Understanding the evolution of HIV-1 Env through computational analysis and visualization of long-read amplicon sequences

Eren, Kemal

2017

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Understanding the evolution of HIV-1 Env through computational analysis and visualization of long-read amplicon sequences

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Kemal Eren

Committee in charge:

Professor Joel Okrent Wertheim, Chair
Professor Siavash Mir Arabbaygi, Co-Chair
Professor Benjamin Sylvester Murrell
Professor Sergei L. Kosakovsky Pond
Professor Douglas D. Richman
Professor Davey Smith

2017

The dissertation of Kemal Eren is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2017

DEDICATION

For Stephen Palmer

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I have been most fortunate in my advisor. Thank you Joel Wertheim, for all your support, assistance, and advice.

A special thanks to Ben Murrell for years of collaboration and mentorship. None of this would have been possible without your guidance. I owe much to your intellectual generosity.

My thanks to Sergei L. Kosakovsky Pond, for first advising me and for continuing to provide invaluable counsel.

I would also like to acknowledge the other members of my committee: Siavash Mir Arabbaygi, Douglas D. Richman, and Davey Smith.

I am indebted to Sasha Murrell. Thank you for all your help and hospitality throughout the years. I am especially grateful to you for tireless copyediting.

Thanks also to my many mentors, especially Arthur G. Palmer, Peter A. Zimmerman, Metin Eren, Joseph Popelka, and Ümit V. Çatalyürek. Your example continues to inspire me.

My siblings have always been a valuable source of motivation. Thank you Taner, Niko, Eleni, Deniz, Emre, Alek, and Myia.

Thank you, mom and dad. For everything.

Finally, thank you Sami for all your support. I will always, always be grateful.

Chapter 1 contains, in part, material that appeared in “Rapid sequencing of complete env genes from primary HIV-1 samples”, Melissa Laird Smith, Ben Murrell, Caroline Ignacio, Elise Landais, Steven Weaver, Pham Phung, Colleen Ludka, Lance Hepler, Gemma Caballero, Tristan Pollner, Yan Guo, Douglas Richman, IAVI Protocol C Investigators & The IAVI African HIV Research Network, Pascal Poignard, Ellen E Paxinos, Sergei L Kosakovsky Pond, Davey M Smith, *Virus Evolution* 2016. The dissertation author was an author of this paper.

Chapter 2, in full, is a reprint of material that has been submitted as “Full-Length Envelope Analyzer (FLEA): A tool for longitudinal analysis of viral amplicons”, Kemal Eren, Steven Weaver, Robert Ketteringham, Morné Valentyn, Melissa Laird Smith, Venkatesh Kumar, Sanjay Mohan, Sergei L Kosakovsky Pond, Ben Murrell. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of material that has been submitted as “RIFRAF: a frame-resolving consensus algorithm”, Kemal Eren, Ben Murrell. The dissertation author was the primary investigator and author of this paper.

Chapter 4 contains, in part, material that appeared in the poster “Full-length *env* deep sequencing in a donor with broadly neutralizing N332 antibodies”, Ben Murrell, Kemal Eren, Lorena S Ver, Nancy Choi, Elise Landais, Pascal Poignard, Sergei L Kosakovsky Pond, and Davey Smith, *Keystone Symposia on Molecular and Cellular Biology* 2016. The dissertation author was an author of this poster.

VITA

2008	B. S. in Biology, University of Michigan, Ann Arbor
2012	M. S. in Computer Science, The Ohio State University, Columbus
2017	Ph. D. in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

Kemal Eren, Ben Murrell, “RIFRAF: a frame-resolving consensus algorithm”. In submission.

Kemal Eren, Steven Weaver, Robert Ketteringham, Morné Valentyn, Melissa Laird Smith, Venkatesh Kumar, Sanjay Mohan, Sergei L Kosakovsky Pond, Ben Murrell, “Full-Length Envelope Analyzer (FLEA): A tool for longitudinal analysis of viral amplicons”. In submission.

