Overcoming Missing Data in Phylogenetic Analysis of Shotgun Sequencing to Detect HIV Adaptation to Immune Response

by

Thuy Nguyen

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

The University of British Columbia
(Vancouver)

August 2016

Abstract

DNA sequencing gives us insight into how viruses adapt to their host immune systems. Studies of viral populations typically employ deep amplicon sequencing with next-generation reads to capture a detailed sample of genetic variation in a population. The high amount of overlapping sites in a multiple sequence alignment of reads from amplicon sequencing form ideal input for phylogenetic reconstruction, a necessary step for studying evolutionary relations in a population. However, the typical short read lengths of $<$600 bp from next generation sequencing technology with the best sequence error rate impose a severe limit on the width of genomic regions for which evolutionary relationships can be analyzed.

Shotgun sequencing, in which DNA is fragmented at random positions, is an efficient alternative to amplicon sequencing for covering wider regions of a genome with sufficient depth. Due to the random staggered positions of shotgun reads in a genome, an extremely high percentage of missing data can result in multiple sequence alignment of shotgun sequencing. The absence of sequence homology across the entire set of short reads makes it impossible to reconstruct a phylogenetic tree, limiting the utility of shotgun data for phylogenetic analysis.

We developed the Umberjack software pipeline, which employs the 'sliding window' approach to minimize the effect of missing data during phylogenetic reconstruction and obtain evolutionary statistics to detect sites under selection.

Using Umberjack to measure a new metric of directional selection I, significant directional selection was detected in treatment-naive HIV populations at sites with previously documented associations with cytotoxic T-lymphocyte (CTL) response. Further, substitutions towards wild-type amino acids were found to occur early within the population's history, but rarely occurred at a site after the appear-

ance of a CTL escape mutation. Measuring the same metric I in drug treated HIV populations, the directional selection due to the constant pressure of drug treatment was much greater than the directional selection from the immune system.

Preface

The words ‘we’ and ‘our’ throughout the manuscript refers to the work of Thuy Nguyen unless otherwise specified.

Dr. Art Poon architected the sliding window strategy for phylogenetic analysis of short reads. Dr. Richard Liang wrote the software for generating ancestral selection graphs. Thuy Nguyen implemented and tested Umberjack, and analyzed patient datasets under the advice of Dr. Art Poon.

Aram Karakas, Dr. Rachel McGovern, and other members of the BC Centre for Excellence in HIV/AIDS prepared and sequenced HIV patient samples.

Ethical approval was obtained from Providence Health Care/University of British Columbia research ethics board (H04-50276) for the untreated BC patient HOMER cohort and the Netherlands national institutional review board for the Maraviroc clinical trial Netherlands cohort.

There are no publications at this time.

Table of Contents

Abstract . . . . . **ii**

Preface . . . . . **iv**

Table of Contents . . . . . **v**

List of Tables . . . . . **ix**

List of Figures . . . . . **x**

Acknowledgments . . . . . **xiii**

1 Introduction **1**

1.1 HIV Evolution 1

1.1.1 HIV 1

1.1.2 HIV Adaptation to Immune System CTL Response 1

1.1.3 HIV Adaptation to Drug Treatment 2

1.2 Phylogenetic Analysis With Short Reads 3

1.2.1 Next Generation Sequencing 3

1.2.2 Phylogeny of a Single Population 4

1.2.3 Multiple Sequence Alignment 5

1.2.4 Difficulties in Phylogenetic Reconstruction of HIV 7

1.2.5 Phylogenetic Reconstruction Techniques 7

1.2.6 Tree Error from Missing Data 10

1.2.7 Current Approaches to Phylogenetic Analysis From Missing Data . . . 11
1.2.8 New Developments to Overcome Missing Data . . . 13
1.3 Detecting Selection . . . 14
1.3.1 Importance of Selection and Evolutionary History . . . 14
1.3.2 Evolutionary Statistics . . . 14

2 Phylogenetic Analysis With Sliding Windows . . . 19
2.1 Proposal of Sliding Windows to Overcome Missing Data . . . 19
2.2 Software Details . . . 20
2.2.1 Umberjack Use Cases . . . 20
2.2.2 Umberjack Computational Requirements . . . 20
2.2.3 Umberjack Availability . . . 21
2.3 Umberjack Implementation . . . 21
2.3.1 Read Processing . . . 21
2.3.2 Window Extraction . . . 23
2.3.3 Phylogenetic Reconstruction . . . 24
2.3.4 Reconstructing Ancestral Sequences . . . 25
2.3.5 Evolutionary Statistics . . . 26
2.4 Method for Validating on Simulated Data . . . 31
2.4.1 Selection of Simulation Parameters . . . 31
2.4.2 Simulated Topology Under Selection . . . 32
2.4.3 Simulated Recombination . . . 32
2.4.4 Simulated Genomes . . . 32
2.4.5 Simulated Shotgun Reads . . . 33
2.4.6 Umberjack settings . . . 33
2.4.7 Measuring Umberjack $dN - dS$ Accuracy . . . 34
2.4.8 Measuring Tree Topology Accuracy . . . 34
2.5 Determining Features that Affect Umberjack Accuracy . . . 35
2.5.1 Features Considered For Predicting Umberjack Accuracy . 35
2.5.2 Univariate Correlation Ranking . . . 37
2.5.3 Regression Random Forest . . . 37
2.5.4 Backwards Feature Selection in Regression Random Forest 39

2.6 Results . . . 39
2.6.1 Weak Linear Relationship Between Features and Δ . . . 39
2.6.2 Advantages of Regression Random Forests . . . 40
2.6.3 Features Affecting Umberjack Accuracy . . . 41
2.6.4 Effect of Recombination . . . 41
2.6.5 Umberjack Accuracy . . . 42

3 Within-host HIV Evolution . . . 47
3.1 Background . . . 47
3.2 Methods . . . 48
3.2.1 Dataset . . . 48
3.2.2 Umberjack Processing . . . 49
3.2.3 Umberjack Cleaning . . . 50
3.2.4 Finding CTL Response Associated Amino Acid Polymorphisms . . . 51
3.2.5 $dN - dS$ Analysis . . . 52
3.2.6 I Analysis . . . 52
3.2.7 Timing Infection . . . 52
3.3 Results . . . 53
3.3.1 Genetic Variation Across Sites . . . 53
3.3.2 Most Subjects in Chronic Stage of Infection . . . 54
3.3.3 Evolving Sites Associated with CTL Response . . . 55
3.3.4 Host HLA Genotype Drives Diversifying Selection . . . 56
3.3.5 Evidence of Selective Sweeps . . . 59
3.3.6 Reversion to Wild Type . . . 64
3.4 Within-host HIV Evolution During Drug Treatment . . . 68
3.4.1 Background . . . 68
3.4.2 Results . . . 69
3.5 Comparing Methods for Detecting Selection . . . 70

4 Conclusions . . . 73
4.1 Future Directions . . . 74

Bibliography . . . 76

A Supporting Materials 88
A.1 Simulated Datasets For Predicting Umberjack Accuracy 88
A.2 Untreated Patient Metadata 102
A.3 HIV Selective Sweeps in Untreated Patients 108
A.4 Effect of Recombination on Umberjack Error 108

List of Tables

Table 2.1 Spearman Correlation Between Feature and Δ 40
Table 2.2 Feature Ranking for Final Random Forest Predicting Umberjack Accuracy. 42

Table 3.1 Empirical $dN - dS$ Datasets 51

Table A.1 Umberjack Simulated Dataset Parameter Ranges 88
Table A.2 Umberjack Simulated Data Parameters 92
Table A.3 Umberjack Recombination Simulated Data Parameters 94
Table A.4 Untreated Patient Datasets 103
Table A.5 Untreated Patient Dataset Sequencing Statistics 104

List of Figures

Figure 1.1 Example of Phylogenetic Trees 5
Figure 1.2 Example of Multiple Sequence Alignment 9
Figure 1.3 Long Branch Attraction Due to Missing Data 10
Figure 1.4 Multiple Sequence Alignment of Typical Shotgun Sequencing Reads 13

Figure 2.1 Missing Sequence Homology in Shotgun Reads 20
Figure 2.2 Effect of Recombination on Δ 44
Figure 2.3 Umberjack Inferred $dN - dS$ vs Expected $dN - dS$ From Simulated Data 45
Figure 2.4 True Umberjack Inaccuracy Versus Random Forest Prediction of Umberjack Inaccuracy 46

Figure 3.1 Summary of Amino Acid Entropy Along HIV *env* 54
Figure 3.2 Estimated Durations of Infection at Baseline by Molecular Clock Analysis 55
Figure 3.3 Shifts in Matched Non-escape and Escape Mutation Frequencies Over Time in HIV *nef* 57
Figure 3.4 Boxplots of $dN - dS$ at HLA Matched Sites and Unmatched Sites by Gene 58
Figure 3.5 Site $dN - dS$ per Sample at Sites Associated with CTL Response for Patient's HLA 60
Figure 3.6 Substitutions Tend to Map to Tips of the Within-host Phylogeny 61

Figure 3.7 Nonsynonymous Substitutions Map to Tips of the Phylogeny More Often Than Synonymous Substitutions 62
Figure 3.8 Distributions of Nonsynonymous and Synonymous Substitutions in the Tree are More Similar at Sites with Known CTL Associations 63
Figure 3.9 Nonsynonymous Substitutions to HLA-matched Amino Acids Tend to Map Deeper in Within-host Phylogenies 65
Figure 3.10 Nonsynonymous Substitutions to Wild-type Residues Occur Deeper in Phylogenies 67
Figure 3.11 Violin Plot Distributions of I_N and I_S for Site Substitutions from Matched Escape Towards Unmatched Wild Type 68
Figure 3.12 Violin and Boxplot of I Statistic of Viral Populations Treated with Maraviroc 70
Figure 3.13 Onset of Directional Selection 72

Figure A.1 Δ vs Window Total Breakpoint Ratio 95
Figure A.2 Δ vs Window-Site Unambiguous Codon Rate 95
Figure A.2 Δ vs Window-Site Amino Acid Depth 96
Figure A.3 Δ vs Window-Site Substitutions 96
Figure A.3 Δ vs Window-Site Nonsynonymous Substitutions 97
Figure A.4 Δ vs Window-Site Synonymous Substitutions 97
Figure A.4 Δ vs Window-Site Expected Nonsynonymous Substitutions Per Branch 98
Figure A.5 Δ vs Window-Site Expected Synonymous Substitutions Per Branch 98
Figure A.5 Δ vs Normalized Window Tree Length 99
Figure A.6 Δ vs Normalized $\overline{WRF}$ 99
Figure A.6 Δ vs Normalized Polytomies 100
Figure A.7 Δ vs Codon Distribution P-value 100
Figure A.7 Δ vs True Site Codon Entropy 101
Figure A.8 Δ vs Window-Site Codon Entropy 101
Figure A.8 Δ vs Window-Site Codon Entropy 102

Figure A.9 Phylogenetic Tree of GAG Consensus Sequence of Samples from Untreated Patients 105
Figure A.10 Phylogenetic Tree of NEF Consensus Sequence of Samples from Untreated Patients 106
Figure A.11 Phylogenetic Tree of ENV Consensus Sequence of Samples from Untreated Patients 107
Figure A.12 Cleaned Umberjack Estimate of Site $dN - dS$ vs True $dN - dS$ 108
Figure A.13 Violin Plots Summarizing the Distributions of Site $I_N - I_S$ . . 109
Figure A.14 Effect of Recombination on Δ 110

Acknowledgments

I would like to thank my supervisor Dr. Art Poon for his unfailing patience and tireless efforts in guiding me through this thesis.

This work was supported by a CIHR operating grant awarded to Dr. Art Poon (HOP 111406), by an administrative supplement to NIH NIDA grant R01-DA011591, and by a grant from Genome Canada (Genomics and Personalized Health).

Chapter 1

Introduction

1.1 HIV Evolution

1.1.1 HIV

Human Immunodeficiency Virus (HIV) is a double stranded RNA retrovirus that targets the human immune system. HIV causes the steady decline of CD4 lymphocytes (white blood cells) through the destruction of infected CD4 cells or stimulating cell death pathways for uninfected CD4 cells [21]. Left untreated, HIV infection will lead to Acquired Immune Deficiency Disorder (AIDS) during which opportunistic infections can cause mortality [15]. Upon entering a cell, HIV integrates its own genome into the genome of the infected cell. Using the cell's natural replication process, it transcribes and translates viral proteins to create progeny viruses that bud out of the infected cell to infect other cells [38]. A full cycle of replication occurs approximately once a day for actively replicating HIV.

1.1.2 HIV Adaptation to Immune System CTL Response

The cytotoxic T-cell (CTL) response is a mechanism from the adaptive immune system that destroys infected cells [19, 36]. CTL-mediated vaccines have been proposed for boosting CTL targeting of HIV infected cells to reduce viral load and thus risk of transmission [36]. In untreated patients, the efficacy of CTL response is

responsible for controlling the viral population after viremia and partially dictates the rate of progression to AIDS [70].

Viral proteins in an infected cell are degraded by normal cellular processes into peptides that are specifically bound by human leukocyte antigen (HLA) molecules and presented on the cell surface, signalling the CTLs to destroy the cell. The HLA molecules are encoded by alleles at the highly variable Major Histocompatibility Complex (MHC) class I loci in the human genome. Each HLA allele binds specific viral peptides that are known as CTL epitopes. However, HIV can evade the CTL response by developing 'escape' mutations within CTL epitopes that prevent it from being recognized and bound by the corresponding HLA molecule [36].

Following the standard nomenclature [11], we refer to HIV mutations known to be statistically associated with an HLA allele as 'matched' if they occur in a host carrying that allele. Matched mutations that are expected to decrease in frequency in that host are referred to as 'non-escape' alleles, and those expected to increase are 'escape' alleles.

With a rapid mutation rate of 10^{-4} mutations/site/day in a genome of size 9.5kbp [36], HIV infection rapidly diversifies and accumulates genetic differences within each host [96]. For an effective CTL-based vaccine, the vaccine epitopes used to induce CTL response must be designed such that mutations in the epitope result in viruses with poor survival fitness or replicative capacity. Detecting sites where CTL imparts a directional selective force on HIV can shed light on how to design vaccine epitopes to prevent evasion of CTL response.

We developed the software package Umberjack to explore HIV adaptation to the CTL response and how well it can be detected through evolutionary statistics.

1.1.3 HIV Adaptation to Drug Treatment

Wild-type HIV-1 enters host cells by binding to CCR5 co-receptors [18]. The drug Maraviroc inhibits HIV entry into host cells by binding to the host CCR5 co-receptor, which alters the 3D conformational structure such that the HIV envelope glycoprotein can no longer attach [18, 22]. Viruses with mutations that allow it to bind CXCR4 co-receptors instead of CCR5 proliferate under Maraviroc treatment, rapidly rendering the drug ineffective. In over half of untreated HIV1-B infected

patients, viral populations switch from CCR5 to CXCR4 co-receptor usage, followed by patient progression towards AIDS [72]. Clinical guidelines typically require that patient HIV populations be tested for prior CXCR4 co-receptor usage before the administration of Maraviroc, and only recommend that the drug be used as an alternative third agent in combination therapy [74].

The portion of the HIV genome responsible for co-receptor binding is predominantly encoded in the V3 region of the *env* gene. There is quite a bit of variance in conditions that lead to co-receptor usage. The few rules-based methods predicting co-receptor usage based on mutations at specific sites have high specificity but low sensitivity, thus the exact sites under directional selection and sequence of mutations can not easily be inferred a priori [50].

We employed the Umberjack software once again to profile within-host evolution in four patients that developed resistance mutations within days of using Maraviroc. The efficacy of Umberjack estimates as predictors for directional selection were compared between the untreated patient dataset and the treated patient dataset.

1.2 Phylogenetic Analysis With Short Reads

1.2.1 Next Generation Sequencing

Next-generation sequencing (NGS) is the massive parallelization of nucleic acid sequencing reactions that allows millions of DNA templates to be processed simultaneously. NGS amplicon sequencing has become a popular method of capturing genetic variation in pathogen populations due to its ability to detect variants at minority frequencies by producing a separate sequence for up to thousands of individuals in a population. Amplicons are DNA fragments isolated from a particular region of the genome and replicated using specific primers. Prior to NGS, viral populations were sequenced with Sanger bulk sequencing, which yielded a single consensus sequence for an entire population where different polymorphisms from different viruses would be indicated as a mixture of bases at a site [54]. Although Sanger sequencing yields longer reads ($\sim$ 1kbp), the consensus loses the evolutionary relationship between viruses since mutations cannot be attributed to any spe-

cific virus. Using NGS to obtain deep coverage of an amplicon can yield reads that cover the isolated region end-to-end, providing ideal input for phylogenetic analysis, the study of evolutionary relationships [84]. However, the typical read lengths with current NGS technologies with acceptable error rates remain short (<600 bp), imposing a severe limit on the width of genomic regions for which variation can be observed. Pyrosequencing, an NGS technology can produce close to 1kbp reads have error rates of higher than 1% and over 3% error rate in homopolymer regions with repeated nucleotides [62]. Third generation reads, such as Single-Molecule Sequencing in Real Time (SMRT) sequencing [25] and nanopore sequencing [7] can generate very long reads up to 25kbp long; however, they suffer from sequencing error rates of more than 15% [4].

Shotgun sequencing, another NGS method in which reads are generated by randomly fragmenting long DNA templates, allows efficient coverage of wide regions of a genome. However, only a fraction of the total reads will cover any particular region of the genome. The absence of sequence homology across the entire set of short reads makes it extremely difficult to reconstruct a phylogenetic tree, limiting the utility of shotgun data for phylogenetic analysis.

1.2.2 Phylogeny of a Single Population

A phylogeny of a single population displays the ancestry of individuals sampled from the population in an acyclic graph. Phylogenies are also referred to as trees. External nodes represent observed individuals within the population, and internal nodes represent ancestral individuals that must be inferred since they no longer exist. In Figure 1.1, the external nodes are A, B, C, D and internal nodes are E, F, G. An edge connecting two nodes, also known as a branch, indicates an ancestor-descendent relationship between individuals. A weighted edge indicates the amount of evolution or time between the individuals.

In a rooted tree, the ancestor of each individual can be inferred, and the root denotes the common ancestor of all individuals in the population. In an unrooted tree, the most common ancestor and thus root node are unspecified (Figure 1.1).

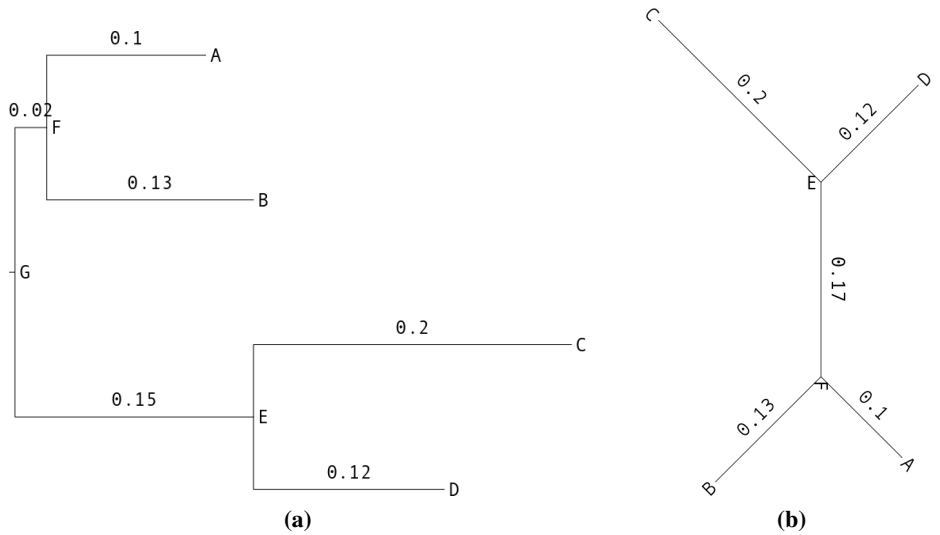


Figure 1.1: Example of phylogenetic trees. Nodes A, B, C, D are individuals observed within the population. Nodes E, F, G are inferred ancestral individuals. Node G represents the common ancestor of A, B, C, D. (a) Rooted phylogenetic tree. Common ancestor G is known. (b) Unrooted phylogenetic tree. Common ancestor G is unknown.

1.2.3 Multiple Sequence Alignment

As a precursor to building a tree, the genetic differences between each individual must be known. Genomic sites can be compared between individuals when the sequences are stacked in a multiple sequence alignment where each row is the sequence of an individual, and each column is a genomic site (Figure 1.2). The sequences are vertically aligned by maximizing the similarity between each sequence at each column. Similar sequences are inferred to be homologous, that is, descended from the same ancestor. It is possible that similar sequences actually descended from separate ancestors but share similarity due to convergent evolution from similar evolutionary pressures or due to chance.

A multiple sequence alignment requires comparing each sequence against all other sequences resulting in $O(N^2L^2)$ complexity, where N is the number of sequences and L is the length of the sequence [57]. One heuristic towards multiple sequence alignment is to align each sequence against a reference. In algorithms

where the reference is hashed, such as BWA-mem, pairwise alignment against a reference is typically much faster with complexities of roughly $O(NL/k)$, where k is the size of a word within the sequence to look up in the reference [52]. This heuristic can be a preferred solution for multiple sequence alignment of NGS libraries, which can contain millions of reads. Using the pairwise alignment positions with respect to the reference, we can obtain 'pseudo' multiple sequence alignments in which each column represents a position in the reference sequence. For example, if a reference sequence is AAA and read sequences are AGG and AGT, the alignments against the reference will be:

Reference	AAA	Reference	AAA
Read1	ACG	Read2	AGT

Using the positions with respect to the reference, we obtain the 'pseudo' multiple sequence alignment of the reads:

Read1	ACG
Read2	AGT

However, pairwise alignments against a reference lose information regarding the multiple sequence alignment of insertions with respect to the reference. For example, if a reference sequence is AAA and read sequences are ACGTAA and AGGAA, pairwise alignment against the reference will be:

Reference	A---AA	Reference	A--AA
Read1	ACGTAA	Read2	AGGAA

The alignments against the reference do not yield any information regarding the positions of the inserted bases with respect to other inserted bases. Thus, we can not discern the best alignment from the potential multiple sequence alignments of the inserted bases:

Read1	ACGTAA	Read1	ACGTAA	Read1	ACGTAA
Read2	AGG-AA	Read2	AG-GAA	Read2	A-GGAA

1.2.4 Difficulties in Phylogenetic Reconstruction of HIV

One of the ways in which HIV populations maintain genetic diversity is through recombination. HIV experiences recombination when its reverse transcriptase jumps between strands during RNA replication. In order for the recombination to be genetically noticeable, either strand must be genetically different from the other in the sections that swap. This only happens in 'hybrid' progeny viruses created by combining an RNA strand from 2 different strains that infect the same cell [10]. In HIV, the rate of genetically noticeable recombinations is reported to be 10^{-5} recombinations/bp/day [77]. Recombination affects reconstruction of evolutionary reconstruction by altering the true topology on either side of a genomic breakpoint. Thus, a tree reconstructed from an alignment containing a breakpoint will be different from the true trees on either side of the breakpoint [94].

Further, the HIV effective population size can be quite large. The number of genetically unique replicating viruses in an HIV population can reach 10^3 to 10^5 within a patient [63, 79], requiring the same number of overlapping sequences to represent each virus in a multiple sequence alignment.

1.2.5 Phylogenetic Reconstruction Techniques

The topology and branch lengths of a phylogenetic tree can be reconstructed by several algorithms such as neighbour joining [93] and maximum likelihood [28]. Although time consuming, maximum likelihood approaches yield far more accurate trees than neighbour joining [17], especially when there is missing sequence data [104].

Neighbour Joining

Neighbour joining is a distance-based approach to phylogenetic reconstruction. It employs a dynamic programming algorithm in which pairs of individuals are recursively clustered together to form a new subtree until all observed individuals and inferred ancestral individuals are included in a single tree. In each iteration the pair of unconnected individuals with the minimum pairwise genetic distance is connected to a new node representing their inferred common ancestor. Branch lengths from the new node to either of the pair is set to a function of the pair's

genetic distance.

Genetic and Evolutionary Distance

The genetic distance is a function of the probability of mutations at each site. There are many mathematical models for calculating genetic distance that take into account various biological biases of mutation. The general time reversible (GTR) nucleotide mutation model estimates a separate parameter for the symmetric mutation rates between each pair of nucleotides [102]. The Jukes Cantor model is a special subset of the GTR which assumes each nucleotide has the same probability and the mutation between each nucleotide occurs at the same rate [44]:

$$P(i|k,t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & i = k \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & i \neq k \end{cases} \tag{1.1}$$

where $P(i|k,t)$ is the probability that nucleotide k mutates to nucleotide i after time t. μ is the nucleotide mutation rate. The Jukes Cantor model is too simplistic to realistically represent HIV evolution, but it allows for quicker calculations in initial stages of phylogenetic reconstruction algorithms.

Maximum Likelihood

The maximum likelihood approach attempts to selects the tree $T = T_{max}$ that maximizes the likelihood of tree T given multiple sequence alignment D. The likelihood of T given D is equivalent in terminology to the probability of D given T, $P(D|T)$.

Given tree T, the likelihood of T given the sequences at site u (also known as the site-likelihood) can be calculated recursively by traversing the tree in post-order (tips-first, where children are visited before parents). The site-likelihood of the subtree rooted at a node is a function of the site-likelihood of the subtrees rooted by its children.

At site u, the probability that node k has sequence character s_k is the sum of the probabilities of mutations from s_k to all possible characters at the child nodes.

$$P(k|s_k) = \sum_{s_{\text{left}}, s_{\text{right}}} P(s_{\text{left}}|s_k, t_{\text{left}}) P(\text{left}|s_{\text{left}}) P(s_{\text{right}}|s_k, t_{\text{right}}) P(\text{right}|s_{\text{right}}) \quad (1.2)$$

where $P(k|s_k)$ is the probability that node k has sequence character s_k at site u. Node k has children *left* and *right* with sequence character states s_{left} and s_{right} respectively. $P(s_{\text{left}}|s_k, t_{\text{left}})$ is the probability that character s_k mutates to child character s_{left} after the time elapsed on branch t_{left} connecting node k and child left. The mutation probability typically follows a mutational model as described in Section 1.2.5.

The site-likelihood of the entire tree is the sum of the probabilities that the root node has sequence character s_k for every possible s_k. Since each site is considered independent of other sites, the likelihood of T given D can be expressed as the product of the likelihoods of T at each site.

The search over all possible trees to find the tree T_{max} that maximizes the likelihood can be computationally expensive. Search times can be shortened with greedy approaches that stop searching when newer trees no longer increase likelihood according to some threshold.

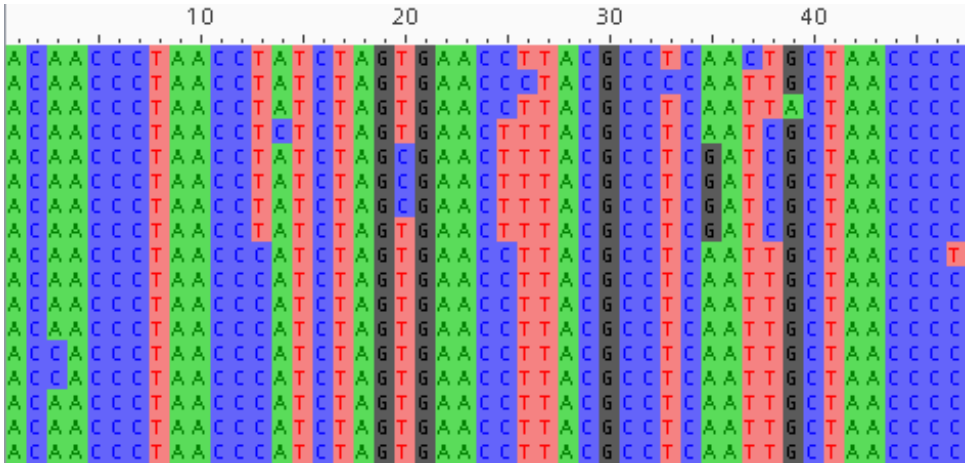


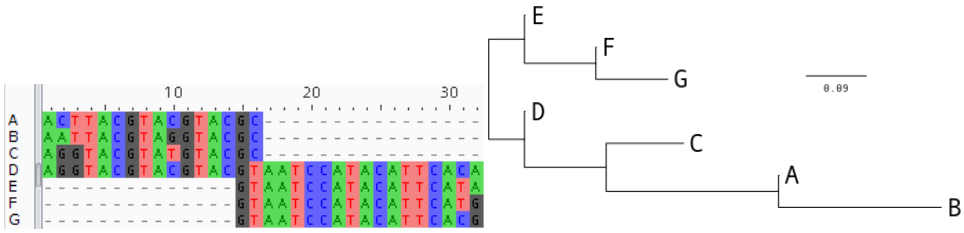
Figure 1.2: Example of multiple sequence alignment. Each row represents the genetic sequence of an individual in the population. Each column is a position in the organism's genome.

