *Nature Reviews Immunology*, 10(1):11, 2010.

[98] George M Shaw and Eric Hunter. HIV transmission. *Cold Spring Harbor perspectives in medicine*, 2(11):a006965, 2012.

[99] Andrew J McMichael and Sarah L Rowland-Jones. Cellular immune responses to HIV. *Nature*, 410(6831):980–987, 2001.

[100] Beatriz Mothe, Anuska Llano, Javier Ibarrondo, Marcus Daniels, Cristina Miranda, Jennifer Zamarreño, Vanessa Bach, Rosario Zuniga, Susana Pérez-Álvarez, Christoph T Berger, et al. Definition of the viral targets of protective HIV-1-specific T cell responses. *Journal of translational medicine*, 9(1):208, 2011.

[101] Henry AF Stephens. HIV-1 diversity versus HLA class I polymorphism. *Trends in immunology*, 26(1):41–47, 2005.

[102] M Huber and A Trkola. Humoral immunity to HIV-1: neutralization and beyond. *Journal of internal medicine*, 262(1):5–25, 2007.

[103] Douglas D Richman, Terri Wrin, Susan J Little, and Christos J Petropoulos. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proceedings of the National Academy of Sciences*, 100(7):4144–4149, 2003.

[104] Maureen Goodenow, Thierry Huet, William Saurin, Shirley Kwok, John Sninsky, and Simon Wain-Hobson. HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. *Journal of acquired immune deficiency syndromes*, 2(4):344–352, 1989.

[105] Penny L Moore, Nthabeleng Ranchobe, Bronwen E Lambson, Elin S Gray, Eleanor Cave, Melissa-Rose Abrahams, Gama Bandawe, Koleka Mlisana, Salim S Abdool Karim, Carolyn Williamson, et al. Limited neutralizing antibody specificities drive neutralization escape in early HIV-1 subtype C infection. *PLoS Pathog*, 5(9):e1000598, 2009.

[106] AJ Leslie, KJ Pfafferott, P Chetty, R Draenert, MM Addo, M Feeney, Y Tang, EC Holmes, T Allen, JG Prado, et al. HIV evolution: CTL escape mutation and reversion after transmission. *Nature medicine*, 10(3):282–289, 2004.

[107] Esteban A Hernandez-Vargas and Richard H Middleton. Modeling the three stages in HIV infection. *Journal of theoretical biology*, 320:33–40, 2013.

[108] Steven G Deeks, Christina MR Kitchen, Lea Liu, Hua Guo, Ron Gascon, Amy B Narváez, Peter Hunt, Jeffrey N Martin, James O Kahn, Jay Levy, et al. Immune activation set point during early HIV infection predicts subsequent CD4+ T-cell changes independent of viral load. *Blood*, 104(4):942–947, 2004.

[109] Victor Appay, Douglas F Nixon, Sean M Donahoe, Geraldine MA Gillespie, Tao Dong, Abigail King, Graham S Ogg, Hans ML Spiegel, Christopher Conlon, Celsa A Spina, et al. HIV-specific CD8+ T cells produce antiviral cytokines but are impaired in cytolytic function. *The Journal of experimental medicine*, 192(1):63–76, 2000.

[110] Anthony S Fauci. HIV and AIDS: 20 years of science. *Nature medicine*, 9(7):839–843, 2003.

[111] Gang Huang, Yasuhiro Takeuchi, and Andrei Korobeinikov. HIV evolution and progression of the infection to AIDS. *Journal of theoretical biology*, 307:149–159, 2012.

[112] Philip JR Goulder and Bruce D Walker. The great escape–AIDS viruses and immune control. *Nature medicine*, 5(11):1233–1235, 1999.

[113] Robin A. Weiss. How Does HIV Cause AIDS? *Science*, 260(5112):1273–1279, 1993. ISSN 0036-8075. URL http://www.jstor.org.proxy1.lib.uwo.ca/stable/2881758.

[114] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Helper T cells and lymphocyte activation. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[115] Erin EH Tran, Mario J Borgnia, Oleg Kuybeda, David M Schauder, Alberto Bartesaghi, Gabriel A Frank, Guillermo Sapiro, Jacqueline LS Milne, and Sriram Subramaniam. Structural mechanism of trimeric HIV-1 envelope glycoprotein activation. *PLoS Pathogens*, 8(7):e1002797, 2012.

[116] Xuegong Zhu, Christoph Borchers, Rachelle J Bienstock, and Kenneth B Tomer. Mass spectrometric characterization of the glycosylation pattern of HIV-gp120 expressed in CHO cells. *Biochemistry*, 39(37):11194–11204, 2000.

[117] Xiping Wei, Julie M Decker, Shuyi Wang, Huxiong Hui, John C Kappes, Xiaoyun Wu, Jesus F Salazar-Gonzalez, Maria G Salazar, J Michael Kilby, Michael S Saag, et al. Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307, 2003.

[118] Penny L Moore, Elin S Gray, C Kurt Wibmer, Jinal N Bhiman, Molati Nonyane, Daniel J Sheward, Tandile Hermanus, Shringkhala Bajimaya, Nancy L Tumba, Melissa-Rose Abrahams, et al. Evolution of an HIV glycan–dependent broadly neutralizing antibody epitope through immune escape. *Nature medicine*, 18(11):1688–1692, 2012.

[119] Antoine Chaillon, Martine Braibant, Thierry Moreau, Suzie Thenin, Alain Moreau, Brigitte Autran, and Francis Barin. The V1V2 domain and a N-linked glycosylation site in the V3 loop of the HIV-1 envelope glycoprotein modulate neutralization sensitivity to the human broadly neutralizing antibody 2G12. *Journal of virology*, 2011.

[120] Cynthia A Derdeyn, Julie M Decker, Frederic Bibollet-Ruche, John L Mokili, Mark Muldoon, Scott A Denham, Marintha L Heil, Francis Kasolo, Rosemary Musonda, Beatrice H Hahn, et al. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science*, 303(5666):2019–2022, 2004.

[121] Manish Sagar, Xueling Wu, Sandra Lee, and Julie Overbaugh. Human immunodeficiency virus type 1 V1-V2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. *Journal of virology*, 80(19):9586–9598, 2006.

[122] Bruno R Starcich, Beatrice H Hahn, George M Shaw, Paul D McNeely, Susanne Modrow, Hans Wolf, Elizabeth S Parks, Wade P Parks, Steven F Josephs, Robert C Gallo,

et al. Identification and characterization of conserved and variable regions in the envelope gene of HTLV-III/LAV, the retrovirus of AIDS. *Cell*, 45(5):637–648, 1986.

[123] W. H. Li, M. Tanimura, and P. M. Sharp. Rates and dates of divergence between AIDS virus nucleotide sequences. *Molecular Biology and Evolution*, 5(4):313–330, July 1988. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040503. URL https://academic.oup.com/mbe/article/5/4/313/1026948.

[124] Colin Anthony, Talita York, Valerie Bekker, David Matten, Philippe Selhorst, Roux-Cil Ferreria, Nigel J Garrett, Salim S Abdool Karim, Lynn Morris, Natasha T Wood, et al. Cooperation between strain-specific and broadly neutralizing responses limited viral escape and prolonged the exposure of the broadly neutralizing epitope. *Journal of virology*, 91(18):e00828–17, 2017.

[125] Rong Rong, Frederic Bibollet-Ruche, Joseph Mulenga, Susan Allen, Jerry L Blackwell, and Cynthia A Derdeyn. Role of V1V2 and other human immunodeficiency virus type 1 envelope domains in resistance to autologous neutralization during clade C infection. *Journal of virology*, 81(3):1350–1359, 2007.

[126] Vlad Novitsky, Rui Wang, Raabya Rossenkhan, Sikhulile Moyo, and M Essex. Intrahost evolutionary rates in HIV-1C env and gag during primary infection. *Infection, Genetics and Evolution*, 19:361–368, 2013.

Chapter 2

Indel Evolution Among Hosts

2.1 Background

My work included in this chapter is reproduced from the published journal article Palmer and Poon [1]. I included this article as a chapter because it contains work conducted during the first 6 months of my masters program and served as the foundation for the primary component of my thesis project in Chapter 3. To reiterate, I produced Table 2.1, Figures 2.4 and 2.5, and Supplementary Figures A1 and A2 during the first 6 months of my masters program, while remaining work was conducted during my BMSc program. Here, I will provide a short overview of crucial concepts and contextual information needed to understand this study.

2.1.1 Context

Foremost, substitution rates have been previously studied in gp120 and the variable regions [2–5], but indel rates have not. It is also widely accepted that indels rapidly accumulate in the gp120 variable loops, particularly V1, V2, V4, and V5, and are suggested to play a role in immune escape in these regions [6–10].

Importantly, this study uses HIV-1 genetic sequence data sampled among hosts in the human population, meaning that a single HIV-1 sequence was selected to represent each patient.

These sequences are often generated using Sanger sequencing methods and therefore, represent the consensus of a patient's highly-diverse HIV-1 population [11, 12]. These Sanger sequences only tend to contain those mutations that have come sufficiently close to reaching fixation (*i.e.* 100% prevalence) in the population and typically does not register lower frequency variants [11–13]. The process of reaching fixation typically involves enduring extensive purifying selection over countless viral generations, thereby making the sampled consensus sequence a highly refined and filtered representation of the HIV-1 population [14]. Given our use of among-host sequence data, this study is specifically measuring the rates at which indels are being fixed in the gp120 variable loops at the population level. It is not measuring the rates of indel accumulation in the virus population within individual hosts, which will be covered in Chapter 3.

Use of among host data also facilitated the collection of sequences worldwide that included several HIV-1 group M subtypes. Recall from Chapter 1 that the HIV-1 group M subtypes exhibit considerable genetic differences which can reach as high as 35% in *env* [15]. Given the substantial differences among HIV-1 group M, we wanted to further test the hypothesis that the HIV-1 subtypes exhibit difference in variable loop indel rates. The term "clade", which describes populations containing all the descendants that originate from a single ancestor, is used throughout this study to accurately describe both subtypes and CRFs together.

2.1.2 Pairwise Alignment

To extract the variable regions, I used a reference sequence describing HIV-1 subtype B called HXB2 (Genbank Acession: K03455), which has been annotated at every nucleotide position for relevant genetic regions, protein structures, and various other characteristics [16]. The process of aligning each gp120 sequence with this reference sequence adds necessary gaps to both sequences to bring their matching features into the same positions, including the variable regions. Therefore, the positions of the variable regions in the HXB2 reference sequence allow me to locate the corresponding variable region locations in the patient-derived sequence.

2.1.3 Phylogenetic Inference

In this study, I used phylogenetic inference methods to estimate the rates of evolution, estimate the time to the most recent common ancestor (tMRCA), and reconstruct trees on a per-subtype basis. I will briefly outline some key concepts involved in this process.

First, phylogenetic trees are initially generated in an unrooted format, which only describe the relatedness of different lineages — individual lines of descent from ancestors — and do not describe the direction of evolution from a common ancestor [17]. Rooting a phylogenetic tree involves inferring the root of the tree — the theoretical common ancestor of all sampled tree tips — which, in turn, also enables reconstruction of evolutionary history relating tree lineages originating from this point [17]. The methods I used to root my phylogenetic trees in this study rely on the assumption of a molecular clock in the data: a hypothesis that biological sequences (*i.e.* DNA or amino acids) evolve or accumulate mutations at a constant rate over time which enables the reconstruction and analysis of evolutionary history [18]. One of the processes I used is called root-to-tip regression, which is a linear model fit between the sampling dates and genetic differences of tip sequences relative to the root [19]. When assuming that only a single underlying substitution rate exists, this hypothesis is more specifically referred to as the strict molecular clock [19]. The goodness of fit of the root-to-tip regression indicates how well the tree adheres to the strict molecular clock hypothesis, and therefore can be used to search for the most appropriate root position [19].

The second method I used to root phylogenetic trees in this study is called least-squares-dating (LSD) [20]. This method utilizes a relaxed molecular clock, which allows the substitution rate to change between branches within a phylogenetic tree based on an underlying rate distribution [20]. It applies a minimization algorithm to the square errors of branch lengths to estimate the substitution rate and further approximates an underlying distribution from which the rate of each tree branch is drawn [20]. Relaxed molecular clock approaches are preferred for rapidly evolving pathogens due to the considerable heterogeneity in evolution between different lineages [21].

I then use the root-to-tip strict clock method and the LSD relaxed clock to estimate the tMRCA of each HIV-1 group M clade. This is done by determining the mean substitution rate and extrapolating back in time to estimate the date of the tree root, or MRCA.

2.1.4 Indel Rate Estimation

In our phylogenetic trees, we compared the lengths of variable loop sequences at every pair of cherries: pairs of sequences related by a common ancestor with no intervening nodes in between. We estimated indel rates in our phylogenetic trees by fitting a customized statistical model to our data and solving for the parameter values that best fit the data using maximum likelihood methods. Indels were detected as true/false outcomes, and therefore warranted a unique approach to determine the underlying indel rate. The model used in this study incorporates a true/false outcome into a Poisson likelihood function, which produces the probability of seeing count data given a particular rate and length of time. In the equation below, Y_i either takes the value of 0 or 1 based on whether the variable loop lengths in cherries were the same or different, respectively. We used the presence of length differences to detect indel events over short time periods in cherries. The time data of the associated cherry sequence pair is provided as t_i, and used to solve for the single underlying indel rate λ in each subtype's phylogenetic tree.

The indel rate was estimated from these data by fitting the following model using maximum likelihood, where the Bernoulli likelihood for the i-th cherry is:

$$L(Y_i|\lambda, t_i) = (1 - Y_i)e^{-\lambda t_i} + Y_i(1 - e^{-\lambda t_i})$$

where $Y_i = 1$ if the sequence lengths differ (implying one or more indels), and is 0 otherwise; t_i is the total branch length, λ is the overall indel rate, and $e^{-\lambda t_i}$ is the Poisson probability of no indels in the cherry. The total log-likelihood across cherries is thus:

$$\log L = \sum_i \log L(Y_i|\lambda, t_i)$$

Phylogenetic measures of indel rate variation among the HIV-1 group M subtypes [1]

2.2 Introduction

Human immunodeficiency virus type 1 (HIV-1) is a rapidly evolving retrovirus with enormous genetic diversity that is divided into four groups (M, N, O and P). The global HIV-1 pandemic that affects approximately 37 million people as of 2017 [22] is largely caused by group M, which is further partitioned into nine subtypes (A-D, F-H, J, K) that can differ by roughly 30% of their genome sequence and have distinct geographic distributions due to historical founder effects [15, 23]. In addition, there are a large number of circulating recombinant forms (CRFs) that are the result of recombination among two or more HIV-1 subtypes that have subsequently become established in particular regions at high prevalence. There is accumulating empirical evidence of significant variation among specific HIV-1 subtypes with respect to pathogenesis, *e.g.*, rates of disease progression, and the evolution of drug resistance, which implies that the HIV-1 subtypes and CRFs are clinically significant [24–27].

In the host cell-derived lipid membrane of every HIV-1 particle, there are numerous virus-encoded envelope glycoprotein complexes composed of three gp41 transmembrane units and three gp120 surface units [28]. The HIV-1 gp120 glycoprotein is a potent surface-exposed antigen that plays a significant role in the recognition and binding of target cell receptors [29]. One reason for the difficulty in immunologically targeting this glycoprotein is the abundance of N-linked glycoslyation sites: sequence motifs that encode the post-translational linkage of glycan groups to asparagine residues [30]. In addition, the HIV-1 gene encoding gp120 has a particularly high rate of evolution, especially within the five hypervariable regions that encode surface-exposed, disordered loop structures. These five variable regions (numbered V1-V5)

[1] Contents in this chapter have been reproduced from: Palmer J, Poon AFY. Virus Evol. 2019 Jul 21;5(2):vez022. doi: 10.1093/ve/vez022. The study was published under the terms of the Creative Commons Attribution Non-Commercial License, which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

can tolerate substantially higher amino acid substitution rates than the rest of the HIV-1 genome [3]. Both the extensive glycosylation and rapid substitution rates in HIV-1 gp120 facilitate the escape of the virus from neutralizing antibodies [8].

There are multiple mechanisms by which mutations arise within the HIV-1 genome including nucleotide substitutions, insertions, and deletions [31]. While substitution rates have been extensively characterized in HIV-1 and specifically in the *env* gene [32, 33], less attention has been given to sequence insertions and deletions (indels). The few studies that examine indels in the HIV-1 genome have focused on the location, behaviour, and clinical significance of specifically recurring indels, such as indels in HIV-1 *pol* associated with drug resistance and indels in *gag* and *vif* associated with disease progression and infectivity [34–36]. Only a small number of comparative studies have examined indel rates in the HIV-1 *env* gene encoding gp120 and gp41. Wood et al. [8], for one, found that indels preferentially accumulate in the variable loops of gp120 compared to the remainder of this sequence, while other studies have suggested that variable loop indels correspond with HIV-1 transmission and modulate coreceptor switching [37, 38].

Despite the significant impact of indels within HIV-1 gp120 on virus transmission and adaptation, the overall rates of indel evolution in gp120 have not yet been measured through a comparative analysis. Furthermore, as previous studies on indels in HIV-1 have tended to focus on defined study populations, we have not found any study that has examined indel rates in a large database covering multiple HIV-1 subtypes and geographical regions. Here we present results from a dated-tip phylogenetic analysis of HIV-1 *env* sequences from a public database. By comparing sequences from different hosts, our analysis focuses on fixed indel differences that are tolerated by the virus; for example, this analysis implicitly excludes indels that induce frameshifts in *env*. Our novel phylogenetic method specifically analyzes ‘cherries’ [39] — pairs of sequences directly descended from a common ancestor with no intervening ancestral nodes — in time-scaled phylogenies to estimate the rates of indel evolution in the gp120 variable loops of seven HIV-1 group M subtypes and CRFs (herein referred to collectively as clades).

We focused on cherries to reduce the exposure of indels to purifying selection, and also the probability of multiple indel events occurring in the same variable region, as the divergence time in a cherry tends to be shorter on average than a random selection of two tips. Using this method, we evaluate the hypothesis that the mean rates of indels significantly vary among the gp120 variable loops and group M clades. Further, we examine the nucleotide composition of indels to assess how this characteristic might be shaped by the virus genome, and quantify the impact of indels on N-linked glycosylation sites in HIV-1 gp120.

2.3 Methods

2.3.1 Data processing

We queried the Los Alamos National Laboratory (LANL) HIV Sequence Database (http://www.hiv.lanl.gov/) for all sequence records covering HIV-1 *env* gp120, limiting the records to one sequence per patient. The 26,359 matching sequences were downloaded with predicted subtype, collection year and GenBank accession number. We parsed the resulting FASTA file and removed sequences that lacked subtype or collection year fields, or were shorter than 1400 nt (roughly 90% of full-length HIV-1 gp120), yielding a final data set of 6605 sequences. To extract the interval encoding gp120 from each sequence and partition the result into the variable and conserved regions, we performed pairwise alignments using an implementation of the Altschul-Erickson [40] modification of the Gotoh algorithm in Python (http://github.com/ArtPoon/gotoh2). Each nucleotide sequence was aligned against the HXB2 (Genbank accession number K03455) gp120 reference sequence with match/mismatch scores of +5/−4, gap open/extension penalties of 30 and 10, respectively, and no terminal gap penalty. The aligned query sequence was cut at the boundaries of the aligned HXB2 reference gene to extract the patient-derived subsequence homologous to gp120. Next, we removed any gaps in this result and then aligned the amino acid translation to the gp120 protein reference sequence using an empirical HIV amino acid scoring matrix (25% divergence [41]) with the same gap penalties, except that terminal gaps were penalized at this stage. Finally, we used the aligned query to insert gap character triplets into the preceding nucleotide sequence as 'in-frame' codon deletions.

Using the HXB2 reference annotations, we extracted the five variable (V1-V5) and five conserved (C1-C5) regions of gp120. The conserved region sequences were concatenated and exported to separate files for phylogenetic reconstruction. We subsequently determined that our method was not reliably extracting the V5 regions, based on the overabundance of multiple gap characters at the 5' end of many outputs. To avoid further problems downstream, we implemented a modified extraction method specific to V5. We first extracted nine extra nucleotides beyond the 5' boundary of the V5 reference to provide conserved sequence coverage outside this hypervariable region. The extended V5 sequence was then translated and aligned to a V5 amino acid reference sequence of matching length as above. Lastly, we used the first non-gap character (a matched amino acid) immediately following the first three conserved residues in the amino acid alignment as the adjusted V5 start position, thereby omitting any gap characters that preceded this first residue.

2.3.2 Phylogenetic analysis

We used the program MAFFT (version 7.271) with the default settings [42] to generate a multiple sequence alignment (MSA) from the concatenated sequences of conserved regions for each subtype. On manual inspection of the resulting MSAs, we found some alignment columns comprised mostly of gaps caused by rare insertions, so we removed all columns with gap characters in more than 95% of sequences. Next, we reconstructed phylogenies for each subtype-specific MSA by approximate maximum likelihood using FastTree2 (version 2.1.8) compiled with double precision [43]. The resulting trees were manually screened for unusually long terminal branches indicative of problematic sequences, which we removed from the corresponding MSA before reconstructing a revised tree.

Effective estimation of indel rates required that all phylogenetic trees be scaled in time.

To rescale the maximum likelihood trees, sequence accession numbers were used to query the GenBank database for more precise collection dates containing month and day fields; otherwise we retained the collection years from the LANL database. The R package *ape* was then used to change each tree into a strictly bifurcating structure and to root the tree using root-to-tip regression [19] based on the associated tip dates. We evaluated the correlation between the time since the inferred root date (x-intercept) and the total branch length (in expected numbers of substitutions) to determine if the data were consistent with a molecular clock [44].

Using the same dates, we employed the least-squares dating (LSD) program [20] to adjust node heights and rescale the tree in time under a relaxed molecular clock model. Dates lacking either month or day fields were specified as bounded intervals. The time-scaled tree outputs from LSD were imported into R to extract the 'cherries': pairs of sequences directly descended from a common ancestor with no intervening ancestral nodes. Focusing on sequences in cherries provides phylogenetically independent observations and minimizes the divergence times, thereby reducing the chance of encountering multiple indel events as well as the effect of purifying selection on indels. Cherries with time-scaled branch lengths totaling zero years were removed from our analysis as they did not provide meaningful indel observations and caused problems for rate estimation.

2.3.3 Indel rate estimation

To estimate the rate of indels in the variable loops, homologous variable regions from each pair of sequences in a cherry were compared for length differences. The presence of a length difference was reported as a binomial outcome implying that an indel event had occurred along these branches; this approach does not account for the possibility of multiple indels causing reversion to the same sequence lengths. Additionally, the total branch lengths comprising the cherry was employed as an estimate of divergence time in years (Supplementary Figure A2). The indel rate was estimated from these data by fitting the following model using maximum likelihood, where the Bernoulli likelihood for the *i*-th cherry is:

$$L(Y_i|\lambda, t_i) = (1 - Y_i)e^{-\lambda t_i} + Y_i(1 - e^{-\lambda t_i})$$

where $Y_i = 1$ if the sequence lengths differ (implying one or more indels), and is 0 otherwise; t_i is the total branch length, λ is the overall indel rate, and $e^{-\lambda t_i}$ is the Poisson probability of no indels in the cherry. The total log-likelihood across cherries is thus:

$$\log L = \sum_i \log L(Y_i|\lambda, t_i)$$

We used the Brent minimization method implemented in the R package 'bbmle' to obtain a maximum likelihood estimate of λ for each clade and variable loop combination. A generalized linear model (GLM) with a logit link function was also applied to these data to evaluate statistical associations of the inferred distribution of indels on clades and variable loops; the model incorporated a divergence time term as a rudimentary adjustment for variation in 'sampling effort'.

2.3.4 Analysis of indels

For every combination of five variable loops and seven clades, we categorized the inferred lengths of indels into three discrete classes: single-codon (3 nt), double-codon (6 nt), and long (9+ nt). Pearson χ^2 residuals were calculated on these distributions to determine if, and in what direction, these observed proportions significantly deviated from their expected values. To further analyze the composition of indels, we generated pairwise alignments for each cherry with discordant sequence lengths to identify and extract indels. From these pairwise alignments, we calculated the proportions of adenine, thymine, guanine, and cytosine (A,C,G,T) nucleotides in the indel and non-indel regions of the gp120 variable loops. In addition, we recorded the positions and numbers of PNGSs in the five gp120 variable loops by scanning the unaligned amino acid sequences with the regular expression 'N[ˆP][ST][ˆP]', where 'ˆP' maps to any symbol except P (proline). We then used these data to investigate how commonly indels tended to change PNGSs in the variable loops. By combining PNGS and indel location data, we searched for instances where an indel overlapped with a PNGS in one of the two sequences of a cherry, indicating either the deletion of a PNGS or the insertion of a sequence containing one. To avoid recording instances of partial indel overlap that leave the PNGS intact, we verified the PNGS was disrupted by scanning it again with a regular expression.

2.4 Results

We collected HIV-1 sequences covering gp120 from the Los Alamos National Laboratory (LANL) HIV Sequence Database (http://www.hiv.lanl.gov/) and filtered these data (as described in Methods) to obtain a final data set of 6,605 sequences. To estimate the rates of indel evolution for different HIV-1 subtypes and circulating recombinant forms (CRFs) in this data set, we reconstructed phylogenies for each of the seven group M clades using maximum likelihood, and then rooted and rescaled each tree based on the sample collection dates under a molecular clock model. Initially we employed a strict clock model in root-to-tip regressions (Figure 2.1) to assess whether the data sets contained sufficient signal to estimate rates of evolution (Table 2.1). Specifically, we confirmed that the lower bounds of the 95% confidence intervals of rate estimates exceeded zero for all clades, which implied a gradual and measurable accumulation of mutations over the sampling time frame. Further, we assessed the model fit with the coefficient of determination (R^2), which was greatest for 01_AE and F1, and lowest for subtype C (Table 2.1). Next, we employed a more robust least-squares dating method [20] to rescale the trees in time. Table 2.1 summarizes the substantial differences between the strict clock and least-squares estimates of the times to the most recent common ancestor (tMRCA) for each clade.