Elise Landais, Ben Murrell, Bryan Briney, Sasha Murrell, Kimmo Rantalainen, Alejandra Ramos, Lalinda Wickramasinghe, Melissa Laird Smith, Kemal Eren, Zachary Berndsen, Natalia De Val, Mengyu Wu, Audrey Cappelletti, Yolanda Lie, Terri Wrin, Paul Algate, Etienne Karita, B Andrew, Ian A Wilson, Dennis R Burton, Davey Smith, L Sergei, Pascal Poignard, Computational Biology, Vaccine Im- munology, La Jolla, Biomedical Informatics, San Diego, Monogram Biosciences, Monogram Biosciences, Theraclone Sciences, and Project San Francisco, “HIV Envelope Glycoform Heterogeneity and Localized Diversity Govern the Initiation and Maturation of a V2 Apex Broadly Neutralizing Antibody Lineage”, *Immunity*, 47, 2017.

Marina Caskey, Till Schoofs, Henning Gruell, Allison Settler, Theodora Karagou- nis, Edward F Kreider, Ben Murrell, Nico Pfeifer, Lilian Nogueira, Thiago Y Oliveira, Gerald H Learn, Yehuda Z Cohen, Clara Lehmann, Daniel Gillor, Irina Shimeliovich, Cecilia Unson-OBrien, Daniela Weiland, Alexander Robles, Tim Ku mmerle, Christoph Wyen, Rebeka Levin, Maggi Witmer-Pack, Kemal Eren, Caroline Ignacio, Szilard Kiss, Anthony P West, Hugo Mouquet, Barry S Zingman, Roy M Gulick, Tibor Keler, Pamela J Bjorkman, Michael S Seaman, Beatrice H Hahn, Gerd Fa tkenheuer, Sarah J Schlesinger, Michel C Nussenzweig, and Flo- rian Klein, “Antibody 10-1074 suppresses viremia in HIV-1-infected individuals”, *Nature Medicine*, 23, 2017.

Daniel T. MacLeod, Nancy M. Choi, Bryan Briney, Fernando Garces, Lorena S. Ver, Elise Landais, Ben Murrell, Terri Wrin, William Kilembe, Chi Hui Liang, Alejandra Ramos, Chaoran B. Bian, Lalinda Wickramasinghe, Leopold Kong, Kemal Eren, Chung Yi Wu, Chi Huey Wong, Sergei L. Kosakovsky Pond, Ian A. Wilson, Dennis R. Burton, Pascal

Poignard, and The IAVI Protocol C Investigators, “Early antibody lineage diversification and independent limb maturation lead to broad HIV-1 neutralization targeting the Env high-mannose patch”, *Immunity*, 44, 2016.

Melissa Laird Smith, Ben Murrell, Caroline Ignacio, Elise Landais, Steven Weaver, Pham Phung, Colleen Ludka, Lance Hepler, Gemma Caballero, Tristan Pollner, Yan Guo, Douglas Richman, IAVI Protocol C Investigators & The IAVI African HIV Research Network, Pascal Poignard, Ellen E Paxinos, Sergei L Kosakovsky Pond, Davey M Smith, “Rapid sequencing of complete env genes from primary HIV-1 samples”, *Virus Evolution*, 2, 2016.

Ben Murrell, Kemal Eren, Lorena S Ver, Nancy Choi, Elise Landais, Pascal Poignard, Sergei Kosakovsky Pond, and Davey Smith. “Full-length env deep sequencing in a donor with broadly neutralizing N332 antibodies”, *Keystone Symposia on Molecular and Cellular Biology*, 2016.

Mehmet Deveci, Onur Küçüktunç, Kemal Eren, Doruk Bozdağ, Kamer Kaya, Ümit V Çatalyürek, “Querying Co-regulated Genes on Diverse Gene Expression Datasets Via Biclustering”, *Microarray Data Analysis: Methods and Applications*, 2015.

Ben Murrell, Steven Weaver, Martin D. Smith, Joel O. Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, Tristan Pollner, Darren P. Martin, Davey M. Smith, Konrad Scheffler, Sergei L. Kosakovsky Pond, “Gene-wide identification of episodic selection”, *Molecular Biology and Evolution*, 2015.