1.2.6 Tree Error from Missing Data

Missing data in multiple sequence alignments is handled differently by different phylogenetic reconstruction approaches. Neighbour joining ignores sites with missing data when calculating pairwise genetic distance. Maximum likelihood sums over the likelihood of all possible characters for missing data [87].

Long branch attraction is an issue common to most phylogenetic reconstruction approaches and occurs when there are wide swaths of missing data for some but not all individuals in a multiple sequence alignment. If there is little overlap in alignment between two sets of individuals, most phylogenetic reconstruction approaches would force either sets of individuals into separate elongated lineages regardless of their true relationship (Figure 1.3) [104].

Past studies of phylogenetic reconstruction on missing data have shown that the percentage of missing data was not as important to tree accuracy as whether there were sites in the multiple sequence alignment with 'informative' sequence from all individuals. Informative sites contain sufficient sequence diversity to distinguish between individuals but evolve slowly enough such that mutations can be captured within the sampling timeframe [105].



(a) (b)

Figure 1.3: Long branch attraction due to missing data. (a) Multiple sequencing alignment with 2 partitions of missing data. (b) Corresponding tree with 2 long lineages.

1.2.7 Current Approaches to Phylogenetic Analysis From Missing Data

Filling and Trimming Alignments

In multiple sequence alignment of genomes between species, unsequenced or biologically missing genes can either be filled in with known sequences from databases, or all genomes trimmed such that only sites common to most species remain [13]. When multiple populations are shotgun sequenced with separate libraries, inter-population evolutionary relationships can be found through phylogenetic reconstruction of population consensus sequences [54]; however, within-population evolutionary relationships will be lost.

Amplicon Sequencing

Studies of viral populations frequently sequence amplicons that are narrow enough to be covered almost entirely by reads, which reduces sites in the multiple sequence alignment with missing data across individuals [68, 84]. In order to measure evolution across a wider range of the genome, overlapping or adjacent amplicons can be sequenced [107].

PCR primers that amplify a particular portion of the genome require regions on either side of the amplicon that are conserved across the population. As a result, amplicon primer design can be a limiting step for diverse regions of the genome or when studying communities that may not share similar genomic features across species or strains [35].

Phylogenetic analysis options for shotgun sequencing are minimal as demonstrated by a recent viral population study in which overlapping amplicons across the genome for an HIV population were shotgun sequenced [107]. Even though sequencing was available across the entire genome, phylogenetic analysis was only available for the amplicons narrow enough to fit a full read.

Inferring Missing Data

ForeSeqs is newly released software that infers missing data in a multiple sequence alignment and corrects branch lengths on phylogenies made from incomplete data

[16]. It requires a multiple sequence alignment, the corresponding phylogeny, and a user-supplied list of site boundaries delineating partitions of missing data in the multiple sequence alignment. As such, ForeSeqs is more suited for situations in which gaps are at the same site for multiple taxa, such as biologically missing genes.

ForeSeqs uses partitions of the alignment with complete data to scale branch lengths for taxa with missing data. Unknown sequence characters are inferred by simulating evolution down the tree or through marginal likelihood ancestral reconstruction in which the sequence at a node is selected to maximize the likelihood of the connecting node.

Since ForeSeqs does not correct the topology of the input tree, any errors from the initial phylogenetic reconstruction due to missing data will have knock on effects on the reconstruction of missing data. This proves problematic in its application to NGS shotgun libraries. An initial phylogenetic reconstruction from shotgun multiple sequence alignments would tend towards a star topology since each read shares little sequence homology with most other reads (Figure 1.4). For example, the multiple sequence alignment of shotgun libraries analyzed in Chapter 3 ranged between 70-85% missing data and none of the sites contained sequence from $>$ 50% of reads.

Inferring Haplotypes From Single Nucleotide Polymorphisms

A haplotype is the sequence along a single DNA strand. By aligning reads to a reference genome, the nucleotide variants with respect to the reference can be used to deconvolute the separate haplotypes within a population. Through reconstruction of the haplotypes, we fill in the missing gaps within a multiple sequence alignment of shotgun reads. Typically, the variants are modelled as a mixture model, and the variants belonging to the same haplotype are linked together by clustering. Existing software accomplish haplotyping for cancer genomics in which tumour cells are made up of multiple subclonal populations, as well as for viral populations [49, 85, 106]. However, haplotype reconstruction software are limited to detecting 10^1 to 10^2 haplotypes within a population, whereas the number of unique replicating HIV haplotypes can reach 10^3 to 10^5 within a patient [63, 79].

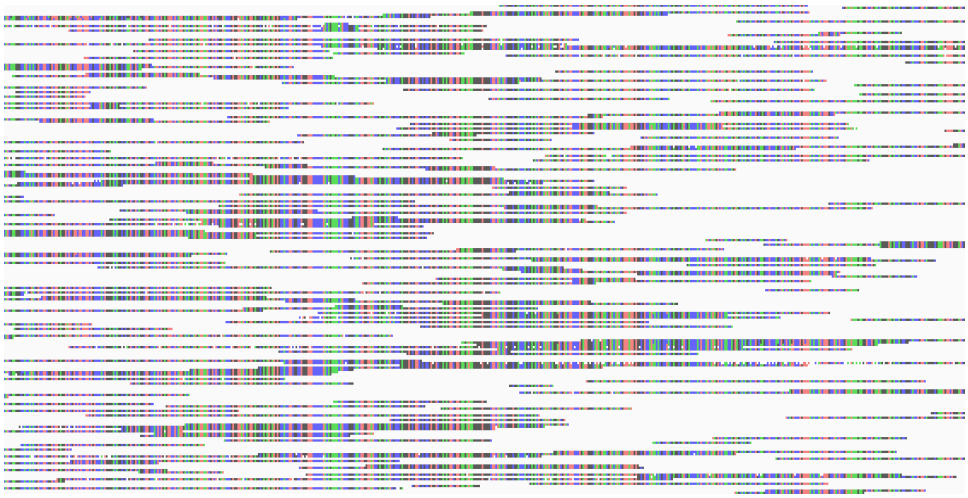


Figure 1.4: Subset of a multiple sequence alignment of typical shotgun sequencing reads from a simulated viral population. Colored blocks represent nucleotide bases within a read. White spaces indicate genomic positions uncovered by the read. This multiple sequence alignment contains over 80% missing bases.

1.2.8 New Developments to Overcome Missing Data

We developed the software pipeline Umberjack to address the problem of excess missing data in shotgun sequencing for phylogenetic analysis. Umberjack uses a sliding window technique to obtain windows of alignment with sufficient sequence overlap and homology at all sites to reconstruct a tree (Figure 2.1). A window refers to a slice of a multiple sequence alignment containing only a subset of sequences and a subset of contiguous sites. Using the reconstructed evolutionary history from each window, Umberjack detects sites under selection by aggregating over evolutionary statistics from overlapping windows. Chapter 2 details Umberjack implementation and its ability to accurately output evolutionary statistics from shotgun sequencing data.

1.3 Detecting Selection

1.3.1 Importance of Selection and Evolutionary History

Selection refers to the deterministic change in frequency of individuals in a population due to their traits that improve or diminish their capacity to reproduce and survive in their environment. At the molecular level, selection operates correspondingly to change the frequency of genetic variants that code for the traits. Selection generally acts more strongly at genomic positions that have a high functional impact. Thus, detecting genetic signatures of selection can provide the basis for diagnostic tools and identify functionally important sites. For example, potential genetic markers for drug resistance in *Mycobacterium tuberculosis* have been found by calculating the strength of selection [27].

Leveraging the evolutionary history of pathogens can identify potential targets for vaccines and treatments. The immune system and drug treatment act as selective forces on pathogens. Antibody-based vaccines induce the immune system to produce proteins called antibodies that bind and neutralize foreign molecules called antigens. In a study of untreated patients that were naturally able to control their Human Immunodeficiency Virus (HIV) infections, potential vaccine antigens were discovered by tracing the series of HIV mutations that subsequently stimulated the production of antibodies that neutralized a broad range of HIV variants [55].

Although taking frequent longitudinal samples is the most accurate way of determining evolutionary history, resource constraints make this infeasible. Phylogenetic and ancestral reconstruction seek to fill in the gaps in evolutionary history when samples are unavailable.

1.3.2 Evolutionary Statistics

Umberjack generates several statistics that describe the rate of selection at a codon site, including $dN - dS$, I, and Shannon entropy. The first two statistics are phylogenetically informed and take into account the evolutionary history of a population, leading to less confounding between demographic events and actual selection.

Selection statistics that are frequency based, such as Shannon Entropy, are af-

fected by demographic events. Fixation of genetic variants may be due to a population size bottleneck during which a single founder individual passes on its genetic variants to its progeny. This founder effect will continue if there is no selection and there has been insufficient time for random mutations to remove or reduce the genetic variant from the population. Further, a growing population that is undergoing neutral evolution with no selection will experience increasing genetic diversity over time simply due to random mutations [91].

Most of the statistics label sites as under diversifying or purifying selection. With regards to within-population evolution at the molecular level, diversifying selection indicates a proliferation of polymorphisms. Purifying selection indicates the removal of variants unfit for survival in the population's environment. Directional selection indicates only genotypes coding for a single specific trait survive.

$dN - dS$

Diversifying or purifying selection at a site can be measured by the $dN - dS$ statistic, where dN is the rate of nonsynonymous substitutions and dS is the rate of synonymous substitutions. A site is diversifying if $dN - dS > 0$, purifying if $dN - dS < 0$, and neutral if $dN - dS = 0$:

$$dN - dS = \frac{N}{E[N]} - \frac{S}{E[S]}$$

The total nonsynonymous substitutions N and total synonymous substitutions S are normalized by their respective expected numbers, $E[N]$ and $E[S]$, based on the genetic code [80]. For example, all possible mutations at each nucleotide position of codon ATG leads to a different amino acid translation. Thus, each nucleotide position in ATG is expected to be 100% nonsynonymous. The $dN - dS$ statistic was chosen instead of the typical dN/dS ratio [34], since the latter is more numerically unstable when dS is small.

I

We developed a new metric of site selection inspired by a statistical test of neutrality proposed by Fu and Li [32]. Their test is based on the distribution of substitution events that have been mapped to a phylogeny. We assume that substitutions on the

tips of the phylogeny are more recent in time. Under purifying selection, substitutions should tend to be observed on the tips because they represent lineages that have not yet been removed by selection. Conversely, if a greater proportion of substitutions are mapped to internal branches of the tree, then the variants from those ancestral lineages have been allowed to proliferate under diversifying selection.

A simple metric of site diversifying and purifying selection can therefore be calculated from the numbers of substitutions at a site that map to internal (S_I) and external branches (S_E), normalized by their respective branch lengths t_I, t_E. This follows a similar approach as Fu and Li; however, we use branch lengths for normalization instead of the Fu and Li mathematical simplifications based on binary phylogenies where each non-root node has at most two child nodes [32]. The metric I ranges in [0, 1], where $I = 0.5$ indicates neutral evolution, $I > 0.5$ indicates diversifying selection, and $I < 0.5$ indicates purifying selection.

$$I = \frac{\frac{S_I}{t_I}}{\frac{S_I}{t_I} + \frac{S_E}{t_E}}$$

Ancestral lineages that maintain a specific amino acid provide evidence of directional selection at the amino acid level. A nonsynonymous substitution to a favoured amino acid in an ancestral individual followed by synonymous substitutions in its descendants could indicate that the lineage is undergoing directional selection. An existing amino acid preserved through synonymous substitutions along a lineage except for nonsynonymous substitutions at the tips indicates that purifying selection has not yet had time to excise the spurious mutations.

Breaking down substitutions as nonsynonymous and synonymous, we can measure directional selection at the amino acid level by calculating the difference between nonsynonymous substitutions and synonymous substitutions at internal versus external branches, $I_N - I_S$. $I_N - I_S$ ranges in [-1, +1], where $I_N - I_S = \pm 1$ implies strong directional selection and $I_N - I_S \approx 0$ implies no or weak selection.

$$I_N - I_S = \frac{\frac{S_{IN}}{t_I}}{\frac{S_{IN}}{t_I} + \frac{S_{EN}}{t_E}} - \frac{\frac{S_{IS}}{t_I}}{\frac{S_{IS}}{t_I} + \frac{S_{ES}}{t_E}}$$

where S_{IN} is the number of nonsynonymous substitutions mapping to internal

branches, S_{IS} is the number of synonymous substitutions mapping to internal branches, S_{EN} is the number of nonsynonymous substitutions mapping to external branches, S_{ES} is the number of synonymous substitutions mapping to external branches.

It is possible that a site indicated as under selection by $I_N - I_S$ is actually neutrally evolving but appears as selected because it is located in a selective sweep in which sites adjacent to selected sites are dragged to fixation. Although a favoured mutation imparting a selective advantage will be passed to progeny and increase in frequency, the entire genomic backbone containing the mutation may not be passed due to recombination. RNA sections are swapped between strands during recombination, breaking up the linkage of genomic sequence on either side of the breakpoint. Since sites close in proximity are less likely to segregate to separate strands, the genomic sequence adjacent to a favoured mutation will be passed to progeny and increase in frequency simply due to genomic linkage. Variants at neutral sites in a selective sweep will remain fixed in frequency until sufficient time has passed for mutation and recombination to occur [40].

Entropy

The distribution of variant frequencies can be used to infer selection without a phylogeny. For example, high frequency of a variant indicates purifying selection, and a wide distribution of variant frequencies indicates diversifying selection. In addition to avoiding computational efforts of phylogenetic and ancestral reconstruction, calculating variant frequencies can be performed on any type of sequencing.

Shannon entropy is a commonly used summary metric to indicate the distribution of variant frequencies at a genomic site. Shannon entropy ranges from [0, ∞), where 0 entropy indicates all individuals have the same variant at a site:

$$H(X) = \sum_{i=1}^{n} P(x_i) log_2 P(x_i)$$

where $H(X)$ is the Shannon Entropy of site X. x_i is ith variant at site X. $P(x_i)$ is the probability of observing variant i at site X. n is the total number of unique variants at site X.

Metric entropy is a normalized version of Shannon entropy with values in [0,

1], which allows comparison of entropy between sites that may be covered by a different number of reads:

$$M(X) = \frac{\sum_{i=1}^{n} P(x_i) log_2 P(x_i)}{log_2 n}$$

where $M(X)$ is the Metric Entropy at site X.

Chapter 2

Phylogenetic Analysis With Sliding Windows

2.1 Proposal of Sliding Windows to Overcome Missing Data

Parameters that require a phylogeny for accurate estimation, such as rates of substitution, can be progressively estimated from a series of phylogenies along an alignment. This approach is similar to smoothing or 'sliding window' methods for estimating a summary statistic over a series of data points, except that data within a window is mapped into tree space from which the statistics are derived. For example, sliding windows of phylogenies along alignments is a popular approach for detecting recombination [66, 94].

The Umberjack software pipeline applies the sliding window approach towards phylogenetic analysis of shotgun reads. To reduce missing data used in phylogenetic reconstruction, windows of multiple sequence alignment are selected such that all sequences have sufficient overlap as defined by the user (Figure 2.1). To evaluate evolutionary statistics at each site along the genome, Umberjack aggregates the statistics from overlapping windows. Although the trees from each window will vary depending on the sequence diversity and individuals included within each window, averaging the site evolutionary statistics over each window replicate

has been tested on simulated data to be robust as long as there is sufficient sequence diversity to provide a phylogenetic signal in the windows (Section 2.6.5).

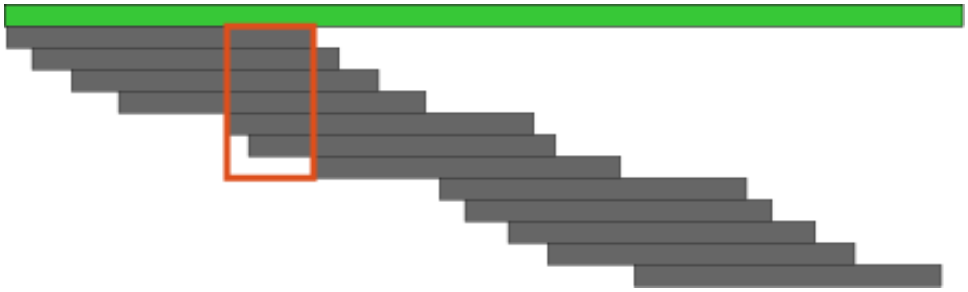


Figure 2.1: Depiction of multiple sequence alignment of shotgun reads. Reference sequence in green. Shotgun reads in grey. Slice of alignment in orange. White spaces indicate missing sequence homology between reads.

2.2 Software Details

2.2.1 Umberjack Use Cases

Umberjack is primarily designed to efficiently process extremely high coverage next-generation sequencing of a population with a reference genome with minimal repetition and known open reading frames. As such, Umberjack is ideal for studying populations of RNA viruses such as HIV, as they typically have little genome duplication but display a large amount of diversity that requires deep coverage to capture [48, 73, 98].

2.2.2 Umberjack Computational Requirements

Umberjack is Message Passing Interface (MPI) enabled, allowing it to be run across multiple nodes in a cluster. Although Umberjack is written in Python and R, which are supported on multiple platforms, it has only been tested on Linux environments.

Since Umberjack's aim is to rapidly process many overlapping windows of phylogenetic analysis simultaneously, its implementation favours speed over accuracy. For example, processing 70 windows for a single sample library for a

genomic region of 3kbp sequenced at 7000x with MiSeq Nextera (shotgun) reads takes roughly 17 minutes on a cluster with 50 CPU cores spread across 4 nodes.

2.2.3 Umberjack Availability

Umberjack and the third party software it is packaged with are publicly available for download under a GNU General Public License (GPL): https://github.com/cfe-lab/Umberjack.

2.3 Umberjack Implementation

The Umberjack pipeline processes reads to form a 'pseudo' multiple sequence alignment and slices windows out of the alignment. Each window is sent for phylogenetic reconstruction, ancestral sequence reconstruction, and calculation of evolutionary statistics. The evolutionary statistics from each window are aggregated into per-site evolutionary statistics. The pipeline is further detailed in the following sections.

2.3.1 Read Processing

As input, Umberjack requires pairwise read alignments against a reference in SAM format [53] or a multiple sequence alignment in Fasta format. Umberjack uses the alignment information specified in the SAM files (positions with respect to the reference, and aligned characters and qualities) to generate a 'pseudo' multiple sequence alignment of the reads (Section 1.2.3).

One technique for ensuring that majority consensus insertions with respect to a reference genome are included in the 'pseudo' multiple sequence alignment is to employ an iterative approach to pairwise alignment against a reference, using the previous iteration's consensus as the reference in the next iteration (Section 3.2.2). The iterations stop when the consensus no longer changes from the previous iteration. We used this iterative approach when applying Umberjack to our datasets. However, Umberjack itself does not perform any alignments.

The majority of evolutionary statistics reported by Umberjack are substitution based. Insertions that do not occur within the majority of the population will have gaps in the multiple sequence alignment for all the individuals lacking the inser-

tion; that is the majority of the population will have missing data at the insertion sites. If there are no substitutions within the insertions, the substitution-based evolutionary statistics will not be able to discern evolution at those sites. However, detecting the total number and locations of insertions can be a way to detect evolution and is worthy of exploring in future releases of the software. Further, we note that alignment errors from indels or otherwise are a potential bias in Umberjack estimates of evolution in that they would show more diversity where there is none.

When paired-end or mate-pair reads are used, Umberjack merges the mates into a single read using the mate alignment positions reported in the SAM file. Both paired-end and mate-pair reads are reads that are sequenced from both ends of a DNA fragment. In the case of paired-end reads, the DNA is sequenced inwards starting from the ends. Mate-pair reads are sequenced outwards starting from a position from within the DNA fragment. SAM alignment positions are taken at face value without consideration that mates may align to duplicate portions of the reference. This simple and rapid merging procedure is sufficient for RNA virus populations which experience minimal genome repetition. To handle organisms that do experience genome repetition, users may use their own multiple sequence alignments as input. Further, reads that contain the exact same sequence are removed, since they could potentially be PCR (Polymerase Chain Reaction) duplicates in which the same DNA fragment from the same individual is amplified, yielding no additional information regarding the population.

All bases that do not meet a minimum Phred-like quality score are masked with N's. A Phred quality score Q is a log transform of the probability of base error determined from internal benchmarking by the sequencing technology [9, 26]:

$$Q = -10 log_{10} P_{\text{err}}$$

where P_{err} is the probability that a sequenced base is erroneous. Discordant bases from overlapping mates are resolved by accepting the base with the highest quality score, and are masked with N's if both mates do not reach the minimum quality score, Q_{min}. Concordant bases from overlapping mates are allowed a lower Phred

quality score before N-masking, Q'_{min}.

$$\begin{aligned} P(M_1 = M_2, M_{1\text{err}}, M_{2\text{err}}) &= P(M_1 = M_2 | M_{1\text{err}}, M_{2\text{err}}) P(M_{1\text{err}}) P(M_{2\text{err}}) \\ &= \frac{3}{9} P(M_{1\text{err}}) P(M_{2\text{err}}) \\ Q'_{\text{min}} &= -10 log_{10}(3) + Q_{\text{min}} <= Q_{M1} + Q_{M2} \end{aligned}$$

where $P(M_1 = M_2, M_{1\text{err}}, M_{2\text{err}})$ indicates the probability that mate 1 and mate 2 are concordant and erroneous. $P(M_1 = M_2 | M_{1\text{err}}, M_{2\text{err}})$ indicates the probability that mate 1 and mate 2 are concordant given that mate 1 and mate 2 are erroneous. $P(M_{1\text{err}})$ indicates the probability that mate 1 is erroneous. Q'_{min} is the new minimum quality threshold for concordant overlapping bases. Q_{min} is the user-defined minimum quality threshold for a base. Q_{M1} is the quality score for mate 1 which is dictated by the probability that mate 1 is erroneous. There are 3 permutations of concordant erroneous overlapping bases and 9 permutations of erroneous overlapping bases.

The revised quality score cutoff makes the assumption that the probability of error at each mate is independent of the other mate within a read and that the probability of error towards each base is the same, which may not necessarily be true depending on the sequencing technology [95]. However, it provides a rough approximation of the increase in confidence of concordant overlapping bases that does not require additional user configuration based on sequencing technology.

2.3.2 Window Extraction

The width of windows and the amount of overlap between windows are user configurable. Sequences that do not meet a user defined threshold for minimum percentage of non-N, non-gap bases within the window are excluded from that window. Windows that do not meet user defined thresholds for minimum number of sequences are excluded from further phylogenetic analysis. Sequencing error has the effect of falsely increasing the rate of nonsynonymous mutations, since the majority of mutations are nonsynonymous. We attempt to reduce the amount of sequencing error making it to the end analysis by masking bases that have low quality scores (as configured by the user). Thus, bases that are likely to be erroneous will

be treated as missing data instead of influencing the mutation counts.

2.3.3 Phylogenetic Reconstruction

Umberjack comes packaged with FastTree2 [88] to perform phylogenetic reconstruction. FastTree2 provides a rapid method of phylogenetic reconstruction that has been benchmarked favourably in terms of speed with minimal drop in accuracy compared to other maximum likelihood phylogenetic reconstruction software [88].

First, FastTree2 calculates the pairwise distance between sequences using a simple Jukes Cantor model that takes into account multiple substitutions per site (Equation 1.1). When FastTree2 encounters an N or a gap, it uses the distance information from the rest of the sequence to infer evolutionary distance [87], and weights sequences by the number of positions without missing data. Using the Jukes Cantor pairwise genetic distances, FastTree2 generates an initial tree using neighbour joining (Section 1.2.5).

The initial tree topology and branch lengths are further refined using an approximate maximum likelihood phylogenetic reconstruction. Instead of finding the tree that yields the maximum likelihood (see Section 1.2.5), FastTree2 uses heuristics that reduce search time by only examining a subset of trees. The tree with the highest likelihood within the subset is selected, but it may not be the maximum likelihood tree amongst all possible trees. FastTree2 performs up to two sequential rounds of optimizations for topology and branch length.

FastTree2 refines tree topology using nearest neighbour interchange [20]. Traversing through the tree in post-order traversal (tips-first, where children are visited before parents), the subtree rooted at the current node is perturbed by swapping the sibling and child nodes. The trees resulting from the swappings are calculated for likelihood (see Equation 1.2).

After optimization of the tree topology, the branch lengths are refined using the GTR nucleotide mutation model (Section 1.2.5). Since different sites might evolve at different speeds, the set of mutation rates at each site are scaled by a site specific evolution scaling factor. Each site specific evolution scaling factor is selected from the category of a discrete 20-category gamma distribution that maximizes the likelihood.

2.3.4 Reconstructing Ancestral Sequences

In order to calculate $dN - dS$ and I, we need to reconstruct the sequences of ancestral individuals to count mutations throughout a population's evolutionary history. Umberjack uses HyPhy to perform maximum likelihood ancestral reconstruction in which the codon sequence of each ancestral individual is inferred from a given tree and multiple sequence alignment [82].

HyPhy first estimates the GTR nucleotide mutation model (Section 1.2.5) that maximizes the likelihood of the sequences given the tree. Using the fixed GTR model, HyPhy estimates the Muse Gaut codon mutation model [76] that maximizes the likelihood of the sequences given the tree (Section 1.2.5). The Muse Gaut codon mutation model (MG94) specifies that the probability of a codon mutating to another codon depends on the nucleotide mutation rate, the ratio of nonsynonymous to synonymous mutations, and whether the mutation is nonsynonymous or synonymous. Stop codons indicating the end of transcription (TAG, TAA, TGA) are excluded from the codon mutation model and disallowed during ancestral reconstruction with the reasoning that a premature stop codon would result in a defective organism. Umberjack has a user configurable option to mask stop codons in window alignments with NNN to allow ancestral reconstruction to proceed.

Since the input phylogenetic tree lengths are in units of nucleotide substitutions per nucleotide site, the tree needs to be scaled to nucleotide substitutions per codon site using the branch length scaling factor B. B is estimated at the same time as the codon model via maximum likelihood.

The ancestral sequences are reconstructed using a joint maximum likelihood approach based on the Pupko dynamic programming algorithm for ancestral reconstruction [89] in which the ancestral sequences are selected to maximize the likelihood of all nodes for a fixed tree.

The ancestral character at each site is calculated independently of other sites. For each site, HyPhy traverses each node starting from the external nodes towards an arbitrarily chosen root node. At each non-root node L_b, it evaluates the likelihood of every possible character s_a at its parent node L_a. It keeps track of the sequence character s_b for the current node L_b that maximizes the likelihood given

that the parent character is s_a.

$$P(L_b|L_a = s_a) = P(s_b|s_a, t_b) \prod_{c \in C} P(L_c|L_b = s_b)$$

where $P(L_b|L_a = s_a)$ denotes the likelihood of node L_b given that the parent node L_a has character s_a. Node L_b has the set of child nodes C. $P(s_b|s_a, t_b)$ denotes the probability that character s_a will mutate to character s_b after time t_b. The time t_b is obtained from the branch length connecting the node L_b to its parent L_a.