We extracted pairs of cherries from these rescaled trees as phylogenetically-independent observations on relatively short time frames. Next, we used these cherries to estimate the mean

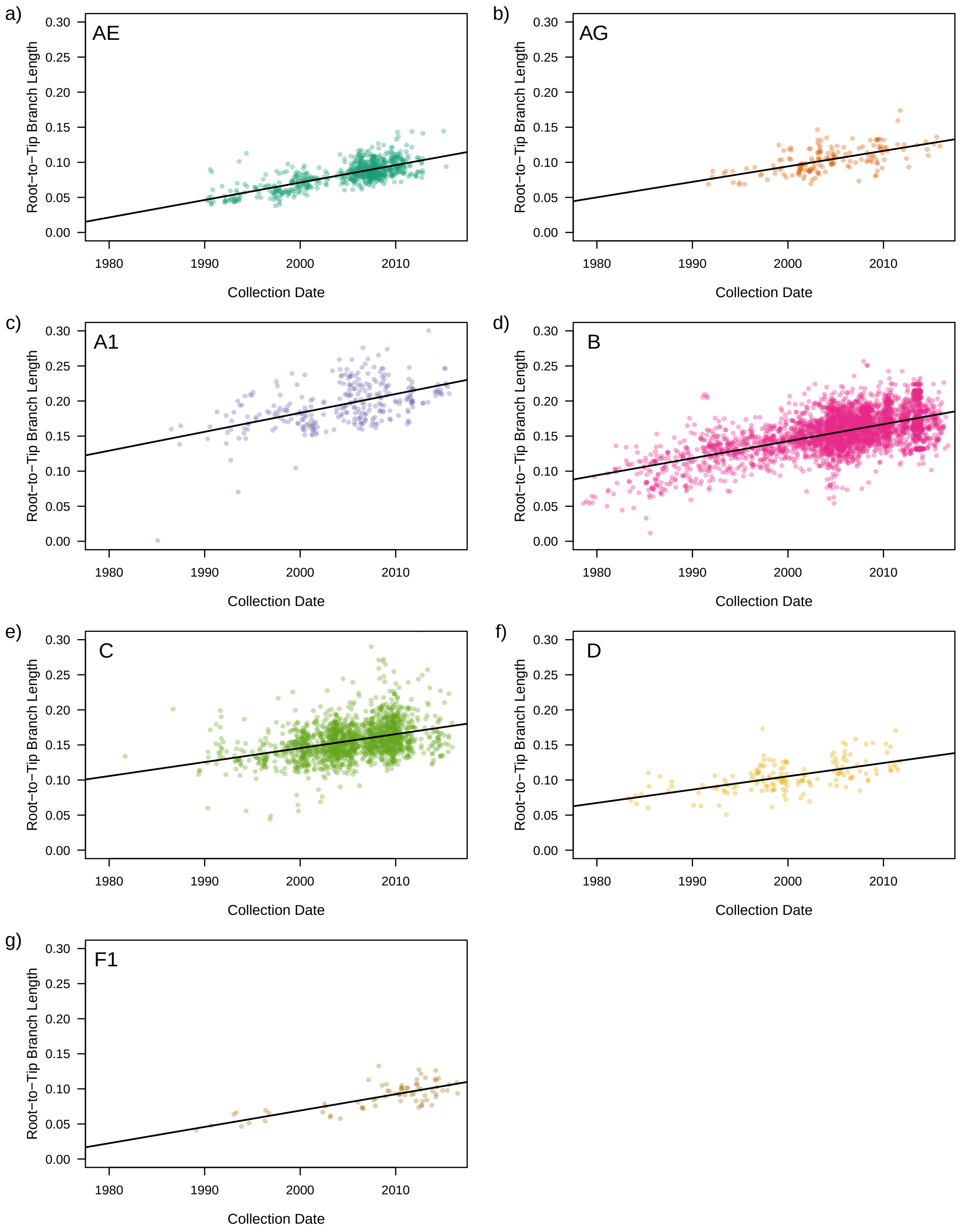

Figure 2.1: The relationship between sequence root-to-tip branch lengths and sequence collection dates in seven clade-wise phylogenetic trees reconstructed from gp120 conserved region (C1-C5) alignments. Each panel is labelled by the clade that its tree represents. All plot axes have been adjusted to the same scales for comparison. Regions of greater color density indicate the clustering of multiple plotted points. The solid line on each plot describes the linear regression of branch lengths on collection dates.

Root-to-tip				LSD	
Clade	Rate ×10^{-3}	tMRCA	R^2	Rate ×10^{-3}	tMRCA
01_AE	2.49 (2.31, 2.67)	1971.4	0.51	1.87 (1.85, 2.10)	1968.6 (1965.5, 1974.6)
02_AG	2.21 (1.75, 2.67)	1957.4	0.35	2.27 (2.10, 2.68)	1961.9 (1957.5, 1969.0)
A1	2.69 (2.17, 3.21)	1932.0	0.26	2.45 (2.32, 2.66)	1966.3 (1964.0, 1969.5)
B	2.43 (2.32, 2.54)	1941.2	0.34	1.47 (1.46, 1.57)	1951.7 (1951.2, 1954.8)
C	1.99 (1.75, 2.23)	1926.9	0.13	1.80 (1.78, 1.96)	1939.8 (1937.5, 1946.7)
D	1.90 (1.47, 2.32)	1944.5	0.33	1.88 (1.68, 2.11)	1957.8 (1952.7, 1962.9)
F1	2.33 (1.85, 2.81)	1970.4	0.57	1.67 (1.34, 2.03)	1956.2 (1943.9, 1965.2)

Table 2.1: Summary of the evolutionary rate estimates, times to most recent common ancestor (tMRCAs), and R^2 values generated by applying root-to-tip and least-squares dating models to our seven clade-specific trees. The 95% confidence intervals for the evolutionary rates of both models and for the tMRCAs estimates of the LSD model are enclosed in brackets. Both models are shown to illustrate the differences between fitting strict (root-to-tip) and relaxed clock models to our sequence data.

indel rates for each variable loop using a binomial-Poisson model, where the probability of detecting an indel event in a cherry increased exponentially with the divergence time. The indel rate estimates across the five variable loops and seven HIV-1 clades in this study ranged between 3.0×10^{-5} to 1.5×10^{-3} indels/nt/year (Figure 2.2). We could not obtain an indel rate estimate for V3 in F1 due to low sample size for this sub-subtype, such that no cherries had discordant sequence lengths in V3. Similarly, we observed wide confidence intervals for the rate estimates for indels within V1 in 02_AG and F1, and for V5 in F1. The frequency of indels was significantly lower in subtype B than the reference clade, 01_AE (binomial GLM, $p < 2 \times 10^{-16}$; Supplementary Table A1). In addition, indels were significantly less frequent in V3 irrespective of clade relative to V1. Estimated interaction effects in the model also indicated that indels were significantly less frequent than expected in V2 within clades B and C.

Under the assumption that differences in sequence lengths of variable loops was caused by a single fixed indel (*i.e.*, no multiple hits), we examined the distribution of indel lengths among variable loops and clades. Cherries with putative indels in the HIV-1 subtype C phylogeny tended to contain significantly longer indels than expected (Figure 2.3). Conversely, the variable loops V1, V2 and V4 tended to contain longer indels than expected irrespective of clade,

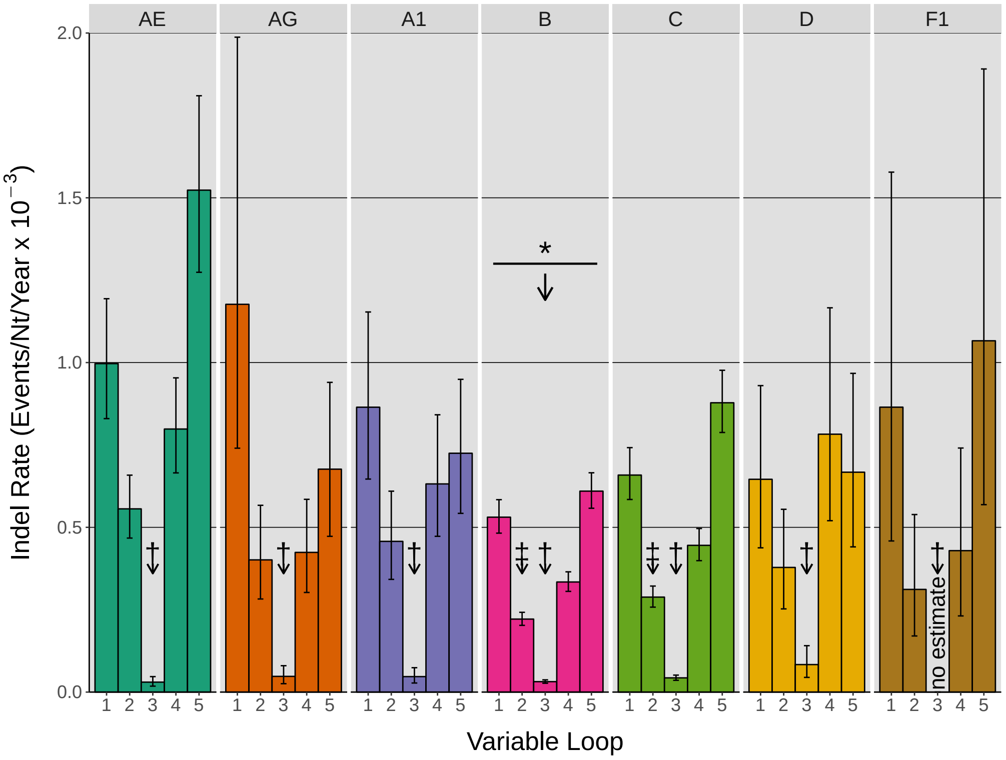

Figure 2.2: Indel rate estimates in the five gp120 variable loops of seven HIV-1 group M clades. Each group of five colored bars describes the indel rates of V1-V5 for one of the seven examined clades. Maximum likelihood estimation was applied to cherry indel outcomes using a binomial-Poisson model to determine the above indel rates. Error bars represent the 95% confidence intervals within which indel rates were estimated. Arrows labeled with a * symbol indicate the presence and direction of significant differences among the mean indel rates of group M clades, relative to the CRF 01_AE reference. Arrows labeled by a † symbol denote significant differences among the variable loops irrespective of clade, relative to V1. Arrows labeled by a ‡ symbol denote individual interactions between variable loop and clade which are significantly different than their predicted value. No meaningful rate estimate was provided for V3 of clade F1 because no indels were detected in this data set.

whereas V3 and V5 tended to contain shorter indels.

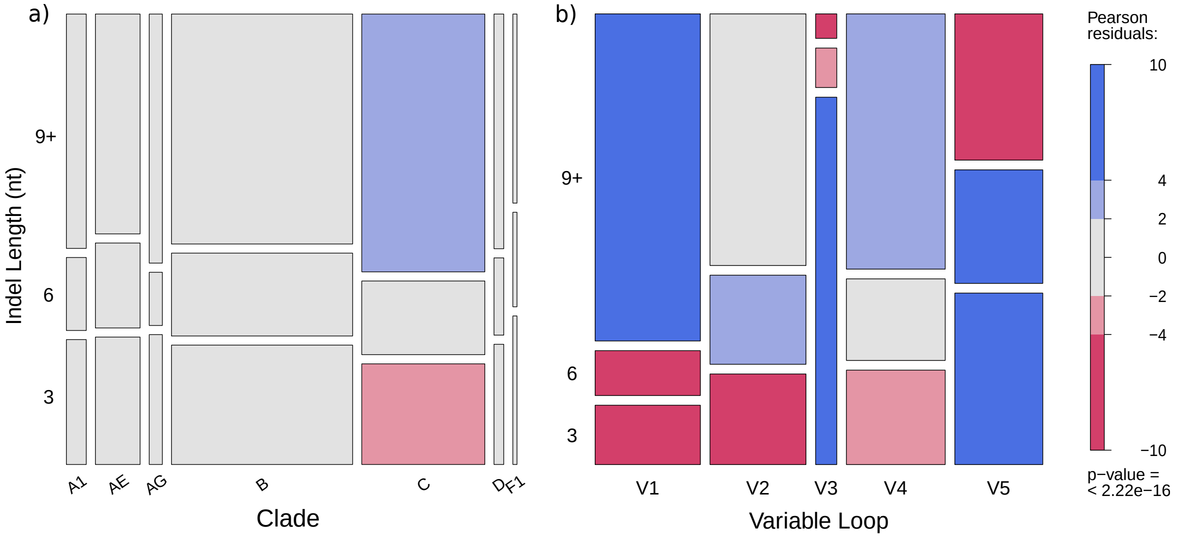

Figure 2.3: The distribution of indel lengths within (a) the seven group M clades and (b) the five gp120 variable loops . Indel lengths, measured in nucleotides, were classified into three categories: 3 nt, 6 nt, and 9 nt or longer. Box heights indicate the proportion of indels belonging to the given length category, while box widths indicate the proportion of indels belonging to (a) each clade or (b) each variable loop . Pearson χ^2 residuals — quantified measures of the difference between observed and expected values — were calculated for every group on these plots to determine if, and in what direction, these proportions significantly deviated from the χ^2 value. Pearson residuals are comparable to the number of standard deviations away from the χ^2 value, meaning that values greater than 2, and especially those greater than 4, describe groups whose proportions significantly deviate from the predicted outcome. Blue shading indicates higher indel counts than expected, while red indicates lower counts.

Next, we examined the frequencies of nucleotides in indel- and non-indel regions of sequences in cherries with putative indels (Figure 2.4). Because these frequencies measured for different clades tended to cluster by variable loop, we treated the clades as rudimentary replicates for this comparison (notwithstanding sample variation associated with variable loop V3 and subtype F1, for example). Overall, we observed that indels tended to contain higher proportions of G and lower proportions of T than the corresponding non-indel regions.

Figure A1 summarizes the numbers of potential N-linked glycosylation sites (PNGS) detected in each variable loop across the clades in our study. The mean counts in loops V1 to V5 were 2.4, 2.1, 0.9, 4.1 and 1.3 PNGSs, respectively. We found significant differences in PNGS

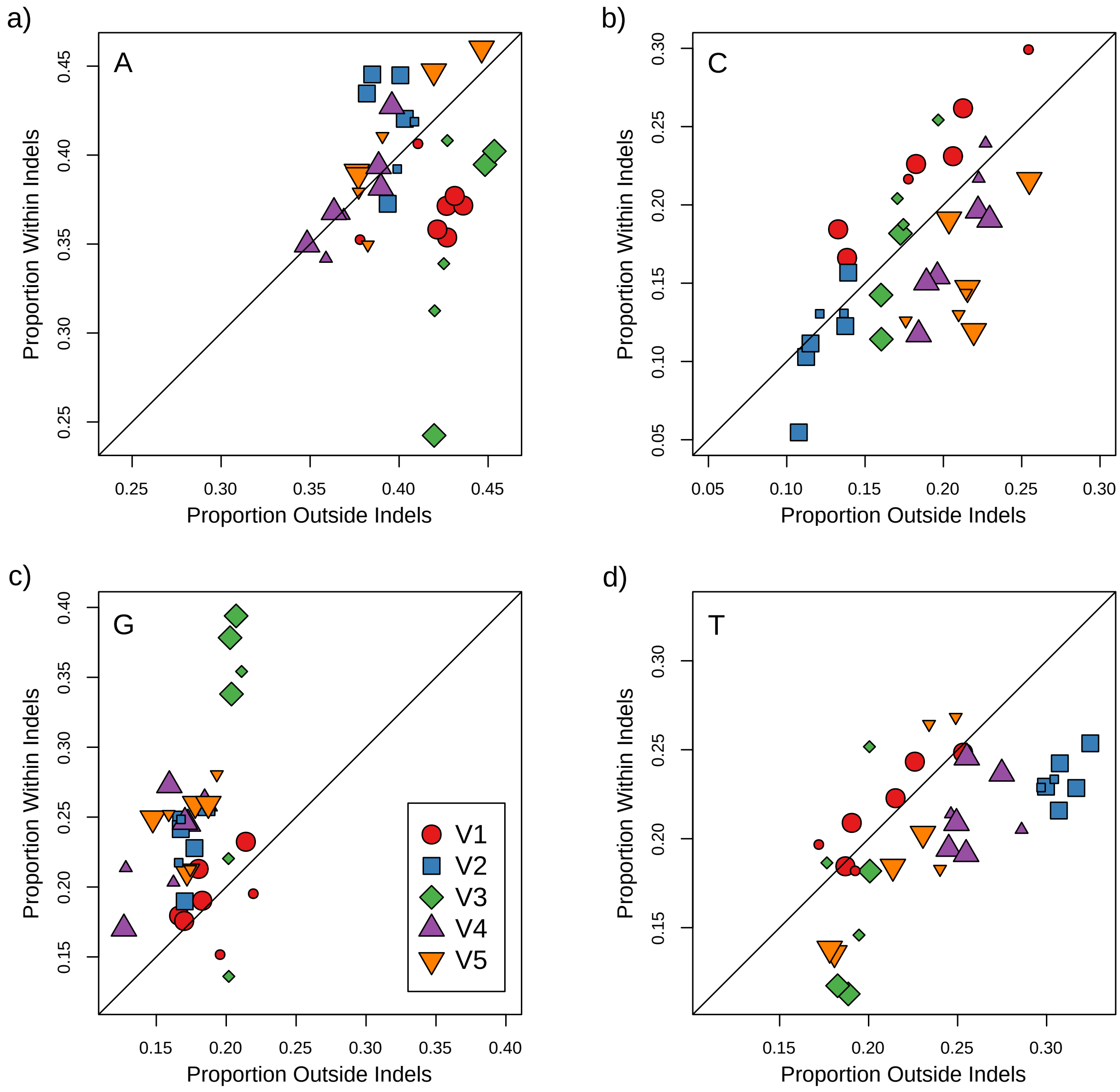

Figure 2.4: Nucleotide proportions in indel sequences relative to flanking non-indel sequences for all examined variable loops and subtypes. Plots (a-d) illustrate these relations in adenine, cytosine, guanine, and thymine nucleotides, respectively. Each group denoted by a colored shape represents one of the five variable loops of gp120 and contains seven data points corresponding to each of the examined group M clades. The plotted line with a slope of 1 (y=x) represents the null result in which sequences inside and outside of indels show no difference in their nucleotide proportions. Plotted points that deviate from this line indicate differences between nucleotide proportions found in indels compared to those found outside indels. Larger data points indicate a significant χ^2 test result testing for a difference between indel and non-indel counts in that particular data set.

counts among variable loops (likelihood ratio test, $p = 2.8^{-13}$) and among clades ($p < 10^{-15}$). For variable loop V1, 01_AE contained significantly more PNGS (mean 2.93 PNGS) than the other clades; the next highest count was obtained for subtype C (2.43 [95% C.I. 2.31, 2.57]). Subtypes B and C had significantly higher numbers of PNGS within V3 on average (0.96 [0.87, 1.05] and 0.95 [0.96, 1.05], respectively) than the reference clade 01_AE (mean 0.81). We observed substantial variation in the numbers of PNGS among clades in variable loop V4. For instance, clades 02_AG, A1 and B had significantly higher numbers, and subtype F1 significantly lower, than the reference clade 01_AE (mean 3.72). Finally, we mapped indels to PNGS in the variable loop sequences to determine how frequently indels were associated with the addition or removal of a PNGS ('disruption', Figure 2.5). V1, V2, and V4 contained the highest proportions of indel-induced PNGS disruption among the five variable loops. Again, we observed that estimates for different clades visibly clustered by variable loop. When we adjusted for the relative proportions of the variable loops occupied by PNGS, only V1 and V2 markedly departed from this expectation.

2.5 Discussion

To our knowledge, these results represent the first comprehensive measurement of indel rates in variable regions of HIV-1 gp120 across major virus subtypes and circulating recombinant forms (CRFs). Surprisingly, one of the only estimates of HIV-1 indel rates we have found dated back to 1995 [2], where Mansky and Temin used an *in vitro* assay of genetic mutations in HIV-1 reverse transcriptase (RT) and reported the observed counts of both nucleotide substitutions and indels. In contrast, our comparative study measures indel rates among different hosts, and as a result will inevitably underestimate these rates due to purifying selection on indels. We chose to focus on cherry sequences derived from between-host data as it provided phylogenetically-independent observations of indel evolution, while reducing the probability of multiple hits and the effects of purifying selection. While the comparison of within-host HIV-1 sequences would

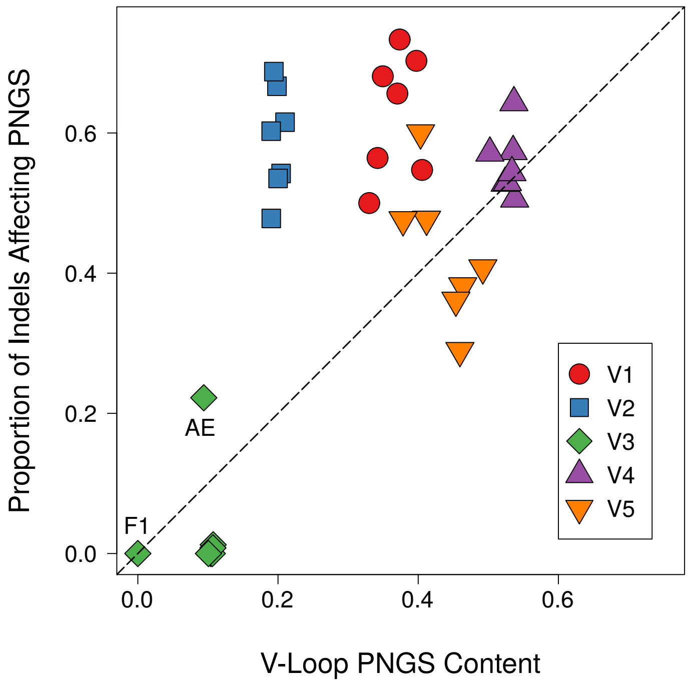

Figure 2.5: The proportion of indels in a variable region that knocked out at least one PNGS, relative to the PNGS content of the given variable loop. Data sets were represented by a colored shape to denote their variable loop and contained seven points derived from the group M clades. The dotted line of slope 1 provides a rough representation of the general trend expected if PNGSs were under no selection. Individual points that did not cluster with their variable loop were labeled with their clade. Specifically, V3 of subtype F1 did not have any records of indel events.

provide even shorter time scales and thereby more accurate measures of indel rates before selection, some HIV-1 subtypes and CRFs remain underrepresented in publicly available, large and longitudinal same-patient sequence data sets. In addition, results from Wood et al. [8] suggest that purifying selection against indels is relaxed in the variable regions of HIV-1 gp120.

To estimate indel rates, we needed to accurately rescale the HIV-1 phylogenies in chronological time. Estimates of the tMRCA can vary by genomic region, and estimates from regions within the HIV-1 *gag* and *pol* genes tend to be more recent than regions in *env* irrespective of subtype [45]. Overall, we determined that the diversity of HIV-1 sequences and sample collection dates were sufficient to fit a strict molecular clock model (Table 1). We note that for the purpose of rescaling the trees after this initial assessment with a strict clock, we employed an implementation of a relaxed clock model that allows for rate variation over time. However, we also observed that the goodness-of-fit used to assess support for the clock model was the lowest for subtype C. We attribute this poor model fit to both the relatively old age of subtype C [46] and the relative lack of HIV-1 C samples collected prior to 1995 (Figure 2.1). Estimates of the times to the most recent common ancestor (tMRCA) from the relaxed clock model implemented in the LSD program were generally comparable to previous estimates in the literature for the corresponding HIV-1 clades [46, 47] except for subtype A1, for which we obtained a more recent range of estimates (1964 to 1970). For instance, Tongo et al. [48] recently estimated that (sub-)subtype A1 originated around 1946 – 1957 from an analysis of full-length genome sequence data. We note that because our estimate relies on the ‘point estimate’ of the phylogeny reconstructed by maximum likelihood, the confidence intervals reported for our tMRCA estimates underestimate the true level of uncertainty and fixing the tree may skew the mean estimate. A Bayesian method would more accurately capture this substantial source of uncertainty, but would also be restricted to substantially reduced numbers of sequences due to the complexity of sampling from tree space. Furthermore, Wertheim et al. [46] postulated that estimates of tMRCA among studies may be inconsistent due to the use of nucleotide substitution models that are an inadequate approximation of past molecular evolution.

Our estimates of region- and subtype-specific indel rates ranged from 3.0×10^{-5} to 1.5×10^{-3} indels/nt/year (Figure 2.2). As expected, our average estimate (5×10^{-4} indels/nt/year) was considerably lower than the rate inferred from Mansky and Temin's *in vitro* experiments (about 1.5×10^{-3} indels/nt/year) [2] , where we used parameter estimates from Perelson and Nelson [49] to convert the observed numbers of indel counts to a rate. Since our study compares HIV-1 sequences isolated from different hosts, the indels have been filtered by purifying selection so that only a subset become fixed within the respective hosts. We found that indel rates in subtype B gp120 were significantly lower than the reference clade 01_AE, and generally lower than the other clades in our study. The significantly lower indel rate estimates in V3 irrespective of HIV-1 clade (Figure 2.2) were consistent with the functional importance of this variable loop. As V3 contributes to HIV entry by binding to the CCR5 or CXCR4 coreceptors, there is substantial purifying selection to conserve its overall structure [50, 51]. This lower tolerance for mutational change, relative to other variable loops of gp120, is consistent with reduced numbers of fixed indels among hosts. The lower indel rates might also be attributed in part to compensatory mutations to preserve structural interactions in V3 [52]. For example, an arginine insertion at position 11 of V3 confers CXCR4 tropism tends to be accompanied by a single amino acid deletion near the C-terminal of V3 [38].

The tendency of HIV-1 subtype C to accumulate longer indels in our analysis is consistent with results previously reported by Derdeyn et al. [37]. By examining HIV-1 heterosexual donor-recipient pairs, Derdeyn et al. [37] first determined that subtype C viruses initially contained shorter V1-V4 sequences upon transmission, which then substantially lengthened by up to 25 amino acids after progressing to chronic late-stage infection. A follow-up study by Chohan et al. [53] provided evidence that this trend was subtype-specific, as it was observed in infections by subtypes A and C, but not subtype B. Similarly, the observed preference for longer indels in V1 and V2 in our data is consistent with the role of these variable loops in facilitating immune evasion. For example, the insertion of five or more amino acids into V1/V2 is associated with reduced sensitivity of HIV-1 gp120 to neutralizing antibodies [54, 55], which is more efficiently achieved by a single long insertion than a series of short insertions. Conversely, the tendency for shorter insertions to accumulate in V3 is consistent with the existence of functional and structural constraints as noted above.

The nucleotide composition of the HIV-1 genome is generally skewed to higher frequencies of A (adenine), in large part due to G-to-A hypermutation induced by host factors [56]. We have not found previous studies that have compared the nucleotide composition of indels to the flanking sequence in the HIV-1 genome. Overall, we observed that indel sequences tended to comprise higher frequencies of G and lower frequencies of T relative to the rest of the variable loop sequence. We note that some frequency estimates had greater sample variation due to limited numbers of indels and sequences in association with V3 and subtypes D and F, for instance. Because the *env* gene generally contains slightly higher proportions of A (40%) and lower proportions of G (18%) than the rest of the HIV-1 genome (35% and 24%, respectively), we propose that this outcome might reflect the derivation of insertions into *env* from outlying sequence. Since we have not individually resolved these numerous indels into insertions or deletions through ancestral reconstruction, however, we cannot determine whether this pattern reflects a tendency for sequence insertions to be G-rich, or whether G-rich sequences are specifically targeted for deletion.

The variable regions of HIV-1 gp120 contribute disproportionately to the mean number of potential N-linked glycosylation sites (11 out of 25), which make up the glycan shield of gp120 [57]. Overall, we found that the numbers of PNGS for each variable loop was fairly consistent across clades, with some significant but minor differences in means (Figure A1). Furthermore, we observed that PNGS were more frequently affected by indels in variable loops V1 and V2 irrespective of clade (Figure 2.5). Put another way, the proportion of indels affecting PNGS in these variable regions was substantially greater than the proportion of the region encoding PNGS. This outcome supports the hypothesis of diversifying selection for the addition or removal of PNGS in V1/V2, where glycosylation plays a major role in mediating immune escape [55] and transmission fitness [37] at different stages in the natural history of HIV-1 infection.

Our estimates of indel rates in the variable regions of HIV-1 gp120 imply that fixed differences in variable loop lengths accumulate between infections on a time scale of about 10-20 years per variable loop with the exception of V3, which accumulates these differences an order of magnitude slower. This time frame is consistent with past observations that HIV-1 gradually 'raises' the glycan shield with insertions in V1/V2 several years post-infection [55]. The accumulation and composition of indels among infections is clearly heterogeneous among HIV-1 clades and variable loops. Some of the more exploratory results in this study, *e.g.*, differences in nucleotide frequencies within indels, are particularly novel and suggest new areas for further research in the molecular evolution of HIV-1 to identify the biological or selective determinants of sequence insertions and deletions.

2.6 Acknowledgments

This study was made possible by the HIV-1 sequence database curated by the Los Alamos National Laboratory. This work was supported in part by the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGI-131), and by grants from the Canadian Institutes of Health Research (CIHR PJT-153391, PJT-155990, and PJT-156178) and the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN-2018-05516).

2.7 Data Availability

Sequence data and scripts used for this study are available at https://github.com/PoonLab/indelrates.

Bibliography

[1] John Palmer and Art FY Poon. Phylogenetic measures of indel rate variation among the hiv-1 group m subtypes. *Virus evolution*, 5(2):vez022, 2019.

[2] Louis M Mansky and Howard M Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, 69(8):5087–5094, 1995.