Kemal Eren, “Application of biclustering algorithms to biological data”, *The Ohio State University*, 2012.

Kemal Eren, Mehmet Deveci, Onur Küçüktunç, Ümit V Çatalyürek, “A comparative analysis of biclustering algorithms for gene expression data”, *Briefings in Bioinformatics*, 14, 2012.

ABSTRACT OF THE DISSERTATION

Understanding the evolution of HIV-1 Env through computational analysis and visualization of long-read amplicon sequences

by

Kemal Eren

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2017

Professor Joel Okrent Wertheim, Chair
Professor Siavash Mir Arabbaygi, Co-Chair

Single-molecule long-read sequencing technology recently reached accuracies useful for studying diverse viral genes and genomes. Challenging error profiles, however, hinder the interpretability of long-read sequencing datasets. Here we develop computational tools for processing such datasets and for visualizing rapidly evolving viral populations.

Our primary biological focus is the HIV-1 envelope protein, which is the only target of neutralizing antibodies. An effective HIV-1 vaccine would be a powerful weapon against the current global epidemic, but progress has been slow because Env is a difficult target. Nevertheless, some hosts develop broadly neutralizing antibodies (bNAbs), which could be protective if they could be elicited by vaccination. Env and bNAb lineages co-evolve, so understanding the Env populations and evolutionary dynamics will likely be critical for understanding how to elicit the desired immune response. Tools developed in this dissertation allow, for the first time, accurate processing of full-length sequencing of HIV-1 *env* populations. Computational challenges in analyzing these sequences include the length of the gene (2.6kb) and the prevalence of indel sequencing errors and extensive biological indel variation which render traditional approaches inaccurate.

`FLEA` is a pipeline for processing circular consensus sequences and providing biological insights into the evolution of *env*. It performs sequence cleaning, infers high-quality consensus sequences, and performs analyses including codon alignment, phylogenetic tree inference, ancestor reconstruction, and selection inference. The `FLEA` pipeline supports multiple cluster and high-performance computing environments. A client-side web application provides interactive visualizations, including a tree viewer, MSA browser, and three-dimensional structure viewer.

`RIFRAF` is a novel multi-objective sequence consensus algorithm. It uses per-base quality scores and uses a reference sequence for frame correction. `RIFRAF` consistently finds consensus sequences that are more accurate and in-frame than those from other methods, even with few reads and a distant reference. It is also uniquely capable of keeping true indels while removing spurious ones.

These tools have been used to study donors from the Protocol C primary infection cohort, resulting in two high-profile journal articles and another in preparation. They have also been used to analyze data from a phase-I clinical trial of an anti-Env monoclonal antibody therapy, published in Nature Medicine. This dissertation reviews those articles, focusing on the results obtained with these tools.

Chapter 1

Background and motivation

1.1 HIV-1

Human immunodeficiency virus (HIV) is a retrovirus that infects CD4-positive cells, such as T-cells and macrophages. After an initial viral load spike, the infection becomes asymptomatic for a period of months to years. Without treatment, this stage usually ends within ten years, causing acquired immunodeficiency syndrome (AIDS) due to the depletion of CD4-positive T cells. The compromised immune system is vulnerable to opportunistic infections or cancers, which are almost always fatal. A worldwide HIV pandemic has been raging for decades, causing millions of deaths.

HIV descends from Simian Immunodeficiency Virus (SIV), which crossed the species barrier into humans from primates [105]. SIV is an old virus, estimated to have been circulating in simian populations for more than 32,000 years, during which time it may have successfully crossed into humans multiple times [156]. However, the ancestor of the current pandemic crossed into humans relatively recently. HIV-1, which is the most prevalent type, is related to SIV in chimpanzees and gorillas. HIV-1 group M, which is responsible for the global pandemic, came from SIVcpz (a strain of HIV that infects chimpanzees), probably in the late 19th or early 20th century in Cameroon or the Democratic Republic of the Congo, as a result of human contact with chimpanzee blood while harvesting bushmeat [143]. Factors that may have contributed to its spread include social changes and expanding transportation networks [34]. An increasing number of cases occurred during the following decades, culminating in the official start of the epidemic in 1981 [120]. Since that time, 76.1 million people have become infected with HIV, and 35.0 million have died. 36.7 million people globally were living with HIV in 2016, of which 1.8 million became newly infected and 1 million died [142].