The root character is selected to maximize likelihood of the root:

$$P(L_{root}|L_{root} = s_{root}) = P(s_{root}) \prod_{c \in C} P(L_c|L_{root} = s_{root})$$

where $P(L_{root}|L_{root} = s_{root})$ is the likelihood of the root given that it has character s_{root}. The root has the set of child nodes C. $P(s_{root})$ is the prior probability of the character s_{root}.

The final set of ancestral characters at each site are found by traversing the tree from root towards the tips, and selecting the node character that yields the maximum likelihood for the parent node.

2.3.5 Evolutionary Statistics

Umberjack uses a custom HyPhy Batch Language implementation of the Single Likelihood Ancestor Counting (SLAC) [78, 81] method for calculating $dN - dS$ (Section 1.3.2). HyPhy allows programmers to utilize its functionality via its APIs (application programming interfaces) written in HyPhy Batch Language. Although HyPhy contains an implementation of SLAC, we created a custom SLAC implementation for cleaner integration with Umberjack's workflows, including constraining tree topologies to those obtained from FastTree2 and ensuring consistent clade names. Most importantly, Umberjack's implementation of SLAC avoids HyPhy's built-in resolution of ambiguous codons, which improved accuracy in $dN - dS$ estimations from NGS shotgun sequencing data which has more N's and gaps than the population mixture Sanger sequencing data that HyPhy was originally designed to handle.

For each codon site, SLAC calculates $dN - dS$ by counting the nonsynonymous

and synonymous substitutions along branches in the phylogeny, and normalizes the counts by the expected nonsynonymous and expected synonymous substitutions.

dN is the rate of nonsynonymous substitutions per expected nonsynonymous substitutions at a codon site. dS is the rate of synonymous substitutions per expected synonymous substitutions at a codon site.

$$dN = \frac{N_T}{E_T[N]}$$

$$dS = \frac{S_T}{E_T[S]}$$

where N_T is the total nonsynonymous substitutions in tree T, and $E_T[N]$ is the total expected nonsynonymous substitutions T. S_T is the total synonymous substitutions T, and $E_T[S]$ is the total expected synonymous substitutions T.

The method to calculate dS is analogous to the method for dN. We will focus only on nonsynonymous substitutions for simplicity.

Counting Nonsynonymous Substitutions

The nonsynonymous substitutions are counted along the branches in the tree. If the path between the parent and child codon of a branch requires more than one nucleotide mutation, the branch nonsynonymous substitutions are calculated as the average nonsynonymous substitutions across all possible paths between the parent and child codon with 1-nucleotide mutations at each step.

$$N_T = \sum_{b \in T} \frac{1}{|M|} \sum_{m \in M} N_m \tag{2.1}$$

where branch b is in tree T. M is the set of paths from the parent to child codon of branch b. Each path m consists of steps of 1-nucleotide mutations, where mutations to stop codons are disallowed. N_m is the total nonsynonymous mutations during path m.

For example, if branch b had parent codon AAG and child codon GAA:

$$\begin{aligned} M =& \{\text{AAG} \xrightarrow{\text{1 synonymous}} \text{AAA} \xrightarrow{\text{1 nonsynonymous}} \text{GAA}, \\ & \text{AAG} \xrightarrow{\text{1 nonsynonymous}} \text{GAG} \xrightarrow{\text{1 synonymous}} \text{GAA}\} \\ N_b =& \frac{1}{|M|}(N_{\text{AAG}\to\text{AAA}\to\text{GAA}} + N_{\text{AAG}\to\text{GAG}\to\text{GAA}}) \\ =& \frac{1}{2}(1+1) \\ =& 1 \end{aligned}$$

Counting Expected Nonsynonymous Substitutions in a Tree

The expected nonsynonymous substitutions along a tree is weighted by each branch length:

$$E_T[N] = \frac{1}{|T|}\sum_{b\in T} t_b E_b[N] \tag{2.2}$$

where branch b has branch length t_b in tree T. $E_b[N]$ is the expected nonsynonymous substitutions in branch b. $|T|$ is the total length of tree T.

Counting Expected Nonsynonymous Substitutions in a Branch

The expected nonsynonymous substitutions along a branch is an average of the expected nonsynonymous mutations across all paths from the parent to child codon. Each path from the parent to child codon consists of steps of 1-nucleotide mutations where stop codons are disallowed.

$$E_b[N] = \frac{1}{|M|}\sum_{m\in M} E_m[N]$$

where $E_b[N]$ is the expected nonsynonymous substitutions in branch b. M is the set of paths from the parent to child codon. $E_m[N]$ is the expected nonsynonymous mutations in path m from the parent codon to child codon.

The expected nonsynonymous substitutions in a path, $E_m[N]$, is an average of

the expected nonsynonymous substitutions across each codon in the path:

$$E_m[N] = \frac{1}{|C|} \sum_{c \in m} E_c[N]$$

where path m is comprised of the set of codons resulting from 1-nucleotide mutation steps from the parent to the child codon, including the parent and child codon themselves.

Counting Expected Nonsynonymous Substitutions in a Codon

The expected nonsynonymous mutations in a codon is the sum of the expected nonsynonymous mutations at each codon position. The expected nonsynonymous mutations at each position ranges in [0, 1]. Thus the total expected nonsynonymous mutations in a codon ranges in [0, 3].

$$E_c[N] = \sum_{i \in c} E_i[N] = \sum_{i \in c} \sum_{j \neq i} w_{ij} N_{ij}$$

where i is a nucleotide at one of the three nucleotide positions in the codon. w_{ij} is a weight assigned to the mutation $i \to j$ based on the mutation probability obtained from the MG94 codon mutation model. N_{ij} is the total nonsynonymous mutations in codon c resulting from mutation $i \to j$, excluding stop codons.

For example, let's say the parent codon is AAG. Mutation AAG $\to$ TAG is not allowed since TAG is a stop codon. Thus, the expected nonsynonymous mutations at the first codon position is:

$$\begin{aligned} E_{\text{i}}[N] &= \sum_{j \neq \text{i}} w_{ij} N_{ij} \\ E_{\text{first A}}[N] &= \sum_{j \neq \text{A}} w_{Aj} N_{Aj} \\ &= w_{\text{AC}} N_{\text{CAG}} + w_{\text{AC}} N_{\text{GAG}} \\ &= w_{\text{AC}}(1) + w_{\text{AC}}(1) \end{aligned}$$

The total expected nonsynonymous mutations in codon AAG is:

$$\begin{aligned}
E_{\mathrm{AAG}}[N] =& \sum_{i \in c} \sum_{j \neq i} w_{ij} N_{ij} \\
=& [w_{\mathrm{AC}} N_{\mathrm{CAG}} + w_{\mathrm{AG}} N_{\mathrm{GAG}}] + \\
& [w_{\mathrm{AC}} N_{\mathrm{ACG}} + w_{\mathrm{AT}} N_{\mathrm{ATG}} + w_{\mathrm{AG}} N_{\mathrm{AGG}}] + \\
& [w_{\mathrm{GA}} N_{\mathrm{AAA}} + w_{\mathrm{GC}} N_{\mathrm{AAC}} + w_{\mathrm{GT}} N_{\mathrm{AAT}}] \\
=& [w_{\mathrm{AC}}(1) + w_{\mathrm{AG}}(1)] + \\
& [w_{\mathrm{AC}}(1) + w_{\mathrm{AT}}(1) + w_{\mathrm{AG}}(1)] + \\
& [w_{\mathrm{GA}}(0) + w_{\mathrm{GC}}(1) + w_{\mathrm{GT}}(1)]
\end{aligned}$$

SLAC Power

As a counting method, SLAC has low power at sites with few substitutions resulting from low diversity [81]. Although maximum likelihood methods exist that have been documented to have more power for populations with fewer sequences or little diversity, such as Fixed Effects Likelihood (FEL) where dN/dS is a parameter to be estimated via maximum likelihood at each site, those methods were benchmarked to be at least 6 times slower than SLAC for alignments with as little as 100 sequences, and at least 10 times slower for alignments with 1000 sequences. Considering viral populations are deep sequenced closer to 10000X and that selection at each site would be recalculated for every overlapping window, likelihood methods were abandoned for the initial release of the Umberjack pipeline. However, incorporating more powerful selection detection methods may be incorporated in future releases.

Internal Versus External Substitutions

I (Section 1.3.2) is calculated similarly to $dN - dS$ in that substitutions are counted along the phylogeny, but substitutions are counted separately along internal branches and external branches. Instead of normalizing by the expected nonsynonymous substitutions and synonymous substitutions, the substitution counts are normalized by the total internal and external branch lengths.

Average Over Windows

As each site can be covered by multiple overlapping windows, site $dN - dS$ is calculated as the average site $dN - dS$ across all windows, weighted by the site depth of unambiguous codons in each window.

Site I is calculated as the average site I across all windows, weighted by the site substitutions (along the phylogeny) in each window. Site I_N and site I_S are calculated analogously to I except that windows are weighted by the window-site's nonsynonymous substitutions and synonymous substitutions.

2.4 Method for Validating on Simulated Data

In order to test the accuracy of substitution counts across the phylogeny, Umberjack's estimates of $dN - dS$ were compared to known values from simulated data. 50 populations were simulated under parameter settings within ranges informed by empirical HIV measurements. Each population contained 1000 observed viruses with genomes of 9kbp. The observed viruses in the simulated populations were sequenced in silico using parameter ranges measured from MiSeq Nextera (shotgun) libraries produced by the BC Centre for Excellence in HIV/AIDS. Umberjack was run on each read library resulting in 82 502 Umberjack window-site $dN - dS$ estimates aggregated into 14 640 site $dN - dS$ estimates.

2.4.1 Selection of Simulation Parameters

In order to generate simulated data that covered the most representative combination of parameters, we sampled the parameters using a Latin hypercube sampling [101]. Latin hypercube sampling creates an N-dimensional grid, where each dimension represents a parameter. For each sample, a point is selected within the grid to maximize the distance between each point. Parameter values for each simulated dataset were taken from the coordinates of the selected points in the Latin hypercube. Parameter ranges can be found in Table A.1. The parameters for simulated datasets used to test Umberjack accuracy can be found in Table A.2.

2.4.2 Simulated Topology Under Selection

The topology of a population under selection was generated using an ancestral selection graph [45]. Under neutral evolution with no selection, the time from 2 individuals to their common ancestor follows an exponential distribution where the time between branching events on the tree increases closer to the root [75]. For lineages under selection, the ancestral selection graph corrects the distribution of time between branching events using the selection rate [45]. Selection rates were set to 0.01 per individual per generation in all simulated datasets. Average selection coefficients in envelope proteins of HIV populations in untreated HIV patients have been estimated to be 0.008-0.02 per individual per generation [77].

2.4.3 Simulated Recombination

Recombination was simulated by altering the tree topology at each recombination breakpoint within the genome by randomly pruning and regrafting a subtree to another location within the tree. In addition to the 50 simulated populations generated to predict Umberjack accuracy, we generated an extra 27 simulated populations to focus solely on the effects of recombination by fixing the parameters for high levels of diversity and high sequence quality. Although the effective HIV recombination rate is 10^{-5} recombinations/bp/generation [77], we wanted to determine the effects of light to heavy recombination using recombination rates in [10^{-4}, 0 recombinations/bp/generation], equivalent to a breakpoint occurring up to every 2 codons in the genome within the time course of the infection. Parameters for the recombination datasets are found in Table A.3.

2.4.4 Simulated Genomes

The branch lengths of the trees created by the ancestral selection graphs were in units of generations. Keeping the topologies, the entire tree lengths were scaled to units of substitutions/site according to the simulated population parameters: tree length = mutation rate x generations.

To simulate the different mutation rates of genes, the genome was randomly sectioned in up to 3 sections. The mutation rate in each gene was selected from the range [2×10^{-5}, 8×10^{-4}] mutations/bp/generation. Mutation rates for HIV

sequenced in plasma have been reported at 10^{-5} - 10^{-4} mutations/bp/generation [14, 65].

Using the trees as input, we simulated genomes for every observed and ancestral individual in a population using INDELible [30]. Codon site dN/dS values followed a discrete gamma distribution with 60 categories (shape=1.5, rate=3, scale=$\frac{1}{3}$). Site dN/dS values in inter-patient HIV datasets [71] have been reported to follow a smooth gamma distribution between [0, 3] with shape parameters in [0.2, 0.6] and rate parameters in [0.27, 1.1] .

2.4.5 Simulated Shotgun Reads

MiSeq shotgun paired-end 2x250bp reads were simulated for the observed individuals using a customized version of ART read simulation software that featured adapter contamination [42]. When read lengths are longer than the DNA fragment being sequenced, the sequencer will also sequence the primers attached to the DNA fragment, leading to adapter contamination in the read.

Empirical MiSeq Nextera (shotgun) sequence quality distributions measured from libraries generated by the BC Centre for Excellence in HIV/AIDS laboratory were used as quality guides. These quality distributions resulted in an average sequence error of 1% and $<$0.01% indel error across all simulated reads. Simulated fragment lengths followed a Gaussian distribution with average fragment length between 104 - 500 bp and standard deviation fixed at 100bp. The libraries generated at the BC Centre for Excellence in HIV/AIDS had mean fragment sizes in 104 -147 bp and standard deviations in 64 -100 bp.

The genome of each observed individual was sequenced between 0x-8x depth coverage. Shotgun libraries in our real datasets were estimated as having 2x -25x depth coverage per virus.

Each library was aligned against the population reference genome sequence using BWA-mem [52].

2.4.6 Umberjack settings

Depending on the Latin hypercube sampling, Umberjack was run on the simulated populations using one of the possible settings:

Setting	Config a	Config b
Window Size	150bp	300bp
Min Window Width Threshold for Read Inclusion in Window	0.7	0.875
Min Window Depth Threshold for Window Inclusion	10	10
Min Phred Score Threshold for N-Masking	15	20

2.4.7 Measuring Umberjack $dN - dS$ Accuracy

The accuracy in Umberjack's estimation of $dN - dS$ at the window-site level was quantified by the squared error:

$$\Delta = (\hat{x} - x)^2 \tag{2.3}$$

where $\hat{x}$ and x were the window-site predicted values of $dN - dS$ and known values of site $dN - dS$, respectively.

2.4.8 Measuring Tree Topology Accuracy

Every tree inferred from a window in the Umberjack pipeline was compared against the true tree(s) of the simulated population. If there were no recombination breakpoints within the window, there was only one true tree to compare against the inferred window tree. If there were recombination breakpoints within the window, the true tree for each recombinant section within the window was compared against the inferred window tree. Even without recombination, topology and branch length differences between the true and inferred tree could occur due to sampling, sequencing error, or heuristics in the phylogenetic reconstruction algorithm.

The similarity between an inferred window tree and a true tree was measured using their weighted Robinson Foulds distance WRF [92]. Lower WRF values indicate higher similarity, with $WRF = 0$ indicating equivalent trees. WRF measures the distance between tree A and tree B by summing up the lengths of branches in A not found in B, the lengths of branches in B not found in A, and the absolute difference in lengths of branches found in both trees. Branch t_A in tree A is equivalent to branch t_B in tree B if the 2 disjoint sets of tip nodes located on either side of branch t_A are the same as the 2 disjoint sets of tip nodes located on either side of branch

t_B.

The overall accuracy of an inferred window tree was calculated as $\overline{WRF}$, the average WRF between the inferred window tree and each true tree, weighted by the width of the window represented by each true tree.

To fairly compare a true tree against an inferred window tree, the true tree was downsampled to include only the individuals captured within the window reads. A copy of each individual was made in the true tree every time that individual was sequenced in the window. The branch lengths in the true tree were scaled according to the diversity found in the slice of true sequences corresponding to the true tree and window. Using FastTree2, the branch lengths of the true tree were chosen to maximize the likelihood of the slice of true sequences given the constrained true tree topology.

2.5 Determining Features that Affect Umberjack Accuracy

In order to determine which features (also known as covariates) affected Umberjack accuracy and rank their effects, we employed feature selection techniques: ranking univariate correlations and backwards feature selection in a regression random forest. Initially, backwards feature selection in general linear regression models was also employed but abandoned due to poor model fits. Further details are given in Section 2.6.1.

We simulated 50 populations with parameters described in Table A.2. Using the known $dN - dS$ values from the simulated data, we measured Umberjack accuracy using the Δ metric (Equation 2.3) and ranked features on their ability to predict Δ.

2.5.1 Features Considered For Predicting Umberjack Accuracy

(A) **Window Total Breakpoint Ratio**: We define the breakpoint ratio as the ratio of bases on either side of the breakpoint within a window, with the lower number as the numerator. The total breakpoint ratio is the sum of breakpoint ratios from all breakpoints within a window. This metric is a summary statistic indicating the number of breakpoints within a window and

how close they are to the middle of the window.

(B) **Window-Site Unambiguous Codon Rate**: Rate of unambiguous codons per read at a codon site in a window.

(C) **Window-Site Amino Acid Depth**: Depth of unambiguous amino acids at a codon site in a window.

(D) **Window-Site Substitutions**: Total nucleotide substitutions at a codon site in a window. Substitutions are counted between ancestral sequences along the phylogeny inferred by the Umberjack pipeline.

(E) **Window-Site Nonsynonymous Substitutions**: Total nonsynonymous substitutions at a codon site in a window. Substitutions are counted between ancestral sequences along the phylogeny inferred by the Umberjack pipeline. This metric is calculating using Equation 2.1 from SLAC Section 2.3.5.

(F) **Window-Site Synonymous Substitutions**: Total synonymous substitutions at a codon site in a window, counted similarly to 'Window-Site Nonsynonymous Substitutions'.

(G) **Normalized Window-Site Expected Nonsynonymous Substitutions**: Expected nonsynonymous substitutions at a codon site in a window, normalized by the total tree length, where the tree length is in units of nucleotide substitutions per codon site. Expected substitutions are counted between ancestral sequences along the phylogeny inferred by the Umberjack pipeline. This metric is equivalent to $\frac{E_T[N]}{\texttt{Tree Length}}$ (see Equation 2.2). The values range in [0, 3] and can be considered similar to the average number of nucleotide positions within a codon that could cause a mutation to a different amino acid.

(H) **Normalized Window-Site Expected Synonymous Substitutions**: This metric is similar to 'Window-Site Expected Nonsynonymous Substitutions Per Branch', except it counts synonymous substitutions instead of nonsynonymous substitutions.

(I) **Normalized Window Tree Length**: Length of the window tree inferred by the Umberjack pipeline, normalized by the total reads in the window. Units are in nucleotide substitutions/nucleotide site.

(J) **Normalized $\overline{WRF}$**: Average distance between window tree inferred by Umberjack pipeline and true trees corresponding to the window, normalized by total reads within the window. Units are in nucleotide substitutions/nucleotide site. See Section 2.4.8.

(K) **Normalized Polytomies**: Total polytomies in the window tree inferred by the Umberjack pipeline, normalized by total reads in the window.

(L) **Codon Distribution P-value**: Probability that the codon frequency distribution at a codon site within a window is the same as the codon frequency distribution at the same codon site within the genomes of all extent individuals in the simulated population. The probability is calculated using the G-test of independence between a set of distributions [67].

(M) **True Site Codon Entropy**: Shannon entropy of codons at a codon site of the genomes of all extent individuals in the simulated population.

(N) **Window-Site Codon Entropy**: Shannon entropy of unambiguous codons at a codon site in a window.

(O) **Window-Site Sequence Error Rate**: Rate of erroneously sequenced bases per read at a codon site in a window.

2.5.2 Univariate Correlation Ranking

The effect of each feature on Δ was ranked by their Spearman's correlation [100], a non-parametric measure of the monotonicity of the relationship between two variables. Since each feature was considered independently of other features, correlation only provided a rough picture of their effect on Δ.

2.5.3 Regression Random Forest

Since features had poor correlations and poor linear model fits to Δ, we used a regression random forest model to obtain more accurate predictions of Δ. A back-

wards feature selection algorithm produced a regression random forest using the best set of features to predict Δ. The features can be found in Table 2.2. A regression random forest is a collection of regression trees. A regression tree decides which linear regression model should be used to predict the response of a particular input datapoint. Each tip of a regression tree represents a separate linear regression model. When predicting the response for an input datapoint, the regression tree is traversed from root to tip. Each node is assigned a feature, and the connected edges represent conditions based on a feature threshold value. The path to the next node is determined by whether the input datapoint satisfies an edge's condition [8].

To train (i.e. generate) a regression tree, training data with known response values are recursively partitioned to form each node. The training data are partitioned based on a feature threshold value, and a linear regression model is fit to the data in each partition. The feature and its threshold value are selected to minimize the residuals of the linear regression models [8].

Random forests avoid overfitting by selecting a feature at each node from a random subset of features, and training each regression tree with a random subset of training data. Further, the random forest uses the average prediction amongst its collection of regression trees as the final prediction [8].

To gauge the importance of a feature, the values of that feature were permuted across observations in each training dataset. For each tree, the mean squared error between the true Δ and the predicted $\hat{\Delta}$, $MSE(\Delta, \hat{\Delta})$, was calculated for each regression tree using permuted and original training data as input for predictions. The feature importance score was calculated as the average increase in $MSE(\Delta, \hat{\Delta})$ due to permuting the feature, normalized by the standard deviation of $MSE(\Delta, \hat{\Delta})$. We used the Breiman (1999) test for feature importance significance, which treats feature importance scores as z-scores. The predictive performance of a random forest was measured as the total $MSE(\Delta, \hat{\Delta})$ from each regression tree.

For all random forest models, we used the R randomForest package [56] with default parameters, with the exception of ntree, which we chose as 501 to break ties:

- mtry = the total random features to select from at each node = floor(one-third of total features)

- ntree = total regression trees = 501
- permutations = 1

2.5.4 Backwards Feature Selection in Regression Random Forest

The backwards feature selection algorithm employed for the regression random forests used cross-validation to decide upon the ideal model size, M, then fit a final random forest model using the highest M ranking features.

The ideal model size was chosen as the model size that yielded the best predictive performance across 5 cross-validations. The training datasets were resampled between cross-validations. In each cross-validation, an initial random forest was trained using all available features. A new random forest was trained for each model size X, using only the X most important features. The feature importance ranks were calculated from the initial random forest and fixed for subsequent models. The predictive performance of the model size was calculated as the total $MSE(\Delta, \hat{\Delta})$ using the test datasets from the corresponding random forest in all cross-validations.

The final set of features were selected according to their revised reranking from all the cross-validations. A final random forest was trained using all the data.

Model selection was performed using R caret package [31] and R randomForest package [56].

2.6 Results

2.6.1 Weak Linear Relationship Between Features and Δ

The features considered for predicting the Umberjack accuracy (Section 2.5.1) demonstrated poor linear relationship with the Δ metric, causing issues when fitting linear models. The Spearman's correlation between each feature and Δ was fairly low, with absolute values ≤ 0.3 (Table 2.1).

Certain default R implementations for generalized linear models (speedglm package, Gamma distributed response GLM) were unable to fit models with colinear features, requiring manual exclusion of colinear features during backwards

feature selection. Some nested models were unable to be fit using default R implementations due to inability to find convergence of fitted values. These issues may have been avoided by using specialized linear regression models that penalize collinear or low importance features such as Lasso or Ridge regression [41, 103]. However, we decided to employ random forests to handle the possible non-linear relations between each feature with Δ.

Visual inspection of scatterplots of Δ versus each feature in A.1 to A.8 indicate fairly noisy regressions with a wide spread of error.

Feature	Corr	Lower	Upper
Diversity			
True Site Codon Entropy	0.16	0.15	0.16
Window-Site Codon Entropy	0.15	0.15	0.16
Normalized Window Tree Length	-0.063	-0.07	-0.056
Window-Site Nonsynonymous Substitutions	0.059	0.052	0.065
Normalized Window-Site Expected Synonymous Substitutions	-0.057	-0.064	-0.05
Window-Site Synonymous Substitutions	0.035	0.028	0.042
Normalized Polytomies	0.021	0.014	0.028
Normalized Window-Site Expected Nonsynonymous Substitutions	-0.013	-0.019	-0.0059
Sampling			
Codon Distribution P-value	-0.34	-0.34	-0.33
Window-Site Amino Acid Depth	-0.038	-0.045	-0.031
Window-Site Unambiguous Codon Rate	-0.0096	-0.016	-0.0028
Reconstruction & Sequence Error			
Window-Site Sequence Error Rate	0.16	0.16	0.17
Normalized $\overline{WRF}$	-0.03	-0.037	-0.023
Window Total Breakpoint Ratio	0.013	0.0067	0.02

Table 2.1: Spearman Correlation Between Feature and Δ; lower and upper 95% confidence interval.

2.6.2 Advantages of Regression Random Forests

We focused on random forests due to their built-in mitigation of feature collinearity. Since each node split is chosen from a random subset of features, random

forests are less likely to train models comprised mostly of collinear features. Regression random forests are also inherently able to model the non-linear relationships between the features through their piecewise linear modeling on partitions of training data. In addition, the built-in ensemble of regression trees reduces overfitting. Given that there were 82502 datapoints in the simulated datasets created to test Umberjack accuracy, the ease of parallelization of each regression tree in the random forest helped reduce the runtime for model fitting.

2.6.3 Features Affecting Umberjack Accuracy

Due to the lack of strong linear relations between features and Δ, the feature ranking results from the univariate correlations approach and linear regression approach were discarded from further analysis. The final regression random forest produced after backwards feature selection was able to explain 77% of the variation in Δ (Figure 2.4).

The model contained 11 features as shown in Table 2.2. Most notably, the ranking of the feature importance tells us that Umberjack is most sensitive to low genetic diversity since it employs a substitution counting technique. Although an approximate p-value was provided for every feature importance, they should only be considered a rough guide. The p-values are based on an approximation of the increase in model MSE due to feature permutation across regression trees with a normal distribution.

2.6.4 Effect of Recombination

A tree reconstructed from a multiple sequence alignment containing a recombination breakpoint will be a combination of the true trees on either side of the breakpoint. Figure 2.2 and Figure A.14 demonstrate how an increase of recombination led to an increase in error of inferred trees (as measured by $\overline{WRF}$) and consequently an increase in error of Umberjack $dN-dS$ estimates (as measured by Δ). However, at the recombination rates typical of HIV ($1.4 \pm 0.6 \times 10^{-5}$ recombinations/bp/generation) [77] or lower, Umberjack estimates with high error were dominated by low diversity (as measured by tree length).

Feature	Feature Importance	Approx p-value
Window-Site Substitutions	7.7	1.59e-14
Window-Site Synonymous Substitutions	6.3	4.08e-10
Normalized Window Tree Length	3.8	1.31e-04
Window-Site Sequence Error Rate	3.7	2.24e-04
Codon Distribution P-value	3.7	2.56e-04
Window-Site Codon Entropy	3.4	6.29e-04
Normalized $\overline{WRF}$	3.4	7.17e-04
Window-Site Unambiguous Codon Rate	3.3	1.02e-03
Window-Site Nonsynonymous Substitutions	3.3	1.08e-03
Window-Site Amino Acid Depth	2.8	4.90e-03
True Site Codon Entropy	2.8	4.96e-03

Table 2.2: Feature Ranking for Final Random Forest Predicting Umberjack Accuracy.

The above results were taken from the 27 simulated population dataset concentrating on recombination (Table A.3). However, similar results occured in the 50 simulated population dataset created with varied parameters chosen from Latin hypercube sampling (Table A.2). Although the rate of recombination simulated in the latter datasets was smaller, low site diversity (as measured by site substitutions) was the most importantly ranked feature affecting Umberjack accuracy (Table 2.2).

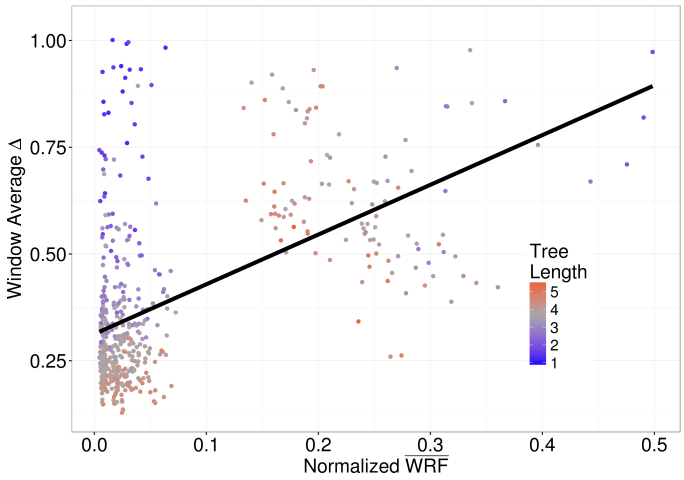
Since Umberjack breaks up the genome into shorter windows, Umberjack can reduce the effect of recombination on estimates of site evolutionary statistics if the average distance between recombination breakpoints is larger than the window size.

