HIV Evolution

Biogeography and Intra-Individual Dynamics

by

Crystal Marie Hepp

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved May 2013 by the
Graduate Supervisory Committee:

Michael Rosenberg, Chair
Philip Hedrick
Ananias Escalante
Sudhir Kumar

ARIZONA STATE UNIVERSITY

August 2013

ABSTRACT

The entire history of HIV-1 is hidden in its ten thousand bases, where information regarding its evolutionary traversal through the human population can only be unlocked with fine-scale sequence analysis. Measureable footprints of mutation and recombination have imparted upon us a wealth of knowledge, from multiple chimpanzee-to-human transmissions to patterns of neutralizing antibody and drug resistance. Extracting maximum understanding from such diverse data can only be accomplished by analyzing the viral population from many angles.

This body of work explores two primary aspects of HIV sequence evolution, point mutation and recombination, through cross-sectional (inter-individual) and longitudinal (intra-individual) investigations, respectively.

Cross-sectional Analysis:

The role of Haiti in the subtype B pandemic has been hotly debated for years; while there have been many studies, up to this point, no one has incorporated the well-known mechanism of retroviral recombination into their biological model. Prior to the use of recombination detection, multiple analyses produced trees where subtype B appears to have first entered Haiti, followed by a jump into the rest of the world. The results presented here contest the Haiti-first theory of the pandemic and instead suggest simultaneous entries of subtype B into Haiti and the rest of the world.

Longitudinal Analysis:

Potential N-linked glycosylation sites (PNGS) are the most evolutionarily dynamic component of one of the most evolutionarily dynamic proteins known to date. While the number of mutations associated with the increase or decrease of PNGS frequency over time is high, there are a set of relatively stable sites that persist within and between longitudinally sampled individuals. Here, I identify the most conserved stable PNGSs and suggest their potential roles in host-virus interplay. In addition, I have identified, for the first time, what may be a gp-120-based environmental preference for N-linked glycosylation sites.

DEDICATION

I dedicate this work to my sweet and beautiful daughter, Adyley Marie Gordon, for motivating me to finish my dissertation in a more timely manner so our family could continue to move forward. I thank her for reminding me to take time to enjoy little things, like watching a baby girl learn to walk. I hope this, in turn, motivates you to go out and accomplish all the things you set your heart and mind to.

In addition to the dedication, I have many friends and family members that I would like to thank. The long hours spent in the lab and on campus would not have been possible if it hadn't been for coffee and conversation breaks with Janine Quijano, Brooke Hjelm, Joanna Malukiewicz, Dr. Charlotte Konikoff and other classmates. I thank my parents, Donna and Jerry Hepp, for always making sure I had everything I ever needed and that I knew that I could achieve anything I ever wanted. My loving boyfriend, Nathaniel Gordon, has been an unwaivering partner throughout this jourey, despite many nights of putting our daughter to bed before I got home from the lab. I additionally thank his family for supporting both of us emotionally since I met them nearly five years ago. I thank the late and lovely Jax Brown for being the greatest companion a girl could ever have wanted; I really loved that dog. Lana and Oatmeal, Jax Brown's loving successors, have been essential in distracting me every once in a while with playtime, and I am grateful for it. I thank all of my other friends and family for playing a special part in my life at some time or another. Finally, I thank all of the HIV-infected individuals, and their family members, that have donated their time and samples to the progress of health and science. Thank you.

ACKNOWELEDGEMENTS

I greatly appreciate the administrative, financial and other forms of support I have received during the pursuit of my Ph.D. degree at Arizona State University. First and foremost, I thank my Ph.D. advisor, Dr. Michael Rosenberg for giving me the opportunity to learn from and work with someone so knoweledgeable about many different facets of evolutionary biology. I am also thankful that he fostered such an independent spirit of research in me, which I will apply to all future scientific endeavors. I also sincerely thank my other committee members, Drs. Ananias Escalante, Philip Hedrick and Suhdir Kumar, who were involved with much of this research, either by offerring helpful suggestions or through general discussion. Drs. Anne Stone, Jay Taylor, Marty Wojciechowski, Fabia Battistuzzi, Kristian Schneider, Nevin Gerek and other experts within the School of Life Sciences additionally provided me opportunities to learn from their scientific expertise. Finally, I am not sure if I would have ever been interested in evolutionary biology had it not been for Dr. Marcie McClure, at Montana State University, for taking me into her lab years ago as a work study undergraduate research assistant and I will be forever grateful. Thank you.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INSIGHTS INTO HIV EVOLUTION: AN INTRODUCTION

Historical Background

Recognition of an epidemic

More than three decades ago, the Centers for Disease Control and Prevention (CDC) started publishing reports of *Pneumocystis carinii*, a rare opportunistic fungus, in men who have sex with men (MSM) living in Los Angeles (CDC 1981c). Shortly thereafter, reports of a rare cancer, Kaposi's sarcoma (KS), emerged and revealed that young MSM on both the east and west coasts of the United States were afflicted (CDC 1981b). Two months later, in August of 1981, the total number of individuals diagnosed with *P. carinii* and/or KS had jumped to 111, now including heterosexual men and women as well as MSM (CDC 1981a). The following year, reports of rare opportunistic infections and KS in a total of 34 Haitians residing in Florida, New York, California, Georgia and New Jersey were released (CDC 1982b). Five days later, the CDC reported that three non-IV drug abusing heterosexual males with hemophilia A were infected with *P. carinii* (CDC 1982a). Within a very short period of time, it was established that the etiologic agent resulting in acquired immunodeficiency syndrome (AIDS) could be transmitted sexually or though transfusion of blood-borne products. From these early reports, four high risk groups were identified: Homosexuals, Heroine addicts, Hemophiliacs and Haitians; collectively given the moniker, the "4 H's" (Gallo 2006). As the validated number of infections grew, borders between the 4 H's and the rest of the population diminished and it became evident that anyone could become infected. De Cock et al. (2011) referred to

these initial reports as, “sentinels for what became one of history’s worst pandemics, with >60 million infections, 30 million deaths, and no end in sight.”

The Etiologic Agent

In 1983, it was determined that the etiologic agent causing AIDS was a retrovirus (Barre-Sinoussi et al. 1983), later named the Human Immunodeficiency Virus (Coffin et al. 1986). Shortly after, a cell line was discovered that could be infected by and continually produce high quantities of the retrovirus for molecular characterization and detection purposes (Popovic et al. 1984). These findings and others that followed were referred to as the period of “Intense Discovery” within the rich history of AIDS (Gallo 2006).

Molecular Characteristics of HIV

In early 1985, the first full-length sequence of the Human Immunodeficiency Virus was published with six fold coverage (Wain-Hobson et al. 1985). It was revealed that the retrovirus was nearly 10,000 bases in length and did include the expected polycistronic retroviral core genes: *gag*, *pol* and *env* (Fig. 1A). Specifically, after the Gag-Pol polyprotein is cleaved, Gag is further broken down into the structural matrix (MA), capsid (CA), nucleocapsid (NC) and p6 proteins. The proteins resulting from the proteolytic cleavage of the enzymatic Pol protein are protease (PR), reverse transcriptase (RT, this includes the ribonuclease H or RH region) and integrase (INT). Cleavage of the Env protein results in two structural proteins; the surface protein (gp120) and the transmembrane protein (gp41). Interestingly, Wain-Hobson et al. (1985) described two additional open reading frames, originally referred to as Q and F, which clearly set HIV apart from other identified retroviruses. Later, it was determined that HIV actually had an additional six accessory genes, bringing the total gene count to nine and the protein count to 15 (Frankel, Young 1998).

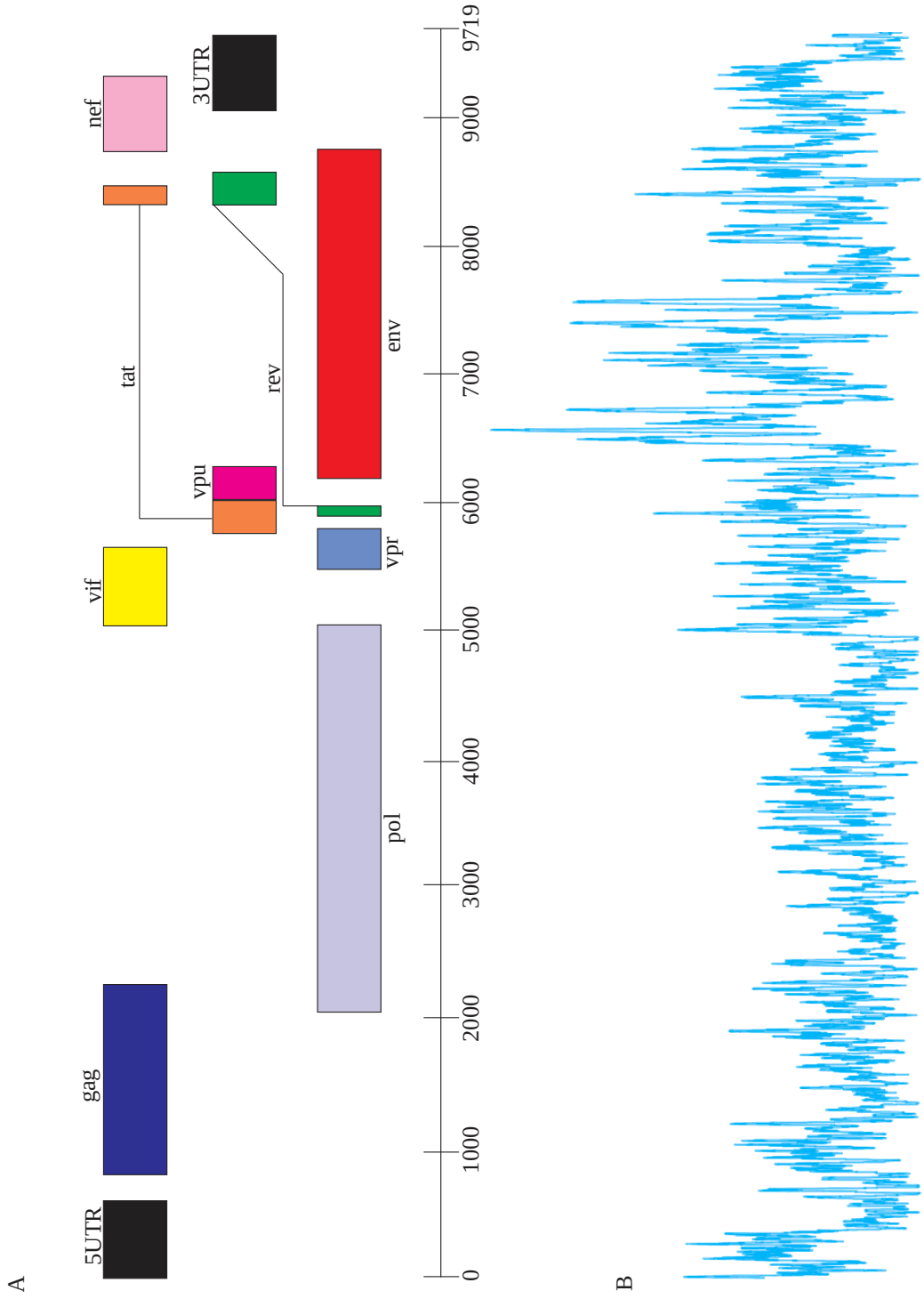


Figure 1. (A) HIV-1 HXB2 genome and (B) the relative evolutionary rate for each position. HXB2 is the HIV-1 subtype B reference sequence in the Los Alamos National Laboratories HIV Database (accession K03455).

The initiation of the HIV lifecycle occurs with the binding of the viral gp120 with the host T-cell CD4 receptor and is quickly followed by that same gp120 also binding to either the host CCR5 or CXCR4 co-receptor (Figs. 2 and 3, step 1) (Frankel, Young 1998; Anastassopoulou 2012). This binding triggers a conformational change in gp41, promoting fusion of the virus and host cell, and subsequent adsorption of the viral core.

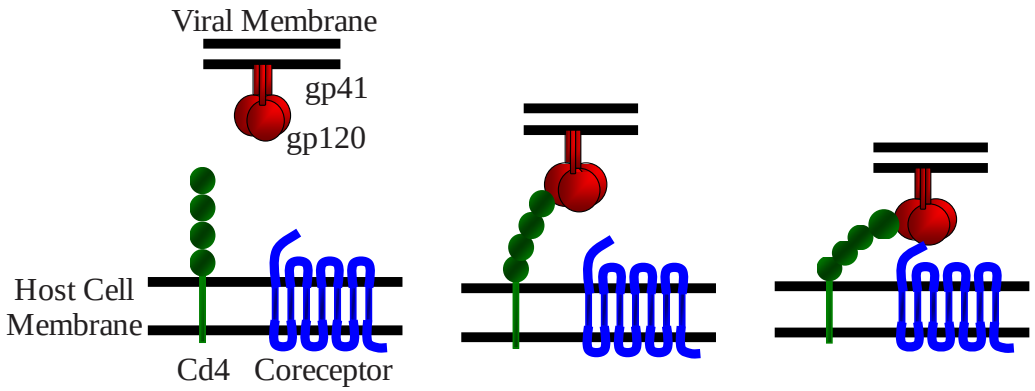


Figure 2. gp120 binding to CD4 and then the coreceptor; either CCR5 or CXCR4. Adapted from Figure 1 in Anastassopoulou (2012).

Next the retroviral RNA is reverse transcribed into cDNA (Fig. 3, step 2). (See below for a detailed explanation of reverse transcription). It should be noted that coat is removed from the viral core at some point between entry into the cellular cytoplasm and reaching the nucleus, although the timing of uncoating is conjectural (Arhel 2010). The new double stranded viral cDNA is transferred to the nucleus and will gain entry to the organelle through a nuclear pore (Fig. 3, step 3). Integration of the cDNA into the host chromosome is catalyzed by the INT protein (Fig 3, step 4). Next, host RNA polymerase transcribes the viral DNA into mRNA transcripts, including full-lengths and spliced forms, which are transported out of the nucleus and into the cytoplasm (Fig. 3, step 5). There, the transcripts are either packaged (full-length) or translated into viral protein products (Fig. 3, step 6). Specifically, the Gag and Gag-Pol polyproteins are localized to

the cellular membrane where assembly of the viral core takes place, while Env is localized to the endoplasmic reticulum to undergo post-translational modifications (not shown). Mature Env proteins are then translocated to the cell surface where the core particle buds, coated in gp41/gp120 complexes. The final step in this process occurs during or right after budding, when the Gag-Pol polyproteins are proteolytically cleaved, resulting in a mature virus (Fig. 3, step 8) (Frankel, Young 1998).

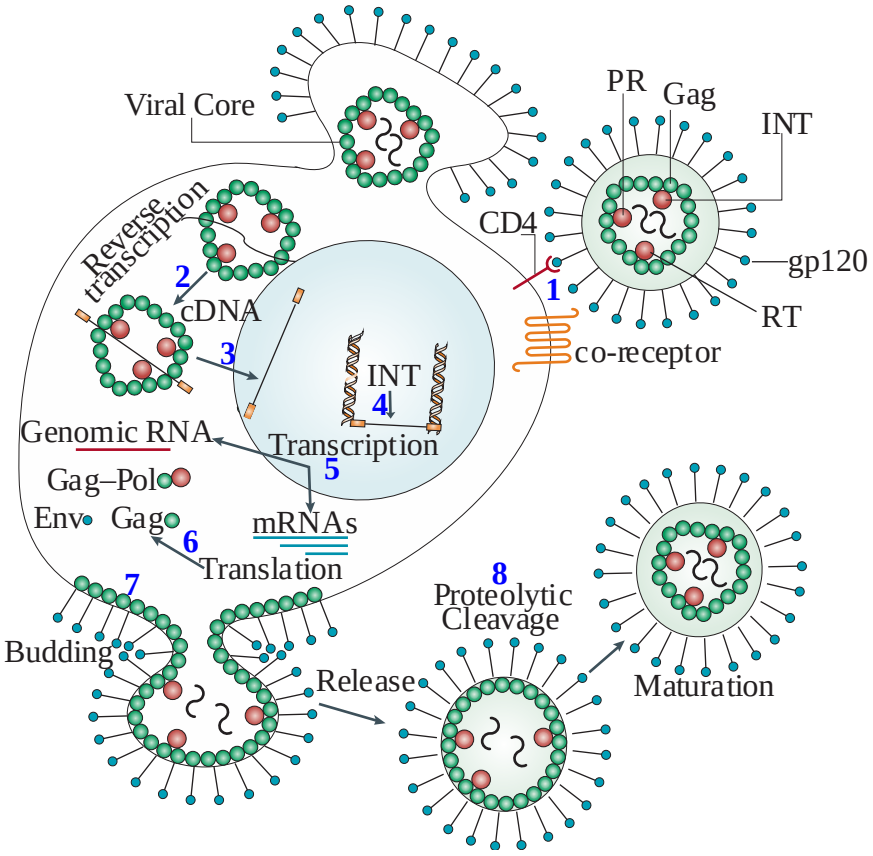


Figure 3.HIV lifecycle. Each step, as is described in the text, is numbered in blue. Adapted from Monini et al. (2004).

Mechanisms of Variability

Perhaps the most interesting step in the HIV lifecycle is that of reverse transcription, not because it completely disregards the established central dogma of transcription, but

because it generates immense diversity through its propensity for inaccuracy. As mentioned above, reverse transcription takes place shortly after entry of the viral core into the cellular cytoplasm (Fig. 3, step 2). The first step of reverse transcription requires that a tRNA is bound to the primer binding site (PBS), to act as a primer for RT (Fig. 4, step 1). RT synthesizes negative strand DNA through the 5' end of the viral RNA and RH degrades the RNA portion of the RNA:DNA duplex. The resulting DNA is referred to as minus-strand strong-stop DNA (-sssDNA) (Fig. 4, step 2). The now free –sssDNA undergoes first strand transfer, where the repeat region (R) is annealed to the complimentary 3'R of the viral RNA (Fig. 4, step 3). Reverse transcription of the negative strand can reinitiate, while RH digests all but the polypurine tract (ppt) of the viral RNA strand (Fig. 4, step 4), which acts as a primer for positive strand synthesis. Positive strand synthesis, which uses the negative stranded DNA as a template, continues through a portion of the tRNA primer. Once the tRNA is in a partial duplex with DNA, RH can degrade the tRNA (Fig. 4, step 5). The partial positive strand DNA is referred to as plus-strand strong-strop DNA (+sssDNA). The second strand transfer occurs when the +sssDNA is transferred such that the PBS regions on the positive and negative strands are annealed (Fig. 4, step 6). RT completes the synthesis of the positive strand, starting at the PBS through the 3' end, yielding double stranded cDNA (Fig. 4, step 7) (Coffin, Hughes, Varmus 1997; Hu, Hughes 2012).

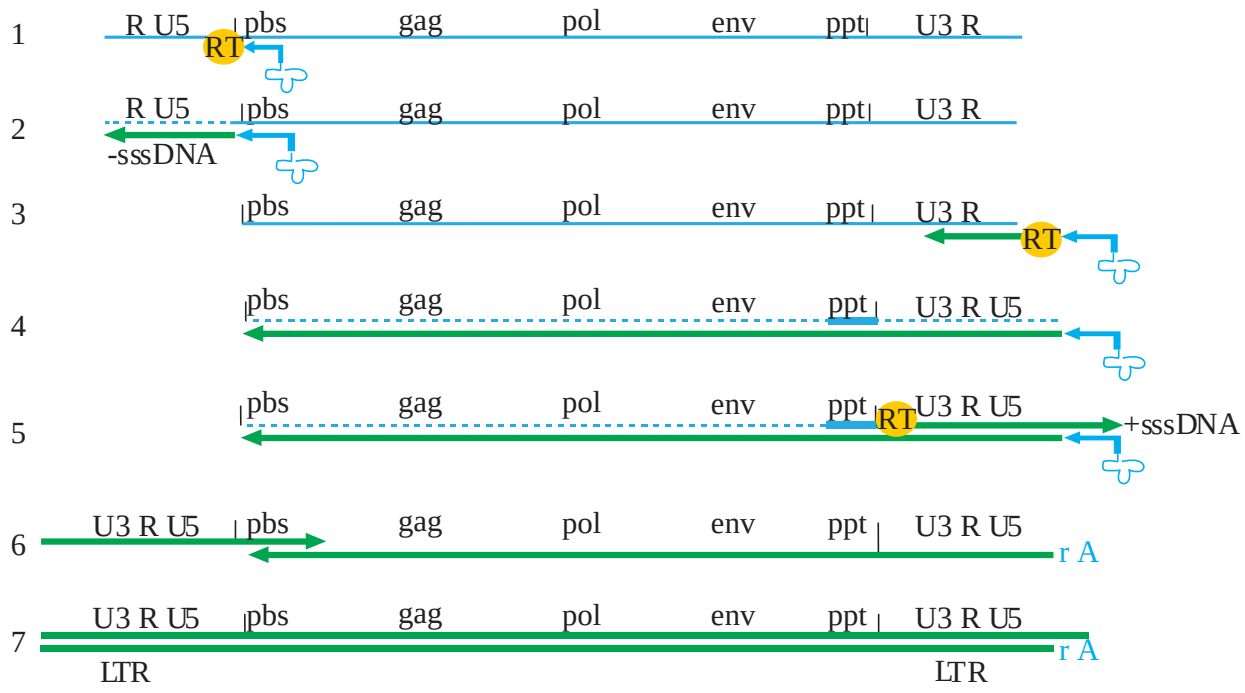


Figure 4. Process of reverse transcription. Blue represents RNA while green represents DNA. The gold circle labeled RT represents the RT molecule performing reverse transcription. All steps are described in the text. This figure was conceptually modified from multiple sources (Coffin, Hughes, Varmus 1997; Hu, Hughes 2012).

Reverse transcriptase is notorious for its tendency to generate mutation throughout the described conversion of RNA to DNA (Fig. 4), with an estimated mutation rate between 10^{-5} and 10^{-4} (Coffin 1995), leading to approximately one mutation per generation. These errors generally take one of two forms: base substitutions and frameshift mutations. Both can be the result of insertions or deletions, while misincorporation can also cause base substitutions. One well-known cause of these errors is slippage of the two DNA strands in a repetitive region (e.g. Fig. 4, step 5, R region). After slippage and the addition the next base, if the strands are able to correctly realign, the slippage will result in misincorporation and a base substitution. Alternatively, if the strands never correctly realign, the result is either a single insertion in the primer strand or deletion in the template strand, leading to a frameshift (Bebenek, Kunkel 1993).

In addition to the low fidelity of reverse transcriptase, recombination is a second mechanism that can generate variability. Retroviral recombination occurs between two strands of viral genomic RNA and occurs at a rate of approximately 2.4×10^{-4} / bp / generation, equivalent to 2.4 crosses per round of replication (Jetzt et al. 2000; Rhodes, Wargo, Hu 2003; Lanciault, Champoux 2006). While recombination likely occurs between identical strands, the process is only detectable between variant strands of RNA. Therefore, detectable recombination is initiated when two variant viruses either simultaneously or sequentially infect a single cell (Fig. 5). Both undergo the same processes explained in Fig. 3, but after each variant has undergone transcription (Fig. 3, step 5), it is possible, through chance alone, that a genomic RNA from each will be co-packaged. When the bi-variant virion infects a new cell, and reverse transcription takes place, it is possible that during strand switching of reverse transcription (Fig. 4, steps 3

and 6), -sssDNA or +sssDNA will anneal to the RNA strand that it was not originally reverse transcribed from or the DNA it should be paired with, respectively (Goodrich, Duesberg 1990; Hu, Temin 1990). Recombination can have an additive fitness effect, in that it can combine adaptive mutations, or the reverse, where combining mutations can prove to lower fitness (Fisher 1930; Muller 1932). Additionally, many mutations are likely to be neutral, their combination having no effect on fitness.

The errors made during reverse transcription, whether mutation or recombination, would not be able to generate nearly as much population variability if it was not for the rapid turnover of viral replication. The generation time of a single virus is 2.6 days on average, from release of a virus to the infection of a new cell through the release of new viral progeny (Perelson et al. 1996). That translates to over 1,400 generations within a person infected for 10 years. Even more astonishing is that the estimated total daily production of 10.3×10^{9} virions per day (Perelson et al. 1996), coupled with the mutation rate, presents the opportunity for every possible single point mutation to occur 10,000 to 100,000 times per day (Coffin 1995), making HIV one of the most variable genomes described to date.

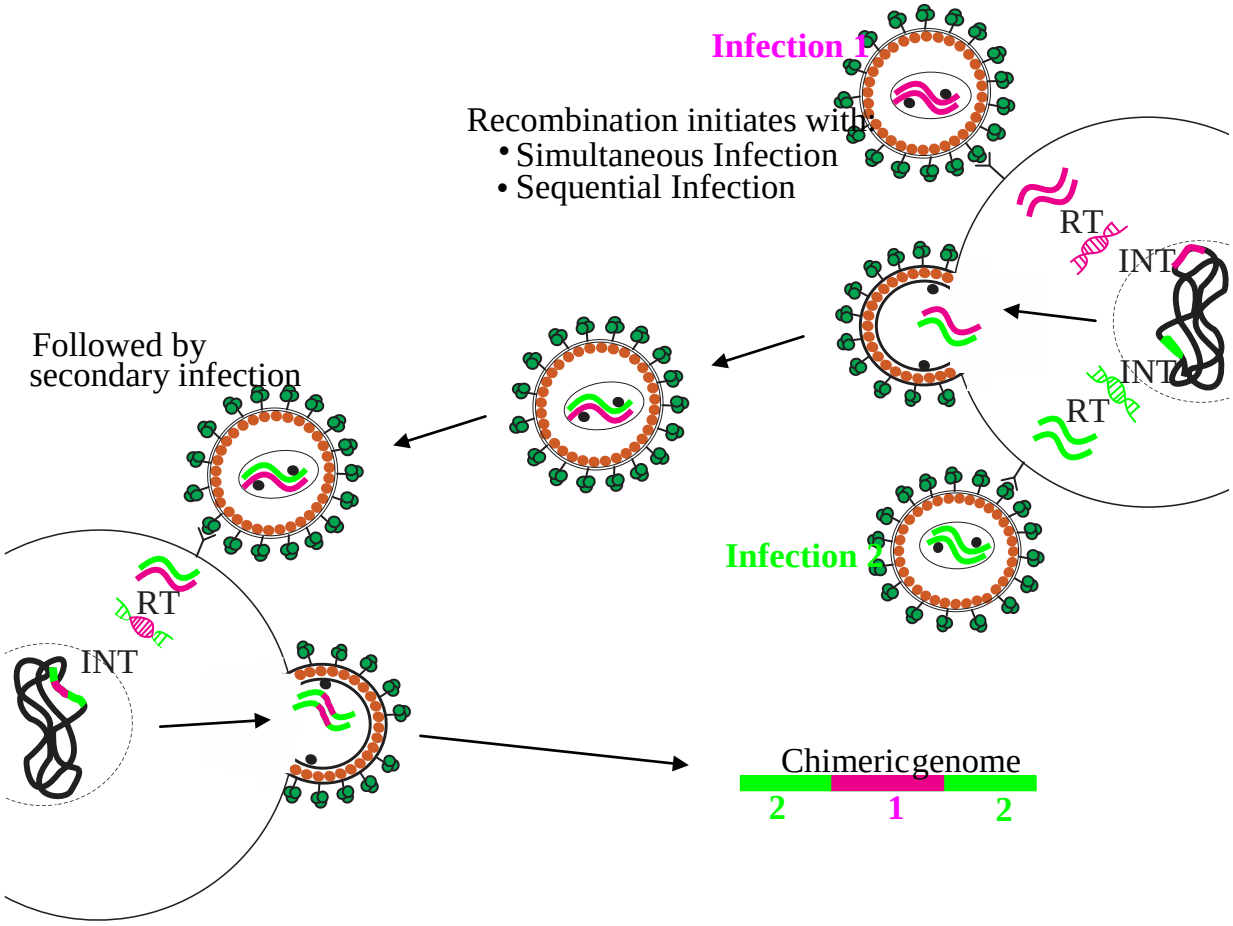


Figure 5. HIV Recombination.

While RT can produce errors anywhere along the genome, detectable errors, which are tolerated such that the virus is able to undergo the entire replication cycle, are highly concentrated in the *env* gene (Fig. 1B). This viral diversity was first noted (Hahn et al. 1985) immediately after the first research groups started sequencing whole HIV genomes from different isolates (Hahn et al. 1985; Ratner, Gallo, Wong-Staal 1985; Ratner et al. 1985; Wong-Staal et al. 1985). Further analyses revealed that not only is there substantial variation between viral isolates sampled from different individuals, but also within a single individual. Initially, comparisons were made using differences among restriction digestion patterns from different clones isolated from the same individual. These patterns differed by 11-44%, resulting in an estimated nucleotide sequence difference of 2-7%. Moreover, *in vitro* isolation and amplification did not yield the same type or number of changes that were seen from patient isolates, indicating that *in vivo* variation is generated rapidly (Saag et al. 1988), likely due to selective pressure exerted by the host immune system on the virus.

Intra-individual evolution

Within a single individual, the HIV population experiences a wide range of selective pressures elicited by both the cellular and humoral immune reponses. Given that the host immune system retains a high level of plasticity at any given time point, the HIV quasispecies undergoes strong episodes of diversifying selection (Lemey, Rambaut, Pybus 2006). Diversifying, or disruptive, selection can be classified as the favoring of several variants at any given moment in time. As a result, given that selection acts on the phenotype rather than the genotype, effective adjustments made during diversifying

selection must result in a “discontinuity (of the current phenotype) at the phenotypic level” (Mather 1955). This discontinuity can easily be observed on a phylogenetic tree, as a ladder-like topology, reconstructed from sequences collected longitudinally from a single HIV-infected individual (Fig. 6). It is readily apparent that the sequences from any single time point have limited diversity, but there is substantial diversity generated between different time points upon branch length comparison (Grenfell et al. 2004).

Concentrated examinations of molecular changes have revealed a plethora of surprising viral evolutionary dynamics within a single individual. For example, normally progressing HIV-infected individuals typically mount a partially robust neutralizing antibody response to a susceptible viral population. After a short period of time, the susceptible variants are replaced by neutralization-resistant variants. While one might expect that mutations would occur at the antibody binding sites on the virus, mutations have actually been found to be sparsely distributed across the *env* gene, and not at binding sites. Rather, mutations more often occur at potential N-linked glycosylation sites (PNGS), which act as a carbohydrate based shield (Wei et al. 2003). PNGSs can also protect the virus against therapeutics directed towards conserved regions of the viral surface.

Valuable information can be gained by comparing differences in mutation, diversity and other characteristics among multiple longitudinally sampled individuals. While HIV progression cannot be exclusively characterized into a few all-encompassing profiles, there are clearly individuals that progress much more rapidly or slowly than others. Comparison of the viral evolutionary dynamics within individuals at different

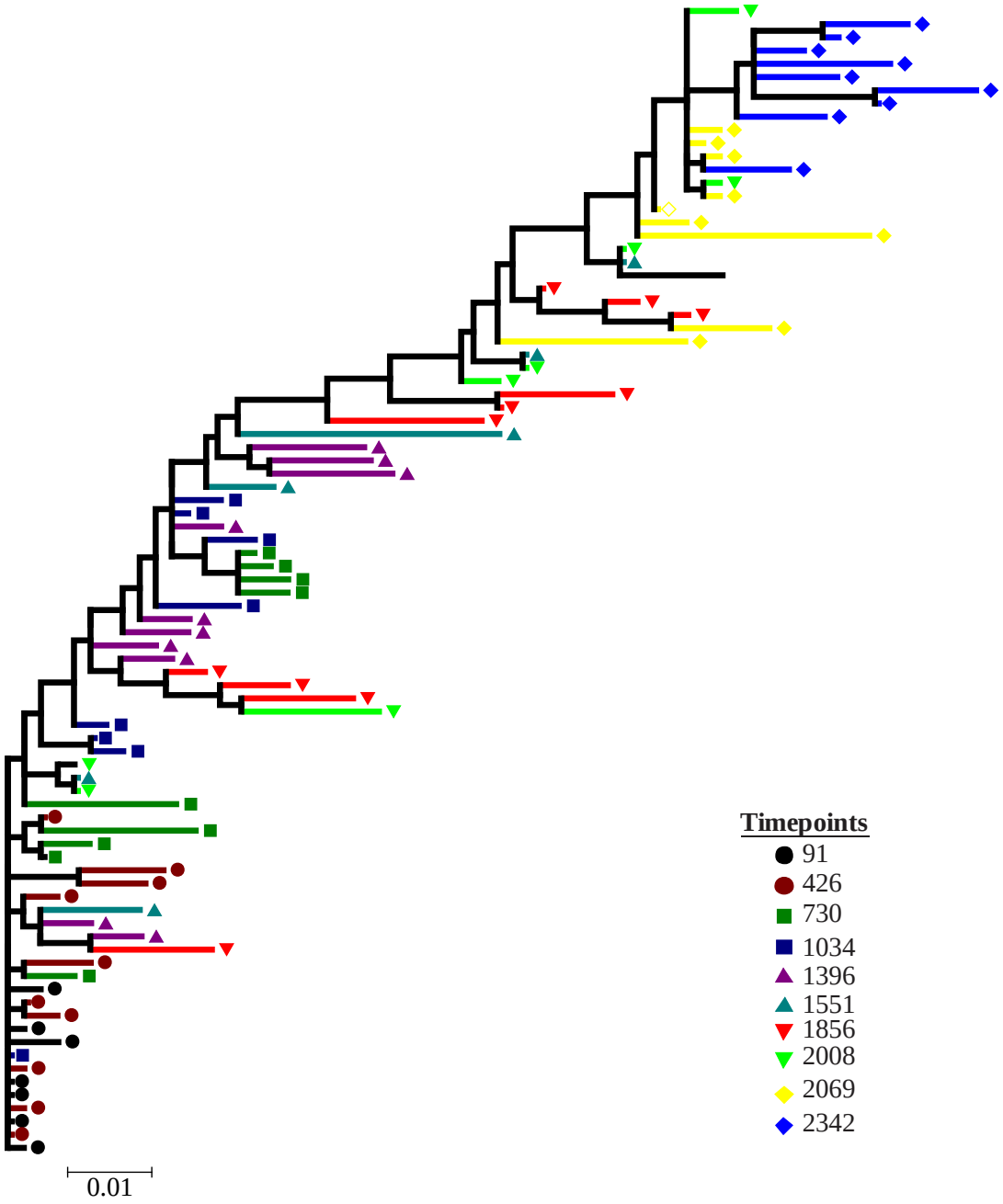


Figure 6. Maximum parsimony tree reconstructed from longitudinally collected sequences. This individual (P1) was originally described in Shankarappa et al. (1999) and corresponding sequences were used as part of the Chapter 3 analyses (Individual 1).

points of the progression spectrum has revealed some surprising discoveries. In a study where samples were taken from normally progressing individuals over a span of 6-12 years, there were consistent shifts within individuals of viral divergence, diversity, and quasispecies towards variants demonstrating a co-receptor switch (Shankarappa et al. 1999). Alternatively, perinatally infected children who experienced different rates of disease progression harbored viral populations with different levels of genetic diversity. Specifically, children with lower viral loads that experience slower progression to AIDS carried a more diverse viral population over time than their more rapidly progressing counterparts. Longer branch lengths on phylogenetic trees were correlated with a greater accrual of non-synonymous substitutions, indicating stronger selective pressures in the slower progressing children (Ganeshan et al. 1997). This finding was also upheld when six HIV infected men were compared over time, where increased diversity of the viral population could be correlated to prolonged survival (Wolinsky et al. 1996).

In the third chapter of this dissertation, I have compared longitudinally sampled sequences within and among HIV-1 subtype B infected individuals. By analyzing sequences within each individual, I was able to identify temporally-consistent potential N-linked glycosylation sites (PNGS) and quantify their impact on divergence of the entire viral population within an individual. Inter-individual comparisons revealed that the stably encoded PNGSs were consistent across infected individuals, implying that these sites are functionally important over the entirety of infection. By mapping the proximity of PNGSs to critical binding sites on the gp120 portion of *env*, I have suggested possible roles that PNGSs play in infection both individually and collectively.

Inter-individual evolution

Aside from longitudinally sampled inter-individual comparisons, a wealth of knowledge has accrued based on single HIV sequences collected from globally distributed individuals. Phylogenetic trees reconstructed from multiple individuals, even over different timepoints, do not demonstrate the ladder-like topology seen with intra-individual trees (Lemey, Rambaut, Pybus 2006). Instead, inter-host phylogenies show that multiple lineages display temporal endurance. Taxa comparison on a tree is less likely to represent host-specific selective regimes; rather it tells a story about the demographic and spatial history of the virus (Grenfell et al. 2004; Lemey, Rambaut, Pybus 2006).

The first network of HIV infected individuals was not drawn based on genetic information, but rather by connecting individuals through sexual contact. The main finding, published shortly after reports of AIDS being caused by a retrovirus, was that forty of the first reported AIDS cases in ten cities were linked by sexual contact. This was consistent with the hypothesis that AIDS could be caused by a sexually transmissible infectious agent (Auerbach et al. 1984).Transmission networks, where the true contacts are known, are rarely encountered. Phylogenetics is a way around this problem, where the evolutionary history can be used to infer transmission networks. A study analyzing sequences sampled from 2,126 London-based HIV-infected individuals, primarily MSM, found six strongly supported MSM clusters using Bayesian Monte Carlo Markov chain phylogenetic analysis (Lewis et al. 2008). They found that 25% of new infections occurred within 6 months of the donor's infection, and that 65% of new infections took place between 1995 and 2000. This is important from an intervention standpoint, as

education about transmissibility early after infection can be given to at-risk groups. Furthermore, they found that in no case did nearest neighbor pairs have the same drug resistance mutations, indicating that drug resistant strains were not being transmitted.

Biogeographic studies of HIV sequences can give insight into the deep history of the virus. Until recently, there was only a single HIV-1 sequence dated prior to 1976. Worobey et al. (2008) obtained tissue blocks dating between 1958 and 1960 from Kinshasa, where the HIV-1 epidemic is thought to have originated, and screened the tissues for HIV-1 RNA. They found a single positive specimen (DRC60), which they use to reconstruct a phylogeny, along with a sequence from the same region dating back to 1959 (ZR59), and 156 other HIV-1 Group M sequences. This revealed that Group M, the most successful HIV-1 group, dated back much later than previously estimated; between 1884 and 1924. Additionally, the fact that DRC60 and ZR59 clustered with different subtypes indicates HIV-1 Group M was already extensively variable in the 1950s (Worobey et al. 2008).

In the next chapter, I examine the relationship between globally distributed HIV-1 subtype B sequences to further ascertain Haiti's role in the subtype B pandemic. Groups had previously examined the relationship between Haitian-derived sequences and those from the rest of the world (Li, Tanimura, Sharp 1988; Gojobori et al. 1990; Korber et al. 2000; Gilbert et al. 2007) and found the global strains to be nested within Haitian-derived strains. I have repeated this analysis, and for the first time incorporated the evolutionary force of recombination detection into the evolutionary model. While the identification of risk groups is not an outcome of this study, the results contribute, for the first time, an alternate possibility to the subtype B pandemic.

References

Anastassopoulou, CG. 2012. Chemokine Receptors as Therapeutic Targets in HIV Infection. In: M K, editor. Immunodeficiency: InTech.

Arhel, N. 2010. Revisiting HIV-1 uncoating. Retrovirology 7:96.

Auerbach, DM, WW Darrow, HW Jaffe, JW Curran. 1984. Cluster of cases of the acquired immune deficiency syndrome. Patients linked by sexual contact. Am J Med 76:487-492.

Barre-Sinoussi, F, JC Chermann, F Rey, et al. 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220:868-871.

Bebenek, K, T Kunkel. 1993. Reverse Transcriptase Fidelity. In: A Skalka, S Goff, editors. Reverse Transcriptase. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. p. 85-102.

CDC. 1981a. Follow-up on Kaposi's Sarcoma and Pneumocystis Pneumonia. Morbidity and Mortality Weekly Report 30:409-410.

CDC. 1981b. Kaposi's sarcoma and Pneumocystis Pneumonia among homosexual men - New York City and California. Morbidity and Mortality Weekly Report 30:305-308.

CDC. 1981c. Pneumocystis pneumonia - Los Angeles. Morbidity and Mortality Weekly Report 30:250-252.

CDC. 1982a. Epidemiologic Notes and Reports Pneumocystis carinii Pneumonai among Persons with Hemophilia A. Morbidity and Mortality Weekly Report 31:365-367.

CDC. 1982b. Opportunistic Infections and Kaposi's Sarcoma among Haitians in the United States. Morbidity and Mortality Weekly Report 31:353-354, 360-361.

Coffin, J, A Haase, JA Levy, et al. 1986. What to call the AIDS virus? Nature 321:10.

Coffin, J, S Hughes, H Varmus. 1997. Overview of Reverse Transcription. In: J Coffin, S Hughes, H Varmus, editors. Retroviruses. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Coffin, JM. 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science 267:483-489.

Fisher, R. 1930. The Genetical Theory of Natural Selection. Dover, New York.

Frankel, AD, JA Young. 1998. HIV-1: fifteen proteins and an RNA. Annu Rev Biochem 67:1-25.

Gallo, RC. 2006. A reflection on HIV/AIDS research after 25 years. Retrovirology 3:72.

Ganeshan, S, RE Dickover, BT Korber, YJ Bryson, SM Wolinsky. 1997. Human immunodeficiency virus type 1 genetic evolution in children with different rates of development of disease. J Virol 71:663-677.

Gilbert, MT, A Rambaut, G Wlasiuk, TJ Spira, AE Pitchenik, M Worobey. 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A 104:18566-18570.

Gojobori, T, EN Moriyama, Y Ina, K Ikeo, T Miura, H Tsujimoto, M Hayami, S Yokoyama. 1990. Evolutionary origin of human and simian immunodeficiency viruses. Proc Natl Acad Sci U S A 87:4108-4111.

Goodrich, DW, PH Duesberg. 1990. Retroviral recombination during reverse transcription. Proc Natl Acad Sci U S A 87:2052-2056.

Grenfell, BT, OG Pybus, JR Gog, JL Wood, JM Daly, JA Mumford, EC Holmes. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327-332.

Hahn, BH, MA Gonda, GM Shaw, M Popovic, JA Hoxie, RC Gallo, F Wong-Staal. 1985. Genomic diversity of the acquired immune deficiency syndrome virus HTLV-III: different viruses exhibit greatest divergence in their envelope genes. Proc Natl Acad Sci U S A 82:4813-4817.

Hu, WS, SH Hughes. 2012. HIV-1 reverse transcription. Cold Spring Harb Perspect Med 2.

Hu, WS, HM Temin. 1990. Retroviral recombination and reverse transcription. Science 250:1227-1233.

Jetzt, AE, H Yu, GJ Klarmann, Y Ron, BD Preston, JP Dougherty. 2000. High rate of recombination throughout the human immunodeficiency virus type 1 genome. J Virol 74:1234-1240.

Korber, B, M Muldoon, J Theiler, F Gao, R Gupta, A Lapedes, BH Hahn, S Wolinsky, T Bhattacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789-1796.

Lanciault, C, JJ Champoux. 2006. Pausing during reverse transcription increases the rate of retroviral recombination. J Virol 80:2483-2494.

Lemey, P, A Rambaut, OG Pybus. 2006. HIV evolutionary dynamics within and among hosts. AIDS Rev 8:125-140.

Lewis, F, GJ Hughes, A Rambaut, A Pozniak, AJ Leigh Brown. 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med 5:e50.

Li, WH, M Tanimura, PM Sharp. 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 5:313-330.

Mather, K. 1955. Polymorphism as an Outcome of Disruptive Selection. Evolution 9:52-61.

Monini, P, C Sgadari, E Toschi, G Barillari, B Ensoli. 2004. Antitumour effects of antiretroviral therapy. Nat Rev Cancer 4:861-875.

Muller, HJ. 1932. Some genetic aspects of sex. American Naturalist 66:118-138.

Perelson, AS, AU Neumann, M Markowitz, JM Leonard, DD Ho. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582-1586.

Popovic, M, MG Sarngadharan, E Read, RC Gallo. 1984. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science 224:497-500.

Ratner, L, RC Gallo, F Wong-Staal. 1985. HTLV-III, LAV, ARV are variants of same AIDS virus. Nature 313:636-637.

Ratner, L, W Haseltine, R Patarca, et al. 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature 313:277-284.

Rhodes, T, H Wargo, WS Hu. 2003. High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication. J Virol 77:11193-11200.

Saag, MS, BH Hahn, J Gibbons, Y Li, ES Parks, WP Parks, GM Shaw. 1988. Extensive variation of human immunodeficiency virus type-1 in vivo. Nature 334:440-444.

Shankarappa, R, JB Margolick, SJ Gange, et al. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73:10489-10502.

Wain-Hobson, S, P Sonigo, O Danos, S Cole, M Alizon. 1985. Nucleotide sequence of the AIDS virus, LAV. Cell 40:9-17.

Wei, X, JM Decker, S Wang, et al. 2003. Antibody neutralization and escape by HIV-1. Nature 422:307-312.

Wolinsky, SM, BT Korber, AU Neumann, M Daniels, KJ Kunstman, AJ Whetsell, MR Furtado, Y Cao, DD Ho, JT Safrit. 1996. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science 272:537-542.

Wong-Staal, F, GM Shaw, BH Hahn, SZ Salahuddin, M Popovic, P Markham, R Redfield, RC Gallo. 1985. Genomic diversity of human T-lymphotropic virus type III (HTLV-III). Science 229:759-762.

Worobey, M, M Gemmel, DE Teuwen, et al. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455:661-664.

CHAPTER 2

RECOMBINATION AFFECTS ORIGIN ESTIMATES: RETHINKING THE HIV-1 SUBTYPE B GLOBAL PANDEMIC

Abstract

Multiple analyses utilizing phylogenetic inference have all but established that Haiti was the final stepping stone for HIV-1 subtype B before the virus achieved pandemic status. A recent study clearly demonstrated a well-supported Haiti-first model, indicating that the worldwide subtype B pandemic is the result of a single migration event out of Haiti. While it is encouraging that multiple studies are in agreement regarding the basal placement of Haiti-originating sequences using a variety of phylogenetic methods, they also collectively share a single major caveat: each assumes a single evolutionary history for the genomic region under scrutiny. Here we show that the majority of positions within the *env* gene support two independently initiated introductions of subtype B epidemics after emergence from Africa; one into Haiti and one into the rest of the world. By using a conservative recombination detection method based on phylogenetic incongruence, we found that multi-partition models fit the *env* dataset better than single-partition models. The paraphyletic and basal clustering of Haiti-originating subtype B sequences was not indicated from the phylogenetic reconstruction of three putative non-recombinant *env* segments. Concatenation of the two longest *env* segments, after putative recombinant removal, also supports two independent introductions of subtype B into Haiti and the rest of the world. Our results not only oppose the Haiti-first model of the most globally widespread HIV-1 subtype, but also indicate that Haiti had little to do with the global pandemic in general.

Introduction

HIV evolution is largely modulated by two dominant forces; mutation and recombination. The high rate of mutation is due to the low-fidelity reverse transcriptase, which allows for rapid expansion across sequence space (Hahn et al. 1986). Fortunately, numerous evolutionary models have been developed and are used in phylogenetic analyses to describe substitution rates and rate variation. Recombination occurs when reverse transcriptase switches RNA strands during cDNA synthesis (Goodrich, Duesberg 1990), combining the genetic information from both RNA strands. While strand switching between two identical strands is likely, recombination is only detectable when variable sites are present between the two strands. State of the art phylogenetic tools implicitly assume a single evolutionary history, and as a result, detection of non-recombinant segments within recombinants must be carried out prior to topological reconstruction.

The earliest phylogenetic trees reconstructed from HIV subtype B sequence data have demonstrated a basal and paraphyletic clustering of Haiti-originating sequences when compared to the rest of the pandemic (Li, Tanimura, Sharp 1988; Gojobori et al. 1990; Korber et al. 2000). The most recent of these studies, using Bayesian methods, corroborated past findings based on a much larger dataset with highly supportive posterior probabilities (Gilbert et al. 2007). The nearly definitive conclusion of research conducted by Gilbert et al. (2007) was that Haiti acted as stepping stone between Africa and the rest of the world. While the agreement between studies using a variety of methods further substantiates the Haiti-first hypothesis of subtype B dispersal, each report overlooks recombination and assumes that all sites share identical ancestry.

The goal of this study is to determine if the incorporation of recombination detection methods in the analysis of the worldwide subtype B pandemic origin has an effect on the basal and paraphyletic clustering of Haiti-originating sequences (Gilbert et al. 2007). It was previously suggested that that intra-subtype recombination could not "plausibly lead to strains from one locality (Haiti-originating) falling basal to all the others (Gilbert et al. 2007)." Admittedly, it does seem improbable that the clear-cut systematic clustering of Haiti-originating taxa, forming a paraphyletic clade basal to all other subtype B sequences, is an artifact of combining multiple independent evolutionary histories. Regardless, given that recombination is a frequently occurring force in HIV that can have extensive effects on topology-based inferences (Robertson, Hahn, Sharp 1995; Schierup, Hein 2000; Posada, Crandall 2002; Woolley, Posada, Crandall 2008; Martin, Lemey, Posada 2011), it is worthwhile to repeat the analysis with the inclusion of recombination detection methods. An analysis of this nature is especially necessary since results from the previous study have not only affected our knowledge of the geographical and temporal spread of HIV-1 subtype B, but also because there are controversial social implications at stake that date back to the beginning of the pandemic (Lambert 1990; Farmer 1992; Carmichael 2007; Pape et al. 2008). Furthermore, the popular press has taken interest in the topic and has popularized the notion of the pandemic subtype B virus coming from Haiti (Bowdler 2007; Carmichael 2007).

Results and Discussion

To address whether or not recombination has an effect on the inferred geographical path of subtype B from a neutral standpoint, we considered four possible outcomes suggested

by Gilbert et al (2007): Haiti-first, Pandemic-first, simultaneous and distinct epidemics, or an unresolved scenario. It should be noted that for either the Haiti or Pandemic-first scenarios that clustering must be both basal and paraphyletic. Basal clustering alone is not sufficient for one clade to be nested within the other (Krell, Cranston 2004; Crisp, Cook 2005).

Reproducing Prior Results

To ensure reproducibility of Gilbert et al. (2007) results, globally distributed datasets spanning the *gag* and *env* regions identical to those used in the prior study were collected (table S1). MrBayes (Huelsenbeck, Ronquist 2001) was used to carry out phylogenetic reconstruction of *gag*1 and *env* datasets prior to recombination detection. The overall topology, indicative of Haiti-originating strains of subtype B falling basal to all others, is replicated here using both *gag1* and *env* datasets (Appendix II, Figs 1A and C). While the average standard deviation of split frequencies did not drop below 0.01 for either dataset (StdDev: *gag1* = 0.03, *env* = 0.19), the log-likelihood values plotted over generation time for each individual run did not follow any specific trend, indicating convergence within chains (Appendix II, Figs. 1B and D). Conversely, the two independent runs of 20 million steps on the *env* dataset failed to converge to the same solution (Appendix II, Fig. 1D), indicating that the number of generations should have been increased or that the phylogeny could not be adequately resolved.

Recombination Breakpoint Detection

In analyzing the initial and reproduced trees, we noticed that a single patient was represented twice in the *gag1* dataset (Appendix II, Fig. 1A, arrows). Both sequences AY247251 and AY268493 are derived from Patient AC_06. We expected that sequences from the same individual should form a monophyletic cluster in the absence of drug resistance mutations, multiple infections or recombination. This finding, coupled with the fact that HIV frequently recombines, prompted the use of the Genetic Algorithms for Recombination Detection (GARD) program in HyPhy (Kosakovsky Pond et al. 2006) to identify recombination breakpoints. This method should be considered conservative in breakpoint assignment, as adjacent partitions must be significantly incongruent, per the Shimodaira-Hasegawa test (Shimodaira, Hasegawa 1999) in order for a breakpoint to be indicated. The *gag1* and *env* datasets each contained two intra-subtype breakpoints, yielding three putatively non-recombinant segments a piece (*S1*, *S2* and *S3*), while the *gag*2 dataset was found to be non-recombinant (Fig. 1 and Appendix I, Tables 2 and 3).

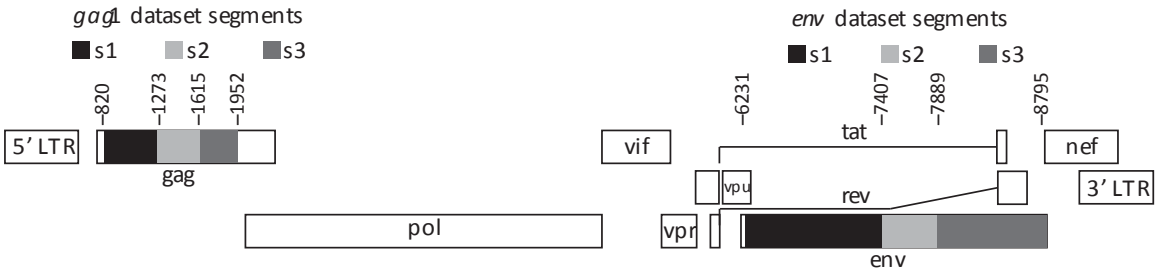


Figure 1. *Recombination breakpoints. LANL HIV-1 Recombinant Mapper (http://www.hiv.lanl.gov/) schematic of identified recombination breakpoints plotted on the HXB2 genome. All breakpoints (vertically numbered) were determined using GARD and validated (within GARD) using the a posteriori Shimodaira and Hasegawa (SH) test for topological incongruence between multiple trees, with the requirement that $P < 0.01$. Inferred coding non-recombinant segments are labeled S1, S2 and S3 for the gag and env genes, respectively. While S1 and S3 of the env gene overlap vpu, rev and tat, only cDNA from the env gene (frame +3) was examined.*

Majority-rule consensus trees (Appendix II, Figs. 2-4) were constructed in MrBayes, using non-recombinant segments for comparison to the trees prior to recombination detection. Phylogenies drawn from the *gag1* and *gag2* datasets are largely unresolved beyond the monophyletic clustering of subtype B, possibly due to the short length of the analyzed segments (Appendix II, Figs. 2 and 3). The reconstructed trees for the *env* segments demonstrated three of the four possible scenarios (Appendix II, Fig. 4). The phylogenetic tree representing the evolutionary history from *envS1* shows that the Haitian sequences form polytomic clades with respect to each other. Although the relationship between the different Haitian taxa is largely unresolved, each cluster of Haitian sequences is sister to the pandemic clade, indicating simultaneous epidemics. The phylogenetic tree reconstructed from the shortest segment, *envS2*, is star-shaped, which may be indicative of a recombination hotspot or high migration (Gilbert et al. 2007). Alternatively, there may not be sufficient signal, due to short length, to make a reliable conclusion. The *envS3* tree clearly demonstrates a Haiti-first scenario, with multiple Caribbean sequences clustering within the pandemic clade.