[3] W. H. Li, M. Tanimura, and P. M. Sharp. Rates and dates of divergence between AIDS virus nucleotide sequences. *Molecular Biology and Evolution*, 5(4):313–330, July 1988. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040503. URL https://academic.oup.com/mbe/article/5/4/313/1026948.

[4] Vlad Novitsky, Rui Wang, Raabya Rossenkhan, Sikhulile Moyo, and M Essex. Intra-host evolutionary rates in hiv-1c env and gag during primary infection. *Infection, Genetics and Evolution*, 19:361–368, 2013.

[5] Samuel Alizon and Christophe Fraser. Within-host and between-host evolutionary rates across the hiv-1 genome. *Retrovirology*, 10(1):49, 2013.

[6] Simon DW Frost, Terri Wrin, Davey M Smith, Sergei L Kosakovsky Pond, Yang Liu, Ellen Paxinos, Colombe Chappey, Justin Galovich, Jeff Beauchaine, Christos J Petropoulos, et al. Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent hiv infection. *Proceedings of the National Academy of Sciences*, 102(51):18514–18519, 2005.

[7] Penny L Moore, Nthabeleng Ranchobe, Bronwen E Lambson, Elin S Gray, Eleanor Cave, Melissa-Rose Abrahams, Gama Bandawe, Koleka Mlisana, Salim S Abdool Karim, Carolyn Williamson, et al. Limited neutralizing antibody specificities drive neutralization escape in early hiv-1 subtype c infection. *PLoS Pathog*, 5(9):e1000598, 2009.

[8] Natasha Wood, Tanmoy Bhattacharya, Brandon F Keele, Elena Giorgi, Michael Liu, Brian Gaschen, Marcus Daniels, Guido Ferrari, Barton F Haynes, Andrew McMichael, et al. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS pathogens*, 5(5):e1000414, 2009.

[9] Rong Rong, Frederic Bibollet-Ruche, Joseph Mulenga, Susan Allen, Jerry L Blackwell, and Cynthia A Derdeyn. Role of v1v2 and other human immunodeficiency virus type 1 envelope domains in resistance to autologous neutralization during clade c infection. *Journal of virology*, 81(3):1350–1359, 2007.

[10] Manish Sagar, Xueling Wu, Sandra Lee, and Julie Overbaugh. Human immunodeficiency virus type 1 v1-v2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. *Journal of virology*, 80(19):9586–9598, 2006.

[11] Caroline F Wright, Marco J Morelli, Gaël Thébaud, Nick J Knowles, Pawel Herzyk, David J Paton, Daniel T Haydon, and Donald P King. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *Journal of virology*, 85(5):2266–2275, 2011.

[12] Vincent Montoya, Andrea Olmstead, Patrick Tang, Darrel Cook, Naveed Janjua, Jason Grebely, Brendan Jacka, Art FY Poon, and Mel Krajden. Deep sequencing increases hepatitis c virus phylogenetic cluster detection compared to sanger sequencing. *Infection, Genetics and Evolution*, 43:329–337, 2016.

[13] Carlos Y Valenzuela, Sergio V Flores, and Javier Cisternas. Fixations of the hiv-1 env

gene refute neutralism: new evidence for pan-selective evolution. *Biological research*, 43 (2):149–163, 2010.

[14] CTT Edwards, EC Holmes, OG Pybus, DJ Wilson, RP Viscidi, EJ Abrams, RE Phillips, and AJ Drummond. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics*, 174(3):1441–1453, 2006.

[15] Denis M Tebit and Eric J Arts. Tracking a century of global expansion and evolution of HIV to drive understanding and to combat disease. *The Lancet infectious diseases*, 11(1): 45–56, 2011.

[16] Lee Ratner, William Haseltine, Roberto Patarca, Kenneth J Livak, Bruno Starcich, Steven F Josephs, Ellen R Doran, J Antoni Rafalski, Erik A Whitehorn, Kirk Baumeister, et al. Complete nucleotide sequence of the aids virus, htlv-iii. *Nature*, 313(6000): 277–284, 1985.

[17] Tonny Kinene, J Wainaina, Solomon Maina, and LM Boykin. Rooting trees, methods for. *Encyclopedia of Evolutionary Biology*, page 489, 2016.

[18] E Zuckerkandl, L Pauling, et al. Horizons in biochemistry. *Horizons in Biochemistry*, pages 97–166, 1962.

[19] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

[20] Thu-Hien To, Matthieu Jung, Samantha Lycett, and Olivier Gascuel. Fast dating using least-squares criteria and algorithms. *Systematic biology*, 65(1):82–97, 2015.

[21] Alexei J Drummond, Simon YW Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4(5):e88, 2006.

[22] World Health Organization. Hiv/aids, 2018.

[23] Bette Korber, Brian Gaschen, Karina Yusim, Rama Thakallapally, Can Kesmir, and Vincent Detours. Evolutionary and immunological implications of contemporary HIV-1 variation. *British medical bulletin*, 58(1):19–42, 2001.

[24] Noah Kiwanuka, Oliver Laeyendecker, Merlin Robb, Godfrey Kigozi, Miguel Arroyo, Francine McCutchan, Leigh Anne Eller, Michael Eller, Fred Makumbi, Deborah Birx, et al. Effect of human immunodeficiency virus Type 1 (HIV-1) subtype on disease progression in persons from Rakai, Uganda, with incident HIV-1 infection. *J Infect Dis*, 197 (5):707–13, 2008.

[25] Ashwin Vasan, Boris Renjifo, Ellen Hertzmark, Beth Chaplin, Gernard Msamanga, Max Essex, Wafaie Fawzi, and David Hunter. Different Rates of Disease Progression of HIV Type 1 Infection in Tanzania Based on Infecting Subtype. *Clinical Infectious Diseases*, 42(6):843–852, 2006. ISSN 1058-4838. URL http://www.jstor.org.proxy1.lib.uwo.ca/stable/4463724.

[26] Mark A Wainberg. HIV-1 subtype distribution and the problem of drug resistance. *Aids*, 18:S63–S68, 2004.

[27] Kevin K Ariën, Guido Vanham, and Eric J Arts. Is HIV-1 evolving to a less virulent form in humans? *Nature Reviews Microbiology*, 5(2):141, 2007.

[28] Erin EH Tran, Mario J Borgnia, Oleg Kuybeda, David M Schauder, Alberto Bartesaghi, Gabriel A Frank, Guillermo Sapiro, Jacqueline LS Milne, and Sriram Subramaniam. Structural mechanism of trimeric HIV-1 envelope glycoprotein activation. *PLoS Pathogens*, 8(7):e1002797, 2012.

[29] Peter D Kwong, Richard Wyatt, James Robinson, Raymond W Sweet, Joseph Sodroski, and Wayne A Hendrickson. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, 393(6686):648, 1998.

[30] Ming Zhang, Brian Gaschen, Wendy Blay, Brian Foley, Nancy Haigwood, Carla Kuiken, and Bette Korber. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology*, 14(12):1229–1246, 2004.

[31] Michael E Abram, Andrea L Ferris, Wei Shao, W Gregory Alvord, and Stephen H Hughes. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *Journal of virology*, 84(19):9864–9878, 2010.

[32] Wilco Keulen, Charles Boucher, and Ben Berkhout. Nucleotide substitution patterns can predict the requirements for drug-resistance of HIV-1 proteins. *Antiviral research*, 31 (1-2):45–57, 1996.

[33] Rasmus Nielsen and Ziheng Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–936, 1998.

[34] Abdelrahim Rakik, Mounir Ait-Khaled, Philip Griffin, Deborah A Thomas, Margaret Tisdale, et al. A novel genotype encoding a single amino acid insertion and five other substitutions between residues 64 and 74 of the HIV-1 reverse transcriptase confers high-level cross-resistance to nucleoside reverse transcriptase inhibitors. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 22(2):139–145, 1999.

[35] Shambhu G. Aralaguppe, Dane Winner, Kamalendra Singh, Stefan G. Sarafianos, Miguel E. Quiñones-Mateu, Anders Sönnerborg, and Ujjwal Neogi. Increased replication capacity following evolution of PYxE insertion in Gag-p6 is associated with enhanced virulence in HIV-1 subtype C from East Africa. *Journal of Medical Virology*, 89 (1):106–111, January 2017. ISSN 1096-9071. doi: 10.1002/jmv.24610.

[36] Louis Alexander, Mary Janette Aquino-DeJesus, Michael Chan, and Warren A Andiman. Inhibition of human immunodeficiency virus type 1 (HIV-1) replication by a two-amino-

acid insertion in HIV-1 Vif from a nonprogressing mother and child. *J Virol*, 76(20): 10533–10539, 2002.

[37] Cynthia A Derdeyn, Julie M Decker, Frederic Bibollet-Ruche, John L Mokili, Mark Muldoon, Scott A Denham, Marintha L Heil, Francis Kasolo, Rosemary Musonda, Beatrice H Hahn, et al. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science*, 303(5666):2019–2022, 2004.

[38] Kiyoto Tsuchiya, Hirotaka Ode, Tsunefusa Hayashida, Junko Kakizawa, Hironori Sato, Shinichi Oka, and Hiroyuki Gatanaga. Arginine insertion and loss of N-linked glycosylation site in HIV-1 envelope V3 region confer CXCR4-tropism. *Scientific reports*, 3:2389, 2013.

[39] Andy McKenzie and Mike Steel. Distributions of cherries for two models of trees. *Mathematical biosciences*, 164(1):81–92, 2000.

[40] Stephen F Altschul and Bruce W Erickson. Optimal sequence alignment using affine gap costs. *Bull Math Biol*, 48(5-6):603–616, 1986.

[41] David C Nickle, Laura Heath, Mark A Jensen, Peter B Gilbert, James I Mullins, and Sergei L Kosakovsky Pond. HIV-specific probabilistic models of protein evolution. *PLoS One*, 2(6):e503, 2007.

[42] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[43] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010.

[44] Alexei Drummond, Oliver G Pybus, and Andrew Rambaut. Inference of viral evolutionary rates from molecular sequences. *Adv Parasitol*, 54:331–358, 2003.

[45] Abayomi S Olabode, Mariano Avino, Tammy Ng, Faisal Abu-Sardanah, David W Dick, and Art FY Poon. Evidence for a recombinant origin of HIV-1 group M from genomic variation. *bioRxiv*, page 364075, 2018.

[46] Joel O Wertheim, Mathieu Fourment, and Sergei L Kosakovsky Pond. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. *Mol Biol Evol*, 29(2):451–456, 2011.

[47] Joris Hemelaar. The origin and diversity of the HIV-1 pandemic. *Trends Mol Med*, 18 (3):182–192, 2012.

[48] Marcel Tongo, Gordon W Harkins, Jeffrey R Dorfman, Erik Billings, Sodsai Tovanabutra, Tulio de Oliveira, and Darren P Martin. Unravelling the complicated evolutionary and dissemination history of HIV-1M subtype A lineages. *Virus Evolution*, 4(1):vey003, 2018.

[49] Alan S Perelson and Patrick W Nelson. Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Rev*, 41(1):3–44, 1999.

[50] X. Liang, S. Munshi, J. Shendure, G. Mark, M. E. Davies, D. C. Freed, D. C. Montefiori, and J. W. Shiver. Epitope insertion into variable loops of HIV-1 gp120 as a potential means to improve immunogenicity of viral envelope protein. *Vaccine*, 17(22):2862–2872, July 1999. ISSN 0264-410X.

[51] Xunqing Jiang, Valicia Burke, Maxim Totrov, Constance Williams, Timothy Cardozo, Miroslaw K Gorny, Susan Zolla-Pazner, and Xiang-Peng Kong. Conserved structural elements in the V3 crown of HIV-1 gp120. *Nature Structural and Molecular Biology*, 17 (8):955, 2010.

[52] Art FY Poon, Fraser I Lewis, Sergei L Kosakovsky Pond, and Simon DW Frost. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput Biol*, 3(11):e231, 2007.

[53] Bhavna Chohan, Dorothy Lang, Manish Sagar, Bette Korber, Ludo Lavreys, Barbra Richardson, and Julie Overbaugh. Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter V1-V2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral RNA levels. *J Virol*, 79(10): 6528–6531, 2005.

[54] Marcel E Curlin, Rafael Zioni, Stephen E Hawes, Yi Liu, Wenjie Deng, Geoffrey S Gottlieb, Tuofu Zhu, and James I Mullins. HIV-1 envelope subregion length variation during disease progression. *PLoS pathogens*, 6(12):e1001228, 2010.

[55] Manish Sagar, Xueling Wu, Sandra Lee, and Julie Overbaugh. Human immunodeficiency virus type 1 V1-V2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. *J Virol*, 80(19):9586–9598, 2006.

[56] Mark T Liddament, William L Brown, April J Schumacher, and Reuben S Harris. APOBEC3F properties and hypermutation preferences indicate activity against HIV-1 in vivo. *Curr Biol*, 14(15):1385–1391, 2004.

[57] Xuegong Zhu, Christoph Borchers, Rachelle J Bienstock, and Kenneth B Tomer. Mass spectrometric characterization of the glycosylation pattern of HIV-gp120 expressed in CHO cells. *Biochemistry*, 39(37):11194–11204, 2000.

Chapter 3

Indel Evolution Within Hosts

3.1 Introduction

HIV-1 demonstrates an incredible capacity to evolve and diversify within hosts due to a variety of factors including its short generation time, high viral load, and rapid accumulation of mutations such as substitutions, indels, and recombinations [1–4]. These factors also produce a unique HIV-1 metapopulation structure within hosts, comprised of numerous smaller subpopulations that exhibit a unique virus genotype [5–7]. Together, these characteristics make the HIV-1 population a highly resilient, rapidly moving target capable of adapting to and overwhelming the host immune system in the vast majority of infected subjects [8, 9]. For instance, the acute phase of HIV-1 infection sees high counts of HIV-1 virus (roughly 10^7) generated in the first 1-3 months, even though initial infection typically originates from just a single founding virus [10, 11]. This is followed by an asymptomatic phase during which virus levels are generally lower, but importantly, still undergo rapid turnover and adaptive evolution to host immune responses [8, 9]. It is this adaptive evolution that enables HIV-1 to induce its final disease outcome of AIDS, demonstrating the importance of studying within host evolutionary processes [8].

3.1.1 Sampling Differences

Importantly, the HIV-1 sampled in sequence data within hosts differs from the HIV-1 being sampled among different hosts (one sequence per patient) in the population. For context, recall that the error-prone nature of HIV-1 replication and roughly 10^{10} new offspring produced each day generate enormous diversity in the within-host population [12, 13]. HIV-1 research conducted among hosts predominantly utilizes Sanger sequencing technology, as it provides a single consensus sequence describing each individual's virus population, in line with its objectives. Sanger sequencing is unable to sample low frequency variants in the HIV-1 population, meaning that mutations are required to reach sufficient prevalence (at least $\sim 20\%$) to be included in the final sequence product [14–16]. Reaching sufficient prevalence involves enduring extensive purifying selection over numerous viral generations, therefore making the final sequence a highly refined and filtered representation of the HIV-1 population.

On the other hand, sequence data sampled within individual hosts over time is predominantly generated using single genome amplification (SGA): a method that involves the serial dilution of virus sample to concentrations low enough that individual virus genomes can be isolated and sequenced. Instead of creating an average or consensus, SGA generates observations of individual viruses which importantly, can include HIV-1 genetic variants present at lower than 20% prevalence in the population. This means that many mutations, including indels, can exist at low prevalence in patient HIV-1 populations and be detected using SGA, but will often not reach high enough prevalence to be included in the final Sanger consensus sequence when sampling among hosts due to purifying selection [16, 17]. For example, Wood et al. [18] found numerous frameshift-inducing indels in the variable regions of gp120 using within-host sequence data, while my previous study examining these same regions among hosts found essentially no cases of frameshifts [19]. These sampling differences provided part of the rationale for our choice to conduct a within-host phylogenetic analysis in this study having performed a similar one among hosts previously (Chapter 2).

3.1.2 Evolutionary Processes

Different processes shape the evolution of HIV-1 populations within hosts compared to that occurring among hosts, which were discussed previously (Chapter 2). Within hosts, HIV-1 populations predominantly undergo adaptive evolution in order to evade host immune responses, a process which involves the successive fixation of advantageous traits driven by positive selection [2, 20–22]. Purifying selection is also present and responsible for filtering out mutations that compromise virus functionality and thus confer a decrease in viral fitness [17].

On an among-host scale, recall that diversity here is strongly shaped by HIV-1 transmission and its associated evolutionary pressures. This primarily involves strong population bottleneck effects, as HIV-1 infection is established by only one or a few T/F variants [9]. In the presence of these stringent bottleneck effects, the process of reaching high prevalence or fixation is essentially the only way a mutation can be transmitted to the new host and thus, plays a significant role in modulating among-host variation [9, 16]. In fact, the relatively small fraction of mutation successfully transmitted to a new host helps explain the higher evolutionary rates of HIV-1 measured within hosts than those measured among different hosts in the population [23, 24].

3.1.3 HIV-1 gp120

The above trends concerning evolution among and within hosts are particularly applicable to the *env* gene that encodes the gp120 glycoprotein, given that it demonstrates the highest rates of evolution and most abundant positive selection relative to other genes in the HIV-1 genome [23, 25, 26]. Briefly recall that the gp120 gene is segmented into five conserved and five variable regions, the latter of which encode five disordered loop structures on the exterior of the protein. Additionally, the gp120 protein sequence contains numerous potential N-linked glycosylation sites (PNGS) that form a dense glycan shield around this protein. In line with among host evolution, there are reports that the glycan shield and variable loop lengths of gp120 follow trends of transmission fitness, as transmitted HIV-1 tends to have shorter gp120 variable loops and fewer N-linked glycosylation sites that lengthen and increase in number over the course of infection, respectively [27]. The gp120 glycoprotein also follows trends of within-host evolution, as demonstrated by its frequent accumulation of mutations that escape the neutralizing antibody and CTL immune responses [28].

Immune Escape

Importantly, immune escape mutations, whether in gp120 or other HIV-1 genes, enable the avoidance of HIV-1 neutralization in the circulation (antibody-mediated) or suppression during infection of cells (CTL-mediated), therefore allowing HIV-1 to undergo replication normally [29, 30]. These escape mutations can occur through multiple different pathways that involve amino acid sites in both the conserved and variable regions of gp120 [31], though they are more frequently found in the latter due to direct targeting by neutralizing antibodies [32, 33]. This is demonstrated by studies finding immune escape mutations in the V1V2 regions [34, 35], V3 regions [21, 36], and V4 regions [33, 35, 37], for example. These escape mutations can come in many forms including nucleotide substitutions, indel events, and changes to the positions of N-linked glycans [26, 29, 34].

Rates

For one, the role of nucleotide substitutions in immune escape has been extensively studied in gp120 [16, 29, 34, 35] and fittingly, so have the rates of substitutions in this gene [23, 38, 39]. The rates of mutations, such as substitutions or indels, are highly important and useful to the study of virus evolution, as they provide an indication of how rapidly a pathogen evolves its molecular sequence and thus, the speed at which it can respond to environmental pressures [23, 40]. Therefore, faster evolution has been suggested to correlate with greater pathogen resilience and quicker disease progression [13, 41, 42]. Furthermore, rates can also be used to estimate evolutionary time scales and reconstruct common ancestry in phylogenetic inference, which can be used to provide insights into a pathogen's diversity, origins, and evolutionary patterns [43, 44].

3.1.4 Indels

Indels have also been shown to play a significant role in the generation of immune escape variants in gp120, especially in the variable regions where they accumulate rapidly [27, 37, 45–47]. Indels produce these escape variants by altering the physical structure of the variable loops in several different ways, which include changes to amino acid compositions, overall variable loop lengths, and changes to the number and positions of PNGS which control the glycan shield [18, 34, 48]. However, despite the implicated role of indels in HIV-1 immune escape, the estimation of indel rates had not yet been conducted within the *env* gene of HIV-1 prior to my study present in Chapter 2. This study addressed this knowledge gap by estimating among-host indel rates in the gp120 variable loops using a phylogenetic analysis of HIV-1 sequence data worldwide [19]. This work is important as it addresses the question of how quickly indels are being fixed in the HIV-1 populations circulating in different people [19]. By nature of the among-host sequence data utilized in this past study however, this study is unable to examine the indels that either get removed from the population by purifying selection or simply do not reach high enough prevalence to be included in the final sequence product [17]. The considerable differences in evolutionary pressures and type of sampled virus within hosts compared to among hosts provided the rationale for this study. The study described in this chapter seeks to address this unexplored topic by estimating gp120 variable loop indel rates within hosts using longitudinally sampled patient data.

3.2 Hypothesis

I hypothesize that the rates of variable loop insertions and deletions can be determined on a within-host scale using a dated phylogenetic analysis. Compared to my previous among-host indel analysis, I postulate that insertion and deletion rates will be higher within hosts because these sampling approaches enable the detection of additional virus that has undergone less purifying selection relative to that sampled among hosts [17, 19]. Given the predicted weaker effects of purifying selection, I further expect to detect more indels that induce frameshift mutations. Finally, I also predict that insertions and deletions will induce changes to N-linked glycosylation site counts in the variable loops over the course of infection.

3.3 Methods

3.3.1 Sequence Retrieval and Processing

I first queried the Los Alamos National Laboratory HIV Database (LANL; http://www.hiv.lanl.gov/) for HIV-1 sequences that covered gp120 and were derived from longitudinally sampled patients containing at least 100 sequences. Sequence data downloaded from LANL contained a total of 11,265 sequences, spanning 29 different research studies and 25 unique patients. In the data, I found cases of redundancy where the same patient appeared in multiple studies. Using a Python script, I compared redundant patient datasets and selected the single largest one (if applicable) for the next stage of processing. At this point, I also checked for sequence that did not cover roughly 90% of full-length gp120 (1400 nt), though no sequences registered positive for this screen. Next, I needed to ensure that all patient sequences were consistently labelled with an appropriate time value. Consistent time-stamping of all sequences was crucial for later analysis, as these values were used to rescale phylogenetic trees in time and provide the basis for indel rate estimates. Naturally, research studies authoring these patient data on LANL had a variety of different aims, meaning that collection date measures were inconsistent between studies. The different time measures found in LANL-derived datasets included: the number of days since first sample, treatment start, infection start, and seroconversion. I am aware that these different scales could describe considerably different periods of HIV-1 infection. However, we deemed any of these four time scales sufficient for this study given that subsequent analysis would be conducting a full reconstruction of HIV-1 evolutionary history within hosts and rescaling time estimates while doing so. For patient datasets annotated with more than one time scale, I chose the time scale that had the widest coverage of sequences, and fewest negative or incomplete entries. I also removed patient datasets that spanned less than 200 days and contained fewer than five unique sampling timepoints to ensure that datasets provided sufficient coverage of HIV-1 infection.

I then needed to check whether any data were derived from DNA sequences because HIV-1 DNA can be an ineffective measure of within-host evolution when a patient is receiving anti-retroviral therapy (ART). When a patient is receiving ART, HIV-1 replication is suppressed and the virus is not able to integrate its viral DNA into the genomes of $CD4^{+}$ cells, thus prohibiting detectable evolution in these sampled sequences. All patient datasets in the study were found to contain HIV RNA sequence data, with the exception of one patient data set which contained only HIV-1 DNA. Upon further investigation of this data set, we found that the corresponding patient was not receiving ART, indicating that the patient's DNA sequences were undergoing evolution and suitable for use in our analysis.

I also received an additional 2,541 HIV-1 gp120 sequences from Dr. Vlad Novitsky at Harvard University. These sequence data were derived from 25 patients also sampled using SGS and provided coverage from V1 to C5 of gp120. Unlike the LANL-derived sequence data, these sequences did not include the first conserved region of gp120 (C1) and therefore, required slight modifications to initial alignment and extraction procedures. All gp120 sequences from the Novitsky Lab were labelled with a consistent date indicating the days past seroconversion, requiring no further screening efforts in this regard.

Nucleotide sequences from both sources were sorted into individual FASTA files on a per-patient basis using unique identifiers in the header of each sequence. This produced an initial data set of 6,717 sequences sampled across 44 patients.

3.3.2 Sequence Slicing and Alignment

The next step of my pipeline removed sequence outside the gp120 gene and recorded the locations of the five gp120 variable regions. To accomplish this, I performed pairwise alignments using a Python-based implementation of the Gotoh algorithm as modified by Altschul and Erickson [49] (http://github.com/ArtPoon/gotoh2). I aligned every gp120 nucleotide sequence in my patient datasets against the gp120 sequence of HXB2 (Genbank Acession: K03455), which recall, is the reference sequence of HIV-1 subtype B. Using the HXB2 annotations, I then cut patient nucleotide sequences at the boundaries of the aligned boundaries of gp120 to remove extraneous flanking regions that did not overlap with the gp120 gene. Next, I removed all gap characters from this result, translated it into its amino acid sequence, and aligned it to the HXB2 gp120 protein sequence. Using the locations of gaps in the protein alignment, I then inserted gap triplets (i.e. – – –) into the corresponding locations of the patient's nucleotide sequence in order to generate codon-based nucleotide sequence alignments. We opted for this approach to address issues in locating the exact nucleotide boundaries of the variable loop sequences, which were believed to be caused in part by the abundance of mutations present. Using the HXB2 reference annotations, I searched our codon-based nucleotide alignment and recorded the sequences and locations of the five gp120 variable regions in every patient sequence for later analysis.

I ran MAFFT (version 7.271) [50] with default settings on each set of patient-derived gp120 sequence data to generate multiple sequence alignments (MSA) describing within-host diversity.

3.3.3 Patient Screening for Phylogenetic Analysis

I then used the Randomized Axelerated Maximum Likelihood program (RAxML) version 8.2.11 to reconstruct ML phylogenetic trees describing within-host gp120 evolution [51]. Recall that a phylogenetic tree is a model of evolutionary history governed by multiple parameters, such as the evolutionary rate and branching order (topology). The next step was to root these

ML phylogenetic trees, as the current unrooted trees do not describe the ancestry or direction of evolution among tree lineages. While multiple tree rooting methods exist, I used the root-to-tip regression function made available in the Analyses of Phylogenetic and Evolution (APE) R package to root my phylogenetic trees [52].

Molecular Clock Signal

Next, I proceeded to gauge the strength of the same molecular clock signal in my patient datasets. An important step in my study involves the generation of time-scaled phylogenetic trees, *i.e.* trees whose branch units have been rescaled to units of time, such as days or years. Scaling a phylogenetic tree in time requires the presence of molecular clock signal in the data and therefore, the inclusion of patients in forthcoming analyses is contingent on their sequence data exhibiting sufficient clock signal. I generated a root-to-tip regression on each within-host ML tree and excluded patient datasets with an R^2 value below 0.30, which removed an additional 13 patients from the analysis.

Hypermutation

The hypermutation of HIV-1 sequences refers to an abundance of G to A nucleotide substitutions and is primarily caused by an intrinsic anti-viral mechanism in human cells involving the protein APOBEC3G [53]. Hypermutation can cause problems in HIV-1 sequence analysis (*i.e.* introduction of premature stop codons) and thus need to be removed [54]. I therefore, screened the 32 remaining patient datasets for sequences containing G to A hypermutation using a local implementation of the Hypermut Analysis Tool made available by LANL [54]. To conduct this analysis, I generated a consensus sequence from the sequence data collected at the first sampling timepoint and used it as the reference for this analysis. We opted to use a consensus generated from the first timepoint because it would approximate a relatively older ancestral sequence and thus allow the directionality of substitutions to be determined (*i.e.* G in ancestor ⟶ A in descendant). In one patient data set (LANL ID: 30651), Hypermut flagged approximately 35% of all sequences as hypermutants. To investigate this result, we first reviewed the metadata describing this patient data set in its study of origin and found that no evidence of hypermutation had been reported. [55]. I then manually examined this patient's multiple sequence alignment, but no grouping of distinct sequences displaying signs of hypermutation could be found. Given the sufficient clock-like signal and lack of any highly divergent sequences, we elected to skip the hypermutation screen in this patient.