2.6.5 Umberjack Accuracy

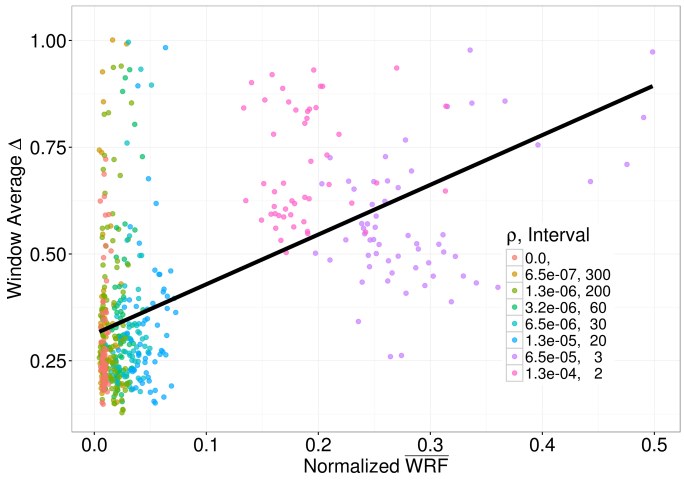
Many non-ideal scenarios were simulated to test the limits of Umberjack's accuracy. The Lin's concordance coefficient between Umberjack's estimate of $dN - dS$ at each window site versus the true site $dN - dS$ for all simulations was fairly low at 0.59 (95%CI = [0.58-0.60]). Lin's concordance coefficient ranges in [-1, +1], where 1 indicates perfect concordance, -1 indicates perfect discordance, and 0 in-

dicates no concordance.

Umberjack concordance was > 0.83 for ideal scenarios representative of good library preparation and sufficient population diversity (**Figure 2.3**). These ideal scenarios had an average 15% adapter contamination, 7 substitutions per site, 0.7x depth coverage per individual, $<1\%$ sequencing error, and branch length of 5.0×10^{-3} substitutions per site.



(a)



(b)

Figure 2.2: Effect of recombination on Δ. Each point represents a window of Umberjack estimates on simulated datasets focused on recombination (Table A.3). Black line indicates linear regression fit. Y-axis limited to Window Average $\Delta \leq 1.0$. **(a)** Points are shaded according the the length of the window tree in units of nucleotide substitutions/site. **(b)** Points are shaded according to the recombination rate ρ of population in units of recombinations/site/generation. Interval denotes average codon distance between breakpoints across the genome.

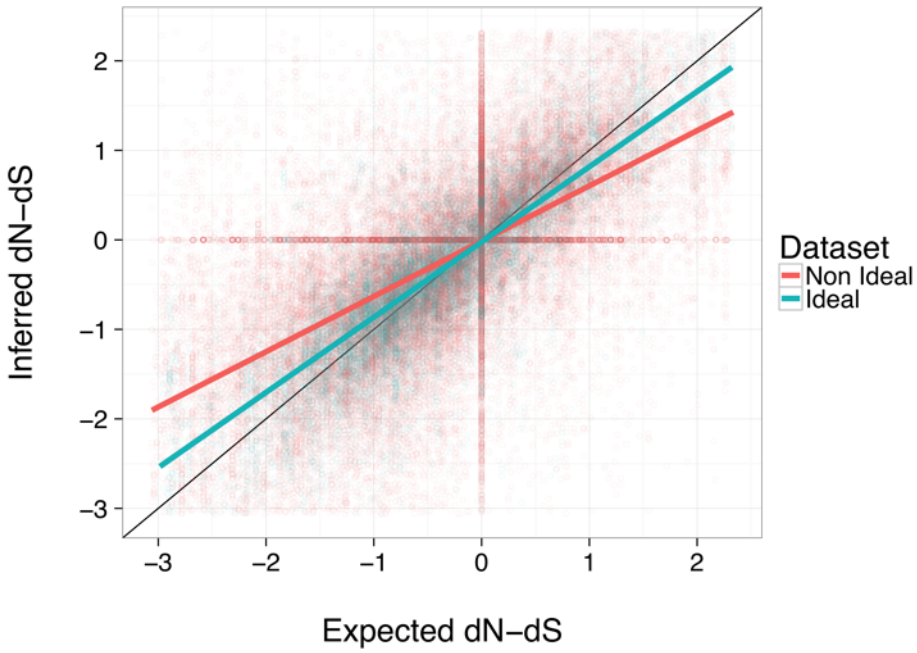


Figure 2.3: Umberjack inferred $dN - dS$ vs expected $dN - dS$ from simulated data. Each point represents a window-site extracted from a simulated dataset. Black line indicates y=x. Red line indicates correlation of Umberjack inferred $dN - dS$ to expected $dN - dS$ for datasets representing non-ideal scenarios such as poor library preparation, insufficient sampling, or low population diversity. Blue line indicates correlation of Umberjack inferred $dN - dS$ to expected $dN - dS$ for datasets representing ideal scenarios.

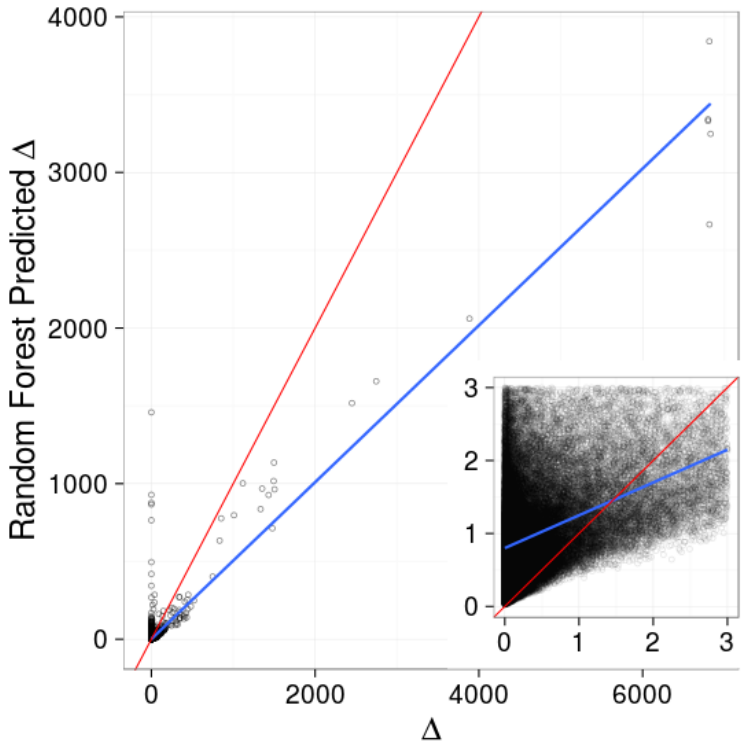


Figure 2.4: True Umberjack inaccuracy versus random forest prediction of Umberjack inaccuracy. Umberjack inaccuracy was measured as the squared error (Δ, Equation 2.3) between Umberjack's estimate of $dN - dS$ and expected $dN - dS$. Each point represents a single site from a simulated dataset. Red line indicates y=x. Blue line indicates fitted line of Δ vs random forest predicted Δ. Inset: expansion of plot in the region of $\Delta = (0,3)$ and random forest predicted $\Delta = (0,3)$.

Chapter 3

Within-host HIV Evolution

3.1 Background

The host-specific adaptive immune response is a primary source of selection that shapes the genetic variation of HIV [86]. For example, a vigorous cytotoxic T lymphocyte (CTL) response (Section 1.1.2) within the first few weeks of infection is associated with a substantial decline in viral load [6]. During the acute stage of HIV infection, which lasts approximately for the first 6 months of infection, viral loads rapidly increase to a peak of around 10^7 copies/mL [36]. As a host CTL cell population targets certain HIV epitopes, new HIV escape mutations emerge and sweep. The CTL population size decreases when the number of infected cells it binds decreases. Another CTL population targeting different sets of epitopes takes its place, resulting in another set of HIV escape mutations that sweep [33]. Transition to a chronic stage of infection is associated with a diminishing rate of CTL escape mutations, whereupon the viral loads decline to a 'set point' level averaging around 3×10^5 copies/mL, with substantial variation among individuals [33, 36]. To evaluate the role of the CTL response in shaping the evolution of HIV within hosts, we processed and analyzed shotgun sequencing of viral RNA isolated from plasma of 31 treatment-naive HIV patients with high viral loads.

3.2 Methods

3.2.1 Dataset

The study used frozen plasma samples from BC HOMER cohort individuals who had provided informed consent to donate and store the remainder of their routinely collected plasma samples for the purposes of research relevant to the pathogenesis and treatment of HIV under the Experimental HIV Monitoring Program. Samples were chosen from individuals on the basis of the following criteria: first, the individuals had at least two samples available before the first therapy start date; second, the samples were collected at least 90 days apart (median 377 days) to increase the chances of measurable evolution occurring between observations; and third, the samples were associated with a minimum viral load of 5×10^4 copies/mL. The last criterion was employed to reduce the probability of template resampling, in which a single template is represented by multiple sequences. Not only does template resampling skew minor variant frequency counts, it also introduces false diversity during phylogenetic reconstruction if there is sequencing error amongst the resampled templates [60]. We note that templating resampling is difficult to discern bioinformatically from true biological proliferation of a variant. There are wet lab methods such as primer IDs that can tag templates to determine how many unique templates there are in a population; however, these methods were not employed for this dataset [108]. Although patient CTL responses were not explicitly measured through 'wet lab' assays, the high viral loads implied adaptation to the host-specific immune responses by the respective HIV populations. Most patients either demonstrated an increase in viral load or a change within assay measurement error. The exceptions were patients BC11 and BC22, for whom a log10-fold decrease in viral load of $<$-0.67 and $<$-0.68 was observed, respectively. Further details on patient measurements can be found in Table A.4.

In addition, the genotypes for HLA genes HLA-A, HLA-B, and HLA-C for these individuals had been previously determined using the procedure documented in [11]. Briefly, the HLA genes were sequenced with Sanger sequencing, and the HLA genotypes classified at the resolution of allele group (coarser) or protein (finer). Allele groups were further imputed down to the protein level using the

software HLA Completion [58] for individuals for whom allele groups from all genes were available. Five patients were missing sequences for HLA-A and HLA-C and thus only their HLA-B allele groups were known. HLA Completion was unable to compute suitable protein resolutions for one patient due to maximum likelihood convergence issues; consequently, the HLA genotypes for this patient could only be resolved at the allele group level.

HIV RNA was extracted from the plasma samples using a NucliSENS easyMAG instrument (bioMérieux, St. Laurent, Quebec, Canada). Nested RT-PCR amplification reactions in which a shorter region was isolated from a wider region in two rounds of PCR were performed as described in previous work [1] using primers specific to regions of the HIV-1 genome encoding *gag* p17, *nef*, and the third variable loop of *env* gp120. The amplified samples were used to generate Illumina Nextera XT libraries and sequenced on an Illumina MiSeq instrument using a paired-end 2 x 250bp reagent kit (version 2), resulting in a random 'shotgun' distribution of short reads across each target region. Median depth of coverage reached 8200-, 10000-, and 65000-fold for genes *gag*, *nef*, and *env*, respectively. Further details on the sequencing libraries can be found in Table A.5.

A phylogenetic tree was constructed using the sample consensus sequences for each patient at each time point in the original dataset. All patients were infected with HIV1B subtype of HIV. Samples that did not phylogenetically cluster together by patient indicated potential cross-contamination and were discarded, forming the final 31 patient dataset (Figure A.9, Figure A.10, Figure A.11).

3.2.2 Umberjack Processing

Reads from each patient sample were iteratively aligned using BWA-mem [52] against the HXB2 reference (GenBank accession K03455.1), where the previous iteration's consensus was used as the reference in the next iteration. Re-alignment was halted after 10 iterations or if the consensus sequence did not change from the previous iteration. When forming the consensus sequence, gaps that induced frameshifts (gaps not in groups of three) were shifted together into groups of three as long as they did not introduce premature stop codons. The SAM files from every patient sample alignment were used as inputs for Umberjack to estimate $dN - dS$

and I per codon site, Section 1.3.2, (window size = 300bp, window overlap = 30bp, read bases with quality score $<$20 masked with N, reads with $>$12.5% missing bases within a window excluded, windows $<$50 sequences excluded).

We note that PCR amplification of populations has been known to introduce artificial recombination that is indistinguishable from true biological recombination [99]. Due to the short average fragment size of the dataset ($<$200bp), we did not employ bioinformatic tools to detect the rate of recombination, as there would be insufficient sequence length to obtain confident estimates. We relied on excluding sites in which random forest predictors deemed that Umberjack produced inaccurate estimates.

3.2.3 Umberjack Cleaning

Random forest predictors were tested for their ability to predict Umberjack error solely using features that could be extracted from real datasets. Features which would most likely be unknown in an experimental context, such as 'Window-Site Sequence Error Rate' and 'Normalized $\overline{\text{WRF}}$', were excluded from backwards feature selection. The final random forest trained on the best performing real-dataset features ('Normalized Window-Site Expected Synonymous Substitutions', 'Window-Site Synonymous Substitutions', 'Normalized Window-Site Expected Nonsynonymous Substitutions', 'Window Unambiguous Codon Rate', 'Normalized Window Tree Length') explained 19% of the variance in error within Umberjack estimates of window-site $dN-dS$.

Although the random forest did not predict error well for $dN-dS$ calculations at the window-site level, it was still useful in removing unreliable window-site $dN-dS$ estimates to improve the accuracy of site $dN-dS$ estimates averaged across windows. Using an empirically derived threshold, Umberjack estimates of window-site $dN-dS$ with an excessive predicted error by the random forest model were excluded from further processing. Umberjack site $dN-dS$ was calculated by averaging estimates across the cleaned windows. Removing dubious window-site estimations reduced the total number of site $dN-dS$ estimates from the simulated dataset by 1%. Based on a simulation analysis in the previous chapter, the R^2 for cleaned estimates of site $dN-dS$ compared to true site $dN-dS$ was 80%

Dataset	Patients	Samples Per Patient	Time Span (post infection)	Total Seq
Shankarappa *et al.* [97]	11	10-14	6-12 years	1300 *env*
Liao *et al.* [55]	1	10	1-4 months	295 *env*
Liu *et al.* [61]	2	2, 23	8 days-4 years	>155 *gag*, >194 *nef*, >165 *env*

Table 3.1: Composition of published HIV-1 sequence data sets used to derive an empirical distribution of site-specific $dN - dS$ estimates.

(Figure A.12).

The error threshold ($\Delta < 10.7$) was based on the IQR (Interquartile Range) of empirical $dN - dS$ values taken from other empirical datasets of longitudinal samples of untreated patients where several HIV viruses were isolated per timepoint and sequenced using conventional capillary-based methods (Table 3.1).

Further, the indel-rich V4 region of the *env* gene (amino acid coordinates 395-413 with respect to ENV protein in HXB2 strain) was excluded from the analysis to avoid erroneous nonsynonymous substitutions that could arise from the spurious alignment of non-homologous insertions.

3.2.4 Finding CTL Response Associated Amino Acid Polymorphisms

HIV amino acid polymorphisms associated with CTL response by Carlson *et al.* in a prior study [11] were located within the multiple sequence alignments and reconstructed ancestral sequences were generated by Umberjack from the patient sample libraries. The CTL-associated HIV polymorphisms were classified as 'Matched Escape' if the polymorphism associated with evading a host-specific CTL response and a selective HLA allele was present in that patient, 'Matched Nonescape' if it was associated with susceptibility to the CTL response and a selective HLA allele was present, or 'Unmatched' if no selective HLA alleles were present.

3.2.5 $dN - dS$ Analysis

Site $dN - dS$ was estimated for each sample separately and across samples (pooled data) for each patient. Sites under significant selection according to $dN - dS$ were determined by performing binomial tests on substitutions counts using 'nonsynonymous' and 'synonymous' as the response categories and the expected synonymous substitutions as the probability of success [80].

3.2.6 I Analysis

Site I and $I_N - I_S$ were calculated for each sample separately and across both timepoint samples for each patient. Sites under significant selection according to I were determined by performing permutation tests in which substitutions were randomly reassigned as either 'internal' or 'external' using the total internal and external branch lengths as weights. The p-value was calculated as the fraction of permutation trials in which the $\hat{I}$ calculated from permuted substitutions was further from 0.5 (the neutral threshold value) than the original I calculated from non-permuted data. Site substitutions were permuted for 10^4 independent trials in each window covering the site. The total site substitutions in each window were kept constant during each trial.

In order to perform multivariate regression on $I_N - I_S$ statistics, $I_N - I_S$ values were transformed from $[-1,+1]$ to $[0,1]$ so that it could be modeled as a zero-one inflated beta distribution. Further, all p-values from each separate zero-one inflated beta regression were Benjamini-Hochberg corrected.

3.2.7 Timing Infection

Since clinical information for the dates of HIV-1 infection was unavailable, we estimated these dates using the software package BEAST (Bayesian Evolutionary Analysis by Sampling Trees) [23]. BEAST employs a Bayesian Markov chain Monte Carlo (MCMC) method to generate a random sample from the joint posterior distribution of rooted trees and rates of evolution in units of time, given sequences labeled with dates of sample collection. Using the most recent common ancestor as a proxy for the infecting virus, the date of infection was set to the date at the root of the tree. Since the Bayesian MCMC procedure was computationally

intensive, the date of infection was only inferred for one window of alignment per patient. Both baseline and followup sequences were pooled into the same alignment and windows were extracted by Umberjack (window size = 300bp, window overlap = 30bp, read bases with Phred quality score < 20 masked with N, reads with $> 12.5\%$ missing bases within a window excluded, windows < 50 sequences excluded). Windows to analyze with BEAST were selected on a score based on a combination of the highest average site entropy across the window, highest average site depth coverage across the window, and lowest percentage of missing bases (N's, gaps). 85% of the selected windows encompassed the V3 variable region in *env*. After running BEAST for two chains of 3×10^8 MCMC samples for each patient, mutual convergence of both chains on the estimated date of infection was evaluated through visual inspection of trace plots displaying estimates. The first 10^7 MCMC samples from each chain were discarded as a 'burn-in' period.

3.3 Results

3.3.1 Genetic Variation Across Sites

As expected, sequences covering HIV *env* gp120, which encodes the surface envelope glycoprotein, tended to be more variable. Using amino acid entropy to quantify variation, there was a significant difference in variability between genes, with entropy in both *nef* and *env* greater than *gag* ($t = 11.9$ and $t = 40.6$, respectively; likelihood ratio test, $P < 10^{-15}$). Entropy in the follow-up samples also varied over the genome in a consistent pattern amongst subjects. For example, the amino acid entropy along the portion of HIV *env* covered by the NGS data tended to be greatest near the disulfide loop regions (V3, V4 and V5), and lowest in regions associated with conserved functional motifs such as the GPGR motif within V3 and the CD4 binding loop (Figure 3.1). Note that we excluded the V4 region from any further analysis regarding selection due to issues with multiple sequence alignment of excessive indels.

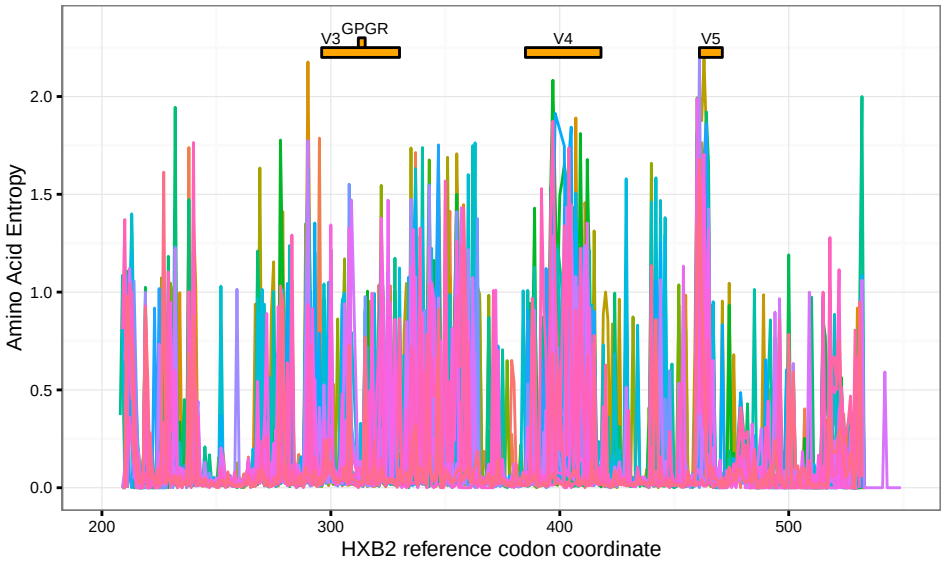


Figure 3.1: Summary of amino acid entropy along HIV *env*. Each trace corresponds to a different individual's follow-up sample.

3.3.2 Most Subjects in Chronic Stage of Infection

Based on BEAST timing estimates of the most recent common ancestor, 23/31 subjects were in the chronic stage of HIV infection by the baseline sampling date, with a median infection date of 2.6 years prior to baseline (Figure 3.2, Table A.4). The timing of infection for 7/31 subjects was $<$9 months, which we classified as close to the boundary of acute/chronic stages of infection. The viral loads and CD4 counts from these subjects were also ambiguously acute/chronic. Their minimum viral load was 5×10^6. The limit of quantification on the viral load assay was 10^6 copies/mL, which several of the patients hit; however, this study population had been selected on the basis of having samples with high viral loads. This can bias results in that patients with high viral loads indicate proliferation of viral populations and adaptation of the viral population to the host immune system. CD4 counts associated with these samples were within 330-541 cells/mm^3. The BEAST runs for one of the patients did not converge on a timing estimate, and their infection timing information was excluded from further analysis. The median estimated rate of evolution across subjects was 8.3×10^{-5} mutations per base per day, which was

consistent with published estimates of the HIV mutation rate (3.5×10^{-5} [64]) and the hypervariability of the V3 region.

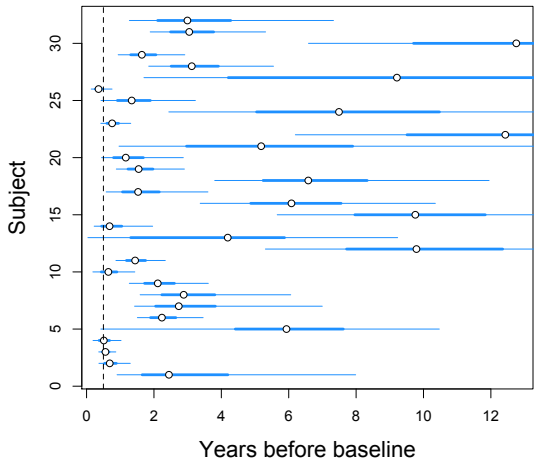


Figure 3.2: Estimated durations of infection at baseline by molecular clock analysis. Each point represents the median estimate for a given subject. The thick line segment indicates the interquartile range, and the thin line indicates the 95% credible interval. A dashed line is drawn at six months to indicate which subjects may have been at an acute or early stage of infection at baseline.

3.3.3 Evolving Sites Associated with CTL Response

To study evolution at the protein level in these data, we extracted the amino acid frequency distributions per site from windows of the NGS alignments for each gene and subject. We used a G-test statistic to evaluate changes in these frequency distributions per site between the baseline and follow-up samples. To adjust for multiple comparisons, we used a simple Bonferroni correction such that $\alpha = 1.9 \times 10^{-6}$ to classify sites with a significant change in the amino acid frequency distribution over time, which will be referred to as 'evolving sites' for brevity. Despite this conservative procedure, 5354 (20%) out of 26186 tests were considered significant. Evolving sites were more likely than not to contain an HIV polymorphism matched with at least one of the subject's HLA alleles, insinuating that CTL medi-

ated response drove selection at those sites in those individuals (Odds Ratio=1.3, $P = 9.7 \times 10^{-7}$).

Matched escape mutations were significantly more likely to increase in frequency than matched non-escape mutations, and this trend was the most pronounced in HIV *nef* ($P = 0.015$). Additionally, the increase of matched escape mutations over time was significantly greater for mutations restricted by HLA-B alleles than alleles at the other HLA loci ($P = 0.023$; Figure 3.3). This is consistent with previous studies that determined that HLA-B alleles were the most effective at viral load suppression [36].

To investigate these associations in greater detail, we fit a generalized linear mixed model on the classification of evolving sites using a binomial link function, with subject as a random effect on model intercepts, rejecting simpler models on the basis of the Akaike information criterion. In addition to terms corresponding to known HLA associations and matched status within subjects, we observed significant fixed effects of genes relative to HIV *gag* (*nef*: $z = 2.2$, $P = 0.02$; *env*: $z = 26.4$, $P < 10^{-12}$); in other words, evolving sites were observed significantly more often in *nef* and *env*. Based on the model and residual deviances, however, the mixed model explained only about 8% of variation in the binomial outcome of being classified as an evolving site. Considering *env* is also known as a highly variable gene in general, and adaptive to the human immune response, it is possible that the increased significance of *env* compared to *nef* is due to the larger number of substitutions found in *env*. Further, it is possible that the majority of the sites found under significant selection were hitchhiking variants caught in a selective sweep; however, it is difficult to disentangle true selection from a mere sweep.

3.3.4 Host HLA Genotype Drives Diversifying Selection

The next objective was to investigate whether the application of phylogenetic methods could improve or supplement the study of HIV evolution within hosts beyond what was attained using frequency distributions alone (section 3.3.3).

Site specific patterns of diversifying selection were analyzed using the $dN - dS$ statistic (Figure 3.4) across samples for each patient using Umberjack. On average, most sites tended to experience purifying selection (median $dN - dS = -0.13$).

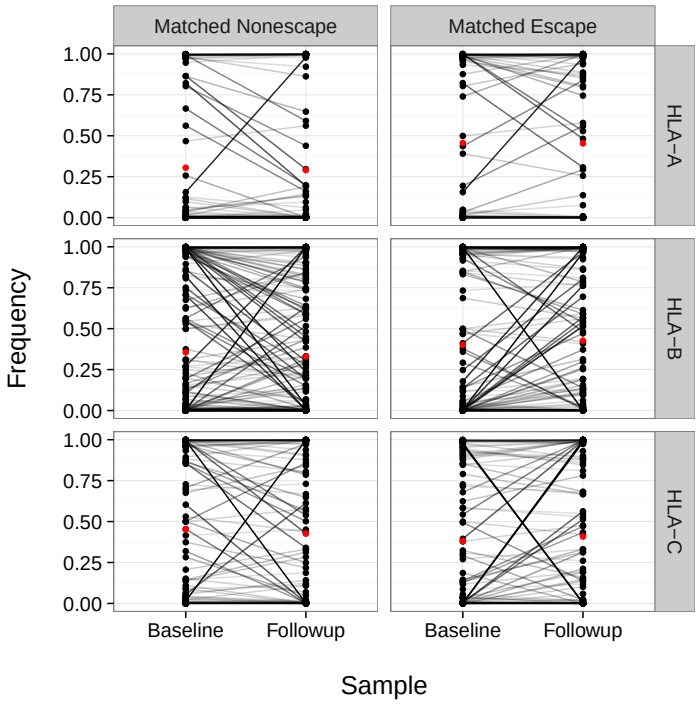


Figure 3.3: Shifts in matched non-escape and escape mutation frequencies over time in HIV *nef* across all subjects. Black lines indicate the change in the frequency of the amino acid polymorphisms from baseline to follow-up. There was a small but significant tendency for non-escape mutations to decline, and for escape mutations to increase, which was the most apparent for mutations restricted by HLA-B alleles. Red dots indicate the overall mean frequency in each group.