In an effort to compare pre-and post-recombination detection dates of the worldwide subtype B epidemic onset as well as to reconstruct phylogenies with greater resolution, we enforced an uncorrelated lognormal relaxed molecular clock in BEAST (Drummond et al. 2012). As in Gilbert et al., (2007), we performed 10 independent analyses per non-recombinant segment, each running for 100,000,000 generations. The only difference between this and the former analysis is that the non-Caribbean clade is no longer limited to the US and a single Canada-originating sequence. The resulting phylogenies representing the six non-recombinant segments and the *gag2* dataset did not

support a Haiti-first topology (Figs. 2-4 and Appendix II, Figs. 5-7). Instead, maximum clade credibility trees reconstructed from the majority of *env* positions present relationships more indicative of independent and simultaneous introductions of subtype B into Haiti and the rest of the world (Figs. 2 and 4). Phylogenetic reconstruction of the shortest *env* segment suggests that subtype B has been introduced into Haiti multiple times during the subtype B pandemic or that the represented region within the *env* gene is a recombination hot spot (Fig. 3). While topological differences between this and the previous study suggest drastically different routes of spread for subtype B from Africa into the global population, the time to the most recent common ancestor (tMRCA) is nearly identical to the previous estimate when comparing the longest *env* segments (Figs. 3-5 and Appendix II, Figs. 7A and C). It should be noted that timing estimates using different genomic regions and where datasets are expanded beyond a single subtype show the tMRCA of subtype B to be well before 1960 (Worobey et al. 2008; Wertheim, Fourment, Kosakovsky Pond 2012).

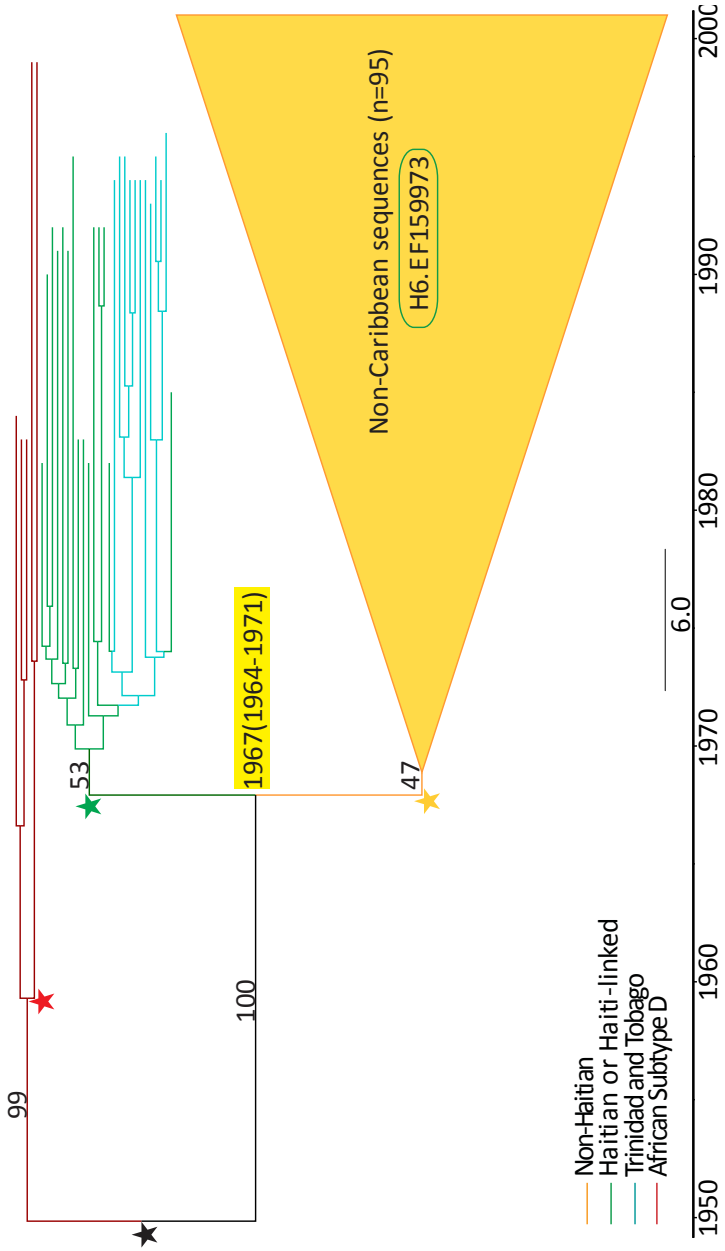


Figure 2. *Maximum clade credibility tree for genome positions 6231-7406. The maximum clade credibility tree reconstructed under a relaxed molecular clock does not support a Haiti-first model for the geographical spread of HIV-1 subtype B. Under a Bayesian Skyline Coalescent tree prior, the tips are representative of the year of sampling. Posterior probabilities are shown where all subtype D, all subtype B, Caribbean subtype B and Non-Caribbean subtype B taxa are clustered. Phylogenetic trees were individually reconstructed for all 127 sequences spanning HXB2 genome positions 6231-7406. Stars represent the TMRCA, where ★ = Subtype B/D: 1949 (1940-1958), ★ = Subtype D: 1959 (1951-1966), ★ = Haiti/Caribbean Subtype B: 1969 (1966-1973) and ★ = Non-Haitian/Pandemic Subtype B: 1968.*

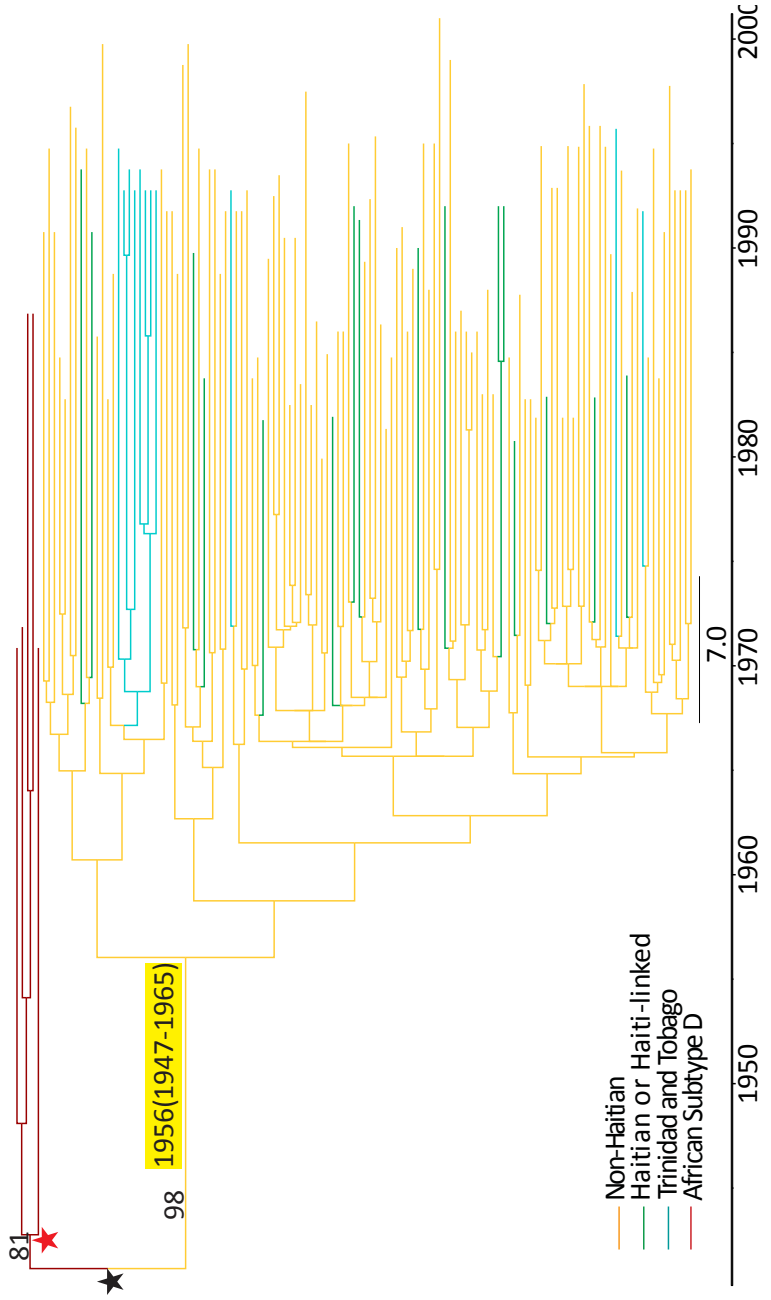


Figure 3. *Maximum clade credibility tree for genome positions 7407-7898. Maximum clade credibility tree reconstructed under a relaxed molecular clock do not support a Haiti-first model for the geographical spread of HIV-1 subtype B. Under a Bayesian Skyline Coalescent tree prior, the tips are representative of the year of sampling. Posterior probabilities are shown where all subtype D, all subtype B, Caribbean subtype B and Non-Caribbean subtype B taxa are clustered. Phylogenetic trees were individually reconstructed for all 127 sequences spanning HXB2 genome positions 7407-7898. Stars represent the TMRCA, where* ★ *= Subtype B/D: 1940 (1921-1958),* ★ *= Subtype D: 1941 (1938-1969).*

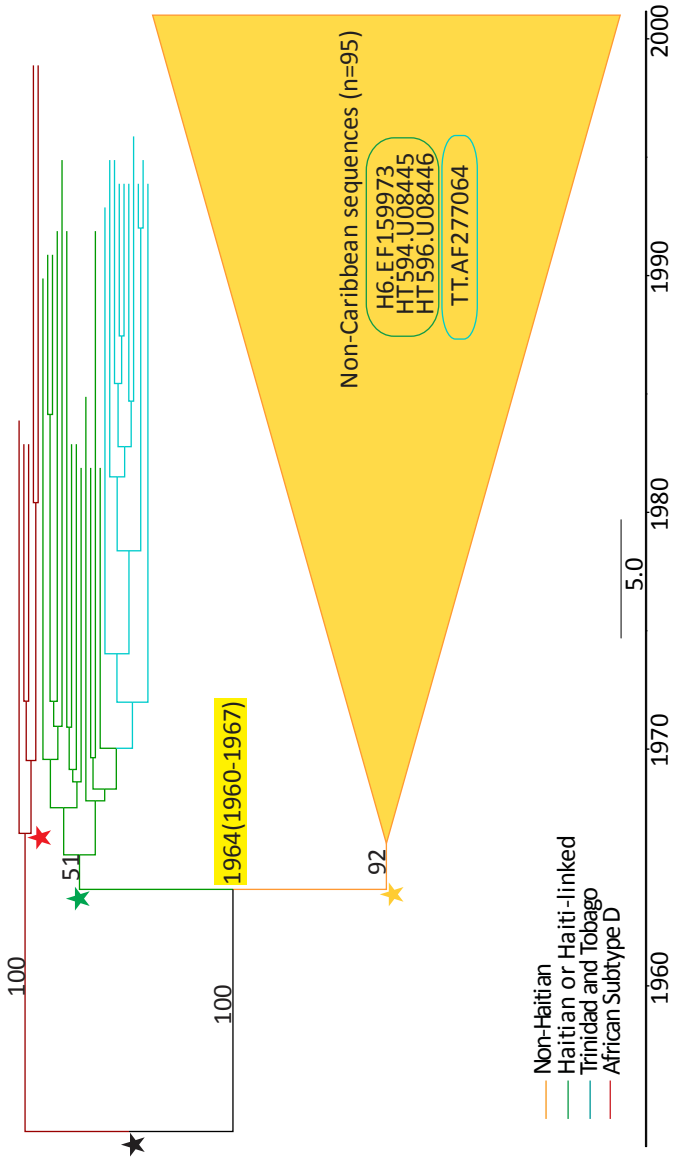


Figure 4. *Maximum clade credibility tree for genome positions 7899-8795. Maximum clade credibility tree reconstructed under a relaxed molecular clock do not support a Haiti-first model for the geographical spread of HIV-1 subtype B. Under a Bayesian Skyline Coalescent tree prior, the tips are representative of the year of sampling. Posterior probabilities are shown where all subtype D, all subtype B, Caribbean subtype B and Non-Caribbean subtype B taxa are clustered. Phylogenetic trees were individually reconstructed for all 127 sequences spanning HXB2 genome positions 7899-8795. Stars represent the TMRCA, where ★ = Subtype B/D: 1953 (1946-1960), ★ = Subtype D: 1966 (1961-1971), ★ = Haiti/Caribbean Subtype B: = 1965 (1961-1969) and ★ = Non-Haitian/Pandemic Subtype B: 1966 (1962-1969).*

Upon comparison of the trees independently reconstructed from *envS1* and *S3*, both of which demonstrated simultaneous and distinct post-Africa epidemics into Haiti and the rest of the world, we observed three sequences that switched between the two major clades in question (Fig. 5). Despite their Caribbean origin, these taxa clustered within the pandemic clade in *S3*. HT594 and HT596 are sequences sampled from HIV-infected pregnant women both of whom were residing in Port-Au-Prince, Haiti's notorious slum, Cite Soleil in 1992 (Ruff et al. 1994). Interestingly, despite the fact that these two sequences switch clades depending on the examined segment, they maintain a strongly supported cluster (Fig. 5A, S7A and C). This indicates that a recombination event involving HIV-1 subtype B from both a Haitian and Pandemic individual must have occurred and the recombinant strain circulated within the population prior to these women becoming infected. A second possibility is that HT594 or HT596 was involved in the recombination event and then the other was subsequently infected. Regardless of the mode of infection, this highlights the existence of cross-clade recombinants sometime between the entrance of subtype B into Haiti (Fig. 5, 1967) and the most recent common ancestor of HT594 and HT596 (1988). Given that Haiti has been identified as a sexual tourism hot spot for vacationers from around the world (Hooper 1999), it is not surprising that we find recombinants that are chimeras between strains from Haiti and other geographic regions.

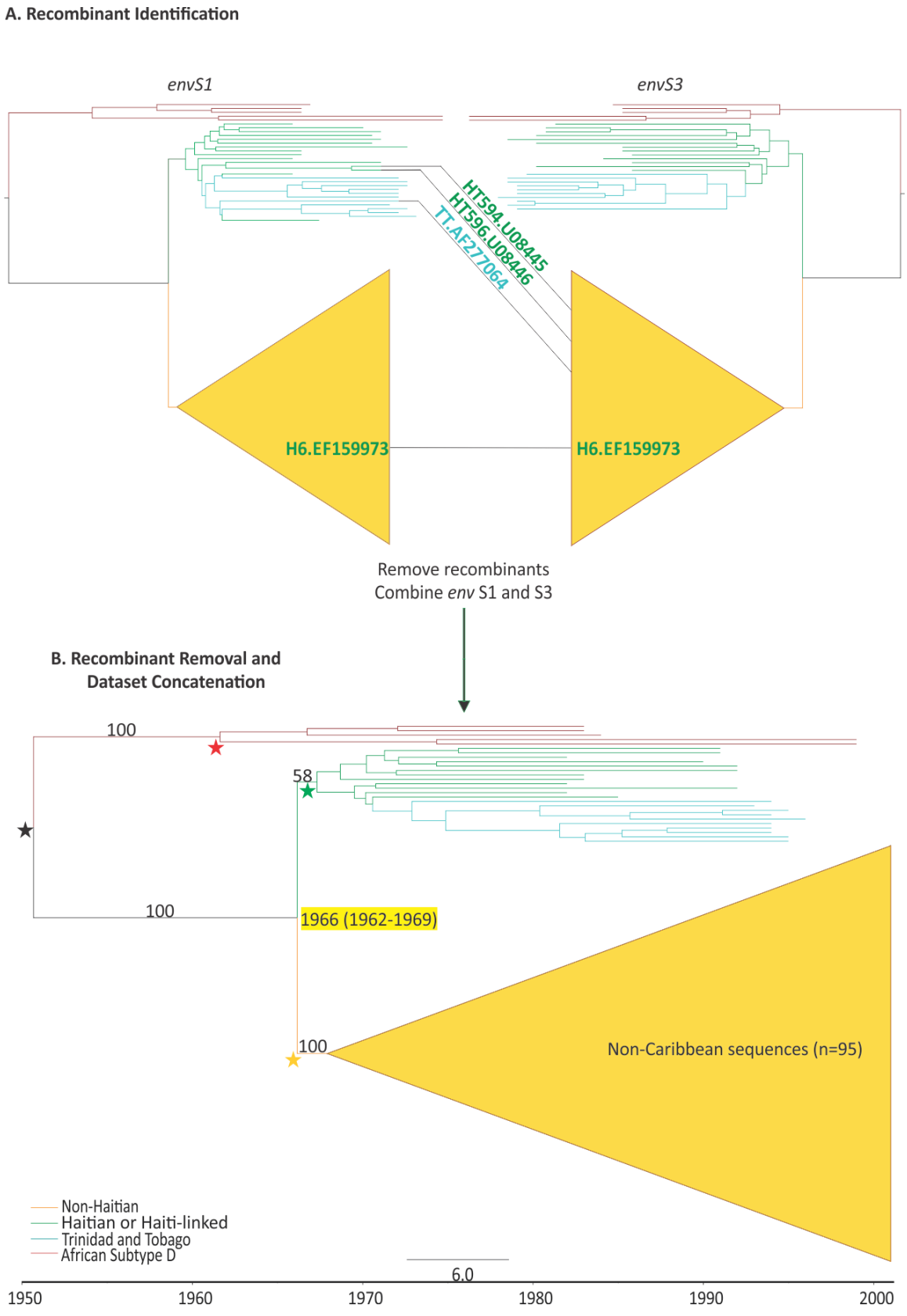

A. Recombinant Identification
envS1
envS3
HT594.U08445
HT596.U08446
TT.AF277064
H6.EF159973
H6.EF159973
Remove recombinants
Combine env S1 and S3
B. Recombinant Removal and
Dataset Concatenation
100
58
100
1966 (1962-1969)
100
Non-Caribbean sequences (n=95)
Non-Haitian
Haitian or Haiti-linked
Trinidad and Tobago
African Subtype D
6.0
1950
1960
1970
1980
1990
2000

Figure 5. *Recombinant Removal and concatenation of envS1 and S3. The simultaneous and distinct post-Africa epidemic scenario is recapitulated when recombinants identified through phylogenetic discordance are removed. A) When compared, three sequences of Caribbean origin that cluster with the Caribbean clade in envS1 switch clades and cluster with the non-Caribbean sequences in envS3. B) The first and third env segments are concatenated and cross-clade recombinants and H6 are removed. The maximum clade credibility tree reconstructed under a relaxed molecular clock does not support a Haiti-first model for the geographical spread of HIV-1 subtype B. Under a Bayesian Skyline Coalescent tree prior, the tips are representative of the year of sampling. Posterior probabilities are shown where all subtype D, all subtype B, Caribbean subtype B and Non-Caribbean subtype B taxa are clustered. Phylogenetic trees were individually reconstructed for all 123 sequences spanning HXB2 genome positions 6231-7406 and 7899-8795. Stars represent the TMRCA, where ★ = Subtype B/D: 1950 (1942-1958),* *= Subtype D: 1961 (1955-1967),* *= Haiti/Caribbean Subtype B: = 1967 (1963-1970) and* *= Non-Haitian/Pandemic Subtype B: 1967(1964-1970).*

Given that *envS1* and *S3* had nearly identical topologies at the level of Caribbean versus Pandemic clustering (Fig. 5A), we concatenated the two segments after removing the identified recombinants to find out if the simultaneous and distinct epidemic result would be replicated (Fig. 5B). Indeed, both the Caribbean and Pandemic sequences form monophyletic clades, respectively. Again, the estimated TMRCA of subtype B is 1966 (Fig. 5), in agreement with Gilbert et al. (2007).

As mentioned previously, GARD is useful for the identification of breakpoints. Through the phylogenetic reconstruction of non-recombinant segments, recombinants that have significant topological effects may be revealed. Sequences that only have short recombinant stretches or where the recombination event has taken place between more similar sequences may be missed. To better pinpoint regions where recombination events occur across the envelope gene, the Recombination Detection Program version 4 (Martin et al. 2010) was used. When looking across the entire analyzed *env* region, it is apparent that recombinants are scattered throughout. Even after we filtered out recombinant hits

that were possibly due to misalignment, the greatest number of recombination events is clustered within the *envS2* region (Fig. 6). This finding is supportive of a recombination hot spot(s) within *S2*. Beyond estimating locations where recombination has occurred and whether there is a cross-clade or intra-clade event, it is difficult to remove putative recombinants and redraw the phylogeny. This is because in the case of most events, the parent and recombinant sequences could not be disentangled. Regardless, the placement of recombinants appears to be roughly in line with the GARD inferred breakpoints.

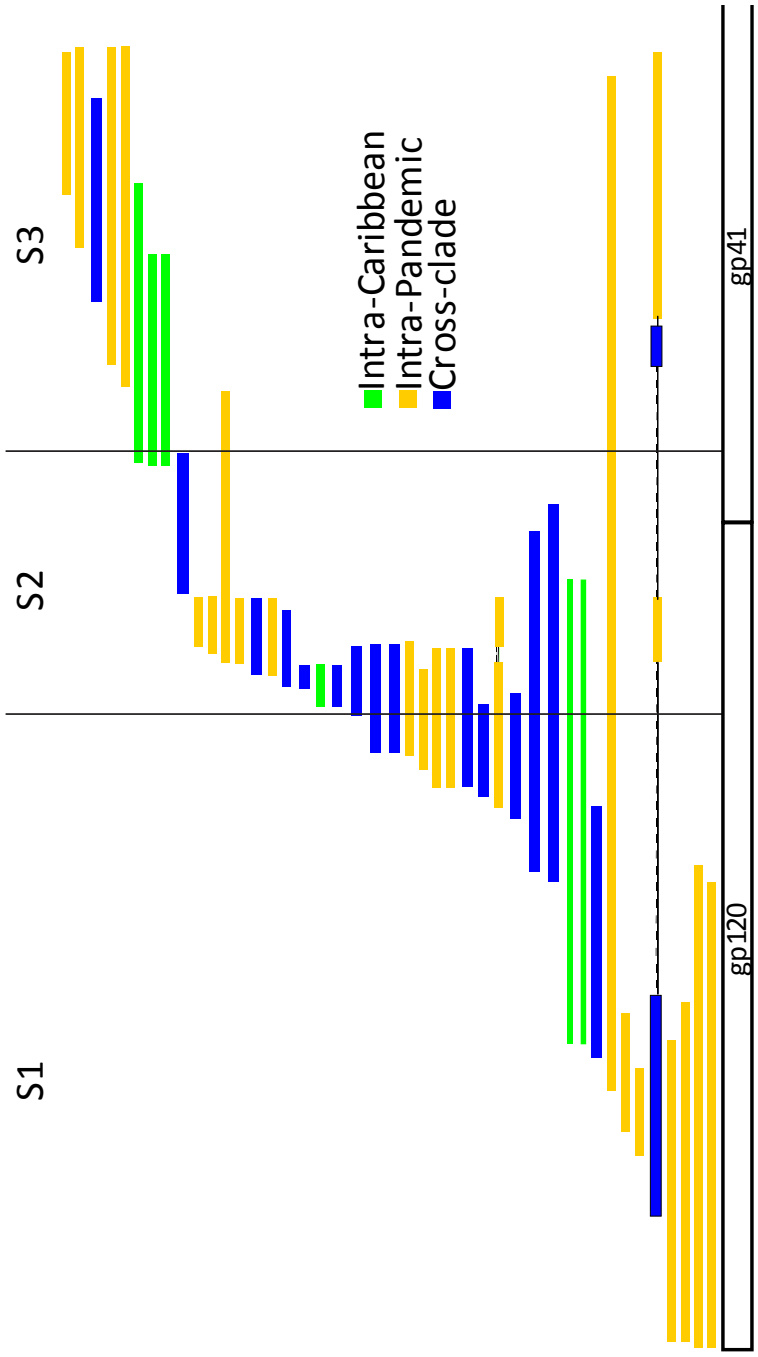


Figure 6: *Recombinants identified using RDP. The presence of recombination is verified using RDP. All RDP identified-recombinant regions within sequences are plotted along the length of the env gene. Recombination events involving: two Caribbean derived strains are designated green, two Pandemic strains are gold or Caribbean and Pandemic strains are blue. Black lines connecting recombinant regions indicate multiple recombination events within the same sequence. GARD-inferred breakpoints are designated by the vertical lines separating the segments.*

Conclusions

Outside of sub-Saharan Africa, the Caribbean has been most devastated by HIV, and until recently, Haiti had both the highest number of people living with HIV and new infections per capita within the region (Joint United Nations Programme on HIV/AIDS., World Health Organization. 2008; Joint United Nations Programme on HIV/AIDS. 2011). It has been repeatedly suggested that the significant Haitian presence in the post-independent Congo region is likely responsible for the entrance of subtype B into the Haiti and our results do not oppose that possibility (Hooper 1999; Pepin 2011). Alternatively, there were a number of individuals from other countries, included in the worldwide subtype B pandemic population, that were also recruited to the Congo for teaching and other employment opportunities (Appendix II, Fig. 8) (Kimpesa 1983). It is known that at least a few Belgian individuals residing in the area were infected with HIV before returning to Belgium (Sonnet et al. 1987). This is not surprising, since Belgium had a strong presence in the Congo river basin area beginning in 1884, and officially established a colony (Belgian Congo) in 1908 through 1960. While it would be interesting to include earlier dating taxa representing other countries where people traveled between their homes and the Congo region, there are no publicly available sequences. Of course, such a study could also reveal that these early infections were dead ends with no connection to the worldwide pandemic.

Here, we have considered the four possible scenarios of subtype B dispersal after Africa, as it applies to Haiti. Our results indicate that after HIV-1 subtype B emerged from Africa, there were two subsequent and nearly simultaneous epidemics into Haiti and

the rest of the world; the latter has become one of the most severe disease pandemics of human history. While the short *gag* segments and *envS2* do not provide much in the way of the geographical path that subtype B took, it is reassuring that the longest two segments, which were independently analyzed, come to the same conclusion. Furthermore, by screening this dataset for the presence of recombination, we were able to identify two events where pandemic and Caribbean strains recombined. Once these evolutionary contaminants were removed from the dataset, the two segments could be joined and displayed, again, the distinct and simultaneous epidemic scenario. The data presented here reveal that phylogenies derived from shorter segments, estimated to be non-recombinant, indicate that Haiti was not a jumping point for HIV-1 subtype B into the rest of the world.

Methods

Sequence Collection

Nucleotide sequences used in the current analyses were downloaded from the Los Alamos National Laboratory HIV Database (http://www.hiv.lanl.gov/). Three datasets were constructed (Table S1): 1) An envelope (*env*) dataset composed of 127 coding nucleotide sequences spanning HXB2 6231-8795, mimicking the dataset in Gilbert et al.(2007) ; 2) A *gag* dataset composed of the same sequences used in Gilbert et al. spanning HXB2 820-1952. This dataset will be referred to as the *gag*1 dataset.; 3) the *gag*2 dataset is composed of all *gag* sequences available from individuals that were represented in the *env* dataset and spans HXB2 1198-1866. (Table S1).

Subtype assignments for all sequences were validated using RIP (http://www.hiv.lanl.gov/content/sequence/RIP/RIP.html) (Siepel et al. 1995). The following parameters were chosen: The HIV subtype consensus alignment was set as background, with a window size set at 400, confidence threshold of 95% and global gap-stripping with all window values plotted. Codon alignments were produced with MUSCLE (Edgar 2004) as implemented in MEGA5.0 (Tamura et al. 2011).

Reproduced Trees

Phylogenetic analyses were completed using MrBayes (Huelsenbeck, Ronquist 2001). The model settings for all constructed trees specified a General Time Reversible Model with a proportion of invariable sites and a gamma-shaped distribution to describe site to site rate variation (GTR+I+Γ). Although the best fitting model for the *gag* datasets is TN93, MrBayes does not have a complementary model built in. Appropriate models were determined using the DNA/Protein Model test in MEGA5 (Tamura et al. 2011).

Trees built under the assumption of no recombination were created from the *env* (Fig S1a) and first *gag* datasets (Fig S1b) to reproduce results in Gilbert et al. The *env* and *gag* analyses were run for 20,000,000 (×2) and 5,000,000 (×2) generations, respectively, with the chains being sampled every 1000 generations. Majority rule consensus trees were constructed after a 25% burn-in was applied and Tracer (Rambaut 2007) was used to ensure convergence within and between chains after a 10% burn-in. FigTree (Rambaut 2009) was used to produce all tree figures.

Intra-subtype Recombination Detection

The scan for intrasubtype recombination was restricted to subtype B sequences, as the inclusion of subtype D or C outgroup sequences would reduce support for subtype B-specific breakpoints. The Single Break-Point (SBP) screen for recombination (http://www.datamonkey.org/) was used as an initial filter for all datasets (Table S2), while GARD (Kosakovsky Pond et al. 2006) was exploited in-house to detect multiple recombination breakpoints (Table S3). The 012345 (REV) and 010040 (TrN93) models were selected to describe the substitution patterns of the *env* and *gag* datsets, respectively, for both SBP and GARD. Site-to-site rate variation was modeled by the General Discrete Distribution with 6 rate classes.

The *env* subtype B dataset was also analyzed with RDP3 (Martin et al. 2010) to verify the presence of recombination. RDP (Martin, Rybicki 2000), GENECONV (Padidam, Sawyer, Fauquet 1999), Bootscan (Martin et al. 2005), MaxChi (Smith 1992), Chimaera (Posada, Crandall 2001), SiSscan (Gibbs, Armstrong, Gibbs 2000), PhylPro (Weiller 1998), LARD (Holmes, Worobey, Rambaut 1999) and 3Seq (Boni, Posada, Feldman 2007) were used to identify recombinants. Any recombinant was accepted as long as it met the additional criteria of causing phylogenetic discordance and was not considered to be a possible misalignment artifact.

Trees Constructed from Non-Recombinant Segments

Phylogenetic analyses were completed using MrBayes (Huelsenbeck, Ronquist 2001) using GTR+I+Γ. Chains for the *env* and both *gag* datasets were run 65,000,000 and 20,000,000 generations, respectively, with sampling every thousand generations.

Majority rule consensus trees and figures were constructed using the same methodology as in the reproduced trees.

Molecular Dating

For the inference of a time-measured phylogeny that would shed light on the emergence of the worldwide HIV subtype B epidemic, molecular dating was estimated using BEAST (Drummond et al. 2012). A random start tree was used for each analysis. An uncorrelated log-normal relaxed molecular clock, as opposed to the rejected strict clock model (*LRT*: X^2=180.855, *df* = 126, $P \leq 0.001$), was used to allow variation of the rate of evolution among the tree branches, where each branch is drawn from an underlying lognormal distribution. A Bayesian Skyline coalescent tree prior was assigned**.** For each individual segment, 10 independent analyses were completed, where each analyses was run for 100,000,000 generations. For the *envS1 + S3* tree, 4 independent analyses, 100,000,000 in length, were completed. A burn-in of 10% was applied (10,000,000 generations discarded per analyses) and the maximum clade credibility tree is displayed along with posterior probabilities. It should be noted here, that since subtype D sequences were clustering within the subtype B clade in MrBayes trees constructed from both *gag1* and *gag2* datasets, we only used subtype C sequences as the outgroup for *gag* datasets for molecular dating. This was to ensure that conclusions about the subtype B clade were not affected by the chosen outgroup.

References

Boni MF, Posada D, Feldman MW. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176:1035-1047.

Bowdler N. 2007. Key HIV strain 'came from Haiti'. BBC News.

Carmichael M. 2007. Haunted by HIV's Origins. Newsweek Magazine. New York City: Newsweek, Inc.

Crisp MD, Cook LG. 2005. Do early branching lineages signify ancestral traits? Trends Ecol Evol 20:122-128.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-1973.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Farmer P. 1992. AIDS and accusation : Haiti and the geography of blame. Berkeley: University of California Press.

Gibbs MJ, Armstrong JS, Gibbs AJ. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16:573-582.

Gilbert MT, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A 104:18566-18570.

Gojobori T, Moriyama EN, Ina Y, Ikeo K, Miura T, Tsujimoto H, Hayami M, Yokoyama S. 1990. Evolutionary origin of human and simian immunodeficiency viruses. Proc Natl Acad Sci U S A 87:4108-4111.

Goodrich DW, Duesberg PH. 1990. Retroviral recombination during reverse transcription. Proc Natl Acad Sci U S A 87:2052-2056.

Hahn BH, Shaw GM, Taylor ME, Redfield RR, Markham PD, Salahuddin SZ, Wong-Staal F, Gallo RC, Parks ES, Parks WP. 1986. Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. Science 232:1548-1553.

Holmes EC, Worobey M, Rambaut A. 1999. Phylogenetic evidence for recombination in dengue virus. Mol Biol Evol 16:405-409.

Hooper E. 1999. The river : a journey to the source of HIV and AIDS. Boston, MA: Little, Brown and Co.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Joint United Nations Programme on HIV/AIDS. 2011. UNAIDS World AIDS day report 2011. Geneva: UNAIDS.

Joint United Nations Programme on HIV/AIDS., World Health Organization. 2008. AIDS outlook/09 World AIDS Day 2008. Geneva: UNAIDS : World Health Organization.

Kimpesa M. 1983. L'operation de l'UNESCO au Congo-Leopoldville et le diagnostic des realities educatives congolaises: 1960-1964. Faculte de Psychologie. Geneva: Universite de Geneve.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R, Lapedes A, Hahn BH, Wolinsky S, Bhattacharya T. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789-1796.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23:1891-1901.

Krell F-T, Cranston PS. 2004. Which side of the tree is more basal? Systematic Entomology 29:279-281.

Lambert B. 1990. Now, No Haitians Can Donate Blood. The New York Times. New York City: The New York TImes COmpany.

Li WH, Tanimura M, Sharp PM. 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 5:313-330.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562-563.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26:2462-2463.

Martin DP, Lemey P, Posada D. 2011. Analysing recombination in nucleotide sequences. Mol Ecol Resour 11:943-955.

Martin DP, Posada D, Crandall KA, Williamson C. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses 21:98-102.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265:218-225.

Pape JW, Farmer P, Koenig S, Fitzgerald D, Wright P, Johnson W. 2008. The epidemiology of AIDS in Haiti refutes the claims of Gilbert et al. Proc Natl Acad Sci U S A 105:E13.

Pepin J. 2011. The origins of AIDS. Cambridge, UK ; New York: Cambridge University Press.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A 98:13757-13762.

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. Journal of molecular evolution 54:396-402.

Rambaut A. 2007. Tracer v1.4.

Rambaut A. 2009. FigTree v1.3.1.

Robertson DL, Hahn BH, Sharp PM. 1995. Recombination in AIDS viruses. J Mol Evol 40:249-259.

Ruff AJ, Coberly J, Halsey NA, et al. 1994. Prevalence of HIV-1 DNA and p24 antigen in breast milk and correlation with maternal factors. J Acquir Immune Defic Syndr 7:68-73.

Schierup MH, Hein J. 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879-891.

Shimodaira H, Hasegawa M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular Biology and Evolution 16:1114.

Siepel AC, Halpern AL, Macken C, Korber BT. 1995. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. AIDS Res Hum Retroviruses 11:1413-1416.

Smith JM. 1992. Analyzing the mosaic structure of genes. J Mol Evol 34:126-129.

Sonnet J, Michaux JL, Zech F, Brucher JM, de Bruyere M, Burtonboy G. 1987. Early AIDS cases originating from Zaire and Burundi (1962-1976). Scand J Infect Dis 19:511-517.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739.

Weiller GF. 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol Biol Evol 15:326-335.

Wertheim JO, Fourment M, Kosakovsky Pond SL. 2012. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. Mol Biol Evol 29:451-456.

Woolley SM, Posada D, Crandall KA. 2008. A comparison of phylogenetic network methods using computer simulation. PLoS One 3:e1913.

Worobey M, Gemmel M, Teuwen DE, et al. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455:661-664.

CHAPTER 3

N-LINKED GLYCOSYLATION SITES: EVOLUTIONARY CONTRIBUTION AND STRUCTURAL CHARACTERIZATION

Abstract

On the HIV gp120 gene, potential N-linked glycosylation sites (PNGSs) are commonly thought to be the positions where mutations are most commonly observed within and among individuals. Although collectively the mutational accumulation appears to be great, glycans that attach to these sites are required for proper folding and subsequent attachment to the primary host cell receptor, CD4. Simultaneously, these glycans provide protection to the virus against host immune defenses. For the first time, the evolutionary contribution of PNGSs has been quantified, and I statistically determined that greater than half of the divergence of gp120 is due to mutations at these sites. Furthermore, from the longitudinally sampled sequences analyzed here, two potential N-linked glycosylation sites are identified that are significantly more conserved than all others. In addition to the conservation analysis, protein structural analyses, where the PNGSs and critical binding sites were mapped to a gp120 structure, were carried out. Results revealed that only four PNGSs are situated such that they are the closest to fifty of fifty-two non-overlapping binding sites. This suggests an evolutionary strategy to maximize the protective abilities of glycans while minimizing their attachment sites. Finally, I found that potential N-linked glycosylation sites tend to be closest to binding sites with the greatest flexibility, potentially suggesting an environmental preference for occurrence on the protein structure.

Introduction

It has been nearly three decades since the causative agent of Acquired Immunodeficiency Syndrome was determined to be a retrovirus, now known as the Human Immunodeficiency Virus (HIV). Three years later, in 1987, a nucleoside analog reverse transcriptase inhibitor, azidithymidine (AZT), became the first drug approved by the Food and Drug Administration to combat HIV (www.fda.gov). Saquinavir, approved in 1995, was the first protease inhibitor brought to market (Naeger et al. 2010). To date, at least 36 antiretrovirals are available, 35 of which target the viral reverse transcriptase, protease, integrase and/or the gp41 portion of the envelope (www.fda.gov). The single commonality among all of these drugs is that they interact with protein component of the virus. Given that a substantial fraction, 60 kD of 120 kD (Lasky et al. 1986; Geyer et al. 1988), of HIV's envelope glycoprotein gp120 is composed of N-linked glycans, a new therapeutic target is apparent.

N-linked glycosylation is a post-translational protein modification that occurs in the rough endoplasmic reticulum (RER), where large oligosaccharide precursors are attached to asparagines residues that are part of a NX[S/T] tripeptide. Further enzymatic processing in the RER yields an N-linked glycan where one mannose and three glucose residues have been cleaved from the oligosaccharide precursor. The glycosylated protein is transported by vesicle to the Golgi apparatus for further processing. If pertinent parts of the attached oligosaccharide are not accessible to mannosidases, which further cleave mannose residues, the protein with its high-mannose oligosaccharide(s) is the final product. More commonly, in vertebrate cells, the protein product will possess complex oligosaccharides. In contrast to the high mannose oligosaccharides, two additional mannose residues are removed and three N-acetylglucosamine, one fucose and three galactose residues as well as three N-acetylneuraminic acids are attached (Lodish et al.

2000). Interestingly, HIV possesses both types of N-linked glycans as well as hybrids between the two, with the vast majority consisting of the incompletely processed simple high-mannose type oligosaccharides (Lasky et al. 1986; Geyer et al. 1988).

The necessity of N-linked glycans was shown with the expression of gp120 in the presence of the antibiotic tunicamycin, which blocks the initial transfer of the large oligosaccharide precursor in the RER and prevents synthesis of all N-linked glycans. Non-glycosylated forms of gp120 were unable to bind to the host CD4. It was further discovered that the non-glycosylated forms of gp120 were misfolded due to the random formation of disulfide bonds. Collectively, this research demonstrated that N-linked glycans are required for the proper folding of gp120, and furthermore, without this folding, the virus cannot interact with the host cell (Li et al. 1993).

In addition to the mediation of correct folding, N-linked glycans also play a major role in neutralizing antibody evasion. Comparison of longitudinally sampled sequences from a single HIV-1 subtype B infected individual revealed that neutralization resistant viral populations primarily contained mutations at potential N-linked glycosylation sites (PNGSs) (Wei et al. 2003). While PNGSs were primarily gained over time, they were also lost at some positions. Moreover, there are paired, and perhaps even more complex interactions between glycans; those within proximity to each other tend to have exclusionary interactions, while those more distant from each other tend to have inclusionary interactions (Poon et al. 2007).

The goal of the current study is to characterize PNGSs through the analysis of sequences derived from eleven normally progressing subtype B infected individuals. While many studies have verified that PNGSs change over time, none have actually quantitatively evaluated the total evolutionary contribution of PNGSs to the viral population within and among individuals. I have

further analyzed a portion of PNGSs that are relatively evolutionarily stable over time in an effort to understand the roles each stable PNGS plays in relationship to each other, critical host protein binding sites and within the general environment where they occur. In contrast to analyses that characterize sites by *in situ* mutation of amino or nucleic acids that are proximally located with respect to primary sequence distance (distance between amino acids of interest by counting intervening amino acids), I have statistically associated PNGSs with binding sites using structural distances. Studies of this nature are more evolution-centric, as selection acts on the phenotype (structure) and results in evolutionary changes to the genotype (sequence) and phenotype frequencies.

Results and Discussion

Effects of PNGSs on Divergence

To estimate the effect of PNGSs on overall evolutionary divergence of the viral population within each individual, two datasets were compared: one with all sites included (ALL) and one where all PNGS were removed (PGlyRem). After phylogenetic reconstruction and calculation of individual root to tip lengths and sum of branch lengths, the impact of PNGS removal was determined through statistical testing.

Individual-specific linear regression analyses were carried out to identify: 1) if a linear relationship exists between individual branch lengths reconstructed from the ALL and PGlyRem datasets, and 2) if divergence of the HIV population is decreased upon removal of PNGS. For this analysis, significance cannot be assessed, because shared phylogenetic history of tip sequences violates the independence assumption of linear regression. Regardless, by plotting all branch lengths from the PGlyRem against ALL dataset, it is apparent that branch length variation due to PNGS exists for the majority of taxa sampled from each individual (Fig. 1). The line of

expectation (slope = 1), where regression lines should fall if no difference exists between per-site branch lengths in the two datasets, has a greater slope than that from any of the individual regression lines drawn (Table 1 and Figure 1).

Individual	Regression Equation	R^2
Null	1x + 0	1
1	0.4003x + 0.0139	0.5193
2	0.2604x + 0.0073	0.8947
3	0.2279x - 0.0009	0.8264
4	0.2135x + 0.0045	0.7947
5	0.3475x + 0.0006	0.7582
6	0.2173x + 0.0073	0.7749
7	0.8223x + 0.003	0.8878
8	0.2987x - 0.0081	0.832
9	0.6242x - 0.022	0.662
10	0.2338x + 0.0116	0.9004
11	0.1948x + 0.002	0.8517

Table 1. Individual analysis of PNGS contribution to per-site divergence. Linear regression equations and regression coefficients are shown for each taxon branch length in the PGlyRem versus ALL dataset, along with the null expectation (Null).

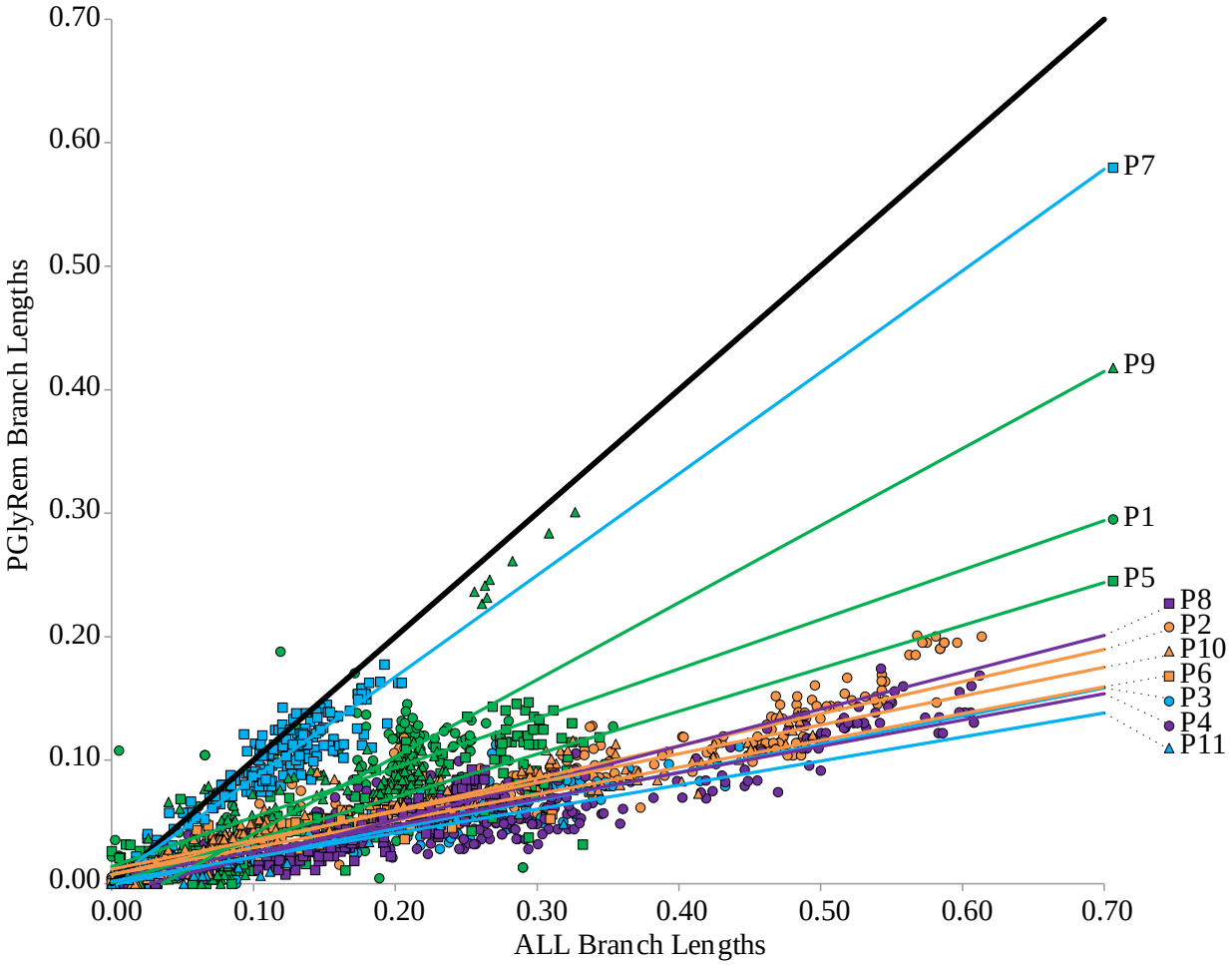


Figure 1. Individual analysis of PNGS contribution to per-site divergence. PGlyRem branch lengths are plotted against ALL branch lengths for taxa sampled from each individual. The thick black line represents the line of expectation (m = 1), where removal of glycosylation sites would not have had any impact on per-site branch lengths.

A second simple linear regression was performed to compare the normalized sum of branch lengths between the ALL and PGlyRem datasets. The sum of branch lengths for the PGlyRem versus ALL (Appendix III, Table 6) dataset was plotted for each individual (Fig. 2), revealing a strong linear relationship. More specifically, nearly 89% of the variation in the normalized sum of branch lengths for the PGlyRem dataset can be explained the regression (Fig 2). An analysis of variance performed within the regression established that the sum of branch lengths in the ALL dataset could statistically significantly predict the PGlyRem sum of branch lengths, $F(1, 9) = 75.4$, $p < 0.0005$. This is expected since the compared phylogenetic trees are equivalent with the exception of the PNGS removed in the PGlyRem dataset. The remaining 12% of variation not explained by the regression of PGlyRem on ALL, must be accounted for by the removal of PNGSs.

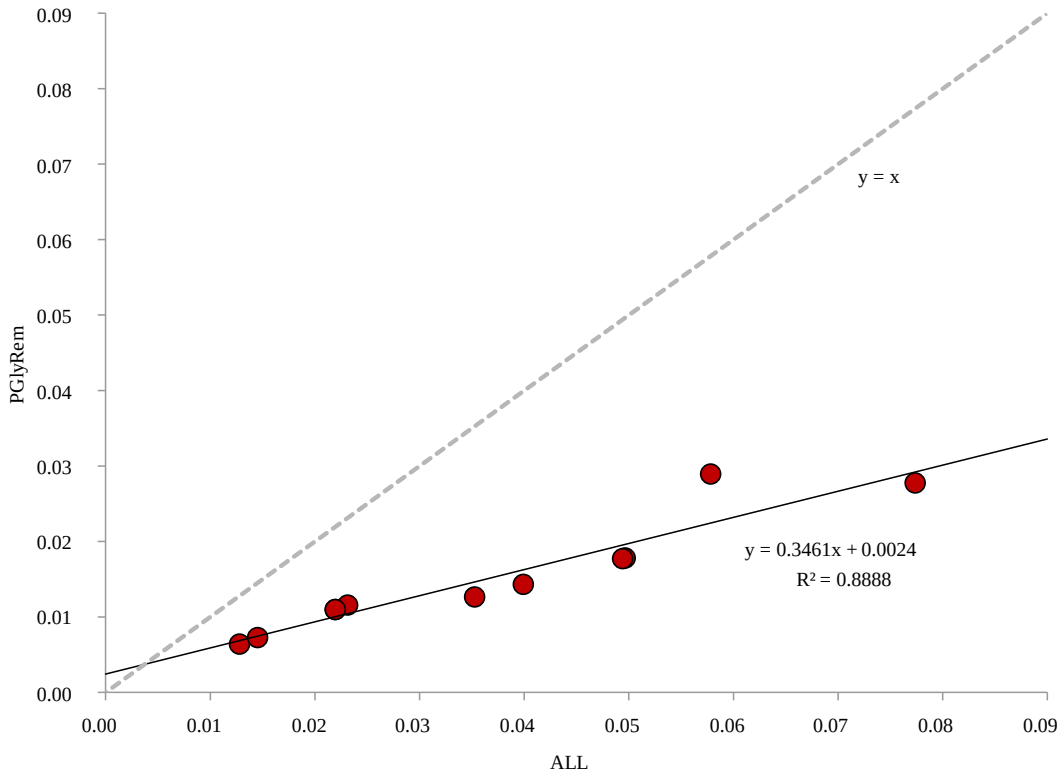


Figure 2. Plotted normalized sum of branch lengths for PGlyRem vs ALL. The dashed grey line with slope of 1 (m = 1) indicates what we would expect if there was no change between the sum of branch lengths reconstructed from the two datasets.

The slope of the regression equation (m = 0.34) indicates that branch lengths are generally shorter when glycosylation sites are removed. To identify whether or not the decrease in divergence, due to the removal of PNGS, is statistically significant, the slope of the linear regression line determined from plotting the PGlyRem sum of branch lengths against those from the ALL datasets was compared to the slope of the line of expectation (m = 1). The comparison revealed that the slope comparing PGlyRem and ALL datasets is statistically significantly reduced, indicating a significant reduction of divergence experienced by the PGlyRem dataset, t (9) = 16.02, $p < 0.0005$.

As a follow-up to the linear regression tests, a paired samples t-test was used to identify if there was statistically significant mean difference between the normalized sum

of branch lengths from PGlyRem and ALL (Appendix III, Tables 2 and 3). A single outlier was detected but inspection of its values revealed that it was not extreme and was therefore retained for the analysis (Appendix IV, Fig. 1). The differences between the sum of branch lengths did not statistically significantly deviate from normality as assessed by the Shapiro-Wilk test (*p = 0.231*, Appendix III, Table 4). Phylogenetic trees reconstructed from the PGlyRem datasets had shorter sum of branch lengths (0.015 ± 0.002 substitutions per site) than those reconstructed from the ALL datasets (0.037 ± 0.006 substitutions per site); a statistically significant decrease of 0.022 (95% CI, 0.013 to 0.031) substitutions per site, $t(10) = 5.391$, $p < 0.0005$, $d = 1.62$. Given that the two datasets used to reconstruct the branch lengths are identical with the exception of PNGS removal in the PGlyRem dataset, PNGS are responsible for a 59% reduction of the mean divergence (Appendix III, Table 3). Taken together, the regression analysis and paired t-test indicate that PNGSs are an important source of variation and that a considerable amount divergence within gp120 occurs at those sites.

The longitudinally sampled paired compartments datasets (Appendix III, Table 5) were analyzed to ascertain whether or not compartmentalization, site position and/or individual has an effect on the raw conservation of 12 stable PNGSs (> 60% conservation for at least 4 of 5 individuals). To compensate for the lack of independence between measurements, the Generalized Estimating Equations (GEE) procedure was used. While individual (Appendix III, Tables 6 and 7) and site (Appendix III, Tables 8 and 9) had an effect on raw conservation ($p < 0.0005$ for both), there was no effect of compartment (Appendix III, Tables 13 and 14). The lack of effect due to compartment indicates that both plasma and PBMC derived sequences can be used to make conclusions about PNGS

conservation for this dataset. The fact that PNGSs are differentially conserved indicates that there may be functional differences between sites. A difference of PNGS conservation among individuals is expected if sites are randomly conserved.

Identification of Differential Conservation Among PNGSs

A randomization study was carried out to find out which, if any, PNGSs were significantly more conserved than one would expect by random chance alone. Comparison of stable PNGSs ranked by conservation revealed that all glycosylation sites are not randomly conserved among individuals. Rather, two glycosylation sites are consistently found to be more conserved than all others (N301, $p = 0.012$ and N448, $p = 0.034$) (Fig. 3 and Appendix III, Table 5). Further investigations also revealed significantly more variation in conservation among observed PNGSs than would be expected by chance alone $(p = 0.0024)$ (Fig. 4); a finding that further supports the hypothesis that PNGSs are differentially conserved.