Superinfection

Next, I manually inspected patient MSAs and phylogenetic trees for signs of co-infection or superinfection. One patient (LANL ID: 111848) showed signs of superinfection based on the presence of two distantly related populations in its phylogenetic tree evolving over the same sampling time points. Upon referencing the original study, we confirmed that this patient had a highly heterogeneous infection and contained extensive recombination [56]. We opted to exclude this patient from the study given the multiple complications that would arise from the two divergent HIV-1 populations and numerous recombinant sequences contained in the data set. The screens for superinfection, hypermutation, and molecular clock signal produced a data set containing 4,573 sequences from 30 patients.

3.3.4 Phylogenetic Tree Reconstruction

I then reconstructed phylogenetic trees using Bayesian Evolutionary Analysis Sampling Trees (BEAST) v1.10.4 [57]: a software framework that uses Bayesian statistical inference to reconstruct phylogenetic trees under a variety of models. Briefly, Bayesian inference in BEAST first involves placing a prior probability distribution (starting estimate or guess) on each of the tree parameters [57]. From here, BEAST implements a Markov chain Monte Carlo (MCMC) algorithm: an algorithm that draws a (typically large) number of samples from a probability distribution, where each step only depends on the previous one, in order to approximate its characteristics [57]. Markov chain Monte Carlo methods are useful for sampling probability distributions in highly complex model systems, often involving multiple parameters. BEAST uses MCMC to sample tree parameter values according to their posterior probability: the product of the value's prior probability and the likelihood of the observed data given the parameter values. As a result, the density of values sampled through MCMC will be proportional to the parameter's posterior distribution (final result). This MCMC sampling procedure is performed for every tree parameter, including the tree topology itself. Importantly, BEAST analysis also incorporates time data to scale phylogenetic trees in time and thus depends on the presence of molecular clock signal which I screened for earlier.

Parameters

I performed two analyses on each patient in BEAST, each of which ran for 100 million MCMC iterations [57]. The use of MCMC in a Bayesian framework allows BEAST to fit numerous models to the data that facilitate the effective reconstruction of time-scaled phylogenies. For one, I fit a TN93 model of evolution to the data, which includes parameters describing nucleotide frequencies and three different rate parameters: one for nucleotide transversions and one for each type of transition [58]. I also fit a Gamma site heterogeneity model, which accounts for differences in the substitution rate between nucleotide positions in the sequence by assuming that the rate at each position is randomly drawn from a Gamma distribution [59]. For the molecular clock, I fit a uncorrelated relaxed clock model, which allows each tree branch to adopt an evolutionary rate independent of other branches (thus uncorrelated) that is drawn from a log-normal probability distribution [60].

Posterior Convergence

In order for BEAST output to be deemed sensible and valid, the posterior distributions of tree parameters were required to exhibit sufficient convergence on a single value (*i.e.* spending the majority of steps close to a single value and not substantially deviated from it). However, roughly 12 patient analyses had abnormally high clock rates often accompanied with large variance, causing poor convergence of multiple other tree parameters. I attempted to fix this by 1) setting narrower prior distributions on the mean and standard deviation parameters governing the relaxed clock model, 2) setting an upper bound on the analysis, and 3) setting the start value closer to evolutionary rates approximated by the root-to-tip regression analysis conducted previously ($1E^{-5} - 1E^{-4}$). These attempts however, only partially fixed convergence problems and so we deemed it necessary to constrain the tree topology during BEAST analysis in order to get working results. Normally, BEAST is able to change the topology, or branching order, of the phylogenetic tree during analysis to explore a wide variety of possible tree configurations. Constraining the topology involves restricting BEAST analysis to sample a single tree structure determined by a preset guide tree, which in this case, was the ML tree generated using RAxML. This greatly improved convergence by reducing analysis complexity while still allowing BEAST to focus on the optimal scaling of branch lengths, which were central to our estimation of indel rates.

Coalescent Model & Model Selection

In BEAST, I also fit a coalescent model to the phylogenetic tree — a model describing how genetically-distinct lineages originate from points of common ancestry [61]. The coalescent model permits estimation of a phylogeny's effective population size (N_e), which is essentially the size of a theoretical population that would demonstrate identical traits to the real one under the model assumptions [62]. The N_e provides an estimate of how population demographics are changing over time and can be used to compare between evolutionary models [62]. The constant coalescent model is the most basic of these models which assumes that the N_e remains constant throughout evolutionary history [61]. In BEAST, there have been several advancements to the coalescent model that, for example, allow the population size to vary between coalescent events and apply a smoothing algorithm over the population trajectory. Of these extensions, I decided to use the recent Skygrid coalescent model. Importantly, this model contains a parameter called the number of population sizes: a user-defined number specifying how many times the N_e to allowed to change throughout the tree. As the most modern development, the Skygrid model provides the greatest flexibility of the existing coalescent model versions.

For each patient's data set, use of the Skygrid model first required justification that this model was preferred over the constant coalescent using model selection criteria, *i.e.* demonstrating the model provided a sufficiently better fit to this patient's data than the constant coalescent while accounting for its higher complexity. For those patients that fit the Skygrid model, I further sought to determine the number of population sizes that best suited their sequence data, as these patient datasets had differing time periods of infection. Accounting for both of these factors, I conducted model selection between the constant coalescent model and Skygrid models containing 10, 20, and 30 population sizes using the Bayes factor: the ratio between the likelihood values (probability of the data given model parameters) of the two model fits [63]. Importantly, there were rare instances where the model selected by the Bayes factor conferred poor posterior convergence of tree parameters. Since posterior convergence is essential to consider results legitimate, we cannot trust the Bayes factor to determine the most appropriate model in these cases, and so I selected the model with the next highest Bayes factor that conferred sufficient posterior convergence.

Skygrid cutoff values — a parameter indicating how far back in time the coalescent model will reconstruct infection — were set to 125% of the estimated RAxML tree root height to cover the entirety of infection. Of the 60 BEAST analyses performed (2 per patient), 24 runs used the constant coalescent while 9, 22, and 5 runs used the Skygrid model containing 10, 20, and 30 population sizes, respectively. Brief tests involving fewer than 10 and more than 30 population sizes were commonly met with poor convergence.

Tree Extraction

I extracted phylogenetic trees from the BEAST analyses in two different ways to address separate objectives in my study. First, I randomly sampled 200 phylogenetic trees from the posterior distribution of each patient's BEAST analysis for the purposes of determining indel rates and timings during infection. Two BEAST analyses were conducted per patient, resulting in a total of 400 trees/pat. We chose this approach because these analyses involved continuous time data and used methods that could incorporate numerous data points to estimate distributions of the results. The other approach involved extracting the single posterior tree selected by a branch length consensus algorithm. A single tree was deemed suitable for the remaining analyses on lengths, compositions, and PNGS changes because these analyses dealt with discrete observations of indel sequences. This branch length consensus (BLC) tree was selected by first generating a distance matrix relating the branch lengths of all posterior trees, determining the centroid position of this matrix (*i.e.* the mean position across all branch length dimensions), and finding the tree with the lowest Euclidean distance to this centroid position. The tree selected using this method will be closest to the theoretical center of the distribution of sampled trees, meaning that it will have one of the highest posterior probabilities.

3.3.5 Ancestral State Reconstruction & Indel Extraction

The next step involved extracting insertions and deletions from the phylogenetic trees sampled from the posterior of BEAST analysis by reconstructing the ancestral state sequences within these trees. For this task, I chose to use the Historian v0.1 ancestral reconstruction software package because of its incorporation of optimized algorithms from other modern software packages, its effective reconstruction of indels, and its accuracy in estimating evolutionary rates [64]. After receiving a MSA as input, Historian attempts to reconstruct a phylogenetic tree to conduct its analysis. Given that I already reconstructed trees in BEAST, I turned off phylogenetic reconstruction in Historian and used my own trees as a guide for analysis. I performed two separate analyses in Historian as my phylogenetic trees were now stratified into two formats: the 400 randomly sampled posterior trees, and the single modified BLC tree from each patient's BEAST run. We encountered errors when running Historian on the four largest patient datasets, which persisted despite tuning parameters to lessen the complexity of the analysis. This prevented 4 patient datasets from completing their analyses in Historian, resulting in a fully finalized data set containing 2,668 sequences from 26 patients. When stratified by HIV-1 group M subtype, there were 2,452 subtype C sequences across 24 patients and 216 subtype B in 2 patients.

The output of Historian is a MSA relating all tip and ancestor sequences in the phylogenetic tree. Therefore, to extract indels, I recursively iterated through the within-host tree to identify pairs of sequences connected by a branch (*i.e.* every ancestor-descendant pair). I then scanned through the aligned variable regions in the Historian output using locations determined in my previous in-frame pairwise alignment with HXB2 gp120. Insertions presented as nucleotides that were only found in the descendant, while deleted nucleotides were only present in the ancestor.

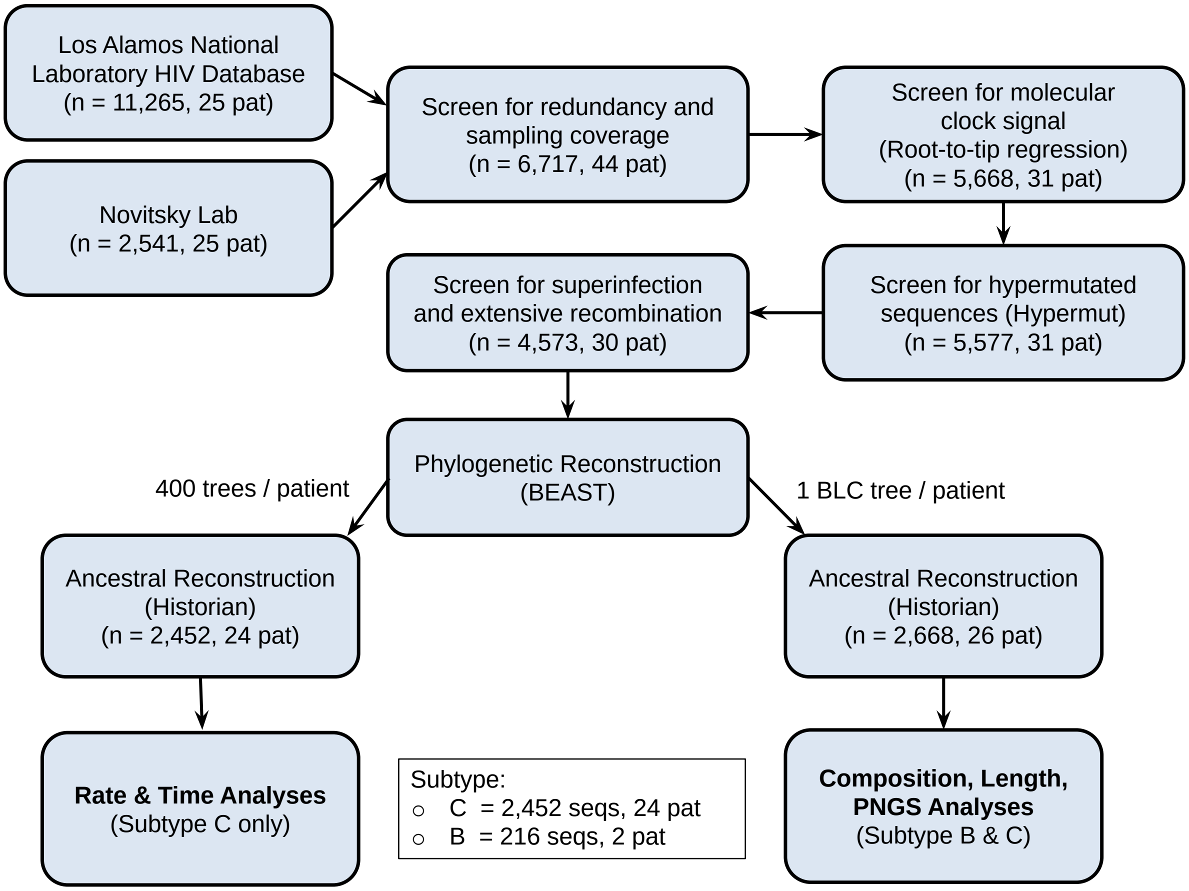
Los Alamos National Laboratory HIV Database (n = 11,265, 25 pat)
Novitsky Lab (n = 2,541, 25 pat)
Screen for redundancy and sampling coverage (n = 6,717, 44 pat)
Screen for molecular clock signal (Root-to-tip regression) (n = 5,668, 31 pat)
Screen for hypermutated sequences (Hypermut) (n = 5,577, 31 pat)
Screen for superinfection and extensive recombination (n = 4,573, 30 pat)
Phylogenetic Reconstruction (BEAST)
400 trees / patient
1 BLC tree / patient
Ancestral Reconstruction (Historian) (n = 2,452, 24 pat)
Ancestral Reconstruction (Historian) (n = 2,668, 26 pat)
Rate & Time Analyses (Subtype C only)
Subtype:
○ C = 2,452 seqs, 24 pat
○ B = 216 seqs, 2 pat
Composition, Length, PNGS Analyses (Subtype B & C)

Figure 3.1: Summary of data sourcing, filtering, and analyses in this study. BLC refers to branch length consensus trees found by determining the tree with the lowest Euclidean distance to the mean across all branch lengths.

3.3.6 Indel Rate Estimation

I first retrieved the indel count and time-scaled length associated with every tree branch in the 400 randomly sampled trees of the 24 subtype C patients (*n*=2,452). I then used the R implementation of Stan (RStan) v2.19.1, a modern statistical inference platform for conducting Bayesian inference, to fit a hierarchical model to this data and thereby estimate indel rates in our within-host data. This hierarchical model consisted of multiple layers to account for the stratification of data in this study. First, the model fits a Poisson generalized linear model to the counts of indels (y) along the branches of a single phylogenetic tree. The Stan framework then samples the mean count parameter as a product of the branch-associated time values (t) and a single indel rate (r) describing the tree. We decided that a Poisson GLM was most appropriate for this analysis given that the indel count data exhibited highly similar mean and variance measures, an intrinsic assumption of the Poisson distribution. The following equation describes the probability of observing indel counts y on branch k in one tree, given the time-scaled branch length t and the overall rate of the tree r.

$$P(y_k \mid r, t_k) = \frac{(rt_k)^{y_k} e^{-(rt_k)}}{y_k!} \tag{3.1}$$

By generating a single indel rate estimate per one tree (r), I accumulate 400 tree-based indel rate estimates within one patient ($r_j ... r_{400}$). The hierarchical model then fits these 400 rates to a normal distribution describing a single patient i and samples the mean (μ) and variance (σ_{pat}) parameters governing this distribution.

$$P(r_j \mid \mu_i, \sigma_{pat}) = \frac{1}{\sigma_{pat}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{r_j - \mu_i}{\sigma_{pat}}\right)^2} \tag{3.2}$$

This process is performed once for each patient in the study to generate 24 patient means ($\mu_i ... \mu_{24}$). Finally, the model then fits these 24 patient means to another normal distribution describing all of subtype C and again, samples the mean (ω) and variance (σ_{sub}) of this subtype-level distribution.

$$P(\mu_i \mid \omega, \sigma_{sub}) = \frac{1}{\sigma_{sub}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mu_i - \omega}{\sigma_{sub}}\right)^2} \tag{3.3}$$

The final result that I am estimating is the mean indel rate ω of HIV-1 subtype C. A summary of the entire hierarchical model is shown below, in which case distributions are simplified in the following manner: $normal(mean, variance)$, $Poisson(mean)$.

$$\mu_i \sim normal(\omega, \sigma_{sub}) \tag{3.4}$$

$$r_j \sim normal(\mu_i, \sigma_{pat}) \tag{3.5}$$

$$y_k \sim Poisson(r_j \times t_k) \tag{3.6}$$

I fit this hierarchical model in RStan, sampling for 10^4 MCMC iterations, in order to estimate insertion and deletion rates separately in each of the five gp120 variable loops (V1-V5). Additionally, I wanted to determine whether indels events accumulate more frequently in the recently-sampled terminal branches of the phylogenies, and so further I stratified each variable loop indel rate by internal and terminal branches in the tree. This produced one internal and terminal rate estimate per variable loop for insertions (2 × 5) and for deletions (2 × 5).

3.3.7 Indel Timings

While extracting indel events along tree branches, I also retrieved the cumulative branch length from the root of the tree to the midpoint of that branch. Phylogenetic inference aims to reconstruct the entire evolutionary history dating back to the root, or start of infection. Therefore, these cumulative length values in units of days represent the time since the start of infection when insertions and deletions occurred.

3.3.8 Nucleotide Composition

Next, I wanted to examine trends in nucleotide compositions within insertions and deletions. For this analysis and forthcoming ones, I applied Historian software to the BLC tree of each BEAST analysis to reconstruct ancestral state sequences and extract indels in these trees. Importantly, this involved use of the full data set which contained 26 patients and 2,668 sequences across subtypes A1, B and C. Using an R script, I collected insertions and deletions, stratified them by variable loop, and compared their nucleotide compositions to their variable loops of origin. To check for significant differences in these results, I performed randomization tests on the insertions and deletions in each variable loop. For every indel, I randomly sampled 100 nucleotide fragments of the same length from the variable loop where the indel occurred. I then determined whether the cumulative nucleotide proportions of the indels (size n) were above or below the 95% quantiles of the proportions of the randomly sampled fragments (size 100n).

3.3.9 Indel Lengths

Using the BLC tree, I examined the distribution of insertion and deletion lengths across the five gp120 variable loops. I wrote a custom R script that further stratified variable loop indels into seven different length categories. To detect significant differences, I utilized the 'mosaic' plot function made available by the 'vcd' R package on my data [65]. This function examines the Pearson chi-squared residuals of every bin in the contingency table (5 variable loops × 7 length categories) to detect counts that are significantly higher or lower than expected.

3.3.10 Changes to Potential N-linked Glycosylation Sites

I then sought to determine whether insertions and deletions added or removed PNGSs in the gp120 variable loops more frequently than expected by chance. Recall that previously, I scanned all branches of the within-host phylogenetic trees to extract the exact sequences and locations of indel events in the variable loops. While this provided information about the indels themselves, I also needed to determine the locations and numbers of PNGSs in the variable loop protein sequences. For every ancestor and tip sequence in each tree, I translated the variable loops to their amino acids sequences and searched them for PNGS motifs using the regular expression ‘N[ˆP][ST][ˆP]’, where ‘[ˆP]’ will match any amino acid except proline. At this point, I had information about all indels on tree branches and the counts of PNGSs in every ancestor and tip sequence.

This analysis required the comparison of two components: the observed and the expected changes in PNGS counts induced by indels. At every indel event in the tree, I first determined the number of PNGSs in the affected variable loop before and after the indel occurred (*i.e.* in the ancestor without the indel and the descendant containing it), and recorded the differences in these counts. Importantly, I made adjustments to ancestor-descendant sequence pairs to isolate for the effects of individual indels. In cases where multiple indels occurred on the same branch, I selected a single indel to analyze, kept all the others in the sequence, and determined PNGS counts with and without the one indel of interest. I chose to include other indels in the sequence because I was specifically interested in examining how each indel contributed to PNGS counts in context of all the others. For example, it might be that one indel does not add or remove a PNGS on its own, but might do so when placed in tandem with a second indel. Additionally, I also controlled for substitution events as a potential source of PNGS creation or deletion by reverting all polymorphisms to match the ancestral sequence before examining the indel. The results were the observed changes to PNGS counts induced by indels.

To estimate the indel-induced PNGS changes expected by chance, I performed a randomization test using the indels collected in the data. This approach involved placing every indel back into its variable loop sequence of origin 500 times in randomized locations, sampling with replacement. Using the same process as before for observed PNGS changes, I then calculated the differences between ancestor and descendant PNGS counts on this new randomized data.

3.4 Results

3.4.1 Indel Rates

I retrieved a total of 2,452 longitudinally-sampled gp120 sequences from LANL and the Novitsky Lab and subject them to phylogenetic reconstruction in BEAST to generate time-scaled trees describing within-host HIV-1 infections. I then reconstructed ancestral state sequences of 400 posterior BEAST trees to extract insertion and deletion events, and fit a Poisson-based hierarchical model to time and indel count data using Bayesian inference in RStan in order to estimate the insertion and deletion rates of the gp120 variable loops. Median insertion rates in terminal branches ranged between 4.3×10^{-5} and 2.0×10^{-3}, while those on internal branches were between 4.3×10^{-5} and 4.1×10^{-3} (Figure 3.2). On the other hand, median deletion rates ranged between 5.6×10^{-5} and 6.4×10^{-3} on the terminal branches, and between 4.3×10^{-5} and 9.6×10^{-3} (Figure 3.2). The rates of both insertions and deletions appeared to be highest in V1 and V5, slightly lower in V2 and V4, and substantially lower in V3. Overall, deletion rates appeared to be higher than insertion rates in V1, V2, and V5 based on the distances between the 95% confidence intervals (Figure 3.2). Rate estimates based on the internal phylogenetic tree branches also appeared higher than those based on the terminal branches in particular cases such as: indel rates in V1, indel rates in V5, and only insertion rates in V4 (Figure 3.2).

3.4.2 Indel Timings

While extracting indels from within-host phylogenies, I also collected the cumulative time-scaled branch length from the tree root to the midpoint of the branch containing each indel. I collected these time values in order to estimate the distribution of insertion and deletion timings over the course of infection, which is displayed in Figure 3.3. The first 800 days of infection sees small and large increases in insertion and deletion counts, respectively (Figure 3.3). Interestingly, deletions reach a frequency above 2.0 events/200 days/patient during this time frame, followed by a gradual decrease and stabilization at approximately 1.0 deletion/200

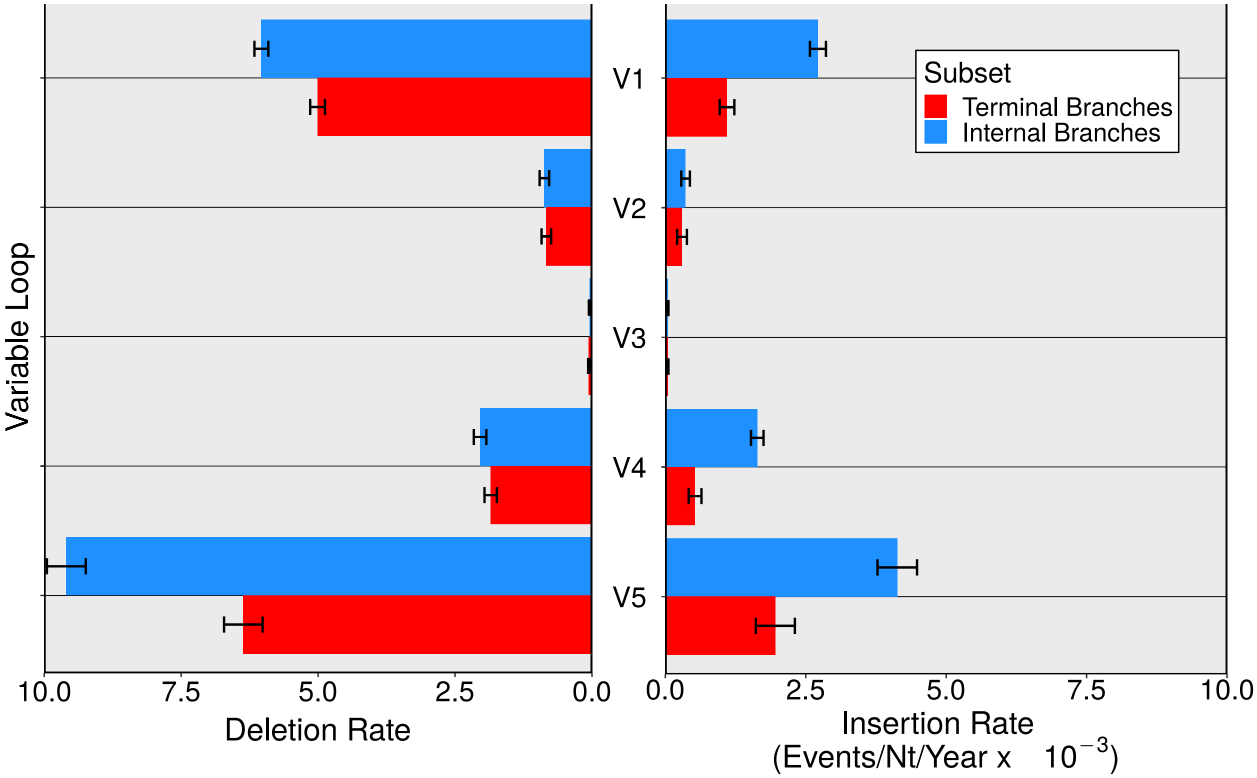

Figure 3.2: Posterior median estimates of within-host insertion and deletion rates in the five gp120 variable loops of HIV-1 subtype C. Rate estimates are further stratified based on data recovered along terminal and internal branches within intrapatient phylogenetic trees. A Poisson-based hierarchical model accounting for 400 trees per patient across 24 patients (400×24) was sampled for 10^4 Hamiltonian Monte Carlo (HMC) iterations using an RStan Bayesian inference framework. Error bars represent the 95% confidence intervals of the indel rate distribution describing HIV-1 subtype C.

days/patient (Figure 3.3). Insertions exhibit lower counts than deletions throughout infection, peaking at just over 1.0 event per 200 days in the first 400 days and subsequently decreasing throughout the rest of infection (Figure 3.3).

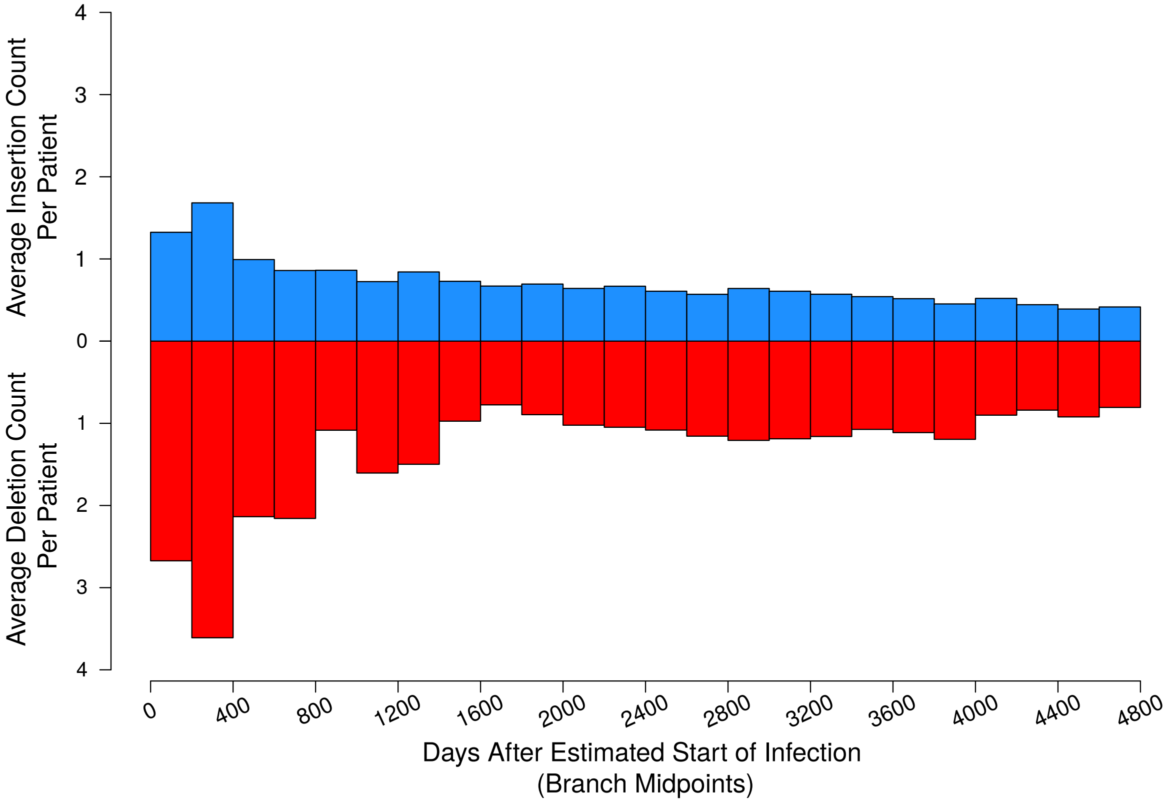

Figure 3.3: The mean counts of insertions (blue) and deletions (red) collected per patient within every 200 day interval since the estimated start of infection (tree root). Associated indel timings were based on the cumulative time-scaled edge distance from the root of the tree to the midpoint of the branch where the indel occurred. To account for the differences in infection duration, I applied a correction factor to higher time intervals as fewer datasets became available.