Roughly 36% of all sites across genes and subjects had values of $dN - dS$ exceeding zero (diversifying selection), of which only about 0.97% were considered significant based on a q-value cutoff of 0.05 (Benjamini-Hochberg correction). $dN - dS$ varied amongst genes, with significantly greater values observed on average for HIV *nef* than either *env* or *gag* ($t = 6.0$, $P = 1.9 \times 10^{-9}$). There was also a significant association between $dN - dS$ and whether that site had a CTL-associated polymorphism matched to the subject's HLA genotype ($t = 5.8$, $P = 7.0 \times 10^{-9}$), which is consistent with the CTL mediated response driving the diversification of HIV within hosts. Finally, $dN - dS$ was marginally associated

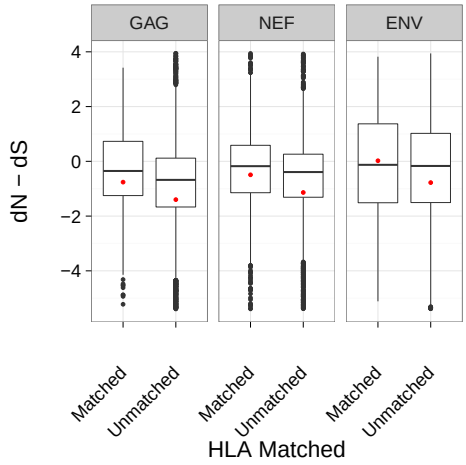


Figure 3.4: Boxplots of $dN - dS$ at HLA Matched Sites and Unmatched Sites by Gene. Red dots indicate means.

with an indicator variable of whether the subject was classified with an acute or early stage of HIV infection (defined here as being within the first year of infection) on the basis of the molecular clock analysis ($t = 2.2$, $P = 0.026$; section 3.3.2). The association between $dN - dS$ and HLA matched sites still existed after stratifying by gene (Mann-Whitney U, *gag*, $P = 1.1 \times 10^{-8}$; *nef*, $P = 3.7 \times 10^{-11}$; *env*, $P = 4.3 \times 10^{-3}$; Figure 3.4).

This last result is consistent with the bulk of CTL mediated selection occurring in the earlier period of an HIV infection. At the same time, it may be an artifact of the transient nature of diversifying selection in the relatively constant environment presented within a single host. Once the host-specific immune response has been stimulated by the infection, the virus population undergoes a number of 'selective sweeps' (Figure 3.3). During the sweep, an elevated amino acid substitution rate is manifested by an elevated $dN - dS$ rate indicating selection. After the conclusion of the selective sweep, however, we do not expect to observe any further evolution at the amino acid level at the sites targeted by the immune response and $dN - dS$ will revert to lower levels consistent with purifying selection. If a sample is taken at the point of a sweep when synonymous substitutions begin to overtake nonsynonymous substitutions to maintain favoured amino acids, $dN - dS$ may even

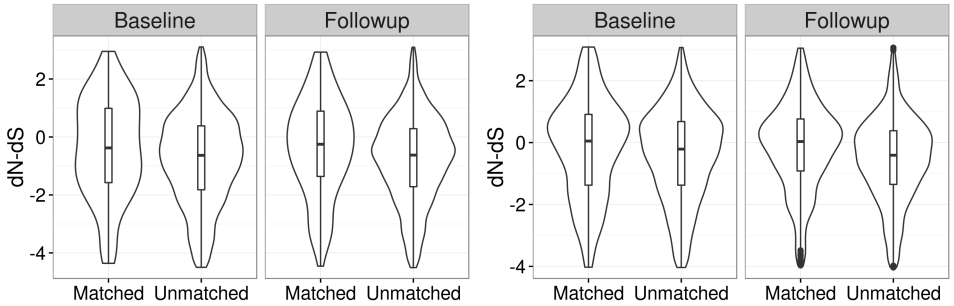
indicate neutral evolution. Consequently, $dN - dS$ alone is not a sufficient metric for quantifying the response to selection by HIV within hosts [46].

The temporal nature of this statistic is exemplified when we examined the per-sample $dN - dS$ values in Figure 3.5 as opposed to the across-samples $dN - dS$ values in Figure 3.4. We compared sites with matched polymorphisms to sites with no matched polymporphisms (Figure 3.5). Here, $dN - dS$ hovers around neutral or purifying at 'Matched' sites and $dN - dS$ is purifying at 'Matched' for *gag* and *nef*, and a slight increase of Matched $dN - dS$ upon followup in *env*. However, the differences between Matched and Unmatched groups is insignificant (P>0.55, linear regression), though we expect that CTL response will imbue selective pressure on Matched sites.

3.3.5 Evidence of Selective Sweeps

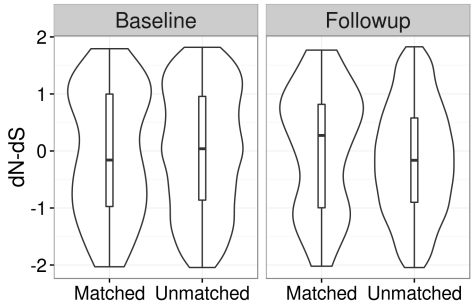
The I statistic suffered from similar issues as $dN - dS$ with respect to aggregating over the substitutions. Similar to the original Fu & Li statistic [32], I does not differentiate between nonsynonymous and synonymous mutations. Consequently, directional selection at the amino acid level may appear neutral. After Benjamini-Hochberg multiple test correction, only 3% of sites were under significant diversifying selection and only 19% of sites were under significant purifying selection. In this section, we demonstrate that $I_N - I_S$ is a better measure of relative directional selection at the amino acid level. Estimates of $I_N - I_S$ (section 1.3.2) were loosely correlated with measures of diversifying selection $dN - dS$ (Kendall's $\rho = 0.42$, $P < 2.2 \times 10^{-16}$).

Across all patients and sites, nonsynonymous substitutions were significantly more likely than synonymous substitutions to be mapped to tips of the phylogeny (Figure 3.6; $I_N < I_S$, Mann-Whitney $U = 2.37 \times 10^8$, $P < 2.2 \times 10^{-16}$). This result is consistent with purifying selection being the dominant mode of selection within hosts. In other words, most nonsynonymous mutations are deleterious. This trend varied significantly among genes. For instance, from Figure 3.7 and Figure A.13, median $I_N - I_S \approx 0$ in *env*, indicating that nonsynonymous and synonymous mutations were similarly distributed between internal and external branches. However, in *gag*, median $I_N - I_S = -0.11$ indicating that *gag* nonsynonymous mutations



(a) *gag*

(b) *nef*



(c) *env*

Figure 3.5: Site $dN - dS$ per sample at sites associated with CTL response for patient's HLA. Site $dN - dS$ calculated separately for baseline and followup samples. Violin plots show density of sample-sites at each value of $dN - dS$. Inner boxplots show IQR and median. 'Matched' groups refer to any site associated with CTL response in study cohort ([11]) sharing patient's HLA. 'Unmatched' groups refer to any site with no CTL association for study cohorts sharing patient's HLA.

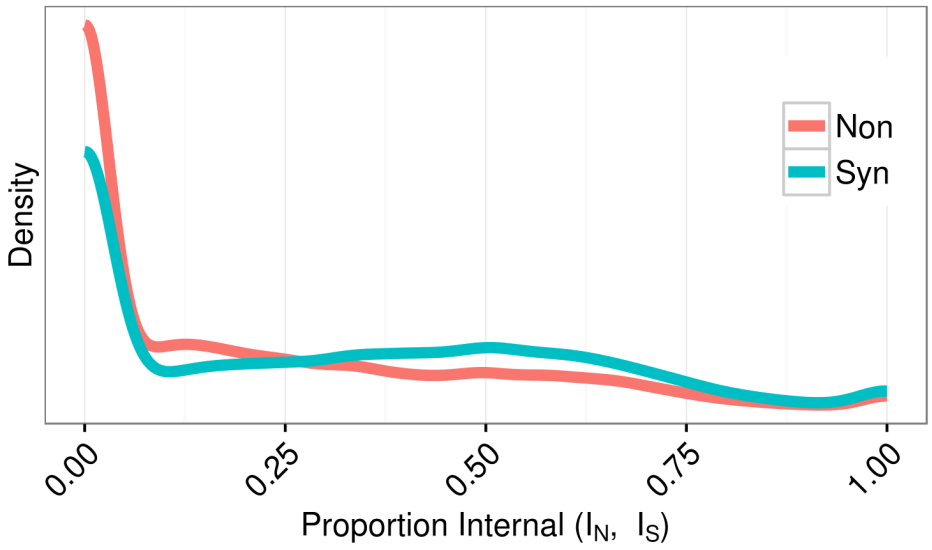


Figure 3.6: Substitutions tend to map to tips of the within-host phylogeny. Density plots comparing how often site nonsynonymous and synonymous substitutions map to internal branches of phylogeny. Red lines indicate site I_N, blue lines indicate site I_S. I_N and I_S range in [0, 1], where 0 means all substitutions were found in the tips, and 1 means all substitutions were found in internal branches.

tended to occur more recently than synonymous substitutions, implying that *gag* was under stronger purifying selection than *env* ($t = 19.6$, $P = 1.5 \times 10^{-83}$, linear regression). *gag* epitopes have been previously reported to have a stronger immunogenic in CTL response than *env* [37]. Alternatively, the higher indel rate in *env* could have induced more alignment errors, driving up artificial mutations seen at the tips of the *env* phylogeny.

To determine the effect of HLA pressure on HIV evolution, we stratified sites by whether they contained a Matched polymorphism at anytime during the population history, as observed at sequencing or reconstructed by the Umberjack pipeline. An entire amino acid site was labeled 'Matched' as long as it contained at least one Matched polymorphism, and 'Unmatched' if it never contained a Matched polymorphism. $I_N - I_S$ at Matched sites increased significantly with median ≈ 0 in all

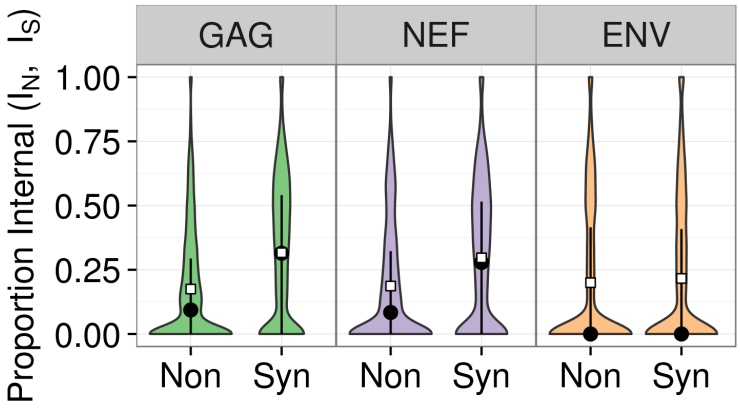


Figure 3.7: Nonsynonymous substitutions map to tips of the phylogeny more often than synonymous substitutions. The violin plots summarize the overall distributions of I_N and I_S statistics across subjects, broken down by gene. A lower I statistic indicates that the substitutions tend to map onto the tips of the phylogeny, which is consistent with purifying selection. Black point ranges indicate median and IQR. White squares indicate mean.

genes, and mean equal to -0.03, -0.02, and 0.08 for *gag*, *nef*, and *env* respectively ($t = 5.4$, $P = 6.0 \times 10^{-8}$). These greater values of $I_N - I_S$ indicate that the nonsynonymous substitutions at these HLA-restricted sites mapped earlier in the phylogenies, which is consistent with recent selective sweeps (Figure 3.8). The difference between site I_N and I_S was significant for *env*, the most variable gene studied in this thesis (*env*: $t = 2.9$, $P = 0.054$). Further, I_N for Matched sites were consistently higher than I_N for Unmatched sites across each gene (*gag*: $t = 9.7$, $P = 2.6 \times 10^{-21}$, *nef*: $t = 6.4$, $P = 6.2 \times 10^{-10}$, *env*: $t = 2.6$, $P = 0.013$), leading to more evidence that CTL pressure was driving directional selection and causing favoured amino acids to overtake the population earlier on.

Since overall site $I_N - I_S$ does not differentiate between amino acids, it is best used as a relative measure of directional selection, as $I_N - I_S = 0$ does not necessarily mean that a site is neutrally evolving. When sites experience early mutations to favoured amino acids at sites that may continue to diversify through synonymous

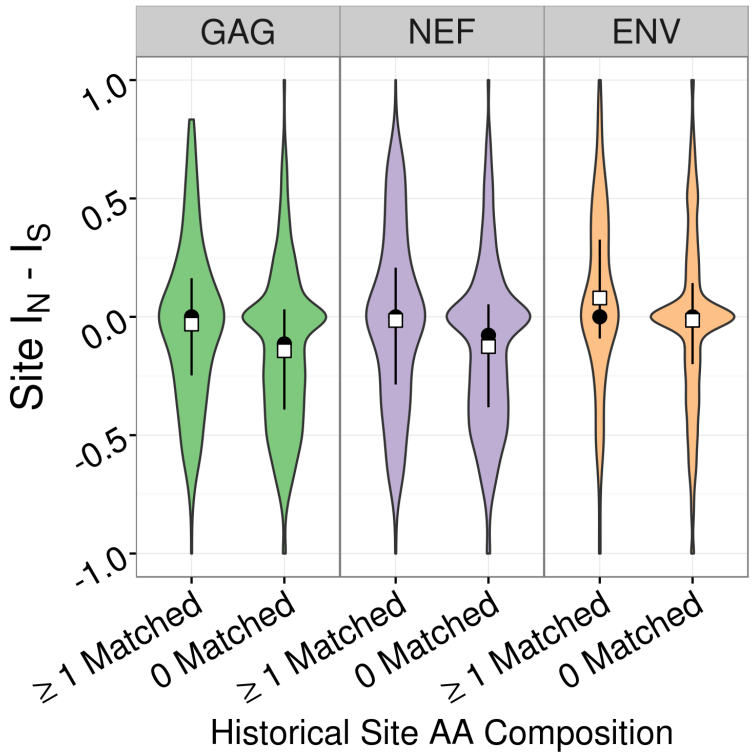
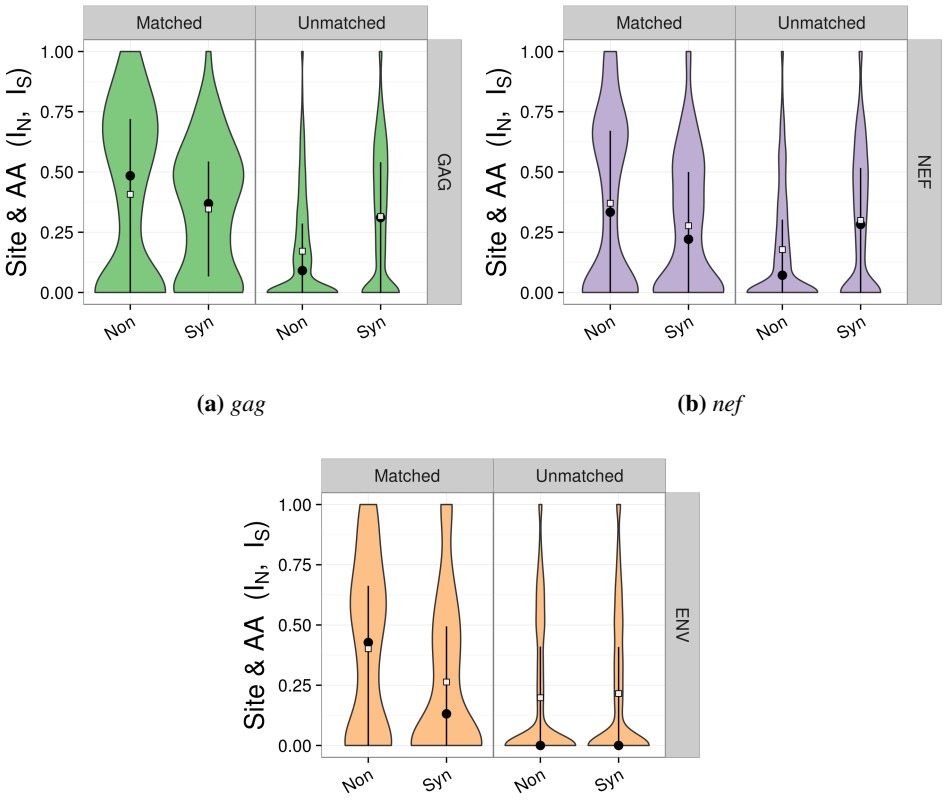


Figure 3.8: Distributions of nonsynonymous and synonymous substitutions in the tree are more similar at sites with known CTL associations. The violin plots summarize the distributions of site $I_N - I_S$ per patient for sites with known CTL associations. Sites are stratified by whether the site contained a Matched amino acid at any point in the reconstructed within-host phylogeny. Each constituent site $I_N - I_S$ represents the difference between proportion of site nonsynonymous and synonymous substitutions mapped to internal branches of the phylogeny containing both baseline and followup sequences from a single patient. As long as the site in the patient samples contained at least 1 Matched amino acid in any of the observed sequences or inferred ancestral sequences, the site was labeled "Matched". A lower I statistic indicates that the substitutions tend to map onto the tips of the phylogeny, which is consistent with purifying selection. The filled circle and whiskers indicate the medians and IQR respectively. White squares indicate means.

substitutions, recent mutations away from the favoured amino acids at the tips will decrease site I_N, driving $I_N - I_S$ towards parity. If the site is truly under directional selection, the unfavoured variants will be purged from the population. Therefore, in order to truly see directional selection by comparing internal vs external substitutions, we need to explicitly label the amino acid polymorphisms at each site. Instead of calculating overall site $I_N - I_S$ and then stratifying sites as Matched/Unmatched, we stratified substitutions at each site as Matched/Unmatched according to the specific amino acid resulting from the substitution. From Figure 3.9, it is evident that nonsynonymous substitutions to Matched amino acids tended to occur earlier than synonymous substitutions, implying that lineages carrying the Matched amino acids proliferated in the ancestral population under positive selection. Conversely, nonsynonymous substitutions to Unmatched amino acids occurred more often at the tips than the synonymous substitutions, implying that existing amino acids conferring a fitness advantage were maintained and new deleterious amino acids were purged before they could proliferate (*gag*: $t = 9.6$, $P = 7.6 \times 10^{-2}$, *nef*: $t = 10.7$, $P = 1.4 \times 10^{-26}$, *env*: $t = 3.6$, $P = 2.6 \times 10^{-3}$).

3.3.6 Reversion to Wild Type

There were no significant differences when the distributions of I_N and I_S shown in Figure 3.9 were further stratified by substitutions to new Matched Escape versus Matched NonEscape amino acids ($P > 0.16$). The Carlson *et al*. study, from which we extracted the HLA allele - HIV polymorphism associations for this analysis, reported that Matched NonEscape amino acids were typically (>80%) the same as the wild-type amino acid, which was defined as the HIV-1 subtype B consensus sequence as described by the Los Alamos HIV Sequence Database [47]. Reversion to wild type epitopes has been noted in patients infected with strains containing mismatched escape mutations [51, 107]. These reversions took years to emerge, which would be consistent with our dataset of chronic patients. Breaking down I_N and I_S by substitutions towards a wild-type or mutant amino acid at each site (Figure 3.10) revealed that reversions to wild-type occurred significantly earlier in the phylogenies (*gag*: $t = 14.6$, $P = 1.5 \times 10^{-83}$, *nef*: $t = 20.4$, $P = 1.2 \times 10^{-90}$, *env*: $t = 15.1$, $P = 2.6 \times 10^{-51}$), implying a selective advantage relative to the mutant residues.



(a) *gag*

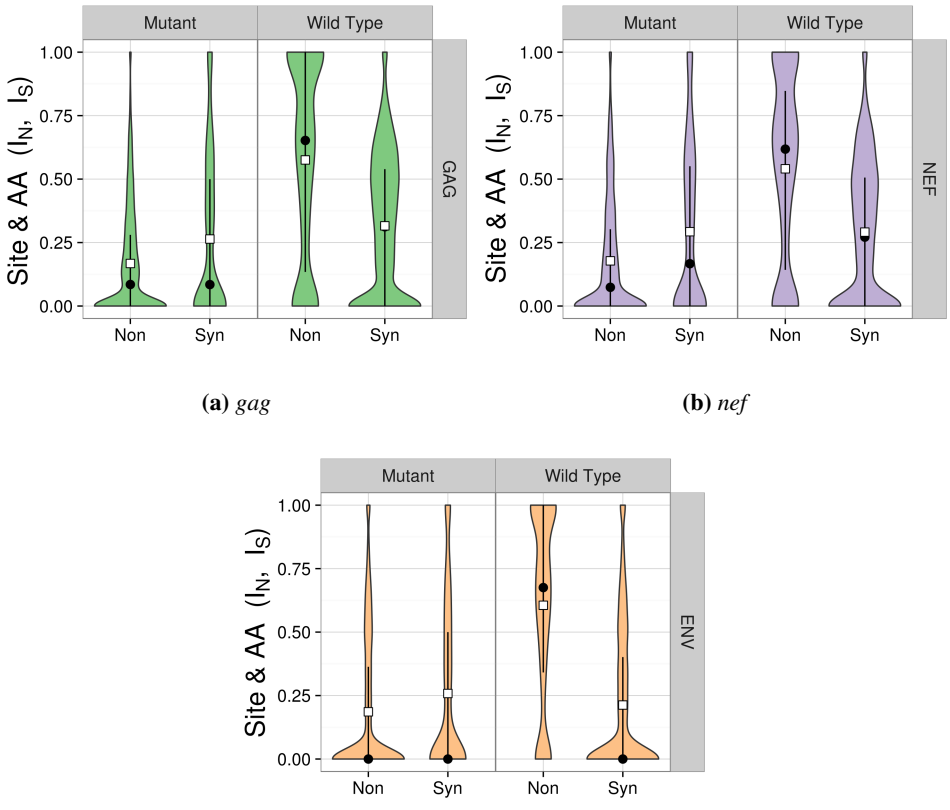
(b) *nef*

(c) *env*

Figure 3.9: Nonsynonymous substitutions to HLA-matched amino acids tend to map deeper in within-host phylogenies. Violin plots summarize the distributions of I_N and I_S, stratified by substitutions *to* Matched or Unmatched amino acids at each site. Each constituent site I_N and I_S respectively represents the proportion of site nonsynonymous and synonymous substitutions (to the specific residues) mapped to internal branches of the phylogeny containing both baseline and followup sequences from a single patient. A substitution was labeled 'Matched' if it resulted in an amino acid Matched with the patient HLA, and 'Unmatched' otherwise. Black point ranges indicate median and IQR. White squares indicate mean. Width of violins represents total matched or unmatched sites.

Although unmatched amino acids typically appeared only at the tips, mutations to unmatched wild-type appeared to be strongly selected in the absence of CTL pressure very early on in the evolutionary history of the infection (Mann-Whitney U, $P < 1.4 \times 10^{-168}$). Wild-type amino acids always appeared earlier in the phylogeny than mutant amino acids, but amongst Matched polymorphisms, only *nef* wild-types were significantly earlier than mutants ($P = 6.1 \times 10^{-39}$). There was no significant difference between the appearance of Matched or Unmatched wild-type amino acids (Mann-Whitney U, $P > 0.29$).

Only 0.038% of amino acid substitutions exhibited a Matched Escape amino acid mutating towards an Unmatched wild-type amino acid, indicating that CTL selective pressure continued and was usually stronger than the wild-type fitness advantage. At sites in which the Matched Escape reverted to an Unmatched wild-type (Figure 3.11), *nef* imparted the strongest drive towards wild-type reversion. Amino acid substitutions to wild-types were significantly maintained by synonymous substitutions ($I_N > I_S$, Mann-Whitney U, Benjamini-Hochberg corrected $P = 5.4 \times 10^{-5}$). In *env*, nonsynonymous substitutions to Unmatched wild type were generally earlier than synonymous substitutions, but there were enough nonsynonymous substitutions at the tips to deem Unmatched wild-type amino acids neutrally selected with a median $I_N = 0.49$ (Mann-Whitney U, Benjamini-Hochberg corrected , $I_N < I_S$, $P = 0.0034$). *gag* had the highest baseline mean frequency of wild-type amino acids (94%) which were maintained by synonymous substitutions deep in the tree, but there were insufficient nonsynonymous substitutions to determine if there was significant active amino acid-level selection towards wild-type (Mann-Whitney U, Benjamini-Hochberg corrected , $I_N < I_S$, *gag* $P = 0.89$). Although the precise timing of each substitution could not be reconstructed due to computational complexity of Bayesian inference, the wild-type variant to which the Matched Escape variant mutated towards already existed either as a major or minor variant at baseline sampling for all genes.



(a) *gag*

(b) *nef*

(c) *env*

Figure 3.10: Nonsynonymous substitutions to wild-type residues occur deeper in the phylogenies. The violin plot distributions of I_N and I_S are stratified at each site by substitutions **to** wildtype amino acid (HIV1-B subtype consensus). Black point ranges indicate median and IQR. White squares indicate mean.

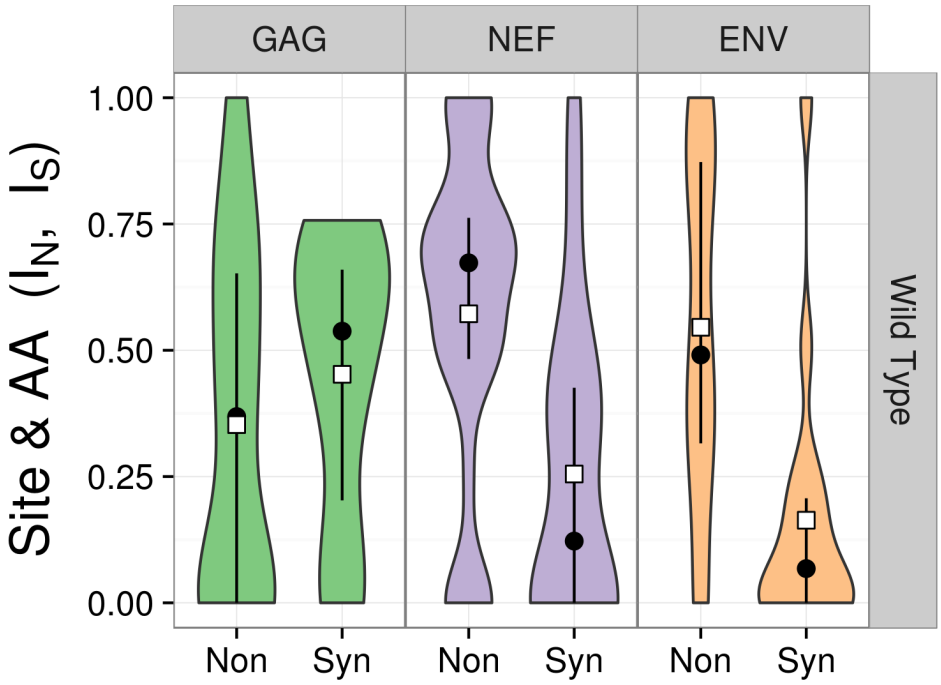


Figure 3.11: Violin plot distributions of I_N and I_S for site substitutions towards Unmatched wild type amino acid from Matched Escape amino acid. HIV1-B subtype consensus used as wild type. Black point ranges indicate median and IQR. White squares indicate mean.

3.4 Within-host HIV Evolution During Drug Treatment

3.4.1 Background

In addition to characterizing putative selection from immune response, we sought to validate how well Umberjack could detect selection in a dataset exhibiting very clear and strong directional selection. As such, we compared the treatment-naive dataset to a previously published Maraviroc clinical drug trial dataset [68].

We employed Umberjack to profile within-host evolution of the 4 patients undergoing HIV drug treatment in this trial. Patients were sampled before and after the addition of Maraviroc to an existing drug regimen. Over a median of 53

days, 4-5 longitudinal samples were taken per patient. Samples were amplicon sequenced with single 454 reads. Although the amplicon sequencing resulted in far more sequence overlap than would be present in shotgun sequencing, low-quality clipping lead to variable length reads with jagged alignments, requiring the use of Umberjack to infer evolutionary statistics at end positions of the alignment. Enhanced Sensitivity Trofile Assays (ESTA) for three out of four patients indicated that their viral populations used both CCR5 and CXCR4 co-receptors before starting Maraviroc treatment. The co-receptor usage for each virus was also inferred with Geno2Pheno [3, 68], a machine learning algorithm that uses viral sequence to build its predictions. Each patient displayed a marked increase in viruses using CXCR4 between the baseline and final sampling.