HIV is transmitted when HIV virions from an infected individual contact a mucosal surface or the bloodstream of an uninfected individual [129]. Rates of infection vary according to the mechanism of transmission, which include sharing needles, sexual transmission, mother to child transmission, and other forms of blood contact. Certain populations are more at risk of infection than others, partially due to differences in rates of these behaviors. The most at-risk populations include drug users, men who have sex with men (MSM), and sex workers [142].

Over the past decades, HIV has been the target of an intensive research effort to find treatments, vaccines, and cures. Thanks to the development of antiretroviral therapy (ART) and its combination into highly active antiretroviral therapy (HAART), the disease is currently manageable [97, 99, 112], but no effective vaccine or cure has yet been found. ART disrupts specific phases of the virus replication cycle, reducing the viral load to below detectable limits in plasma and reducing transmission rates [3, 28]. However, the virus persists inside latently inside cells such as CD4+ lymphocytes, macrophages, and monocytes [2] and in tissue reservoirs [154], and viremia returns if the patient stops therapy or if their viral population develops drug resistance. HIV rapidly acquires resistance to individual therapies, which is why HAART is the current standard of care; a combination of therapies that target different parts of the HIV life cycle make escape more difficult. Access to HAART has greatly increased the life expectancy, especially in high-income countries, and life expectancy continues to improve [139]. ART may also be used as a pre-exposure prophylactic (PrEP) for high-risk populations [55, 135, 160]. PrEP could even be used on an intervention basis to identify and treat high-risk individuals [73]. However, most HIV-positive individuals are not receiving treatment. Of the 36.7 million people infected with HIV in 2016, only 19.5 million people were accessing antiretroviral therapy because of factors such as expense, lack of access to medical care, social stigma of being HIV-positive or of accessing ART, and insufficient health and sex education [142].

Despite its success, HAART has its downsides. It is not curative, so a course of therapy must be taken for life, administered daily, and monitored for effectiveness. Low-level viremia may still have unknown long-term consequences. ARTs do not penetrate nervous tissue, so replication and neuropathology occurs unchecked in those tissues. It also modulates lipid and glucose metabolism, with potentially deleterious effects. The long-term effects of HAART itself are unknown, and HIV-positive individuals on HAART still experience excessive morbidity and mortality [118].

HAART and public health programs have contributed to the decline of AIDS-related deaths [142]. However, these programs have so far failed to completely halt the spread of HIV. Crucially, no cure or vaccine against HIV has ever been developed. The search for a vaccine is particularly important because models suggest that a vaccine that is only 50% effective would still make a huge contribution to halting the HIV pandemic, preventing 6.3 million new infections by 2035 [86]. HIV is difficult to treat and has resisted efforts to develop vaccines and cures for a number of reasons [119]. It evolves quickly, because it lacks proofreading machinery and has a fast generation time (as fast as 1.5 days). Viral populations therefore acquire large amounts of genetic diversity within a single host in a short amount of time [10, 18, 20], which allows it to acquire

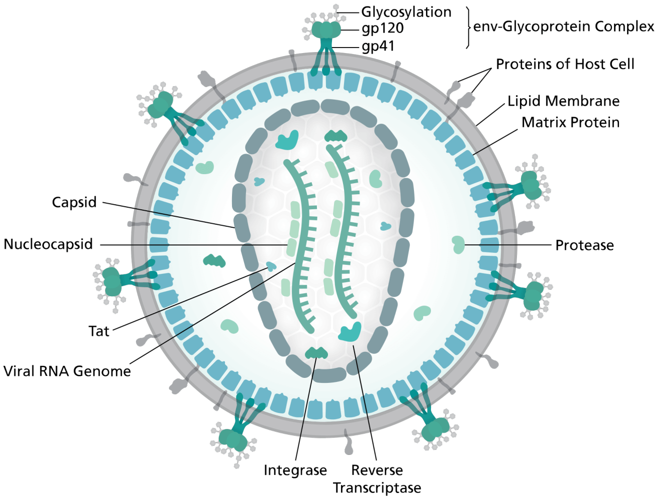


Figure 1.1: HIV-1 virion structure. Scale: approximately 120 nm in diameter. Credit Thomas Splettstoesser (www.scistyle.com).