3.4.3 Indel Lengths

For remaining analyses, I extracted insertions and deletions from the BLC tree selected from the posterior distribution of each patient's BEAST analysis. Upon collecting these indels, I first examined trends in the length distributions of insertions and deletions as shown in Figure 3.4.

Interestingly, there were significantly elevated counts of insertions and deletions that were a) longer than 9 nucleotides in length in V1, b) 6 nucleotides in length in V2, and c) 3 nucleotide in length in V5 (Figure 3.4). Relatively long insertions (> 9 nt) were significantly elevated in V4, while long insertions and deletions were significantly less frequent in V5 (Figure 3.4). We also detected negligible counts of insertions or deletions in V3 with lengths longer than 3 nucleotides as expected. Frameshift-inducing lengths were very rare in both insertions and deletions, comprising 4.2% and 4.5% of these populations, respectively (Figure 3.4).

3.4.4 Indel Nucleotide Compositions

I then estimated the nucleotide proportions of insertions and deletions extracted from each patient's BLC tree to investigate whether indels differed in compositions from the variable loop sequences where they occur. Nucleotide compositions of both insertions and deletions showed little to no deviation from those of their variable loops of origin (Figure 3.5). Similar results were observed when investigating indel dinucleotide proportions, which also essentially matched those in their variable loops (Supplementary Figure B2). I noticed a relatively small increase in proportions of G within insertion sequences relative to the variable loops, but a randomization test did not find this increase to be significant (Figure 3.5). Further stratification of nucleotide proportions by variable loop revealed that these higher proportions of G in insertions are primarily found in V2 and V5 (Supplementary Figure B1a). Guanine proportions in insertions were also higher than those in deletions, though we did not investigate this further given the small magnitude of this difference (Figure 3.5).

3.4.5 Indel-Induced PNGS Changes

Using regular expressions, I recovered the positions of PNGSs within the variable loop sequences throughout BLC trees and using this information, determined how frequently indels induced the addition or removal of these sites. I then compared the observed indel-induced PNGS changes to those expected under stochastic processes to test the hypothesis that indels

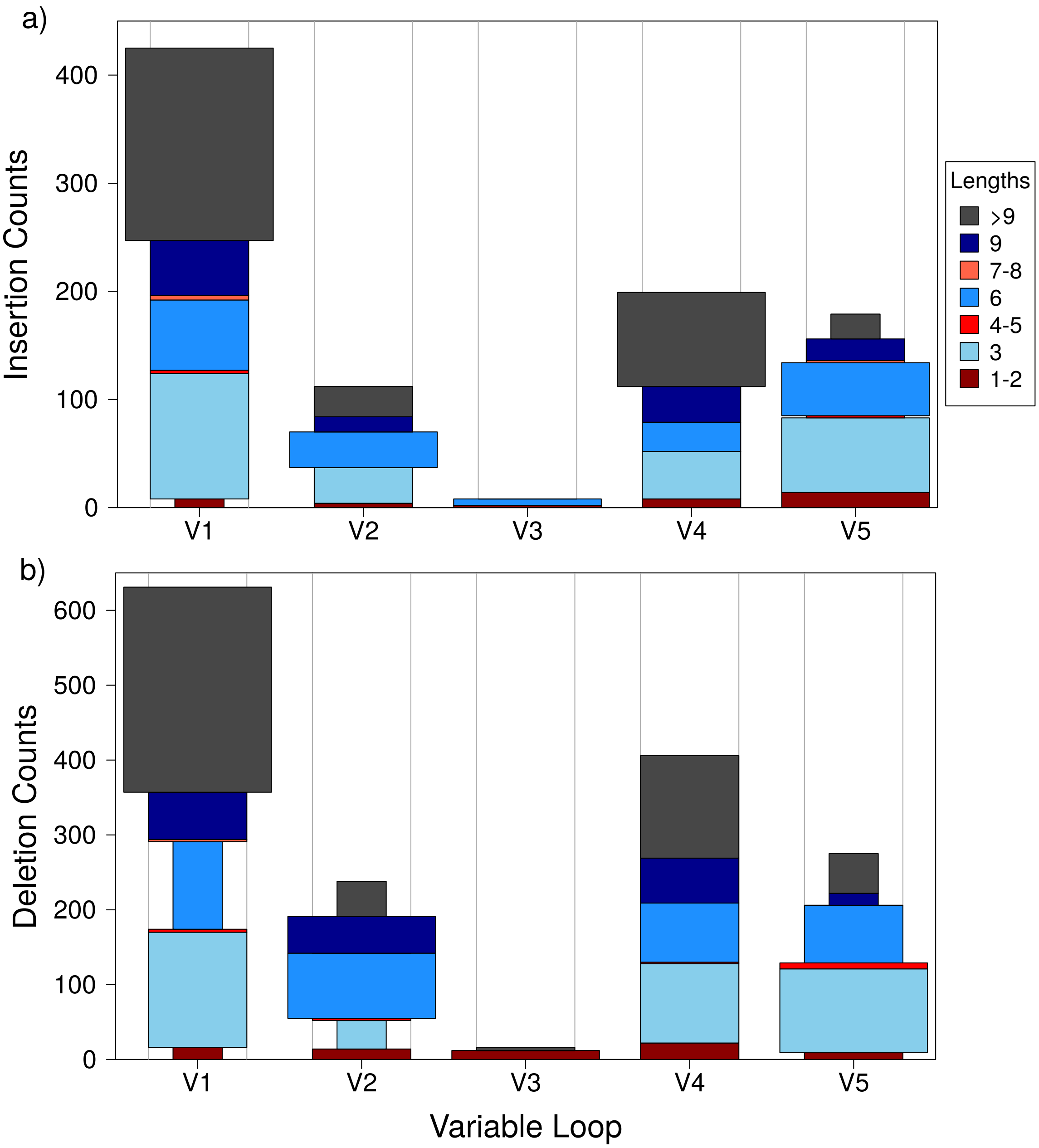

Figure 3.4: The lengths of insertions (a) and deletions (b) across the five variable loops of HIV-1 gp120. Indel counts are stratified into seven length categories: blue-colored bins describe in-frame lengths (3, 6, 9), red-colored bins describe frameshift-inducing lengths (not a multiple of 3), and dark gray bins describe very long indels (> 9 nt). Pearson chi-squared residuals — standardized differences between observed and expected counts — were generated for every length category in both plots. Bins that are wider and narrower that the outlined column margins indicate counts that are significantly higher and lower than expected, respectively, based on Pearson chi-squared residuals. Bins matching the column margins in width showed no significant differences.

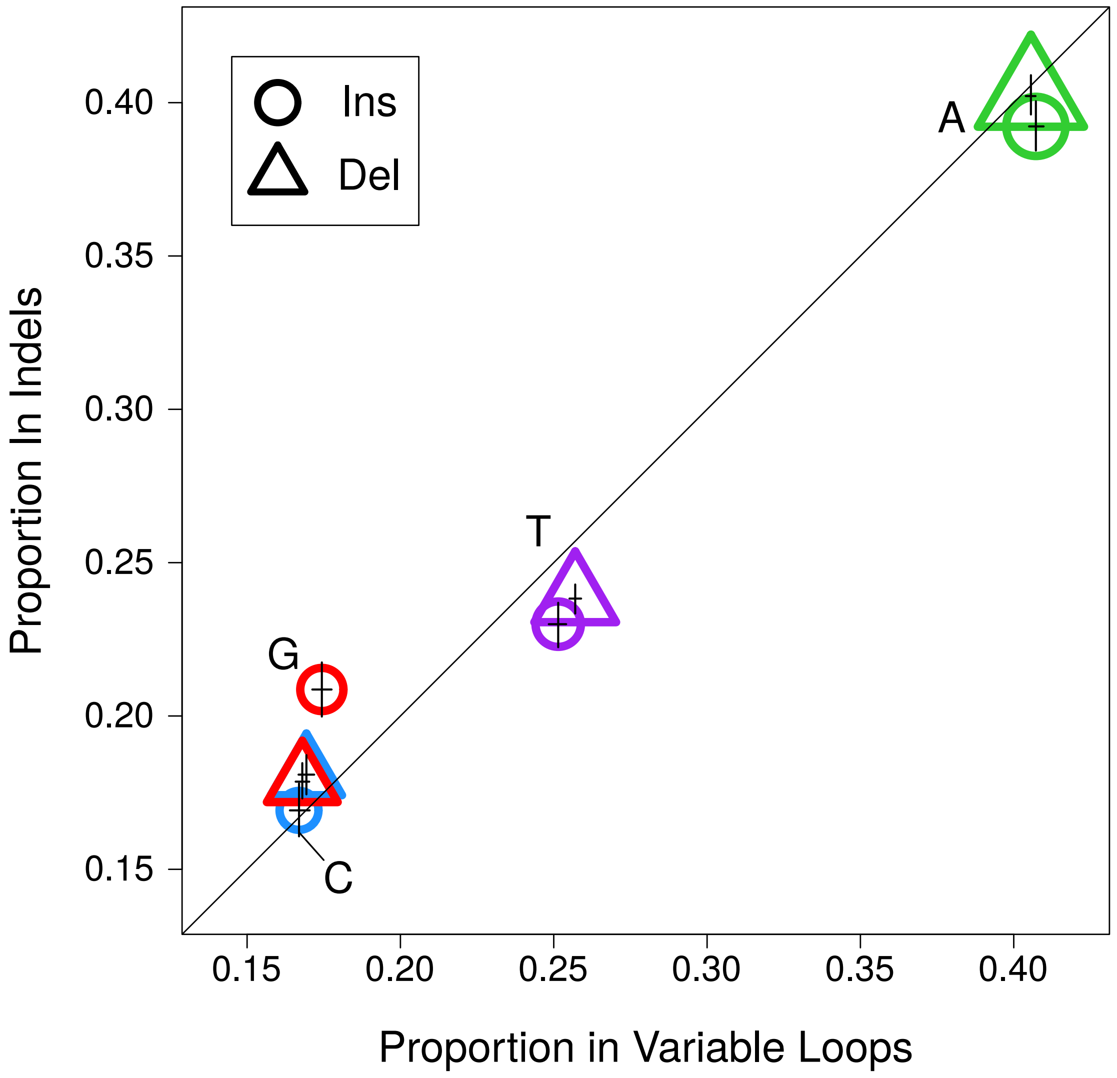

Figure 3.5: Nucleotide compositions of variable loop insertions (○) and deletions (△) relative to their surrounding non-indel loop sequence. Error bars represent the 95% confidence intervals for estimates of nucleotide proportions, acquired by regenerating nucleotide proportions in randomly sampled subsets of the data 1000 times. Point sizes are proportional to the square root of nucleotide counts detected within the given mutation event. As x and y axes are identical, points that deviate above and below the line indicate higher and lower proportions of the given nucleotide in indels, respectively. I tested for significant differences using a randomization test, which involved 1) sampling each variable loop sequence 100 times for substrings matching the size of the indel, 2) pooling these substrings together, 3) calculating the nucleotide proportions on this null distribution of variable loop samples, 4) and testing whether observed insertion/deletion nucleotide proportions were above or below the 95% quantiles of this distribution. No nucleotide proportions in either insertions or deletions registered as significantly higher or lower based on this test.

tend to add or remove PNGSs more often than expected by chance (Figure 3.6). On average, insertions in V1 and V4 increased PNGS counts in their variable loops significantly more often than expected by chance. Insertions in V2 also demonstrated a similar significant trend, albeit to a lesser extent. Specifically, insertions in V1, V2, and V4 removed 0.53, 0.23 and 0.32 PNGSs on average when placed randomly, while instead causing the addition of roughly 0.32, 0.09 and 0.40 PNGS when observed in the data (Figure 3.6). Deletions in V1 and V4 frequently removed PNGSs, though these changes were roughly comparable to those expected from random placement (Figure 3.6). Insertions in V5, and deletions in V2, V3, and V5, showed less prominent tendencies to add or remove PNGS, which also showed little difference relative to the changes generated at random. Insertion in V3 did not induce any observed changes to PNGS counts and were thus excluded from this result.

3.5 Discussion

3.5.1 Indel Rates

Here, I present the first estimates of insertion and deletion rates in the gp120 variable loops of HIV-1 subtype C measured within hosts. Our approach involved the inference of time-scaled phylogenetic trees and subsequent reconstruction of tree ancestral sequences in order to study insertions and deletions over the course of HIV-1 infection. While Mansky and Temin [4] provided an indel mutation rate of HIV-1, rates of indel evolution had not yet been studied in *env*, nor using time-scaled phylogenetic approaches prior to my study in Chapter 2 [19]. This study addressed this knowledge gap by providing the first estimates of gp120 variable loop indel rates in seven group M subtypes of HIV-1 using sequence data sampled among hosts [19]. Unlike my previous among-host study, this study estimates insertion and deletion rates independently of each other and utilizes indel events retrieved throughout the entire phylogenetic tree, instead of just the tip sequences [19]. I hypothesized that within-host insertion and deletion rates would be higher than those measured among hosts due to lessened purifying selection that has acted

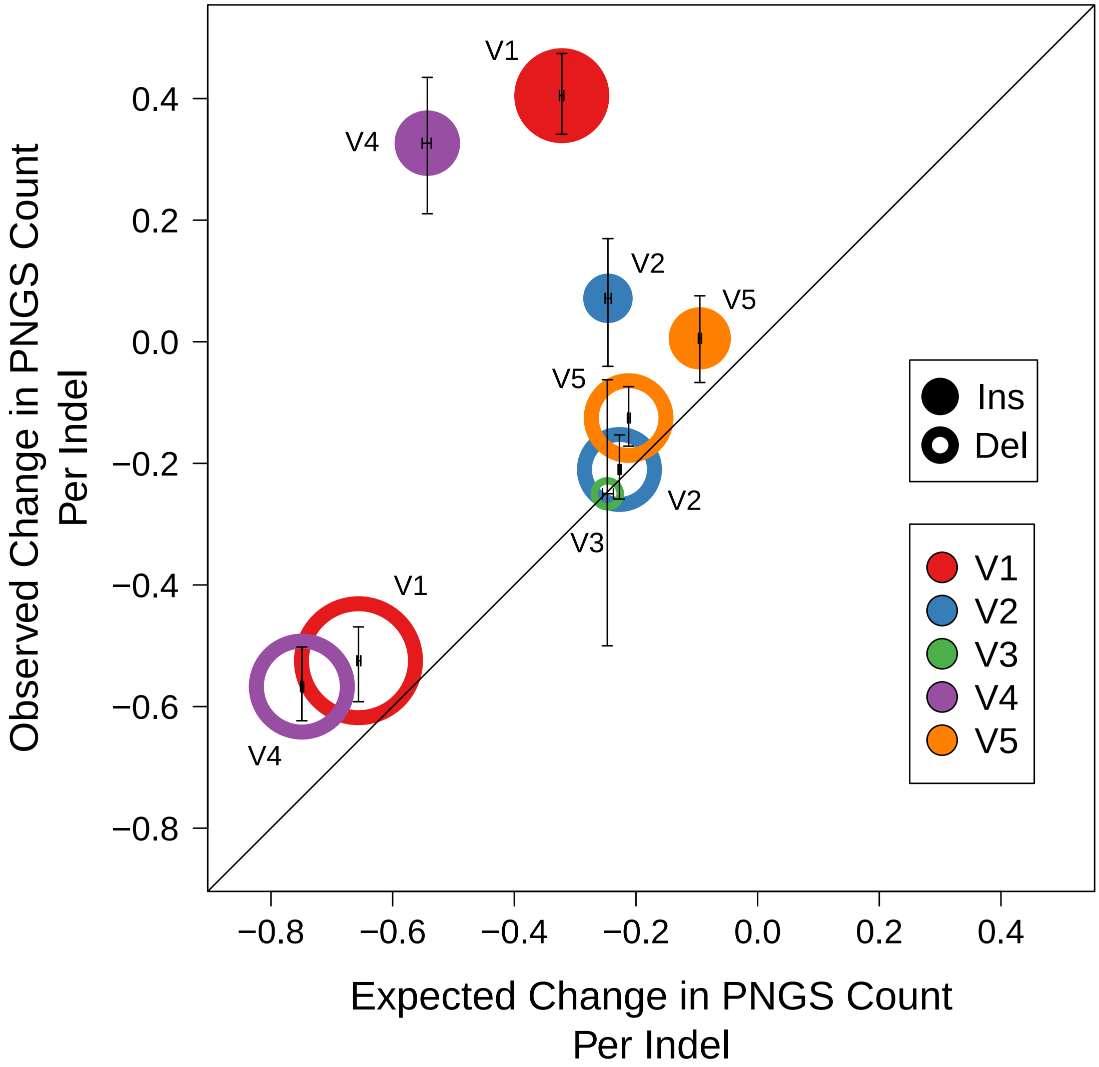

Figure 3.6: Changes to the counts of potential N-linked glycosylation sites (PNGSs) in the variable loops induced by insertions and deletions. On the x axis, I estimated changes to PNGS counts expected by chance by placing every indel back into its variable loop sequence of origin 500 times in randomized locations. The y axis on the other hand, describes the changes to PNGS counts observed in the collected data. Given the identical axes, points above and below the line indicate that indels from specific variable loops tend to increase and decrease the PNGS counts more often than expected by chance, respectively. Error bars represent the 95% confidence intervals acquired by randomly sampling from the final distribution of observed and expected changes 1000 times. Point sizes are scaled to the square root of the number of indel events in the given category. No estimates could be generated for the insertions of V3 due to insufficient data.

on these sequence data. I found support for my hypothesis, as within-host indel rates in subtype C appeared considerably higher in V1, V2, V4, and V5 than those measured among hosts given the distinct separation from the 95% confidence intervals (Figure 3.2, Figure 2.2). Additionally, I noticed that both insertion and deletion rates in this study exhibit the same trends across variable loops as in my among-host study: indel rates in V5 and V1 were relatively high, those in V2 and V4 were relatively moderate, and rates in V3 were relatively low (Figure 3.2, Figure 2.2). This almost identical trend in indel rates among variable loops does not offer support for our hypothesis that within-host sequence data has undergone far less purifying selection. If we were looking past purifying selection, we would expect to see higher indel rates in loops intolerant to mutation (*i.e.* V3) due to the removal of this strong filtering effect, and even higher rates in V1 and V2 due to their heavy involvement in immune escape [34, 66].

When comparing deletions to insertions, deletion rates in V1, V2, and V5 appear significantly higher than insertion rates based on the distinct separation of the 95% confidence intervals (Figure 3.2). Interestingly, Cheynier et al. [67] reported collecting four times as many deletions than insertions in V1 of a strain of SIV, showing a similar trend to the ratio between our deletion and insertion rates in this variable loop (5.0:1.1 in terminal branches, and 6.0:2.7 in internal branches; Figure 3.2) . This suggests that SIV may hold similarities regarding the accumulation of insertions and deletions in the variable loops of its surface protein. There are also interesting implications of higher deletion rates (Figure 3.2) when considering the similar length distributions of insertions and deletions (Figure 3.4). Based on the trends displayed in Figure 3.4, insertions and deletions appear to be roughly the same length on average, meaning that these findings together suggest that the variable loops are in fact shrinking over time. A possible explanation for this phenomenon is the selection of shorter variable loops to enhance the efficiency of HIV-1 binding to susceptible immune cells during acute infection [68]. Moreover, the patients examined in this study may have severely weakened immune systems which would further exacerbate this trend [66]. A weak immune system would exert minimal selective pressure on the gp120 variable loops to lengthen, causing them to instead shrink due to the fitness advantages associated with more efficient fusion [66].

When stratifying indel rates by internal and terminal branches, we expected to find higher indel rates in the terminal branches on average based on our hypothesis. We hypothesized that our within-host sequence data has undergone relatively little purifying selection and on this basis, terminal branches should contain additional indels from malfunctioning or non-functional viruses that have not been filtered out by purifying selection. Internal branches on the other hand, are required to have successfully replicated in order to be sampled, likely reducing the potential to detect indels. However, we found higher indel rates in the internal tree branches relative to those in the terminal branches in V1, V5, and to a limited extent in V4, which contradicted our expectations. As discussed previously with deletion rates, this could be explained by selective pressures driving the accumulation of indels during early stages of infection when the HIV-1 population undergoes more rapid evolution [13]. Alternatively, in a population dynamics context, this could be caused by the HIV-1 population undergoing periods of exponential growth during early infection, which would present as very short internal branches in our time-scaled trees that may overestimate rates [69]. In fact, I observed this exact trend in my finalized trees sampled from the posterior of BEAST analysis, supporting this as a possible cause for the elevated internal branch indel rates. While these offer plausible biological explanations for our findings, we also cannot exclude the possibility of intrinsic bias towards shorter internal branch lengths introduced by RAxML and BEAST.

3.5.2 Indel Timings

Using the same posterior trees, I also estimated the time at which variable loop insertions and deletions occurred using the cumulative time-scale branch lengths of the phylogenetic trees. The slight and large increases in variable loop insertion and deletion counts, respectively, during the first 1500 days of infection shows concordance with existing knowledge of the evolutionary processes occurring during this time period of HIV-1 infection. The asymptomatic phase of HIV-1 infection often starts 2 to 10 weeks after the start of HIV-1 infection (< 100 days) and typically lasts 7 to 10 years (2550 - 3650 days) [8, 9]. Recall that HIV-1 undergoes substantial adaptive evolution during this phase, which involves the positive selection of immune escape variants [8, 9, 30]. In context of HIV-1 infection timings, these results appear to support the notion that variable loop deletions, and insertions to a lesser extent, are involved in the adaptive evolution of HIV-1 in the asymptomatic phase (100 - 2500 days), possibly through the generation of immune escape variants (Figure 3.3) [34, 45, 47]. Lee et al. [13] has demonstrated that HIV-1 substitution rates undergo a similar trend to indel frequencies in Figure 3.3, in that these rates often decline in the later years of HIV-1 infection (i.e. 4-10 years) when CD4$^+$ cell counts decrease in number. The concordance in these findings suggest that indel rates are correlated with substitution rates and possibly even correlated with CD4$^+$ cell count, though further research will be needed to confirm this [13]. These findings however, still have notable uncertainty and may in fact, be artifacts generated by biases in phylogenetic reconstruction, as mentioned previously.

3.5.3 Indel Lengths

Using the single BLC from each patient's BEAST analysis, I then examined trends in insertion and deletion lengths. I found remarkably few cases of frameshift-inducing insertion (4.2%) and deletion (4.5%) lengths within hosts, which suggest that purifying selection exerts substantial effects within hosts (Figure 3.4). Interestingly, I found a relatively and significantly increased proportion of 6-nucleotide insertions in V3, which is surprising given the stringent conservation of this loop due to its functional involvement in coreceptor binding [70]. Aside from this, the seeming intolerance of insertions or deletions longer than 6 nucleotides in this loop still aligned with our expectations and corresponded with its previously established conserved nature [70]. The significantly elevated counts of long insertions (> 9 nt) in V1 corresponds with previous reports that the V1V2 loop lengthens and gains PNGS over the course of infection in response to neutralizing antibody responses (Figure 3.4) [27, 45, 46]. The length of these insertions matches the length increases reported in the V1V2 loop structure and importantly, are also of sufficient length to add entire PNGSs to the loop sequence (3-5 amino acids) (Figure 3.4) [27, 45, 71]. It is however, surprising that the same abundance of long insertions is not observed in V2, given that the same trends are also expected to affect this loop (Figure 3.4) [45, 71]. Fascinatingly, we see a remarkable number of long deletions in V1 that even exceeds the number of long insertions in this variable loop. To reconcile this finding with the trends of V1V2 growth during infection, I postulate that these large deletions could occur relatively early in infection as shown in Figure 3.3, or undergo removal or reversion driven by purifying selection, thereby allowing V1V2 to lengthen later in infection using insertions. Furthermore, the remarkably lower tolerance for long deletions (> 9 nt) observed in V2 relative to V1 is surprising, given that it is the longest variable region (120 nt) and has been reported to accumulate numerous indels throughout infection [34, 35, 45, 71].

3.5.4 Indel Nucleotide Compositions

I also found that the nucleotide compositions of both insertions and deletions essentially matched those found in their variable loops of origin (Figure 3.5). These findings loosely suggest that the nucleotides comprising insertions originate from somewhere in the HIV-1 genome, given the distinctly high proportions of A and low proportions of C found in both sources (Figure 3.5) [72]. Regarding deletions, these findings suggest that there are no relative differences in the fitness costs of maintaining each of four nucleotides in the variable loop sequences, since all nucleotides have a roughly equal propensity to be deleted (Figure 3.5). The slightly higher proportions of A (~ 0.40) and lower proportions of G (~ 0.17) relative to those found in both the subtype C consensus sequence (GenBank Accession: U46016) and reports by van der Kuyl and Berkhout [72] (A: 0.36, G: 0.24) suggest that some G to A hypermutation was undetected by our Hypermut analysis screen [72].

3.5.5 Indel-Induced PNGS Changes

I found that insertions in V1, V4, and to a lesser extent, V2 tended to create new PNGSs, and also demonstrated a tendency to do so more frequently than expected by chance (Figure 3.6). These findings align with previous reports that indels in V1, V2 and V4 tend to add and change the positions of PNGS to generate glycan shield variants to escape neutralizing antibody responses over the course of infection [27, 37, 45, 46]. Moreover, the stark differences in observed and expected changes in V1, V2, and V4 suggest that insertions are under strong selection to avoid disrupting existing PNGS and instead add new sites (Figure 3.6). Deletions in V1 and V4 demonstrate tendencies to remove PNGSs, but unlike insertions, show minimal signs of selection acting on their placement given that they induce changes that align with stochastic processes (Figure 3.6). These results also contextualize our findings of significantly elevated proportions of long insertions in V1 and V4 (Figure 3.4), suggesting that these PNGS changes are primarily driven by relatively long insertions that may add partial or entire PNGSs at a time [27, 37, 45].

3.5.6 Limitations

An important limitation to consider in this study is our constraint of tree topology during BEAST analysis. Different tree topologies, which would have been explored in typical BEAST analyses, may change the distributions of indel counts and timings associated with tree branches, both of which are used in the estimation of indel rates. Therefore, searching multiple tree topologies favoured by posterior probabilities would provide a more comprehensive estimation of indel rate variation and uncertainty. However, recall that there was insufficient molecular clock signal in my data to generate sampling convergence for the posterior distributions in BEAST in this more complex tree parameter space. Since posterior convergence is essential to consider BEAST outcomes valid, we deemed constraining the tree topology necessary in order to generate working results. At the same time, this constraint of topology is likely causing an underestimation of the uncertainty surrounding indel rates and may further be systematically biasing these estimates in some way.

3.5.7 Conclusions

This study contributes to the understanding of indels in the gp120 variable loops on a within-host scale, which comprise an important evolutionary mechanism driving HIV-1 adaptation through immune escape. We found that indel rate estimates appeared higher on a within-host scale than those estimated among hosts in a previous study, suggesting support for our hypothesis (Figure 3.2) [19]. However, further investigation into the caveat of constraining tree topologies during BEAST analysis will be needed to confirm this finding. Interestingly, we find rare instances of frameshift-inducing insertion and deletion lengths (4.2%, 4.5% respectively) in our within-host analyses, demonstrating that purifying selection is highly prevalent within hosts (Figure 3.4). Nucleotide compositions of insertions and deletions were essentially indistinguishable from those recorded in the surrounding variable loop sequences, loosely suggesting that insertions do not have a distinctly different origin and that nucleotides are not being selectively deleted (Figure 3.5). Finally, we find that insertions in V1, V2, and V4 appear to undergo selection for those that introduce new PNGS sites, while deletions tend to remove PNGS in a rather stochastic manner (Figure 3.6).