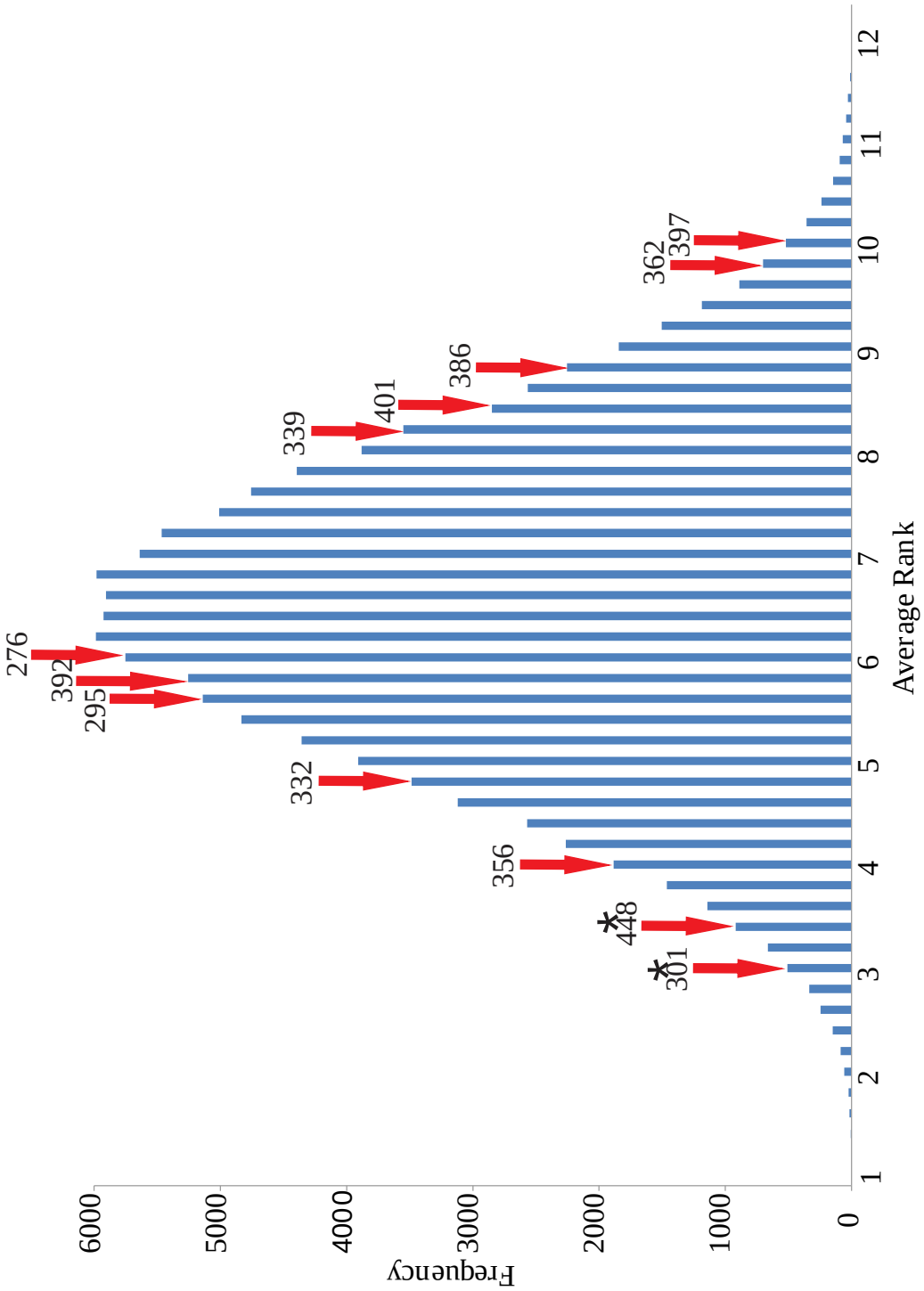


Figure 3. Histogram of average rank frequencies. The observed average ranks are indicated for each stable PNGS. Sites marked with an asterisk are statistically significantly more conserved than other stable glycosylation sites using a one-tailed test.

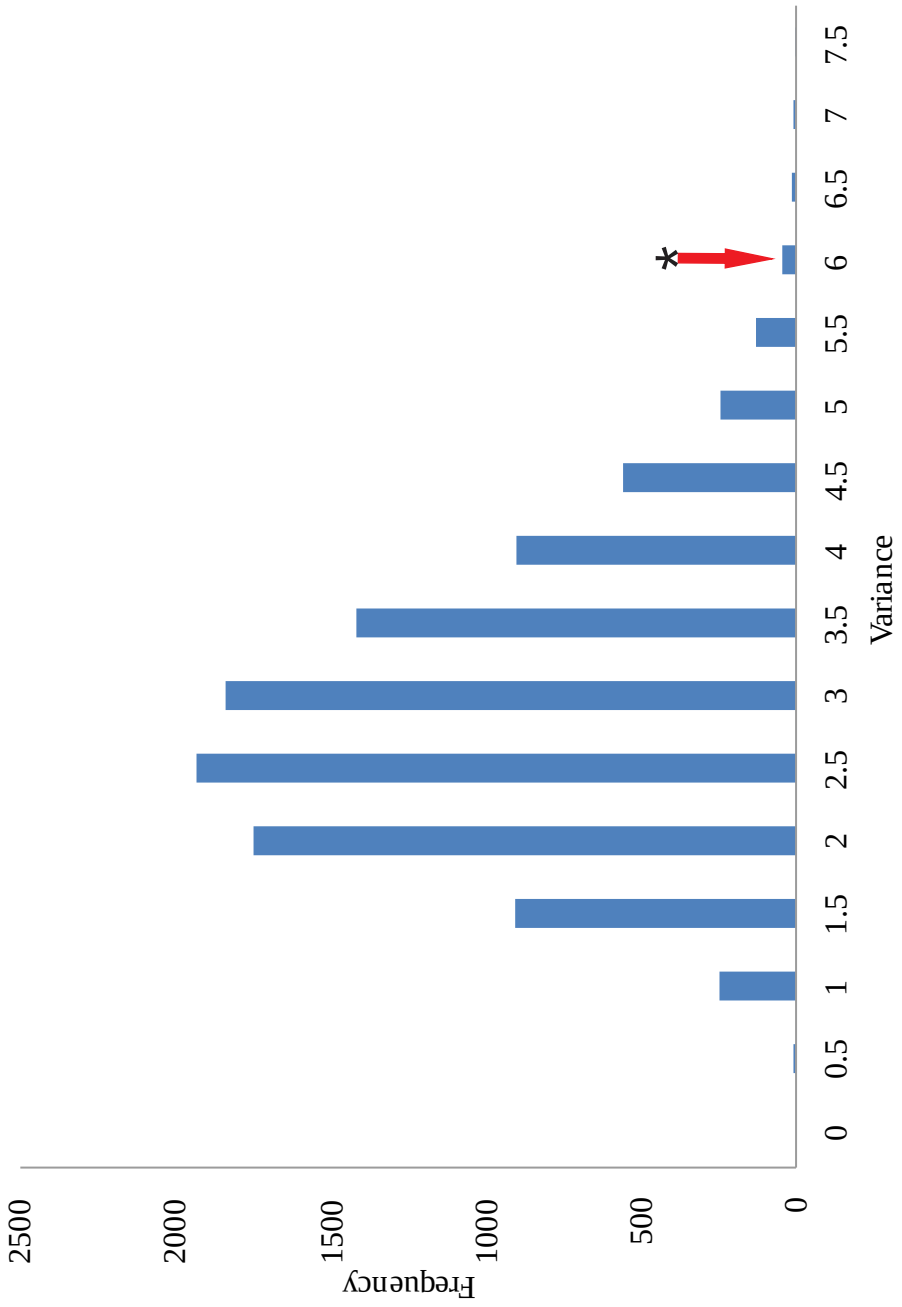


Figure 4. Histogram of average rank variances. The variance for each set of 12 randomly generated average ranks was plotted, versus the frequency of that variance. The variation in conservation for the observed PNGSs is shown with a red arrow and is denoted as statistically significant with an asterisk.

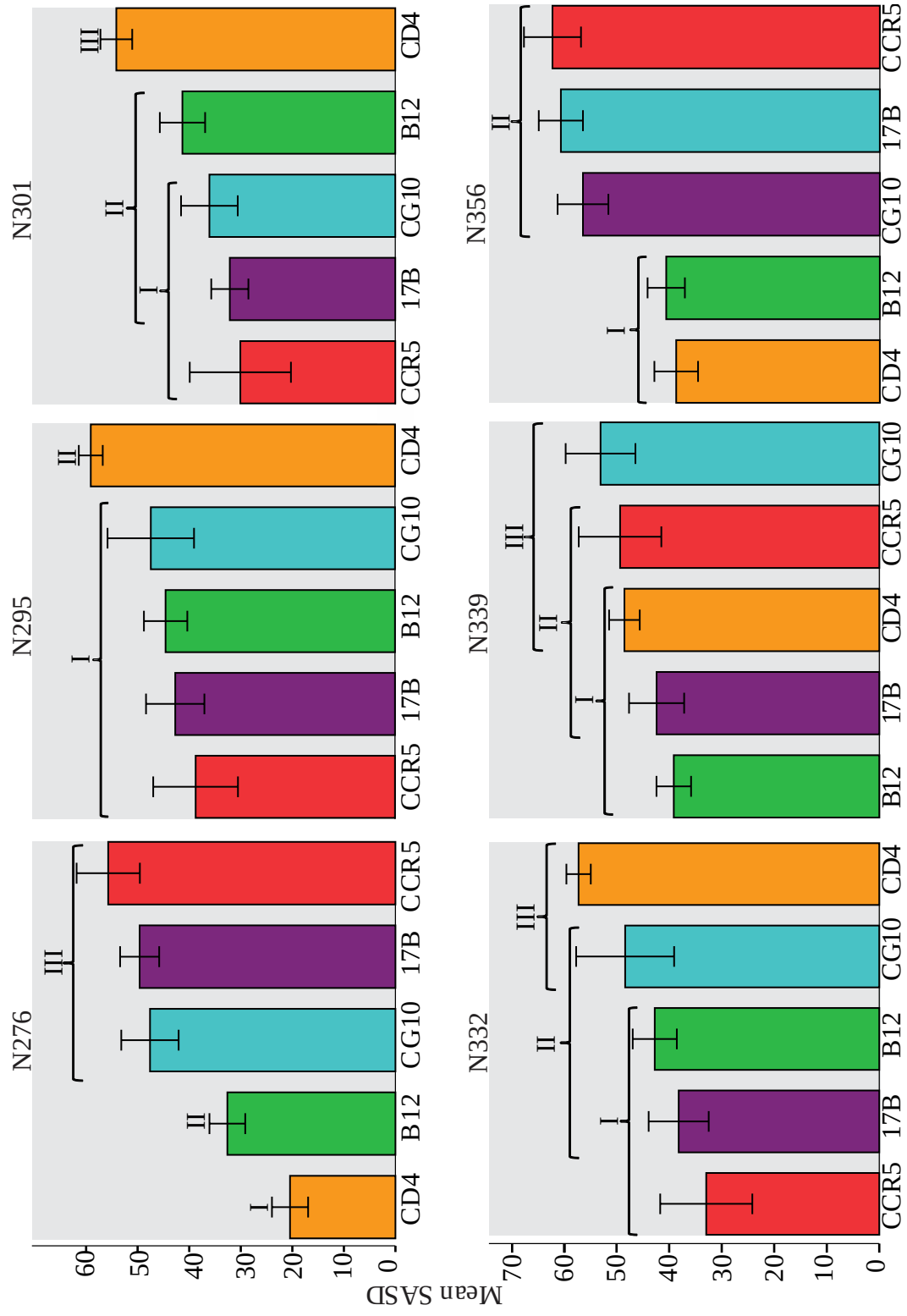
Characterization Stable PNGS by SASD to Critical Binding Sites

In an effort to functionally characterize PNGSs, the solvent accessible surface distance was calculated between each PNGS and linearly discontinuous critical binding sites on gp120 (Appendix III, Table 12). It should be noted that three of the PNGSs could not be mapped to 3TGQ and were removed from the structural analyses, reducing the total number of analyzed PNGSs to nine. Statistical analyses revealed that mean SASDs between a PNGS and the different binding site types (17B, B12, CG10, CD4 and/or CCR5) were statistically significantly different (Table 2).

Site	Levene's Test		ANOVA (or Welch's ANOVA)				Power
	F	*p*	F	$df_{between}$	df_{within}	*p*	
N276	1.42	0.240	47.8	4	74	< 0.005	1
N295*	2.50	0.050	19.3	4	20.6	< 0.005	1
N301*	3.82	0.018	15.0	4	26.5	< 0.005	1
N332*	2.90	0.027	22.9	4	21.7	< 0.005	1
N339	2.30	0.660	6.7	4	74	< 0.005	1
N356*	2.79	0.032	30.2	4	25.6	< 0.005	1
N362*	3.79	0.007	21.9	4	23.4	< 0.005	1
N386*	8.64	< 0.005	15.6	4	22.7	< 0.005	1
N448*	3.81	0.007	73.8	4	21.8	< 0.005	1

Table 2. Levene's test of homogeneity of variances and ANOVA or Welch's ANOVA results. Analyses using the non-parametric Welch's ANOVA in the case of heteroskedasticity are indicated with an asterisk.

Tukey's HSD test for multiple comparisons exposed which binding types individual PNGSs are positioned significanlty closer to or farther from (Figure 5). In only a single case did a PNGS cluster most closely with a single binding type (N276 with CD4) while most PNGSs were equidistantly spaced between multiple binding types. This is not surprising, as some binding sites partially overlap.


N276
N295
N301
N332
N339
N356
Mean SASD
0
10
20
30
40
50
60
70
I
II
III
CD4
B12
CG10
17B
CCR5

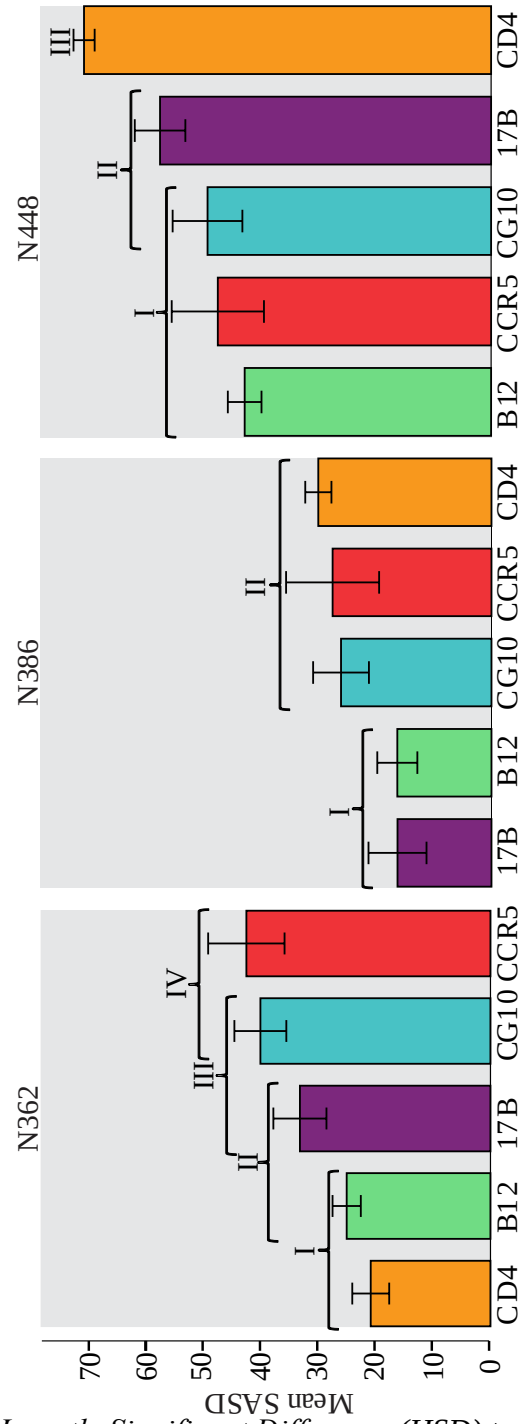


Figure 5. Tukey's Honestly Significant Difference (HSD) test for each PNGS. Tukey's HSD was performed for all PNGS sites since means significantly differed between different binding types. Binding types clustered together under the same roman numeral indicate that mean SASDs did not differ significantly among the clustered types. Alternatively, binding types categorized by different roman numerals did have statistically significantly mean differences between the indicated PNGS and binding types.

Two additional Welch's ANOVAs were carried out to determine if the average SASD among: 1) a PNGS and all binding types significantly differed for any of the PNGSs and 2) a binding type and all PNGSs significantly differed for any of the binding types. In both cases, there was a statistically significant difference between groups ($p < 0.05$, Table 3). The Games-Howell post-hoc tests for multiple comparisons revealed that the PNGSs and binding types could be clustered into seven (Fig. 6) and two (Fig. 7) overlapping groups based on mean SASD from all binding sites or all PNGSs, repectively.

Site Comparison	Levene's Test		Welch's ANOVA				Power
	F	*p*	F	$df_{between}$	df_{within}	*p*	
PNGS	4.9	< 0.005	56.1	8	291.3	< 0.005	1
Binding Type	14.7	< 0.005	13.4	4	223.6	< 0.005	1

Table 3. Levene's test of homogeneity of variances and Welch's ANOVA output to detect mean differences in SASDs. Differences in mean SASDs existed for different PNGSs and binding types.

Next, the SASD between each binding site and the closest PNGS to that binding site was examined (Appendix III, Table 13). Interestingly, only six of the nine PNGSs were among the closest to a binding site (N276, N295, N301, N362, N332 and N386). Four of the sites, N276, N301, N362 and N386, accounted for a noteworthy 96% of the closest PNGSs to a binding site.

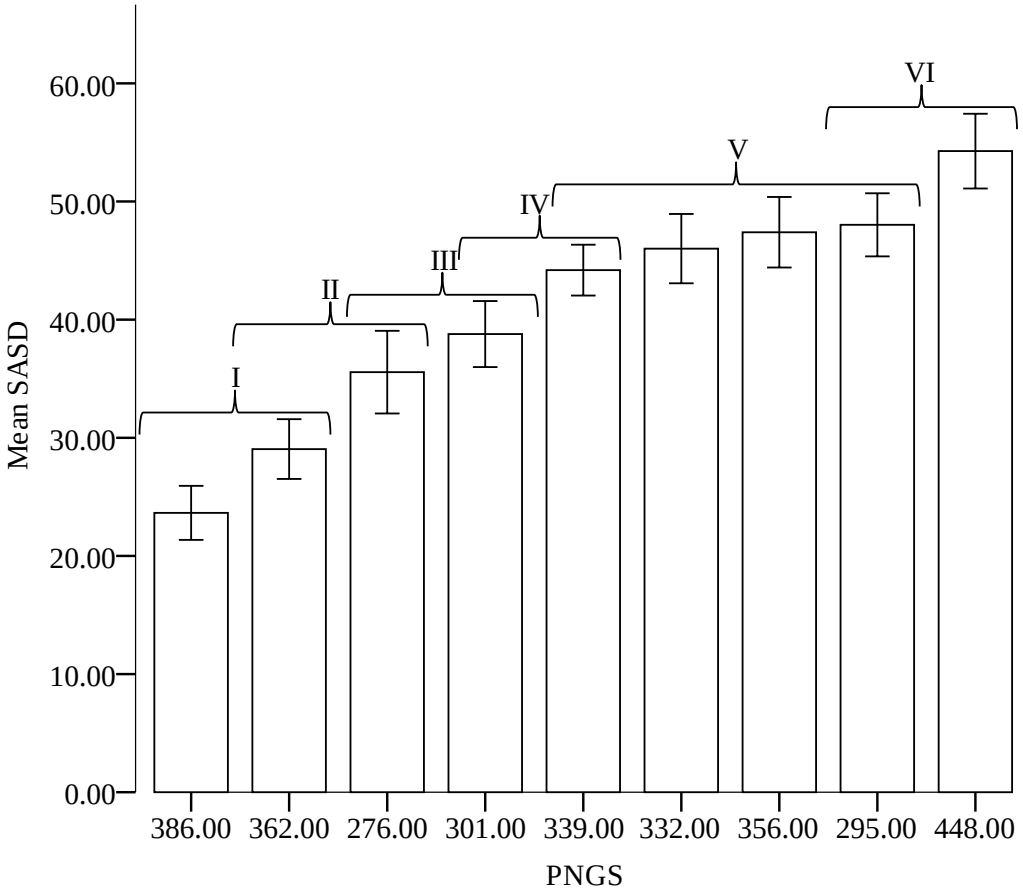


Figure 6. A Games-Howell post-hoc test for multiple comparisons of PNGSs based on SASD. This test was performed to identify which PNGSs had significantly different mean SASDs from all binding sites. PNGSs clustered together under the same roman numeral indicate that mean SASDs did not differ significantly among the clustered sites. Alternatively, PNGSs categorized by different roman numerals did have statistically significantly mean differences between the indicated PNGS(s) at the 0.05 level.

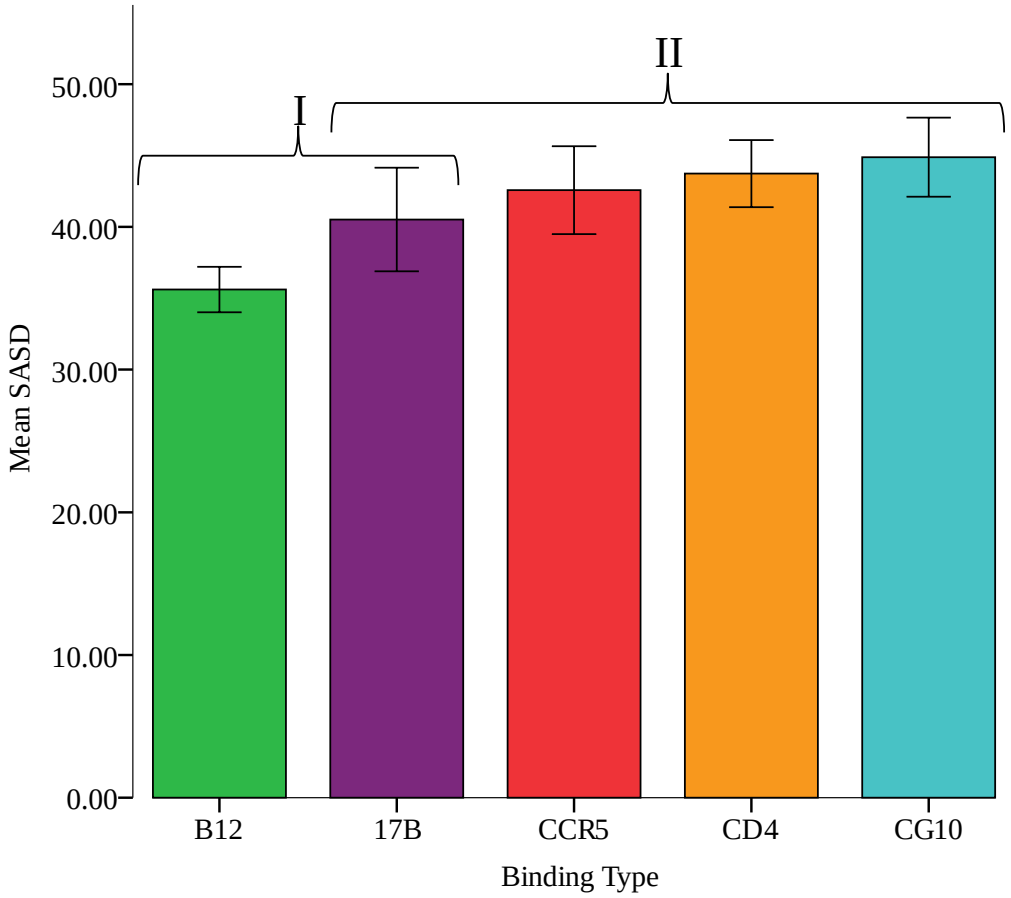


Figure 7. A Games-Howell post-hoc test for multiple comparisons for binding types based on SASD. This test was performed to identify which binding types had significantly different mean SASDs from all PNGSs. Binding types clustered together under the same roman numeral indicate that mean SASDs did not differ significantly among the clustered types. Alternatively, binding types categorized by different roman numerals did have significantly different mean SASDs from PNGSs at the 0.05 level.

Critical Binding Site Flexibility and PNGS proximity

The dynamic flexibility index and solvent accessible surface area were determined for each residue included in chain D of 3TGQ to determine if flexibility or accessibility play a role in the placement of stable PNGSs. A histogram of the plotted %DFI for each residue revealed a left-skewed, or negatively skewed, distribution (Fig. 8 green bars). All stable PNGSs cluster within the rightmost portion of the histogram as their %DFI ranges from 81.3% to 99.8% (Appendix III, Table 5). Alternatively, there doesn't appear to be a preference for any particular level of solvent accessibility for gp120, as residues appear to have uniformly distributed %ASA (Fig. 8, mauve bars).

%DFI and %ASA of binding sites were more closely inspected to identify trends that may indicate a relationship between binding site flexibility or solvent accessibility and proximity to PNGSs. The shortest SASD between each binding site and the closest stable PNGS was plotted against %DFI (Figure 9) or %ASA (Appendix IV, Fig. 2). Remarkably, %DFI and SASD clearly exhibit a strong negative correlation ($r = 0.86$), strongly implying that the most flexible binding sites are situated closest to PNGSs. Alternatively, no relationship beyond what one would expect randomly could be discerned between the shortest SASDs from binding sites to a PNGS ($r = 0.04$).

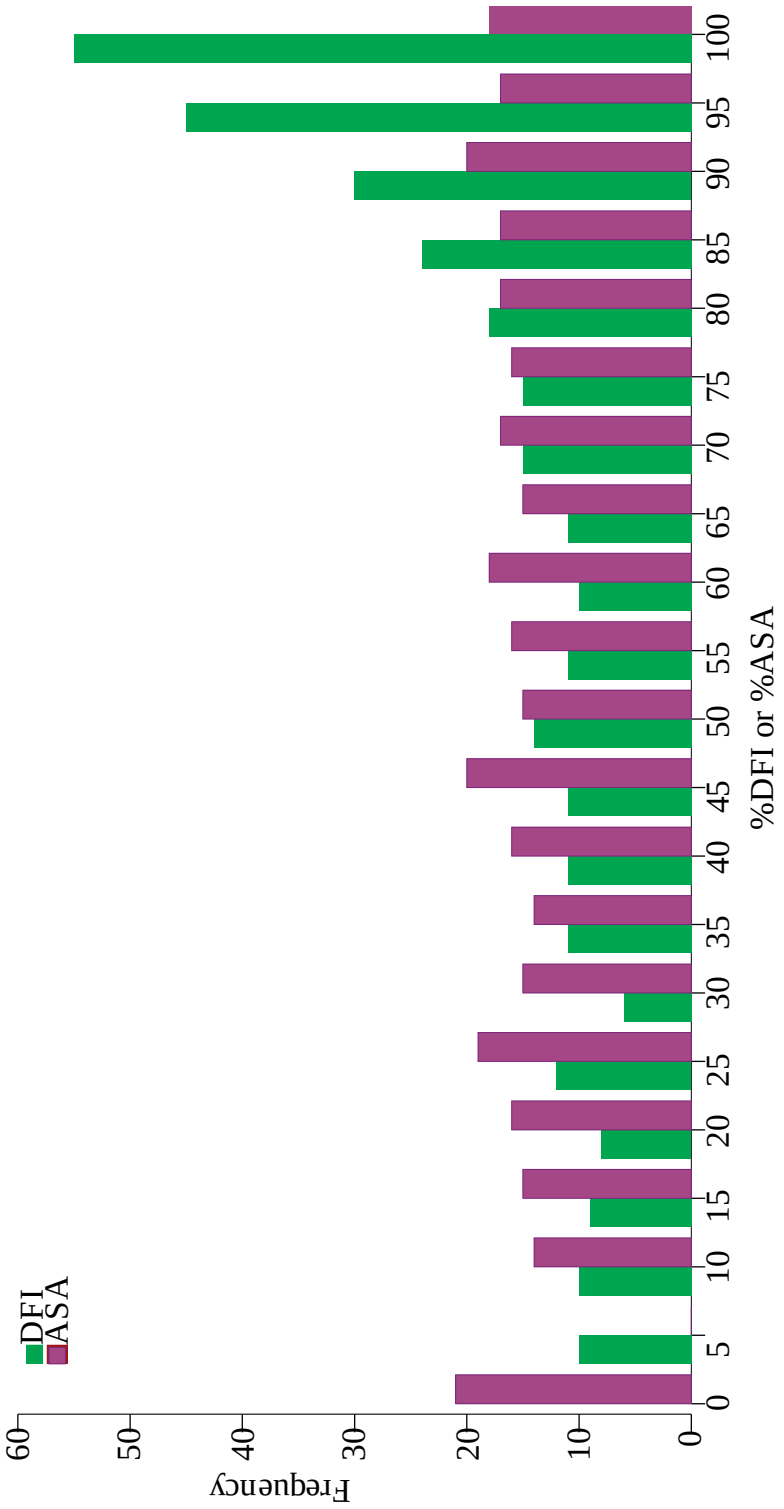


Figure 8. %DFI and %ASA histogram. The frequency of all sites on 3TGQ, chain D, with a given %DFI (green bars) or %ASA (mauve bars) is shown.

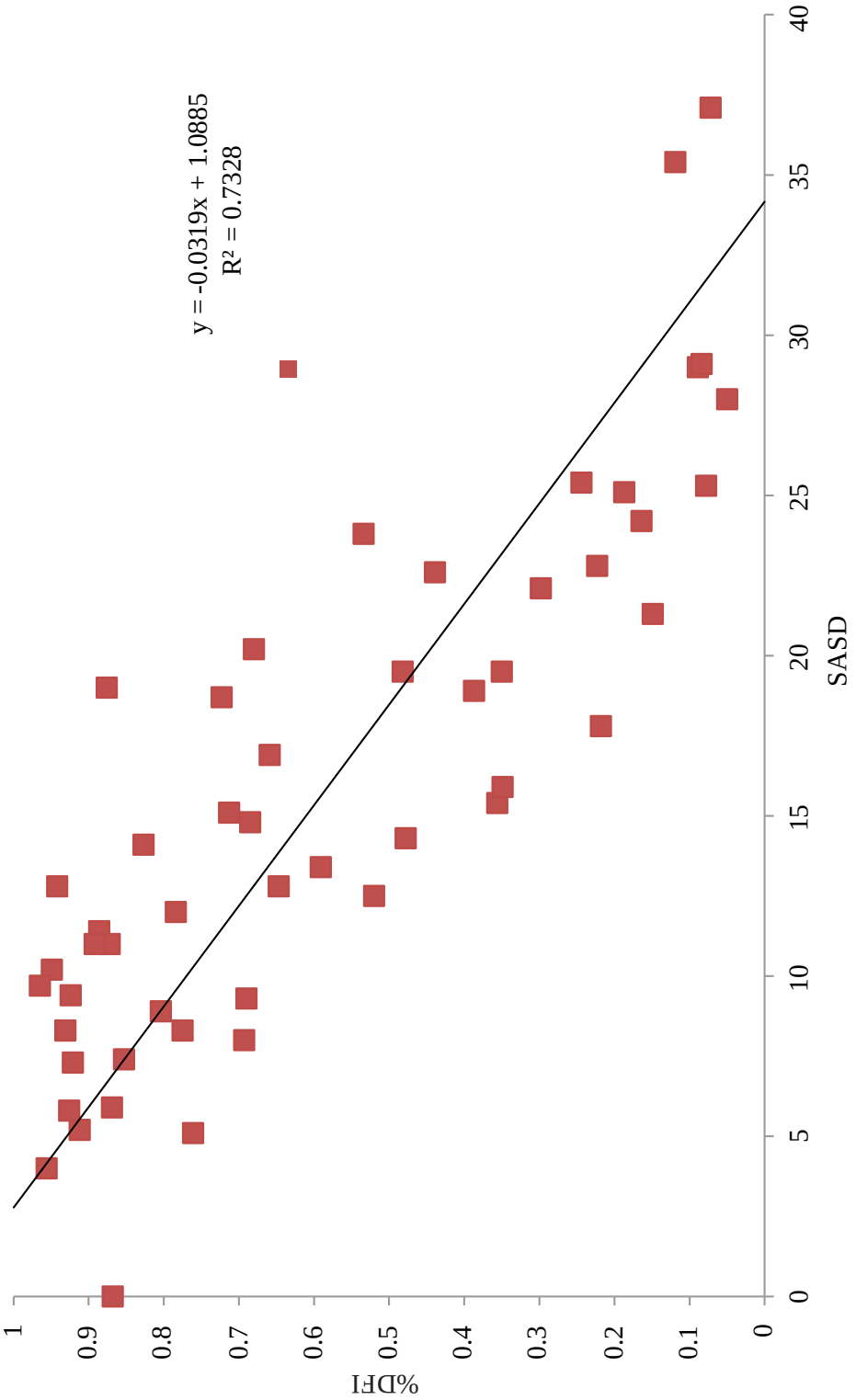


Figure 9. The shortest SASD for each of the 52 binding sites to the closest PNGS plotted against %DFI.

Conclusions

Differential Conservation of PNGSs and Evolutionary Contribution

The gp120 glycoprotein of HIV is a prime target for therapeutic development, given its intimate relationship with the immune system and the fact that it is involved in the initial stage of infection; cellular attachment. Despite the appeal of developing drugs that interfere with the earliest stages of infection, there are fundamental obstacles that must be overcome. Firstly, gp120 experiences an incredibly rapid rate of intra-host evolution, 8.18E-3 substitutions per site per year (Lemey, Rambaut, Pybus 2006), resulting in a robust and diverse quasispecies population. The very crux of a quasispecies is the ability to swiftly respond to natural (e.g. neutralizing antibodies) or artificial (e.g. therapeutics) selective pressures. In addition, and not mutually exclusive from the former point, an evolving glycan shield has been found to offer substantial protection to otherwise accessible regions of gp120 less prone to mutation (Wei et al. 2003). In the current study, stable PNGSs have been systematically characterized in relationship to each other as well as to critical binding sites on gp120.

It has been established previously that the number of PNGSs in gp120 tends increase over time, although the positions they occur at often differ at any given time point (Wei et al. 2003). Despite that fact, many PNGSs are relatively conserved throughout infection (Blay et al. 2006). This study establishes and quantifies, for the first time, the significant effect of PNGSs on the divergence of gp120 in normally progressing HIV-1 subtype B infected individuals. This was accomplished by comparing branch lengths of identical maximum likelihood-generated topologies before and after removal of positions where PNGSs occur. Strikingly, a 59% reduction in mean divergence

(Appendix III, Table 5) is experienced by the PNGS devoid dataset (PGlyRem), indicating that, at least in normally progressing individuals, positions encoding PNGSs are a major source of variation.

Since co-evolutionary toggling (i.e. positional switching) of PNGSs has been recognized (Poon et al. 2007), a goal of this study was to identify whether stable PNGSs tend to be randomly conserved over the entire course of infection or if some tended to be significantly more conserved than others. A randomization analysis was employed to find the probability of a PNGS achieving a given average conservation rank among all 12 PNGSs. Two sites, N301 (average rank of 3) and N448 (average rank of 3.6), were found to be highly and non-randomly conserved in both an intra- and inter-individual context (Figure 3, Appendix III, Table 5). Furthermore, the variability in conservation of the observed PNGSs was significantly higher than what would be expected if the relative conservation of the PNGSs was random (Fig. 4). This implies that PNGSs with the highest and lowest observed ranks are truly outliers.

In order to identify whether or not a structural versus protective role was played by each stable PNGS, comparisons between critical binding sites and PNGSs were carried out. These analyses were accomplished with a metric referred to as the solvent accessible surface distance (SASD) which provides a more accurate estimate of the distance between given sites than popular, yet antiquated, methods of counting amino acids between sites or even calculating the Euclidean distance in the three-dimensional structure (Kahraman, Malmstrom, Aebersold 2011). A crucial assumption of this study is that in order for a glycan attached to a PNGS to convey protection from the immune system, or aid in the binding of a particular site, proximity to the glycan is required.

Given that average length of an N-linked glycan is 30 angstroms, that length is assumed to be a near maximum breadth of protective coverage here (Rudd et al. 1999). In accordance with that length, it should not be surprising then that SASDs calculated for fifty of fifty-two binding sites (excluding T123 and K117) were within 30 angstroms of a stable PNGSs (Appendix III, Table 13). Remarkably, only four of the nine PNGSs included in the structural analyses accounted for the closest PNGSs to 96% of the binding sites. The placement of only four stable PNGSs within a proximal and putatively protective distance from fifty critical binding sites is an extraordinarily clever strategy which may minimize mutational accumulation while maximizing protective potential. Upon visual inspection, it is apparent that these four PNGSs form are situated at the anterior-most boundary of the binding sites mapped here. In addition, fewer glycans near critical binding sites may result in a decrease of steric hindrance of cellular receptors (Fig. 10). While N301 is the most conserved stable PNGS, N362 is actually one of the least conserved. It is possible that in the absence of the PNGS at N362 that PNGSs at N276 and N386, or more transiently occurring PNGSs, could collectively compensate for the loss. Expanded studies focusing on these binding site-proximal PNGSs may reveal co-evolutionary processes in the face of glycan-specific selective pressures. Future mutational analyses should take into account these SASD findings rather than sequence distance to better account for interactions between binding sites and PNGSs.

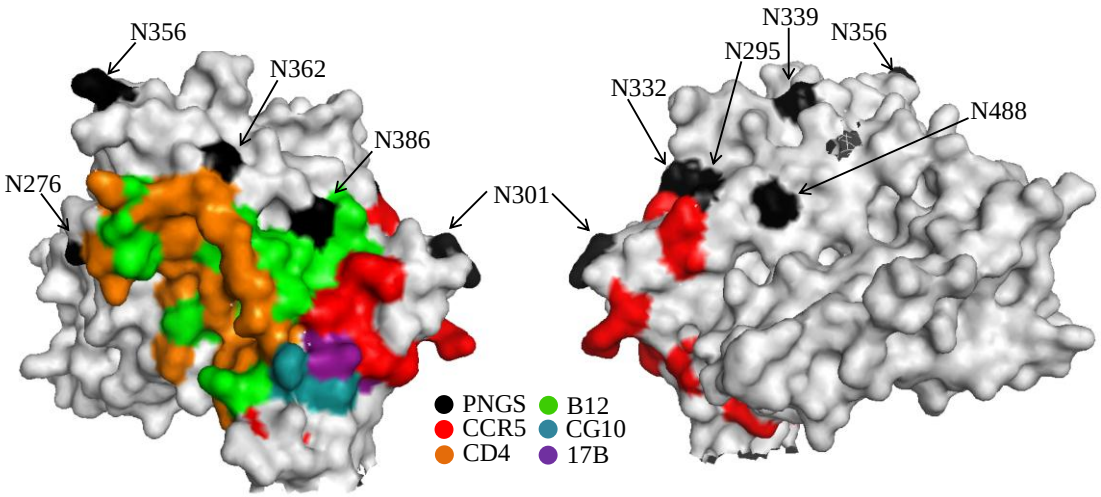


Figure 10. Stable PNGSs and other sites of interest mapped on the 3TGQ protein structure. CG10 binding sites are colored in teal, 17B in purple, CCR5 in red, and the PNGSs are colored in black. The right structure is rotated around the y-axis to show all stable PNGSs and binding sites.

Included in the four PNGSs that were found to be closest to the most binding sites, is N301, which I also determined to be significantly more conserved than eleven other PNGSs (Fig. 3). In a prior mutational analysis, where the asparagine at position 301 was conservatively mutated to a glutamine, it was revealed that the mutant virus was unable to replicate in PBMCs (Ogert et al. 2001). Interestingly, N301 was only conserved in 64% of sequences sampled from individual 7 (analyzed here) despite a persistent viral load (Shankarappa et al. 1999). SLAC results indicate that variants harboring non-synonymous mutations to aspartic acid, glutamic acid, glycine, lysine and threonine were positively selected in favor of variants with asparagine in this individual (Appenidix IV, Fig. 3). Since the viral population sampled from individual 7 was collectively dual-tropic (able to use either co-receptor) from nearly the onset of infection (Shankarappa et al. 1999), it is possible that the maintenance of viral load was due to a co-receptor switch.

In addition to the conservation analyses, N301 was found to be most closely situated with and equidistantly placed in proximity to CCR5, 17B and CG10 binding sites (Figs. 5, 10 and 11) which do have some overlap. CCR5 is considered to be the primary co-receptor, as most strains utilize CCR5 at the onset of infection (Coakley, Petropoulos, Whitcomb 2005). 17B and CG10 are CD4 induced antibodies, as binding is enhanced in the presence of soluble CD4, indicating that a conformational shift exposes their binding sites after CD4 is attached (Zhang et al. 1999). These results, taken together with the high conservation of N301 and its close proximity to binding sites suggest this PNGS is both critical for co-receptor binding activity and may play protective role against 17B and CG10.

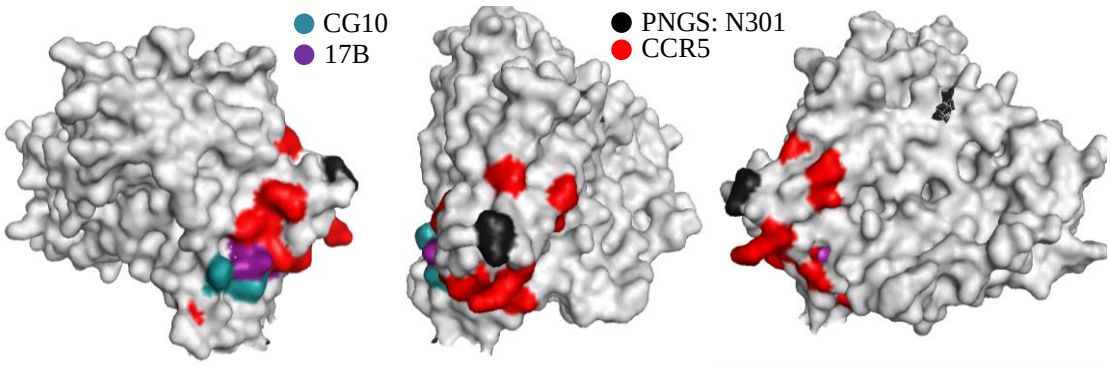


Figure 11. N301 and other sites of interest mapped on the 3TGQ protein structure. CG10 binding sites are colored in teal, 17B in purple, CCR5 in red, and the PNGS site, N301, is colored in black. The structure is rotated around the y-axis to show the binding sites centered around N301.

A second PNGS, N448, was also found to be significantly more conserved than nine other stable PNGSs. In contrast to the SASD results seen with N301, N448 is located farthest from all critical binding sites analyzed here (Figs. 5 and 10 and Appendix III, Table 12). Interestingly, a study testing the neutralizing activity of a prokaryotic lectin, actinohivin, also examined PNGS deletion in response to the selective pressure.

Actinohivin preferentially binds to high mannose sugars, which N448 may be occupied by, rather than complex sugars. Despite persistent sub-cultivations for actinohivin-resistant strains of HIV-1 subtype B infected CEM T-cells, with escalating concentrations of actinohivin, N448 persisted (Hoorelbeke et al. 2010). Furthermore, N448 is located directly proximal to another PNGS, N262 (Fig. 12), which was not included in this study due to lack of sequence coverage. Deletion of N262 results in a significantly hindered ability to attach to CD4, due to lowered overall expression of gp120 in the viral particle (Francois, Balzarini 2011). Given the proximity of N448 to N262, along with the fact that a non-synonymous mutation was not observed in the face of a strong selective pressure against this type of PNGS, and that it is more than 50 angstroms away from any of the binding sites, it is likely that the highly conserved nature of this PNGS is not to maintain immunological protection, but may instead play a role in structural integrity.

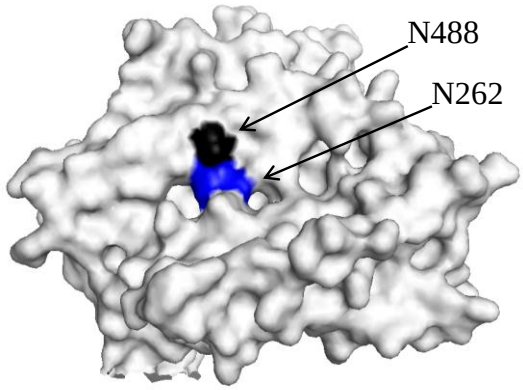


Figure 12. N262 (blue) and N448 (black) mapped on the 3TGQ protein structure.

As a final attempt to determine characteristics of binding sites that PNGSs cluster closest to, the dynamic flexibility index (DFI) was calculated as a comparative percentage (to the rest of the molecule) for each binding site. I found that binding sites falling closer to a PNGS were more flexible than those clustering farther away (Fig. 9).

This presents two possibilities. First, it is possible that flexible binding sites are more susceptible to neutralization than there rigid counterparts. Alternatively, it may be that PNGSs preferentially occur in areas where the surrounding amino acids are more flexible, possibly to allow for the attachment of a large glycan moiety. Future studies will include the elucidation of DFI for all the amino acids immediately surrounding the PNGSs.

Methods

Sequence Selection

Previously published HIV-1 subtype B partial *env* nucleotide sequences, longitudinally sampled from peripheral blood mononuclear cells (PBMC) and/or plasma from eleven infected individuals (Shankarappa et al. 1999), were collected from the Los Alamos National Laboratories HIV database (http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html). In addition, descriptive information for each sequence was also collected, and included sampled compartment, time since seroconversion, CD4 count, viral load and treatment history. Sequences were codon aligned in MEGA 5.10 (Tamura et al. 2011) using MUSCLE (Edgar 2004a; Edgar 2004b).

Datasets

Sequences were organized according to the statistical test being performed (Table 4). For each individual, the following datasets were created: 1) A paired compartment dataset where time points were only represented if sequences from both PBMC and plasma compartments were present, 2) A PNGS-void dataset (PGlyRem) where any position in

an alignment was removed if a PNGS sequon was present for any sequence and 3) A full dataset (ALL) where all sequence with all positions were included.

Individual	Available		Paired	PGlyRem	ALL
	PBMC	Plasma			
1	87	50	NA	137	137
2	133	150	193	283	283
3	106	24	NA	130	130
4	Mixed		NA	250	250
5	191	44	138	235	235
6	97	32	59	129	129
7	107	100	63	207	207
8	119	82	57	201	201
9	121	16	NA	137	137
10	Mixed		NA	202	202
11	0	52	NA	52	52

Table 4. Count of sequences in each dataset. For some individuals, sequences originating from both compartments were not available, resulting in the individual being non-applicable for the paired dataset.

Selection Detection

The Single-Likelihood Ancestor Counting (SLAC) method was use to identify sites under operating under selective regimes other than neutral (Kosakovsky Pond, Frost 2005). A p-value of 0.1 was used to ascertain significance, as SLAC has been determined to be a conservative method that returns fewer false positive than one would expect at a p-value of 0.05. A general reversible model was used to account for codon substitutions.

Phylogenetic Trees

Maximum likelihood trees were reconstructed for the PGlyRem and ALL datasets in MEGA 5.1. The best fitting model for the *env* datasets is the General Time Reversible (GTR) + invariant sites + Γ model as determined by the model selection tool in MEGA.

The Γ parameter was calculated for each individual's entire collection of sequences, also in MEGA. For each of the eleven ALL datasets, a maximum likelihood phylogenetic tree was reconstructed in MEGA and the Newick files with branch lengths were saved. The topology of each maximum likelihood tree was retained and taxa from PGlyRem datasets were mapped to the branches. New branch lengths were calculated for each tree through the Analyze User Tree function in MEGA. This effectively produced two trees with identical topologies where branch lengths differ only due to the presence (ALL) or absence (PGlyRem) of PNGS. For each tree, an outgroup consisting of sequences sampled most closely to seroconversion was chosen and branches were reorganized. For paired trees, the same outgroup was always chosen.

Phylogenetic trees reconstructed from the PGlyRem and ALL datasets were uploaded to TreeRate at http://www.hiv.lanl.gov/content/sequence/TREERATEv2/treerate.html. The outgroup rooting method was used to extract the distance from the rootmost node to each tip as well as the sum of branch lengths for each tree (Appendix III Table 1) (Maljkovic Berry et al. 2007; Maljkovic Berry et al. 2009). The sum of branch lengths was normalized to the number of sequences in each dataset.

Glycosylation Site Identification

Glycosylation sequons (NX[ST]) were identified within each dataset using the N-Glycosite tool at http://www.hiv.lanl.gov/content/sequence/GLYCOSITE/glycosite.html (Zhang et al. 2004). Default parameters were used, which includes disregarding any sequon with a proline in the central (X) position. A raw conservation metric was defined by counting the total number of predicted glycosylation sites at any given position and

dividing that count by the total number of sequences in the dataset. As an example, for one PNGS position in a dataset composed of 10 sequences, if 5 sequences have the sequon NQS and the other 5 have NST, the % raw conservation is 100%. Only stable glycosylation sites (Appendix III Table 5), where greater than 60% of the sequences in four out of the five individuals, were retained for statistical testing of site-specific conservation.

Stable sites were ranked by % raw conservation, for each individual due to the issue of dependence where multiple sites are nested within an individual. Ties between sites where % raw conservation is equal in an individual were broken by determining the underlying nucleotide genetic distance at the sequon positions using a JC69 substitution model. This model was chosen in favor of those with more parameters to avoid over-fitting, as the number of nucleotide sites is small per sequence (n = 9). Those sites with a shorter genetic distance (i.e. more conserved) were given a lower rank (Appendix Table 1).

Structural Analysis

The unliganded HIV-1 subtype B gp120 structure, 3TGQ (chain D) (Kwon et al. 2012), was obtained from the Protein Data Bank. Binding sites on gp120 were determined from a literature search for the cellular receptors, CD4 (Wu et al. 2009) and CCR5 (Rizzuto et al. 1998), and also the neutralizing antibodies, B12 (Wu et al. 2009), 17B and CG10 (Rizzuto et al. 1998). The solvent accessible surface distance (SASD) and Euclidean distance between the beta carbon of the asparagine for each stable PNGS and the beta carbon of each binding site were determined with XWalk (Kahraman, Malmstrom,

Aebersold 2011) (Appendix III Table 12) and graphical representations were generated with The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.

Quantification of the structural fluctuation of each protein residue within 3TGQ was carried out by using the dynamic flexibility index (Nevin Gerek, Kumar, Banu Ozkan 2013) (Appendix III Table 14). Accessible surface area was calculated using Surface Racer (Tsodikov, Record, Sergeev 2002).

Statistical Tests

For the comparison of sum of phylogenetic tree branch lengths in the PGlyRem versus ALL dataset, both paired t-test and regression analyses were carried out in an attempt to identify whether or not there is an effect of PNGS on divergence. Specifically, both tests were used here to determine if the differences between branch lengths reconstructed from the two datasets were statistically significant. Testing was carried out in SPSS version 21.

Slope comparison was carried out using the slope of the line where the sum of all branch lengths for each individual in the PGlyRem dataset was regressed on those for each individual in the ALL dataset. To compare the observed slope versus expected slope, where $m_{observed} = 0.3433$ and $m_{expected} = 1$, a Student's t-test was conducted in EXCEL.

$$t = \frac{\beta_{estimated} - \beta_{hypothetical}}{s_b}$$

$\beta_{estimated}$, or the parameter value estimate, is equivalent to the observed slope for the regression line for the total sum of branch lengths of PGlyRem to ALL. $\beta_{hypothetical}$, or the hypothesized parameter value, is the slope that is expected if there is no change in the

sum of branch lengths after PNGS are removed. S_b is the standard error of the parameter value estimate and was found to be 0.066.

Longitudinal analyses were executed on the paired compartments datasets to assess whether or not there is an effect of patient, compartment or site on the % raw conservation of PNGS. The generalized estimating equations (GEE) procedure, implemented in SPSS version 21, assuming an unstructured correlation matrix was used to account for the lack of independence for sites nested within each sequence, sequences nested within each time point, and time points nested within each individual. Rather than model the within-subject covariance structure, the GEE procedure treats it as a nuisance parameter and the mean response is modeled instead.

A randomization analysis was carried out on the paired compartments aligned datasets to test the hypothesis that certain stable glycosylation sites are conserved more than others versus the null hypothesis that all glycosylation sites are randomly conserved. Briefly, the number of sites to be scrutinized (12) and the number of individuals we had observations for (5) were used as input. The 12 ranks were shuffled for each individual and the average rank for all individuals was calculated 9,999 times. This resulted in 119,988 average ranks (9999 for each site). The 12 observed average ranks were added to yield a total of 120,000 average ranks.

A one-way analysis of variance (ANOVA) or Welch's ANOVA (in the case of heteroscedasticity) was performed for each stable PNGS, where SASDs were grouped by type of binding site (17B, B12, CCR5, CD4 or CG10), to identify whether or not there was an effect of binding site type on the distances. If the null hypothesis of homogeneity of variances was rejected by way of Levene's test (Table 3), the non-parametric Welch's

ANOVA was used. If between group differences were found to be significant, ANOVAs were followed up with the Tukey's Honestly Significant Difference (Tukey's HSD) post hoc test for multiple comparisons while Welch's ANOVAs were followed up with both Tukey's HSD and the Games-Howell test for multiple comparisons. One important note here is that the few outliers present in the data were not removed as the binding sites are the true binding sites rather than samples to reflect the binding site population.

References

Blay, WM, S Gnanakaran, B Foley, NA Doria-Rose, BT Korber, NL Haigwood. 2006. Consistent patterns of change during the divergence of human immunodeficiency virus type 1 envelope from that of the inoculated virus in simian/human immunodeficiency virus-infected macaques. J Virol 80:999-1014.

Coakley, E, CJ Petropoulos, JM Whitcomb. 2005. Assessing chemokine co-receptor usage in HIV. Curr Opin Infect Dis 18:9-15.

Edgar, RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edgar, RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Francois, KO, J Balzarini. 2011. The highly conserved glycan at asparagine 260 of HIV-1 gp120 is indispensable for viral entry. J Biol Chem 286:42900-42910.

Geyer, H, C Holschbach, G Hunsmann, J Schneider. 1988. Carbohydrates of human immunodeficiency virus. Structures of oligosaccharides linked to the envelope glycoprotein 120. J Biol Chem 263:11760-11767.

Hoorelbeke, B, D Huskens, G Ferir, KO Francois, A Takahashi, K Van Laethem, D Schols, H Tanaka, J Balzarini. 2010. Actinohivin, a Broadly Neutralizing Prokaryotic Lectin, Inhibits HIV-1 Infection by Specifically Targeting High-Mannose-Type Glycans on the gp120 Envelope. Antimicrobial Agents and Chemotherapy 54:3287-3301.

Kahraman, A, L Malmstrom, R Aebersold. 2011. Xwalk: computing and visualizing distances in cross-linking experiments. Bioinformatics 27:2163-2164.

Kosakovsky Pond, SL, SD Frost. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208-1222.

Kwon, YD, A Finzi, X Wu, et al. 2012. Unliganded HIV-1 gp120 core structures assume the CD4-bound conformation with regulation by quaternary interactions and variable loops. Proc Natl Acad Sci U S A 109:5663-5668.

Lasky, LA, JE Groopman, CW Fennie, PM Benz, DJ Capon, DJ Dowbenko, GR Nakamura, WM Nunes, ME Renz, PW Berman. 1986. Neutralization of the AIDS retrovirus by antibodies to a recombinant envelope glycoprotein. Science 233:209-212.

Lemey, P, A Rambaut, OG Pybus. 2006. HIV evolutionary dynamics within and among hosts. AIDS Rev 8:125-140.

Li, Y, L Luo, N Rasool, CY Kang. 1993. Glycosylation is necessary for the correct folding of human immunodeficiency virus gp120 in CD4 binding. J Virol 67:584-588.