3.4.2 Results

Overall, sites from the Maraviroc treated HIV populations were under slightly purifying selection (median $I = 0.47$). Even though there were several longitudinal samples per patient, I calculated from the substitutions accumulated from baseline at each sample was not a predictive measure of whether the following sample would experience a sweep, since sweeps occurred at different rates amongst sites and patients. That is, $I > 0$ at a site in the previous sample would not mean that the next sample would not incur a sweep. However, when examining I across all samples, sites in which a variant swept throughout the population in the drug treated dataset were associated with higher I-statistics ($P = 7.8 \times 10^{-16}$, Mann Whitney U) (Figure 3.12).

Further, the stronger directional selection due to drug pressure is evident in the higher I of the drug treated dataset (median I=0.47) compared to the untreated dataset (median I = 0.22), indicating that drug resistance mutations proliferated in the treated population more so than CTL escape mutations in the untreated population. Drug treatment imparts a constant selection pressure, whereas the CTL immune response varies depending on the CTL cell population size within the patient, leading to variable and weaker drive towards CTL escape [5].

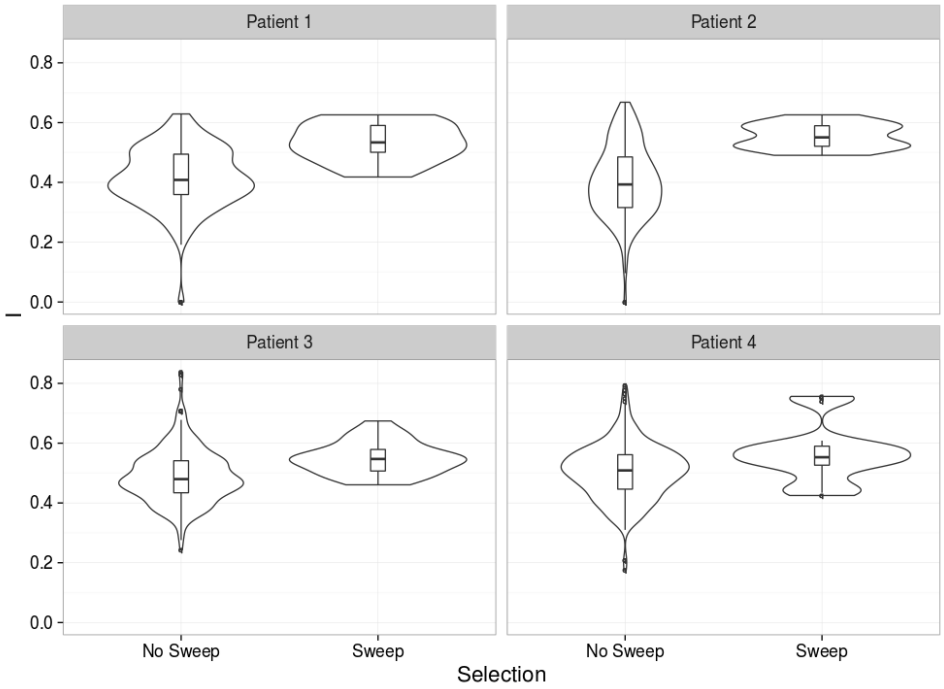


Figure 3.12: Violin and boxplot of I statistic for sites in ENV gene (V3 region) of viral populations treated with Maraviroc.
'Sweep' and 'No sweep' refer to sites in which a variant proliferated throughout the entire population by the final sample and sites in which a variant did not proliferate throughout the entire population by the final sample, respectively.

3.5 Comparing Methods for Detecting Selection

Frequency based methods of detecting selection can confound selection with demographic events such as contraction and expansion of population size [91]. However, site frequency is easily calculated and does not require long reads to obtain evolutionary relationships. Phylogenetic methods will take evolutionary relationships into account to avoid demographic confounding [2, 91]. However, when counting methods are used to evaluate the distribution of mutations in the phylogeny, there is low power to detect selection at sites with low diversity [80]. Moreover, in the untreated patient dataset examined in this study, only 7% of reads were long enough (>263bp) to sufficiently fill a window for phylogenetic analysis due

to a shorter median fragment size of 109bp (IQR 63-189bp). The median depth of coverage of sites that passed window-inclusion thresholds was 362-fold whereas the median depth of coverage across all sites was 7306-fold. This resulted in drastically lower percentages of sites considered under significant selection in comparison to a frequency-based method. Thus, while it seems intuitively true that most sites would not be under directional selection, it is possible that a reduced power of the method, combined with fewer observations per site, prevented the phylogenetic method from detecting sites under weak directional selection.

A disadvantage of aggregating site substitutions across all internal and external branches in I, I_N, and I_S statistics is that we lose the connection of substitutions along lineages. If our assumption that selection is similar across all lineages at a site is inaccurate, then it is possible for lineages exposed to purifying selection or directional selection in different parts of the phylogeny will cancel each other out. However, the lower resolution of the I selection statistic is more robust to sampling and topological errors during window phylogenetic construction, as well as limited amino acid diversity. For example, the $I_N - I_S$ statistic computed for subject BC27, window 271-570bp in HIV-1 *gag*, codon site 147 was -0.01, which is basically indistinguishable from a neutrally-evolving site. Thus, we fail to detect strong directional selection towards the HLA-matched amino acid (leucine), to which substitutions were mapped onto multiple lineages in the within-host phylogeny (Figure 3.13). Breaking down the window-site by substitutions to Matched amino acids shows that the Matched mutations appear early during the population history and are maintained deep within the tree by synonymous substitutions (Matched $I_N = 0.71$, $I_S = 0.66$). Further, all substitutions to unmatched amino acids only occur at the tips with $I_{\text{N,Unmatched}} = I_{\text{S,Unmatched}} = 0$.

Retaining our focus on this patient-window-site combination, we note the utility of deep sequencing compared to conventional bulk Sanger sequencing that predominates the study of HIV adaptation to the CTL-mediated immune response. With deep sequencing, we were able to pick up the change in amino acid consensus sequence from baseline to followup as isoleucene (I) $\rightarrow$ leucine (L). In the baseline sample, 79% of sequences carried the codon ATA encoding I at this position. Subsequently in the follow-up sample, only 44% of sequences carried ATA, while 56% encoded L (TTA 37%, CTA 18%, CTG & CTT & TTG $<$1%). Depend-

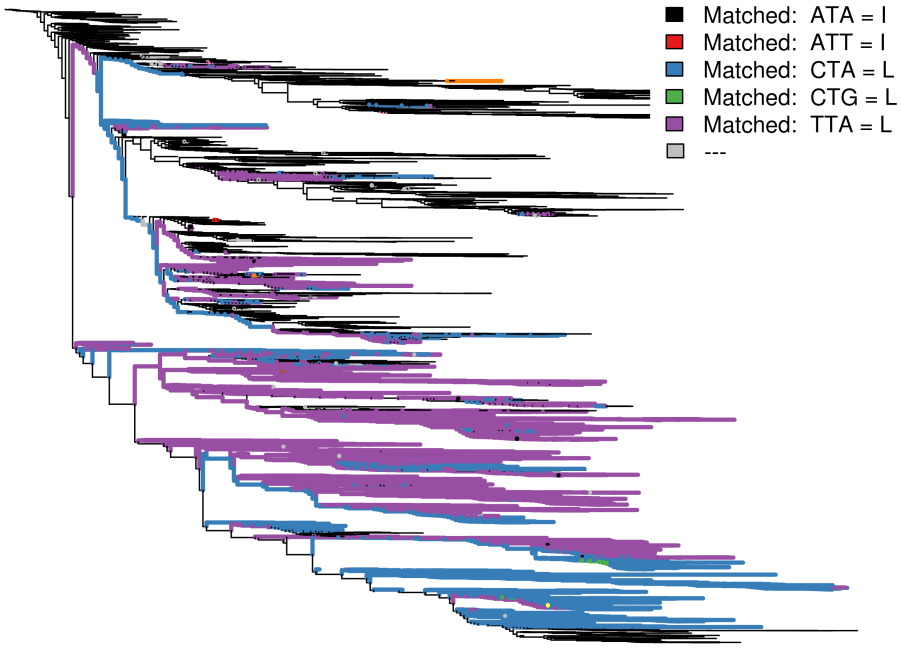


Figure 3.13: Onset of directional selection. All colors not listed in legend are Unmatched amino acid lineages. Phylogeny created from window of alignment containing baseline and followup sequences from patient BC27. Window coordinates 271-570bp, 1-based with respect to HXB2 *gag*. Majority baseline sampling tips code for isoleucine (I) and get replaced by blue or purple tips at follow sampling, which code for leucine (L).

ing on the mixture calling algorithm, bulk Sanger sequencing may not have been able to differentiate between I and L amino acids at followup.

From the various statistics implemented and examined, we recommend the use of I_N and I_S for detection of directional selection for within-host populations over $dN - dS$ and frequency based methods. Given a longitudinal sequencing dataset with sufficient fragment lengths and genetic diversity, it will be superior to $dN - dS$ in detecting temporal changes in nonsynonymous and synonymous substitutions, and will be able to determine if high prevalence variants were simply inherited during genetic bottlenecks.

Chapter 4

Conclusions

Umberjack (Chapter 2) overcomes the phylogenetic challenges of missing homology in shotgun sequencing and has been validated on simulated datasets representing a wide range of sequence quality parameters and sampling scenarios of HIV infections.

After successfully applying Umberjack to quantifying HIV evolution within hosts in the context of the CTL immune response (Chapter 3), significant associations with patient HLA using the phylogenetic approaches were found in addition to a frequency based approach. Although *gag* and *nef* were $>95\%$ conserved at most sites, Umberjack was able to detect directional selection in their sites known to be under CTL pressure. Further, we found evidence that the HIV populations from all patients experienced reversions to the wild type HIV1-B consensus sequence close to the most recent ancestor of the sampled populations. These wild type amino acids were further maintained through synonymous substitutions. However, very little reversion to unmatched wild-type occurred from matched escape variants, indicating that sites seldom experienced switches in selective pressures between general fitness requirements and evasion of the CTL response. In drug treatment datasets (Section 3.4), the phylogenetic statistics generated by Umberjack were able to highlight sites undergoing directional selection, as well as showcase the stronger selective pressures of drug treatment compared to the immune response at the chronic stage of infection.

4.1 Future Directions

Although second generation high throughput short reads can reveal the effect of evolutionary pressures on viral populations, their shortcomings in read length require additional processing such as Umberjack in order to be useful. With third generation reads coming into play, such as 25kbp Single Molecule Real Time reads [25] and Oxford Nanopore reads [90], we have the ability to cover the entire genomes of entire viruses such as HIV [29]. Although their high error rates of over 15% for Single Molecule Real Time reads and over 30% for Oxford Nanopore reads makes them currently prohibitive for phylogenetic analysis [43], they pose a promising method to connect distant regions of the genome, which we can leverage to examine genetic properties such as linkage disequilibrium. New techniques to reduce the error rate of Single Molecular Real Time reads have come forward, such as circularized consensus sequencing in which DNA is circularized and sequenced multiple times in a loop. Using this technique, error rates have been reported to be reduced to 2.5%, albeit with a reduction in read length to 2.5kbp [43].

Using phylogenetic profiling, we can examine regions of viral genomes that remain conserved through various evolutionary pressures, allowing us to design better vaccines. At present, there have been no successful antibody-based or CTL-based vaccines. The most successful antibody-based vaccine thus far, the Thai RV144 vaccine, demonstrated 31% effectiveness in prevention of HIV infection [39]. The CTL-based Step vaccine demonstrated no overall protection against HIV infection, and only small reductions in viral load during acute phase within certain patient HLA groups. Comparison of genetic distance between the Step vaccine epitope inserts and the HIV populations of the vaccinated and placebo patients revealed that vaccinated populations were more distant from the vaccine insert, suggesting a higher rate of escape mutations amongst vaccinated patients [24].

CTL based vaccine designs follow either a mosaic design which uses multiple common epitope variants as vaccine inserts to induce a broad CTL response, or conserved design which uses conserved epitopes to reduce mutations for CTL escape [69]. A successful vaccine design should not only take into account evolution within the HIV population, but evolution within the host immune system as well. Different lines of T-cells recognize different epitopes, and competition ex-

ists amongst these lineages such that T-cells that successfully target more epitopes proliferate to become the dominant T-cells within the immune system. Dominant T-cell lines have been reported to target highly mutating epitopes more than conserved epitopes, especially during acute infection, allowing escape mutations to rapidly expand within the HIV population. Supporters for conserved design vaccines argue that dominance of T-cell lines that prefer highly mutating epitopes becomes less of an issue if vaccine epitopes are all conserved, since all T-cell responses would focus on conserved regions, reducing early escape [59]. Further, mosaic designs would be ineffective against infections in which matched CTL escape mutations are transmitted between patients [12]. Using phylogenetic analysis to confirm that epitopes are conserved due to fitness as opposed to founder effects helps ensure the best epitopes are selected for a successful vaccine.

Bibliography

[1] C. S. Alexander, W. Dong, K. Chan, N. Jahnke, M. V. O'Shaughnessy, T. Mo, M. A. Piaseczny, J. S. Montaner, and P. R. Harrigan. Hiv protease and reverse transcriptase variation and therapy outcome in antiretroviral-naive individuals from a large north american cohort. *AIDS*, 15(5):601–7, Mar 2001. → pages 49

[2] N. H. Barton. Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 355(1403):1553–1562, 2000. → pages 70

[3] N. Beerenwinkel, M. Däumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter. Geno2pheno: estimating phenotypic drug resistance from hiv-1 genotypes. *Nucleic acids research*, 31(13): 3850–3855, 2003. → pages 69

[4] C. Bleidorn. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity*, 14(1):1–8, 2016. → pages 4

[5] S. Bonhoeffer, R. M. May, G. M. Shaw, and M. A. Nowak. Virus dynamics and drug therapy. *Proceedings of the National Academy of Sciences*, 94 (13):6971–6976, 1997. → pages 69

[6] P. Borrow, H. Lewicki, B. H. Hahn, G. M. Shaw, and M. B. Oldstone. Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J Virol*, 68(9):6103–10, Sep 1994. → pages 47

[7] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, et al. The potential and challenges of nanopore sequencing. *Nature biotechnology*, 26(10): 1146–1153, 2008. → pages 4

[8] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. → pages 38

[9] W. Brockman, P. Alvarez, S. Young, M. Garber, G. Giannoukos, W. L. Lee, C. Russ, E. S. Lander, C. Nusbaum, and D. B. Jaffe. Quality scores and snp detection in sequencing-by-synthesis systems. *Genome research*, 18(5): 763–770, 2008. → pages 22

[10] D. S. Burke. Recombination in hiv: an important viral evolutionary strategy. *Emerging infectious diseases*, 3(3):253, 1997. → pages 7

[11] J. M. Carlson, C. J. Brumme, E. Martin, J. Listgarten, M. A. Brockman, A. Q. Le, C. Chui, L. A. Cotton, D. J. Knapp, S. A. Riddler, et al. Correlates of protective cellular immunity revealed by analysis of population-level immune escape pathways in HIV-1. *Journal of virology*, pages JVI–01998, 2012. → pages 2, 48, 51, 60

[12] J. M. Carlson, V. Y. Du, N. Pfeifer, A. Bansal, V. Y. Tan, K. Power, C. J. Brumme, A. Kreimer, C. E. DeZiel, N. Fusi, et al. Impact of pre-adapted hiv transmission. *Nature medicine*, 22(6):606–613, 2016. → pages 75

[13] K. Chen and L. Pachter. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol*, 1(2):e24, 2005. → pages 11

[14] J. M. Cuevas, R. Geller, R. Garijo, J. López-Aldeguer, and R. Sanjuán. Extremely high mutation rate of hiv-1 in vivo. *PLoS Biol*, 13(9):e1002251, 2015. → pages 33

[15] A. G. Dalgleish, P. Beverley, P. R. Clapham, D. H. Crawford, M. F. Greaves, and R. A. Weiss. The cd4 (t4) antigen is an essential component of the receptor for the aids retrovirus. *Nature*, 312(5996):763–767, 1983. → pages 1

[16] D. Darriba, M. Weiß, and A. Stamatakis. Prediction of missing sequences and branch lengths in phylogenomic data. *Bioinformatics*, page btv768, 2016. → pages 12

[17] A. De Bruyn, D. P. Martin, and P. Lefeuvre. Phylogenetic reconstruction methods: an overview. *Molecular Plant Taxonomy: Methods and Protocols*, pages 257–277, 2014. → pages 7

[18] H. Deng, R. Liu, W. Ellmeier, S. Choe, D. Unutmaz, M. Burkhart, P. D. Marzio, S. Marmon, R. E. Sutton, C. M. Hill, et al. Identification of a major co-receptor for primary isolates of hiv-1. 1996. → pages 2

[19] K. Deng, M. Pertea, A. Rongvaux, L. Wang, C. M. Durand, G. Ghiaur, J. Lai, H. L. McHugh, H. Hao, H. Zhang, et al. Broad ctl response is required to clear latent hiv-1 due to dominance of escape mutations. *Nature*, 517(7534):381–385, 2015. → pages 1

[20] R. Desper and O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of computational biology*, 9(5):687–705, 2002. → pages 24

[21] G. Doitsh, N. L. Galloway, X. Geng, Z. Yang, K. M. Monroe, O. Zepeda, P. W. Hunt, H. Hatano, S. Sowinski, I. Muñoz-Arias, et al. Cell death by pyroptosis drives cd4 t-cell depletion in hiv-1 infection. *Nature*, 505(7484): 509–514, 2014. → pages 1

[22] P. Dorr, M. Westby, S. Dobbs, P. Griffin, B. Irvine, M. Macartney, J. Mori, G. Rickett, C. Smith-Burchnell, C. Napier, et al. Maraviroc (uk-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor ccr5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrobial agents and chemotherapy*, 49(11):4721–4732, 2005. → pages 2

[23] A. J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol*, 7:214, 2007. doi:10.1186/1471-2148-7-214. → pages 52

[24] P. T. Edlefsen, P. B. Gilbert, and M. Rolland. Sieve analysis in hiv-1 vaccine efficacy trials. *Current opinion in HIV and AIDS*, 8(5), 2013. → pages 74

[25] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, et al. Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009. → pages 4, 74

[26] B. Ewing, L. Hillier, M. C. Wendl, and P. Green. Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome research*, 8 (3):175–185, 1998. → pages 22

[27] M. R. Farhat, B. J. Shapiro, K. J. Kieser, R. Sultana, K. R. Jacobson, T. C. Victor, R. M. Warren, E. M. Streicher, A. Calver, A. Sloutsky, D. Kaur,

J. E. Posey, B. Plikaytis, M. R. Oggioni, J. L. Gardy, J. C. Johnston, M. Rodrigues, P. K. C. Tang, M. Kato-Maeda, M. L. Borowsky, B. Muddukrishna, B. N. Kreiswirth, N. Kurepina, J. Galagan, S. Gagneux, B. Birren, E. J. Rubin, E. S. Lander, P. C. Sabeti, and M. Murray. Genomic analysis identifies targets of convergent positive selection in drug-resistant mycobacterium tuberculosis. *Nat Genet*, 45(10):1183–9, Oct 2013. doi:10.1038/ng.2747. → pages 14

[28] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981. → pages 7

[29] E. B. Fichot and R. S. Norman. Microbial phylogenetic profiling with the pacific biosciences sequencing platform. *Microbiome*, 1(1):1, 2013. → pages 74

[30] W. Fletcher and Z. Yang. Indelible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, 26(8):1879–1888, 2009. → pages 33

[31] M. K. C. from Jed Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, and C. Candan. *caret: Classification and Regression Training*, 2016. URL http://CRAN.R-project.org/package=caret. R package version 6.0-64. → pages 39

[32] Y.-X. Fu and W.-H. Li. Statistical tests of neutrality of mutations. *Genetics*, 133(3):693–709, 1993. → pages 15, 16, 59

[33] V. V. Ganusov, N. Goonetilleke, M. K. Liu, G. Ferrari, G. M. Shaw, A. J. McMichael, P. Borrow, B. T. Korber, and A. S. Perelson. Fitness costs and diversity of the cytotoxic t lymphocyte (ctl) response determine the rate of ctl escape during acute and chronic phases of hiv infection. *Journal of virology*, 85(20):10518–10528, 2011. → pages 47

[34] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution*, 11(5): 725–736, 1994. → pages 15

[35] J. M. Gonzalez, M. C. Portillo, P. Belda-Ferre, and A. Mira. Amplification by pcr artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS One*, 7(1):e29973, 2012. → pages 11

[36] P. J. Goulder and D. I. Watkins. HIV and SIV CTL escape: implications for vaccine design. *Nature Reviews Immunology*, 4(8):630–640, 2004. → pages 1, 2, 47, 56

[37] P. J. Goulder and D. I. Watkins. Impact of mhc class i diversity on immune control of immunodeficiency virus replication. *Nature Reviews Immunology*, 8(8):619–630, 2008. → pages 61

[38] Y. Han, M. Wind-Rotolo, H.-C. Yang, J. D. Siliciano, and R. F. Siliciano. Experimental approaches to the study of hiv-1 latency. *Nature Reviews Microbiology*, 5(2):95–106, 2007. → pages 1

[39] B. F. Haynes, P. B. Gilbert, M. J. McElrath, S. Zolla-Pazner, G. D. Tomaras, S. M. Alam, D. T. Evans, D. C. Montefiori, C. Karnasuta, R. Sutthent, et al. Immune-correlates analysis of an hiv-1 vaccine efficacy trial. *New England Journal of Medicine*, 366(14):1275–1286, 2012. → pages 74

[40] J. Hermisson and P. S. Pennings. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4): 2335–52, Apr 2005. doi:10.1534/genetics.104.036947. → pages 17

[41] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. → pages 40

[42] W. Huang, L. Li, J. R. Myers, and G. T. Marth. Art: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2012. → pages 33

[43] X. Jiao, X. Zheng, L. Ma, G. Kutty, E. Gogineni, Q. Sun, B. T. Sherman, X. Hu, K. Jones, C. Raley, et al. A benchmark study on error assessment and quality control of ccs reads derived from the pacbio rs. *Journal of data mining in genomics & proteomics*, 4(3), 2013. → pages 74

[44] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. *Mammalian protein metabolism*, 3(21):132, 1969. → pages 8

[45] S. M. Krone and C. Neuhauser. Ancestral processes with selection. *Theoretical population biology*, 51(3):210–237, 1997. → pages 32

[46] S. Kryazhimskiy and J. B. Plotkin. The population genetics of dN/dS. *PLoS Genet*, 4(12):e1000304, Dec 2008. doi:10.1371/journal.pgen.1000304. → pages 59

[47] C. Kuiken, T. Leitner, B. Foley, B. Hahn, P. Marx, F. McCutchan, S. Wolinsky, B. Korber, G. Bansal, and W. Abfalterer. Hiv sequence compendium 2009. *Los Alamos, New Mexico: Los Alamos National Laboratory, Theoretical Biology and Biophysics*, 2009. → pages 64

[48] A. S. Lauring, J. Frydman, and R. Andino. The role of mutational robustness in rna virus evolution. *Nature reviews Microbiology*, 11(5): 327–336, 2013. → pages 20

[49] J. Lee, P. Müller, S. Sengupta, K. Gulukota, and Y. Ji. Bayesian feature allocation models for tumor heterogeneity. In *Statistical Analysis for High-Dimensional Data*, pages 211–232. Springer, 2016. → pages 12

[50] T. Lengauer, O. Sander, S. Sierra, A. Thielen, R. Kaiser, et al. Bioinformatics prediction of hiv coreceptor usage. *Nature biotechnology*, 25(12):1407–1410, 2007. → pages 3

[51] A. Leslie, K. Pfafferott, P. Chetty, R. Draenert, M. Addo, M. Feeney, Y. Tang, E. Holmes, T. Allen, J. Prado, et al. Hiv evolution: Ctl escape mutation and reversion after transmission. *Nature medicine*, 10(3): 282–289, 2004. → pages 64

[52] H. Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013. → pages 6, 33, 49

[53] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009. → pages 21

[54] B. Liang, M. Luo, J. Scott-Herridge, C. Semeniuk, M. Mendoza, R. Capina, B. Sheardown, H. Ji, J. Kimani, B. T. Ball, et al. A comparison of parallel pyrosequencing and sanger clone-based sequencing and its impact on the characterization of the genetic diversity of hiv-1. *PloS one*, 6 (10):e26745, 2011. → pages 3, 11

[55] H.-X. Liao, R. Lynch, T. Zhou, F. Gao, S. M. Alam, S. D. Boyd, A. Z. Fire, K. M. Roskin, C. A. Schramm, Z. Zhang, et al. Co-evolution of a broadly neutralizing hiv-1 antibody and founder virus. *Nature*, 496(7446):469–476, 2013. → pages 14, 51

[56] A. Liaw and M. Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002. → pages 38, 39

[57] D. J. Lipman, S. F. Altschul, and J. D. Kececioglu. A tool for multiple sequence alignment. *Proceedings of the National Academy of Sciences*, 86 (12):4412–4415, 1989. → pages 5

[58] J. Listgarten, Z. Brumme, C. Kadie, G. Xiaojiang, B. Walker, M. Carrington, P. Goulder, and D. Heckerman. Statistical resolution of ambiguous hla typing data. *PLoS Comput Biol*, 4(2):e1000016, 2008. → pages 49

[59] M. K. Liu, N. Hawkins, A. J. Ritchie, V. V. Ganusov, V. Whale, S. Brackenridge, H. Li, J. W. Pavlicek, F. Cai, M. Rose-Abrahams, et al. Vertical t cell immunodominance and epitope entropy determine hiv-1 escape. *The Journal of clinical investigation*, 123(1):380–393, 2013. → pages 75

[60] S. L. Liu, A. G. Rodrigo, R. Shankarappa, G. H. Learn, L. Hsu, O. Davidov, L. P. Zhao, and J. I. Mullins. HIV quasispecies and resampling. *Science*, 273(5274):415–6, Jul 1996. → pages 48

[61] Y. Liu, J. McNevin, J. Cao, H. Zhao, I. Genowati, K. Wong, S. McLaughlin, M. D. McSweyn, K. Diem, C. E. Stevens, et al. Selection on the human immunodeficiency virus type 1 proteome following primary infection. *Journal of virology*, 80(19):9519–9529, 2006. → pages 51

[62] C. Luo, D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis. Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community dna sample. *PloS one*, 7(2):e30087, 2012. → pages 4

[63] F. Maldarelli, M. Kearney, S. Palmer, R. Stephens, J. Mican, M. A. Polis, R. T. Davey, J. Kovacs, W. Shao, D. Rock-Kress, et al. Hiv populations are large and accumulate high genetic diversity in a nonlinear fashion. *Journal of virology*, 87(18):10313–10323, 2013. → pages 7, 12

[64] 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. *J Virol*, 69(8):5087–94, Aug 1995. → pages 55

[65] 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. → pages 33

[66] D. Martin, D. Posada, K. Crandall, and C. Williamson. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Research & Human Retroviruses*, 21 (1):98–102, 2005. → pages 19

[67] J. H. McDonald. *Handbook of biological statistics*, volume 3. Sparky House Publishing Baltimore, MD, 2014. → pages 37

[68] R. A. McGovern, J. Symons, A. F. Poon, P. R. Harrigan, S. F. van Lelyveld, A. I. Hoepelman, P. M. van Ham, W. Dong, A. M. Wensing, and M. Nijhuis. Maraviroc treatment in non-r5-hiv-1-infected patients results in the selection of extreme cxcr4-using variants with limited effect on the total viral setpoint. *Journal of Antimicrobial Chemotherapy*, page dkt153, 2013. → pages 11, 68, 69