By estimating indel rates, we provide a quantified measure of the evolution introduced by indels in the variable loops, which was previously unknown and can be used in future comparative studies. Given the suggested role of variable loop indels in immune escape, indel rates may also correlate with HIV-1 disease progression or CD4$^+$ cell decline, in a manner similar to substitution rates [13, 41, 42]. Future research on indel rates however, will be needed to investigate these putative correlations. We hope that our quantification of variable loop indel rates will facilitate further research into HIV-1 indel rates and lead to the further incorporation of indels into evolutionary analyses.

Bibliography

[1] Michael E Abram, Andrea L Ferris, Wei Shao, W Gregory Alvord, and Stephen H Hughes. Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *Journal of virology*, 84(19):9864–9878, 2010.

[2] Richard A Neher and Thomas Leitner. Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol*, 6(1):e1000660, 2010.

[3] Alan S Perelson, Avidan U Neumann, Martin Markowitz, John M Leonard, and David D Ho. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science*, 271(5255):1582–1586, 1996.

[4] Louis M Mansky and Howard M Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, 69(8):5087–5094, 1995.

[5] Simon DW Frost, Marie-Jeanne Dumaurier, Simon Wain-Hobson, and Andrew J Leigh Brown. Genetic drift and within-host metapopulation dynamics of HIV-1 infection. *Proceedings of the National Academy of Sciences*, 98(12):6975–6980, 2001.

[6] Andreas Meyerhans, Rémi Cheynier, Jan Albert, Martina Seth, Shirley Kwok, John Sninsky, Linda Morfeldt-Månson, Birgitta Asjö, and Simon Wain-Hobson. Temporal fluctuations in HIV quasispecies in vivo are not reflected by sequential HIV isolations. *Cell*, 58 (5):901–910, 1989.

[7] Li Yin, Li Liu, Yijun Sun, Wei Hou, Amanda C Lowe, Brent P Gardner, Marco Salemi, Wilton B Williams, William G Farmerie, John W Sleasman, et al. High-resolution deep sequencing reveals biodiversity, population structure, and persistence of HIV-1 quasispecies within host ecosystems. *Retrovirology*, 9(1):1–9, 2012.

[8] Gang Huang, Yasuhiro Takeuchi, and Andrei Korobeinikov. HIV evolution and progression of the infection to AIDS. *Journal of theoretical biology*, 307:149–159, 2012.

[9] Andrew Rambaut, David Posada, Keith A Crandall, and Edward C Holmes. The causes and consequences of HIV evolution. *Nature Reviews Genetics*, 5(1):52–61, 2004.

[10] Esteban A Hernandez-Vargas and Richard H Middleton. Modeling the three stages in HIV infection. *Journal of theoretical biology*, 320:33–40, 2013.

[11] Tae-Wook Chun, Lucy Carruth, Diana Finzi, Xuefei Shen, Joseph A DiGiuseppe, Harry Taylor, Monika Hermankova, Karen Chadwick, Joseph Margolick, Thomas C Quinn, et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. *Nature*, 387(6629):183–188, 1997.

[12] Esteban Domingo, Cristina Escarmís, Luis Menéndez-Arias, and John J Holland. Viral quasispecies and fitness variations. In *Origin and evolution of viruses*, pages 141–161. Elsevier, 1999.

[13] Ha Youn Lee, Alan S Perelson, Su-Chan Park, and Thomas Leitner. Dynamic correlation between intrahost HIV-1 quasispecies evolution and disease progression. *PLoS Comput Biol*, 4(12):e1000240, 2008.

[14] Caroline F Wright, Marco J Morelli, Gaël Thébaud, Nick J Knowles, Pawel Herzyk, David J Paton, Daniel T Haydon, and Donald P King. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *Journal of virology*, 85(5):2266–2275, 2011.

[15] Vincent Montoya, Andrea Olmstead, Patrick Tang, Darrel Cook, Naveed Janjua, Jason Grebely, Brendan Jacka, Art FY Poon, and Mel Krajden. Deep sequencing increases hepatitis C virus phylogenetic cluster detection compared to Sanger sequencing. *Infection, Genetics and Evolution*, 43:329–337, 2016.

[16] Carlos Y Valenzuela, Sergio V Flores, and Javier Cisternas. Fixations of the HIV-1 env gene refute neutralism: new evidence for pan-selective evolution. *Biological research*, 43 (2):149–163, 2010.

[17] CTT Edwards, EC Holmes, OG Pybus, DJ Wilson, RP Viscidi, EJ Abrams, RE Phillips, and AJ Drummond. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics*, 174(3):1441–1453, 2006.

[18] Natasha Wood, Tanmoy Bhattacharya, Brandon F Keele, Elena Giorgi, Michael Liu, Brian Gaschen, Marcus Daniels, Guido Ferrari, Barton F Haynes, Andrew McMichael, et al. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS pathogens*, 5(5):e1000414, 2009.

[19] John Palmer and Art FY Poon. Phylogenetic measures of indel rate variation among the HIV-1 group M subtypes. *Virus evolution*, 5(2):vez022, 2019.

[20] David A Price, Philip JR Goulder, Paul Klenerman, Andrew K Sewell, Philippa J Easterbrook, Maxine Troop, Charles RM Bangham, and Rodney E Phillips. Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proceedings of the National Academy of Sciences*, 94(5):1890–1895, 1997.

[21] Rasmus Nielsen and Ziheng Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–936, 1998.

[22] Lamei Chen, Alla Perlina, and Christopher J Lee. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug re-

sistance and positive fitness mutations in HIV protease and reverse transcriptase. *Journal of virology*, 78(7):3722–3732, 2004.

[23] Samuel Alizon and Christophe Fraser. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology*, 10(1):49, 2013.

[24] Sarah B Joseph, Ronald Swanstrom, Angela DM Kashuba, and Myron S Cohen. Bottlenecks in HIV-1 transmission: insights from the study of founder viruses. *Nature reviews Microbiology*, 13(7):414–425, 2015.

[25] W. H. Li, M. Tanimura, and P. M. Sharp. Rates and dates of divergence between AIDS virus nucleotide sequences. *Molecular Biology and Evolution*, 5(4):313–330, July 1988. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040503. URL https://academic.oup.com/mbe/article/5/4/313/1026948.

[26] Marc Choisy, Christopher H Woelk, Jean-François Guégan, and David L Robertson. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *Journal of Virology*, 78(4):1962–1970, 2004.

[27] Cynthia A Derdeyn, Julie M Decker, Frederic Bibollet-Ruche, John L Mokili, Mark Muldoon, Scott A Denham, Marintha L Heil, Francis Kasolo, Rosemary Musonda, Beatrice H Hahn, et al. Envelope-constrained neutralization-sensitive HIV-1 after heterosexual transmission. *Science*, 303(5666):2019–2022, 2004.

[28] Douglas D Richman, Terri Wrin, Susan J Little, and Christos J Petropoulos. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. *Proceedings of the National Academy of Sciences*, 100(7):4144–4149, 2003.

[29] Xiping Wei, Julie M Decker, Shuyi Wang, Huxiong Hui, John C Kappes, Xiaoyun Wu, Jesus F Salazar-Gonzalez, Maria G Salazar, J Michael Kilby, Michael S Saag, et al. Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307, 2003.

[30] Andrew J McMichael and Sarah L Rowland-Jones. Cellular immune responses to HIV. *Nature*, 410(6831):980–987, 2001.

[31] BA Watkins, MS Reitz, CA Wilson, K Aldrich, AE Davis, and M Robert-Guroff. Immune escape by human immunodeficiency virus type 1 from neutralizing antibodies: evidence for multiple pathways. *Journal of virology*, 67(12):7493–7500, 1993.

[32] ES Gray, PL Moore, IA Choge, JM Decker, F Bibollet-Ruche, H Li, N Leseka, F Treurnicht, K Mlisana, GM Shaw, et al. Neutralizing antibody responses in acute human immunodeficiency virus type 1 subtype C infection. *Journal of virology*, 81(12):6187–6196, 2007.

[33] Penny L Moore, Elin S Gray, Isaac A Choge, Nthabeleng Ranchobe, Koleka Mlisana, Salim S Abdool Karim, Carolyn Williamson, Lynn Morris, et al. The c3-v4 region is a major target of autologous neutralizing antibodies in human immunodeficiency virus type 1 subtype C infection. *Journal of virology*, 82(4):1860–1869, 2008.

[34] Penny L Moore, Nthabeleng Ranchobe, Bronwen E Lambson, Elin S Gray, Eleanor Cave, Melissa-Rose Abrahams, Gama Bandawe, Koleka Mlisana, Salim S Abdool Karim, Carolyn Williamson, et al. Limited neutralizing antibody specificities drive neutralization escape in early HIV-1 subtype C infection. *PLoS Pathog*, 5(9):e1000598, 2009.

[35] Simon DW Frost, Terri Wrin, Davey M Smith, Sergei L Kosakovsky Pond, Yang Liu, Ellen Paxinos, Colombe Chappey, Justin Galovich, Jeff Beauchaine, Christos J Petropoulos, et al. Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection. *Proceedings of the National Academy of Sciences*, 102(51):18514–18519, 2005.

[36] Alan R Templeton, Rebecca A Reichert, Anton E Weisstein, Xiao-Fang Yu, and Richard B Markham. Selection in context: patterns of natural selection in the glycopro-

tein 120 region of human immunodeficiency virus 1 within infected individuals. *Genetics*, 167(4):1547–1561, 2004.

[37] Erika Castro, Manon Bélair, Gian Paolo Rizzardi, Pierre A. Bart, Giuseppe Pantaleo, and Cecilia Graziosi. Independent evolution of hypervariable regions of HIV-1 gp120: V4 as a swarm of N-Linked glycosylation variants. *AIDS research and human retroviruses*, 24 (1):106–113, January 2008. ISSN 0889-2229. doi: 10.1089/aid.2007.0139.

[38] Vlad Novitsky, Rui Wang, Raabya Rossenkhan, Sikhulile Moyo, and M Essex. Intra-host evolutionary rates in HIV-1C env and gag during primary infection. *Infection, Genetics and Evolution*, 19:361–368, 2013.

[39] Michael J Dapp, Kord M Kober, Lennie Chen, Dylan H Westfall, Kim Wong, Hong Zhao, Breana M Hall, Wenjie Deng, Thomas Sibley, Suvankar Ghorai, et al. Patterns and rates of viral evolution in HIV-1 subtype B infected females and males. *PloS one*, 12(10): e0182443, 2017.

[40] Philippe Lemey, Andrew Rambaut, and Oliver G Pybus. HIV evolutionary dynamics within and among hosts. *AIDs Rev*, 8(3):125–140, 2006.

[41] José M Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely high mutation rate of HIV-1 in vivo. *PLoS Biol*, 13(9):e1002251, 2015.

[42] RAJ Shankarappa, Joseph B Margolick, Stephen J Gange, Allen G Rodrigo, David Upchurch, Homayoon Farzadegan, Phalguni Gupta, Charles R Rinaldo, Gerald H Learn, XI He, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *Journal of virology*, 73(12):10489–10502, 1999.

[43] Art FY Poon, Fraser I Lewis, Sergei L Kosakovsky Pond, and Simon DW Frost. An evolutionary-network model reveals stratified interactions in the V3 loop of the HIV-1 envelope. *PLoS Comput Biol*, 3(11):e231, 2007.

[44] Abayomi S Olabode, Mariano Avino, Tammy Ng, Faisal Abu-Sardanah, David W Dick, and Art FY Poon. Evidence for a Recombinant Origin of HIV-1 group M from Genomic Variation. *bioRxiv*, page 364075, 2018.

[45] Manish Sagar, Xueling Wu, Sandra Lee, and Julie Overbaugh. Human immunodeficiency virus type 1 V1-V2 envelope loop sequences expand and add glycosylation sites over the course of infection, and these modifications affect antibody neutralization sensitivity. *Journal of virology*, 80(19):9586–9598, 2006.

[46] Marit J van Gils, Evelien M Bunnik, Brigitte D Boeser-Nunnink, Judith A Burger, Marijke Terlouw-Klein, Naomi Verwer, and Hanneke Schuitemaker. Longer V1V2 region with increased number of potential N-linked glycosylation sites in the HIV-1 envelope glycoprotein protects against HIV-specific neutralizing antibodies. *Journal of virology*, pages JVI–00268, 2011.

[47] Rong Rong, Frederic Bibollet-Ruche, Joseph Mulenga, Susan Allen, Jerry L Blackwell, and Cynthia A Derdeyn. Role of V1V2 and other human immunodeficiency virus type 1 envelope domains in resistance to autologous neutralization during clade C infection. *Journal of virology*, 81(3):1350–1359, 2007.

[48] Manish Sagar, Oliver Laeyendecker, Sandra Lee, Jordyn Gamiel, Maria J Wawer, Ronald H Gray, David Serwadda, Nelson K Sewankambo, James C Shepherd, Jonathan Toma, et al. Selection of HIV variants with signature genotypic characteristics during heterosexual transmission. *The Journal of infectious diseases*, 199(4):580–589, 2009.

[49] Stephen F Altschul and Bruce W Erickson. Optimal sequence alignment using affine gap costs. *Bull Math Biol*, 48(5-6):603–616, 1986.

[50] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[51] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[52] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35:526–528, 2018.

[53] Ann M Sheehy, Nathan C Gaddis, Jonathan D Choi, and Michael H Malim. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–650, 2002.

[54] Patrick P Rose and Bette T Korber. Detecting hypermutations in viral sequences with an emphasis on G→ A hypermutation. *Bioinformatics*, 16(4):400–401, 2000.

[55] M-R Abrahams, Jeffrey A Anderson, EE Giorgi, Cathal Seoighe, K Mlisana, L-H Ping, GS Athreya, Florette K Treurnicht, Brandon F Keele, N Wood, et al. Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *Journal of virology*, 83(8):3556–3567, 2009.

[56] Hongshuo Song, Elena E Giorgi, Vitaly V Ganusov, Fangping Cai, Gayathri Athreya, Hyejin Yoon, Oana Carja, Bhavna Hora, Peter Hraber, Ethan Romero-Severson, et al. Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nature communications*, 9(1):1–15, 2018.

[57] Alexei J Drummond and Andrew Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7(1):214, 2007.

[58] Koichiro Tamura and Masatoshi Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution*, 10(3):512–526, 1993.

[59] Ziheng Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular biology and evolution*, 10(6):1396–1401, 1993.

[60] Alexei J Drummond, Simon YW Ho, Matthew J Phillips, and Andrew Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4(5):e88, 2006.

[61] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[62] Masatoshi Nei and Naoyuki Takahata. Effective population size, genetic diversity, and coalescence time in subdivided populations. *Journal of Molecular Evolution*, 37(3):240–244, 1993.

[63] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[64] Ian H Holmes. Historian: accurate reconstruction of ancestral sequences and evolutionary rates. *Bioinformatics*, 33(8):1227–1229, 2017.

[65] David Meyer, Achim Zeileis, and Kurt Hornik. *vcd: Visualizing Categorical Data*, 2020. R package version 1.4-7.

[66] Marcel E Curlin, Rafael Zioni, Stephen E Hawes, Yi Liu, Wenjie Deng, Geoffrey S Gottlieb, Tuofu Zhu, and James I Mullins. HIV-1 envelope subregion length variation during disease progression. *PLoS pathogens*, 6(12):e1001228, 2010.

[67] Rémi Cheynier, Laurens Kils-Hütten, Andreas Meyerhans, and Simon Wain-Hobson. Insertion/deletion frequencies match those of point mutations in the hypervariable regions of the simian immunodeficiency virus surface envelope gene. *Journal of General Virology*, 82(7):1613–1619, 2001.

[68] Marielle Cavrois, Jason Neidleman, Mario L Santiago, Cynthia A Derdeyn, Eric Hunter, and Warner C Greene. Enhanced fusion and virion incorporation for hiv-1 subtype c envelope glycoproteins with compact v1/v2 domains. *Journal of Virology*, 88(4):2083–2094, 2014.

[69] Alexei J Drummond, Andrew Rambaut, BETH Shapiro, and Oliver G Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192, 2005.

[70] Xunqing Jiang, Valicia Burke, Maxim Totrov, Constance Williams, Timothy Cardozo, Miroslaw K Gorny, Susan Zolla-Pazner, and Xiang-Peng Kong. Conserved structural elements in the V3 crown of HIV-1 gp120. *Nature Structural and Molecular Biology*, 17 (8):955, 2010.

[71] Bhavna Chohan, Dorothy Lang, Manish Sagar, Bette Korber, Ludo Lavreys, Barbra Richardson, and Julie Overbaugh. Selection for human immunodeficiency virus type 1 envelope glycosylation variants with shorter V1-V2 loop sequences occurs during transmission of certain genetic subtypes and may impact viral RNA levels. *Journal of virology*, 79(10):6528–6531, 2005.

[72] Antoinette C van der Kuyl and Ben Berkhout. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. *Retrovirology*, 9(1):92, 2012.

Chapter 4

Insertion Modelling

4.1 Background

In this chapter, I will describe the future directions for my phylogenetic analyses on indels, as well as my current work on these ideas to date. Having extracted numerous insertion and deletion events from the variable regions of patient-derived gp120 sequences, we wanted to take further advantage of this opportunity by building an empirical model that effectively describes these indels. An empirical model is a simplified statistical representation of a real-world process or mechanism which is importantly, based on observed data [3, 4]. At its core, an empirical model is a set mathematical relationships between one or more variables that capture a set of statistical assumptions and as a whole, form an overarching hypothesis about the system [4]. Being a pragmatic approximation of reality, an optimized empirical model can recreate the outcomes of relatively complex mechanisms and contribute to a more fundamental understanding of the modelled system. Thus, modeling of insertions and deletions has potential to improve understanding of the underlying mechanisms that generate these mutational processes and their contributions to sequence evolution.

Insertions and deletions are capable of inducing more genetic change on a per-nucleotide scale than substitution events and are therefore, suggested to have a crucial role in sequence evolution [5]. Despite their implicated importance, indels have been investigated and utilized to a lesser extent than nucleotide substitutions, likely due in part to the inherent difficulty surrounding their reconstruction [1]. This difficulty stems from the fact that indels cannot be directly observed like nucleotide substitutions; their positions can only be inferred, which leads to greater uncertainty [1, 11]. For example, while substitutions form the basis of evolutionary models and have been studied extensively over the last five decades, the use of indels for evolutionary analysis remains a newly expanding area of research, though some groups have made promising recent advances [5, 7, 8, 12]. This demonstrates the need for additional research on indels in multiple areas which importantly, is not solely limited to their application to evolutionary models. This includes research into the underlying mechanisms that control indel generation in molecular sequences, which can be facilitated through the empirical modelling and includes my modelling efforts here.

Generally, there are two different hypothesized mechanisms of indel generation in molecular sequences. For one, the strand slippage mechanism involves the shift of the template or copying strand during replication, leading to their misalignment and the misincorporation of one or more bases [6]. A slip in the template will cause one or more nucleotides to form a loop which cannot be recognized by the polymerase enzyme, leading to the deletion of base pairs in the loop [6]. Contrarily, if the copying strand slips backward to form its own loop, it will copy redundant nucleotides and result in the insertion of new base pairs [6]. A second proposed mechanism, which has been proposed to occur frequently in HIV-1, occurs during the recombination between two genetic sequences [2, 9, 10]. Recall that in this process, a polymerase enzyme will dissociate from its attached template in the middle of replication, attach to another nearby template and continue copying on the new template [2, 9, 10]. If the polymerase does not attach to the new strand at the same location that it was copying on the original strand, it can lead to the deletion or insertion of nucleotides if it attaches in front or behind its original position, respectively [2, 9].

Here, I describe my work developing an empirical model describing indels in the variable loops of HIV-1 gp120. To accomplish this, I used Bayesian inference to fit various model structures and parameters to the indel data collected during my within-hosts study (Chapter 3). By testing, comparisons, and refining different models in this framework, I hypothesize that the trends in indel lengths, compositions, and frequencies can be recreated and different hypothetical indel generation mechanisms can be compared. In its current state, this chapter constitutes my efforts to model the strand slippage mechanism pertaining solely to insertions. I chose to focus on modeling insertions first due to a unique set of findings displayed in Figure 4.1. Interestingly, I found that both insertions and deletions were frequently flanked by an identical, or mostly identical, sequence immediately upstream or downstream of the event (Figure 4.1). Insertions followed this trend to a greater extent than deletions did, with 20.6% of recovered insertions having an adjacent exact match (Figure 4.1a) and 36.6% having a more lenient match where 1 in 9 nucleotides are allowed to differ (Figure 4.1c). Given these findings and being relatively new to empirical modeling, I decided that modeling insertions as a product of replication slippage was an appropriate starting point.

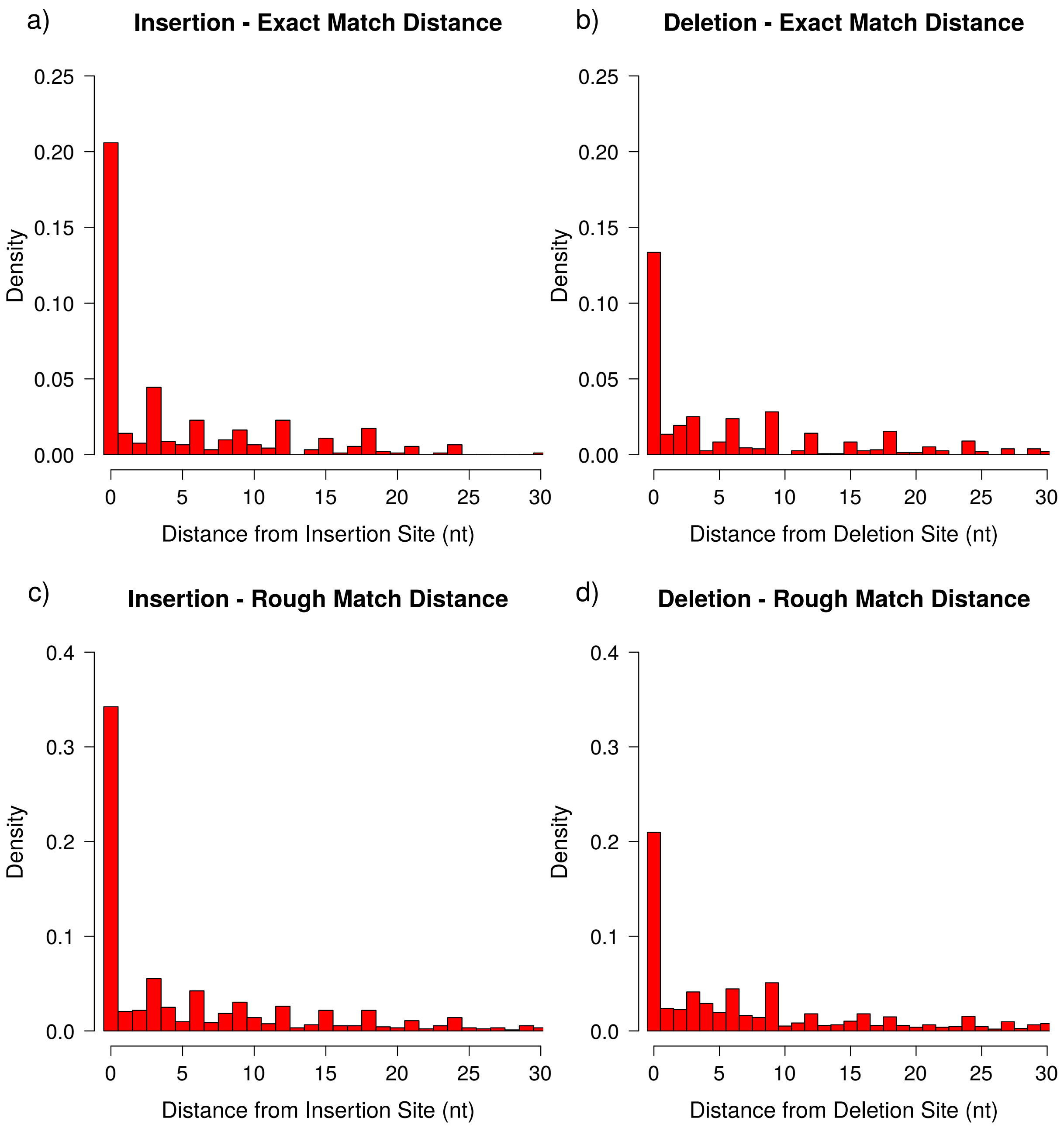

Figure 4.1: The minimum distance, in both the 5' and 3' directions (upstream and downstream), to the next perfect sequence match for insertion (a) and deletion (b) sequences. The same scan was repeated for "rough" matches for both insertions (c) and deletions (d), in which 1 in every 9 nucleotides was allowed to be a mismatch. Results with a distance of 0 nt importantly indicate that a match was detected immediately adjacent to the indel, either in the 5' or 3' direction. Densities indicate the percent of matches found at the given distance, and account for indels with no match found.

4.2 Methods

4.2.1 Insertion Data

For this modelling task, I used the insertion sequence data retrieved from the branch length consensus trees of my within hosts study in Chapter 3. Recall that a single consensus tree was selected from the posterior distribution of each patient's analysis in BEAST by generating a distance matrix of branch lengths and selecting the tree closest to the centroid, or central point describing the mean across all branch lengths. Insertions were retrieved by recursively iterating over every branch of a phylogenetic tree and performing pairwise comparisons between each ancestor and descendant sequence. Importantly, I recorded the insertion sequence, its location, the full descendant sequence containing the insertion, and the full ancestor sequence lacking it. This yielded 923 insertion sequences from 64 different phylogenetic trees.

4.2.2 Bayesian Framework

In order to fit models to my data in a modular and customizable fashion, I opted to code a custom Metropolis-Hastings Markov Chain Monte Carlo (MCMC) algorithm in R to conduct Bayesian inference. Recall that MCMC sampling can be used to sample from a target probability distribution to approximate its characteristics which, in this case, is the posterior distribution of each empirical model parameter. In this framework, I utilize the Metropolis-Hastings algorithm, which first randomly samples a new proposed value (or step) and then chooses to accept or reject this proposal. The algorithm always accepts proposals with a higher probability than the current position and accepts lower probability ones with a random chance proportional to their decrease in probability. The algorithm I coded here is functionally equivalent to that utilized in BEAST to sample the posterior distribution of phylogenetic tree parameters, albeit in a more simplified format. Use of this statistical framework provided great flexibility when fitting more complex models in subsequent model advancements. Generally, model experimentation involves determining the model parameters, specifying the appropriate prior distribution on each parameter, specifying the details of the proposal algorithm (that is, the code responsible for generating a new proposed change to parameter values), and finally, testing the model on simulated data generated using a known process (*i.e.* to test a geometric model, I will simulate data using a randomized geometric process).

4.2.3 Fundamental Assumptions

To model insertions as the products of replication slippage, I first needed to make certain assumptions about how this model system will work. I should note that additional assumptions were required for subsequent model developments, but here I will highlight the core ideas that remained constant throughout the modelling process. Firstly, I assumed that every insertion sequence of length *n* is the product of *n* ‘slip events’, each of which involve the polymerase and copying strand slipping backward on the template sequence by 1 nucleotide. Correspondingly, standard genetic sequence outside of insertions of size *n* was also assumed to be the product of *n* successful ‘copy events’ that add 1 nucleotide at a time, in line with standard nucleotide replication mechanisms. We opted to deconstruct insertions and standard sequences into single nucleotide events in order to account for all possibilities of insertion lengths observed in the data. Also, we referenced replication literature and found no basis that might suggest the slippage of a nucleotide strand would demonstrate a tendency to occur in units larger than 1 nucleotide [6]. Under these assumptions, I could convert the variable loop sequences containing insertions into a sequence of copy and slip events. Since I was focused on modeling the counts of slip events, all copy events were converted into an absence of slip events and slip events were pooled together as shown below (i.e. 0 slips).

```
a) Copy / Slip Events:   C C C S S S C C C S S S C C C
b) Slip Counts:          0 0 0 3     0 0 0 3     0 0 0
```

Figure 4.2: Format of data used as input for empirical models of insertion.