Lodish, H, A Berk, S Zipursky, P Matsudaira, D Baltimore, J Darnell. 2000. Protein Glycosylation in the ER and Golgi Complex. Molecular Cell Biology. New York: Freeman W.H.

Maljkovic Berry, I, G Athreya, M Kothari, M Daniels, WJ Bruno, B Korber, C Kuiken, RM Ribeiro, T Leitner. 2009. The evolutionary rate dynamically tracks changes in HIV-1 epidemics: application of a simple method for optimizing the evolutionary rate in phylogenetic trees with longitudinal data. Epidemics 1:230-239.

Maljkovic Berry, I, R Ribeiro, M Kothari, G Athreya, M Daniels, HY Lee, W Bruno, T Leitner. 2007. Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. J Virol 81:10625-10635.

Naeger, LK, KA Struble, JS Murray, DB Birnkrant. 2010. Running a tightrope: regulatory challenges in the development of antiretrovirals. Antiviral Res 85:232-240.

Nevin Gerek, Z, S Kumar, S Banu Ozkan. 2013. Structural dynamics flexibility informs function and evolution at a proteome scale. Evolutionary Applications 6:423-433.

Ogert, RA, MK Lee, W Ross, A Buckler-White, MA Martin, MW Cho. 2001. N-linked glycosylation sites adjacent to and within the V1/V2 and the V3 loops of dualtropic human immunodeficiency virus type 1 isolate DH12 gp120 affect coreceptor usage and cellular tropism. J Virol 75:5998-6006.

Poon, AF, FI Lewis, SL Pond, SD Frost. 2007. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. PLoS Comput Biol 3:e11.

Rizzuto, CD, R Wyatt, N Hernandez-Ramos, Y Sun, PD Kwong, WA Hendrickson, J Sodroski. 1998. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science 280:1949-1953.

Rudd, PM, MR Wormald, RL Stanfield, M Huang, N Mattsson, JA Speir, JA DiGennaro, JS Fetrow, RA Dwek, IA Wilson. 1999. Roles for glycosylation of cell surface receptors involved in cellular immune recognition. J Mol Biol 293:351-366.

Shankarappa, R, JB Margolick, SJ Gange, et al. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73:10489-10502.

Tamura, K, D Peterson, N Peterson, G Stecher, M Nei, S Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739.

Tsodikov, OV, MT Record, Jr., YV Sergeev. 2002. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. J Comput Chem 23:600-609.

Wei, X, JM Decker, S Wang, et al. 2003. Antibody neutralization and escape by HIV-1. Nature 422:307-312.

Wu, X, T Zhou, S O'Dell, RT Wyatt, PD Kwong, JR Mascola. 2009. Mechanism of human immunodeficiency virus type 1 resistance to monoclonal antibody B12 that effectively targets the site of CD4 attachment. J Virol 83:10892-10907.

Zhang, M, B Gaschen, W Blay, B Foley, N Haigwood, C Kuiken, B Korber. 2004. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. Glycobiology 14:1229-1246.

Zhang, W, G Canziani, C Plugariu, R Wyatt, J Sodroski, R Sweet, P Kwong, W Hendrickson, I Chaiken. 1999. Conformational changes of gp120 in epitopes near the CCR5 binding site are induced by CD4 and a CD4 miniprotein mimetic. Biochemistry 38:9405-9416.

CHAPTER 4

CONCLUSIONS AND FUTURE DIRECTIONS

The human immunodeficiency virus (HIV) has proven to be a resilient opponent in the arms race against the host immune system, as it appears to have efficiently perfected its strategy to continually infect host cells despite a long lived dynamic host response. An analogy for such an arms race was provided by Dawkins and Krebs (1979), using the predator-prey relationship of the fox and rabbit. Individually, a fox chases a rabbit, and the fox may catch the rabbit, or the rabbit may escape. Over the evolutionary timescale, this exercise occurs many times, and while the fox lineage may accrue adaptations for hunting, the rabbit lineage may also evolve adaptive mechanisms for escape. This biological arms race, where both lineages are faced with the ultimate selective pressure of survival, results in antagonistic coevolution. The interplay between HIV and the infected host is similar to the fox and rabbit paradigm; the host immune system is in constant pursuit of the viral population, and while both are continually changing, they also equivalently persist to a point. Normally progressing hosts typically have robust T-cell and antibody responses, constantly adjusting to the viral variants at a given time. This is apparent in the ability of contemporaneous sera sampled from the infected person to neutralize contemporaneous and earlier viral variants, but not viral isolates from later time points (Bunnik et al. 2008). Likewise, the viral quasispecies is constantly shifting towards the most-fit variants given the host selective pressures (Kohler, Goudsmit, Nara 1992; Bunnik et al. 2008). The predator and prey continue the chase until the host immune system fails, typically after 10 years in drug-naïve normally progressing individuals (Coffin 1995).

The driving force of HIV's successful campaign against the immune system is primarily the low-fidelity of reverse transcriptase (Preston, Poiesz, Loeb 1988), which is the source of both recombination (Hu, Temin 1990; Jetzt et al. 2000) and mutation (Coffin 1995). While each of these mechanisms plays a role in the viral population's evolutionary dynamics within and among individuals, the conclusions made from each type of study are often quite different.

Results from cross-sectional studies, where viral strains are sampled from multiple infected individuals over temporal or geographical space, may reveal the timing and location of epidemic origins as well as the spread (Lemey, Rambaut, Pybus 2006). This approach has informed us of myriad historical periods in the HIV evolutionary timeline, including, 1) multiple cross-species transmissions have introduced HIV into humans (Hahn et al. 2000), and 2) that HIV likely originated in west equatorial Africa, which is the only place in the world where HIV-1 groups M, N and O co-circulate (Nkengasong et al. 1994; Delaporte et al. 1996; Takehisa et al. 1998). In addition, it was determined that chimpanzees living in the same region are infected with close relatives to HIV (Gao et al. 1999; Hahn et al. 2000). Inter-individual phylogenies have also been employed for public health purposes, such as solving criminal cases (Machuca et al. 2001; Metzker et al. 2002; Lemey et al. 2005; Scaduto et al. 2010; van der Kuyl et al. 2011) or to identify susceptible individuals and develop intervention strategies (Lewis et al. 2008; Dennis et al. 2012).

In the second chapter of this dissertation, I performed a cross-sectional analysis to better understand Haiti's role in the HIV-1 subtype B pandemic. Previous analyses published phylogenies demonstrating that Haiti-originating strains of HIV clustered

ancestrally to subtype B strains originating elsewhere across the globe (Li, Tanimura, Sharp 1988; Gojobori et al. 1990; Gilbert et al. 2007). While each of these studies incorporated appropriate substitution and site rate models to model mutation, they disregarded the possibility of recombination. A considerable amount of controversy has been centered on Haiti with regards to the subtype B pandemic, starting when Haitians were recognized as a risk group (CDC 1982) and refueled when Haiti was labeled as the jumping point for subtype B into the rest of the world (Gilbert et al. 2007).

The goal of this project was to first identify if recombination could be detected in the Gilbert (2007) dataset, and then to determine if phylogenies reconstructed from putatively non-recombinant segments implied an evolutionary history different from those previously proposed. The intent of this study was not to impact HIV from a medical perspective, but rather to more accurately inform the historical record by using all possible tools to model the evolutionary history.

Recombination detection within HIV subtypes is a difficult task in comparison to inter-subtype recombination, but is becoming more feasible with increasing computational power. As stated previously, the parents of the recombinant sequence must be divergent enough such that recombination can be detected in the viral progeny. Regardless, recombination breakpoints were identified with both GARD and RDP. Phylogenetic reconstruction of the three putatively non-recombinant segments identified in GARD does not support a Haiti-first model of subtype B dispersal. Instead, the two longest segments, which are displaced by an intervening segment, support near-simultaneous entries of subtype B into Haiti and the rest of the world. A secondary phylogenetic analysis, after the removal of cross-clade recombinants and concatenation

of the two largest segments, again revealed a topology indicative of simultaneous entrances into Haiti and the rest of the world. While the posterior probabilities on the Caribbean cluster are lower than the confidence intervals obtained in Gilbert (2007), it is encouraging that topologies from the longest two segments, individually and combined, support the simultaneous entrances of subtype B into Haiti and the rest of the world.

Given the temporal and geographic breadth of sequences included in the described cross-sectional study, it is likely that the datasets used here and previously are highly representative of all publicly available subtype B sequences. Unfortunately, despite the fact that there are early documented cases of HIV in Belgian individuals who were residing (either permanently or temporarily) in the area where HIV almost certainly entered the human population (Sonnet et al. 1987), there are no representative sequences available. The addition of some of these and other earlier strains to the currently constructed datasets would likely increase the accuracy of the historical record.

Longitudinal studies, where the viral population from the same individual(s) is studied temporally, are useful for understanding host-pathogen coevolutionary dynamics. For example, administration of therapeutics targeting a particular region of HIV results in an increase of escape mutations within an infected individual, where drug-resistant variants can be observed within 14 days (Wei et al. 1995). Similarly, variants in the HIV population, harboring neutralizing antibody escape mutations, tend to increase in frequency in response to the particular neutralizing antibody elicited at a particular time point (Albert et al. 1990). Interestingly, while intuition would lead us to suspect that resistance-conferring mutations would occur at the antibody binding sites, a longitudinal

study revealed that the majority of mutations actually occur at potential N-linked glycosylation sites (PNGSs) (Wei et al. 2003).

The third chapter of this dissertation is based on a longitudinal analysis where sequences sampled from 11 subtype B-infected individuals over a period of 6-12 years were analyzed in an attempt to functionally characterize viral PNGSs (Shankarappa et al. 1999). PNGSs are the most evolutionarily dynamic sites within one of the most evolutionarily dynamic proteins identified to date (Wei et al. 2003). The attachment of carbohydrates to the PNGSs allows the virus to go undetected by antibodies specific for particular protein targets on the gp120 surface. After identifying PNGSs that are highly conserved on gp120, I estimated their solvent accessible surface distance to critical receptor and antibody binding sites. Customarily, site-directed mutagenesis targets, to identify sites of importance, are highly based on the sequence distance between sites. While it may not be the intent, this practice intrinsically and incorrectly assumes a positive correlation between the distance between sites within the sequence and the distance between the same sites on the protein surface. For example, two PNGSs, N262 and N448, are 226 amino acids away from each other, yet they are positioned directly next to each other on the protein structure. Given that N262 appears to be indispensible for replication (Francois, Balzarini 2011), coupled with the lack of proposed functionality for N448 despite its high conservation in the presence of strong selective pressures (Hoorelbeke et al. 2010), it may be worthwhile to investigate the interplay between the two sites. Is N448 also important for replication? Does N448 play a protective role for N262? These are the types of questions one can take into account when considering the spatial proximity of PNGSs. The same type of consideration should be given to the four

PNGSs that run along the anterior border of the critical binding sites on gp120. Can PNGSs N386 and N276 compensate when the viral variant is missing the more centrally positioned yet less conserved N362? Is the highly conserved PNGS N301, which is centrally located between all CCR5 binding sites, decreasingly conserved in infected individuals where a high frequency of viral variants have undergone a co-receptor switch to CXCR4? Could decreased frequency of N301 be used as an early indicator that the viral population within an individual will soon undergo a co-receptor switch, resulting in resistance to the CCR5-antagonist they are taking? These are all questions that can be asked when taking into account the co-evolutionary dynamics between the host immune system and viral PNGSs.

It is apparent from earlier research as well as the body of work presented here, that multiple evolution-driven approaches must be taken to understand what makes HIV "tick." The cross-sectional analysis contributes to the socio-economic and historical perspectives of HIV while the longitudinal analysis offers some direction to medical research. As such, the work performed over the course of this dissertation further validates the potential of evolutionary analysis to investigate a complex study system.

References

Albert, J, B Abrahamsson, K Nagy, E Aurelius, H Gaines, G Nystrom, EM Fenyo. 1990. Rapid Development of Isolate-Specific Neutralizing Antibodies after Primary Hiv-1 Infection and Consequent Emergence of Virus Variants Which Resist Neutralization by Autologous Sera. Aids 4:107-112.

Bunnik, EM, L Pisas, AC van Nuenen, H Schuitemaker. 2008. Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. Journal of Virology 82:7932-7941.

CDC. 1982. Opportunistic Infections and Kaposi's Sarcoma among Haitians in the United States. Morbidity and Mortality Weekly Report 31:353-354, 360-361.

Coffin, JM. 1995. Hiv Population-Dynamics in-Vivo - Implications for Genetic-Variation, Pathogenesis, and Therapy. Science 267:483-489.

Dawkins, R, JR Krebs. 1979. Arms Races between and within Species. Proceedings of the Royal Society B-Biological Sciences 205:489-511.

Delaporte, E, W Janssens, M Peeters, et al. 1996. Epidemiological and molecular characteristics of HIV infection in Gabon, 1986-1994. Aids 10:903-910.

Dennis, AM, S Hue, CB Hurt, S Napravnik, J Sebastian, D Pillay, JJ Eron. 2012. Phylogenetic insights into regional HIV transmission. Aids 26:1813-1822.

Francois, KO, J Balzarini. 2011. The highly conserved glycan at asparagine 260 of HIV-1 gp120 is indispensable for viral entry. J Biol Chem 286:42900-42910.

Gao, F, E Bailes, DL Robertson, et al. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 397:436-441.

Gilbert, MT, A Rambaut, G Wlasiuk, TJ Spira, AE Pitchenik, M Worobey. 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A 104:18566-18570.

Gojobori, T, EN Moriyama, Y Ina, K Ikeo, T Miura, H Tsujimoto, M Hayami, S Yokoyama. 1990. Evolutionary origin of human and simian immunodeficiency viruses. Proc Natl Acad Sci U S A 87:4108-4111.

Hahn, BH, GM Shaw, KM De Cock, PM Sharp. 2000. AIDS - AIDS as a zoonosis: Scientific and public health implications. Science 287:607-614.

Hoorelbeke, B, D Huskens, G Ferir, KO Francois, A Takahashi, K Van Laethem, D Schols, H Tanaka, J Balzarini. 2010. Actinohivin, a Broadly Neutralizing Prokaryotic Lectin, Inhibits HIV-1 Infection by Specifically Targeting High-

Mannose-Type Glycans on the gp120 Envelope. Antimicrobial Agents and Chemotherapy 54:3287-3301.

Hu, WS, HM Temin. 1990. Retroviral Recombination and Reverse Transcription. Science 250:1227-1233.

Jetzt, AE, H Yu, GJ Klarmann, Y Ron, BD Preston, JP Dougherty. 2000. High rate of recombination throughout the human immunodeficiency virus type 1 genome. Journal of Virology 74:1234-1240.

Kohler, H, J Goudsmit, P Nara. 1992. Clonal Dominance - Cause for a Limited and Failing Immune-Response to Hiv-1 Infection and Vaccination. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 5:1158-1168.

Lemey, P, A Rambaut, OG Pybus. 2006. HIV evolutionary dynamics within and among hosts. AIDS Rev 8:125-140.

Lemey, P, S Van Dooren, K Van Laethem, Y Schrooten, I Derdehnckx, P Goubau, F Brun-Vezinet, D Vaira, AM Vandamme. 2005. Molecular testing of multiple HIV-1 transmissions in a criminal case. Aids 19:1649-1658.

Lewis, F, GJ Hughes, A Rambaut, A Pozniak, AJ Leigh Brown. 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med 5:e50.

Li, WH, M Tanimura, PM Sharp. 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 5:313-330.

Machuca, R, LB Jorgensen, P Theilade, C Nielsen. 2001. Molecular investigation of transmission of human immunodeficiency virus type 1 in a criminal case. Clinical and Diagnostic Laboratory Immunology 8:884-890.

Metzker, ML, DP Mindell, XM Liu, RG Ptak, RA Gibbs, DM Hillis. 2002. Molecular evidence of HIV-1 transmission in a criminal case. Proceedings of the National Academy of Sciences of the United States of America 99:14292-14297.

Nkengasong, JN, W Janssens, L Heyndrickx, et al. 1994. Genotypic Subtypes of Hiv-1 in Cameroon. Aids 8:1405-1412.

Preston, BD, BJ Poiesz, LA Loeb. 1988. Fidelity of Hiv-1 Reverse-Transcriptase. Science 242:1168-1171.

Scaduto, DI, JM Brown, WC Haaland, DJ Zwickl, DM Hillis, ML Metzker. 2010. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. Proceedings of the National Academy of Sciences of the United States of America 107:21242-21247.

Shankarappa, R, JB Margolick, SJ Gange, et al. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. Journal of Virology 73:10489-10502.

Sonnet, J, JL Michaux, F Zech, JM Brucher, M de Bruyere, G Burtonboy. 1987. Early AIDS cases originating from Zaire and Burundi (1962-1976). Scandinavian Journal of Infectious Diseases 19:511-517.

Takehisa, J, L Zekeng, E Ido, I Mboudjeka, H Moriyama, T Miura, M Yamashita, LG Gurtler, M Hayami, L Kaptue. 1998. Various types of HIV mixed infections in Cameroon. Virology 245:1-10.

van der Kuyl, AC, S Jurriaans, NKT Back, HG Sprenger, TS van der Werf, F Zorgdrager, B Berkhout, M Cornelissen. 2011. Unusual Cluster of HIV Type 1 Dual Infections in Groningen, The Netherlands. Aids Research and Human Retroviruses 27:429-433.

Wei, X, JM Decker, S Wang, et al. 2003. Antibody neutralization and escape by HIV-1. Nature 422:307-312.

Wei, XP, SK Ghosh, ME Taylor, et al. 1995. Viral Dynamics in Human-Immunodeficiency-Virus Type-1 Infection. Nature 373:117-122.

REFERENCES

Albert, J, B Abrahamsson, K Nagy, E Aurelius, H Gaines, G Nystrom, EM Fenyo. 1990. Rapid Development of Isolate-Specific Neutralizing Antibodies after Primary Hiv-1 Infection and Consequent Emergence of Virus Variants Which Resist Neutralization by Autologous Sera. Aids 4:107-112.

Anastassopoulou, CG. 2012. Chemokine Receptors as Therapeutic Targets in HIV Infection. In: M K, editor. Immunodeficiency: InTech.

Arhel, N. 2010. Revisiting HIV-1 uncoating. Retrovirology 7:96.

Auerbach, DM, WW Darrow, HW Jaffe, JW Curran. 1984. Cluster of cases of the acquired immune deficiency syndrome. Patients linked by sexual contact. Am J Med 76:487-492.

Barre-Sinoussi, F, JC Chermann, F Rey, et al. 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). Science 220:868-871.

Bebenek, K, T Kunkel. 1993. Reverse Transcriptase Fidelity. In: A Skalka, S Goff, editors. Reverse Transcriptase. Cold Spring Harbor: Cold Spring Harbor Laboratory Press. p. 85-102.

Blay, WM, S Gnanakaran, B Foley, NA Doria-Rose, BT Korber, NL Haigwood. 2006. Consistent patterns of change during the divergence of human immunodeficiency virus type 1 envelope from that of the inoculated virus in simian/human immunodeficiency virus-infected macaques. J Virol 80:999-1014.

Boni, MF, D Posada, MW Feldman. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176:1035-1047.

Bowdler, N. 2007. Key HIV strain 'came from Haiti'. BBC News.

Bunnik, EM, L Pisas, AC van Nuenen, H Schuitemaker. 2008. Autologous neutralizing humoral immunity and evolution of the viral envelope in the course of subtype B human immunodeficiency virus type 1 infection. Journal of Virology 82:7932-7941.

Carmichael, M. 2007. Haunted by HIV's Origins. Newsweek Magazine. New York City: Newsweek, Inc.

CDC. 1981a. Follow-up on Kaposi's Sarcoma and Pneumocystis Pneumonia. Morbidity and Mortality Weekly Report 30:409-410.

CDC. 1981b. Kaposi's sarcoma and Pneumocystis Pneumonia among homosexual men - New York City and California. Morbidity and Mortality Weekly Report 30:305-308.

CDC. 1981c. Pneumocystis pneumonia - Los Angeles. Morbidity and Mortality Weekly Report 30:250-252.

CDC. 1982a. Epidemiologic Notes and Reports Pneumocystis carinii Pneumonai among Persons with Hemophilia A. Morbidity and Mortality Weekly Report 31:365-367.

CDC. 1982b. Opportunistic Infections and Kaposi's Sarcoma among Haitians in the United States. Morbidity and Mortality Weekly Report 31:353-354, 360-361.

Coakley, E, CJ Petropoulos, JM Whitcomb. 2005. Assessing chemokine co-receptor usage in HIV. Curr Opin Infect Dis 18:9-15.

Coffin, J, A Haase, JA Levy, et al. 1986. What to call the AIDS virus? Nature 321:10.

Coffin, J, S Hughes, H Varmus. 1997. Overview of Reverse Transcription. In: J Coffin, S Hughes, H Varmus, editors. Retroviruses. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Coffin, JM. 1995a. Hiv Population-Dynamics in-Vivo - Implications for Genetic-Variation, Pathogenesis, and Therapy. Science 267:483-489.

Coffin, JM. 1995b. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science 267:483-489.

Crisp, MD, LG Cook. 2005. Do early branching lineages signify ancestral traits? Trends Ecol Evol 20:122-128.

Dawkins, R, JR Krebs. 1979. Arms Races between and within Species. Proceedings of the Royal Society B-Biological Sciences 205:489-511.

Delaporte, E, W Janssens, M Peeters, et al. 1996. Epidemiological and molecular characteristics of HIV infection in Gabon, 1986-1994. Aids 10:903-910.

Dennis, AM, S Hue, CB Hurt, S Napravnik, J Sebastian, D Pillay, JJ Eron. 2012. Phylogenetic insights into regional HIV transmission. Aids 26:1813-1822.

Drummond, AJ, MA Suchard, D Xie, A Rambaut. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969-1973.

Edgar, RC. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edgar, RC. 2004b. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797.

Farmer, P. 1992. AIDS and accusation : Haiti and the geography of blame. Berkeley: University of California Press.

Fisher, R. 1930. The Genetical Theory of Natural Selection. Dover, New York.

Francois, KO, J Balzarini. 2011. The highly conserved glycan at asparagine 260 of HIV-1 gp120 is indispensable for viral entry. J Biol Chem 286:42900-42910.

Frankel, AD, JA Young. 1998. HIV-1: fifteen proteins and an RNA. Annu Rev Biochem 67:1-25.

Gallo, RC. 2006. A reflection on HIV/AIDS research after 25 years. Retrovirology 3:72.

Ganeshan, S, RE Dickover, BT Korber, YJ Bryson, SM Wolinsky. 1997. Human immunodeficiency virus type 1 genetic evolution in children with different rates of development of disease. J Virol 71:663-677.

Gao, F, E Bailes, DL Robertson, et al. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 397:436-441.

Geyer, H, C Holschbach, G Hunsmann, J Schneider. 1988. Carbohydrates of human immunodeficiency virus. Structures of oligosaccharides linked to the envelope glycoprotein 120. J Biol Chem 263:11760-11767.

Gibbs, MJ, JS Armstrong, AJ Gibbs. 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16:573-582.

Gilbert, MT, A Rambaut, G Wlasiuk, TJ Spira, AE Pitchenik, M Worobey. 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc Natl Acad Sci U S A 104:18566-18570.

Gojobori, T, EN Moriyama, Y Ina, K Ikeo, T Miura, H Tsujimoto, M Hayami, S Yokoyama. 1990. Evolutionary origin of human and simian immunodeficiency viruses. Proc Natl Acad Sci U S A 87:4108-4111.

Goodrich, DW, PH Duesberg. 1990. Retroviral recombination during reverse transcription. Proc Natl Acad Sci U S A 87:2052-2056.

Grenfell, BT, OG Pybus, JR Gog, JL Wood, JM Daly, JA Mumford, EC Holmes. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303:327-332.

Hahn, BH, MA Gonda, GM Shaw, M Popovic, JA Hoxie, RC Gallo, F Wong-Staal. 1985. Genomic diversity of the acquired immune deficiency syndrome virus HTLV-III: different viruses exhibit greatest divergence in their envelope genes. Proc Natl Acad Sci U S A 82:4813-4817.

Hahn, BH, GM Shaw, KM De Cock, PM Sharp. 2000. AIDS - AIDS as a zoonosis: Scientific and public health implications. Science 287:607-614.

Hahn, BH, GM Shaw, ME Taylor, RR Redfield, PD Markham, SZ Salahuddin, F Wong-Staal, RC Gallo, ES Parks, WP Parks. 1986. Genetic variation in HTLV-III/LAV over time in patients with AIDS or at risk for AIDS. Science 232:1548-1553.

Holmes, EC, M Worobey, A Rambaut. 1999. Phylogenetic evidence for recombination in dengue virus. Mol Biol Evol 16:405-409.

Hooper, E. 1999. The river : a journey to the source of HIV and AIDS. Boston, MA: Little, Brown and Co.

Hoorelbeke, B, D Huskens, G Ferir, KO Francois, A Takahashi, K Van Laethem, D Schols, H Tanaka, J Balzarini. 2010. Actinohivin, a Broadly Neutralizing Prokaryotic Lectin, Inhibits HIV-1 Infection by Specifically Targeting High-Mannose-Type Glycans on the gp120 Envelope. Antimicrobial Agents and Chemotherapy 54:3287-3301.

Hu, WS, SH Hughes. 2012. HIV-1 reverse transcription. Cold Spring Harb Perspect Med 2.

Hu, WS, HM Temin. 1990. Retroviral recombination and reverse transcription. Science 250:1227-1233.

Huelsenbeck, JP, F Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Jetzt, AE, H Yu, GJ Klarmann, Y Ron, BD Preston, JP Dougherty. 2000. High rate of recombination throughout the human immunodeficiency virus type 1 genome. J Virol 74:1234-1240.

Joint United Nations Programme on HIV/AIDS. 2011. UNAIDS World AIDS day report 2011. Geneva: UNAIDS.

Joint United Nations Programme on HIV/AIDS., World Health Organization. 2008. AIDS outlook/09 World AIDS Day 2008. Geneva: UNAIDS : World Health Organization.

Kahraman, A, L Malmstrom, R Aebersold. 2011. Xwalk: computing and visualizing distances in cross-linking experiments. Bioinformatics 27:2163-2164.

Kimpesa, M. 1983. L'operation de l'UNESCO au Congo-Leopoldville et le diagnostic des realities educatives congolaises: 1960-1964. Faculte de Psychologie. Geneva: Universite de Geneve.

Kohler, H, J Goudsmit, P Nara. 1992. Clonal Dominance - Cause for a Limited and Failing Immune-Response to Hiv-1 Infection and Vaccination. Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology 5:1158-1168.

Korber, B, M Muldoon, J Theiler, F Gao, R Gupta, A Lapedes, BH Hahn, S Wolinsky, T Bhattacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789-1796.

Kosakovsky Pond, SL, SD Frost. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208-1222.

Kosakovsky Pond, SL, D Posada, MB Gravenor, CH Woelk, SD Frost. 2006. Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23:1891-1901.

Krell, F-T, PS Cranston. 2004. Which side of the tree is more basal? Systematic Entomology 29:279-281.

Kwon, YD, A Finzi, X Wu, et al. 2012. Unliganded HIV-1 gp120 core structures assume the CD4-bound conformation with regulation by quaternary interactions and variable loops. Proc Natl Acad Sci U S A 109:5663-5668.

Lambert, B. 1990. Now, No Haitians Can Donate Blood. The New York Times. New York City: The New York TImes COmpany.

Lanciault, C, JJ Champoux. 2006. Pausing during reverse transcription increases the rate of retroviral recombination. J Virol 80:2483-2494.

Lasky, LA, JE Groopman, CW Fennie, PM Benz, DJ Capon, DJ Dowbenko, GR Nakamura, WM Nunes, ME Renz, PW Berman. 1986. Neutralization of the AIDS retrovirus by antibodies to a recombinant envelope glycoprotein. Science 233:209-212.

Lemey, P, A Rambaut, OG Pybus. 2006. HIV evolutionary dynamics within and among hosts. AIDS Rev 8:125-140.

Lemey, P, S Van Dooren, K Van Laethem, Y Schrooten, I Derdehnckx, P Goubau, F Brun-Vezinet, D Vaira, AM Vandamme. 2005. Molecular testing of multiple HIV-1 transmissions in a criminal case. Aids 19:1649-1658.

Lewis, F, GJ Hughes, A Rambaut, A Pozniak, AJ Leigh Brown. 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med 5:e50.

Li, WH, M Tanimura, PM Sharp. 1988. Rates and dates of divergence between AIDS virus nucleotide sequences. Mol Biol Evol 5:313-330.

Li, Y, L Luo, N Rasool, CY Kang. 1993. Glycosylation is necessary for the correct folding of human immunodeficiency virus gp120 in CD4 binding. J Virol 67:584-588.

Lodish, H, A Berk, S Zipursky, P Matsudaira, D Baltimore, J Darnell. 2000. Protein Glycosylation in the ER and Golgi Complex. Molecular Cell Biology. New York: Freeman W.H.

Machuca, R, LB Jorgensen, P Theilade, C Nielsen. 2001. Molecular investigation of transmission of human immunodeficiency virus type 1 in a criminal case. Clinical and Diagnostic Laboratory Immunology 8:884-890.

Maljkovic Berry, I, G Athreya, M Kothari, M Daniels, WJ Bruno, B Korber, C Kuiken, RM Ribeiro, T Leitner. 2009. The evolutionary rate dynamically tracks changes in HIV-1 epidemics: application of a simple method for optimizing the evolutionary rate in phylogenetic trees with longitudinal data. Epidemics 1:230-239.

Maljkovic Berry, I, R Ribeiro, M Kothari, G Athreya, M Daniels, HY Lee, W Bruno, T Leitner. 2007. Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. J Virol 81:10625-10635.

Martin, D, E Rybicki. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562-563.

Martin, DP, P Lemey, M Lott, V Moulton, D Posada, P Lefeuvre. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26:2462-2463.

Martin, DP, P Lemey, D Posada. 2011. Analysing recombination in nucleotide sequences. Mol Ecol Resour 11:943-955.

Martin, DP, D Posada, KA Crandall, C Williamson. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses 21:98-102.

Mather, K. 1955. Polymorphism as an Outcome of Disruptive Selection. Evolution 9:52-61.

Metzker, ML, DP Mindell, XM Liu, RG Ptak, RA Gibbs, DM Hillis. 2002. Molecular evidence of HIV-1 transmission in a criminal case. Proceedings of the National Academy of Sciences of the United States of America 99:14292-14297.

Monini, P, C Sgadari, E Toschi, G Barillari, B Ensoli. 2004. Antitumour effects of antiretroviral therapy. Nat Rev Cancer 4:861-875.

Muller, HJ. 1932. Some genetic aspects of sex. American Naturalist 66:118-138.

Naeger, LK, KA Struble, JS Murray, DB Birnkrant. 2010. Running a tightrope: regulatory challenges in the development of antiretrovirals. Antiviral Res 85:232-240.

Nevin Gerek, Z, S Kumar, S Banu Ozkan. 2013. Structural dynamics flexibility informs function and evolution at a proteome scale. Evolutionary Applications 6:423-433.

Nkengasong, JN, W Janssens, L Heyndrickx, et al. 1994. Genotypic Subtypes of Hiv-1 in Cameroon. Aids 8:1405-1412.

Ogert, RA, MK Lee, W Ross, A Buckler-White, MA Martin, MW Cho. 2001. N-linked glycosylation sites adjacent to and within the V1/V2 and the V3 loops of dualtropic human immunodeficiency virus type 1 isolate DH12 gp120 affect coreceptor usage and cellular tropism. J Virol 75:5998-6006.

Padidam, M, S Sawyer, CM Fauquet. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265:218-225.

Pape, JW, P Farmer, S Koenig, D Fitzgerald, P Wright, W Johnson. 2008. The epidemiology of AIDS in Haiti refutes the claims of Gilbert et al. Proc Natl Acad Sci U S A 105:E13.

Pepin, J. 2011. The origins of AIDS. Cambridge, UK ; New York: Cambridge University Press.

Perelson, AS, AU Neumann, M Markowitz, JM Leonard, DD Ho. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582-1586.

Poon, AF, FI Lewis, SL Pond, SD Frost. 2007. Evolutionary interactions between N-linked glycosylation sites in the HIV-1 envelope. PLoS Comput Biol 3:e11.

Popovic, M, MG Sarngadharan, E Read, RC Gallo. 1984. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science 224:497-500.

Posada, D, KA Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci U S A 98:13757-13762.

Posada, D, KA Crandall. 2002. The effect of recombination on the accuracy of phylogeny estimation. Journal of molecular evolution 54:396-402.

Preston, BD, BJ Poiesz, LA Loeb. 1988. Fidelity of Hiv-1 Reverse-Transcriptase. Science 242:1168-1171.

Rambaut, A. 2007. Tracer v1.4.

Rambaut, A. 2009. FigTree v1.3.1.

Ratner, L, RC Gallo, F Wong-Staal. 1985. HTLV-III, LAV, ARV are variants of same AIDS virus. Nature 313:636-637.

Ratner, L, W Haseltine, R Patarca, et al. 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature 313:277-284.

Rhodes, T, H Wargo, WS Hu. 2003. High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication. J Virol 77:11193-11200.

Rizzuto, CD, R Wyatt, N Hernandez-Ramos, Y Sun, PD Kwong, WA Hendrickson, J Sodroski. 1998. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science 280:1949-1953.

Robertson, DL, BH Hahn, PM Sharp. 1995. Recombination in AIDS viruses. J Mol Evol 40:249-259.

Rudd, PM, MR Wormald, RL Stanfield, M Huang, N Mattsson, JA Speir, JA DiGennaro, JS Fetrow, RA Dwek, IA Wilson. 1999. Roles for glycosylation of cell surface receptors involved in cellular immune recognition. J Mol Biol 293:351-366.

Ruff, AJ, J Coberly, NA Halsey, et al. 1994. Prevalence of HIV-1 DNA and p24 antigen in breast milk and correlation with maternal factors. J Acquir Immune Defic Syndr 7:68-73.

Saag, MS, BH Hahn, J Gibbons, Y Li, ES Parks, WP Parks, GM Shaw. 1988. Extensive variation of human immunodeficiency virus type-1 in vivo. Nature 334:440-444.

Scaduto, DI, JM Brown, WC Haaland, DJ Zwickl, DM Hillis, ML Metzker. 2010. Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. Proceedings of the National Academy of Sciences of the United States of America 107:21242-21247.

Schierup, MH, J Hein. 2000. Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879-891.

Shankarappa, R, JB Margolick, SJ Gange, et al. 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73:10489-10502.

Shimodaira, H, M Hasegawa. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. Molecular Biology and Evolution 16:1114.

Siepel, AC, AL Halpern, C Macken, BT Korber. 1995. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. AIDS Res Hum Retroviruses 11:1413-1416.

Smith, JM. 1992. Analyzing the mosaic structure of genes. J Mol Evol 34:126-129.

Sonnet, J, JL Michaux, F Zech, JM Brucher, M de Bruyere, G Burtonboy. 1987. Early AIDS cases originating from Zaire and Burundi (1962-1976). Scand J Infect Dis 19:511-517.

Takehisa, J, L Zekeng, E Ido, I Mboudjeka, H Moriyama, T Miura, M Yamashita, LG Gurtler, M Hayami, L Kaptue. 1998. Various types of HIV mixed infections in Cameroon. Virology 245:1-10.

Tamura, K, D Peterson, N Peterson, G Stecher, M Nei, S Kumar. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731-2739.

Tsodikov, OV, MT Record, Jr., YV Sergeev. 2002. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. J Comput Chem 23:600-609.

van der Kuyl, AC, S Jurriaans, NKT Back, HG Sprenger, TS van der Werf, F Zorgdrager, B Berkhout, M Cornelissen. 2011. Unusual Cluster of HIV Type 1 Dual Infections in Groningen, The Netherlands. Aids Research and Human Retroviruses 27:429-433.

Wain-Hobson, S, P Sonigo, O Danos, S Cole, M Alizon. 1985. Nucleotide sequence of the AIDS virus, LAV. Cell 40:9-17.

Wei, X, JM Decker, S Wang, et al. 2003. Antibody neutralization and escape by HIV-1. Nature 422:307-312.

Wei, XP, SK Ghosh, ME Taylor, et al. 1995. Viral Dynamics in Human-Immunodeficiency-Virus Type-1 Infection. Nature 373:117-122.

Weiller, GF. 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. Mol Biol Evol 15:326-335.

Wertheim, JO, M Fourment, SL Kosakovsky Pond. 2012. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. Mol Biol Evol 29:451-456.

Wolinsky, SM, BT Korber, AU Neumann, M Daniels, KJ Kunstman, AJ Whetsell, MR Furtado, Y Cao, DD Ho, JT Safrit. 1996. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science 272:537-542.

Wong-Staal, F, GM Shaw, BH Hahn, SZ Salahuddin, M Popovic, P Markham, R Redfield, RC Gallo. 1985. Genomic diversity of human T-lymphotropic virus type III (HTLV-III). Science 229:759-762.

Worobey, M, M Gemmel, DE Teuwen, et al. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455:661-664.

Wu, X, T Zhou, S O'Dell, RT Wyatt, PD Kwong, JR Mascola. 2009. Mechanism of human immunodeficiency virus type 1 resistance to monoclonal antibody B12 that effectively targets the site of CD4 attachment. J Virol 83:10892-10907.

Zhang, M, B Gaschen, W Blay, B Foley, N Haigwood, C Kuiken, B Korber. 2004. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. Glycobiology 14:1229-1246.

Zhang, W, G Canziani, C Plugariu, R Wyatt, J Sodroski, R Sweet, P Kwong, W Hendrickson, I Chaiken. 1999. Conformational changes of gp120 in epitopes near the CCR5 binding site are induced by CD4 and a CD4 miniprotein mimetic. Biochemistry 38:9405-9416.

APPENDIX I

SUPPORTING TABLES FOR CHAPTER 2

Table 1

			GAG		ENV	
Country	Sub type	Patient ID	Accession	RIP	Accession	RIP
IN	C	93IN904	AF067157	95		
CD	D	ELI	A07108**	NS	A07108	95
CD	D	NDK	A34828**	95	M27323	95
CD	D	84ZR085	U88822**	95	U88822	95
UG	D	99UGK09958	AF484499**	95	AF484499	95
UG	D	99UGE13613	AF484515**	95	AF484515	95
US	B	H6	EF362776**	95	EF159973	95
US	B	H3	EF362774**	95	EF159971	95
US	B	H2	EF362773**	95	EF159970	95
US	B	H5	EF362775**	95	EF159972	95
US	B	H7	EF362777	95	EF159974	95
Us[a]	B	WMJ	K03457*	95	M12507	95
US[a]	B	US4	AY173955*	95	AY173955	95
US	B	RF	M17451*	95	M17451	95
Kr[a]	B	KR5086			AJ417429	95
Br[a]	B	92BR020			AY669718	95
HT	B	593			AY669721	95
HT	B	HT594			U08445	95
HT	B	HT596			U08446	95
HT	B	HT599			U08447	95
HT	B	HT651			U08441	95
HT	B	HT652			U08443	95
TT	B	QH0679			AF277064	95
TT	B	QH0060			AF277055	95
TT	B	QH0016			AF277059	95
TT	B	QH1420			AF277075	95
TT	B	QH0515			AF277061	95
TT	B	QH1116			AF277074	95
TT	B	QH0020			AF277056	95
TT	B	QH0791			AF277068	95
TT	B	QH0908			AF277072	95
TT	B	QH0605			AF277063	95
TT	B	QH0605			AF277057	95
US	B	US1	AY173952**	95	AY173952	95
FR	B	HXB2LAIIIIBBRU	K03455**	95	CS793683	95
US	B	DH12_3	AF069140**	95	AF069139	95
GA	B	OYI_397	M26727**	95	M26727	95
GB	B	GB8	AJ271445**	95	AJ271445	95
US	B	NY5CG	M38431**	95	M38431	95

US	B	WCIPR	U69584**	95	U69584	95
US	B	US3	AY173954**	95	AY173954	95
US	B	US2	AY173953**	95	AY173953	95
BR	B	BZ167	AY173956**	95	AY173956	95
ES	B	89SP061	AJ006287**	95		
US	B	MNCG	M17449**	95	M17449	95
DE	B	D31	U43096**	95	AY247223	95
US	B	AD87/ADA	AF004394**	95	AY426119	95
DE	B	HAN	U43141**	95	U43141	95
US	B	YU_2	M93258**	95	M93258	95
US	B	SF33	AY352275**	95	AY352275	95
CO	B	PCM039	AY561239*	95	AY561239	95
AR	B	ARCH054	AY037268*	95	AY037268	95
NL	B	ACH3202A21	U34604*	95	U34604	95
AU	B	MBC18	AF042102*	95	AF042102	95
US	B	WEAU160	AY037282*	95	AY223743	95
US	B	P896	AF042102*	95	U39362	95
CO	B	PCM001	AY561236*	95	AY561236	95
US	B	BC_BCSG3	L02317*	95	L02317	95
US	B	CDC451-b	M13136*	95	M13137	95
US	B	MBC925	AF042101*	95	AF042101	95
US	B	MBC200	AF042100*	95	AF042100	95
AR	B	ARMA132	AY037282*	95	AY037282	95
EC	B	EC102	AY173960*	95	AY173960	95
AR	B	ARMS008	AY037269*	95	AY037269	95
ES	B	89SP061	AJ006287*	95	AJ006287	95
CO	B	PCM013	AY561237*	95	AY561237	95
US	**B**	**Ac_06**	**AY268493**	**95**		
US	**B**	**Ac_06**	**AY247251**	**95**		
KR	B	WK	AF224507	95		
US	B	JR	U63632/M38429	95	AY426125	95
US	B	WR27	AF286365	95		
US	B	1057_01	AY331292	95		
US	B	1299_d22	AY308761	95		
US	B	98USHVTN941c1	AY560110	95		
RU	B	04RU128005	AY682547	95		
US	B	1012_08	AY331285	95		
US	B	1013_03	AY331287	95		
TH	B	BK132	AY173951	95		
US	B	ARES2	AB078005	95		
CA	B	WC10C_10	AY314061	95		
JP	B	JH31	M21137	95		
US	B	P896	U39362	95		
RU	B	04RU129005	AY751406	95		

US	B	BC_BCSG3	L02317	95		
MM	B	mSTD101	AB097870	95		
CN	B	02HNsmx2	AY275557	95		
GB	B	CAM1	D10112	95		
RU	B	04RU139095	AY819715	95		
TW	B	TWCYS_LM49	AF086817	95		
US	B	SF2	K02007	95		
CA	B	CANA1FULL	AY779564	95		
CA	B	CANC2FULL	AY779559	95		
CA	B	CANB3FULL	AY779553	95		
ZA	B	99ZASM1			AY505010	95
US	B	1304_d31			AY308762	95
KR	B	KR3026_C1			AJ417426	95
US	B	C16			U84841	95
US	B	C08			U84799	95
US	B	C09B			U84831	95
US	B	C13D			U84833	95
FR	B	PIH160			AF041131	95
FR	B	PHI159			AF041128	95
US	B	93US073			U79721	95
KR	B	KR2057			AJ417423	95
TH	B	92TH014C_n			U08801	95
US	B	92US657			U04908	95
US	B	SFMHS19			AF025762	95
US	B	SFMHS6			AF025754	95
US	B	SFMHS7			AF025755	95
US	B	SFMHS4			AF025752	95
JP	B	JH32			M21138	95
US	B	81CA2			AY247219	95
FR	B	PIH374			AF041133	95
FR	B	133-L-1			AY535433	95
US	B	92US727			U79720	95
GB	B	876CD372			AJ535615	95
US	B	056			AY669719	95
FR	B	PIH155			AF041130	95
US	B	C04B			U84821	95
US	B	R2			AF128126	95
US	B	81NJ			AY247221	95
US	B	81CA1			AY247218	95
US	B	81NY3			AY247224	95
US	B	81NY1			AY247222	95
US	B	SFMHS20			AF025763	95
KR	B	KR5058			AJ417411	95
KR	B	KR3042			AJ417409	95

US	B	SFMHS2			AF025750	95
US	B	SFMHS8			AF025756	95
CA	B	82CAN			AY247225	95
US	B	SFMHS18			AF025761	95
US	B	SFMHS3			AF025751	95
US	B	81NY2			AY247223	95
US	B	C11			U84806	95
US	B	C14			U84850	95
US	B	C10			U84800	95
GB	B	822CD341			AJ535610	95
FR	B	PIH373			AF041134	95
US	B	C12			PIH373	95
US	B	92US716			U08452	95
US	B	712			AY669725	95
US	B	92US715_6			U08451	95
US	B	C17B			U84814	95
US	B	SF2B13			L07422	95
GB	B	749CD352			AJ535607	95
US	B	C05D			U84823	95
FR	B	PHI153			AF041129	95
NL	B	ACH320			AF069524	95
US	B	C15A			U84811	95
GB	B	747CD321			AJ535599	95
FR	B	PIH309			AF041132	95
US	B	81GA			AY247220	95
FR	B	PHI120			AF041125	95
GB	B	817CD306			AJ535619	95
US	B	CC1_85_start			EF367186	95

Summary characteristics for all analyzed sequences. Throughout this study, non-subtype B sequences are represented in red, assumed Haiti-originating sequences in green, Trinidad and Tobago-based sequences in blue and all "pandemic clade" sequences in gold.

[a]: indicates a non-Haitian patient annotated as being infected in Haiti or the Dominican Republic
The lack of an asterisk next to a gag sequence accession number indicates that the sequence is restricted to the first gag dataset only.
***: indicates a gag sequence included in both gag datasets*
**: indicates a gag sequence included only in the gag2 dataset*
H7 is underlined because the Gag sequence has three in-frame stop codons.
Both patient IDs labeled AC_06 are bolded because they are gag sequences from the same patient.

Table 2

SBP Analyses				
Dataset	Recombination	AIC_c Improvement	Breakpoint Location	Model Avg Support
Gag1	Yes	57.5	221	100%
Gag2	No	N/A	N/A	0%
Env	Yes	954	1738	100%

Recombination report for the SBP analyses. SBP was used as an initial filter for the identification of a single most likely breakpoint, if present, in each dataset. The goodness of fit of a dual- over single partition was measured via the AIC_c. Breakpoint location is relative to the segment rather than HXB2 numbering.

Table 3

A

Gag Dataset 1				
Breakpoint	LHS Raw p	LHS adjusted p	RHS Raw p	RHS adjusted p
471	0.0005	0.002	0.0001	0.0004
812	0.0007	0.0028	0.0012	0.0048
At p = 0.01 there are 2 significant breakpoints				
At p = 0.05 there are 2 significant breakpoints				
At p = 0.1 there are 2 significant breakpoints				

B

Gag Dataset 2				
Breakpoint	LHS Raw p	LHS adjusted p	RHS Raw p	RHS adjusted p
NA	NA	NA	NA	NA
At p = 0.01 there are 0 significant breakpoints				
At p = 0.05 there are 0 significant breakpoints				
At p = 0.1 there are 0 significant breakpoints				

C

Envelope Dataset				
Breakpoint	LHS Raw p	LHS adjusted p	RHS Raw p	RHS adjusted p
1416	0.0009	0.0036	0.0001	0.0004
1990	0.0001	0.0004	0.0001	0.0004
At p = 0.01 there are 2 significant breakpoints				
At p = 0.05 there are 2 significant breakpoints				
At p = 0.1 there are 2 significant breakpoints				

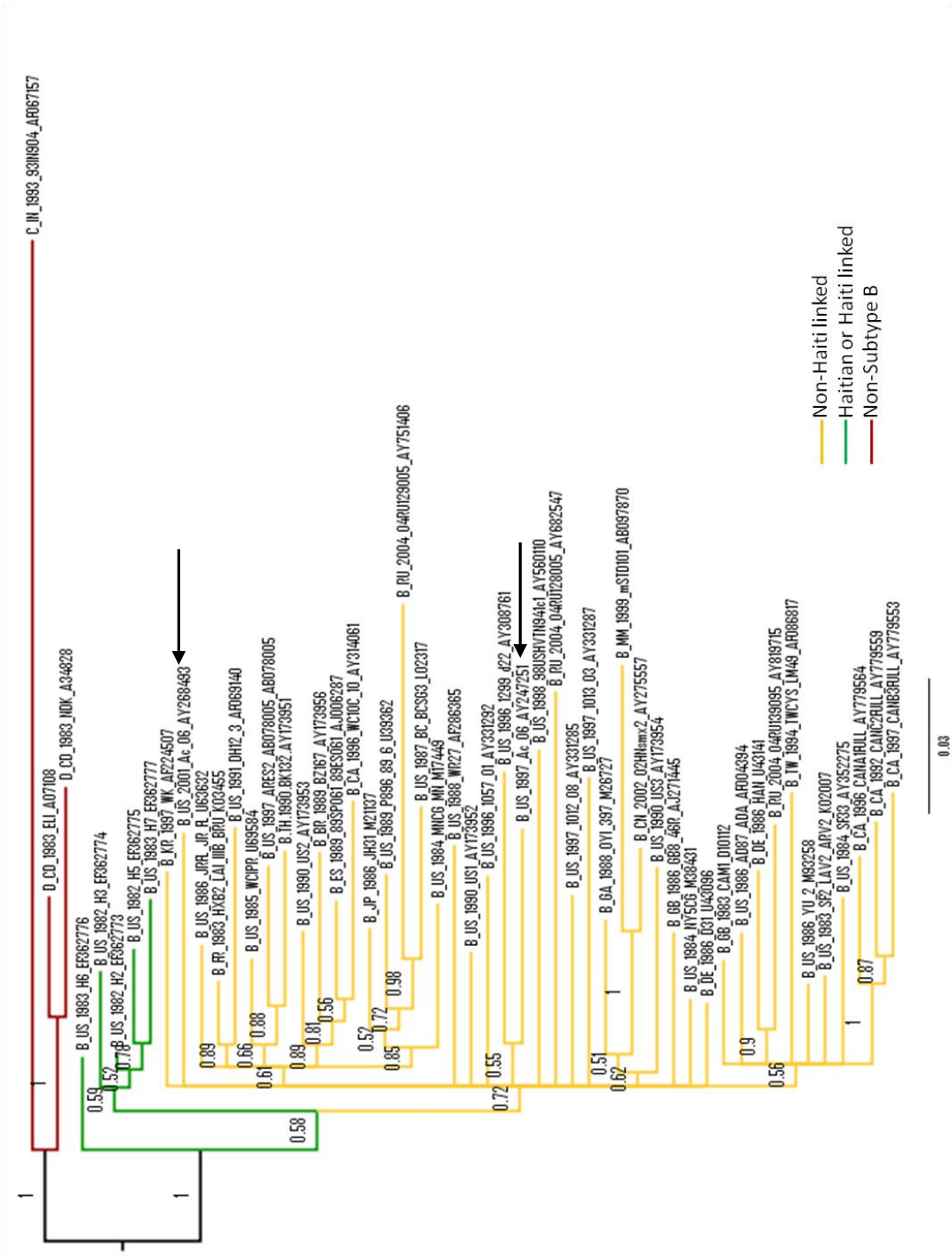
Recombination report from GARD analyses. GARD was employed to identify the approximate location of multiple recombination breakpoints in each dataset. The goodness of fit of the multi- over single partition model is measured by AIC_c. Discordant phylogenies were tested during the processing of GARD results using the Shimodaira and Hasegawa test which tests whether adjacent segments (right handed segment (RHS) vs left handed segment (LHS)) show a statistically significant difference in tree topologies. The adjusted p-values are corrected for multiple tests. Breakpoint location is relative to the segment rather than HXB2 numbering.