drug resistance and evade the body's immune response. Latent viral DNA in host cells is not treatable by drugs or other therapies in the bloodstream. Reservoirs of virus exist intracellularly and within multiple tissues – such as nervous tissue – that are not accessible to current therapies [132]. In addition to these traits, specific factors relating to the structure and function of Env and the details of immune response make vaccines difficult to develop; these details are covered in Sections 1.1.2 and 1.1.3.

Latency makes it a challenge to develop a cure for HIV [78]. Researchers draw a distinction between a *sterilizing* cure, which would clear all trace of the infection from all tissues and cells, and a *functional* cure, which would suppress viremia and allow the host to live a normal life, despite the presence of reservoirs of latent virus [23].

1.1.1 Genome, structure, and replication

The HIV virion is shown in Figure 1.1. Its outermost membrane is a lipid bilayer which is dervived from the host cell when the virion buds from its surface. The membrane is studded with embedded viral envelope proteins, as well as host proteins from the host

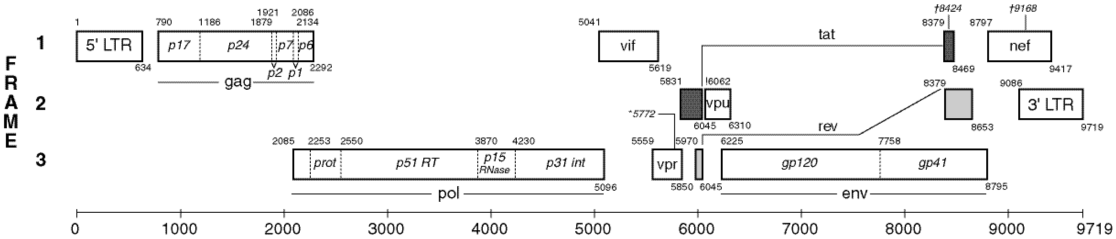


Figure 1.2: Landmarks of the HIV-1 genome, with genes in all three reading frames. From https://www.hiv.lanl.gov/content/sequence/HIV/MAP/landmark.html

cell. Inside the envelope is a matrix of protein p17, which has both a structural function and regulates various parts of the viral life cycle. The matrix contains copies of other viral proteins, and a capsid of viral protein p24. Inside the capsid are two copies of RNA, which encode the virus's genome, plus other viral proteins such as reverse transcriptase and integrase. Figure 1.2 shows the layout of the genes in the HIV-1 genome. Genes are encoded in different, sometimes overlapping, reading frames.

The virion reproduces by infecting host cells, reverse transcribing its RNA and integrating it into the host's DNA, inducing the host into producing more copies of itself. gp160 gets cleaved into gp41 and gp120, which are expressed on the surface of the host cell. The rest of the viral proteins are assembled into a new capsid, which buds from the surface of the cell as a new virion. The full process is shown in Figure 1.3. CD4-positive T cells die after infection; other types of cells do not [19]. Each part of the life cycle is a potential target for anti-retroviral therapy: entry inhibitors prevent binding and fusion; reverse transcriptase inhibitors inhibit the reverse transcription of viral RNA to DNA, either by binding to reverse transcriptase or by incorporation into and termination of the DNA strand; integrase inhibitors prevent the viral DNA from being integrated into the host DNA; protease inhibitors prevent protease from cleaving precursor proteins.