4.3 Results

4.3.1 Single Parameter Models

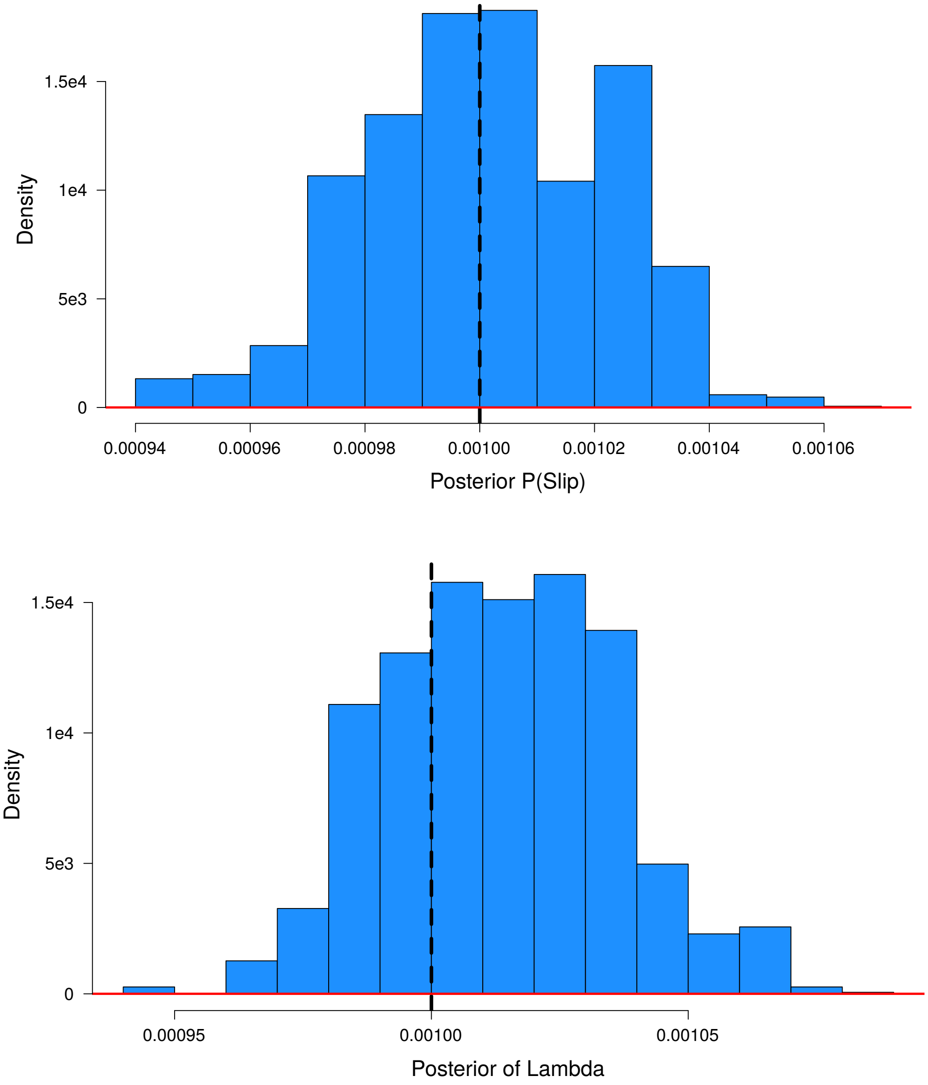

Figure 4.3: Posterior distributions of two different single-parameter models, both sampled for 10^6 MCMC iterations on simulated data in a custom Bayesian framework. In the model shown on top, slip events are assumed to be geometrically distributed based on a single underlying parameter describing the probability of slipping: $P(slip)$. Normal copy events are considered successes while slip events are considered failures, so that the probability of success is $1 - P(slip)$. In the bottom model, slip events are assumed to follow a Poisson distribution governed by a single parameter “lambda” (λ) describing the mean count of slip events per nucleotide. In both models, the dashed black line represents the true value used to generate the simulated data, while the red line describes the prior distribution used for the analysis.

My initial attempts to model insertion sequences involved fitting models containing a single parameter to my insertion data. The first model assumed that there was a constant underlying probability of slipping at each nucleotide which was described by a geometric distribution. Equation 4.1 describes the geometric distribution which captures the probability of seeing x number of slip events, given s the probability of slipping in a sequence (modified so that $s = 1 - p$ where p is the probability of successfully copying). I generated the corresponding log-likelihood function of s shown in Equation 4.2 — which describes the total likelihood of seeing the observed counts x_1 through x_n given the slip probability s — to evaluate the posterior distribution of this parameter using Bayesian inference. The results of this simulated test are shown in Figure 4.3 (top).

$$f(x \mid s) = (1 - s)s^x \tag{4.1}$$

$$logL(s) = n\ln(1 - s) + \left(\sum_{i=1}^{n} x_i - n\right)\ln(s) \tag{4.2}$$

My second single-parameter model assumed that the number of slip events at each nucleotide position was generated through a Poisson process. Equation 4.3 refers to the Poisson distribution that describes the probability of seeing x slip events, given the mean count of slip events across all nucleotides: λ. I then utilized the Poisson log-likelihood function shown in Equation 4.4 to determine the posterior distribution of λ which is shown in Figure 4.3 (bottom).

$$f(x \mid \lambda) = \frac{\lambda^x e^\lambda}{x!} \tag{4.3}$$

$$logL(s) = \sum_{i=1}^{n} x_i \ln(\lambda) - n\lambda \tag{4.4}$$

The two single parameter models showed strong convergence; however, there were notable shortcomings due to their simplicity as expected. When trying to simulate new insertions using the geometric or Poisson processes, the distribution of simulated lengths was strongly biased towards very short lengths of 1, 2 and 3 (not shown), which vastly differed from my observed insertion data (Figure 4.4). Both the geometric and Poisson distributions assume that the probability of each event (a slip in this case) occurs independently from other events and therefore, having 3 slip events occur at the same nucleotide has the probability of $(1-s)\times s\times s\times s$ (Equation 4.1). From here, we were confident that the model needed additional complexity to simultaneously account for the very low frequency of starting a slip event and the long insertion lengths that can be generated, as shown in Figure 4.4.

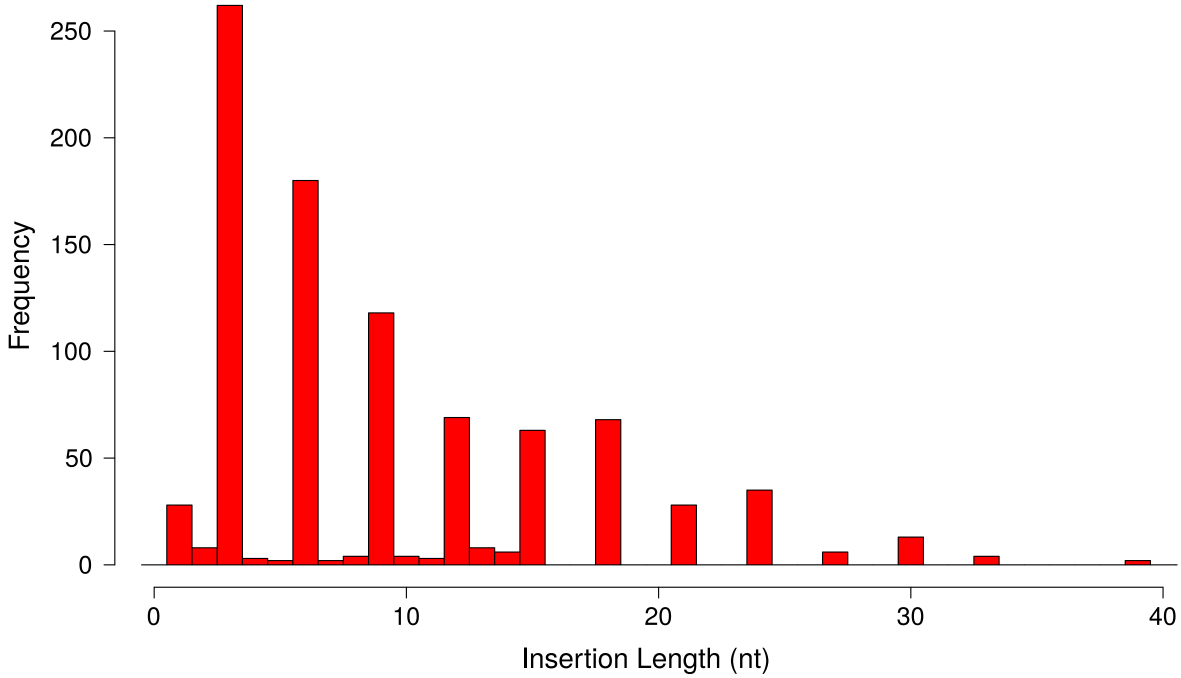

Figure 4.4: The lengths of insertions retrieved in the five variable loops of HIV-1 gp120 across 26 patients. For each patient, insertions were recovered from a branch length consensus tree selected from the posterior distribution of BEAST analysis.

4.3.2 Two State Model

For subsequent modelling, I opted to utilize a model that has two different states: a copy state during which a nucleotide is replicated normally, and a slip state that inserts an additional nucleotide into the sequence (Figure 4.5). This model has two parameters governing movement between the states: a probability of entering the slip state (enter) and a probability of staying in it (stay; Figure 4.5).

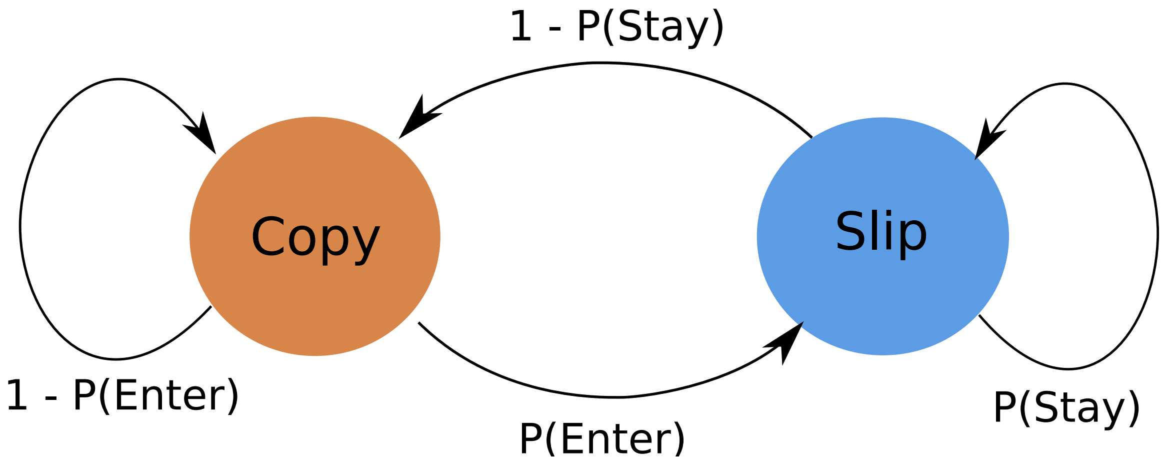

Figure 4.5: Schematic depicting a two state model mechanism governed by two parameters: the probabilities of entering and staying in the slip state. The probability of staying in the current state and moving to the other can differ based on the current state.

The independence of these two parameters addresses the primary shortcoming of the single-parameter models by permitting the frequency of insertions to be low, while generating relatively long insertion lengths. Under this model, the descendant sequence containing the insertion is deconstructed so that every nucleotide is assigned to one of the two model states based on whether it is part of the standard variable loop sequence or an inserted sequence. Importantly, these states then dictate the series of sequential steps through the two state model system (Figure 4.5). For example, if the sequence of states is “copy, copy, slip, copy”, then we know that the steps taken were 1) staying in the copy state, 2) entering the slip state, and 3) leaving the slip state. Since every step through the model system has an associated probability, I can then calculate the likelihood for this series of events given the model parameters. The log-likelihood of a given descendant sequence containing an insertion (x) is given by the following formula:

$$logL(x \mid E, S) = a\ln(1-E) + b\ln(E) + b\ln(1-S) + c\ln(S) \qquad (4.5)$$

where E is the probability of entering the slip state, S is the probability of staying in the slip state, a is the number of regular copying events, b is the number of insertions, and c is the cumulative number of nucleotides in insertions excluding the first nucleotide (*i.e.* insertion length -1).

Nucleotide Substitutions

I then added the capability to account for nucleotide substitutions. Up to this point, I was fitting my two-state model to the descendant sequences of tree branches that had just acquired a new insertion sequence. To account for nucleotide substitutions in the model however, I needed to compare each descendant sequence to that of its immediate ancestor. This data was readily available considering that I saved both the ancestor and descendant sequences along each branch of the tree containing an insertion. Importantly, this does not affect the two state model mechanism as it continues to read the descendant sequence containing the insertion as normal. Inclusion of ancestral sequences allowed me to incorporate a Felsenstein ‘81 model of evolution into my empirical model in order to calculate the likelihood of the observed nucleotide substitutions. Although the Felsenstein ‘81 model has four parameters (three describing nucleotide frequencies and one rate; π_A, π_C, π_G, μ), I opted to fix the nucleotide frequencies to those estimated in my data set and allow the remaining substitution rate to become the third parameter in my model system. I generated simulated sequence data using a rate for substitutions and a two state system for insertions (Figure 4.6).

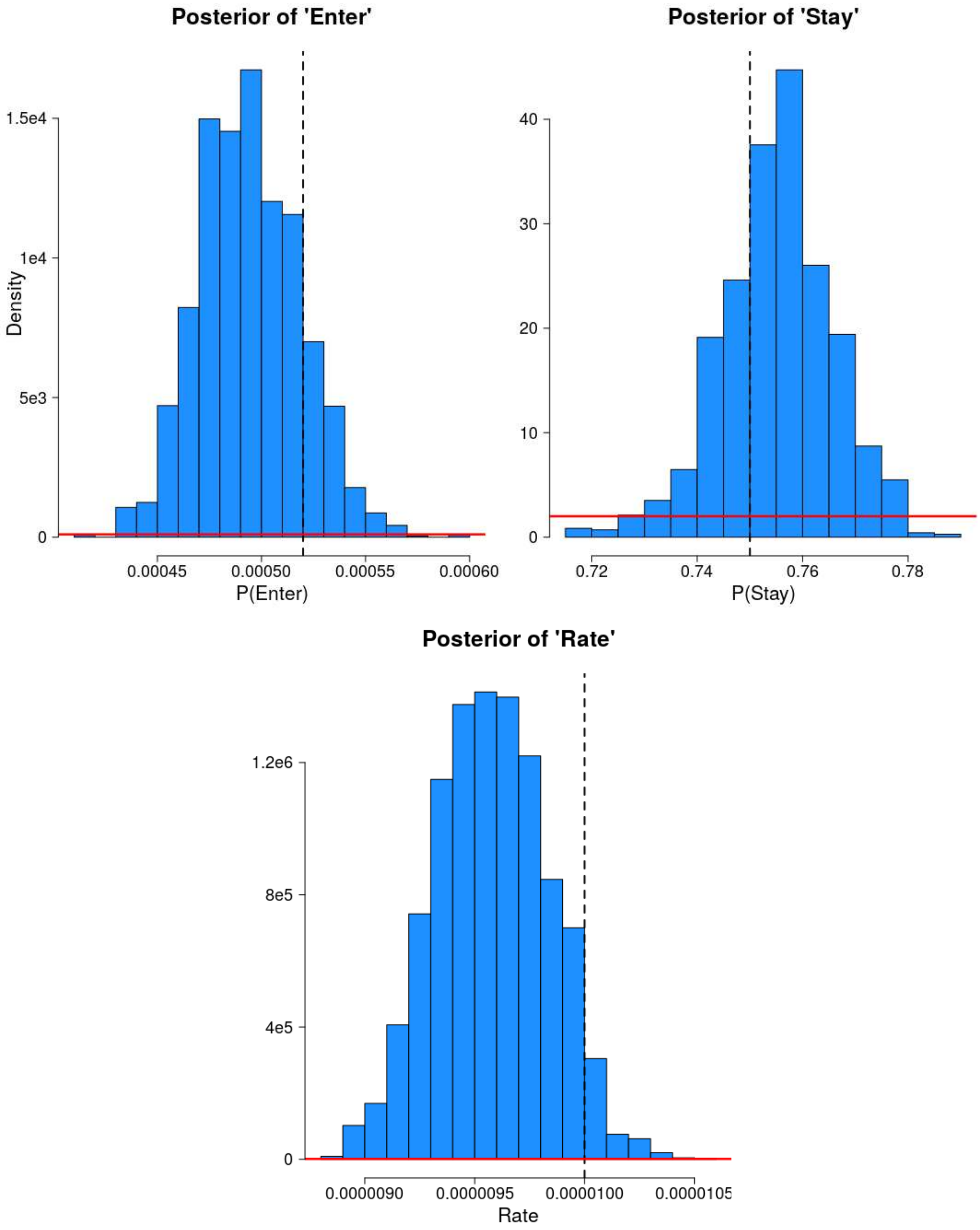

Figure 4.6: Posterior probabilities of three parameters describing the empirical model of insertion sequences. The three parameters shown here are 1) the probability of entering the slip state, 2) probability of staying in the slip state, and 3) the rate of nucleotide substitutions. The model was run for 2×10^5 MCMC iterations on simulated data in a customized Bayesian framework. The dashed vertical line indicates the true value used to simulate the data, while the red horizontal line indicates the prior probability set on each model parameter.

Length Bias

I then sought to account for the strong bias towards inframe insertion lengths in the observed data by including a probability parameter called ‘fixation’ (Figure 4.4). Firstly, this parameter operates under the assumption that inframe insertions are under no purifying selection, while those that induce frameshifts experience strong pressures. This fixation parameter describes the estimated proportion of frameshift-inducing insertions observed in the data, or in other words, the proportion of frameshifts that are successfully reach fixation and are not removed by purifying selection. To incorporate this parameter, we devised an algorithm that uses the model parameters and the number of inframe insertions (assumed to be unaffected) to calculate the theoretical total number of frameshift-inducing insertions that would exist in the data if purifying selection was fully absent. I then use the observed frameshift insertion count and this theoretical total to sample the binomial probability of observing a frameshift in the data (i.e. observed frameshifts / total theoretical frameshifts).

Random Walk of Slip Events

Given that indels cannot be observed directly and only reconstructed through inference, there is always the possibility that the reconstructed positions, lengths or counts of indels in an alignment does not match those in reality. Knowing this, I added functionality to my model that allowed it to change the positions of slip events within genetic sequences to explore other configurations of insertions. For example, with 3 slip events in the same location, the model can move one slip event to evaluate the likelihood of having two insertions of length 1 and 2 a few nucleotides apart. This algorithm essentially lets the model treat the positions of slip events as additional parameters that it can adjust and re-evaluate. Importantly, this mechanism relies on the likelihood calculated by the Felsenstein ‘81 substitution model, as the movement of slip events will shift the sequence and greatly affect the number of perceived substitutions. Proposed slip positions will cause favourable increases in F81 likelihood if they facilitate the alignment of more nucleotides and minimize substitutions. Likewise, positions that induce misalignments that generate more substitution events will be unfavourable due to a corresponding drop in likelihood. This mechanism added considerable complexity to the model, as it essentially sought to realign sequences by moving inserted nucleotides.

The current model implements a two-state mechanism, substitution rate parameter, fixation probability, and a random walk of slip events. Simulated results run for $2e^5$ MCMC iterations are shown in Figure 4.7.

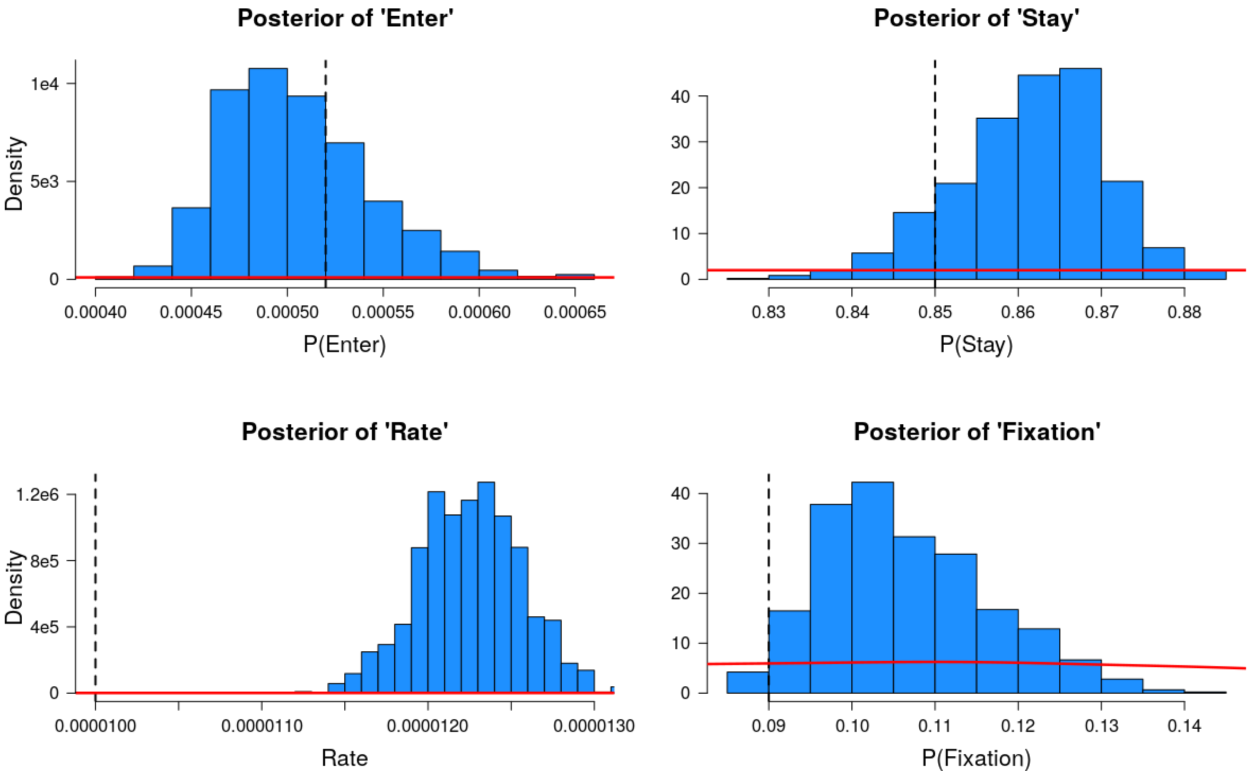

Figure 4.7: Posterior probabilities of four parameters describing the finalized empirical model of insertion sequences. The four parameters shown here are 1) the probability of entering the slip state, 2) probability of staying in the slip state, 3) the rate of nucleotide substitutions, and 4) the probability of seeing a frameshift-inducing insertion length (fixation). The model was run for 2×10^5 MCMC iterations on simulated data in a customized Bayesian framework. The dashed vertical line indicates the true value used to simulate the data, while the red horizontal line indicates the prior probability set on each model parameter. Importantly, this model run involved the initial shuffling and subsequent reconstruction of independent slip events.

4.4 Discussion

Indels are an important mutational mechanism given their significant contributions to molecular sequence evolution [5]. Though indels bring inherent difficulties in their reconstruction, they have been shown to provide substantial benefit to phylogenetic inference methods, especially those pertaining to rapidly evolving pathogens [11]. Here, I present my efforts to generate an empirical model describing the insertions sequences retrieved in the variable loops of HIV-1 gp120 as the products of a replication slippage mechanism [6]. The resulting model, demonstrated in Figure 4.7, represents my work up to this point. Overall, the model is able to simulate nucleotide substitutions and insertions in sequences in a manner that reasonably reflects the trends observed in patient-derived sequence data. The use of a two-state model system provided the flexibility needed to capture the unique trends of very rare insertion events that were simultaneously quite long (Figure 4.4). Importantly, my model is able to account for the strong bias against frameshift-inducing insertion lengths, through the incorporation of a penalty parameter (fixation) (Figure 4.7). Although this mechanism still requires tuning, my model is also able to explore additional insertion configurations by moving independent slip events (Figure 4.7). It has demonstrated a moderate capacity to reconstruct the positions of insertions observed in the data as well (not shown). Importantly, this work should be considered with its scope in mind, as it focuses on cases of indels in a specific genetic region of a particular virus. That being said, I believe my work will contribute to better understanding of indel characteristics and enable more in-depth research into the mechanisms that govern their creation.

Bibliography

[1] H. Ashkenazy, O. Cohen, T. Pupko, and D. Huchon. Indel reliability in indel-based phylogenetic inference. *Genome biology and evolution*, 6(12):3199–3209, 2014.

[2] E. V. Ball, P. D. Stenson, S. S. Abeysinghe, M. Krawczak, D. N. Cooper, and N. A. Chuzhanova. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local dna sequence complexity. *Human mutation*, 26(3):205–213, 2005.

[3] K. A. Clarke and D. M. Primo. *A model discipline: Political science and the logic of representations*. Oxford University Press, 2012.

[4] D. R. Cox. *Principles of statistical inference*. Cambridge university press, 2006.

[5] K. Ezawa. General continuous-time Markov model of sequence evolution via insertions/deletions: are alignment probabilities factorable? *BMC bioinformatics*, 17(1):304, 2016.

[6] M. Garcia-Diaz and T. A. Kunkel. Mechanism of a genetic glissando*: structural biology of indel mutations. *Trends in biochemical sciences*, 31(4):206–214, 2006.

[7] G. Hickey and M. Blanchette. A probabilistic model for sequence alignment with context-sensitive indels. In *International Conference on Research in Computational Molecular Biology*, pages 85–103. Springer, 2011.

[8] I. H. Holmes. Solving the master equation for indels. *BMC bioinformatics*, 18(1):255, 2017.

[9] S. T. Lovett. Template-switching during replication fork repair in bacteria. *DNA repair*, 56:118–128, 2017.

[10] A. Löytynoja and N. Goldman. Short template switch events explain mutation clusters in the human genome. *Genome research*, 27(6):1039–1049, 2017.

[11] B. D. Redelings and M. A. Suchard. Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC evolutionary biology*, 7(1):40, 2007.

[12] O. Westesson, G. Lunter, B. Paten, and I. Holmes. Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. *PLoS One*, 7(4):e34572, 2012.

Chapter 5

Conclusions

5.1 Overview

Overall, indels remain a relatively understudied topic in HIV-1 research, despite having profound effects on sequence variation in the HIV-1 genome. My work in this dissertation contributes to this topic by estimating the rates and characteristics of indels in the variable loops of HIV-1 gp120, and generating an empirical model to describe these mutations. Of the many investigations performed in this work, the estimation of indel rates formed the core of this study. While Mansky and Temin [4] estimated an *in vitro* indel generation rate across the HIV-1 genome, rates of indel evolution had not yet been estimated in HIV-1 prior to this work. Moreover, no studies have estimated rates of indel evolution in the variable loops of gp120 where indels are abundant and play a biologically significant role in the generation of immune escape variants. I address this particular knowledge gap here by providing the first estimates of indel rates and other characteristics in the gp120 variable loops using dated phylogenetic analyses of patient-derived HIV-1 sequence data. Furthermore, I conducted these analyses on both an among-host and within-host scale by utilizing different sequence data and methods. We had reasonable basis to expect different results when examining indels on these two different scales because of the considerable differences in evolutionary forces and sampling techniques among hosts compared to those within hosts. Finally, I extend my efforts by developing an empirical model that describes the observed indel data collected in my within-hosts study.