APPENDIX II

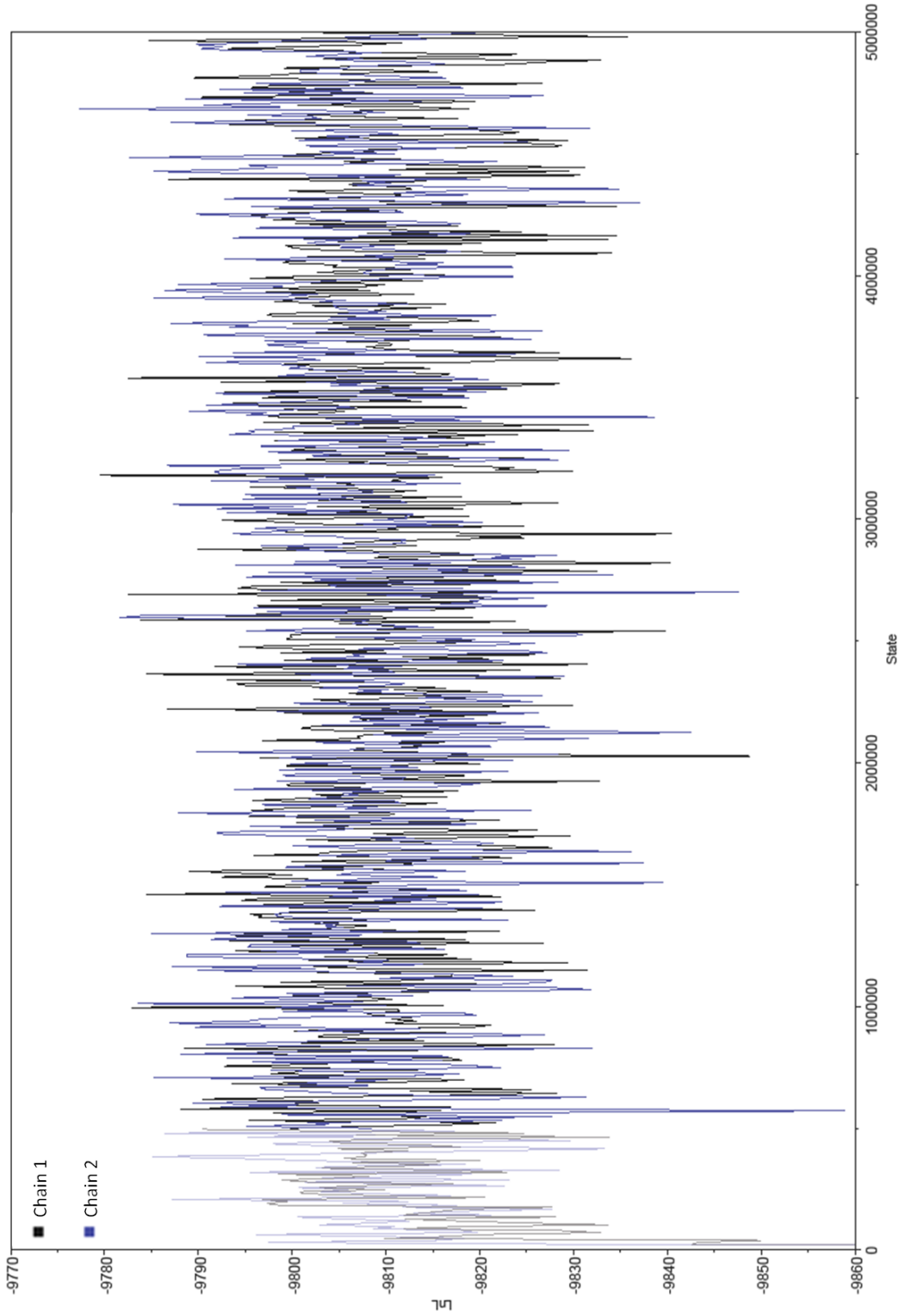
SUPPORTING FIGURES FOR CHAPTER 2

Figure 1. *Majority rule consensus trees constructed before recombination detection analysis*. *Phylogenetic trees shown here depict the topology and Bayesian convergence results, respectively, of the gag1 (A and B) and env (C and D) datasets under the assumption that all input sequences are representative of a single evolutionary history. Again, non-subtype B sequences are represented in red, Haiti-originating sequences in green, Trinidad and Tobago-based sequences in blue and all non-Caribbean sequences in gold (A and C). For the Bayesian convergence results (B and D), individual chains are colored blue or black. Posterior probabilities have been labeled on corresponding branches. It should be noted that upon termination of the env run, MrBayes(1) suggested that the analysis should be run for a greater number of generations.*

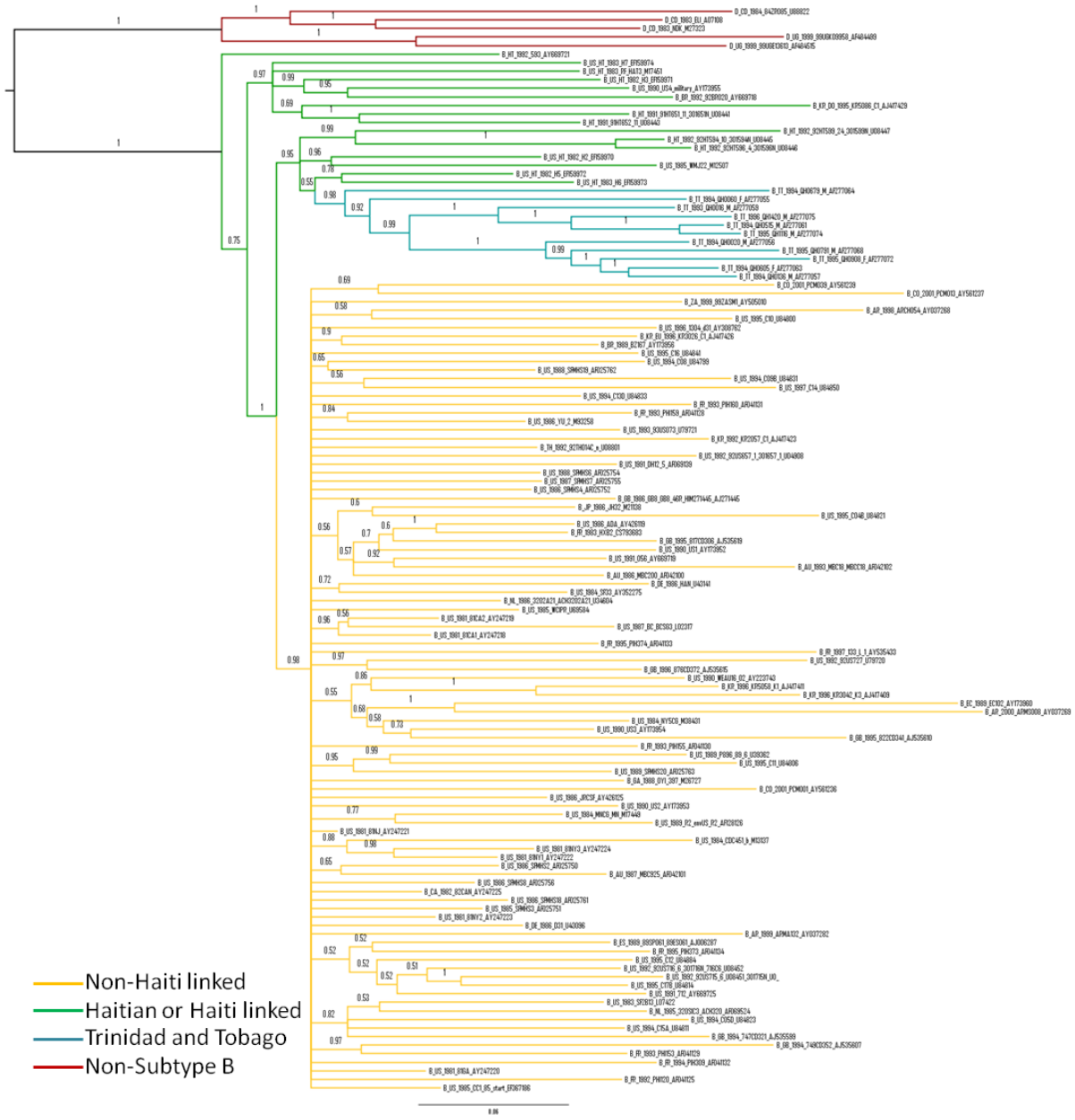
A


C_IN_1993_93IN9904_AF067157
O_CD_1983_ELI_A07108
O_CD_1983_NDK_A34828
B_US_1983_H6_EF362776
B_US_1982_H3_EF362774
B_US_1982_H2_EF362773
B_US_1982_H5_EF362775
B_US_1983_H7_EF362777
B_KP_1997_WK_AF224507
B_US_2001_Ac_06_AY289493
B_US_1986_JOR_JR_A_U63632
B_IR_1983_HXB2_LAI_IIIB_BRU_K03455
B_US_1991_DH12_3_AF069140
B_US_1985_WOPR_U69584
B_US_1997_ARES2_AB078005_AB078005
B_TH_1990_BK132_AY173951
B_US_1990_US2_AY173953
B_BR_1989_BZ167_AY173956
B_ES_1989_89SP061_89ES061_AJ006287
B_CA_1986_MC10C_10_AY314061
B_JP_1986_JH31_M21137
B_US_1989_P896_89_6_U39362
B_RU_2004_04RU129005_AY751406
B_US_1987_BC_BC363_L02317
B_US_1984_MNCG_MN_M17449
B_US_1988_WR27_AF286365
B_US_1990_US1_AY173952
B_US_1986_1057_01_AY331292
B_US_1996_1299_d22_AY308761
B_US_1997_Ac_06_AY247251
B_US_1988_98USHVTN944c1_AY560110
B_RU_2004_04RU128005_AY682547
B_US_1987_1012_08_AY331285
B_US_1997_1013_03_AY331287
B_GA_1988_OYI_397_M26727
B_MM_1999_mSTD101_AB097870
B_CN_2002_02HNsmx2_AY275557
B_US_1990_US3_AY173954
B_GB_1986_GB8_468_AJ271445
B_US_1984_NYSCG_M38431
B_DE_1986_DG1_U43096
B_GB_1983_CAM1_D10112
B_US_1986_AD87_ADA_ARD04394
B_DE_1986_HAN_U43141
B_RU_2004_04RU139095_AY819715
B_TW_1994_TWCYS_LM48_AR088817
B_US_1986_YU_2_M93258
B_US_1983_SF2_LAV2_ARV2_K02007
B_US_1984_SF33_AY352275
B_CA_1996_CANAB1LL_AY779564
B_CA_1992_CANC2R1LL_AY779559
B_CA_1997_CANB3R1LL_AY779553
0.03
Non-Haiti linked
Haitian or Haiti linked
Non-Subtype B

B


Chain 1
Chain 2
-9770
-9780
-9790
-9800
-9810
-9820
-9830
-9840
-9850
-9860
LL
0
1000000
2000000
3000000
4000000
5000000
State

C


Non-Haiti linked
Haitian or Haiti linked
Trinidad and Tobago
Non-Subtype B

D

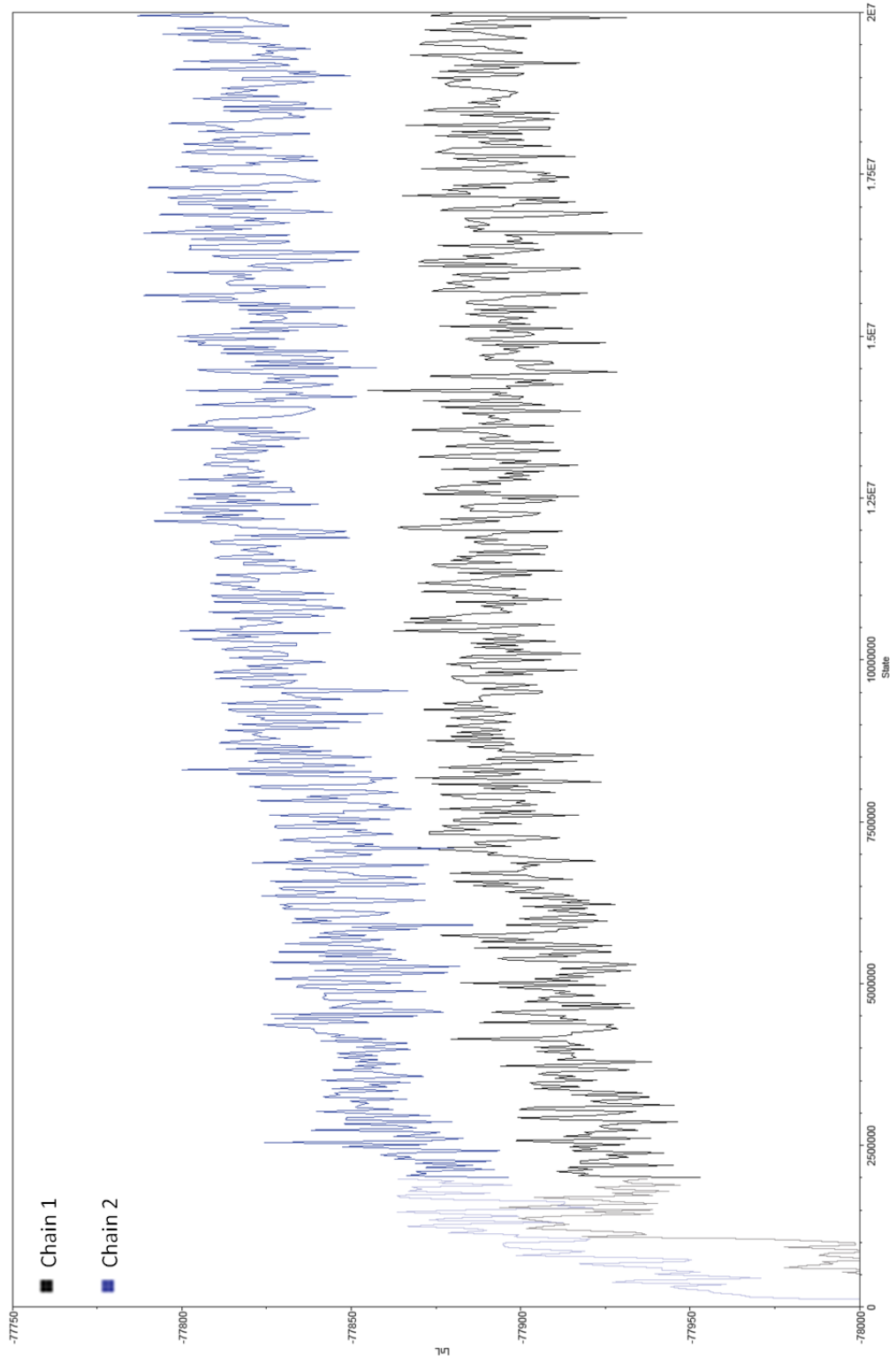
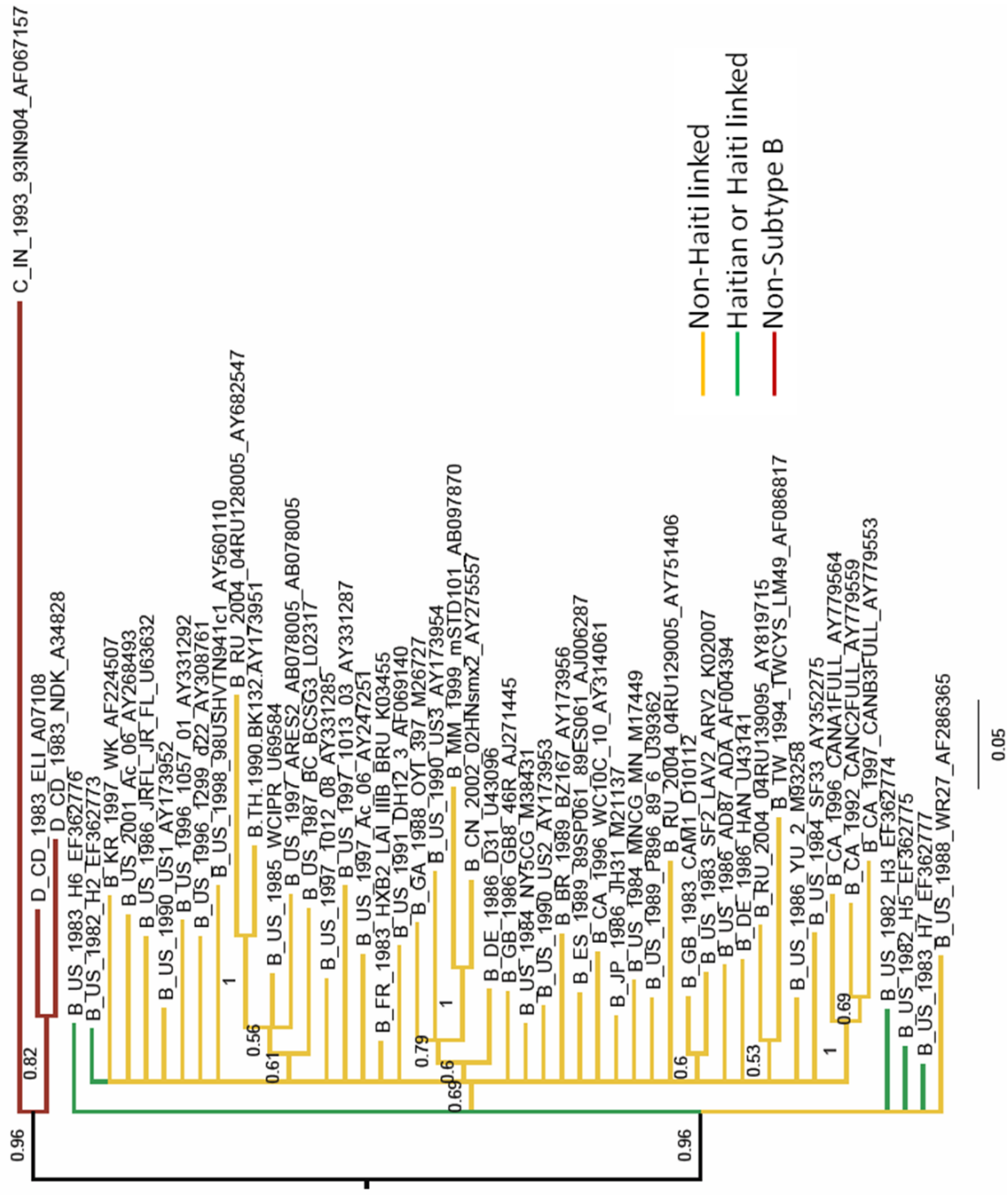
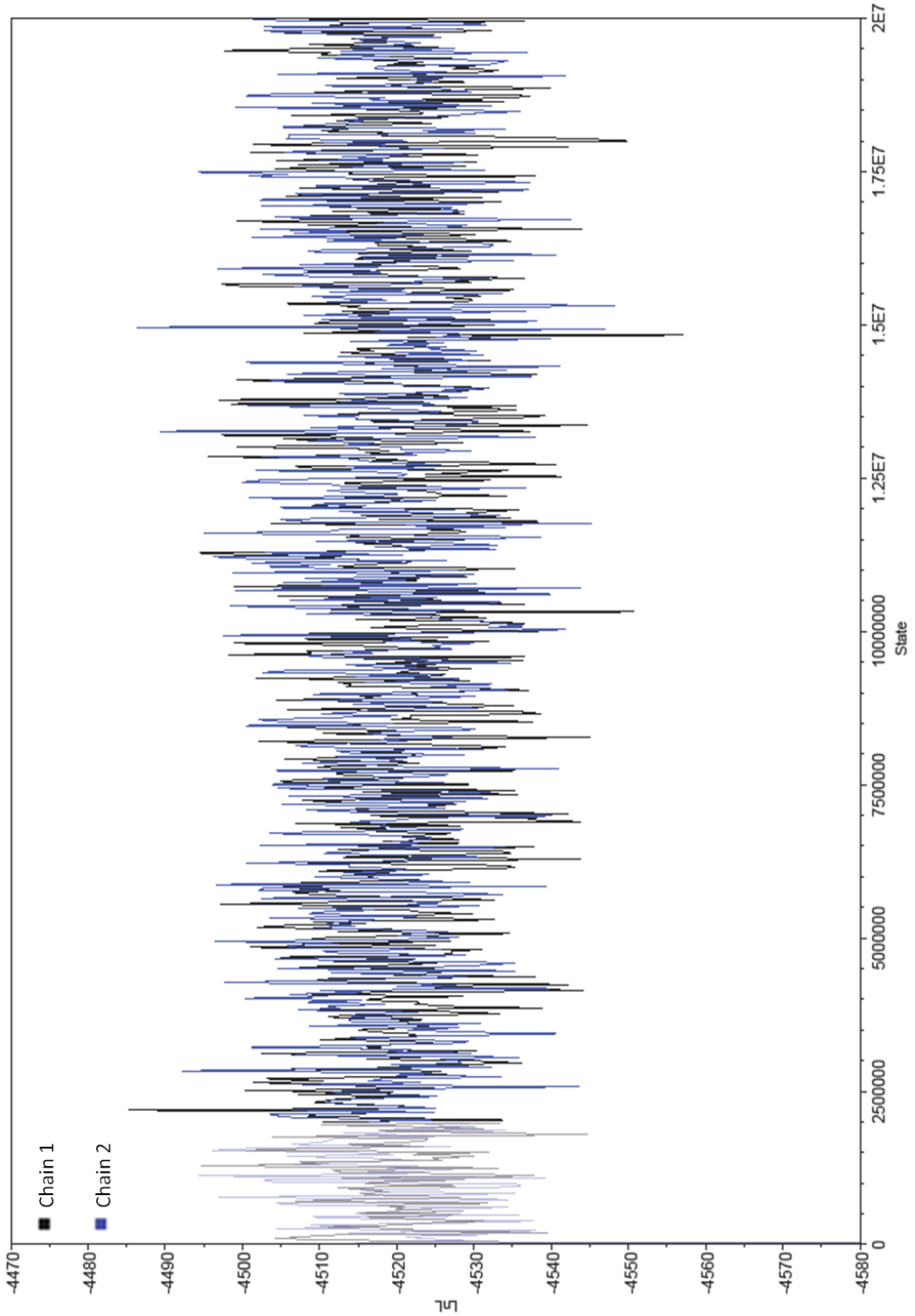

Chain 1
Chain 2
lnL
State
-77750
-77800
-77850
-77900
-77950
-78000
0
2500000
5000000
7500000
10000000
1.25E7
1.5E7
1.75E7
2E7

Figure 2. *Majority rule consensus trees for gag constructed after recombination detection analysis. Phylogenetic trees, reconstructed in MrBayes (1), depict the topology (A, C and E) and Bayesian convergence results (B, D and F) of GARD-inferred non-recombinant segments within the gag1 dataset. Trees were constructed from HXB2 positions (A) 820-1272, (C) 1273-1614 and (E) 1615-1952. Non-subtype B sequences are represented in red, Haiti-originating and Haiti-linked sequences in green and all non-Caribbean subtype B sequences in gold.*

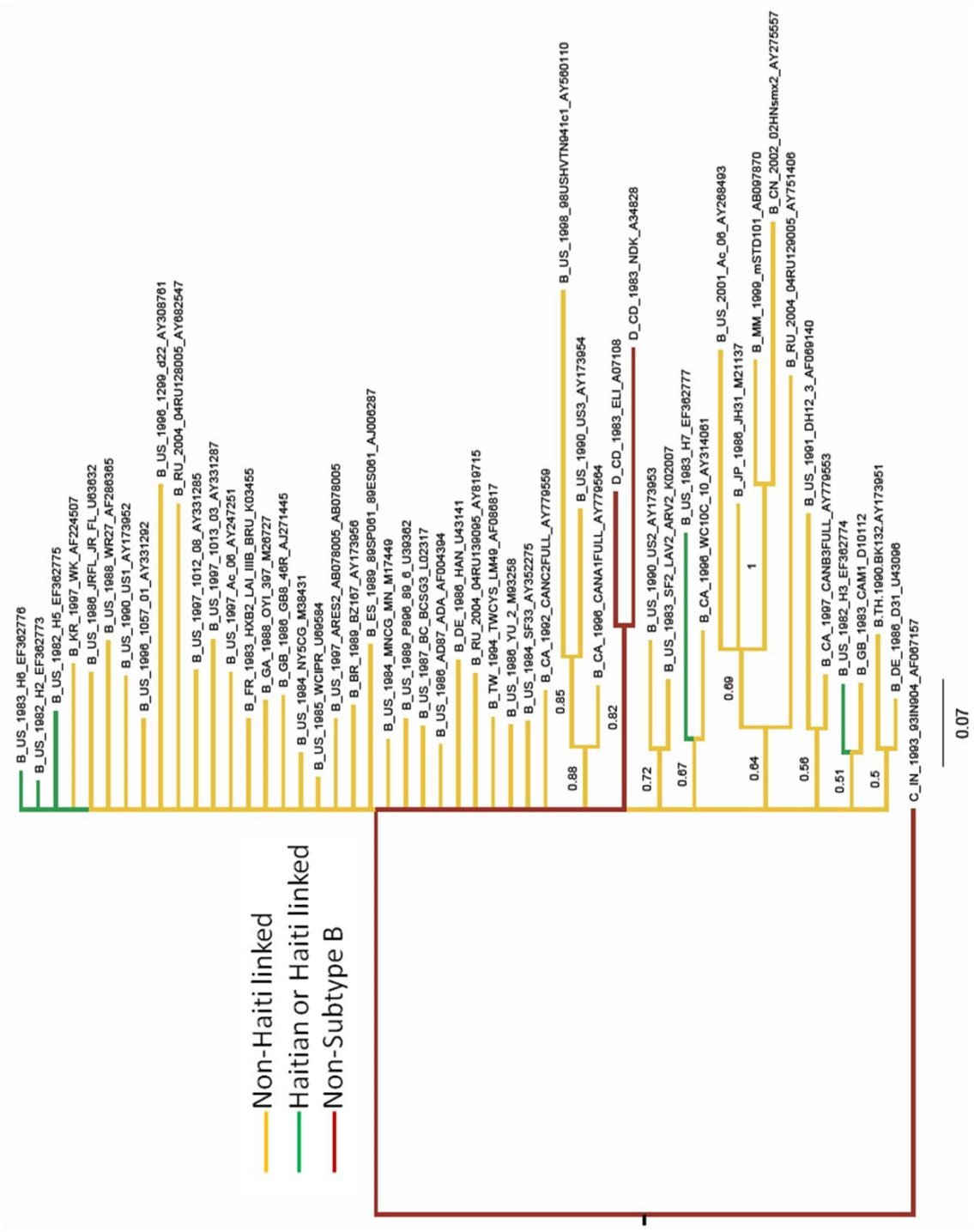
A


C_IN_1993_93IN904_AF067157
D_CD_1983_ELI_A07108
D_CD_1983_NDK_A34828
B_US_1983_H6_EF362776
B_US_1982_H2_EF362773
B_KR_1997_WK_AF224507
B_US_2001_Ac_06_AY268493
B_US_1986_JRFL_JR_FL_U63632
B_US_1990_US1_AY173952
B_US_1996_1057_01_AY331292
B_US_1996_1299_d22_AY308761
B_US_1998_98USHVTN941c1_AY560110
B_RU_2004_04RU128005_AY682547
B.TH.1990.BK132.AY173951
B_US_1985_WCIPR_U69584
B_US_1997_ARES2_AB078005_AB078005
B_US_1987_BC_BCSG3_L02317
B_US_1997_1012_08_AY331285
B_US_1997_1013_03_AY331287
B_US_1997_Ac_06_AY247251
B_FR_1983_HXB2_LAI_IIIB_BRU_K03455
B_US_1991_DH12_3_AF069140
B_GA_1988_OYI_397_M26727
B_US_1990_US3_AY173954
B_MM_1999_mSTD101_AB097870
B_CN_2002_02HNsmx2_AY275557
B_DE_1986_D31_U43096
B_GB_1986_GB8_46R_AJ271445
B_US_1984_NY5CG_M38431
B_US_1990_US2_AY173953
B_BR_1989_BZ167_AY173956
B_ES_1989_89SP061_89ES061_AJ006287
B_CA_1996_WC10C_10_AY314061
B_JP_1986_JH31_M21137
B_US_1984_MNCG_MN_M17449
B_US_1989_P896_89_6_U39362
B_RU_2004_04RU129005_AY751406
B_GB_1983_CAM1_D10112
B_US_1983_SF2_LAV2_ARV2_K02007
B_US_1986_AD87_ADA_AF004394
B_DE_1986_HAN_U43141
B_RU_2004_04RU139095_AY819715
B_TW_1994_TWCYS_LM49_AF086817
B_US_1986_YU_2_M93258
B_US_1984_SF33_AY352275
B_CA_1996_CANA1FULL_AY779564
B_CA_1992_CANC2FULL_AY779559
B_CA_1997_CANB3FULL_AY779553
B_US_1982_H3_EF362774
B_US_1982_H5_EF362775
B_US_1983_H7_EF362777
B_US_1988_WR27_AF286365
0.96
0.82
1
0.56
0.61
0.79
0.6
0.69
1
0.6
0.53
1
0.69
0.96
0.05
Non-Haiti linked
Haitian or Haiti linked
Non-Subtype B

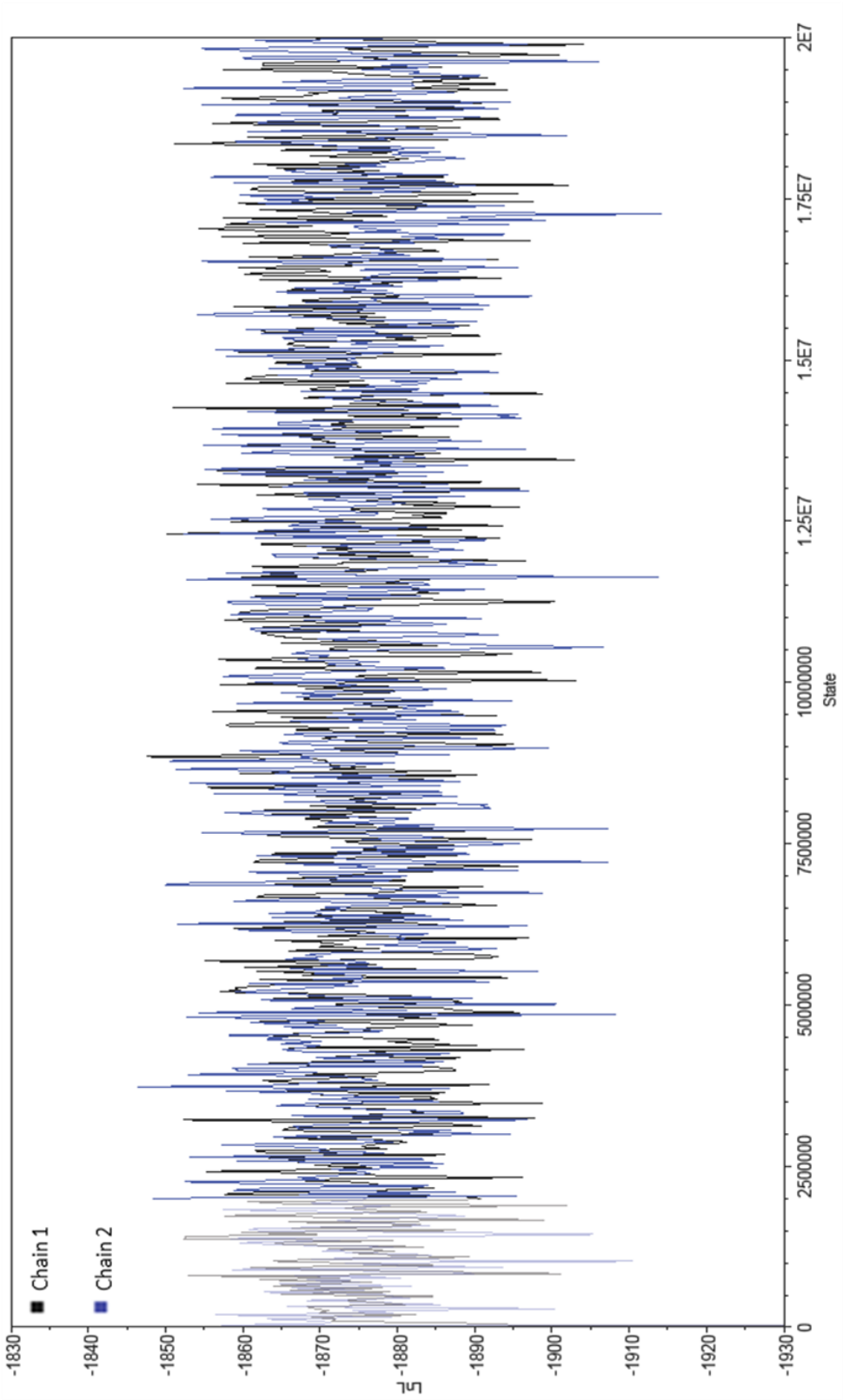
B


Chain 1
Chain 2
lnL
State
-4470
-4480
-4490
-4500
-4510
-4520
-4530
-4540
-4550
-4560
-4570
-4580
0
2500000
5000000
7500000
10000000
1.25E7
1.5E7
1.75E7
2E7

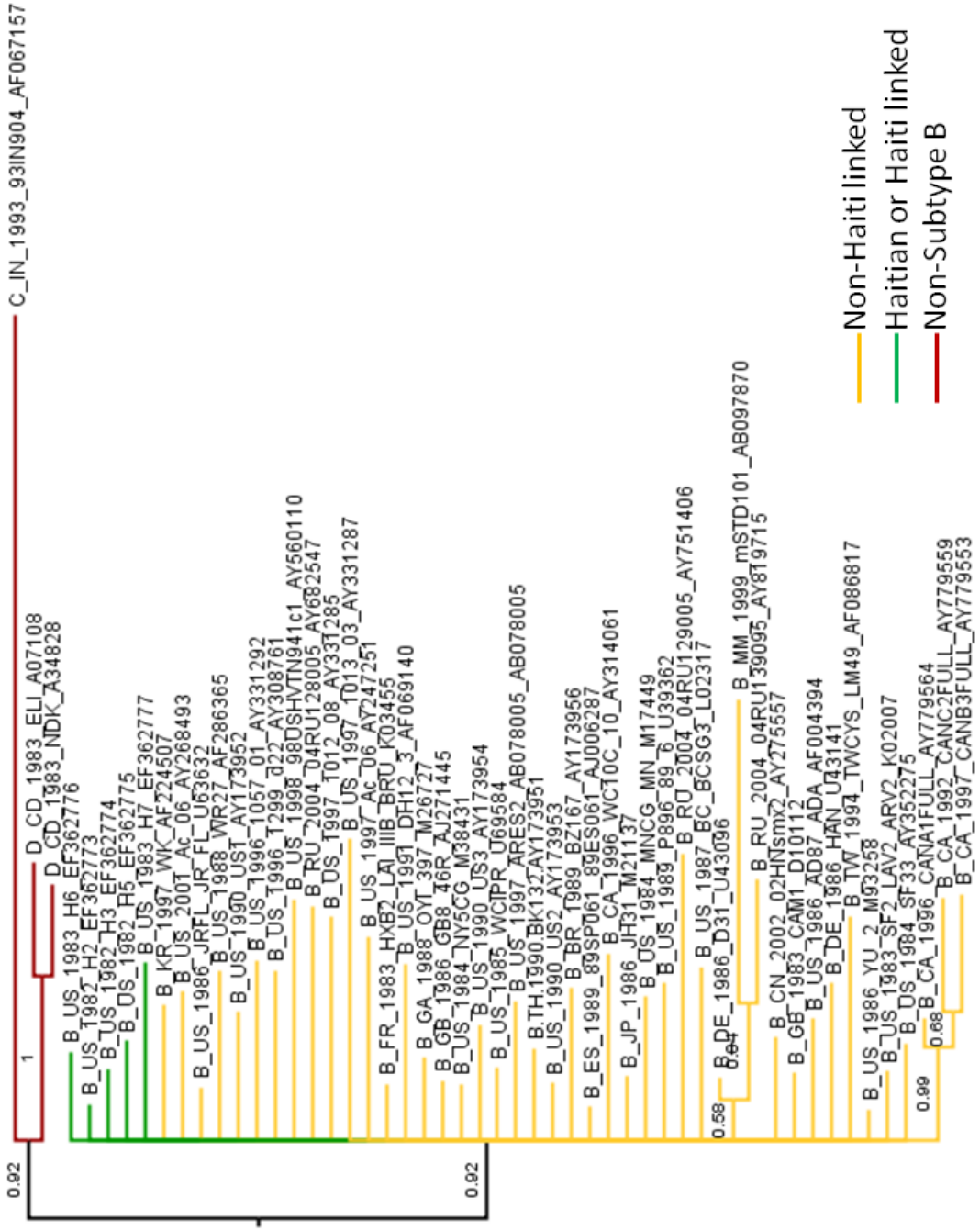
C


B_US_1983_H0_EF362776
B_US_1982_H2_EF362773
B_US_1982_H5_EF362775
B_KR_1997_WK_AF224507
B_US_1980_JRFL_JR_FL_U63632
B_US_1988_WR27_AF286385
B_US_1990_US1_AY173952
B_US_1990_1057_01_AY331292
B_US_1996_1290_d22_AY308761
B_RU_2004_04RU128005_AY682547
B_US_1997_1012_08_AY331285
B_US_1997_1013_03_AY331287
B_US_1997_Ac_06_AY247251
B_FR_1983_HXB2_LAI_IIIB_BRU_K03455
B_GA_1988_OY1_397_M26727
B_GB_1986_GB8_46R_AJ271445
B_US_1984_NY5CG_M38431
B_US_1985_WCIPR_U60584
B_US_1997_ARES2_AB078005_AB078005
B_BR_1989_BZ167_AY173956
B_ES_1989_89SP061_89ES061_AJ006287
B_US_1984_MNCG_MN_M17449
B_US_1989_P896_89_6_U38382
B_US_1987_BC_BCSG3_L02317
B_US_1986_AD87_ADA_AF004394
B_DE_1986_HAN_U43141
B_RU_2004_04RU139005_AY818715
B_TW_1994_TWCYS_LM49_AF086817
B_US_1988_YU_2_M93258
B_US_1984_SF33_AY352275
B_CA_1992_CANC2FULL_AY779559
B_US_1998_98USHVTN941c1_AY560110
B_US_1990_US3_AY173954
B_CA_1996_CANA1FULL_AY779564
D_CD_1983_ELI_A07108
D_CD_1983_NDK_A34828
B_US_1990_US2_AY173953
B_US_1983_SF2_LAV2_ARV2_K02007
B_US_1983_H7_EF362777
B_CA_1996_WC10C_10_AY314061
B_US_2001_Ac_06_AY268493
B_JP_1988_JH31_M21137
B_MM_1999_mSTD101_AB097870
B_CN_2002_02HNsmx2_AY275557
B_RU_2004_04RU129005_AY751406
B_US_1991_DH12_3_AF069140
B_CA_1997_CANB3FULL_AY779553
B_US_1982_H3_EF362774
B_GB_1983_CAM1_D10112
B.TH.1990.BK132.AY173951
B_DE_1986_D31_U43096
C_IN_1993_93IN904_AF067157
0.85
0.88
0.82
0.72
0.67
0.69
1
0.64
0.58
0.51
0.5
0.07
Non-Haiti linked
Haitian or Haiti linked
Non-Subtype B

D


Chain 1
Chain 2
lnL
State
-1830
-1840
-1850
-1860
-1870
-1880
-1890
-1900
-1910
-1920
-1930
0
2500000
5000000
7500000
10000000
1.25E7
1.5E7
1.75E7
2E7

E


C_IN_1993_93IN904_AF067157
D_CD_1983_ELI_A07108
D_CD_1983_NDK_A34828
B_US_1983_H6_EF362776
B_US_1982_H2_EF362773
B_US_1982_H3_EF362774
B_DS_1982_H5_EF362775
B_US_1983_H7_EF362777
B_KR_1997_WK_AF224507
B_US_2001_Ac_06_AY268493
B_US_1986_JRFL_JR_FL_U63632
B_US_1988_WR27_AF286365
B_US_1990_UST_AY173952
B_US_1996_1057_01_AY331292
B_US_1996_1299_d22_AY308761
B_US_1998_98DSHVTN941c1_AY560110
B_RU_2004_04RU128005_AY682547
B_DS_1997_T012_08_AY331285
B_US_1997_1013_03_AY331287
B_US_1997_Ac_06_AY247251
B_FR_1983_HXB2_LAI_IIIB_BRU_K03455
B_US_1991_DH12_3_AF069140
B_GA_1988_OYT_397_M26727
B_GB_1986_GB8_46R_AJ271445
B_US_1984_NY5CG_M38431
B_US_1990_US3_AY173954
B_US_1985_WCTPR_U69584
B_US_1997_ARES2_AB078005_AB078005
B.TH.1990.BK132_AY173951
B_US_1990_US2_AY173953
B_BR_1989_BZ167_AY173956
B_ES_1989_89SP061_89ES061_AJ006287
B_CA_1996_WC10C_10_AY314061
B_JP_1986_JH31_M21137
B_US_1984_MNCG_MN_M17449
B_US_1989_P896_89_6_U39362
B_RU_2004_04RU129005_AY751406
B_US_1987_EC_ECSG3_L02317
B_DE_1986_D31_U43096
B_MM_1999_mSTD101_AB097870
B_RU_2004_04RU139095_AY819715
B_CN_2002_02HNsmx2_AY275557
B_GB_1983_CAM1_D10112
B_US_1986_AD87_ADA_AF004394
B_DE_1986_HAN_U43141
B_TW_1994_TWCYS_LM49_AF086817
B_US_1986_YU_2_M93258
B_US_1983_SF2_LAV2_ARV2_K02007
B_DS_1984_SF33_AY352275
B_CA_1996_CANA1FULL_AY779564
B_CA_1992_CANC2FULL_AY779559
B_CA_1997_CANB3FULL_AY779553
1
0.92
0.92
0.58
0.99
0.68
0.04
Non-Haiti linked
Haitian or Haiti linked
Non-Subtype B

F

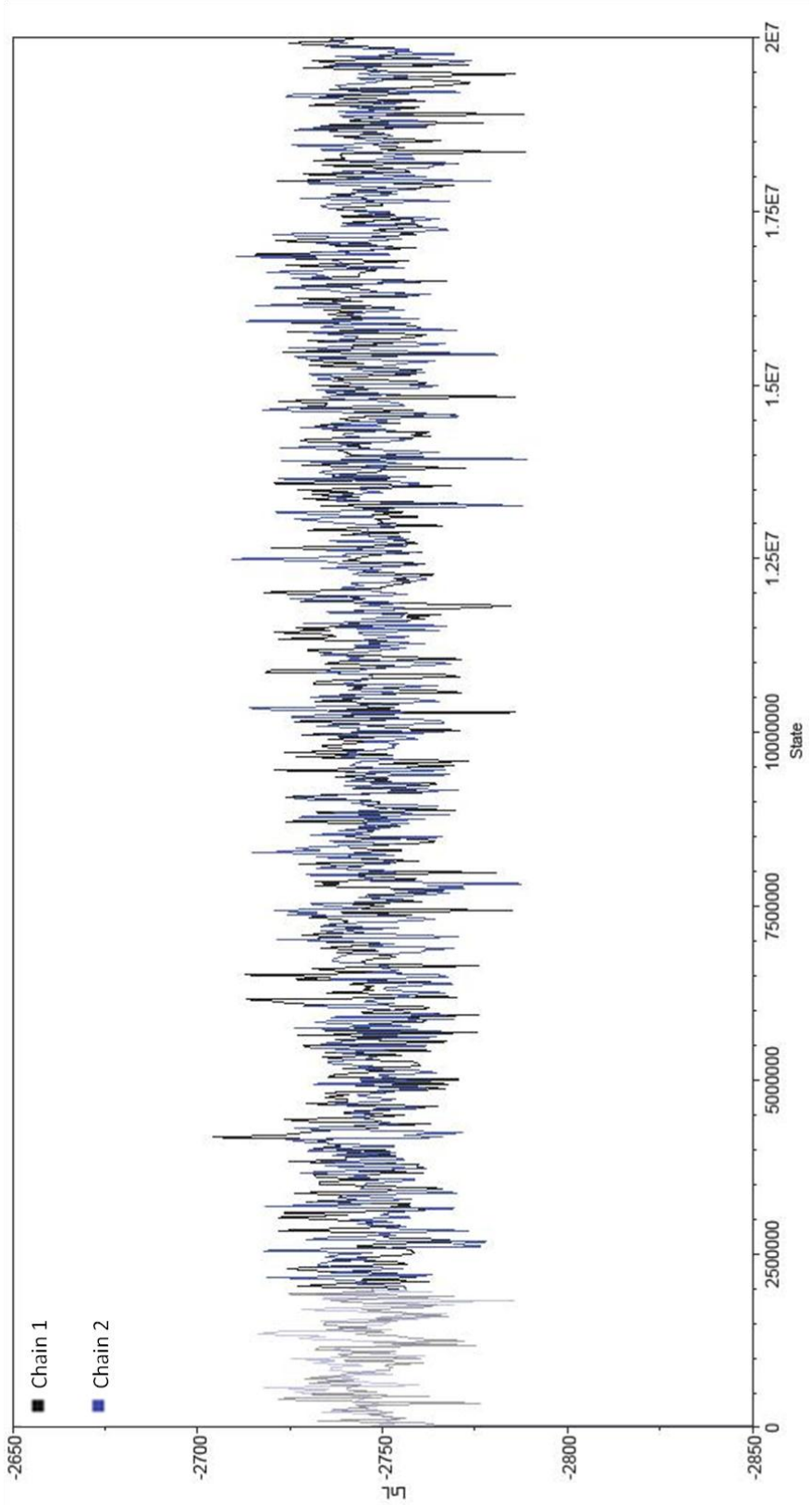
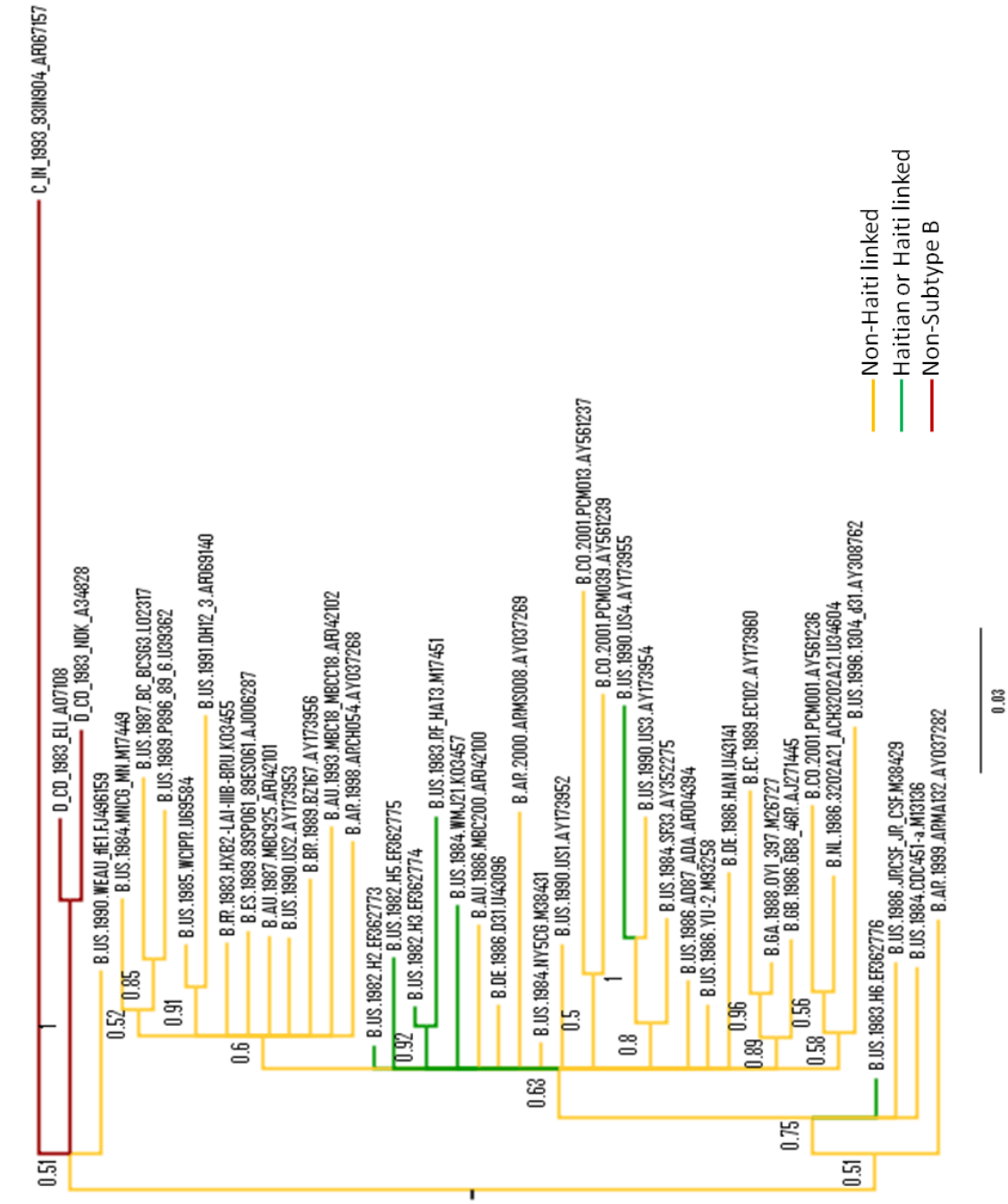

Chain 1
Chain 2
-2650
-2700
-2750
-2800
-2850
lnL
0
2500000
5000000
7500000
10000000
1.25E7
1.5E7
1.75E7
2E7
State

Figure 3. *Majority rule consensus trees for gag2 dataset constructed after recombination detection analysis. Phylogenetic trees, reconstructed in MrBayes, depict the topology (A) and Bayesian convergence results (B) of the non-recombinant gag2 dataset. Trees were constructed from HXB2 genomic positions 1198-1866. Non-subtype B sequences are represented in red, Haiti-originating and Haiti-linked sequences in green and all non-Caribbean subtype B sequences in gold.*

A


C_IN.1993_93IN904_AR067157
O_CD.1983_ELI_A07108
O_CD.1983_NDK_A34828
B.US.1990.WEAU_dE1.FJ496159
B.US.1984.MNCG_MN.M17449
B.US.1987.BC_BCS63.L02317
B.US.1989.P896_89_6.U39362
B.US.1985.WCIPP2.U69584
B.US.1991.DH12_3.AR069140
B.FR.1983.HXB2-LAI-IIIB-BRU.K03455
B.ES.1989.89SP061_89ES061.AJ006287
B.AU.1987.MBC925.AR042101
B.US.1990.US2.AY173953
B.BR.1989.BZ167.AY173956
B.AU.1993.MBC18_MBCC18.AR042102
B.AR.1998.ARCH054.AY037268
B.US.1982.H2.EF362773
B.US.1982.H5.EF362775
B.US.1982.H3.EF362774
B.US.1983.RF_HAT3.M17451
B.US.1984.WMJ21.K03457
B.AU.1986.MBC200.AR042100
B.DE.1986.D31.U43096
B.AR.2000.ARMS008.AY037269
B.US.1984.NY5CG.M38431
B.US.1990.US1.AY173952
B.CO.2001.PCM013.AY561237
B.CO.2001.PCM039.AY561239
B.US.1990.US4.AY173955
B.US.1990.US3.AY173954
B.US.1984.SF33.AY352275
B.US.1986.AD87_ADA.AR004394
B.US.1986.YU-2.M93258
B.DE.1986.HAN.U43141
B.EC.1989.EC102.AY173960
B.GA.1988.OYI_397.M26727
B.GB.1986.GB8_467.AJ271445
B.CO.2001.PCM001.AY561236
B.NL.1986.3202A21_ACH3202A21.U34604
B.US.1986.1304_d31.AY308762
B.US.1983.H6.EF362776
B.US.1986.JRCSF_JR_CSF.M38429
B.US.1984.CDC451-a.M13136
B.AR.1999.ARMA192.AY037282
0.03
Non-Haiti linked
Haitian or Haiti linked
Non-Subtype B

B

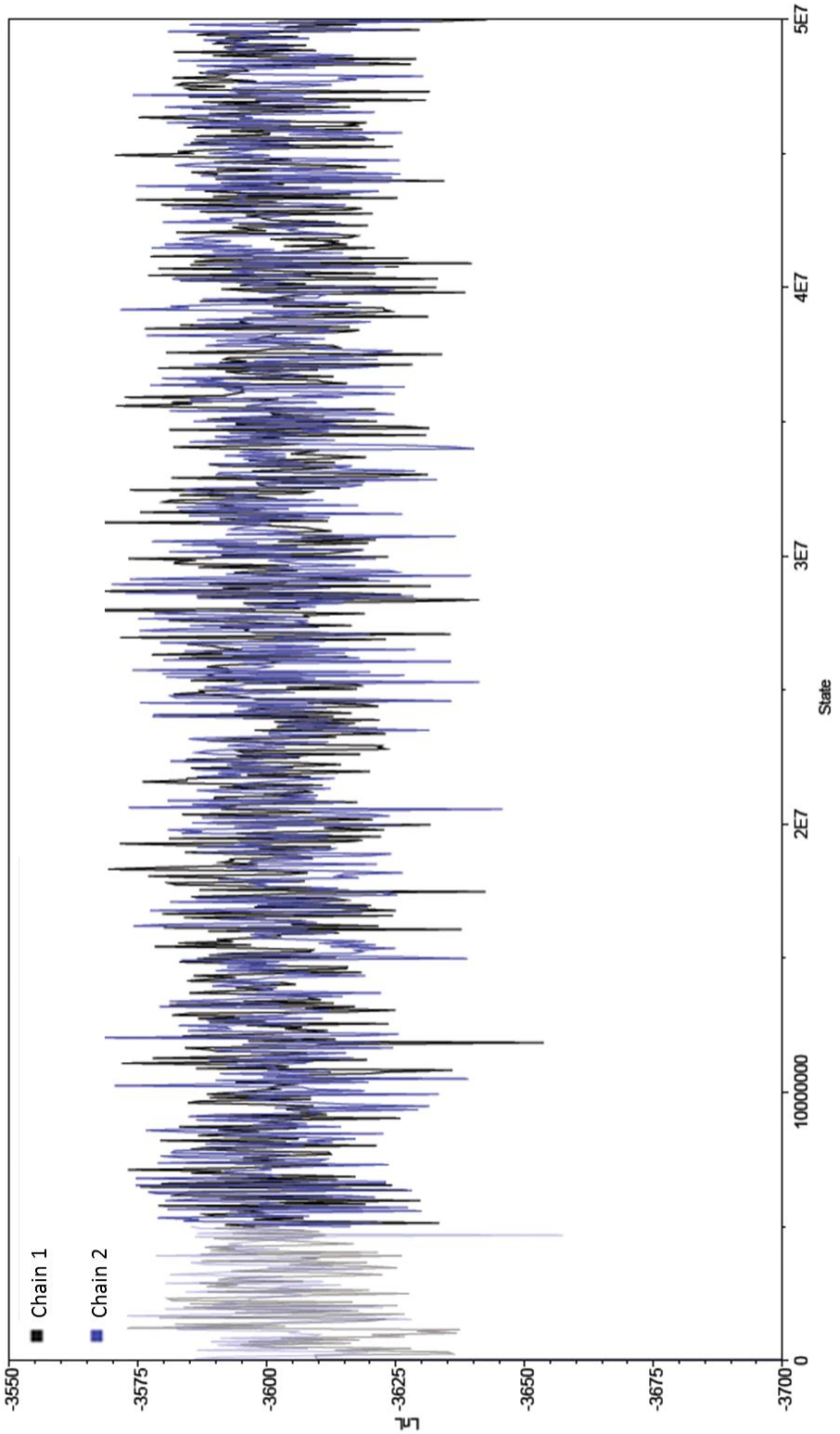
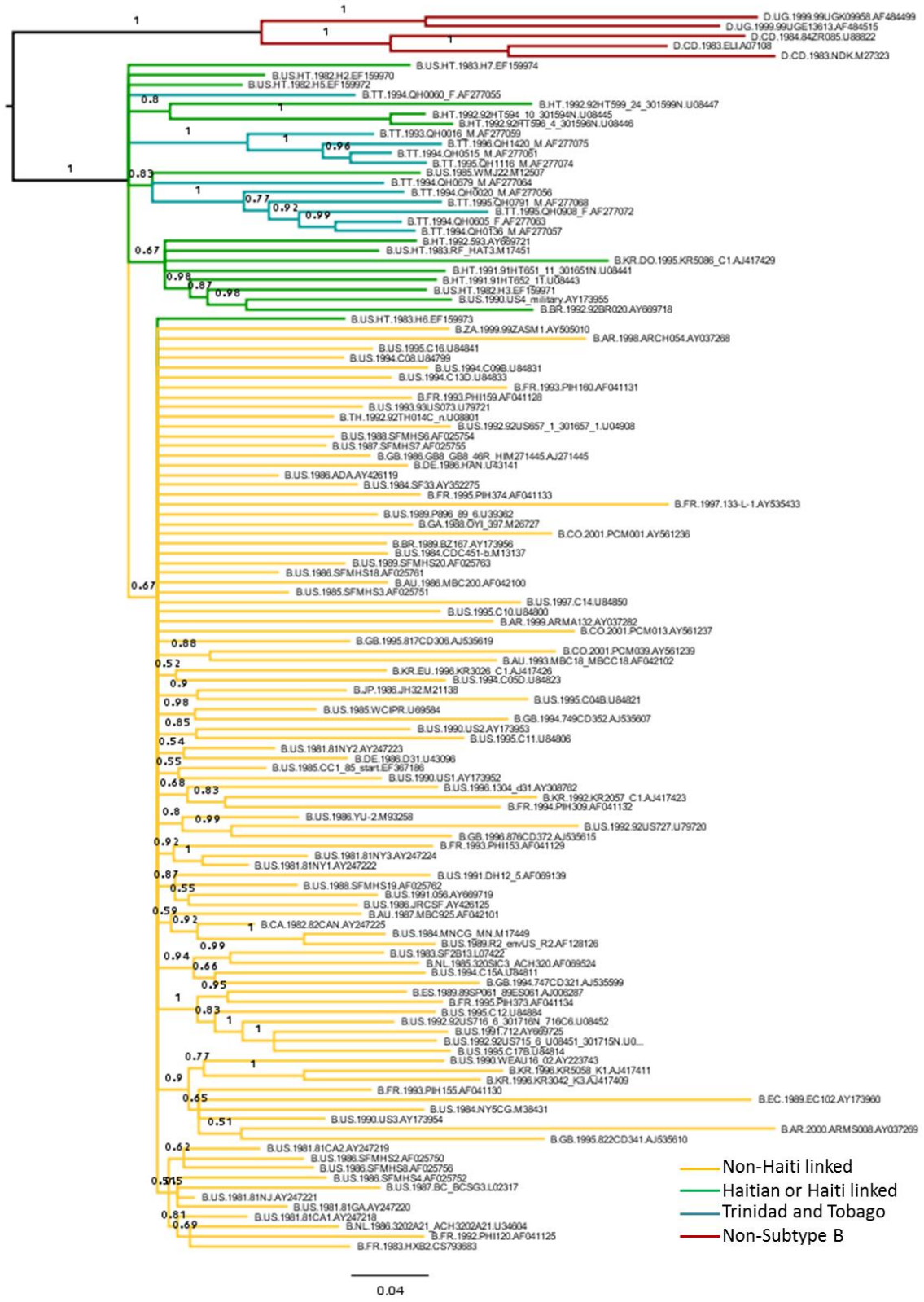
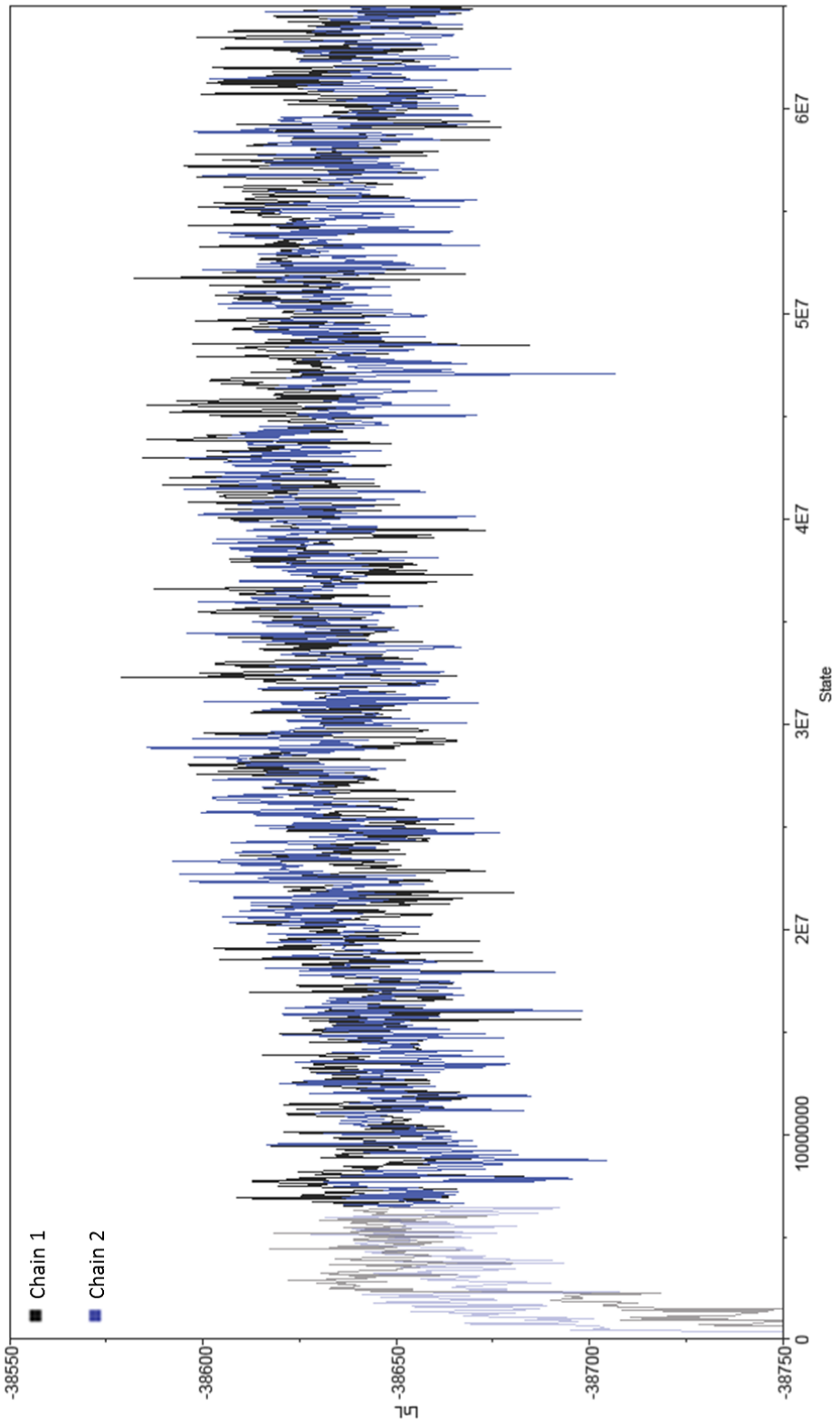