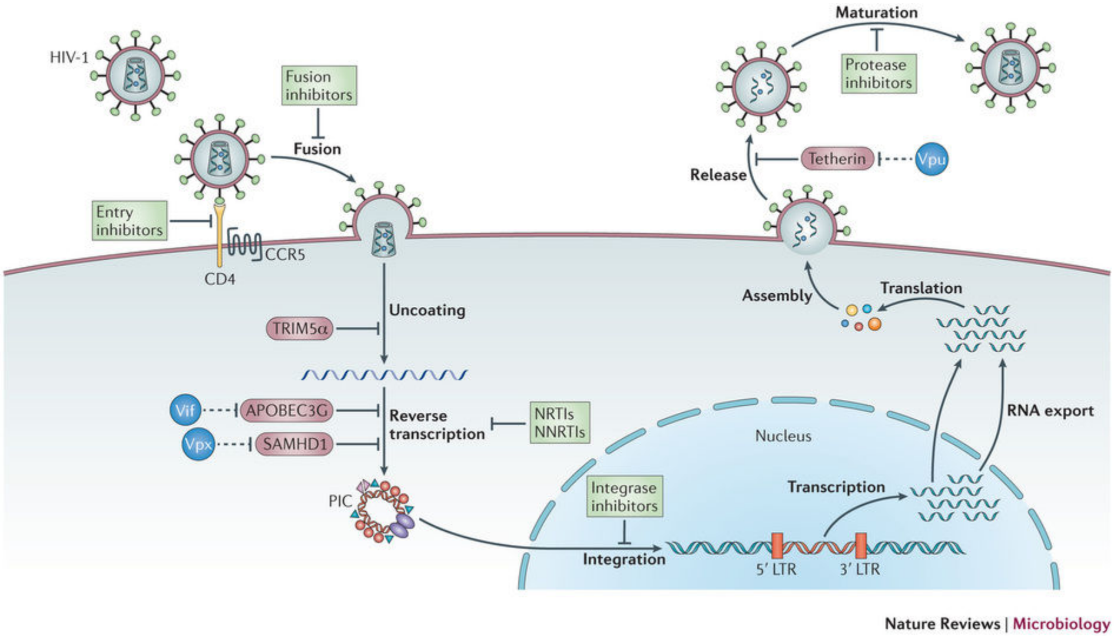


Figure 1.3: Replication cycle of HIV. HIV-1 infects a CD4-positive cell, reverse-transcribes its RNA to DNA, and integrates it into the host DNA. The host transcribes and translates the viral genes, producing proteins which are cleaved and packaged into new virions that bud from the surface of the cell and go on to infect more cells. From [4].

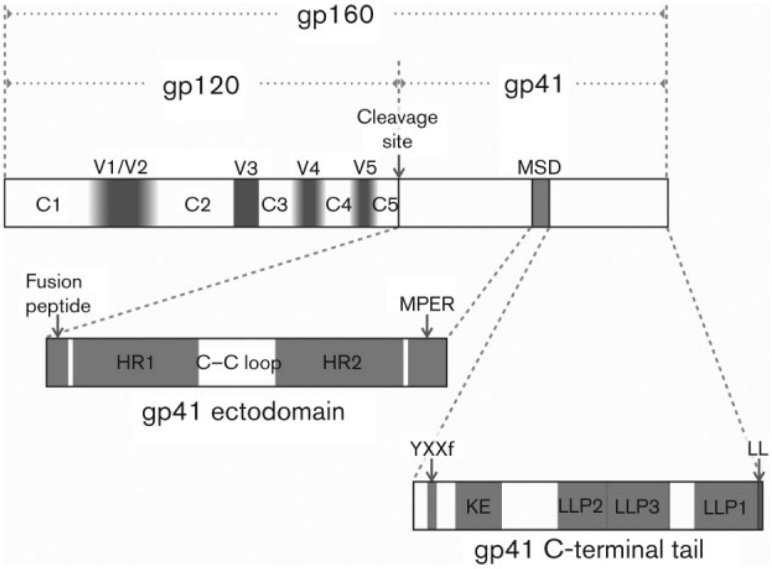
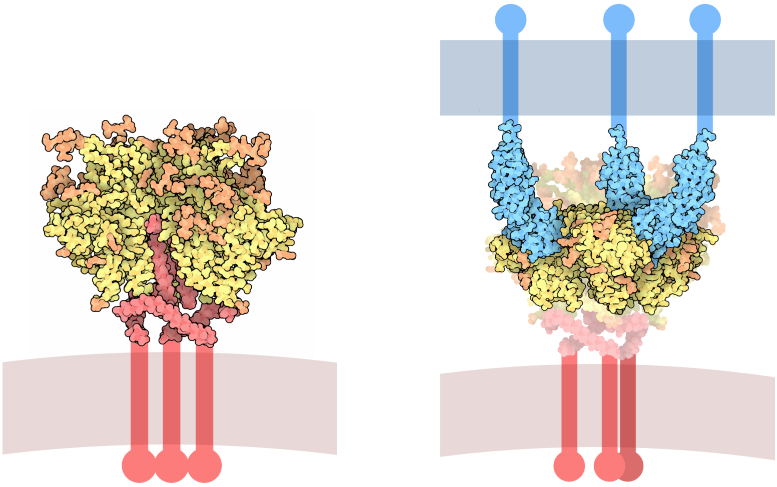