[69] A. J. McMichael and W. C. Koff. Vaccines that stimulate t cell immunity to hiv-1: the next step. *Nature immunology*, 15(4):319, 2014. → pages 74

[70] A. J. McMichael and S. L. Rowland-Jones. Cellular immune responses to HIV. *Nature*, 410(6831):980–987, 2001. → pages 2

[71] A. G. Meyer and C. O. Wilke. The utility of protein structure as a predictor of site-wise dn/ds varies widely among hiv-1 proteins. *Journal of The Royal Society Interface*, 12(111):20150579, 2015. → pages 33

[72] M. Mild, R. R. Gray, A. Kvist, P. Lemey, M. M. Goodenow, E. M. Fenyö, J. Albert, M. Salemi, J. Esbjörnsson, and P. Medstrand. High intrapatient hiv-1 evolutionary rate is associated with ccr5-to-cxcr4 coreceptor switch. *Infection, Genetics and Evolution*, 19:369–377, 2013. → pages 3

[73] R. Mills, M. Rozanov, A. Lomsadze, T. Tatusova, and M. Borodovsky. Improving gene annotation of complete viral genomes. *Nucleic acids research*, 31(23):7041–7055, 2003. → pages 20

[74] J. Montaner. Therapeutic guidelines for antiretroviral (arv) treatment of adult hiv infection, 2015. → pages 3

[75] P. A. P. Moran et al. The statistical processes of evolutionary theory. *The statistical processes of evolutionary theory.*, 1962. → pages 32

[76] S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, 1994. → pages 25

[77] R. A. Neher and T. Leitner. Recombination rate and selection strength in hiv intra-patient evolution. *PLoS Comput Biol*, 6(1):e1000660, 2010. → pages 7, 32, 41

[78] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 3(5):418–426, 1986. → pages 26

[79] P. S. Pennings, S. Kryazhimskiy, and J. Wakeley. Loss and recovery of genetic diversity in adapting populations of hiv. *PLoS Genet*, 10(1): e1004000, 2014. → pages 7, 12

[80] S. L. K. Pond and S. D. Frost. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, 22(5):1208–1222, 2005. → pages 15, 52, 70

[81] S. L. K. Pond and S. D. Frost. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular biology and evolution*, 22(5):1208–1222, 2005. → pages 26, 30

[82] S. L. K. Pond and S. V. Muse. Hyphy: hypothesis testing using phylogenies. In *Statistical methods in molecular evolution*, pages 125–181. Springer, 2005. → pages 25

[83] A. F. Poon, L. C. Swenson, E. M. Bunnik, D. Edo-Matas, H. Schuitemaker, A. B. van't Wout, and P. R. Harrigan. Reconstructing the dynamics of hiv evolution within hosts from serial deep sequence data. *PLoS Comput Biol*, 8(11):e1002753, 2012. → pages 102

[84] A. F. Y. Poon, L. C. Swenson, W. W. Y. Dong, W. Deng, S. L. Kosakovsky Pond, Z. L. Brumme, J. I. Mullins, D. D. Richman, P. R. Harrigan, and S. D. W. Frost. Phylogenetic analysis of population-based and deep sequencing data to identify coevolving sites in the nef gene of HIV-1. *Mol Biol Evol*, 27(4):819–32, Apr 2010. doi:10.1093/molbev/msp289. → pages 4, 11

[85] S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth. Hiv haplotype inference using a propagating dirichlet process mixture model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(1):182–191, 2014. → pages 12

[86] D. A. Price, P. J. Goulder, P. Klenerman, A. K. Sewell, P. J. Easterbrook, M. Troop, C. R. Bangham, and R. 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. → pages 47

[87] M. N. Price, P. S. Dehal, and A. P. Arkin. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, 2009. → pages 10, 24

[88] M. N. Price, P. S. Dehal, and A. P. Arkin. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490, 2010. → pages 24

[89] T. Pupko, I. Pe'er, R. Shamir, and D. Graur. A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol*, 17(6): 890–6, Jun 2000. → pages 25

[90] J. Quick, A. R. Quinlan, and N. J. Loman. A reference bacterial genome dataset generated on the minion portable single-molecule nanopore sequencer. *Gigascience*, 3(1):1, 2014. → pages 74

[91] A. Ramírez-Soriano, S. E. Ramos-Onsins, J. Rozas, F. Calafell, and A. Navarro. Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics*, 179(1):555–567, 2008. → pages 15, 70

[92] D. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1):131–147, 1981. → pages 34

[93] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4): 406–425, 1987. → pages 7

[94] M. Salminen, J. Carr, D. Burke, and F. McCutchan. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS research and human retroviruses*, 11(11):1423, 1995. → pages 7, 19

[95] M. Schirmer, R. D'Amore, U. Z. Ijaz, N. Hall, and C. Quince. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17(1):1, 2016. → pages 23

[96] R. Shankarappa, J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins. Consistent viral evolutionary changes associated with the

progression of human immunodeficiency virus type 1 infection. *J Virol*, 73 (12):10489–502, Dec 1999. → pages 2

[97] R. Shankarappa, J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. 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. → pages 51

[98] E. Simon-Loriere and E. C. Holmes. Gene duplication is infrequent in the recent evolutionary history of rna viruses. *Molecular biology and evolution*, page mst044, 2013. → pages 20

[99] R. Smyth, T. Schlub, A. Grimm, V. Venturi, A. Chopra, S. Mallal, M. Davenport, and J. Mak. Reducing chimera formation during pcr amplification to ensure accurate genotyping. *Gene*, 469(1):45–51, 2010. → pages 50

[100] C. Spearman. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904. → pages 37

[101] M. Stein. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151, 1987. → pages 31

[102] S. Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17:57–86, 1986. → pages 8

[103] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. → pages 40

[104] J. J. Wiens. Missing data and the design of phylogenetic analyses. *Journal of biomedical informatics*, 39(1):34–42, 2006. → pages 7, 10

[105] J. J. Wiens and M. C. Morrill. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Systematic Biology*, page syr025, 2011. → pages 10

[106] O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel. Shorah: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC bioinformatics*, 12(1):1, 2011. → pages 12

[107] F. Zanini, J. Brodin, L. Thebo, C. Lanz, G. Bratt, J. Albert, and R. A. Neher. Population genomics of intrapatient hiv-1 evolution. *eLife*, 4: e11282, 2016. → pages 11, 64

[108] S. Zhou, C. Jones, P. Mieczkowski, and R. Swanstrom. Primer id validates template sampling depth and greatly reduces the error rate of next-generation sequencing of hiv-1 genomic rna populations. *Journal of virology*, 89(16):8540–8555, 2015. → pages 48

Appendix A

Supporting Materials

A.1 Simulated Datasets For Predicting Umberjack Accuracy

Feature	Min	Max
Generations	128	8192
Recombination Rate (recombinations/bp/generation)	0	5.1e-5
Mutation Rate 1 (mutations/bp/generation)	5.6e-7	8.4e-4
Mutation Rate 2 (mutations/bp/generation)	5.6e-7	8.4e-4
Mutation Rate 3 (mutations/bp/generation)	5.6e-7	8.4e-4
Umberjack Config	a	b
Read Coverage Per Individual	0	3
Mean Sequence Fragment Size (bp)	104	500

Table A.1: Parameter ranges for Latin hypercube sampling to generate simulated datasets to test Umberjack accuracy. Three randomly allocated contiguous sections of the genome were assigned a different mutation rate selected from the Latin hypercube sampling to simulate different mutation rates per gene. Umberjack Configuration a = [Window size = 150bp, Min window width coverage = 0.7, Min window read depth = 10, Min phred quality score = 15]. Umberjack Configuration b = [Window size = 300bp, Min window width coverage = 0.875, Min window read depth = 10, Min phred quality score = 20].

Table A.2

Generations	Mutation Rate 1	Mutation Rate 2	Mutation Rate 3	Recomb. Rate	Breakpoints	Coverage	Mean Fragment Size
283	1.03e-05	3.63e-05	1.87e-05	3.53e-05	3	5	135
1230	6.27e-05	5.54e-06	1.44e-05	1.08e-05	4	0	250
307	4.62e-05	6.33e-05	6.39e-05	2.17e-05	2	2	353
412	2.94e-05	5.37e-05	6.53e-05	2.43e-05	3	5	220
784	2.73e-05	4.26e-05	5.61e-05	1.70e-05	4	0	277
4491	5.68e-05	3.15e-06	7.43e-05	2.23e-06	3	1	120
474	7.05e-06	3.04e-05	4.82e-05	0.00e+00	0	2	425
2343	4.92e-05	5.68e-05	1.31e-05	5.69e-06	4	2	115
3614	2.13e-05	1.42e-05	3.91e-05	2.77e-06	3	3	245
675	3.32e-05	7.56e-05	8.17e-05	9.88e-06	2	2	487
502	3.43e-05	1.67e-05	5.03e-05	2.66e-05	4	6	192
881	8.34e-05	2.43e-05	1.12e-05	7.57e-06	2	0	170
1060	5.35e-05	8.26e-06	7.34e-06	3.14e-06	1	0	403
1174	4.20e-05	5.15e-05	4.43e-05	1.14e-05	4	2	205
4817	3.81e-05	6.25e-05	5.75e-05	1.38e-06	2	3	164
6284	3.45e-06	3.17e-05	3.78e-05	5.30e-07	1	0	184
2620	5.27e-05	1.06e-05	3.41e-05	1.27e-06	1	3	141

Continued on next page

Table A.2 – *Continued from previous page*

Generations	Mutation Rate 1	Mutation Rate 2	Mutation Rate 3	Recomb. Rate	Breakpoints	Coverage	Mean Fragment Size
775	7.14e-05	3.76e-05	4.54e-05	1.29e-05	3	4	434
1813	1.19e-05	2.92e-05	7.80e-05	5.52e-06	3	1	289
714	6.57e-05	5.12e-06	1.02e-05	1.40e-05	3	2	214
1308	6.39e-05	6.76e-05	3.20e-05	5.10e-06	2	0	358
1000	8.05e-06	4.55e-05	5.90e-05	6.67e-06	2	4	377
367	6.69e-05	4.47e-05	6.96e-05	0.00e+00	0	3	409
2788	7.09e-05	7.42e-05	6.91e-06	3.59e-06	3	1	285
339	4.14e-05	5.84e-05	2.74e-05	1.97e-05	2	6	382
1597	7.64e-05	1.33e-05	6.10e-05	6.26e-06	3	1	158
1525	3.92e-05	2.23e-05	5.35e-05	4.37e-06	2	7	391
5044	5.92e-05	8.30e-05	7.28e-05	1.32e-06	2	3	317
317	2.36e-05	7.84e-05	2.12e-05	2.10e-05	2	0	343
3424	7.32e-05	6.94e-05	7.95e-05	1.95e-06	2	2	370
547	7.76e-05	5.56e-05	7.70e-05	1.83e-05	3	6	493
5833	8.19e-05	9.96e-06	2.65e-05	5.71e-07	1	7	178
7772	3.58e-05	1.82e-05	1.84e-05	4.29e-07	1	1	483
268	4.80e-05	2.75e-05	8.41e-05	1.24e-05	1	0	335

Continued on next page

Table A.2 – *Continued from previous page*

Generations	Mutation Rate 1	Mutation Rate 2	Mutation Rate 3	Recomb. Rate	Breakpoints	Coverage	Mean Fragment Size
5668	2.61e-05	6.49e-05	1.54e-05	5.88e-07	1	1	261
417	1.79e-05	7.20e-05	4.34e-06	0.00e+00	0	4	302
7164	7.92e-05	1.96e-05	7.12e-05	0.00e+00	0	1	440
1396	2.31e-05	5.07e-05	2.71e-06	0.00e+00	0	0	329
3116	5.04e-05	3.91e-05	2.20e-05	3.21e-06	3	1	456
6981	6.78e-05	8.22e-05	6.16e-05	0.00e+00	0	1	231
1963	5.05e-06	4.69e-05	2.51e-05	5.09e-06	3	0	105
2076	6.09e-05	3.40e-05	4.28e-05	1.61e-06	1	4	323
2441	1.94e-05	2.59e-05	4.10e-05	1.37e-06	1	0	471
1753	1.47e-05	2.05e-05	3.55e-05	1.90e-06	1	5	415
4360	3.15e-05	6.05e-05	4.74e-05	7.65e-07	1	1	150
3024	5.55e-05	4.00e-05	6.63e-05	2.20e-06	2	1	266
3865	1.26e-05	7.65e-05	3.09e-05	8.62e-07	1	1	298
561	1.60e-05	6.93e-05	2.95e-05	1.19e-05	2	0	228
916	7.54e-05	8.08e-05	6.78e-05	0.00e+00	0	5	461
624	4.40e-05	4.88e-05	5.17e-05	1.07e-05	2	1	450

Continued on next page

Table A.2 – *Continued from previous page*

Generations	Mutation Rate 1	Mutation Rate 2	Mutation Rate 3	Recomb. Rate	Breakpoints	Coverage	Mean Fragment Size

Table A.2: Parameters for generating simulated datasets to predict Umberjack accuracy. Each row represents a simulated population and its simulated paired-end MiSeq sequence library. Recombination rate units in recombinations/bp/generation. Mutation rate units in mutations/bp/generation. Read coverage is per extent individual in the population. Sequencing fragment size in bp. Genome size = 900bp. Selection rate = 0.01/generation. Extent population size = 1000. 2x250bp MiSeq paired-end reads. Sequencing fragment size standard deviation = 100bp. Umberjack Configuration a = Window size = 150bp, Min window width coverage = 0.7, Min window read depth = 10, Min phred quality score = 15, Umberjack Configuration b = Window size = 300bp, Min window width coverage = 0.875, Min window read depth = 10, Min phred quality score = 20.

Table A.3

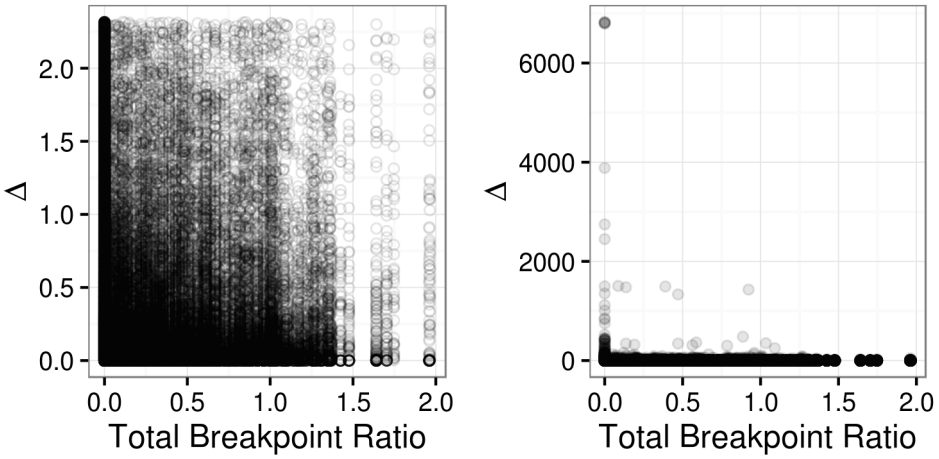
Name	Recombination Rate
Recombo1	1.29e-06
Recombo2	1.29e-06
Recombo3	1.29e-06
Recombo4	1.29e-06
Recombo5	1.29e-06
Recombo6	1.29e-06
Recombo7	0.00e+00
Recombo8	0.00e+00
Recombo9	0.00e+00
Recombo10	6.45e-07
Recombo11	6.45e-07
Recombo12	6.45e-07
Recombo13	3.23e-06
Recombo14	3.23e-06
Recombo15	3.23e-06
Recombo16	6.45e-06
Recombo17	6.45e-06
Recombo18	6.45e-06
Recombo19	6.45e-05

Continued on next page

Table A.3 – *Continued from previous page*

Name	Recombination Rate
Recombo20	6.45e-05
Recombo21	6.45e-05
Recombo22	1.29e-04
Recombo23	1.29e-04
Recombo24	1.29e-04
Recombo25	1.29e-05
Recombo26	1.29e-05
Recombo27	1.29e-05

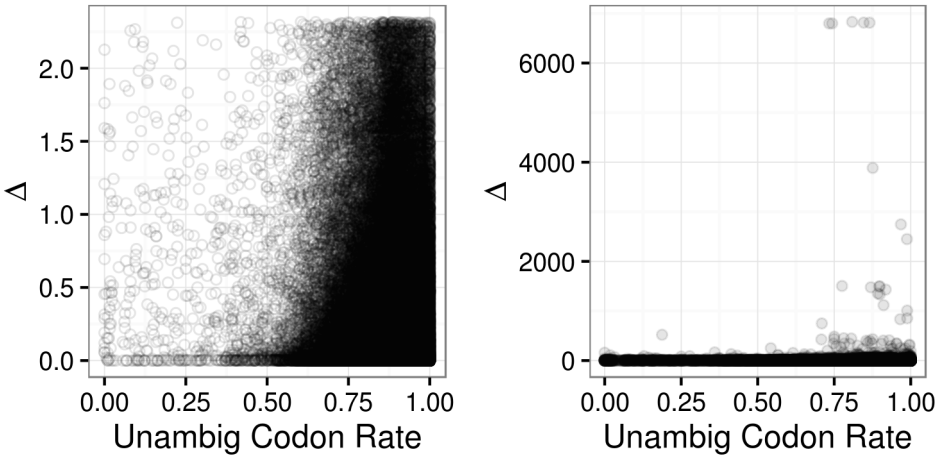
Table A.3: Parameters for generating simulated datasets to predict Umberjack accuracy under varying recombination. Each row represents a simulated population and its simulated paired-end MiSeq sequence library. Recombination rate units in recombinations/bp/generation. Mutation rate units in mutations/bp/generation. Read coverage is per extent individual in the population. Sequencing fragment size in bp. Genome size = 930bp. Selection rate = 0.01/generation. Mutation Rate = 4e-5 mutations/bp/generation. Generations = 5000. Extent population size = 100. 2x250bp MiSeq paired-end reads. Read coverage per individual = 2x. Sequencing fragment size mean = 375bp. Sequencing fragment size standard deviation = 75bp. Umberjack Configuration b = Window size = 300bp, Min window width coverage = 0.875, Min window read depth = 10, Min phred quality score = 20.



(a) Zoomed in to 1.5 × IQR of Δ

(b) All data

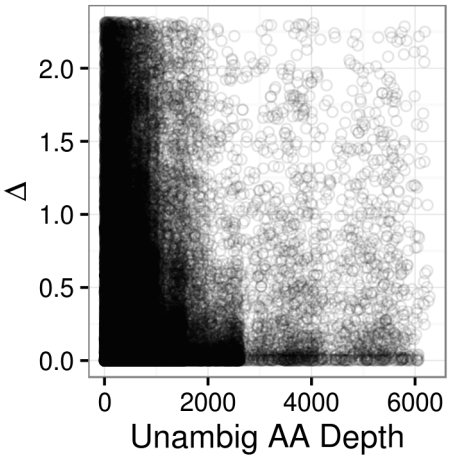
Figure A.1: Δ vs Window Total Breakpoint Ratio. Refer to Feature (A) for feature description.



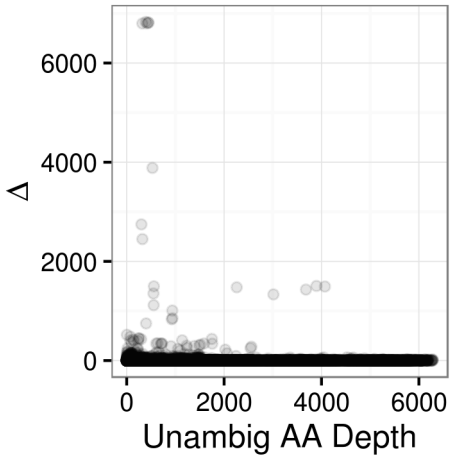
(a) Zoomed in to 1.5 × IQR of Δ

(b) All data

Figure A.2: Δ vs Window-Site Unambiguous Codon Rate. Refer to Feature (B) for feature description.

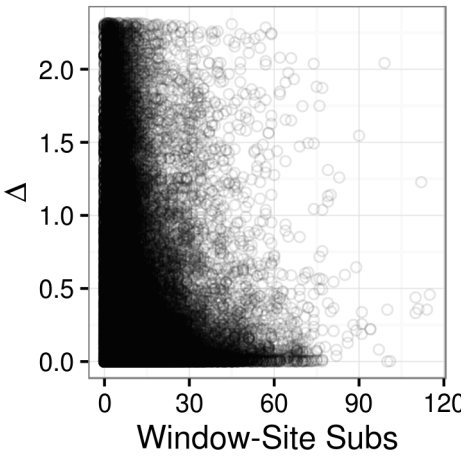


(c) Zoomed in to 1.5 × IQR of Δ

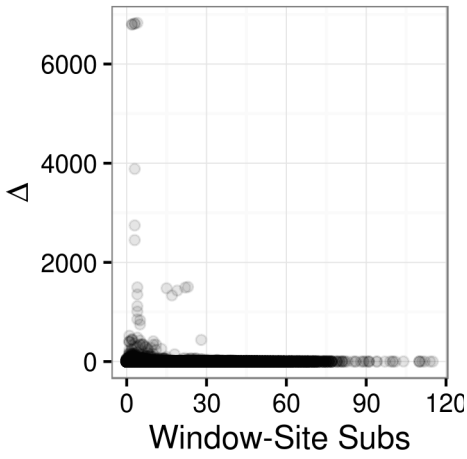


(d) All data

Figure A.2: Δ vs Window-Site Amino Acid Depth. Refer to Feature (C) for feature description.

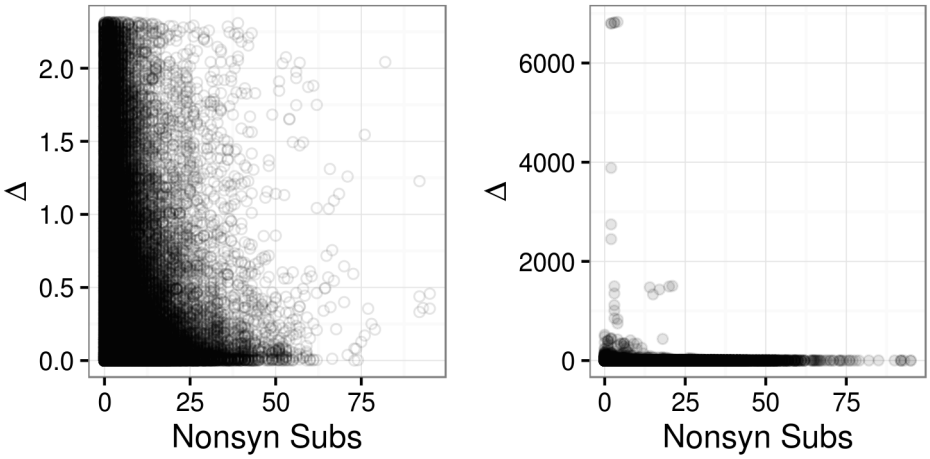


(a) Zoomed in to 1.5 × IQR of Δ



(b) All data

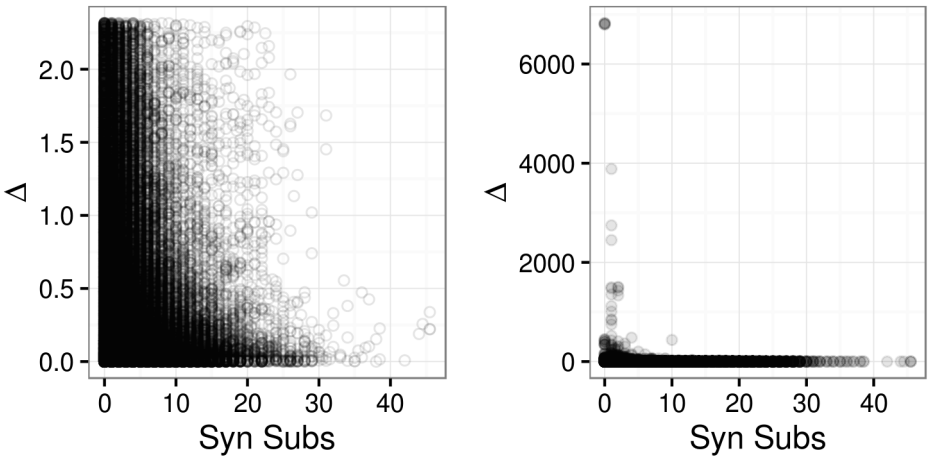
Figure A.3: Δ vs Window-Site Substitutions. Refer to Feature (D) for feature description.



(c) Zoomed in to 1.5 × IQR of Δ

(d) All data

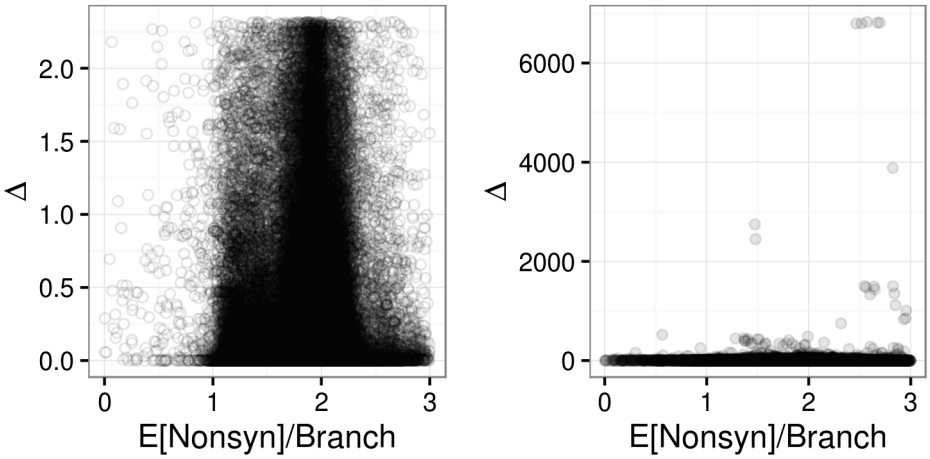
Figure A.3: Δ vs Window-Site Nonsynonymous Substitutions. Refer to Feature (E) for feature description.



(a) Zoomed in to 1.5 × IQR of Δ

(b) All data

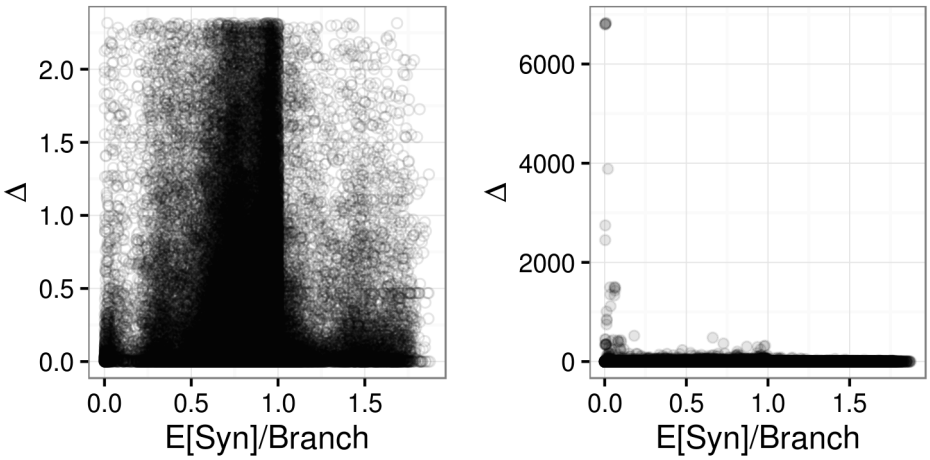
Figure A.4: Δ vs Window-Site Synonymous Substitutions. Refer to Feature (F) for feature description.