Chain 1
Chain 2
-3550
-3575
-3600
-3625
-3650
-3675
-3700
lnL
0
10000000
2E7
3E7
4E7
5E7
State

Figure 4. *Majority rule consensus trees for env constructed after recombination detection analysis. Phylogenetic trees, reconstructed in MrBayes, depict the topology (A, C and E) and Bayesian convergence results (B, D and F) of GARD-inferred non-recombinant segments within the env dataset. Trees were constructed from HXB2 genomic positions (A) 6231-7406, (C) 7407-7888 and (E) 7889-8795.Non-subtype B sequences are represented in red, Trinidad and Tobago-based sequences in blue, Haiti-originating and Haiti-linked sequences in green and all non-Caribbean subtype B sequences in gold.*

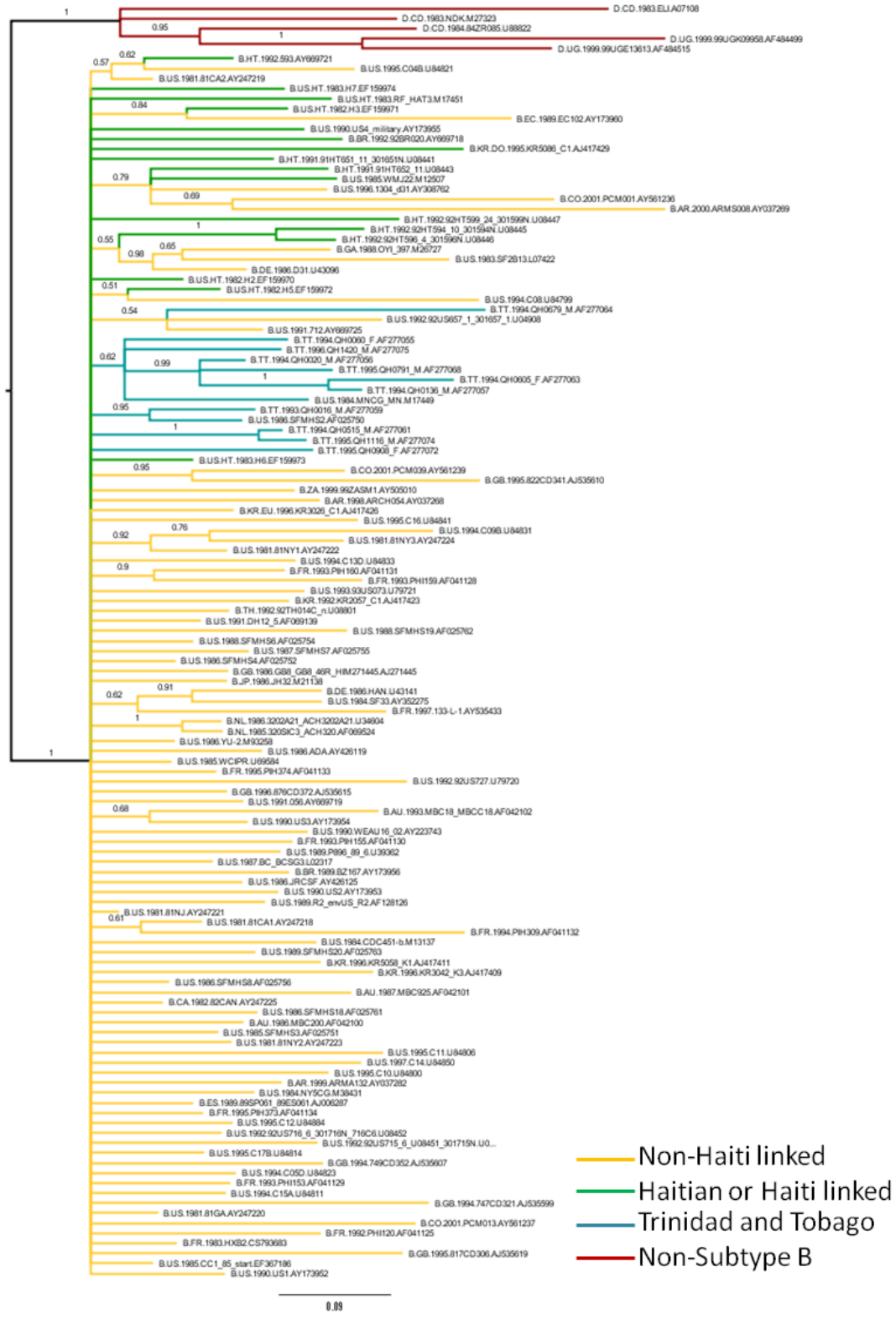
A


D.UG.1999.99UGK09958.AF484499
D.UG.1999.99UGE13613.AF484515
D.CD.1984.84ZR085.U88822
D.CD.1983.ELI.A07108
D.CD.1983.NDK.M27323
B.US.HT.1983.H7.EF159974
B.US.HT.1982.H2.EF159970
B.US.HT.1982.H5.EF159972
B.TT.1994.QH0060_F.AF277055
B.HT.1992.92HT599_24_301599N.U08447
B.HT.1992.92HT594_10_301594N.U08445
B.HT.1992.92HT596_4_301596N.U08446
B.TT.1993.QH0016_M.AF277059
B.TT.1996.QH1420_M.AF277075
B.TT.1994.QH0515_M.AF277061
B.TT.1995.QH1116_M.AF277074
B.US.1985.WMJ22.M12507
B.TT.1994.QH0679_M.AF277064
B.TT.1994.QH0020_M.AF277056
B.TT.1995.QH0791_M.AF277068
B.TT.1995.QH0908_F.AF277072
B.TT.1994.QH0605_F.AF277063
B.TT.1994.QH0136_M.AF277057
B.HT.1992.593.AY699721
B.US.HT.1983.RF_HAT3.M17451
B.KR.DO.1995.KR5086_C1.AJ417429
B.HT.1991.91HT651_11_301651N.U08441
B.HT.1991.91HT652_1T.U08443
B.US.HT.1982.H3.EF159971
B.US.1990.US4_military.AY173955
B.BR.1992.92BR020.AY669718
B.US.HT.1983.H6.EF159973
B.ZA.1999.99ZASM1.AY505010
B.AR.1998.ARCH054.AY037268
B.US.1995.C16.U84841
B.US.1994.C08.U84799
B.US.1994.C09B.U84831
B.US.1994.C13D.U84833
B.FR.1993.PIH160.AF041131
B.FR.1993.PIH159.AF041128
B.US.1993.93US073.U79721
B.TH.1992.92TH014C_n.U08801
B.US.1992.92US657_1_301657_1.U04908
B.US.1988.SFMHS6.AF025754
B.US.1987.SFMHS7.AF025755
B.GB.1986.GB8_GB8_46R_HIM271445.AJ271445
B.DE.1986.HAN.U43141
B.US.1986.ADA.AY426119
B.US.1984.SF33.AY352275
B.FR.1995.PIH374.AF041133
B.FR.1997.133-L-1.AY535433
B.US.1989.P896_89_6.U39362
B.GA.1988.OYI_397.M26727
B.CO.2001.PCM001.AY561236
B.BR.1989.BZ167.AY173956
B.US.1984.CDC-451-b.M13137
B.US.1989.SFMHS20.AF025763
B.US.1986.SFMHS18.AF025761
B.AU.1986.MBC200.AF042100
B.US.1985.SFMHS3.AF025751
B.US.1997.C14.U84850
B.US.1995.C10.U84800
B.AR.1999.ARMA132.AY037282
B.CO.2001.PCM013.AY561237
B.GB.1995.817CD306.AJ535619
B.CO.2001.PCM039.AY561239
B.AU.1993.MBC18_MBCC18.AF042102
B.KR.EU.1996.KR3026_C1.AJ417426
B.US.1994.C05D.U84823
B.JP.1986.JH32.M21138
B.US.1995.C04B.U84821
B.US.1985.WCIPR.U69584
B.GB.1994.749CD352.AJ535607
B.US.1990.US2.AY173953
B.US.1995.C11.U84806
B.US.1981.81NY2.AY247223
B.DE.1986.D31.U43096
B.US.1985.CC1_85_start.EF367186
B.US.1990.US1.AY173952
B.US.1996.1304_d31.AY308762
B.KR.1992.KR2057_C1.AJ417423
B.FR.1994.PIH309.AF041132
B.US.1986.YU-2.M93258
B.US.1992.92US727.U79720
B.GB.1996.876CD372.AJ535615
B.FR.1993.PIH153.AF041129
B.US.1981.81NY3.AY247224
B.US.1981.81NY1.AY247222
B.US.1991.DH12_5.AF069139
B.US.1988.SFMHS19.AF025762
B.US.1991.056.AY669719
B.US.1986.JRCSF.AY426125
B.AU.1987.MBC925.AF042101
B.CA.1982.82CAN.AY247225
B.US.1984.MNCG_MN.M17449
B.US.1989.R2_envUS_R2.AF128126
B.US.1983.SF2613.L07422
B.NL.1985.3203IC3_ACH320.AF069524
B.US.1994.C15A.U84811
B.GB.1994.747CD321.AJ535599
B.ES.1989.89SP061_89ES061.AJ006287
B.FR.1995.PIH373.AF041134
B.US.1995.C12.U84884
B.US.1992.92US716_6_301716N_716C6.U08452
B.US.1991.712.AY669725
B.US.1992.92US715_6_U08451_301715N.U0...
B.US.1995.C17B.U84814
B.US.1990.WEAU16_02.AY223743
B.KR.1996.KR5058_K1.AJ417411
B.KR.1996.KR3042_K3.AJ417409
B.FR.1993.PIH155.AF041130
B.EC.1989.EC102.AY173960
B.US.1984.NY5CG.M38431
B.US.1990.US3.AY173954
B.AR.2000.ARMS008.AY037269
B.GB.1995.822CD341.AJ535610
B.US.1981.81CA2.AY247219
B.US.1986.SFMHS2.AF025750
B.US.1986.SFMHS8.AF025756
B.US.1986.SFMHS4.AF025752
B.US.1987.BC_BCSG3.L02317
B.US.1981.81NJ.AY247221
B.US.1981.81GA.AY247220
B.US.1981.81CA1.AY247218
B.NL.1986.3202A21_ACH3202A21.U34604
B.FR.1992.PHI120.AF041125
B.FR.1983.HXB2.CS793683
Non-Haiti linked
Haitian or Haiti linked
Trinidad and Tobago
Non-Subtype B
0.04

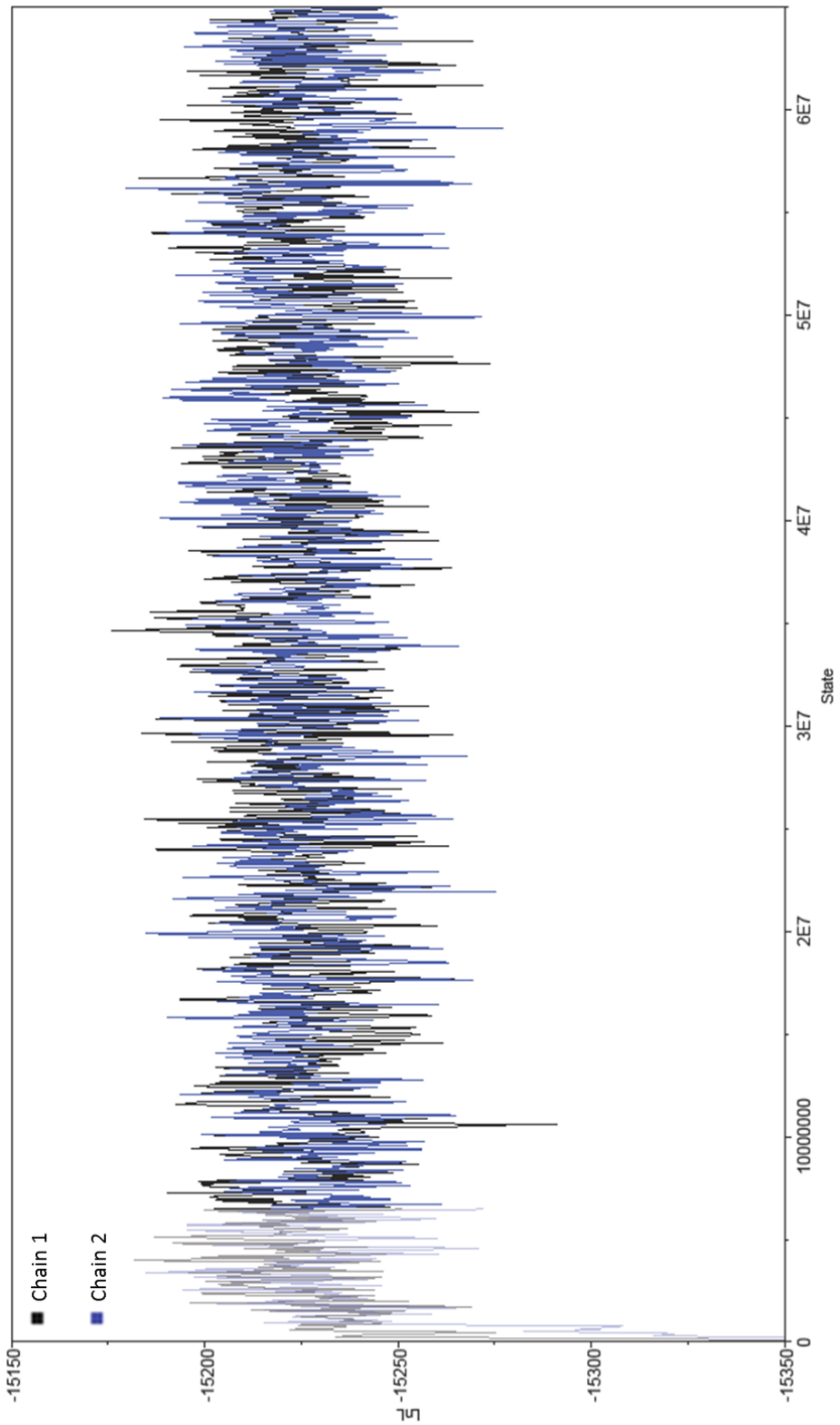
B


Chain 1
Chain 2
-38550
-38600
-38650
-38700
-38750
lnL
0
10000000
2E7
3E7
4E7
5E7
6E7
State

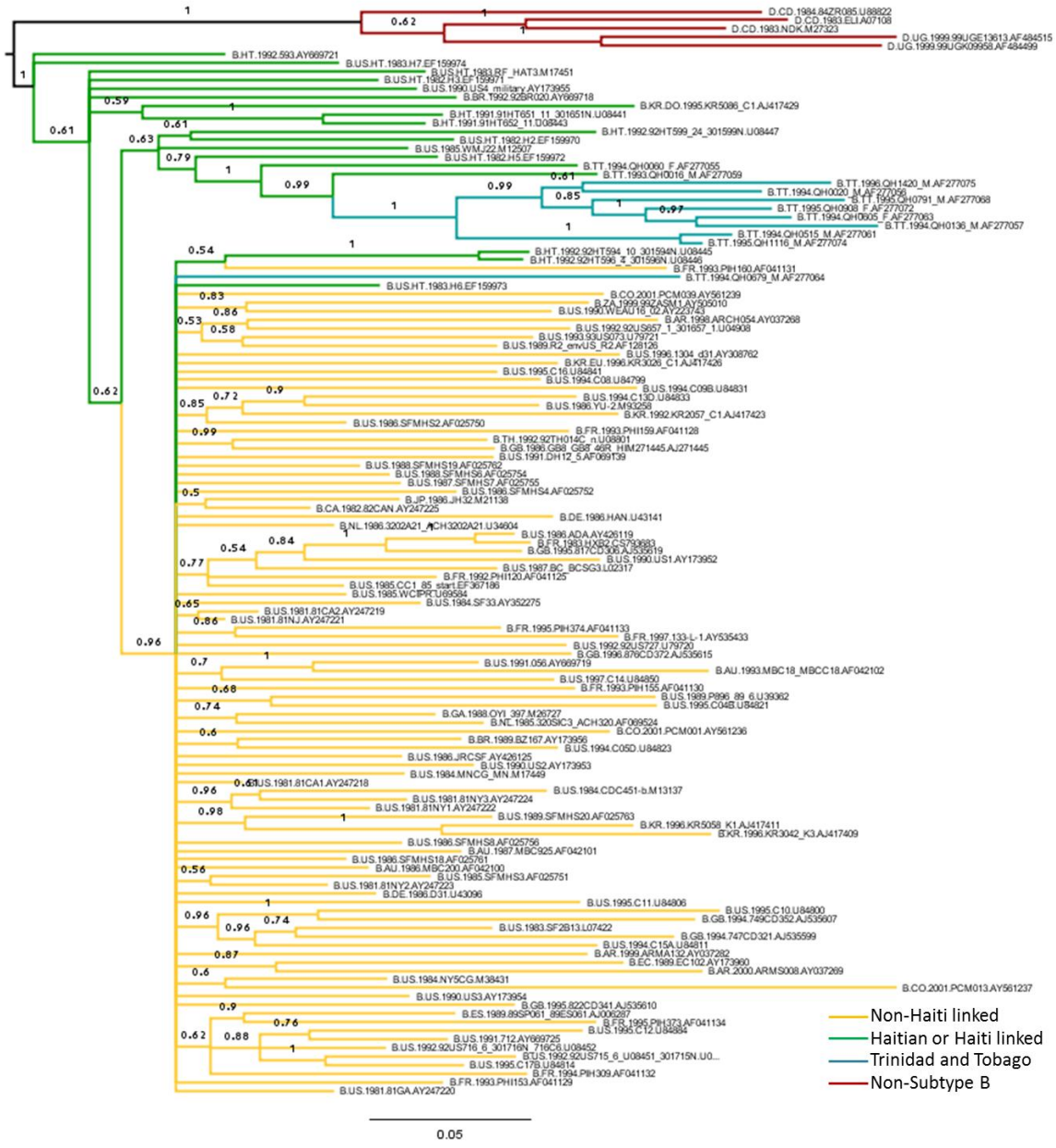
C


D.CD.1983.ELIA07108
D.CD.1983.NDK.M27323
D.CD.1984.84ZR085.U88822
D.UG.1999.99UGK09958.AF484499
D.UG.1999.99UGE13613.AF484515
B.HT.1992.593.AY669721
B.US.1995.C04B.U84821
B.US.1981.81CA2.AY247219
B.US.HT.1983.H7.EF159974
B.US.HT.1983.RF_HAT3.M17451
B.US.HT.1982.H3.EF159971
B.EC.1989.EC102.AY173960
B.US.1990.US4_military.AY173955
B.BR.1992.92BR020.AY669718
B.KR.DO.1995.KR5086_C1.AJ417429
B.HT.1991.91HT651_11_301651N.U08441
B.HT.1991.91HT652_11.U08443
B.US.1985.WM.J22.M12507
B.US.1996.1304_d31.AY308762
B.CO.2001.PCM001.AY561236
B.AR.2000.ARMS008.AY037269
B.HT.1992.92HT599_24_301599N.U08447
B.HT.1992.92HT594_10_301594N.U08445
B.HT.1992.92HT596_4_301596N.U08446
B.GA.1988.OYI_397.M26727
B.US.1983.SF2B13.L07422
B.DE.1986.D31.U43096
B.US.HT.1982.H2.EF159970
B.US.HT.1982.H5.EF159972
B.US.1994.C08.U84799
B.TT.1994.QH0679_M.AF277064
B.US.1992.92US657_1_301657_1.U04908
B.US.1991.712.AY669725
B.TT.1994.QH0060_F.AF277055
B.TT.1996.QH1420_M.AF277075
B.TT.1994.QH0020_M.AF277056
B.TT.1995.QH0791_M.AF277068
B.TT.1994.QH0605_F.AF277063
B.TT.1994.QH0136_M.AF277057
B.US.1984.MNCG_MN.M17449
B.TT.1993.QH0016_M.AF277059
B.US.1986.SFMHS2.AF025750
B.TT.1994.QH0515_M.AF277061
B.TT.1995.QH1116_M.AF277074
B.TT.1995.QH0908_F.AF277072
B.US.HT.1983.H6.EF159973
B.CO.2001.PCM039.AY561239
B.GB.1995.822CD341.AJ535610
B.ZA.1999.99ZASM1.AY505010
B.AR.1998.ARCH054.AY037268
B.KR.EU.1996.KR3026_C1.AJ417426
B.US.1995.C16.U84841
B.US.1994.C09B.U84831
B.US.1981.81NY3.AY247224
B.US.1981.81NY1.AY247222
B.US.1994.C13D.U84833
B.FR.1993.PIH160.AF041131
B.FR.1993.PHI159.AF041128
B.US.1993.93US073.U79721
B.KR.1992.KR2057_C1.AJ417423
B.TH.1992.92TH014C_n.U08801
B.US.1991.DH12_5.AF009139
B.US.1988.SFMHS19.AF025762
B.US.1988.SFMHS6.AF025754
B.US.1987.SFMHS7.AF025755
B.US.1986.SFMHS4.AF025752
B.GB.1986.GB8_GB8_46R_HIM271445.AJ271445
B.JP.1986.JH32.M21138
B.DE.1986.HAN.U43141
B.US.1984.SF33.AY352275
B.FR.1997.133-L-1.AY535433
B.NL.1986.3202A21_ACH3202A21.U34604
B.NL.1985.320SIC3_ACH320.AF009524
B.US.1986.YU-2.M93258
B.US.1986.ADA.AY426119
B.US.1985.WCIPR.U69584
B.FR.1995.PIH374.AF041133
B.US.1992.92US727.U79720
B.GB.1996.876CD372.AJ535615
B.US.1991.056.AY669719
B.AU.1993.MBC18_MBCC18.AF042102
B.US.1990.US3.AY173954
B.US.1990.WEAU16_02.AY223743
B.FR.1993.PIH155.AF041130
B.US.1989.P896_89_6.U39362
B.US.1987.BC_BCSG3.L02317
B.BR.1989.BZ167.AY173956
B.US.1986.JRCSF.AY426125
B.US.1990.US2.AY173953
B.US.1989.R2_envUS_R2.AF128126
B.US.1981.81NJ.AY247221
B.US.1981.81CA1.AY247218
B.FR.1994.PIH309.AF041132
B.US.1984.CDC451-b.M13137
B.US.1989.SFMHS20.AF025763
B.KR.1996.KR5058_K1.AJ417411
B.KR.1996.KR3042_K3.AJ417409
B.US.1986.SFMHS8.AF025756
B.AU.1987.MBC925.AF042101
B.CA.1982.82CAN.AY247225
B.US.1986.SFMHS18.AF025761
B.AU.1986.MBC200.AF042100
B.US.1985.SFMHS3.AF025751
B.US.1981.81NY2.AY247223
B.US.1995.C11.U84806
B.US.1997.C14.U84850
B.US.1995.C10.U84800
B.AR.1999.ARMA132.AY037282
B.US.1984.NY5CG.M38431
B.ES.1989.89SP061_89ESO61.AJ006287
B.FR.1995.PIH373.AF041134
B.US.1995.C12.U84894
B.US.1992.92US716_6_301716N_716C6.U08452
B.US.1992.92US715_6_U08451_301715N.U0...
B.US.1995.C17B.U84814
B.GB.1994.749CD352.AJ535607
B.US.1994.C05D.U84823
B.FR.1993.PHI153.AF041129
B.US.1994.C15A.U84811
B.GB.1994.747CD321.AJ535599
B.US.1981.81GA.AY247220
B.CO.2001.PCM013.AY561237
B.FR.1992.PHI120.AF041125
B.FR.1983.HXB2.CS793683
B.GB.1995.817CD306.AJ535619
B.US.1985.CC1_85_start.EF367186
B.US.1990.US1.AY173952
0.09
Non-Haiti linked
Haitian or Haiti linked
Trinidad and Tobago
Non-Subtype B

D


Chain 1
Chain 2
-15150
-15200
-15250
-15300
-15350
lnL
0
10000000
2E7
3E7
4E7
5E7
6E7
State

E


Non-Haiti linked
Haitian or Haiti linked
Trinidad and Tobago
Non-Subtype B
0.05

F

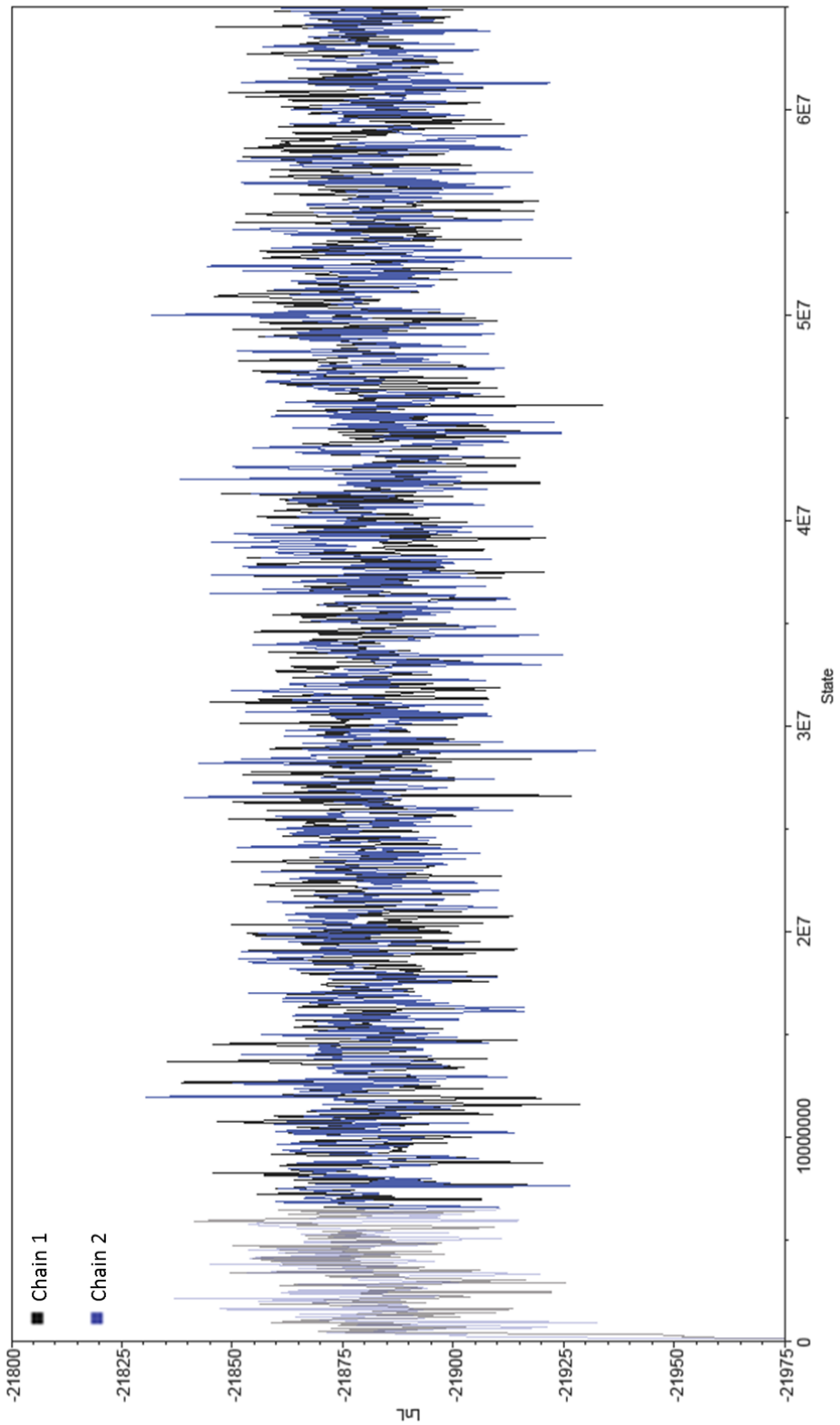
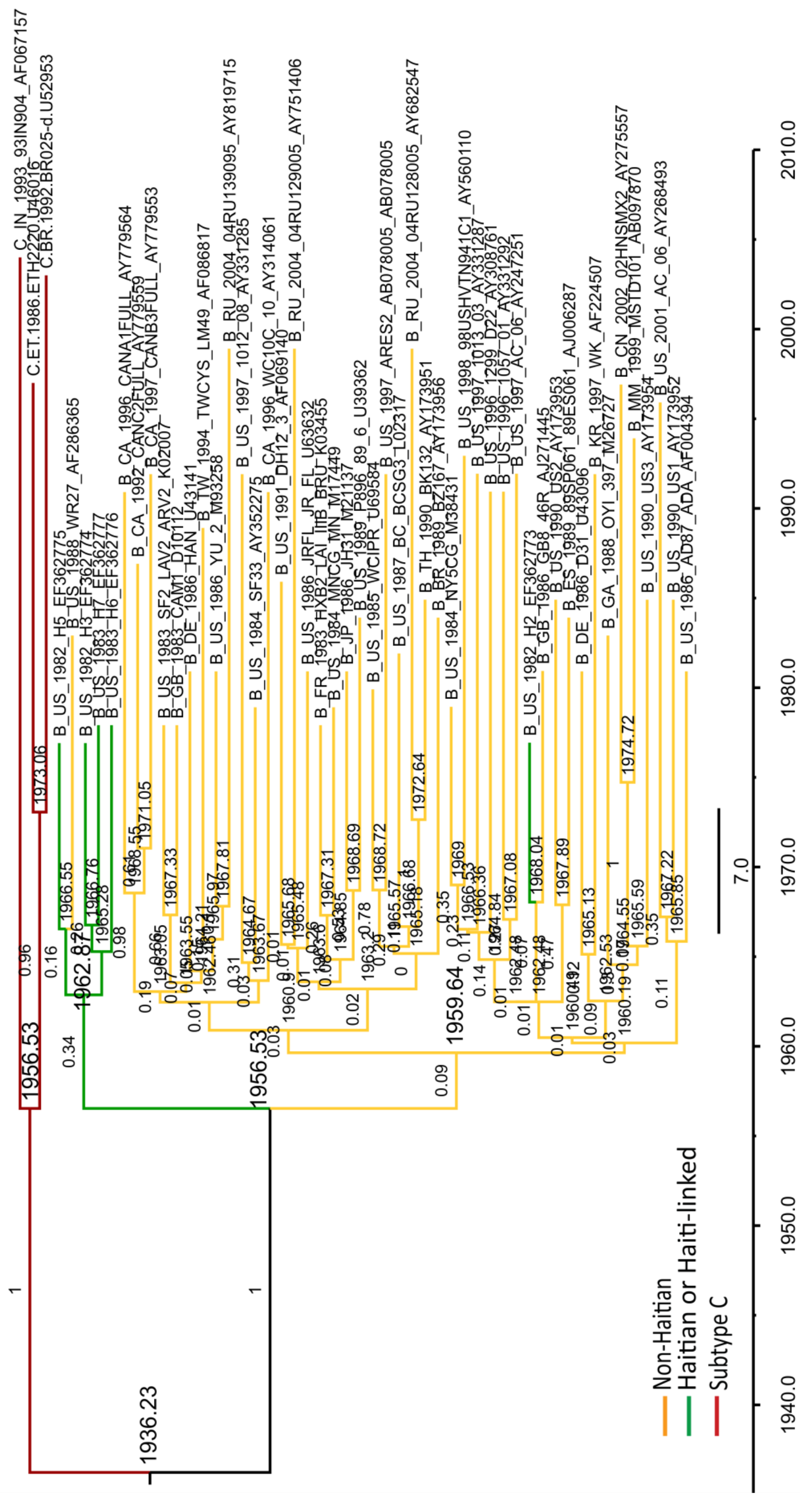

Chain 1
Chain 2
lnL
State
-21800
-21825
-21850
-21875
-21900
-21925
-21950
-21975
0
10000000
2E7
3E7
4E7
5E7
6E7

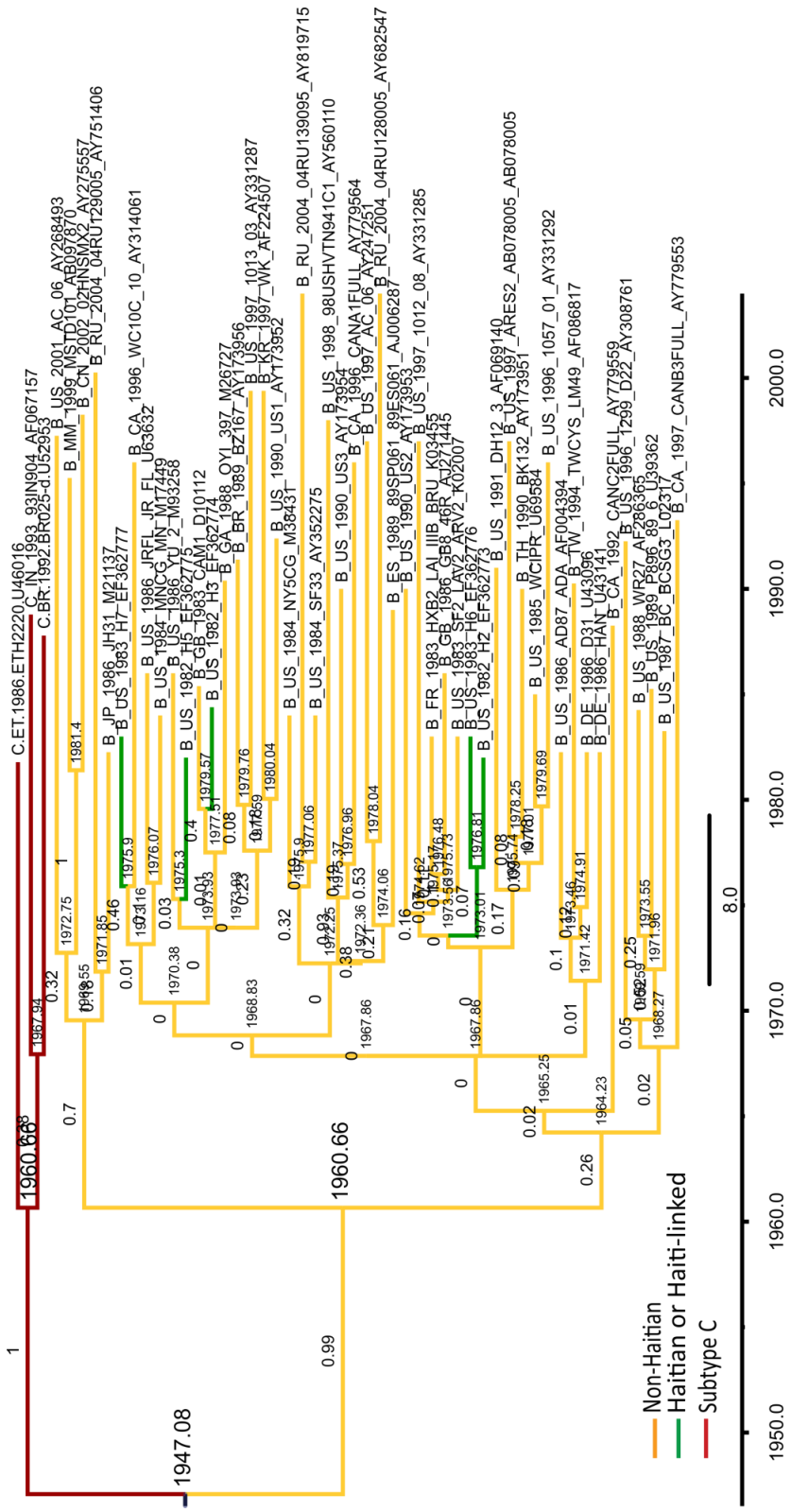
Figure 5. *Maximum clade credibility trees constructed using BEAST. Phylogenetic trees shown here depict the topology (A, B and C) of GARD-inferred non-recombinant segments within the gag1 dataset. Subtype C sequences were substituted for subtype D sequences as the subtype D sequences were clustering within the subtype B clade in trees constructed using MrBayes (1). Trees were constructed from HXB2 positions (A) 820-1272, (B) 1273-1614 and (C) 1615-1952. Non-subtype B sequences are represented in red, Haiti-originating and Haiti-linked sequences in green and all non-Caribbean subtype B sequences in gold.*

A



Non-Haitian
Haitian or Haiti-linked
Subtype C
7.0
1940.0
1950.0
1960.0
1970.0
1980.0
1990.0
2000.0
2010.0

B


C.ET.1986.ETH2220.U46016
C_IN_1993_93IN904_AF067157
C.BR.1992.BR025-d.U52953
B_US_2001_AC_06_AY268493
B_MM_1999_MSTD101_AB097870
B_CN_2002_02HNSMX2_AY275557
B_RU_2004_04RU129005_AY751406
B_JP_1986_JH31_M21137
B_US_1983_H7_EF362777
B_CA_1996_WC10C_10_AY314061
B_US_1986_JRFL_JR_FL_U63632
B_US_1984_MNCG_MN_M17449
B_US_1986_YU2_M93258
B_US_1982_H5_EF362775
B_GB_1983_CAM1_D10112
B_US_1982_H3_EF362774
B_GA_1988_OYI_397_M26727
B_BR_1989_BZ167_AY173956
B_US_1997_1013_03_AY331287
B_US_1997_WK_AF224507
B_US_1990_US1_AY173952
B_RU_2004_04RU139095_AY819715
B_US_1984_NY5CG_M38431
B_US_1998_98USHVTN941C1_AY560110
B_US_1984_SF33_AY352275
B_US_1990_US3_AY173954
B_CA_1996_CANA1FULL_AY779564
B_US_1997_AC_06_AY247251
B_RU_2004_04RU128005_AY682547
B_ES_1989_89SP061_89ESP061_AJ006287
B_US_1990_US2_AY173953
B_US_1997_1012_08_AY331285
B_FR_1983_HXB2_LAI_IIIB_BRU_K03455
B_GB_1986_GB8_46R_AJ271445
B_US_1983_SF2_LAV2_ARV2_K02007
B_US_1983_H6_EF362776
B_US_1982_H2_EF362773
B_US_1991_DH12_3_AF069140
B_US_1997_ARES2_AB078005_AB078005
B_TH_1990_BK132_AY173951
B_US_1985_WCIPR_U69584
B_US_1996_1057_01_AY331292
B_US_1986_AD87_ADA_AF004394
B_TW_1994_TWCYS_LM49_AF086817
B_DE_1986_D31_U43096
B_DE_1986_HAN_U43141
B_CA_1992_CANC2FULL_AY779559
B_US_1996_1299_D22_AY308761
B_US_1988_WR27_AF286365
B_US_1989_P896_89_6_U39362
B_US_1987_BC_BCSG3_LO2317
B_CA_1997_CANB3FULL_AY779553
1960.66
1947.08
Non-Haitian
Haitian or Haiti-linked
Subtype C
8.0
1950.0
1960.0
1970.0
1980.0
1990.0
2000.0

C

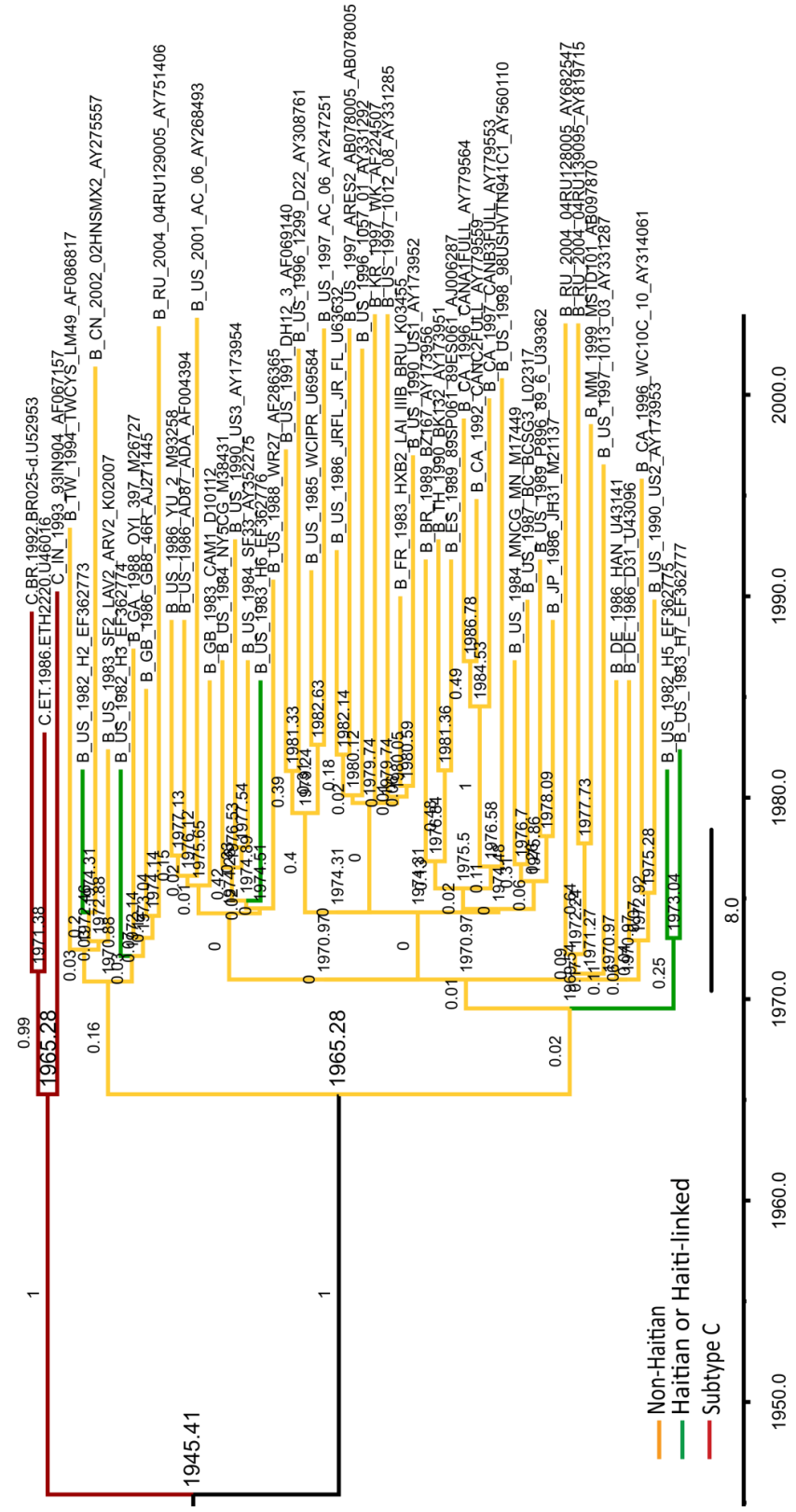

C.BR.1992.BR025-d.U52953
C.ET.1986.ETH2220.U46016
B_TW_1994_TWCYS_LM49_AF086817
B_CN_2002_02HNSMX2_AY275557
B_US_1982_H2_EF362773
B_US_1983_SF2_LAV2_ARV2_K02007
B_US_1982_H3_EF362774
B_GA_1988_OYI_397_M26727
B_GB_1986_GB8_46R_AJ271445
B_RU_2004_04RU129005_AY751406
B_US_1986_YU_2_M93258
B_US_1986_AD87_ADA_AF004394
B_US_2001_AC_06_AY268493
B_GB_1983_CAM1_D10112
B_US_1984_NY5CG_M38431
B_US_1990_US3_AY173954
B_US_1984_SF33_AY352276
B_US_1983_H6_EF362776
B_US_1988_WR27_AF286365
B_US_1991_DH12_3_AF069140
B_US_1996_1299_D22_AY308761
B_US_1985_WCIPR_U69584
B_US_1997_AC_06_AY247251
B_US_1986_JRFL_JR_FL_U63632
B_US_1997_ARES2_AB078005_AB078005
B_US_1996_1057_01_AY331284
B_US_1997_WK_AF224503
B_US_1997_1012_08_AY331285
B_FR_1983_HXB2_LAI_IIIB_BRU_K03455
B_BR_1989_BZ167_AY173956
B_TH_1990_BK132_AY173951
B_ES_1989_89SP061_89ES061_AJ006287
B_CA_1996_CANA1FULL_AY779564
B_CA_1992_CANC2FUL_AY779559
B_CA_1997_CANB3FULL_AY779553
B_US_1998_98USHVTN941C1_AY560110
B_US_1984_MNCG_MN_M17449
B_US_1987_BC_BCSG3_L02317
B_US_1989_P896_89_6_U39362
B_JP_1986_JH31_M21137
B_RU_2004_04RU128005_AY682547
B_RU_2004_04RU139095_AY819715
B_MM_1999_MSTD101_AB097870
B_US_1997_1013_03_AY331287
B_DE_1986_HAN_U43141
B_DE_1986_D31_U43096
B_US_1990_US2_AY173953
B_CA_1996_WC10C_10_AY314061
B_US_1982_H5_EF362775
B_US_1983_H7_EF362777
1945.41
1965.28
1965.28
1971.38
1970.88
1974.31
1972.88
1974.14
1974.04
1977.13
1976.12
1975.65
1974.61
1977.54
1976.53
1974.89
1981.33
1982.63
1982.14
1980.12
1979.74
1980.59
1980.05
1970.97
1974.31
1976.84
1981.36
1975.5
1984.53
1986.78
1976.58
1974.18
1976.7
1978.09
1977.73
1971.27
1970.97
1972.92
1975.28
1973.04
0.99
1
0.03
0.2
0.16
0.01
0.42
0.39
0.4
0.18
0.02
0.49
1
0.11
0.06
0.09
0.25
Non-Haitian
Haitian or Haiti-linked
Subtype C
8.0
1950.0
1960.0
1970.0
1980.0
1990.0
2000.0

Figure 6. *Maximum clade credibility tree for the gag2 dataset constructed from the BEAST analysis. Trees were constructed from HXB2 positions 1198-1866. Non-subtype B sequences are represented in red, Haiti-originating and Haiti-linked sequences in green and all non-Caribbean subtype B sequences in gold.*