Figure 1.4: Regions of *gp160*, which codes for both gp120 and gp41. Three copies of each make up the viral spike Env. From Figure 2 [136].



(a) The structure of unbound Env.

(b) Env bound to CD4 receptor.

Figure 1.5: Structure of Env in unbound state, and bound to a CD4 receptor. Structural visualizations of Env from the Protein Data Bank [7] (www.rcsb.org).

1.1.2 Env

This work is chiefly concerned with Env, because it is the primary target of neutralizing antibodies. The *env* gene (Figure 1.4) is a 2.6kb gene that codes for the gp160 polyprotein. After transcription and translation, the gp160 polyprotein is cleaved into gp120 and gp41 [44]. The fully-assembled protein, also known as the “viral spike”, is a trimer of gp120/gp41 heterodimers: three copies of gp41 form the transmembrane protein, and three copies of gp120 that form the surface protein (Figure 1.5a). The viral spike is thickly shielded with glycans, hiding it from the immune system until it is close to a host cell and begins to bind to a CD4 receptor (Figure 1.5b). Env then undergoes a complex series of conformational changes as it binds to the CD4 receptor and to a chemokine co-receptor. The virion envelope fuses with the host cell’s membrane, releasing the payload into the cell and setting the rest of the replication cycle into motion (Figure 1.6).

Different strains of HIV use different chemokine co-receptors during envelope-

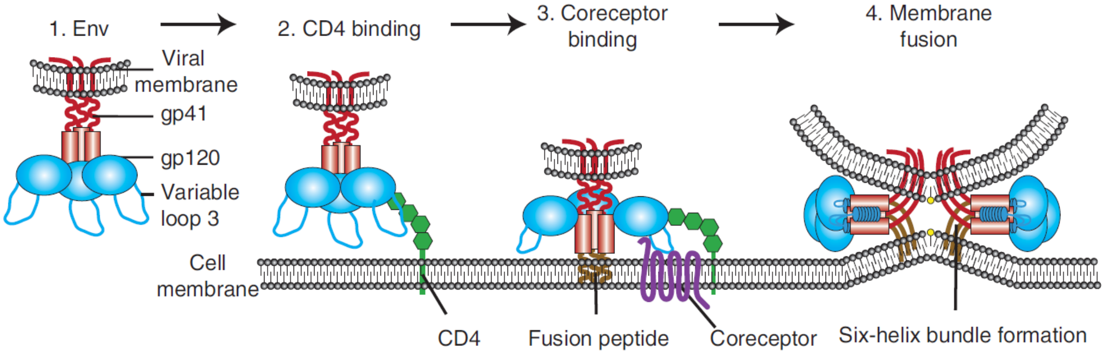


Figure 1.6: CD4-mediated entry. Env binds to a CD4 receptor and a chemokine co-receptor. It undergoes a conformational change that brings the viral envelope into close proximity with the host cell membrane, causing them to fuse. From [153].