5.2 Summary of Findings

Foremost, this work reports unique and novel findings of gp120 variable loop indel rates within and among hosts. On an among-host scale in Chapter 2, we detected that HIV-1 subtype B exhibited significantly lower variable loop indel rates than the reference clade AE and generally lower than the other examined clades (Figure 2.2). Within-host indel rate estimates in Chapter 3 appeared significantly higher than those estimated among hosts in all variable loops except V3 (Figures 2.2 and 3.2). In our within-host study, we also found that deletion rates appeared significantly higher than insertion rates in these same variable loops: V1, V2, V4, and V5 (Figure 3.2). The apparent shrinking of these variable loops could be explained by selection for higher fusion efficiency, which could be further enhanced by relatively weak immune systems in the examined patients. Additionally, we found that within-host variable loop indels tended to accumulate more frequently during the earlier stages of infection based on the higher indel rates of internal branches (Figure 3.2) and the dating of indels earlier in infection (Figure 3.3). This suggests that indels undergo strong selection during the acute stages of HIV-1 infection, but could also be the result of bias within our analyses.

Next, I found interesting trends on the topic of indel lengths. No frameshift-inducing indel lengths could be detected on an among-host scale, while 4.2% and 4.5% of insertions and deletions, respectively, adopted these lengths on a within-host scale (Figure 3.4). The lack of frameshifts on an among host scale was expected due to sequences having undergone substantial purifying selection at this level. However, the low proportions of frameshifts on a within-host scale suggests that there is still notable purifying selection acting on intrapatient datasets and thus contradicts our hypothesis.

I also report interesting findings regarding indel-induced changes to the number of PNGSs in the variable loops of gp120. On an among-host scale, we find that indels in V1 and V2 either added or removed PNGS more frequently than expected (Figure 2.5). A similar, more detailed analysis within hosts further revealed that deletions affected PNGS in a relatively random manner, while insertions in V1, V2, and V4 tended to add PNGS more often than expected by chance (Figure 3.6). This corroborates existing reports on indels by showing that insertions appear to undergo positive selection to add PNGSs in the variable loops, likely causing them to increase in prevalence within HIV-1 populations [1, 2, 6]. Overall, these findings better characterize the role of indels in changing the glycan shield of gp120, and in the case of insertions specifically, provide quantified estimates that reflect the selective pressures they experience (Figure 3.6).

5.3 Concluding Hypotheses

We hypothesized that sequence data sampled within hosts using SGS would permit the detection of indels that have not yet been subject to substantial purifying selection. However, this original hypothesis does not appear to be supported by our findings. As mentioned previously, the minimal frameshift-inducing indel lengths found within hosts suggests strong purifying selection still present at this scale. In further support of this, we find lower indel rates in the terminal branches of phylogenetic trees demonstrating that lineages free from the constraint of virus replication (*i.e.* terminals) still experience notable filtering of indel events, also implying strong purifying selection.

My findings do however, support the "store and retrieve" hypothesis associated with HIV-1 evolution. This hypothesis describes the preferential transmission of an HIV-1 variant more similar to the original founder virus than the highly-adapted variants in circulation, which is a potential explanation for why HIV-1 exhibits faster evolution within hosts than among hosts [5]. For instance, this apparent reversion of HIV-1 genetic diversity accounts for the lower substitution rates observed among hosts relative to those within hosts [3, 7]. Similar to substitution rates, I postulate that this hypothesis also applies to my estimates of variable loop indel rates. I hypothesize this based on the significantly higher rates within hosts, and importantly, the identical trend in indel rates between variable loops (*i.e.* V1 and V5 being highest, V2 and V4 next highest, and V3 lowest). Specifically, this identical trend in rates suggests that within and among host sequence samples may undergo similar selection pressures and patterns of indel accumulation, only that among-host rates exhibit lower magnitudes due to selective "reversion" to an older, less mutated sequence upon transmission.

Finally, I have demonstrated that a two-state model system can be used to effectively capture and recreate trends in insertion sequence length and frequency. Specifically, I generated an empirical model of insertions as the products of replication slippage that shows promise for further expansion. In terms of effectiveness, my model is capable of simulating indels with reasonably similar trends to observed data.

5.4 Significance & Contributions

This work provides an important contribution to the field of HIV-1 evolution by presenting the first reported estimates of gene-specific indel evolution rates. For one, these rate estimates describe biologically significant mutations in the gp120 variable loops, which are responsible for the generation of immune escape variants and drive the adaptation of HIV-1 to human hosts. Therefore, these rates can improve understanding of immune escape processes and may hold correlations with disease progression, similar to evolutionary rates of substitutions. Indel rates also quantify the contributions of indel to genetic sequence evolution in the gp120 variable loops, allowing for comparisons to be made relative to nucleotide substitutions, for example. Furthermore, our analysis of indel rates also marks an important step toward investigating the potential for indels to be incorporated into phylogenetic inference, as current methods do not account for this source of genetic information.

Beyond indel rates, my reports on indel-induced changes to variable loop PNGS provide novel quantified insights into the indel modulation of the glycan shield. Additionally, I contribute an empirical model capable of recreating the trends in variable loop insertion sequences, further demonstrating the viability and effectiveness of modelling indel mutations.

5.5 Future Directions

Future studies may consider further investigating indel timings and their association with immune escape by examining correlations between gp120 variable loop characteristics and patient clinical outcomes. Such an approach could test the hypothesis that variable loop indels and their rates of accumulation are in fact significant in a clinical context. Additionally, having demonstrated the tractability of indel rate estimation among and within hosts, applying these analyses to other areas of the HIV-1 genome is a sensible next step. The documentation of additional indel rates could facilitate the development of efficient models of indel evolution that allow indels to be utilized in phylogenetic inference methods. Moreover, by improving understanding of the evolutionary landscape in gp120, my indel rate estimates might be useful in future attempts at HIV-1 vaccine development, as these approaches would certainly need to account for all aspects of evolution in gp120.

There is considerable work that can expand upon my empirical model of insertion sequences presented in Chapter 4. For one, the current mechanism of replication slippage can be expanded to include deletion events in addition to insertions, possibly using a “jump ahead” mechanism. Also, this model can be changed entirely to describe indels as products of a template-switching mechanism instead. The slippage and template-switching models could then be compared using model selection methods to determine which fits the data more effectively, thereby offering insights into the likely underlying mechanisms driving indel generation.

In conclusion, I present my contributions to the topic of evolution in HIV-1 gp120 with hopes to open up areas of future research into the rates, characteristics, and modelling of indels.

Bibliography

[1] E. Castro, M. Bélair, G. P. Rizzardi, P. A. Bart, G. Pantaleo, and C. Graziosi. Independent evolution of hypervariable regions of HIV-1 gp120: V4 as a swarm of N-Linked glycosylation variants. *AIDS research and human retroviruses*, 24(1):106–113, Jan. 2008. ISSN 0889-2229. doi: 10.1089/aid.2007.0139.

[2] M. E. Curlin, R. Zioni, S. E. Hawes, Y. Liu, W. Deng, G. S. Gottlieb, T. Zhu, and J. I. Mullins. HIV-1 envelope subregion length variation during disease progression. *PLoS pathogens*, 6(12):e1001228, 2010.

[3] K. A. Lythgoe and C. Fraser. New insights into the evolutionary rate of hiv-1 at the within-host and epidemiological levels. *Proceedings of the Royal Society B: Biological Sciences*, 279(1741):3367–3375, 2012.

[4] L. M. Mansky and H. M. Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, 69(8):5087–5094, 1995.

[5] A. D. Redd, A. N. Collinson-Streng, N. Chatziandreou, C. E. Mullis, O. Laeyendecker, C. Martens, S. Ricklefs, N. Kiwanuka, P. H. Nyein, T. Lutalo, et al. Previously transmitted HIV-1 strains are preferentially selected during subsequent sexual transmissions. *The Journal of infectious diseases*, 206(9):1433–1442, 2012.

[6] M. Sagar, X. Wu, S. Lee, and J. Overbaugh. Human immunodeficiency virus type 1 V1-V2 envelope loop sequences expand and add glycosylation sites over the course of infection,

and these modifications affect antibody neutralization sensitivity. *Journal of virology*, 80 (19):9586–9598, 2006.

[7] B. Vrancken, A. Rambaut, M. A. Suchard, A. Drummond, G. Baele, I. Derdelinckx, E. Van Wijngaerden, A.-M. Vandamme, K. Van Laethem, and P. Lemey. The genealogical population dynamics of hiv-1 in a large transmission chain: bridging within and among host evolutionary rates. *PLoS Comput Biol*, 10(4):e1003505, 2014.

Appendix A

Chapter 2 Supplementary Material

A.1 Supplementary Figures

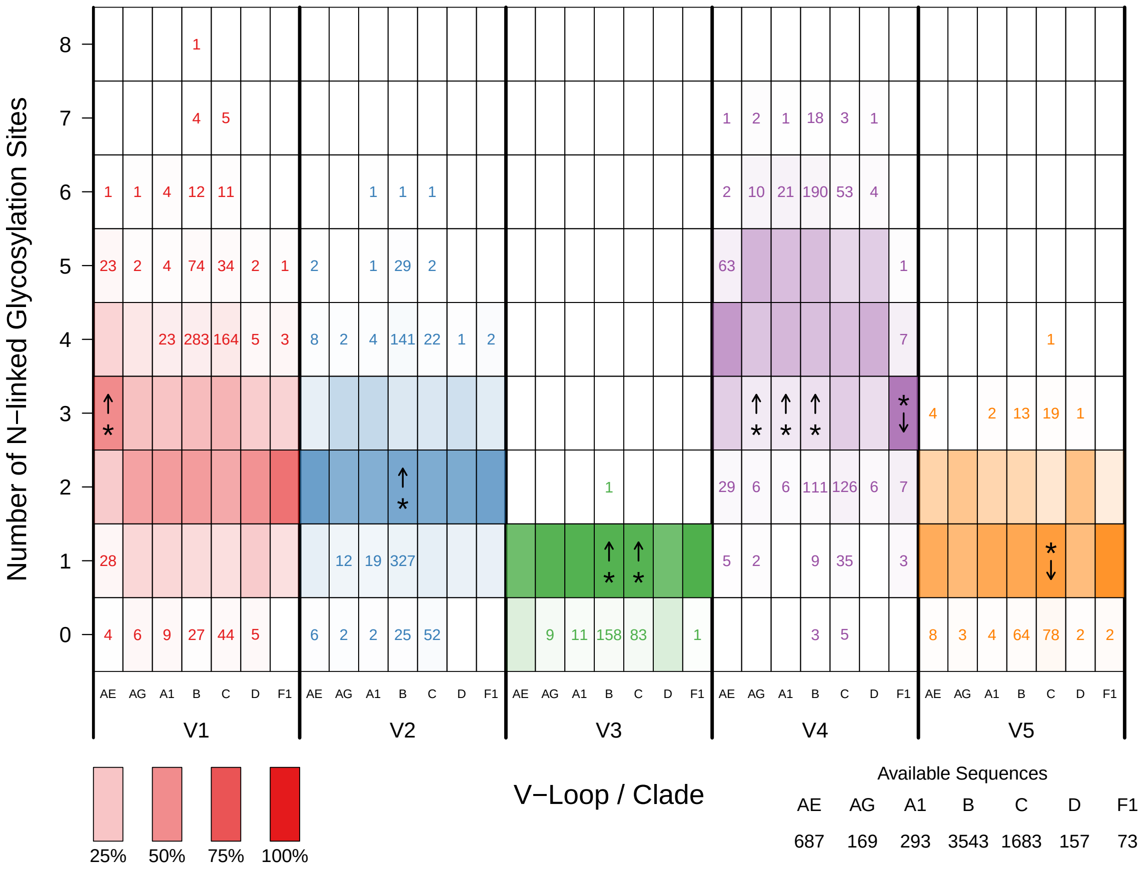

Figure A1: The number of N-linked glycosylation sites in the variable loops of gp120. Columns within each variable loop section describe the counts retrieved for one specific group M clade. The color density of each box indicates the proportion of sequences containing the given count. Examples of color densities and their corresponding proportions are provided by the four boxes to the bottom left of the figure. Boxes containing a number highlight PNGS counts that are above zero, but contained less than 10% of the distribution and therefore, did not generate a noticeable color. Asterisks (*) and arrows denote the presence and direction of significantly different PNGS counts among clades within a given variable loop (Poisson GLM).

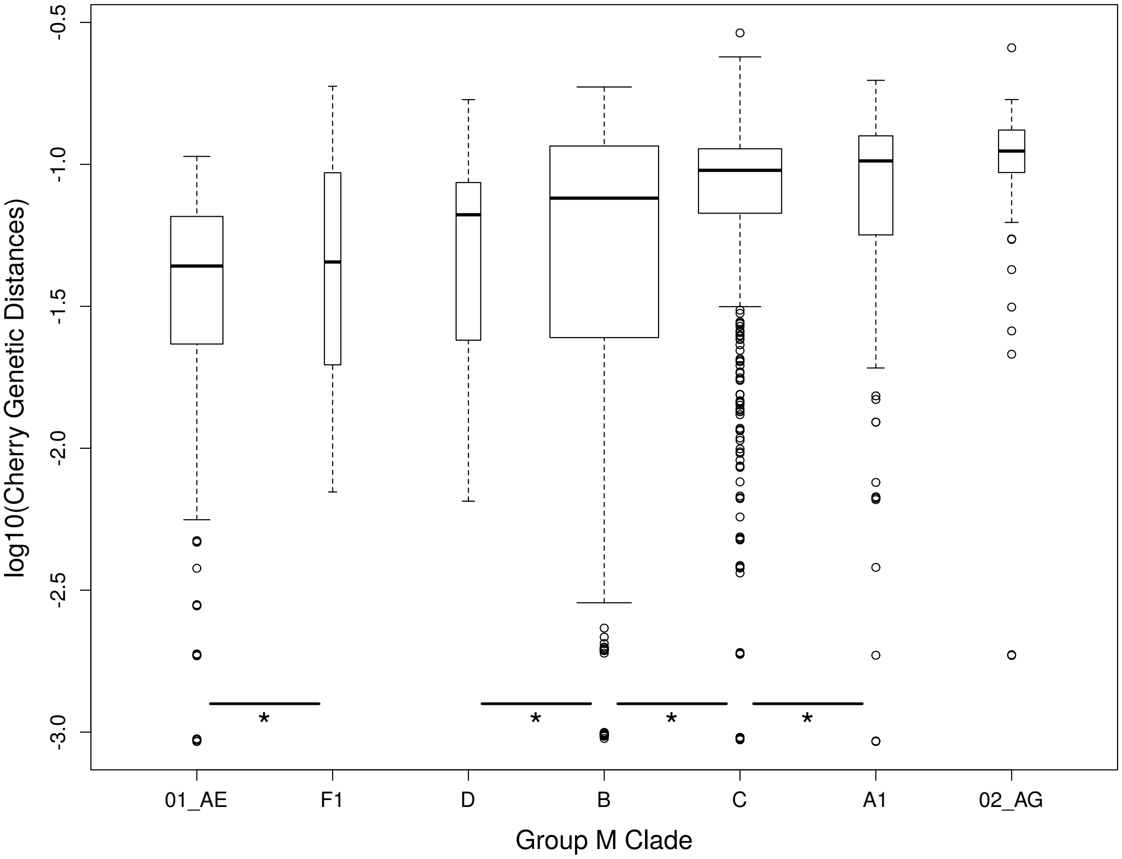

Figure A2: Log$_{10}$-transformed genetic distances of cherries stratified by group M clade. Box widths are representative of the number of cherries retrieved from the seven clades. Cherries with zero genetic distance were excluded from this visualization. For each clade in the order shown, counts of these filtered cherry pairs were 8, 0, 3, 203, 12, 6, and 1, respectively. Significant differences in adjacent group means were determined using a Wilcoxon rank sum test after adjusting for multiple comparisons and are indicated by * symbols. There were 68 instances of cherries with a combined branch length of zero that contained an indel.

A.2 Supplementary Tables

	Estimate	P	Signif.
Time	0.001	$< 10^{-15}$	*
Clade			
01_AE	(reference)		
02_AG	1.946	0.058	
A1	-0.103	0.741	
B	-1.156	1.6×10^{-10}	*
C	-0.372	0.065	
D	-0.639	0.070	
F1	-0.158	0.780	
Variable Region			
V1	(reference)		
V2	-0.578	0.012	
V3	-4.448	$< 10^{-15}$	*
V4	-0.438	0.044	
V5	-0.042	0.844	
Interactions			
02_AG:V2	-2.349	0.039	
A1:V2	-0.457	0.313	
B:V2	-0.918	4.1×10^{-4}	*
C:V2	-1.115	1.1×10^{-4}	*
D:V2	-0.326	0.527	
Interactions			
F1:V2	-0.860	0.285	
02_AG:V3	-2.073	0.063	
A1:V3	0.127	0.799	
B:V3	-0.381	0.255	
C:V3	-0.109	0.752	
D:V3	1.363	0.017	
F1:V3	-13.565	0.926	
02_AG:V4	-2.441	0.027	
A1:V4	-0.099	0.819	
B:V4	-0.208	0.396	
C:V4	-0.215	0.428	
D:V4	0.718	0.164	
F1:V4	-0.577	0.439	
02_AG:V5	-2.455	0.022	
A1:V5	-0.569	0.151	
B:V5	0.194	0.407	
C:V5	0.246	0.345	
D:V5	-0.156	0.740	
F1:V5	-0.188	0.797	

Table A1: Statistical comparisons generated by applying a generalized linear model to cherry indel analysis. Comparisons made between clades and between variable regions were in relation to a reference group (01_AE and V1, respectively). Effects of clade and variable region interactions were compared to predicted mean values to detect significant differences. Groups with * symbols denoted statistically significant differences based on a Bonferroni-corrected threshold suited for multiple comparisons (α = 0.05/n).

Appendix B

Chapter 3 Supplementary Material

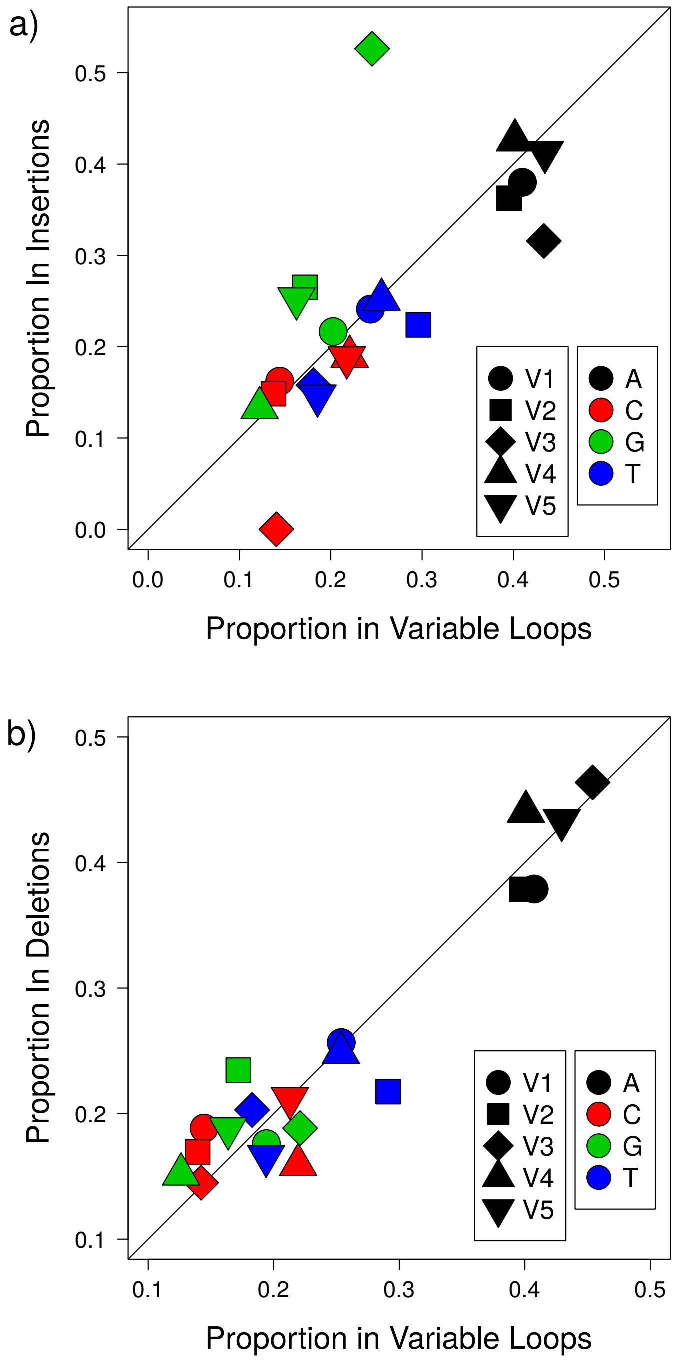

Figure B1: Nucleotide proportions in insertions (a) and deletions (b) relative to the variable loop sequences where they were recovered. Proportions of the four nucleotides, represented by colors, are further stratified across the five gp120 variable loops as denoted by the different shapes. Scales of the x and y axes are identical, meaning that points found above and below the center line represent nucleotide proportions that are higher and lower in indels relative to the variable loops, respectively.

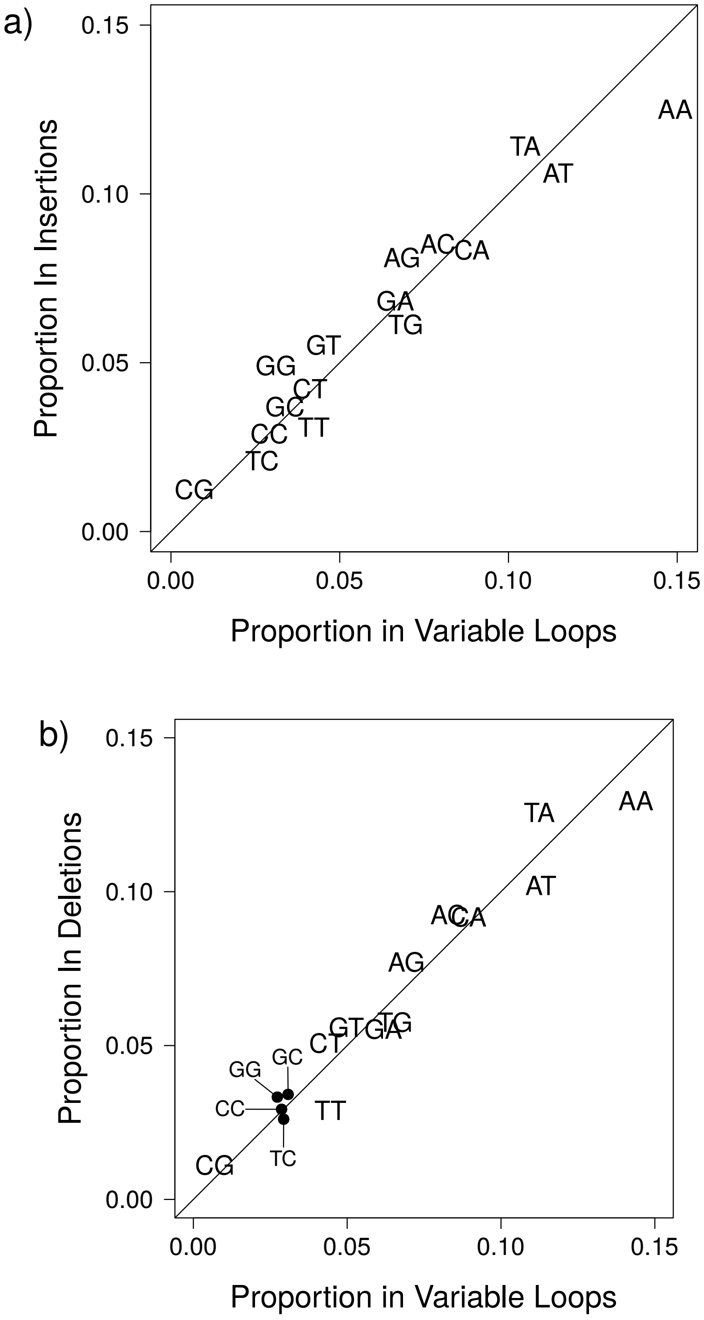

Figure B2: Dinucleotide proportions in insertions (a) and deletions (b) relative to their variable loop sequences of origin. Proportions of all sixteen possible dinucleotides ($A/C/G/T$ + $A/C/G/T$) pooled across all five gp120 variable loops are labelled by their name. As x and y axes are identical, points above and below the line indicate dinucleotide proportions that are higher and lower in indels relative to the variable loops, respectively.

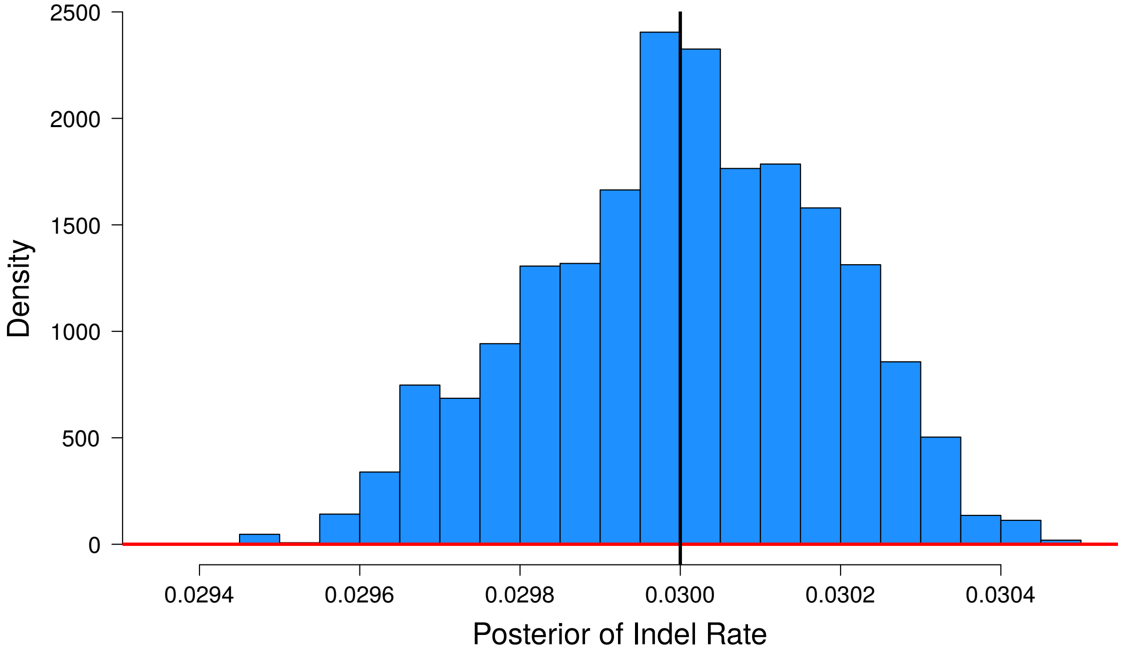

Figure B3: Posterior density of indel rate analysis on simulated data. Analysis was run for 10^5 MCMC iterations in the Bayesian statistical framework RStan to test accuracy of results. The red line indicates the prior distribution while the black line at 0.03 indicates the true value used to generate the simulated data.

Curriculum Vitae

Name:	John Lawrence Palmer
Post-Secondary Education and Degrees:	University of Western Ontario London, ON 2014 - 2018 BMSc.
Honours and Awards:	CIHR Canadian Graduate Scholarship - Masters 2019-2020
	Dr. Frederick Winnett Luney Graduate Research Award 2019
Related Work Experience:	Graduate Teaching Assistant University of Western Ontario - Pathology and Laboratory Medicine Sept 2019 - Apr 2020
	Pre-Graduate Researcher University of Western Ontario - Dr. Art Poon's Lab May 2018 - Aug 2018

Publications:

Palmer J, Poon AF. Phylogenetic measures of indel rate variation among the HIV-1 group M subtypes. Virus evolution. 2019 Jul;5(2):vez022