A

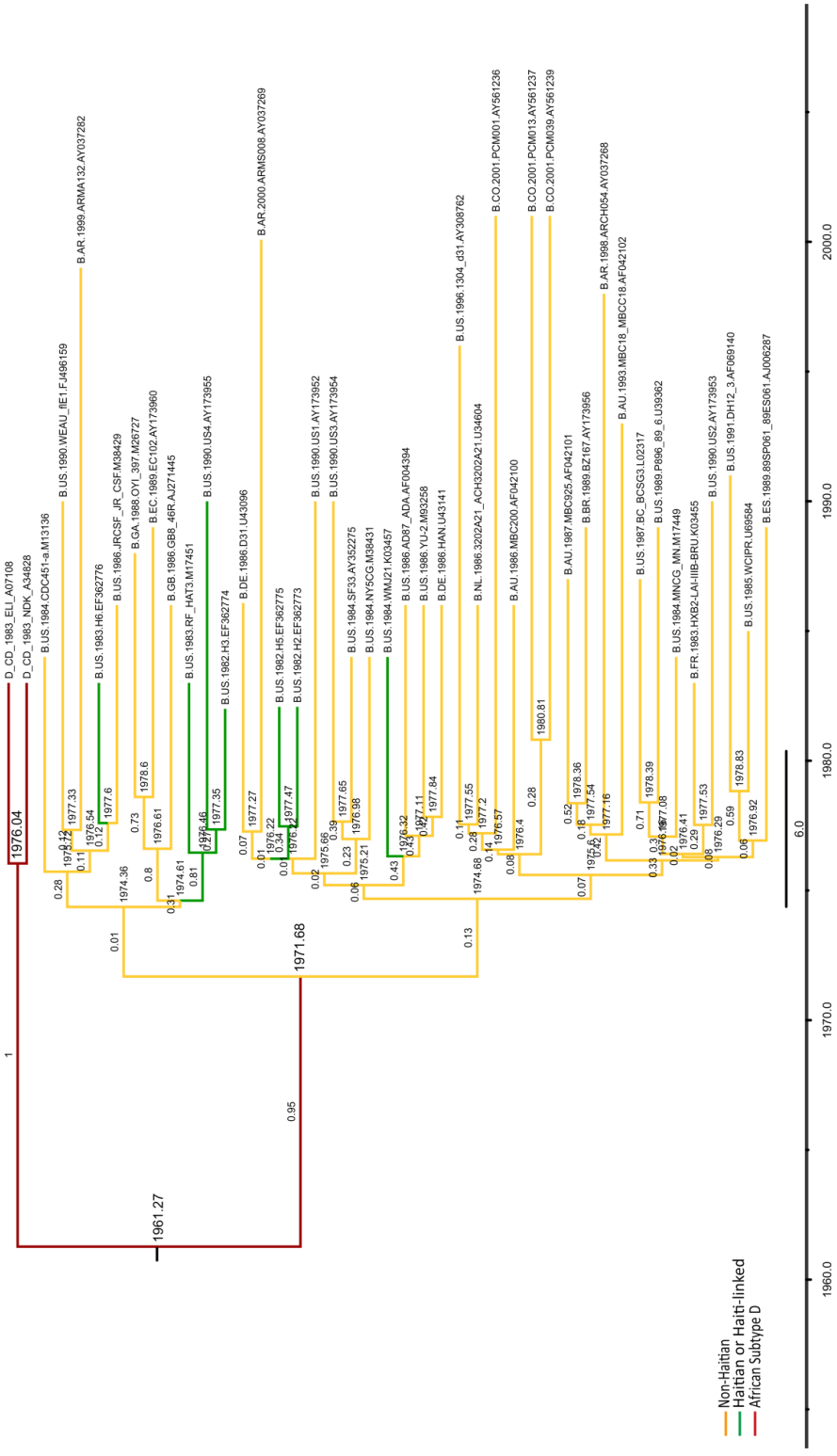
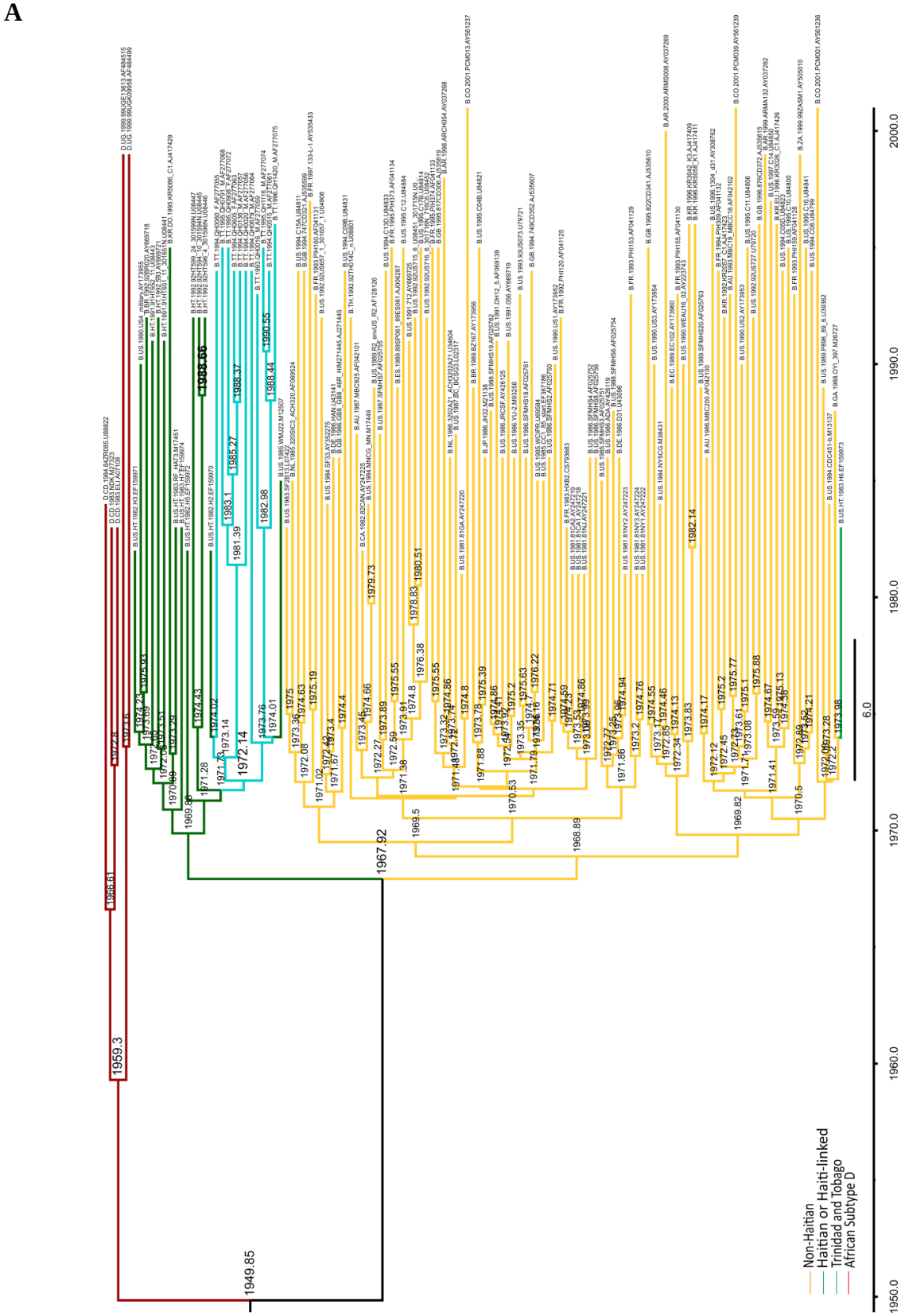
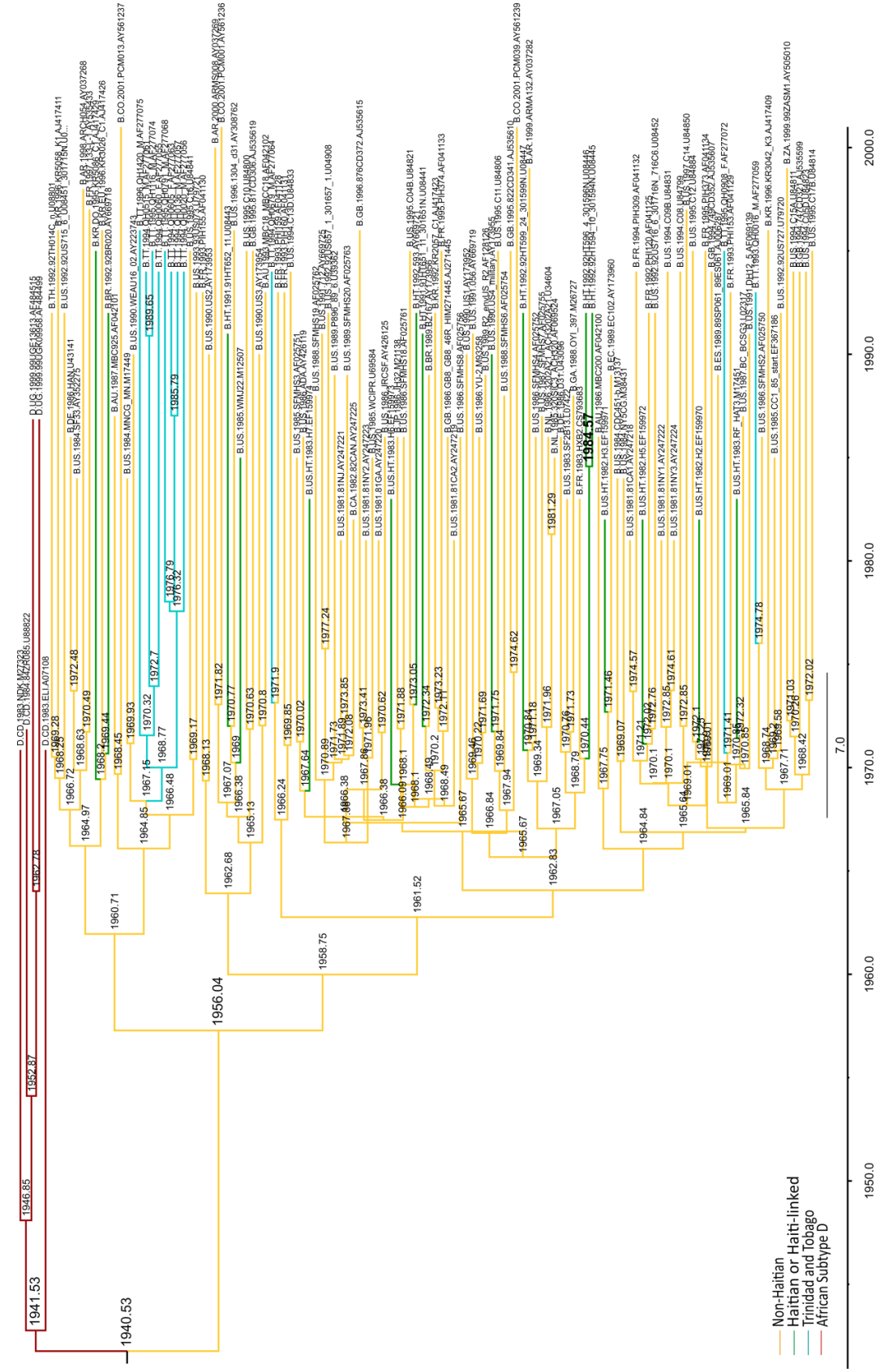

D_CD_1983_ELI_A07108
D_CD_1983_NDK_A34828
B.US.1984.CDC451-a.M13136
B.US.1990.WEAU_fIE1.F.U496159
B.AR.1999.ARMA132.AY037282
B.US.1983.H6.EF362776
B.US.1986.JRCSF_JR_CSF.M38429
B.GA.1988.OYI_397.M26727
B.EC.1989.EC102.AY173960
B.GB.1986.GB8_46R.AJ271445
B.US.1983.RF_HAT3.M17451
B.US.1990.US4.AY173955
B.US.1982.H3.EF362774
B.DE.1986.D31.U43096
B.AR.2000.ARMS008.AY037269
B.US.1982.H5.EF362775
B.US.1982.H2.EF362773
B.US.1990.US1.AY173952
B.US.1990.US3.AY173954
B.US.1984.SF33.AY352275
B.US.1984.NY5CG.M38431
B.US.1984.WMJ21.K03457
B.US.1986.AD87_ADA.AF004394
B.US.1986.YU-2.M93258
B.DE.1986.HAN.U43141
B.US.1996.1304_d31.AY308762
B.NL.1986.3202A21_ACH3202A21.U34604
B.CO.2001.PCM001.AY561236
B.CO.2001.PCM013.AY561237
B.CO.2001.PCM039.AY561239
B.AU.1986.MBC200.AF042100
B.AU.1987.MBC925.AF042101
B.BR.1989.BZ167.AY173956
B.AR.1998.ARCH054.AY037268
B.AU.1993.MBC18_MBCC18.AF042102
B.US.1987.BC_BCSG3.L02317
B.US.1989.P896_89_6.U39362
B.US.1984.MNCG_MN.M17449
B.FR.1983.HXB2-LAI-IIIB-BRU.K03455
B.US.1990.US2.AY173953
B.US.1991.DH12_3.AF069140
B.US.1985.WCIPR.U69584
B.ES.1989.89SP061_89ES061.AJ006287
1976.04
1971.68
1961.27
1960.0
1970.0
1980.0
1990.0
2000.0
6.0
Non-Haitian
Haitian or Haiti-linked
African Subtype D

Figure 7. *Maximum Clade Credibility trees for the env dataset constructed from the BEAST analysis. Phylogenetic trees shown here depict the topology (A, B and C) of GARD-inferred non-recombinant segments within the env dataset. Trees were constructed from HXB2 positions (A) 6231-7406, (B) 7407-7888 and (C) 7889-8795. Non-subtype B sequences are represented in red, Trinidad and Tobago-based sequences in blue, Haiti-originating and Haiti-linked sequences in green and all non-Caribbean subtype B sequences in gold.*

A


1950.0
1960.0
1970.0
1980.0
1990.0
2000.0
6.0
Non-Haitian
Haitian or Haiti-linked
Trinidad and Tobago
African Subtype D

B


Non-Haitian
Haitian or Haiti-linked
Trinidad and Tobago
African Subtype D
1950.0
1960.0
1970.0
1980.0
1990.0
2000.0
7.0

C

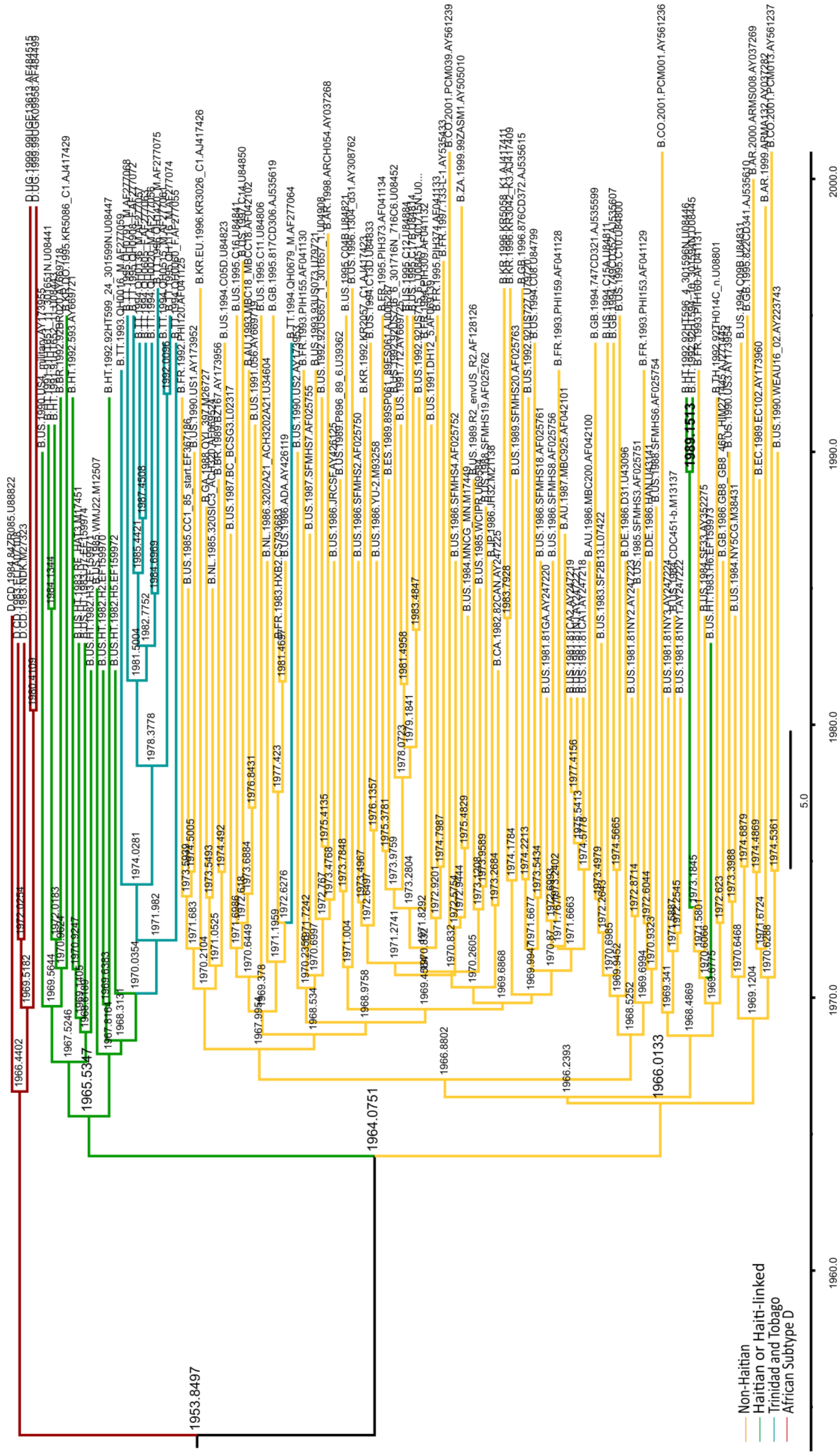
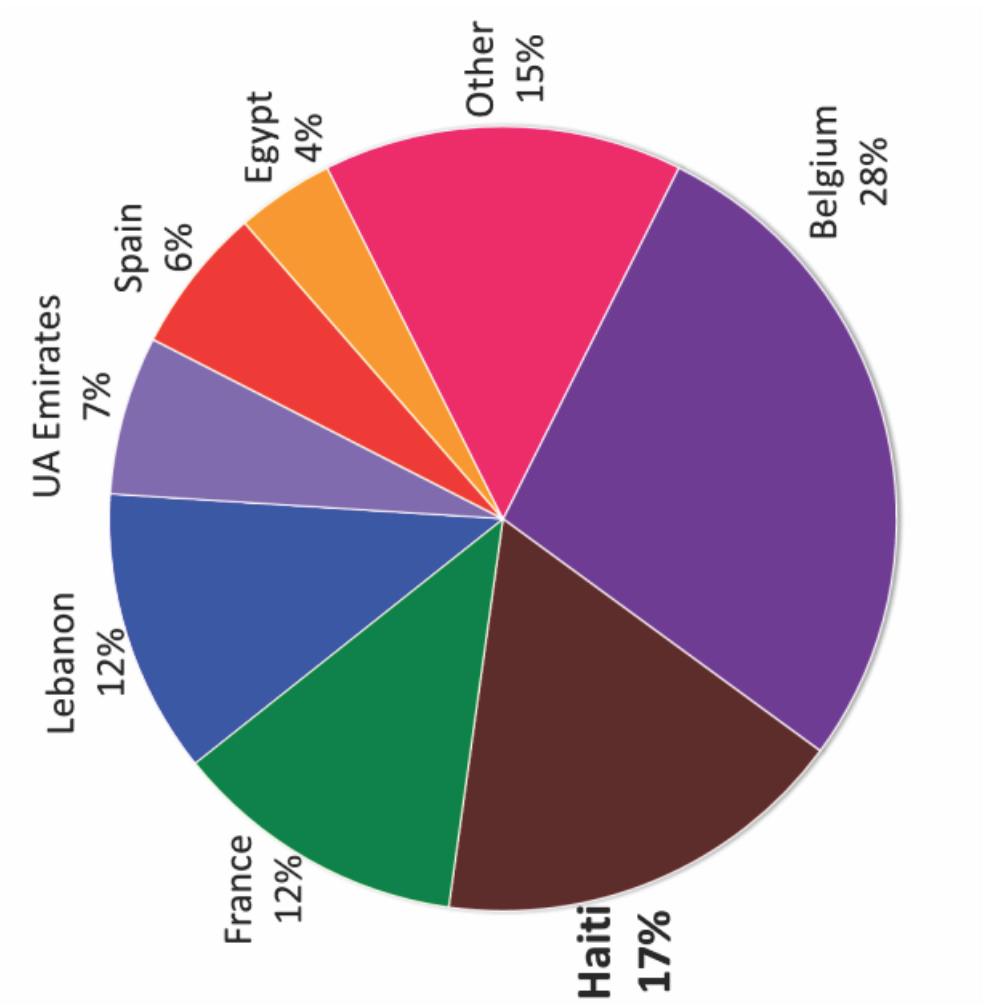

Non-Haitian
Haitian or Haiti-linked
Trinidad and Tobago
African Subtype D
1960.0
1970.0
1980.0
1990.0
2000.0
5.0

Figure 8. *Distribution of UNESCO teachers by nationality recruited to the Congo region between 1960-1964. Data used to construct this table and chart were extracted from the UNESCO archival database.*



Countries	# Teachers
Belgium	220
Haiti	136
France	96
Lebanon	93
United Arab Emirates	52
Spain	48
Egypt	32
Italy	25
Canada	22
Syria	20
China	10
Greece	9
Vietnam	7
Switzerland	4
Hungary	3
Netherlands	3
Afghanistan	2
Chile	2
United Kingdom	2
Germany	1
Columbia	1
Luxembourg	1
Mexico	1
Norway	1
Rwanda	1
Sweden	1
USA	1
Total	794

APPENDIX III

SUPPORTING TABLES FOR CHAPTER 3

Table 1

Individual	ALL	PGlyRem	Dif	ALL^N (X)	$PGlyRem^N$ (Y)
1	3.006425	1.50471	1.501716	0.022	0.011
2	9.979484	3.580432	6.399052	0.035	0.013
3	4.003991	1.004603	2.999388	0.023	0.012
4	7.605062	3.026138	4.578924	0.040	0.014
5	6.407755	2.522747	3.885008	0.013	0.006
6	3.13315	0.911929	2.221221	0.077	0.028
7	4.09977	2.735111	1.364659	0.015	0.007
8	4.436364	1.891936	2.544428	0.050	0.018
9	3.415556	1.793339	1.622217	0.022	0.011
10	3.743879	2.114634	1.629246	0.049	0.018
11	1.5977	0.515621	1.082079	0.058	0.029

Sum of branch lengths of ML trees for each individual reconstructed from the PGlyRem versus ALL dataset and the difference between them. The first set of ALL and PGlyRem values are the raw values from TreeRate.

[N] The second set of ALL and PGlyRem values are the raw values normalized by the number of sequences in each dataset, which effectively yields an average branch length. This second set of values was used to in the simple linear regression.

Table 2

Individual	ALL (X)	PGlyRem (Y)	Dif
1	3.006425	1.50471	1.501716
2	9.979484	3.580432	6.399052
3	4.003991	1.004603	2.999388
4	7.605062	3.026138	4.578924
5	6.407755	2.522747	3.885008
6	3.13315	0.911929	2.221221
7	4.09977	2.735111	1.364659
8	4.436364	1.891936	2.544428
9	3.415556	1.793339	1.622217
10	3.743879	2.114634	1.629246
11	1.5977	0.515621	1.082079

Sum of branch lengths of ML trees for each individual reconstructed from the PGlyRem or ALL dataset. The difference (Dif) between the sum of branch lengths for PGlyRem and ALL is also shown.

Table 3

Dataset	Mean	Std. Dev	N
PGlyRem	1.964	0.951	11
ALL	4.675	2.401	11
Dif	2.712	0.493	
PGlyRem^{N}	0.015	0.008	11
ALL^{N}	0.037	0.020	11
Dif^{N}	0.022	0.004	

Descriptive statistics for the PGlyRem and ALL sum of branch lengths and the difference (DIF).
^{N} The second set of ALL, PGlyRem and Dif values are the raw values normalized by the number of sequences in each dataset, which effectively yields an average branch length.

Table 4

	Statistic	df	Sig.
Dif	0.908	11	0.231

Testing normality for the paired t-test. Shapiro-Wilk test of normality using the differences between the normalized sum of branch lengths of PGlyRem versus ALL trees. Since the value p = 0.231 is greater than p = 0.05, the differences, between the sum of branch lengths between the ALL and PGlyRem dataset, do not statistically significantly deviate from normality.

Table 5

Individual	Stable Glycosylation Site Conservation											
	N276	N295	N301	N332	N339	N356	N362	N386	N392	N397	N401	N448
2	95.34	97.41	100	98.96	79.27	98.96	92.75	95.85	90.67	75.65	84.97	99.48
5	95.65	98.55	100	100	99.27	99.27	95.65	92.02	99.27	97.1	97.1	99.27
6	98.31	69.49	100	100	96.61	100	98.305	94.91	94.91	69.49	93.22	100
7	96.83	98.41	63.43	73.01	92.06	100	61.9	79.37	96.83	96.83	92.06	96.83
8	98.25	98.25	100	96.49	73.68	87.72	0	80.7	100	38.6	98.25	96.49
	Ranked Glycosylation Sites											
2	7	5	1	3	11	4	8	6	9	12	10	2
5	10	7	1	2	6	4	11	12	5	9	8	3
6	5	10	1	2	7	3	6	8	9	12	11	4
7	5	2	11	10	7	1	12	9	4	6	8	3
8	3	4	1	7	10	8	12	9	2	11	5	6
Total Ranks	30	28	15	24	41	20	49	44	29	50	42	18

Stable PNGSs found in gp120 of the individuals sampled. The top part of the table represents the raw % conservation, which is the number of NX[ST] sequons at any given position per number of sequences. This is different from % identity, which would be lower for these sites, since the central sequon position can be highly variable and the last position can be serine or threonine. Sites were ranked for each individual, from 1 to 12, where 1 is the most conserved and 12 is the least conserved. Ties were broken by calculating the genetic distance for each site where there was a tie. Smaller genetic distances received the lower rank, indicating a more conserved position.

Table 6

Patient	Mean	Std. Error	95% Wald Confidence Interval	
			Lower	Upper
2	94.1478	1.49292	91.2667	97.1198
5	97.6590	1.54860	94.6705	100.7419
6	88.8119	1.40831	86.0941	91.6155
7	83.0645	1.31715	80.5226	85.6866
8	90.6778	1.54183	87.7057	93.7506

Descriptive statistics for conservation of PNGSs per individual.

Table 7

Individuals (I, J)		Mean Difference (I-J)	Std. Error	df	Sequential Bonferroni Sig.	95% Wald Confidence Interval for Difference	
						Lower	Upper
2	5	-3.5112[a]	.05568	1	0.000	-3.6675	-3.3549
	6	5.3359[a]	.08461	1	0.000	5.1013	5.5705
	7	11.0833[a]	.17577	1	0.000	10.6027	11.5639
	8	3.4700[a]	.06457	1	0.000	3.2963	3.6437
5	2	3.5112[a]	.05568	1	0.000	3.3549	3.6675
	6	8.8471[a]	.14029	1	0.000	8.4770	9.2172
	7	14.5945[a]	.23145	1	0.000	13.9983	15.1907
	8	6.9812[a]	.04349	1	0.000	6.8726	7.0898
6	2	-5.3359[a]	.08461	1	0.000	-5.5705	-5.1013
	5	-8.8471[a]	.14029	1	0.000	-9.2172	-8.4770
	7	5.7474[a]	.09116	1	0.000	5.5292	5.9656
	8	-1.8659[a]	.13966	1	0.000	-2.1789	-1.5529
7	2	-11.0833[a]	.17577	1	0.000	-11.5639	-10.6027
	5	-14.5945[a]	.23145	1	0.000	-15.1907	-13.9983
	6	-5.7474[a]	.09116	1	0.000	-5.9656	-5.5292
	8	-7.6133[a]	.22814	1	0.000	-8.0604	-7.1661
8	2	-3.4700[a]	.06457	1	0.000	-3.6437	-3.2963
	5	-6.9812[a]	.04349	1	0.000	-7.0898	-6.8726
	6	1.8659[a]	.13966	1	0.000	1.5529	2.1789
	7	7.6133[a]	.22814	1	0.000	7.1661	8.0604

Conservation comparison for individuals. Pairwise comparisons of estimated marginal means for individuals based on the original scale of dependent variable conservation.
[a] The mean difference is significant at the 0.05 level.

Table 8

Site	Mean	Std. Error	95% Wald Confidence Interval	
			Lower	Upper
276	97.1521	3.02652	91.3977	103.2688
295	91.9554	2.82524	86.5814	97.6629
301	90.9035	.00000	90.9035	90.9035
332	92.7889	.35081	92.1039	93.4790
339	87.6483	.00718	87.6343	87.6624
356	97.1573	2.46384	92.4463	102.1083
362	85.2495	.00000	85.2495	85.2495
386	88.5089	2.26993	84.1699	93.0716
392	96.5763	2.70295	91.4214	102.0220
397	71.6228	.00000	71.6228	71.6228
401	93.1493	1.43041	90.3875	95.9955
448	98.6089	3.17706	92.5745	105.0366

Descriptive statistics for conservation of PNGSs per site.

Table 9

Site (I, J)		Mean Difference (I-J)	Std. Error	df	Sequential Bonferroni Sig.	95% Wald Confidence Interval for Difference	
						Lower	Upper
	295	5.1967[a]	.23236	1	0.000	4.4142	5.9793
	301	6.2486	3.02652	1	.662	-2.7517	15.2489
	332	4.3632	2.75496	1	1.000	-3.2590	11.9854
	339	9.5038	3.03344	1	.050	-.0032	19.0107
	356	-.0052	.99352	1	1.000	-1.9542	1.9439
276	362	11.9027[a]	3.02652	1	.003	2.2270	21.5783
	386	8.6432[a]	1.35162	1	.000	4.2445	13.0419
	392	.5758	.93980	1	1.000	-1.5025	2.6541
	397	25.5293[a]	3.02652	1	0.000	15.3492	35.7094
	401	4.0028	1.67696	1	.374	-1.1154	9.1210
	448	-1.4568[a]	.43091	1	.022	-2.8157	-.0979
	276	-5.1967[a]	.23236	1	0.000	-5.9793	-4.4142
	301	1.0519	2.82524	1	1.000	-4.8891	6.9928
	332	-.8335	2.54642	1	1.000	-6.1400	4.4729
	339	4.3070	2.83220	1	1.000	-3.4127	12.0268
	356	-5.2019[a]	.78423	1	.000	-7.7591	-2.6447
295	362	6.7059	2.82524	1	.374	-1.8861	15.2979
	386	3.4465	1.13651	1	.065	-.0915	6.9844
	392	-4.6210[a]	.76503	1	.000	-7.1057	-2.1363
	397	20.3326[a]	2.82524	1	.000	11.1025	29.5626
	401	-1.1939	1.45777	1	1.000	-4.5680	2.1801
	448	-6.6535[a]	.44853	1	0.000	-8.1603	-5.1467
	276	-6.2486	3.02652	1	.662	-15.2489	2.7517
	295	-1.0519	2.82524	1	1.000	-6.9928	4.8891
	332	-1.8854[a]	.35081	1	.000	-3.0200	-.7508
	339	3.2552[a]	.00718	1	0.000	3.2311	3.2793
	356	-6.2538	2.46384	1	.267	-13.8377	1.3301
301	362	5.6541[a]	.00000	1	0.000	5.6541	5.6541
	386	2.3946	2.26993	1	1.000	-3.1504	7.9396
	392	-5.6728	2.70295	1	.645	-13.7582	2.4125
	397	19.2807[a]	.00000	1	0.000	19.2807	19.2807
	401	-2.2458	1.43041	1	1.000	-6.1905	1.6989
	448	-7.7054	3.17706	1	.352	-17.4443	2.0335
332	276	-4.3632	2.75496	1	1.000	-11.9854	3.2590
	295	.8335	2.54642	1	1.000	-4.4729	6.1400

	301	1.8854[a]	.35081	1	.000	.7508	3.0200
	339	5.1406[a]	.35669	1	0.000	3.9487	6.3324
	356	-4.3684	2.13832	1	.662	-10.6928	1.9560
	362	7.5395[a]	.35081	1	0.000	6.3689	8.7100
	386	4.2800	1.93287	1	.536	-1.5637	10.1237
	392	-3.7874	2.37811	1	1.000	-10.3808	2.8059
	397	21.1661[a]	.35081	1	0.000	19.9972	22.3350
	401	-.3604	1.12248	1	1.000	-2.6966	1.9758
	448	-5.8200	2.87878	1	.662	-14.2891	2.6491
	276	-9.5038	3.03344	1	.050	-19.0107	.0032
	295	-4.3070	2.83220	1	1.000	-12.0268	3.4127
	301	-3.2552[a]	.00718	1	0.000	-3.2793	-3.2311
	332	-5.1406[a]	.35669	1	0.000	-6.3324	-3.9487
	356	-9.5089[a]	2.47070	1	.004	-17.3876	-1.6303
339	362	2.3989[a]	.00718	1	0.000	2.3750	2.4228
	386	-.8606	2.27664	1	1.000	-5.6534	3.9323
	392	-8.9280[a]	2.70984	1	.030	-17.4477	-.4083
	397	16.0255[a]	.00718	1	0.000	16.0017	16.0494
	401	-5.5010[a]	1.43743	1	.004	-10.0726	-.9293
	448	-10.9605[a]	3.18406	1	.018	-21.0312	-.8899
	276	.0052	.99352	1	1.000	-1.9439	1.9542
	295	5.2019[a]	.78423	1	.000	2.6447	7.7591
	301	6.2538	2.46384	1	.267	-1.3301	13.8377
	332	4.3684	2.13832	1	.662	-1.9560	10.6928
	339	9.5089[a]	2.47070	1	.004	1.6303	17.3876
356	362	11.9078[a]	2.46384	1	.000	3.9743	19.8413
	386	8.6484[a]	.37578	1	0.000	7.4018	9.8949
	392	.5809	.26820	1	.576	-.2258	1.3876
	397	25.5345[a]	2.46384	1	0.000	17.3739	33.6950
	401	4.0080[a]	1.04834	1	.004	.6785	7.3375
	448	-1.4516	.86096	1	1.000	-3.9602	1.0570
	276	-11.9027[a]	3.02652	1	.003	-21.5783	-2.2270
	295	-6.7059	2.82524	1	.374	-15.2979	1.8861
	301	-5.6541[a]	.00000	1	0.000	-5.6541	-5.6541
	332	-7.5395[a]	.35081	1	0.000	-8.7100	-6.3689
362	339	-2.3989[a]	.00718	1	0.000	-2.4228	-2.3750
	356	-11.9078[a]	2.46384	1	.000	-19.8413	-3.9743
	386	-3.2595	2.26993	1	1.000	-9.3236	2.8047
	392	-11.3269[a]	2.70295	1	.001	-19.9894	-2.6644
	397	13.6267[a]	.00000	1	0.000	13.6266	13.6267

	401	-7.8999[a]	1.43041	1	.000	-12.5360	-3.2637
	448	-13.3594[a]	3.17706	1	.001	-23.5658	-3.1531
	276	-8.6432[a]	1.35162	1	.000	-13.0419	-4.2445
	295	-3.4465	1.13651	1	.065	-6.9844	.0915
	301	-2.3946	2.26993	1	1.000	-7.9396	3.1504
	332	-4.2800	1.93287	1	.536	-10.1237	1.5637
	339	.8606	2.27664	1	1.000	-3.9323	5.6534
386	356	-8.6484[a]	.37578	1	0.000	-9.8949	-7.4018
	362	3.2595	2.26993	1	1.000	-2.8047	9.3236
	392	-8.0674[a]	.52824	1	0.000	-9.8114	-6.3234
	397	16.8861[a]	2.26993	1	.000	9.4565	24.3158
	401	-4.6404[a]	.90754	1	.000	-7.5692	-1.7116
	448	-10.1000[a]	1.21345	1	.000	-14.0789	-6.1210
	276	-.5758	.93980	1	1.000	-2.6541	1.5025
	295	4.6210[a]	.76503	1	.000	2.1363	7.1057
	301	5.6728	2.70295	1	.645	-2.4125	13.7582
	332	3.7874	2.37811	1	1.000	-2.8059	10.3808
	339	8.9280[a]	2.70984	1	.030	.4083	17.4477
392	356	-.5809	.26820	1	.576	-1.3876	.2258
	362	11.3269[a]	2.70295	1	.001	2.6644	19.9894
	386	8.0674[a]	.52824	1	0.000	6.3234	9.8114
	397	24.9535[a]	2.70295	1	0.000	16.0444	33.8627
	401	3.4270	1.28203	1	.188	-.5347	7.3888
	448	-2.0325	.71275	1	.113	-4.2434	.1783
	276	-25.5293[a]	3.02652	1	0.000	-35.7094	-15.3492
	295	-20.3326[a]	2.82524	1	.000	-29.5626	-11.1025
	301	-19.2807[a]	.00000	1	0.000	-19.2807	-19.2807
	332	-21.1661[a]	.35081	1	0.000	-22.3350	-19.9972
	339	-16.0255[a]	.00718	1	0.000	-16.0494	-16.0017
397	356	-25.5345[a]	2.46384	1	0.000	-33.6950	-17.3739
	362	-13.6267[a]	.00000	1	0.000	-13.6267	-13.6266
	386	-16.8861[a]	2.26993	1	.000	-24.3158	-9.4565
	392	-24.9535[a]	2.70295	1	0.000	-33.8627	-16.0444
	401	-21.5265[a]	1.43041	1	0.000	-26.2333	-16.8197
	448	-26.9861[a]	3.17706	1	0.000	-37.4222	-16.5499
	276	-4.0028	1.67696	1	.374	-9.1210	1.1154
	295	1.1939	1.45777	1	1.000	-2.1801	4.5680
401	301	2.2458	1.43041	1	1.000	-1.6989	6.1905
	332	.3604	1.12248	1	1.000	-1.9758	2.6966
	339	5.5010[a]	1.43743	1	.004	.9293	10.0726

	356	-4.0080[a]	1.04834	1	.004	-7.3375	-.6785
	362	7.8999[a]	1.43041	1	.000	3.2637	12.5360
	386	4.6404[a]	.90754	1	.000	1.7116	7.5692
	392	-3.4270	1.28203	1	.188	-7.3888	.5347
	397	21.5265[a]	1.43041	1	0.000	16.8197	26.2333
	448	-5.4596	1.76592	1	.056	-10.9758	.0567
	276	1.4568[a]	.43091	1	.022	.0979	2.8157
	295	6.6535[a]	.44853	1	0.000	5.1467	8.1603
	301	7.7054	3.17706	1	.352	-2.0335	17.4443
	332	5.8200	2.87878	1	.662	-2.6491	14.2891
	339	10.9605[a]	3.18406	1	.018	.8899	21.0312
448	356	1.4516	.86096	1	1.000	-1.0570	3.9602
	362	13.3594[a]	3.17706	1	.001	3.1531	23.5658
	386	10.1000[a]	1.21345	1	.000	6.1210	14.0789
	392	2.0325	.71275	1	.113	-.1783	4.2434
	397	26.9861[a]	3.17706	1	0.000	16.5499	37.4222
	401	5.4596	1.76592	1	.056	-.0567	10.9758

Conservation comparison for sites. Pairwise comparisons of estimated marginal means for site based on the original scale of dependent variable PNGS conservation.
[a] The mean difference is significant at the 0.05 level.

Table 10

Compartment	Mean	Std. Error	95% Wald Confidence Interval	
			Lower	Upper
PBMC	89.4518	1.59227	86.3848	92.6277
Plasma	92.0416	1.77381	88.6298	95.5847

Descriptive statistics for conservation of PNGSs per compartment

Table 11

Compartment (I, J)		Mean Difference (I-J)	Std. Error	df	Sequential Bonferroni Sig.	95% Wald Confidence Interval for Difference[a]	
						Lower	Upper
PBMC	Plasma	-2.5898	1.68631	1	.125	-5.8949	.7153
Plasma	PBMC	2.5898	1.68631	1	.125	-.7153	5.8949

Conservation comparison for compartments. Pairwise comparisons of estimated marginal means for compartment based on the original scale of dependent variable Conservation
[a]Confidence interval bounds are approximate.

Table 12

Atom 1	Atom 2	Binding Type	SeqD	EucD	SASD
ASN276	LYS121	17B	82	34.8	-3
ASN276	ILE423	17B	111	34.3	44.4
ASN276	LYS421	17B	109	33.4	47.4
ASN276	ARG419	17B	107	34.6	49.1
ASN276	GLN422	17B	110	37.8	49.5
ASN276	TYR435	17B	123	37	53.3
ASN276	ILE420	17B	108	36.9	54
ASN276	ALA281	B12	5	8.4	14.3
ASN276	ASN280	B12	4	9	19.4
ASN276	ASP474	B12	159	15.1	21.2
ASN276	ARG456	B12	144	10.5	23.6
ASN276	THR455	B12	143	12	23.9
ASN276	ASP457	B12	145	14.9	24.6
ASN276	MET475	B12	160	20.2	27.1
ASN276	SER365	B12	67	19.7	27.3
ASN276	ILE371	B12	73	19.9	29.1
ASN276	VAL430	B12	118	24.8	30
ASN276	SER364	B12	66	21.2	30.9
ASN276	ASP368	B12	70	24.2	31
ASN276	PRO470	B12	155	19.1	32.6
ASN276	GLU370	B12	72	24	32.7
ASN276	VAL372	B12	74	24.4	33.2
ASN276	THR257	B12	19	21.2	33.5
ASN276	LEU453	B12	141	15.6	35.9
ASN276	PRO369	B12	71	27.7	36.8
ASN276	THR373	B12	75	26.8	37.8
ASN276	SER256	B12	20	23.8	39.2
ASN276	TYR384	B12	86	30.1	40.7
ASN276	ASN386	B12	88	30.1	41.4
ASN276	LYS432	B12	120	31.6	42.8
ASN276	ARG419	B12	107	34.6	45.5
ASN276	CYS418	B12	106	33.7	46
ASN276	PRO417	B12	105	34.7	47.1
ASN276	THR123	CCR5	80	31.4	36.8
ASN276	LYS121	CCR5	82	34.8	43
ASN276	ARG419	CCR5	107	34.6	49.1
ASN276	LYS421	CCR5	109	33.4	50.1
ASN276	GLN422	CCR5	110	37.8	51
ASN276	ILE420	CCR5	108	36.9	54

ASN276	LYS117	CCR5	86	38.8	56.2
ASN276	PRO437	CCR5	125	43.4	57.4
ASN276	PRO438	CCR5	126	40.7	57.8
ASN276	HIS330	CCR5	32	37.1	60.1
ASN276	LYS207	CCR5	69	42	67.5
ASN276	ARG440	CCR5	128	47.7	68.5
ASN276	ARG444	CCR5	132	39.6	72.7
ASN276	ASN279	CD4	3	3.6	4
ASN276	LYS282	CD4	6	4.7	5.9
ASN276	ALA281	CD4	5	8.4	9.4
ASN276	ASN280	CD4	4	9	10.2
ASN276	THR283	CD4	7	10.2	12
ASN276	ARG456	CD4	144	10.5	13.7
ASN276	THR455	CD4	143	12	14.9
ASN276	ASP477	CD4	162	14	15.4
ASN276	ASP474	CD4	159	15.1	15.9
ASN276	ASP457	CD4	145	14.9	16.8
ASN276	ARG476	CD4	161	16	17.8
ASN276	ARG469	CD4	154	17.3	20.1
ASN276	MET475	CD4	160	20.2	21.3
ASN276	ILE371	CD4	73	19.9	22
ASN276	SER365	CD4	67	19.7	22.6
ASN276	TRP427	CD4	115	23.3	25.3
ASN276	GLU370	CD4	72	24	26.2
ASN276	ASP368	CD4	70	24.2	26.4
ASN276	GLN428	CD4	116	26.3	28
ASN276	VAL430	CD4	118	24.8	28.7
ASN276	SER375	CD4	77	25.9	28.9
ASN276	GLU429	CD4	117	26.1	29
ASN276	ASN425	CD4	113	26.6	29.6
ASN276	MET426	CD4	114	27.7	31
ASN276	LYS432	CD4	120	31.6	36
ASN276	THR123	CG10	80	31.4	36.8
ASN276	LYS432	CG10	120	31.6	42.8
ASN276	LYS121	CG10	82	34.8	43
ASN276	ILE423	CG10	111	34.3	44.4
ASN276	LYS421	CG10	109	33.4	47.4
ASN276	GLN422	CG10	110	37.8	49.5
ASN276	MET434	CG10	122	37.4	49.6
ASN276	TYR435	CG10	123	37	52.8
ASN276	LYS207	CG10	69	42	62.2

ASN295	LYS121	17B	101	31.4	-3
ASN295	ARG419	17B	88	16.9	34.2
ASN295	ILE420	17B	89	16	39.5
ASN295	LYS421	17B	90	19.9	40.4
ASN295	GLN422	17B	91	23.4	44.8
ASN295	ILE423	17B	92	25.7	45.1
ASN295	TYR435	17B	104	23.2	49.5
ASN295	PRO417	B12	86	14.5	24.8
ASN295	CYS418	B12	87	11.5	28
ASN295	ASN386	B12	69	16.2	29.5
ASN295	ARG419	B12	88	16.9	31.4
ASN295	TYR384	B12	67	16.6	33.9
ASN295	THR373	B12	56	17.3	34.2
ASN295	SER256	B12	39	15	35.7
ASN295	VAL372	B12	55	22.1	38.3
ASN295	PRO369	B12	52	21.8	39.4
ASN295	SER364	B12	47	22.6	40.6
ASN295	THR257	B12	38	18.5	41.3
ASN295	GLU370	B12	53	21.6	42.1
ASN295	ILE371	B12	54	22.9	43.4
ASN295	PRO470	B12	136	19.9	44
ASN295	ASP368	B12	51	25.6	45.2
ASN295	SER365	B12	48	28.1	45.5
ASN295	LEU453	B12	122	20.9	45.8
ASN295	THR455	B12	124	27	51.9
ASN295	LYS432	B12	101	29.4	54.4
ASN295	ASP457	B12	126	31.6	54.9
ASN295	VAL430	B12	99	32.8	55.6
ASN295	ASP474	B12	140	27.2	56.1
ASN295	ARG456	B12	125	30.2	57.1
ASN295	ALA281	B12	14	34.1	57.2
ASN295	ASN280	B12	15	33.7	57.5
ASN295	MET475	B12	141	23.7	58.9
ASN295	LYS121	CCR5	101	31.4	-3
ASN295	ARG444	CCR5	113	7.2	19
ASN295	HIS330	CCR5	13	9.4	20.1
ASN295	ARG440	CCR5	109	20	30.8
ASN295	LYS207	CCR5	88	21.7	33.7
ASN295	ARG419	CCR5	88	16.9	34.2
ASN295	ILE420	CCR5	89	16	39.5
ASN295	LYS421	CCR5	90	19.9	40.4

ASN295	PRO437	CCR5	106	23.5	41.3
ASN295	PRO438	CCR5	107	18.1	43.3
ASN295	GLN422	CCR5	91	23.4	44.8
ASN295	LYS117	CCR5	105	28.2	44.8
ASN295	THR123	CCR5	99	35.1	68.1
ASN295	SER365	CD4	48	28.1	47.6
ASN295	ASN425	CD4	94	25.6	49.9
ASN295	ARG469	CD4	135	24.8	50.3
ASN295	ASP368	CD4	51	25.6	50.7
ASN295	LYS432	CD4	101	29.4	53.7
ASN295	GLU370	CD4	53	21.6	54.2
ASN295	ASP457	CD4	126	31.6	55.3
ASN295	ILE371	CD4	54	22.9	55.7
ASN295	THR455	CD4	124	27	56.4
ASN295	MET426	CD4	95	27.2	56.7
ASN295	SER375	CD4	58	17.1	57
ASN295	TRP427	CD4	96	24.2	58.8
ASN295	ARG456	CD4	125	30.2	59
ASN295	VAL430	CD4	99	32.8	59
ASN295	ALA281	CD4	14	34.1	59.6
ASN295	ASN280	CD4	15	33.7	60.1
ASN295	THR283	CD4	12	26.7	60.7
ASN295	GLN428	CD4	97	28.2	61.8
ASN295	GLU429	CD4	98	31.9	62.1
ASN295	MET475	CD4	141	23.7	63.4
ASN295	LYS282	CD4	13	32.2	64.4
ASN295	ASP474	CD4	140	27.2	64.9
ASN295	ASP477	CD4	143	24.7	65.9
ASN295	ASN279	CD4	16	36	66
ASN295	ARG476	CD4	142	28.1	68.9
ASN295	LYS121	CG10	101	31.4	-3
ASN295	LYS207	CG10	88	21.7	33.7
ASN295	TYR435	CG10	104	23.2	42.1
ASN295	LYS421	CG10	90	19.9	42.5
ASN295	GLN422	CG10	91	23.4	44.8
ASN295	ILE423	CG10	92	25.7	45.1
ASN295	MET434	CG10	103	27.9	48.3
ASN295	LYS432	CG10	101	29.4	51
ASN295	THR123	CG10	99	35.1	68.1
ASN301	LYS121	17B	107	30.8	-3
ASN301	ARG419	17B	82	15.3	24.8

ASN301	GLN422	17B	85	18.8	27.4
ASN301	LYS421	17B	84	18.5	29
ASN301	ILE423	17B	86	23	30.9
ASN301	ILE420	17B	83	13.1	31.1
ASN301	TYR435	17B	98	20.9	33.8
ASN301	PRO417	B12	80	15.9	18.5
ASN301	CYS418	B12	81	14.3	22
ASN301	ARG419	B12	82	15.3	23.4
ASN301	ASN386	B12	63	20.1	24.3
ASN301	TYR384	B12	61	19	27.5
ASN301	THR373	B12	50	22.2	28.2
ASN301	VAL372	B12	49	26.6	31.4
ASN301	PRO369	B12	46	23.9	31.7
ASN301	GLU370	B12	47	26.4	35.1
ASN301	SER364	B12	41	29.4	35.6
ASN301	THR257	B12	44	27.1	36.6
ASN301	ILE371	B12	48	29.6	36.8
ASN301	PRO470	B12	130	29.8	37.7
ASN301	ASP368	B12	45	28.9	37.8
ASN301	LYS432	B12	95	28.8	38.1
ASN301	SER365	B12	42	34.4	40.5
ASN301	SER256	B12	45	25.6	41.3
ASN301	LEU453	B12	116	32.6	43.9
ASN301	THR455	B12	118	37.2	44.8
ASN301	VAL430	B12	93	36.1	48.3
ASN301	ARG456	B12	119	41.3	48.5
ASN301	ASP457	B12	120	40.7	48.7
ASN301	ASP474	B12	134	36.2	49.1
ASN301	ASN280	B12	21	43.9	51.5
ASN301	ALA281	B12	20	43.4	52.2
ASN301	MET475	B12	135	32.5	52.8
ASN301	ARG440	CCR5	103	10.4	11.4
ASN301	HIS330	CCR5	7	12.2	14.8
ASN301	PRO437	CCR5	100	15.1	16.9
ASN301	PRO438	CCR5	101	11.9	18.7
ASN301	LYS207	CCR5	94	18.7	22.6
ASN301	ARG444	CCR5	107	13.1	23.3
ASN301	GLN422	CCR5	85	18.8	23.4
ASN301	ILE420	CCR5	83	13.1	23.5
ASN301	ARG419	CCR5	82	15.3	24.1
ASN301	LYS421	CCR5	84	18.5	27.4

ASN301	LYS117	CCR5	111	27.5	37.1
ASN301	THR123	CCR5	105	36	54.2
ASN301	LYS121	CCR5	107	30.8	61.7
ASN301	ASN425	CD4	88	27.9	38.1
ASN301	LYS432	CD4	95	28.8	38.1
ASN301	ASP368	CD4	45	28.9	40.1
ASN301	GLU370	CD4	47	26.4	41.9
ASN301	MET426	CD4	89	30	43.4
ASN301	SER365	CD4	42	34.4	43.5
ASN301	SER375	CD4	52	23	44.3
ASN301	TRP427	CD4	90	30.3	45.9
ASN301	ILE371	CD4	48	29.6	46.8
ASN301	VAL430	CD4	93	36.1	47.4
ASN301	ARG469	CD4	129	33.9	47.9
ASN301	GLN428	CD4	91	33.1	48.5
ASN301	GLU429	CD4	92	35.8	48.5
ASN301	ASP457	CD4	120	40.7	50.8
ASN301	MET475	CD4	135	32.5	51.1
ASN301	ASP474	CD4	134	36.2	52.6
ASN301	THR455	CD4	118	37.2	52.8
ASN301	ARG456	CD4	119	41.3	55.1
ASN301	ASN280	CD4	21	43.9	55.8
ASN301	ALA281	CD4	20	43.4	56.8
ASN301	ASP477	CD4	137	36	56.9
ASN301	ARG476	CD4	136	38.2	56.9
ASN301	THR283	CD4	18	37.9	57.2
ASN301	LYS282	CD4	19	43.4	61.2
ASN301	ASN279	CD4	22	47	61.5
ASN301	LYS207	CG10	94	18.7	24.5
ASN301	GLN422	CG10	85	18.8	27.4
ASN301	LYS421	CG10	84	18.5	29
ASN301	ILE423	CG10	86	23	30.9
ASN301	MET434	CG10	97	24.3	31.9
ASN301	TYR435	CG10	98	20.9	33
ASN301	LYS432	CG10	95	28.8	36.8
ASN301	LYS121	CG10	107	30.8	38.5
ASN301	THR123	CG10	105	36	46.2
ASN332	LYS121	17B	116	33.3	-3
ASN332	ARG419	17B	73	15.9	29.7
ASN332	ILE420	17B	74	16.7	35.7
ASN332	LYS421	17B	75	20.2	35.9

ASN332	ILE423	17B	77	26.1	41.5
ASN332	GLN422	17B	76	24.2	41.6
ASN332	TYR435	17B	89	25.1	44.7
ASN332	PRO417	B12	71	11.4	22.9
ASN332	CYS418	B12	72	10.6	26.2
ASN332	ASN386	B12	54	13.8	28.3
ASN332	ARG419	B12	73	15.9	28.4
ASN332	TYR384	B12	52	16.3	32
ASN332	THR373	B12	41	16.2	32.8
ASN332	PRO369	B12	37	21.1	36.3
ASN332	SER256	B12	54	17	37
ASN332	VAL372	B12	40	20.8	37.2
ASN332	SER364	B12	32	21	39.7
ASN332	GLU370	B12	38	22	40.7
ASN332	ILE371	B12	39	22.7	41.3
ASN332	ASP368	B12	36	25.4	42
ASN332	THR257	B12	53	19.2	42.5
ASN332	PRO470	B12	121	18.8	42.9
ASN332	SER365	B12	33	26.3	43.8
ASN332	LEU453	B12	107	21.5	48.4
ASN332	THR455	B12	109	26.4	49.9
ASN332	LYS432	B12	86	30.3	51.8
ASN332	ASP457	B12	111	29.8	52.4
ASN332	VAL430	B12	84	33.9	53.8
ASN332	ASP474	B12	125	28.4	54.1
ASN332	ARG456	B12	110	29.2	56.3
ASN332	ALA281	B12	29	33.8	56.3
ASN332	ASN280	B12	30	32.8	57
ASN332	MET475	B12	126	25.8	57.2
ASN332	HIS330	CCR5	2	7	7.3
ASN332	ARG419	CCR5	73	15.9	20.8
ASN332	ARG444	CCR5	98	10.8	21.5
ASN332	ILE420	CCR5	74	16.7	26.8
ASN332	LYS421	CCR5	75	20.2	27
ASN332	ARG440	CCR5	94	22	29.8
ASN332	PRO438	CCR5	92	19.3	30.8
ASN332	GLN422	CCR5	76	24.2	32.7
ASN332	PRO437	CCR5	91	24.6	34.7
ASN332	LYS207	CCR5	103	24.7	34.8
ASN332	LYS117	CCR5	120	31.2	44.3
ASN332	THR123	CCR5	114	36.7	54.9

ASN332	LYS121	CCR5	116	33.3	62.7
ASN332	SER365	CD4	33	26.3	45.2
ASN332	ASP368	CD4	36	25.4	47.8
ASN332	ARG469	CD4	120	23.2	49.6
ASN332	ASN425	CD4	79	26.2	51.1
ASN332	ASP457	CD4	111	29.8	52.4
ASN332	LYS432	CD4	86	30.3	52.5
ASN332	GLU370	CD4	38	22	52.8
ASN332	ILE371	CD4	39	22.7	54.4
ASN332	THR455	CD4	109	26.4	54.9
ASN332	SER375	CD4	43	17.7	55.7
ASN332	ARG456	CD4	110	29.2	56.9
ASN332	MET426	CD4	80	28.8	57.4
ASN332	ASN280	CD4	30	32.8	57.8
ASN332	VAL430	CD4	84	33.9	58
ASN332	TRP427	CD4	81	26	58.1
ASN332	ALA281	CD4	29	33.8	58.5
ASN332	THR283	CD4	27	27.2	59.6
ASN332	GLU429	CD4	83	33.7	61.6
ASN332	GLN428	CD4	82	30.5	62.1
ASN332	ASP474	CD4	125	28.4	62.8
ASN332	MET475	CD4	126	25.8	63
ASN332	ASN279	CD4	31	35.9	63.2
ASN332	LYS282	CD4	28	32.4	63.3
ASN332	ASP477	CD4	128	26.2	64.8
ASN332	ARG476	CD4	127	30.1	67.4
ASN332	LYS207	CG10	103	24.7	34.8
ASN332	LYS421	CG10	75	20.2	40.8
ASN332	TYR435	CG10	89	25.1	41.5
ASN332	GLN422	CG10	76	24.2	42.8
ASN332	ILE423	CG10	77	26.1	43
ASN332	MET434	CG10	88	29.2	47.1
ASN332	LYS432	CG10	86	30.3	48.8
ASN332	THR123	CG10	114	36.7	64.4
ASN332	LYS121	CG10	116	33.3	72.2
ASN356	LYS121	17B	140	50.1	-3
ASN356	ARG419	17B	49	40.6	50.1
ASN356	LYS421	17B	51	42.5	54.7
ASN356	ILE423	17B	53	45.4	55.9
ASN356	ILE420	17B	50	44.6	56.6
ASN356	GLN422	17B	52	48	60.3