(c) Zoomed in to 1.5 × IQR of Δ

(d) All data

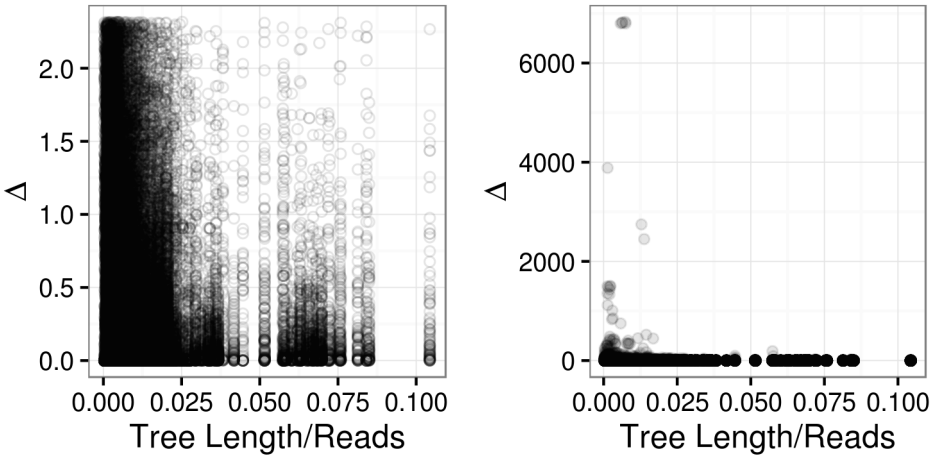
Figure A.4: Δ vs Window-Site Expected Nonsynonymous Substitutions Per Branch. Refer to Feature (G) for feature description.



(a) Zoomed in to 1.5 × IQR of Δ

(b) All data

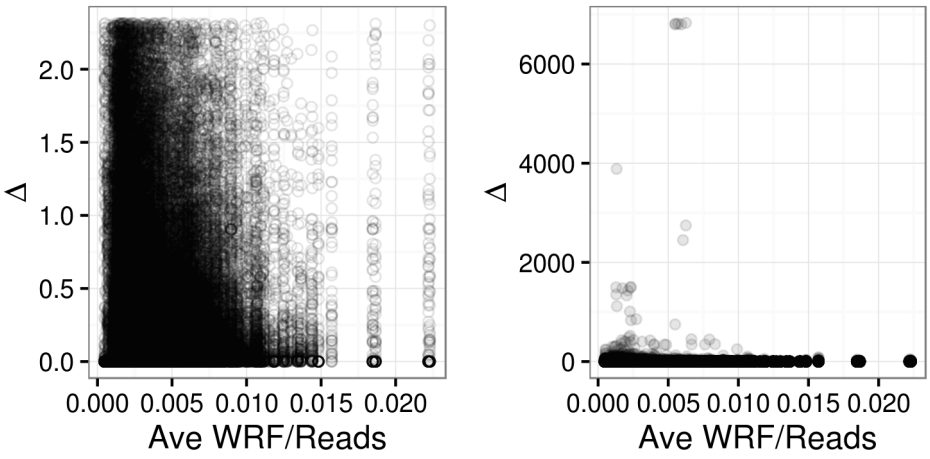
Figure A.5: Δ vs Window-Site Expected Synonymous Substitutions Per Branch. Refer to Feature (H) for feature description.



(c) Zoomed in to 1.5 × IQR of Δ

(d) All data

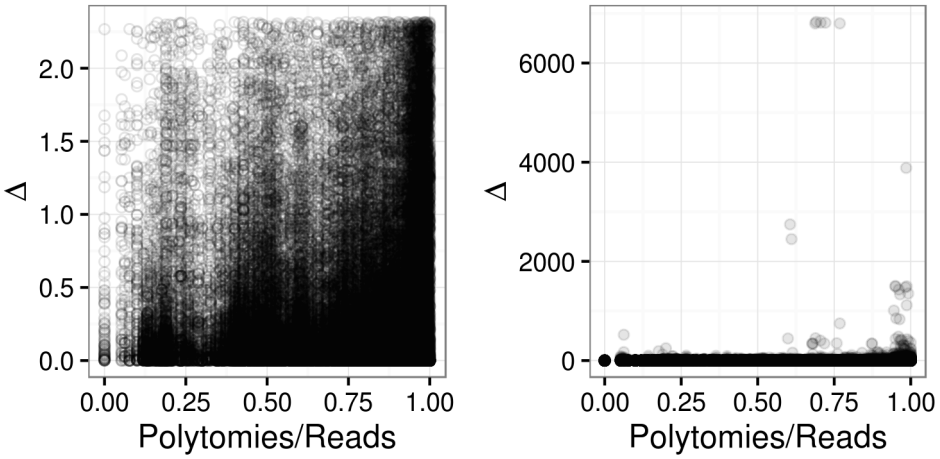
Figure A.5: Δ vs Normalized Window Tree Length. Refer to Feature (I) for feature description.



(a) Zoomed in to 1.5 × IQR of Δ

(b) All data

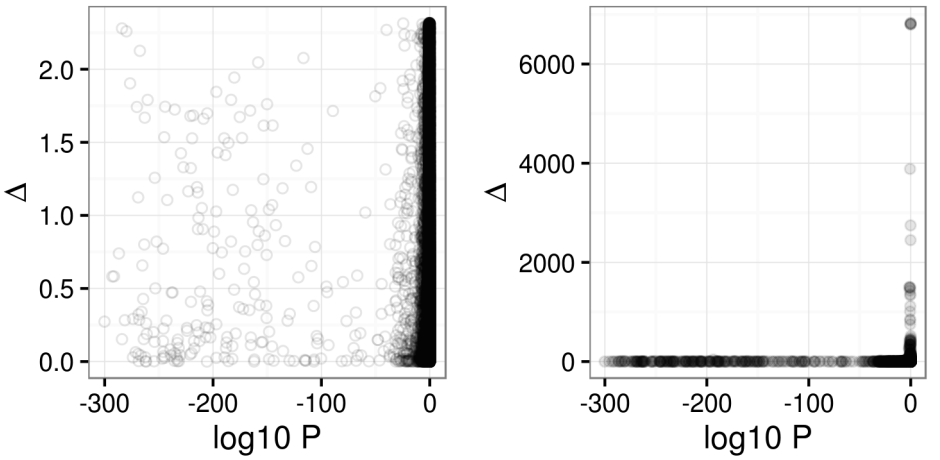
Figure A.6: Δ vs Normalized $\overline{WRF}$. Refer to Feature (J) for feature description.



(c) Zoomed in to 1.5 × IQR of Δ

(d) All data

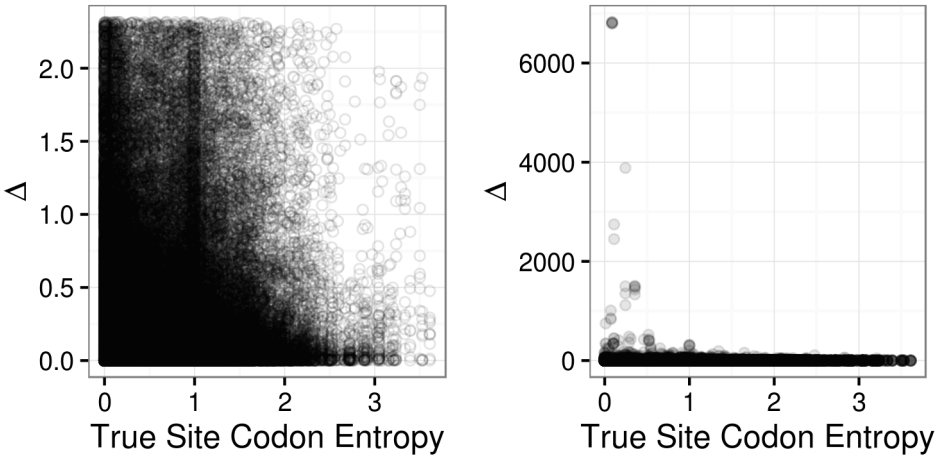
Figure A.6: Δ vs Normalized Polytomies. Refer to Feature (K) for feature description.



(a) Zoomed in to 1.5 × IQR of Δ

(b) All data

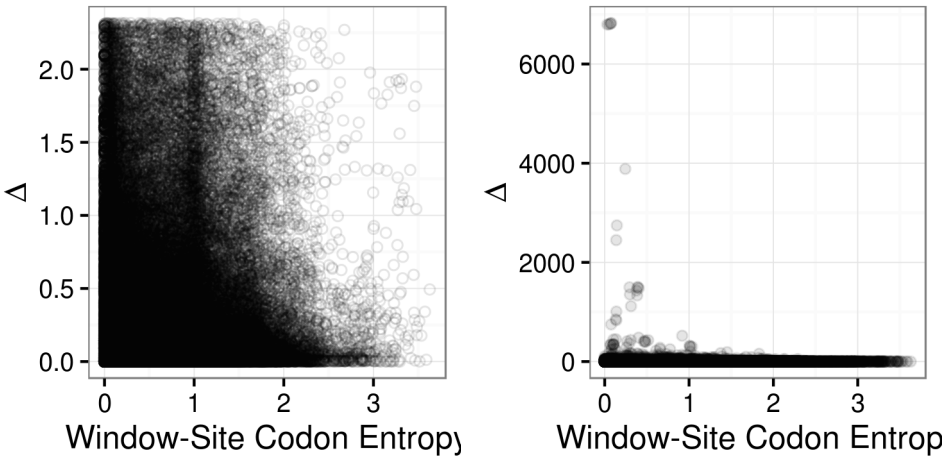
Figure A.7: Δ vs Codon Distribution P-value. P-values on x-axis are log_10 transformed. Refer to Feature (L) for feature description.



(c) Zoomed in to 1.5 × IQR of Δ

(d) All data

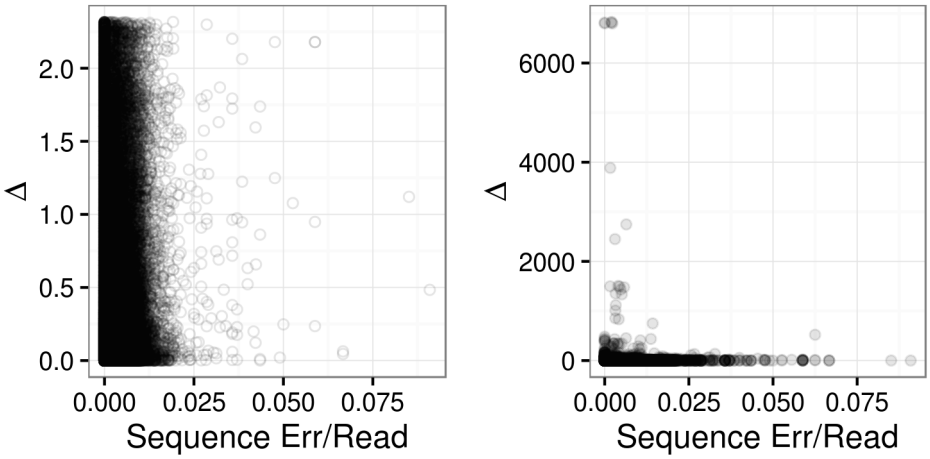
Figure A.7: Δ vs True Site Codon Entropy. Refer to Feature (M) for feature description.



(a) Zoomed in to 1.5 × IQR of Δ

(b) All data

Figure A.8: Δ vs Window-Site Codon Entropy. Refer to Feature (N) for feature description.



(c) Zoomed in to 1.5 × IQR of Δ

(d) All data

Figure A.8: Δ vs Window-Site Codon Entropy. Refer to Feature (O) for feature description.

A.2 Untreated Patient Metadata

We estimated that 1.5% of the RNA templates contained in a patient plasma sample made it to sequencing for NEF and GAG, and 4.4% for ENV. The estimates are based on the fraction of volume of patient plasma used in viral RNA extraction and PCR amplification using a similar procedure specified in [83]. The percentage of NEF and GAG templates sent to PCR amplification and sequencing was one third of the templates sent for ENV.

Table A.4

	Min	Mean	Max	25%	50%	75%
Baseline Viral Load (copies/mL)	2.1×10^4	1.2×10^5	7.5×10^5	9.0×10^4	1.0×10^5	1.0×10^5
Followup Viral Load (copies/mL)	2.1×10^4	1.1×10^4	4.2×10^5	7.8×10^4	1.0×10^5	1.0×10^5
Baseline CD4 Counts (cells/mm^3)	190	388	560	325	405	473
Followup CD4 Counts (cells/mm^3)	20.0	251	460	185	300	345
Years Infected at Baseline	0.35	3.8	13	1.2	2.6	5.7
Months Between Samples	3.0	16	51	8.4	12	17

Table A.4: Untreated Patient Measurement Quantiles. 32 of the patient sample viral loads hit a viral load assay measurement upper limit of 10^5 copies/mL, and 1 patient sample viral load hit another assay limit of 7.5×10^5 copies/mL. None were remeasured using assays with a higher limits. Viral loads that hit upper measurement limits were set to the limit in quantile calculations. CD4 samples were taken within 30 days of viral load samples. Baseline CD4 counts were missing for 14 patients, and followup CD4 counts were missing for 12 patients. 'Years Infected at Baseline' are estimated from BEAST timing of the most recent common ancestor of sequencing reads. Patient BC11 is missing an estimate for 'Years Infected at Baseline' since the BEAST runs did not converge.

Table A.5

	gag	*nef*	*env*
Primer Positions	734 - 1833 bp	8776 - 9593 bp	6945 - 7373 bp, 8325 - 8775bp
Min Templates	310	310	920
Mean Templates	1700	1700	5200
Max Templates	11000	11000	33000
25% Gene Depth Coverage	4500X	6200X	3400X
50% Gene Depth Coverage	8200X	10000X	6500X
75% Gene Depth Coverage	13000X	14000X	13000X
25% Template Depth Coverage	2.8X	4.1X	0.71X
50% Template Depth Coverage	5.6X	7.4X	1.6X
75% Template Depth Coverage	10X	11X	3.3X
Mean Fragment Size	137bp	147bp	104bp
Std Dev Fragment Size	94.5bp	100bp	64.7bp

Table A.5: Sequencing Statistics for Untreated Patient Samples. Primer positions are with respect to HXB2 reference strain (Accession K03455.1)

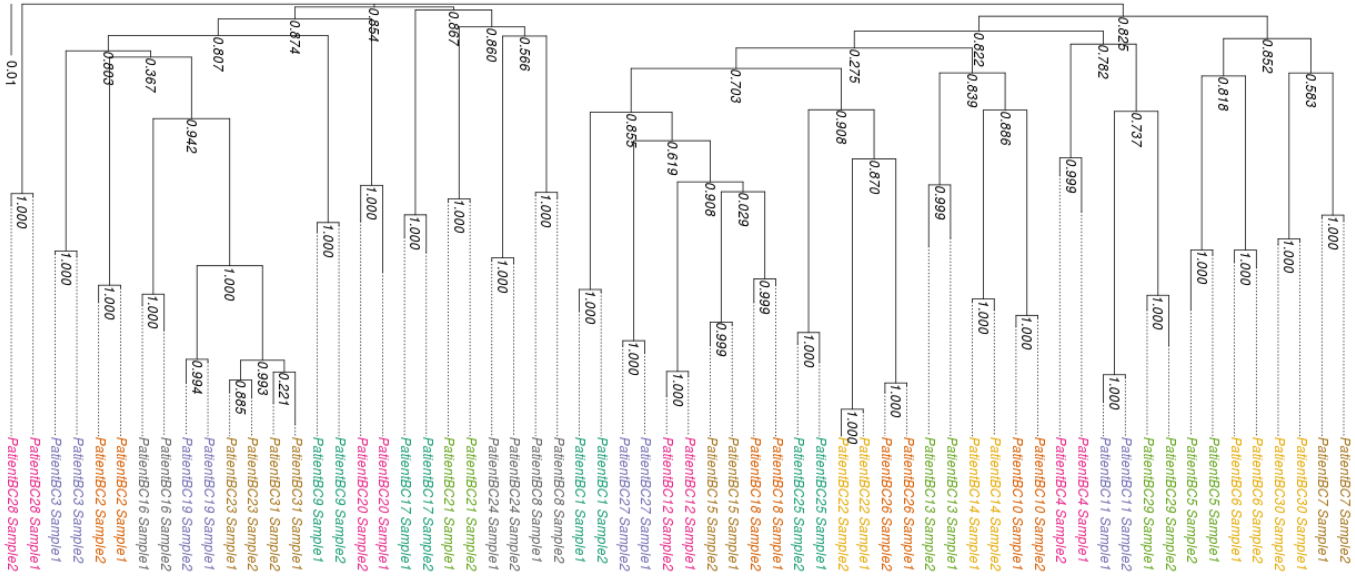


Figure A.9: Phylogenetic tree of GAG consensus sequence of samples from untreated patients. Node values indicate bootstrap support. Branch length units in nucleotide substitutions per site.

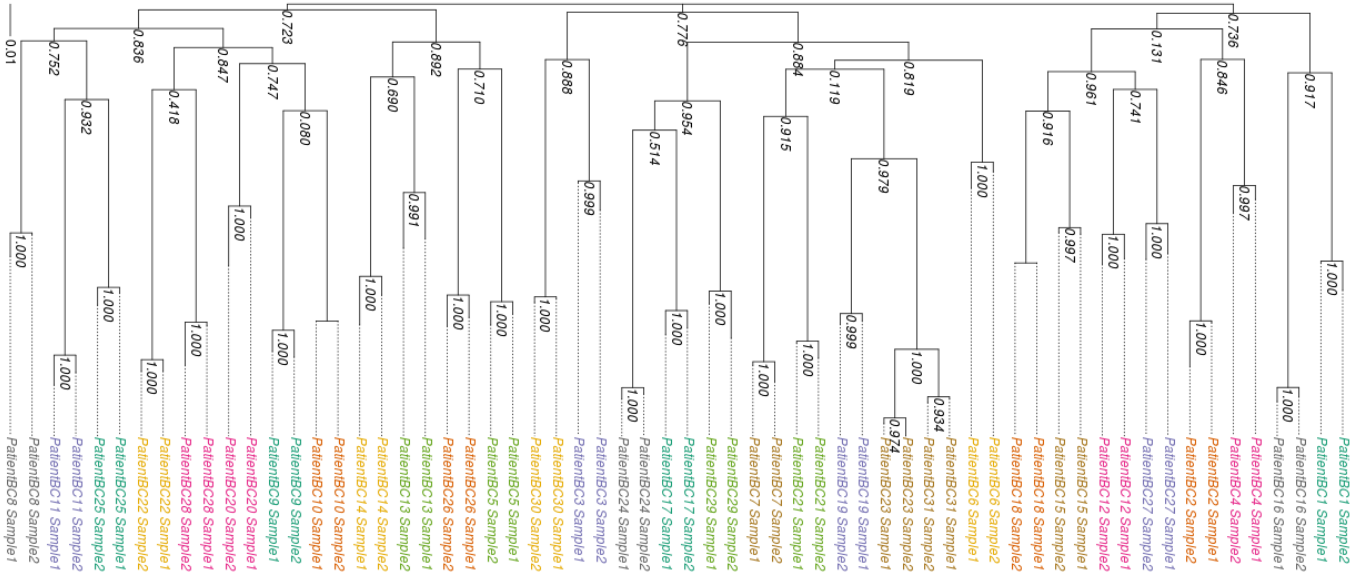


Figure A.10: Phylogenetic tree of NEF consensus sequence of samples from untreated patients. Node values indicate bootstrap support. Branch length units in nucleotide substitutions per site.

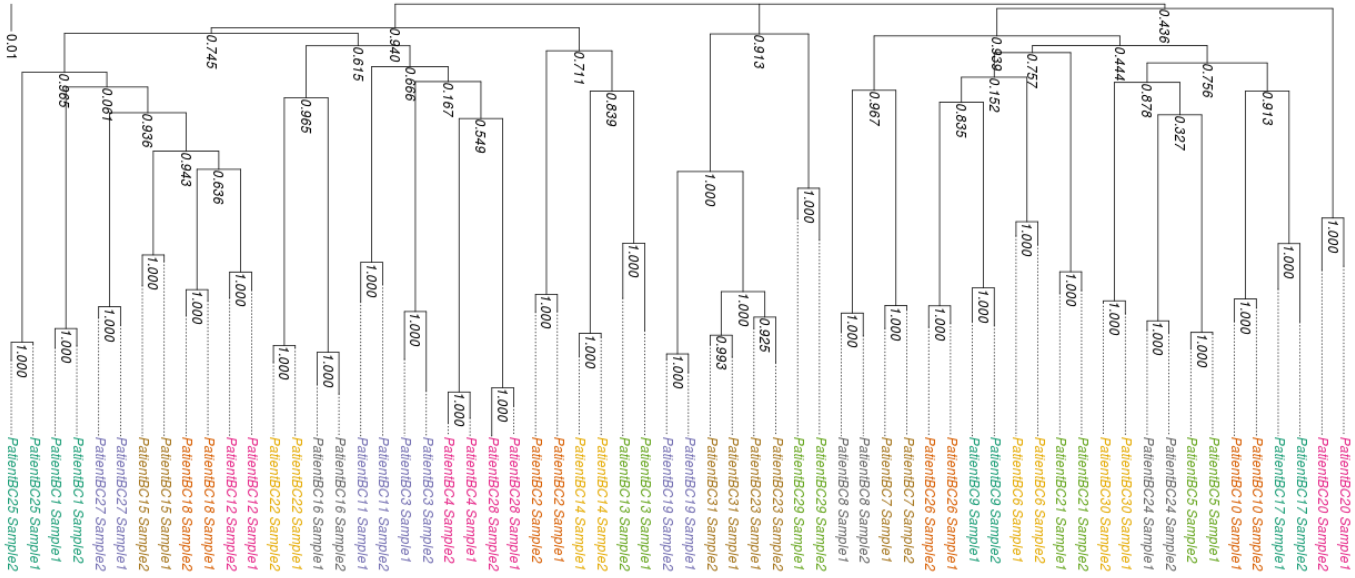


Figure A.11: Phylogenetic tree of ENV consensus sequence of samples from untreated patients. Node values indicate bootstrap support. Branch length units in nucleotide substitutions per site.

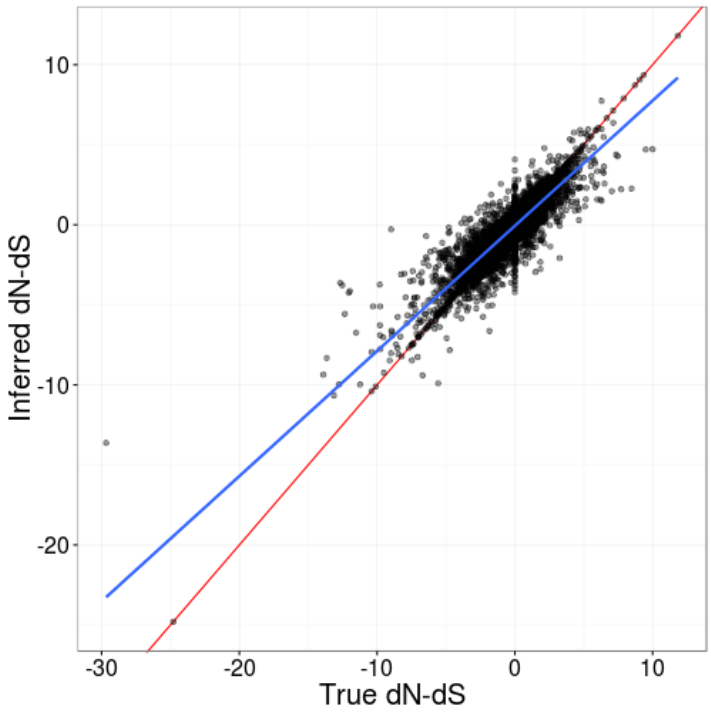


Figure A.12: Cleaned Umberjack estimate of site $dN - dS$ vs true $dN - dS$. Each point represents a site estimate of $dN - dS$. Sites for which the Random Forest predictor deemed the Umberjack estimate inaccurate are excluded. Red line is y=x. Blue line is fitted line.

A.3 HIV Selective Sweeps in Untreated Patients

A.4 Effect of Recombination on Umberjack Error

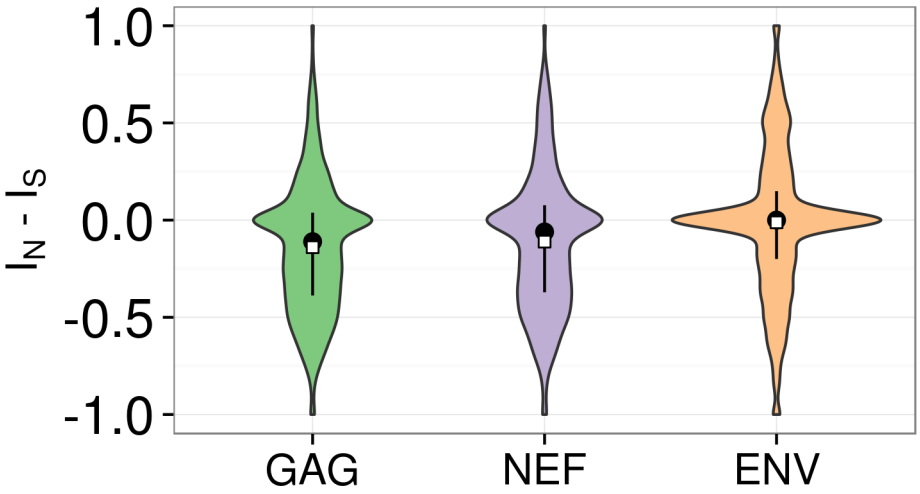
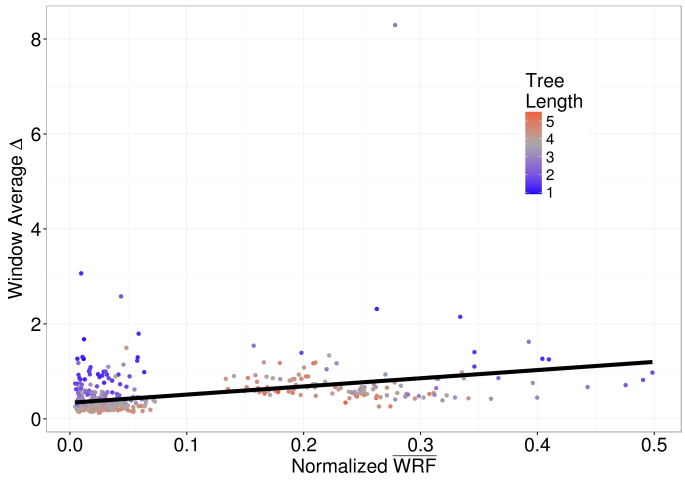
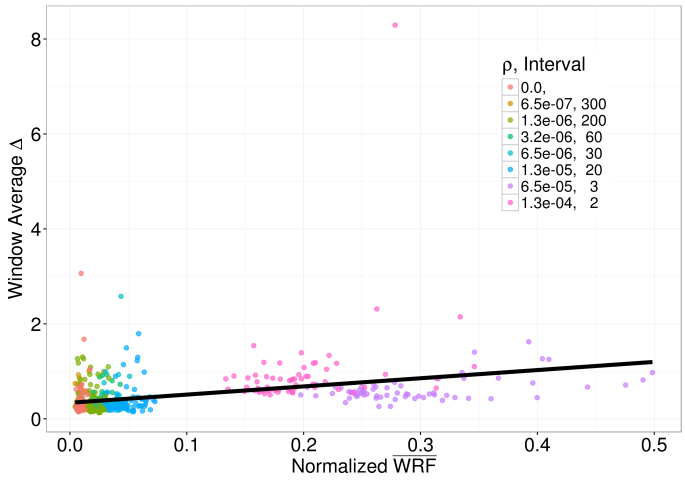


Figure A.13: Violin plots summarizing the distributions of site $I_N - I_S$ statistic across subjects, broken down by gene. A $I_N - I_S < 0$ indicates that nonsynonymous substitutions occur later than synonymous substitutions, which is consistent with purifying selection. Black point ranges indicate median and IQR. White squares indicate mean.



(a)



(b)

Figure A.14: Effect of recombination on Δ. Each point represents a window of Umberjack estimates on simulated datasets focused on recombination (Table A.3). Black line indicates linear regression fit. **(a)** Points are shaded according the the length of the window tree in units of nucleotide substitutions/site. **(b)** Points are shaded according to the recombination rate ρ of population in units of recombinations/site/generation. Interval denotes average codon distance between breakpoints across the genome.