ASN356	TYR435	17B	65	48.9	63.5
ASN356	ASN280	B12	54	17.8	22.4
ASN356	ASP457	B12	87	17.1	24.1
ASN356	ARG456	B12	86	16	25.9
ASN356	ALA281	B12	53	23.2	27.3
ASN356	THR455	B12	85	21.2	29.9
ASN356	SER365	B12	9	25	32.3
ASN356	SER364	B12	8	26.8	37.1
ASN356	ILE371	B12	15	30.3	38.8
ASN356	VAL372	B12	16	31.4	40.1
ASN356	PRO470	B12	97	24.7	40.4
ASN356	ASP474	B12	101	31.6	40.7
ASN356	THR257	B12	77	31.5	43.6
ASN356	ASP368	B12	12	35.3	44
ASN356	ASN386	B12	30	33.2	44.2
ASN356	GLU370	B12	14	35.2	44.3
ASN356	PRO417	B12	47	36.7	44.5
ASN356	THR373	B12	17	33.2	44.8
ASN356	MET475	B12	102	35.6	44.8
ASN356	LEU453	B12	83	26.1	46.2
ASN356	PRO369	B12	13	36.2	46.2
ASN356	VAL430	B12	60	41.1	47.6
ASN356	TYR384	B12	28	37.6	48.9
ASN356	SER256	B12	78	33.8	49.3
ASN356	ARG419	B12	49	40.6	50.1
ASN356	CYS418	B12	48	38.6	50.8
ASN356	LYS432	B12	62	45	55.3
ASN356	ARG419	CCR5	49	40.6	50.1
ASN356	HIS330	CCR5	26	39.5	51.3
ASN356	LYS421	CCR5	51	42.5	54.7
ASN356	ILE420	CCR5	50	44.6	56.6
ASN356	THR123	CCR5	138	47.9	57.5
ASN356	GLN422	CCR5	52	48	60.8
ASN356	PRO438	CCR5	68	49	60.8
ASN356	PRO437	CCR5	67	52.8	65
ASN356	LYS121	CCR5	140	50.1	65.4
ASN356	ARG444	CCR5	74	45.1	66.1
ASN356	ARG440	CCR5	70	55.5	71.4
ASN356	LYS207	CCR5	127	53.2	75.4
ASN356	LYS117	CCR5	144	53.3	79.9
ASN356	ASN280	CD4	54	17.8	22.4

ASN356	ASP457	CD4	87	17.1	24.1
ASN356	ARG456	CD4	86	16	25.9
ASN356	ASN279	CD4	55	20.4	26.4
ASN356	ALA281	CD4	53	23.2	27.3
ASN356	LYS282	CD4	52	22.1	28.9
ASN356	THR455	CD4	85	21.2	29.9
ASN356	ARG469	CD4	96	21.2	30.7
ASN356	SER365	CD4	9	25	32.3
ASN356	THR283	CD4	51	24.3	32.9
ASN356	ASP477	CD4	104	29.1	37.9
ASN356	ILE371	CD4	15	30.3	38.8
ASN356	ASP474	CD4	101	31.6	40.4
ASN356	ARG476	CD4	103	33.7	42.4
ASN356	ASP368	CD4	12	35.3	44
ASN356	GLU370	CD4	14	35.2	44.3
ASN356	MET475	CD4	102	35.6	44.7
ASN356	SER375	CD4	19	35.4	47.1
ASN356	VAL430	CD4	60	41.1	47.6
ASN356	ASN425	CD4	55	39.3	47.8
ASN356	TRP427	CD4	57	38.1	48.4
ASN356	MET426	CD4	56	42.6	52.1
ASN356	GLU429	CD4	59	43.3	52.1
ASN356	GLN428	CD4	58	42.8	52.4
ASN356	LYS432	CD4	62	45	54.1
ASN356	ILE423	CG10	53	45.4	55.9
ASN356	LYS432	CG10	62	45	56.3
ASN356	LYS421	CG10	51	42.5	57.4
ASN356	THR123	CG10	138	47.9	57.5
ASN356	GLN422	CG10	52	48	60.3
ASN356	MET434	CG10	64	49.9	62.4
ASN356	LYS121	CG10	140	50.1	62.5
ASN356	TYR435	CG10	65	48.9	63.9
ASN356	LYS207	CG10	127	53.2	73.5
ASN362	LYS121	17B	146	34	-3
ASN362	ARG419	17B	43	21.5	27
ASN362	LYS421	17B	45	23.8	30.9
ASN362	ILE420	17B	44	26.3	32.5
ASN362	ILE423	17B	47	26.9	33.3
ASN362	GLN422	17B	46	29.4	36.6
ASN362	TYR435	17B	59	31.4	39.7
ASN362	SER365	B12	3	8.2	13.7

ASN362	ASP457	B12	81	8.3	17
ASN362	SER364	B12	2	8.4	17.4
ASN362	VAL372	B12	10	12.6	18.7
ASN362	ASN386	B12	24	14.1	19.7
ASN362	THR373	B12	11	14.6	21.2
ASN362	PRO470	B12	91	8.7	21.5
ASN362	THR455	B12	79	10.5	22.5
ASN362	ILE371	B12	9	14	22.5
ASN362	PRO369	B12	7	17.4	22.6
ASN362	ARG456	B12	80	11	23.4
ASN362	ASN280	B12	60	13.6	23.7
ASN362	ASP368	B12	6	18	24.7
ASN362	PRO417	B12	41	18.2	24.7
ASN362	TYR384	B12	22	19	24.9
ASN362	ALA281	B12	59	16.7	25.2
ASN362	GLU370	B12	8	18.3	25.2
ASN362	THR257	B12	83	16.1	26.9
ASN362	ARG419	B12	43	21.5	27
ASN362	CYS418	B12	42	20.7	27.6
ASN362	LEU453	B12	77	14.8	28.8
ASN362	ASP474	B12	95	20.7	31.5
ASN362	SER256	B12	84	19.9	32.5
ASN362	VAL430	B12	54	26.9	34.4
ASN362	MET475	B12	96	23.5	36.3
ASN362	LYS432	B12	56	27.9	40.1
ASN362	ARG419	CCR5	43	21.5	27
ASN362	LYS421	CCR5	45	23.8	30.9
ASN362	HIS330	CCR5	32	22.5	31.1
ASN362	ILE420	CCR5	44	26.3	32.5
ASN362	GLN422	CCR5	46	29.4	36.6
ASN362	PRO438	CCR5	62	30.8	37.6
ASN362	PRO437	CCR5	61	34.1	41.9
ASN362	THR123	CCR5	144	33	45.4
ASN362	LYS121	CCR5	146	34	51.7
ASN362	LYS207	CCR5	133	36.4	51.8
ASN362	ARG440	CCR5	64	37.9	51.8
ASN362	ARG444	CCR5	68	30	54.1
ASN362	LYS117	CCR5	150	37.5	62.6
ASN362	ARG469	CD4	90	5.2	5.2
ASN362	SER365	CD4	3	8.2	8.3
ASN362	ASP457	CD4	81	8.3	9.7

ASN362	THR455	CD4	79	10.5	11
ASN362	ARG456	CD4	80	11	12.8
ASN362	ASN280	CD4	60	13.6	14.9
ASN362	ILE371	CD4	9	14	16.1
ASN362	THR283	CD4	57	15.8	16.7
ASN362	ALA281	CD4	59	16.7	17
ASN362	LYS282	CD4	58	17.9	19.4
ASN362	ASN279	CD4	61	19	19.7
ASN362	ASP368	CD4	6	18	21.2
ASN362	GLU370	CD4	8	18.3	21.6
ASN362	ASP477	CD4	98	19.8	21.6
ASN362	ASP474	CD4	95	20.7	22.1
ASN362	SER375	CD4	13	18.3	23.8
ASN362	ASN425	CD4	49	22.4	25.5
ASN362	ARG476	CD4	97	24.7	26.2
ASN362	MET475	CD4	96	23.5	27
ASN362	TRP427	CD4	51	23.8	27.3
ASN362	VAL430	CD4	54	26.9	29.7
ASN362	MET426	CD4	50	27.2	30.9
ASN362	LYS432	CD4	56	27.9	31.1
ASN362	GLN428	CD4	52	29.2	32.3
ASN362	GLU429	CD4	53	29.7	32.4
ASN362	ILE423	CG10	47	26.9	33.3
ASN362	LYS421	CG10	45	23.8	34.1
ASN362	GLN422	CG10	46	29.4	36.7
ASN362	LYS432	CG10	56	27.9	36.9
ASN362	MET434	CG10	58	32	38.6
ASN362	TYR435	CG10	59	31.4	41.3
ASN362	LYS121	CG10	146	34	44.9
ASN362	THR123	CG10	144	33	45.4
ASN362	LYS207	CG10	133	36.4	51.1
ASN386	LYS121	17B	170	26.2	-3
ASN386	ARG419	17B	19	7.9	8.9
ASN386	LYS421	17B	21	12.1	13.4
ASN386	ILE420	17B	20	13	15.1
ASN386	ILE423	17B	23	17	18.9
ASN386	GLN422	17B	22	17.6	19.5
ASN386	TYR435	17B	35	20.3	22.1
ASN386	ASN386	B12	0	0	0
ASN386	THR373	B12	13	5.1	5.1
ASN386	PRO417	B12	17	4.9	5.8

ASN386	CYS418	B12	18	7.4	7.4
ASN386	TYR384	B12	2	7.4	8
ASN386	VAL372	B12	14	8.3	8.3
ASN386	ARG419	B12	19	7.9	8.9
ASN386	PRO369	B12	17	9.3	9.3
ASN386	SER364	B12	22	9.7	11
ASN386	GLU370	B12	16	12.5	12.5
ASN386	ILE371	B12	15	12.6	12.8
ASN386	PRO470	B12	67	11.1	14.1
ASN386	THR257	B12	107	12.3	14.3
ASN386	ASP368	B12	18	14	14.8
ASN386	SER365	B12	21	14.6	16.4
ASN386	SER256	B12	108	14.5	19.5
ASN386	LEU453	B12	53	16.6	20.2
ASN386	THR455	B12	55	18.3	20.9
ASN386	VAL430	B12	30	24.2	25.1
ASN386	ASP457	B12	57	20.7	25.2
ASN386	ASP474	B12	71	21.5	25.8
ASN386	ASN280	B12	84	24.5	26.9
ASN386	ARG456	B12	56	22.1	27.1
ASN386	ALA281	B12	83	24.8	27.6
ASN386	LYS432	B12	32	20.6	28.1
ASN386	MET475	B12	72	20.9	29.8
ASN386	ARG419	CCR5	19	7.9	8.9
ASN386	LYS421	CCR5	21	12.1	13.4
ASN386	ILE420	CCR5	20	13	15.1
ASN386	HIS330	CCR5	56	9.9	17.5
ASN386	PRO438	CCR5	38	17.4	19.5
ASN386	GLN422	CCR5	22	17.6	19.5
ASN386	PRO437	CCR5	37	21	23.4
ASN386	ARG440	CCR5	40	24.2	32.8
ASN386	LYS207	CCR5	157	24.4	33.7
ASN386	THR123	CCR5	168	27.9	38.1
ASN386	ARG444	CCR5	44	18.4	40.6
ASN386	LYS121	CCR5	170	26.2	46.2
ASN386	LYS117	CCR5	174	27.8	49.8
ASN386	SER365	CD4	21	14.6	20.1
ASN386	ASP368	CD4	18	14	20.4
ASN386	ARG469	CD4	66	14.2	22.2
ASN386	ASN425	CD4	25	16.3	22.8
ASN386	GLU370	CD4	16	12.5	24.9

ASN386	ILE371	CD4	15	12.6	25.4
ASN386	LYS432	CD4	32	20.6	26.1
ASN386	ASP457	CD4	57	20.7	27.3
ASN386	THR455	CD4	55	18.3	27.9
ASN386	SER375	CD4	11	10.3	28.2
ASN386	MET426	CD4	26	21.1	29.1
ASN386	VAL430	CD4	30	24.2	29.5
ASN386	TRP427	CD4	27	19.3	31
ASN386	ASN280	CD4	84	24.5	31.4
ASN386	ARG456	CD4	56	22.1	31.5
ASN386	ALA281	CD4	83	24.8	31.9
ASN386	THR283	CD4	81	20.8	32.6
ASN386	GLU429	CD4	29	25.9	34.5
ASN386	GLN428	CD4	28	24.4	34.6
ASN386	ASP474	CD4	71	21.5	34.7
ASN386	MET475	CD4	72	20.9	34.9
ASN386	LYS282	CD4	82	25.6	36.3
ASN386	ASN279	CD4	85	28.4	37.6
ASN386	ASP477	CD4	74	21.3	37.7
ASN386	ARG476	CD4	73	25.1	39
ASN386	LYS421	CG10	21	12.1	18.1
ASN386	ILE423	CG10	23	17	19.8
ASN386	GLN422	CG10	22	17.6	20.9
ASN386	LYS432	CG10	32	20.6	24.2
ASN386	TYR435	CG10	35	20.3	25.4
ASN386	MET434	CG10	34	21.9	25.4
ASN386	LYS121	CG10	170	26.2	31.9
ASN386	LYS207	CG10	157	24.4	34
ASN386	THR123	CG10	168	27.9	35.4
ASN448	LYS121	17B	218	31.1	-3
ASN448	ARG419	17B	29	22.4	50.3
ASN448	LYS421	17B	27	23.4	56.5
ASN448	ILE420	17B	28	21.6	56.9
ASN448	GLN422	17B	26	27.3	58.4
ASN448	ILE423	17B	25	28.2	60.6
ASN448	TYR435	17B	13	25.5	62.4
ASN448	SER256	B12	156	11.5	25.7
ASN448	THR257	B12	155	16.2	31.1
ASN448	GLU370	B12	64	20.9	34.7
ASN448	LEU453	B12	5	15.4	36.7
ASN448	ILE371	B12	63	21.2	36.8

ASN448	THR373	B12	61	19.3	37.7
ASN448	PRO470	B12	19	17.8	38
ASN448	PRO369	B12	65	23.6	38.5
ASN448	ASP368	B12	66	25.4	38.9
ASN448	VAL372	B12	62	23	39.8
ASN448	TYR384	B12	50	19.8	40.1
ASN448	SER364	B12	70	22.2	40.7
ASN448	PRO417	B12	31	21.2	42.3
ASN448	ASN386	B12	48	20.4	43.6
ASN448	ARG419	B12	29	22.4	44.3
ASN448	SER365	B12	69	27.2	44.9
ASN448	CYS418	B12	30	17.7	45.5
ASN448	ASP474	B12	23	21.7	45.9
ASN448	VAL430	B12	18	30.5	46
ASN448	THR455	B12	7	22.8	47.3
ASN448	MET475	B12	24	18.4	48.1
ASN448	ALA281	B12	131	29	51.2
ASN448	ARG456	B12	8	25.5	53
ASN448	ASN280	B12	132	29	53.9
ASN448	ASP457	B12	9	28.6	54
ASN448	LYS432	B12	16	30	54.4
ASN448	ARG444	CCR5	4	14.9	24
ASN448	LYS207	CCR5	205	24.9	33.6
ASN448	ARG440	CCR5	8	27.1	37
ASN448	HIS330	CCR5	104	18.1	38.2
ASN448	PRO437	CCR5	11	29	44
ASN448	LYS117	CCR5	222	28.3	45.7
ASN448	PRO438	CCR5	10	24	47
ASN448	GLN422	CCR5	26	27.3	50.2
ASN448	ARG419	CCR5	29	22.4	50.3
ASN448	ILE420	CCR5	28	21.6	51.1
ASN448	LYS421	CCR5	27	23.4	55.1
ASN448	THR123	CCR5	216	33.6	68.8
ASN448	LYS121	CCR5	218	31.1	72.3
ASN448	SER365	CD4	69	27.2	63.2
ASN448	ASP457	CD4	9	28.6	64.6
ASN448	ASN279	CD4	133	29.7	65
ASN448	ASN280	CD4	132	29	65.8
ASN448	ARG469	CD4	18	22.7	67.4
ASN448	LYS432	CD4	16	30	67.4
ASN448	GLN428	CD4	20	24.8	67.6

ASN448	LYS282	CD4	130	25.7	67.8
ASN448	GLU429	CD4	19	28.8	68
ASN448	ASP368	CD4	66	25.4	68.1
ASN448	ASN425	CD4	23	25.3	69
ASN448	ARG456	CD4	8	25.5	69.8
ASN448	MET426	CD4	22	25.6	69.9
ASN448	ALA281	CD4	131	29	70.1
ASN448	THR455	CD4	7	22.8	71.4
ASN448	GLU370	CD4	64	20.9	71.5
ASN448	ILE371	CD4	63	21.2	72.6
ASN448	TRP427	CD4	21	21.1	73.8
ASN448	THR283	CD4	129	20.7	74.4
ASN448	VAL430	CD4	18	30.5	74.9
ASN448	SER375	CD4	59	17.3	75.4
ASN448	MET475	CD4	24	18.4	76.4
ASN448	ASP474	CD4	23	21.7	77.3
ASN448	ARG476	CD4	25	21.2	78
ASN448	ASP477	CD4	26	17.9	78.6
ASN448	LYS207	CG10	205	24.9	33.6
ASN448	TYR435	CG10	13	25.5	42.6
ASN448	GLN422	CG10	26	27.3	47.3
ASN448	MET434	CG10	14	30.1	47.9
ASN448	ILE423	CG10	25	28.2	50.3
ASN448	LYS421	CG10	27	23.4	50.8
ASN448	LYS432	CG10	16	30	55.1
ASN448	LYS121	CG10	218	31.1	55.2
ASN448	THR123	CG10	216	33.6	60.6
GLU339	LYS121	17B	123	40.8	-3
GLU339	ARG419	17B	66	24.9	33.7
GLU339	LYS421	17B	68	28.7	38.7
GLU339	ILE420	17B	67	28	39.3
GLU339	GLN422	17B	69	34.1	43.4
GLU339	ILE423	17B	70	33.9	43.8
GLU339	TYR435	17B	82	35.2	47.5
GLU339	PRO417	B12	64	19	24.2
GLU339	ASN386	B12	47	18.1	24.6
GLU339	CYS418	B12	65	20.8	29.2
GLU339	THR373	B12	34	20.3	29.3
GLU339	SER364	B12	25	19.4	30.5
GLU339	SER365	B12	26	22.6	31.3
GLU339	VAL372	B12	33	22.3	32.2

GLU339	TYR384	B12	45	23.4	32.5
GLU339	ARG419	B12	66	24.9	33
GLU339	PRO369	B12	30	25.5	34.6
GLU339	ASP457	B12	104	21.8	35.5
GLU339	PRO470	B12	114	16	35.9
GLU339	ILE371	B12	32	23.3	37.6
GLU339	GLU370	B12	31	25.7	37.7
GLU339	ASP368	B12	29	28	38.5
GLU339	THR257	B12	60	21.4	39.4
GLU339	ARG456	B12	103	20.9	40.1
GLU339	THR455	B12	102	20.9	40.8
GLU339	ASN280	B12	37	25.1	42.2
GLU339	LEU453	B12	100	19.4	43.8
GLU339	SER256	B12	61	21.2	44.1
GLU339	ALA281	B12	36	28.3	44.9
GLU339	ASP474	B12	118	28.2	48.5
GLU339	VAL430	B12	77	36.8	49.8
GLU339	LYS432	B12	79	36.3	51.9
GLU339	MET475	B12	119	28.3	53.4
GLU339	HIS330	CCR5	9	19.4	27.8
GLU339	ARG419	CCR5	66	24.9	33.7
GLU339	LYS421	CCR5	68	28.7	38.7
GLU339	ILE420	CCR5	67	28	39.3
GLU339	ARG444	CCR5	91	24.8	43.3
GLU339	GLN422	CCR5	69	34.1	43.4
GLU339	PRO438	CCR5	85	31.8	44.4
GLU339	PRO437	CCR5	84	36.7	47.9
GLU339	ARG440	CCR5	87	36.3	48.8
GLU339	LYS207	CCR5	110	37	56.9
GLU339	THR123	CCR5	121	41.8	60.2
GLU339	LYS121	CCR5	123	40.8	67
GLU339	LYS117	CCR5	127	40.9	70
GLU339	SER365	CD4	26	22.6	31.3
GLU339	ASP457	CD4	104	21.8	35.5
GLU339	ARG469	CD4	113	17.6	36.1
GLU339	ARG456	CD4	103	20.9	40.1
GLU339	THR455	CD4	102	20.9	40.7
GLU339	ILE371	CD4	32	23.3	41.5
GLU339	ASN280	CD4	37	25.1	42.2
GLU339	ASP368	CD4	29	28	43.5
GLU339	ALA281	CD4	36	28.3	44.9

GLU339	THR283	CD4	34	23.2	46
GLU339	ASN425	CD4	72	30.8	46.8
GLU339	GLU370	CD4	31	25.7	47
GLU339	ASN279	CD4	38	28.7	48
GLU339	LYS282	CD4	35	26.5	48.8
GLU339	SER375	CD4	36	23	49.4
GLU339	LYS432	CD4	79	36.3	49.9
GLU339	ASP474	CD4	118	28.2	50
GLU339	ASP477	CD4	121	24.9	51.9
GLU339	VAL430	CD4	77	36.8	52
GLU339	TRP427	CD4	74	30	52.3
GLU339	MET426	CD4	73	34.3	52.7
GLU339	MET475	CD4	119	28.3	54.8
GLU339	ARG476	CD4	120	30.2	55.7
GLU339	GLN428	CD4	75	35.4	56.8
GLU339	GLU429	CD4	76	37.8	57
GLU339	LYS421	CG10	68	28.7	41.9
GLU339	GLN422	CG10	69	34.1	44
GLU339	ILE423	CG10	70	33.9	44.2
GLU339	LYS432	CG10	79	36.3	48.6
GLU339	MET434	CG10	81	38.1	49.8
GLU339	TYR435	CG10	82	35.2	50
GLU339	LYS207	CG10	110	37	56.9
GLU339	THR123	CG10	121	41.8	60.2
GLU339	LYS121	CG10	123	40.8	67

SASD between each PNGS and each binding site. Atom 1 is the position of a given PNGS while Atom 2 is the position of a binding site, both on the protein structure 3TGQ. The third column describes the binding site type of Atom 2. SeqD is the distance between Atom 1 and Atom 2 on the primary sequence of 3TGQ chain D by counting intervening amino acids. The EucD and SASD are the shortest Euclidean and Solvent Accessible Surface distances between Atom 1 and Atom 2, respectively. SASDs of -3 indicate that the second atom was not surface accessible.

Table 13

Atom 1	Atom 2	Binding Type	SeqD	EucD	SASD
ASN386	ASN386	B12	0	0	0
ASN276	ASN279	CD4	3	3.6	4
ASN386	THR373	B12	13	5.1	5.1
ASN362	ARG469	CD4	90	5.2	5.2
ASN386	PRO417	B12	17	4.9	5.8
ASN276	LYS282	CD4	6	4.7	5.9
ASN332	HIS330	CCR5	2	7	7.3
ASN386	CYS418	B12	18	7.4	7.4
ASN386	TYR384	B12	2	7.4	8
ASN386	VAL372	B12	14	8.3	8.3
ASN362	SER365	CD4	3	8.2	8.3
ASN386	ARG419	17B	19	7.9	8.9
ASN386	PRO369	B12	17	9.3	9.3
ASN276	ALA281	CD4	5	8.4	9.4
ASN362	ASP457	CD4	81	8.3	9.7
ASN276	ASN280	CD4	4	9	10.2
ASN362	THR455	CD4	79	10.5	11
ASN386	SER364	B12	22	9.7	11
ASN301	ARG440	CCR5	103	10.4	11.4
ASN276	THR283	CD4	7	10.2	12
ASN386	GLU370	B12	16	12.5	12.5
ASN386	ILE371	B12	15	12.6	12.8
ASN362	ARG456	CD4	80	11	12.8
ASN386	LYS421	17B	21	12.1	13.4
ASN386	PRO470	B12	67	11.1	14.1
ASN386	THR257	B12	107	12.3	14.3
ASN386	ASP368	B12	18	14	14.8
ASN386	ILE420	17B	20	13	15.1
ASN276	ASP477	CD4	162	14	15.4
ASN276	ASP474	CD4	159	15.1	15.9
ASN301	PRO437	CCR5	100	15.1	16.9
ASN276	ARG476	CD4	161	16	17.8
ASN301	PRO438	CCR5	101	11.9	18.7
ASN386	ILE423	17B	23	17	18.9
ASN295	ARG444	CCR5	113	7.2	19
ASN386	SER256	B12	108	14.5	19.5
ASN386	GLN422	17B	22	17.6	19.5
ASN386	LEU453	B12	53	16.6	20.2
ASN276	MET475	CD4	160	20.2	21.3

ASN386	TYR435	17B	35	20.3	22.1
ASN301	LYS207	CCR5	94	18.7	22.6
ASN386	ASN425	CD4	25	16.3	22.8
ASN362	SER375	CD4	13	18.3	23.8
ASN386	LYS432	CG10	32	20.6	24.2
ASN386	VAL430	B12	30	24.2	25.1
ASN276	TRP427	CD4	115	23.3	25.3
ASN386	MET434	CG10	34	21.9	25.4
ASN276	GLN428	CD4	116	26.3	28
ASN276	GLU429	CD4	117	26.1	29
ASN386	MET426	CD4	26	21.1	29.1
ASN386	THR123	CG10	168	27.9	35.4
ASN301	LYS117	CCR5	111	27.5	37.1

Shortest SASD between a binding site and its closest PNGS. Atom 1 is the position of a given PNGS while Atom 2 is the position of a binding site, both on the protein structure 3TGQ. The third column describes the binding site type of Atom 2. SeqD is the distance between Atom 1 and Atom 2 on the primary sequence of 3TGQ chain D by counting amino acids. The EucD and SASD are the shortest Euclidean and Solvent Accessible Surface distances between Atom 1 and Atom 2, respectively. The table is sorted from shortest to longest SASD between any binding site and PNGS.

Table 14

RESIDUE	Chain_Name	AA	ASA	%ASA	dfi	%dfi
45	D	W	121	0.941	0.001085	0.88
46	D	K	170	0.979	0.001044	0.854
47	D	E	101	0.867	0.000918	0.766
48	D	A	36	0.511	0.000876	0.736
49	D	T	103	0.88	0.000752	0.579
50	D	T	24	0.424	0.000698	0.499
51	D	T	114	0.919	0.000601	0.354
52	D	L	23	0.414	0.000573	0.308
53	D	F	78	0.774	0.000583	0.327
54	D	C	4	0.161	0.000557	0.285
55	D	A	1	0.062	0.000659	0.446
56	D	S	5	0.184	0.000696	0.497
57	D	D	98	0.857	0.000802	0.65
58	D	A	21	0.39	0.000753	0.58
59	D	K	98	0.857	0.000845	0.699
60	D	A	88	0.824	0.000846	0.699
61	D	Y	183	0.989	0.000914	0.764
62	D	D	62	0.677	0.000835	0.688
63	D	T	112	0.914	0.000778	0.618
64	D	E	6	0.206	0.000703	0.505
65	D	V	15	0.332	0.000583	0.328
66	D	H	4	0.161	0.000551	0.273
67	D	N	15	0.332	0.000662	0.451
68	D	V	67	0.718	0.00064	0.415
69	D	W	10	0.262	0.000525	0.227
70	D	A	2	0.107	0.000577	0.314
71	D	T	65	0.705	0.00068	0.475
72	D	H	150	0.97	0.00061	0.37
73	D	A	49	0.615	0.000535	0.243
74	D	C	20	0.378	0.000603	0.357
75	D	V	64	0.697	0.000725	0.54
76	D	P	120	0.941	0.000833	0.683
77	D	T	33	0.489	0.000835	0.687
78	D	D	76	0.758	0.000963	0.801
79	D	P	103	0.88	0.001077	0.875
80	D	N	115	0.922	0.001124	0.903
81	D	P	73	0.744	0.001022	0.842
82	D	Q	143	0.96	0.001087	0.882

83	D	E	92	0.841	0.001043	0.852
84	D	V	81	0.793	0.001112	0.895
85	D	K	87	0.816	0.001142	0.909
86	D	L	30	0.472	0.001167	0.92
87	D	E	151	0.971	0.001296	0.952
88	D	N	171	0.981	0.001345	0.959
89	D	V	35	0.5	0.00124	0.94
90	D	T	79	0.781	0.001203	0.932
91	D	E	39	0.536	0.001109	0.894
92	D	N	112	0.914	0.001079	0.877
93	D	F	6	0.206	0.000952	0.793
94	D	N	70	0.731	0.000912	0.763
95	D	M	1	0.062	0.000791	0.632
96	D	W	18	0.354	0.000835	0.685
97	D	K	140	0.958	0.000841	0.694
98	D	N	4	0.161	0.000724	0.536
99	D	N	46	0.597	0.000653	0.433
100	D	M	3	0.142	0.000558	0.287
101	D	V	0	0	0.000517	0.215
102	D	E	62	0.677	0.000502	0.186
103	D	Q	42	0.56	0.000455	0.102
104	D	M	0	0	0.000392	0.06
105	D	H	7	0.222	0.000359	0.046
106	D	E	104	0.888	0.000348	0.039
107	D	D	45	0.593	0.000333	0.032
108	D	I	1	0.062	0.000303	0.019
109	D	I	22	0.402	0.00028	0.013
110	D	S	41	0.55	0.000314	0.023
111	D	L	1	0.062	0.000342	0.034
112	D	W	9	0.247	0.000328	0.029
113	D	D	64	0.697	0.000333	0.031
114	D	Q	105	0.89	0.000405	0.064
115	D	S	11	0.282	0.000438	0.09
116	D	L	13	0.308	0.000414	0.071
117	D	K	79	0.781	0.000414	0.072
118	D	P	10	0.262	0.00043	0.084
119	D	C	27	0.449	0.000509	0.204
120	D	V	49	0.615	0.000489	0.161
121	D	K	30	0.472	0.000416	0.074

122	D	L	75	0.754	0.000467	0.125
123	D	T	19	0.365	0.000465	0.119
124	D	G	91	0.836	0.000549	0.269
198	D	G	78	0.774	0.000578	0.318
199	D	S	42	0.56	0.000529	0.237
200	D	V	77	0.763	0.000506	0.198
201	D	I	33	0.489	0.00046	0.11
202	D	T	27	0.449	0.000497	0.172
203	D	Q	37	0.52	0.000502	0.185
204	D	A	54	0.638	0.000531	0.24
205	D	C	34	0.493	0.00052	0.219
206	D	P	86	0.811	0.000613	0.376
207	D	K	74	0.75	0.000656	0.439
208	D	V	56	0.648	0.00068	0.475
209	D	S	66	0.711	0.00072	0.532
210	D	F	16	0.344	0.000645	0.422
211	D	E	140	0.958	0.000684	0.48
212	D	P	22	0.402	0.000608	0.366
213	D	I	19	0.365	0.000657	0.441
214	D	P	40	0.545	0.000661	0.448
215	D	I	3	0.142	0.000577	0.315
216	D	H	39	0.536	0.000625	0.392
217	D	Y	4	0.161	0.000578	0.317
218	D	C	0	0	0.000667	0.46
219	D	A	8	0.236	0.00072	0.531
220	D	P	30	0.472	0.000775	0.614
221	D	A	101	0.867	0.000903	0.756
222	D	G	37	0.52	0.000972	0.81
223	D	F	43	0.574	0.000874	0.734
224	D	A	17	0.349	0.000855	0.711
225	D	I	2	0.107	0.000795	0.641
226	D	L	4	0.161	0.000866	0.725
227	D	K	37	0.52	0.00087	0.731
228	D	C	2	0.107	0.000943	0.784
229	D	N	41	0.55	0.001022	0.842
230	D	D	44	0.583	0.001084	0.88
231	D	K	114	0.919	0.001118	0.9
232	D	K	188	0.99	0.001128	0.905
233	D	F	12	0.293	0.001038	0.848

234	D	N	89	0.827	0.001034	0.848
235	D	G	0	0	0.000938	0.781
236	D	T	53	0.632	0.001027	0.845
237	D	G	28	0.459	0.001126	0.904
238	D	P	84	0.803	0.001151	0.912
239	D	C	0	0	0.001112	0.897
240	D	T	82	0.797	0.001197	0.929
241	D	N	51	0.625	0.001132	0.906
242	D	V	0	0	0.001047	0.858
243	D	S	0	0	0.001013	0.836
244	D	T	32	0.485	0.000948	0.79
245	D	V	13	0.308	0.000891	0.747
246	D	Q	102	0.874	0.000847	0.702
247	D	C	26	0.44	0.000751	0.576
248	D	T	0	0	0.000691	0.489
249	D	H	55	0.643	0.000728	0.545
250	D	G	23	0.414	0.000686	0.484
251	D	I	2	0.107	0.000588	0.338
252	D	R	99	0.86	0.000584	0.331
253	D	P	4	0.161	0.000467	0.122
254	D	V	17	0.349	0.000528	0.236
255	D	V	18	0.354	0.000481	0.147
256	D	S	1	0.062	0.000596	0.35
257	D	T	1	0.062	0.000684	0.478
258	D	Q	0	0	0.000822	0.671
259	D	L	0	0	0.000821	0.668
260	D	L	8	0.236	0.000732	0.551
261	D	L	11	0.282	0.00076	0.597
262	D	N	42	0.56	0.000792	0.635
263	D	G	24	0.424	0.00075	0.575
264	D	S	43	0.574	0.000826	0.673
265	D	L	64	0.697	0.000949	0.79
266	D	A	9	0.247	0.000995	0.828
267	D	E	128	0.95	0.001109	0.893
268	D	E	106	0.894	0.001198	0.93
269	D	E	111	0.911	0.001195	0.928
270	D	I	11	0.282	0.001115	0.898
271	D	V	20	0.378	0.001041	0.85
272	D	I	17	0.349	0.001028	0.845

273	D	R	7	0.222	0.000974	0.812
274	D	S	5	0.184	0.001012	0.836
275	D	E	75	0.754	0.001045	0.856
276	D	N	72	0.737	0.001205	0.933
277	D	F	28	0.459	0.001255	0.946
278	D	T	114	0.919	0.001408	0.971
279	D	N	55	0.643	0.001323	0.956
280	D	N	62	0.677	0.001285	0.949
281	D	A	77	0.763	0.001182	0.924
282	D	K	77	0.763	0.001069	0.869
283	D	T	25	0.432	0.000943	0.784
284	D	I	4	0.161	0.000968	0.807
285	D	I	2	0.107	0.000848	0.704
286	D	V	1	0.062	0.000922	0.771
287	D	Q	9	0.247	0.000904	0.757
288	D	L	2	0.107	0.001019	0.84
289	D	N	57	0.652	0.001114	0.897
290	D	E	137	0.955	0.001182	0.923
291	D	S	45	0.593	0.001112	0.895
292	D	V	5	0.184	0.001166	0.918
293	D	V	57	0.652	0.001161	0.915
294	D	I	2	0.107	0.001123	0.902
295	D	N	46	0.597	0.001164	0.917
296	D	C	2	0.107	0.001081	0.879
297	D	T	20	0.378	0.001148	0.911
298	D	R	7	0.222	0.001132	0.907
299	D	P	44	0.583	0.00128	0.948
300	D	N	102	0.874	0.001311	0.954
301	D	N	152	0.974	0.001411	0.973
324	D	G	106	0.894	0.00137	0.965
325	D	D	61	0.67	0.001236	0.939
326	D	I	90	0.832	0.001122	0.901
327	D	R	72	0.737	0.001024	0.844
328	D	Q	121	0.941	0.001132	0.908
329	D	A	5	0.184	0.001126	0.903
330	D	H	53	0.632	0.001171	0.921
331	D	C	0	0	0.001161	0.916
332	D	N	45	0.593	0.001291	0.95
333	D	L	4	0.161	0.001303	0.953

334	D	S	26	0.44	0.001445	0.976
335	D	K	79	0.781	0.0015	0.984
336	D	T	87	0.816	0.001577	0.989
337	D	Q	104	0.888	0.001452	0.976
338	D	W	2	0.107	0.001351	0.964
339	D	E	101	0.867	0.001462	0.979
340	D	N	76	0.758	0.001464	0.98
341	D	T	12	0.293	0.001302	0.953
342	D	L	6	0.206	0.00133	0.957
343	D	E	103	0.88	0.001455	0.978
344	D	Q	79	0.781	0.00138	0.967
345	D	I	1	0.062	0.001265	0.947
346	D	A	7	0.222	0.00139	0.968
347	D	I	78	0.774	0.001444	0.975
348	D	K	43	0.574	0.001318	0.955
349	D	L	1	0.062	0.001347	0.96
350	D	K	71	0.733	0.001503	0.985
351	D	E	117	0.928	0.001458	0.979
352	D	Q	87	0.816	0.001399	0.97
353	D	F	39	0.536	0.001514	0.986
354	D	G	30	0.472	0.001645	0.991
355	D	N	118	0.933	0.001719	0.996
356	D	N	156	0.976	0.001829	0.998
357	D	K	49	0.615	0.001703	0.994
358	D	T	62	0.677	0.001684	0.992
359	D	I	2	0.107	0.00152	0.987
360	D	I	48	0.609	0.001492	0.983
361	D	F	4	0.161	0.00135	0.963
362	D	N	43	0.574	0.001329	0.956
363	D	P	41	0.55	0.001214	0.935
364	D	S	25	0.432	0.001109	0.892
365	D	S	97	0.856	0.001199	0.931
366	D	G	44	0.583	0.001101	0.888
367	D	G	60	0.664	0.000977	0.814
368	D	D	93	0.844	0.000834	0.685
369	D	P	42	0.56	0.000839	0.69
370	D	E	19	0.365	0.000713	0.52
371	D	I	39	0.536	0.000799	0.647
372	D	V	50	0.621	0.00093	0.775

373	D	T	19	0.365	0.000911	0.761
374	D	H	1	0.062	0.000848	0.703
375	D	S	12	0.293	0.000723	0.534
376	D	F	12	0.293	0.000656	0.44
377	D	N	26	0.44	0.000635	0.408
378	D	C	0	0	0.000757	0.588
379	D	G	32	0.485	0.000788	0.628
380	D	G	10	0.262	0.000666	0.459
381	D	E	5	0.184	0.000712	0.518
382	D	F	15	0.332	0.000669	0.463
383	D	F	0	0	0.000794	0.638
384	D	Y	12	0.293	0.000841	0.693
385	D	C	0	0	0.000964	0.804
386	D	N	62	0.677	0.001064	0.868
387	D	S	4	0.161	0.001096	0.885
388	D	T	60	0.664	0.001244	0.941
389	D	Q	101	0.867	0.001339	0.959
390	D	L	0	0	0.001231	0.938
391	D	F	2	0.107	0.001248	0.943
392	D	T	91	0.836	0.001414	0.973
393	D	W	24	0.424	0.001489	0.982
394	D	N	89	0.827	0.001616	0.99
395	D	D	107	0.898	0.001694	0.993
396	D	T	146	0.967	0.001851	1
411	D	G	109	0.903	0.001802	0.997
412	D	R	178	0.988	0.001702	0.994
413	D	N	89	0.827	0.001547	0.988
414	D	I	15	0.332	0.00141	0.972
415	D	T	63	0.69	0.001349	0.962
416	D	L	0	0	0.001203	0.932
417	D	P	67	0.718	0.001191	0.926
418	D	C	2	0.107	0.001043	0.853
419	D	R	107	0.898	0.000964	0.804
420	D	I	1	0.062	0.000856	0.713
421	D	K	62	0.677	0.000758	0.591
422	D	Q	43	0.574	0.000685	0.482
423	D	I	72	0.737	0.000623	0.387
424	D	I	13	0.308	0.000542	0.256
425	D	N	27	0.449	0.000522	0.223

426	D	M	21	0.39	0.000429	0.084
427	D	W	42	0.56	0.000422	0.078
428	D	Q	31	0.478	0.000372	0.05
429	D	E	88	0.824	0.000437	0.089
430	D	V	139	0.956	0.000503	0.187
431	D	G	20	0.378	0.000482	0.15
432	D	K	93	0.844	0.000493	0.164
433	D	A	0	0	0.000458	0.107
434	D	M	50	0.621	0.000536	0.244
435	D	Y	21	0.39	0.000564	0.298
436	D	A	23	0.414	0.000681	0.476
437	D	P	73	0.744	0.000809	0.659
438	D	P	11	0.282	0.000863	0.723
439	D	I	68	0.723	0.000935	0.777
440	D	R	231	0.999	0.001096	0.886
441	D	G	26	0.44	0.001206	0.934
442	D	Q	114	0.919	0.001192	0.927
443	D	I	7	0.222	0.001052	0.863
444	D	R	121	0.941	0.001079	0.876
445	D	C	38	0.529	0.001015	0.837
446	D	S	60	0.664	0.001074	0.874
447	D	S	9	0.247	0.000967	0.807
448	D	N	61	0.67	0.000977	0.813
449	D	I	0	0	0.00094	0.781
450	D	T	6	0.206	0.000903	0.755
451	D	G	1	0.062	0.000849	0.707
452	D	L	1	0.062	0.000861	0.717
453	D	L	0	0	0.000831	0.68
454	D	L	6	0.206	0.000976	0.813
455	D	T	37	0.52	0.001073	0.872
456	D	R	25	0.432	0.001244	0.942
457	D	D	53	0.632	0.001368	0.965
458	D	G	36	0.511	0.001503	0.985
459	D	G	115	0.922	0.001635	0.991
463	D	N	195	0.993	0.00184	0.999
464	D	G	22	0.402	0.001812	0.997
465	D	T	77	0.763	0.001705	0.995
466	D	E	20	0.378	0.001531	0.988
467	D	I	41	0.55	0.001428	0.974

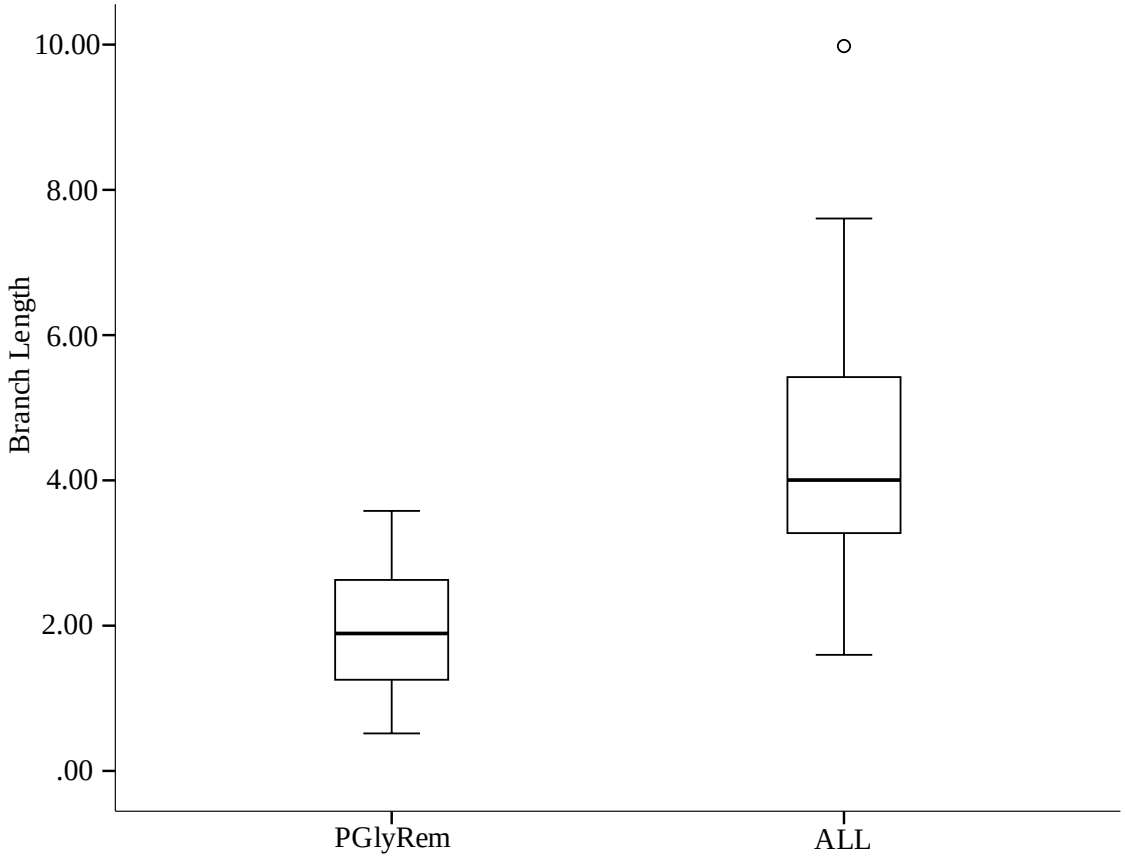
468	D	F	1	0.062	0.00126	0.947
469	D	R	52	0.628	0.001154	0.912
470	D	P	5	0.184	0.000994	0.827
471	D	G	10	0.262	0.000888	0.743
472	D	G	21	0.39	0.000729	0.549
473	D	G	35	0.5	0.000652	0.431
474	D	D	66	0.711	0.000596	0.349
475	D	M	10	0.262	0.000482	0.149
476	D	R	66	0.711	0.000519	0.218
477	D	D	15	0.332	0.000603	0.356
478	D	N	10	0.262	0.000531	0.239
479	D	W	3	0.142	0.000497	0.176
480	D	R	23	0.414	0.000608	0.367
481	D	S	9	0.247	0.00066	0.447
482	D	E	18	0.354	0.000626	0.393
483	D	L	0	0	0.000617	0.381
484	D	Y	13	0.308	0.000725	0.539
485	D	K	35	0.5	0.00079	0.631
486	D	Y	9	0.247	0.000747	0.572
487	D	K	14	0.322	0.000763	0.6
488	D	V	8	0.236	0.000769	0.606
489	D	V	0	0	0.000883	0.74
490	D	K	86	0.811	0.000955	0.797
491	D	I	72	0.737	0.001048	0.859
492	D	E	218	0.997	0.001166	0.919

Dynamic flexibility index output. The residue position is relative to 3TGQ positioning (which conveniently matches up with HXB2 positional naming). The D chain was used for all estimates (out of four identical chains). The one letter amino acid naming convention is used to identify the amino acid at each position. The solvent Accessible Surface Area, calculated via Surface Racer, is shown in ångstroms^2.The %ASA is calculated by dividing the ASA for a given residue by the largest ASA measurement. The dynamic flexibility index (DFI) is the average displacement of a given residue resulting from disturbances to the other residues in the chain. A higher DFI indicates increased flexibility relative to sites with lower DFI values, which are thought to be relatively rigid in comparison. Similar to %ASA, %DFI is calculated by dividing DFI at a given residue by the highest DFI value for the entire structure.

APPENDIX IV

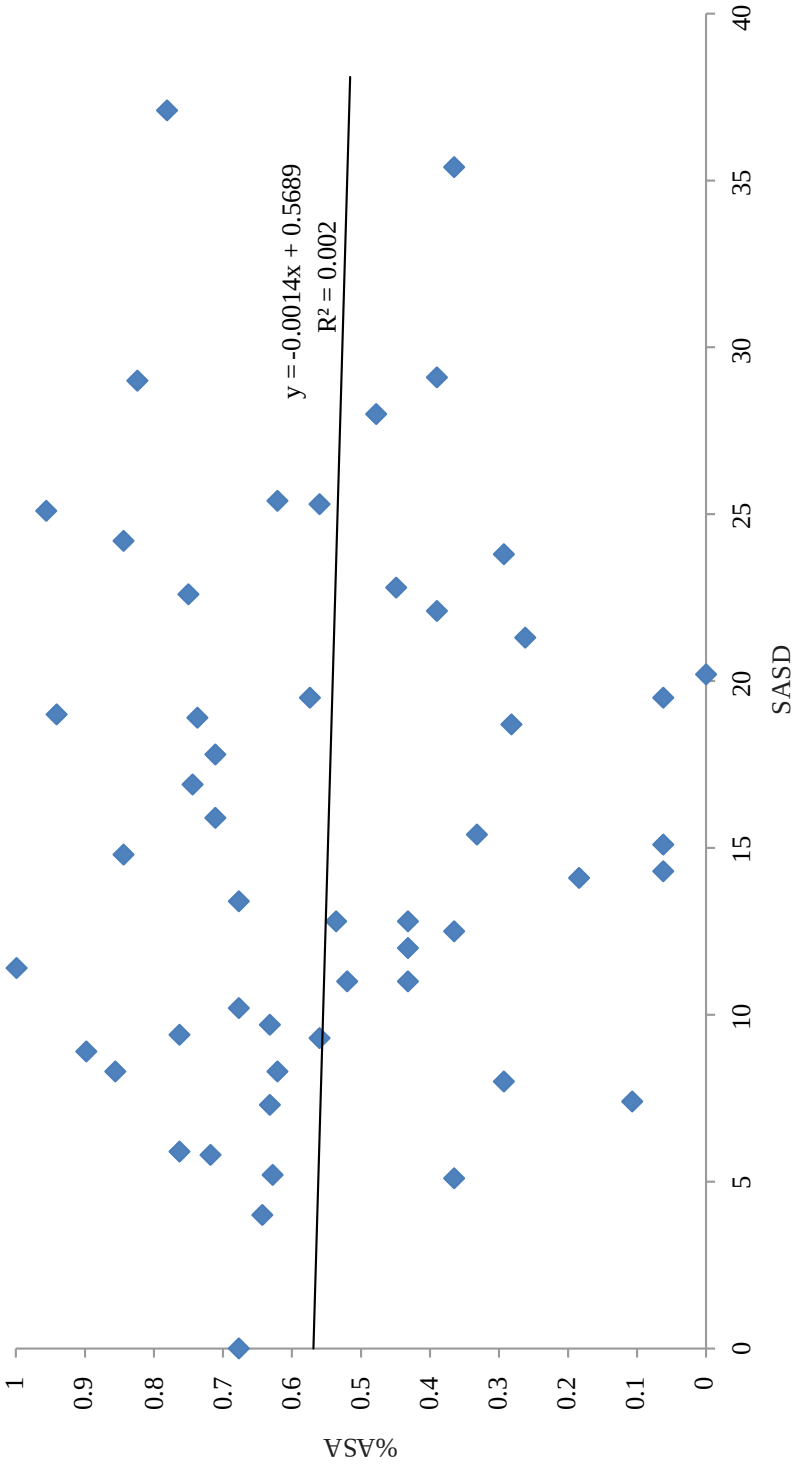
SUPPORTING FIGURES FOR CHAPTER 3

Figure 1



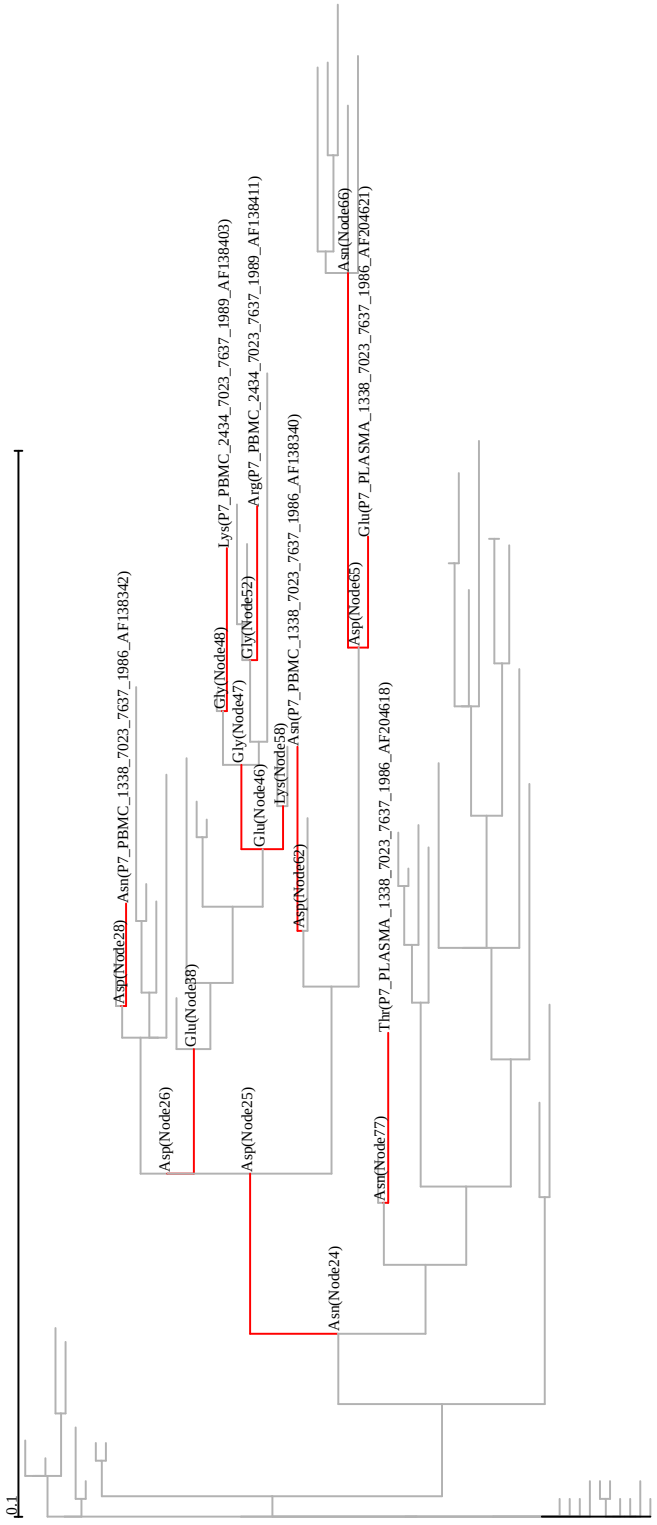
Box and whisker plot comparing the sum of branch lengths for PGlyRem and ALL datasets.

Figure 2



The shortest SASD for each of the 52 binding sites to the closest PNGS plotted against %ASA.

Figure 3



SLAC results for position 301. Red branches indicate a lineage where a non-synonymous mutation occurred.

BIOGRAPHICAL SKETCH

Crystal Hepp was born on February 12, 1981 in Great Falls, MT. She graduated from Conrad High School, in Conrad, MT, in 1999. Crystal attended Montana State University, in Bozeman, MT, where she received her Bachelor of Science in Microbiology. In 2013, Crystal received her Doctor of Philosophy degree in Molecular and Cellular Biology. She specializes in the evolution of infectious disease.