mediated entry into the host cell. The ability of a strain to use a particular coreceptor is known as tropism. R5 viruses use the beta-chemokine CCR5 receptor, which is expressed on the surfaces of macrophages and T-cells. Almost all HIV subtypes use this coreceptor, and it has been extensively studied because of its effect on disease progression [6, 54, 84, 162]. For reasons yet unknown, most transmissions involve R5 virus [129]. X4 viruses use the alpha-chemokine CXCR4 receptor, which is expressed on T-cells. X4R5 viruses can use both. Moreover, some strains can infect CD4-negative cells via CXCR4 alone. HIV may also be transmitted directly from cell to cell *in vitro*, though whether it occurs *in vivo* remains a matter of debate [1].

1.1.3 Broadly-neutralizing antibodies

Antibodies are proteins produced or secreted by B cells that recognize and bind antigens. They either tag pathogens, which are then recognized and attacked by other components of the immune system, or they can directly neutralize the target. Antibodies are composed of two identical heavy chains and two light chains, each of which contains variable and constant regions, as shown in Figure 1.7. The amino acid sequence of the variable regions determines which antigens are recognized by a particular antibody. B

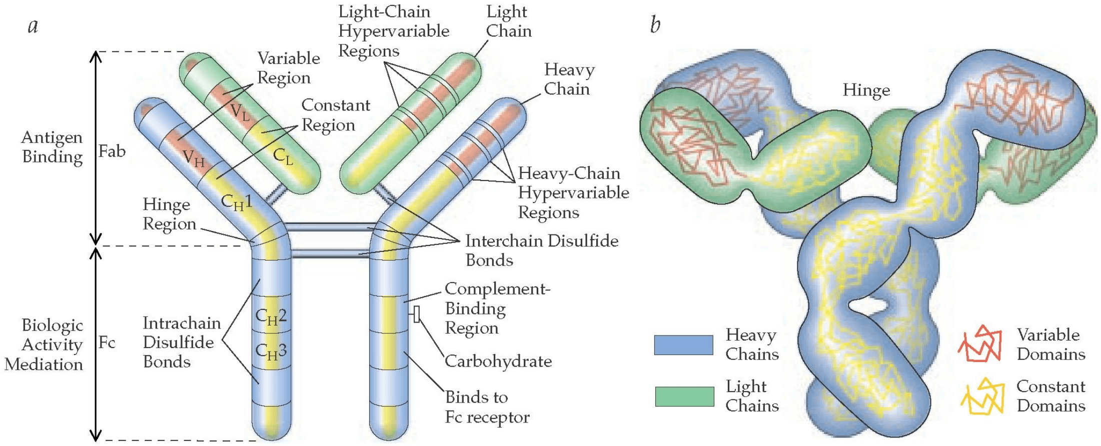


Figure 1.7: Antibody structure. From http://what-when-how.com/acp-medicine/adaptive-immunity-antigens-antibodies-and-t-cell-and-b-cell-receptors-part-1/.

cells code for a huge number of possible antibodies from a limited genetic repertoire via recombination of multiple V, D, and J genes [22, 92, 149], as shown in Figure 1.8. Mutations can further increase their variability. The circulating naive B cell repertoire is capable of recognizing $\sim 10^9$ antigens [26, 89]. The naive repertoire becomes further specialized by evolution in response to exposure to antigens. This process, known as affinity maturation, causes naive B cells that have been exposed to their particular antigen to undergo clonal selection, leading to lineages of evolved B-cell repertoires against specific threats. In the case of a rapidly-evolving pathogen like HIV, this evolution proceeds in a complex series of interactions between the evolving B-cell repertoire and the evolving HIV population.

The host immune system produces antibodies that bind multiple HIV proteins; however, Env is the only target for *neutralizing* antibodies. When targeted by neutralizing antibodies, the viral population quickly evolves escape mutations. This arms race between the immune system and the viral population is one of the main drivers of the explosion of genetic diversity after the initial infection, which usually begins with a single founder sequence [58]. 10-25% of people with HIV develop broadly-neutralizing antibodies

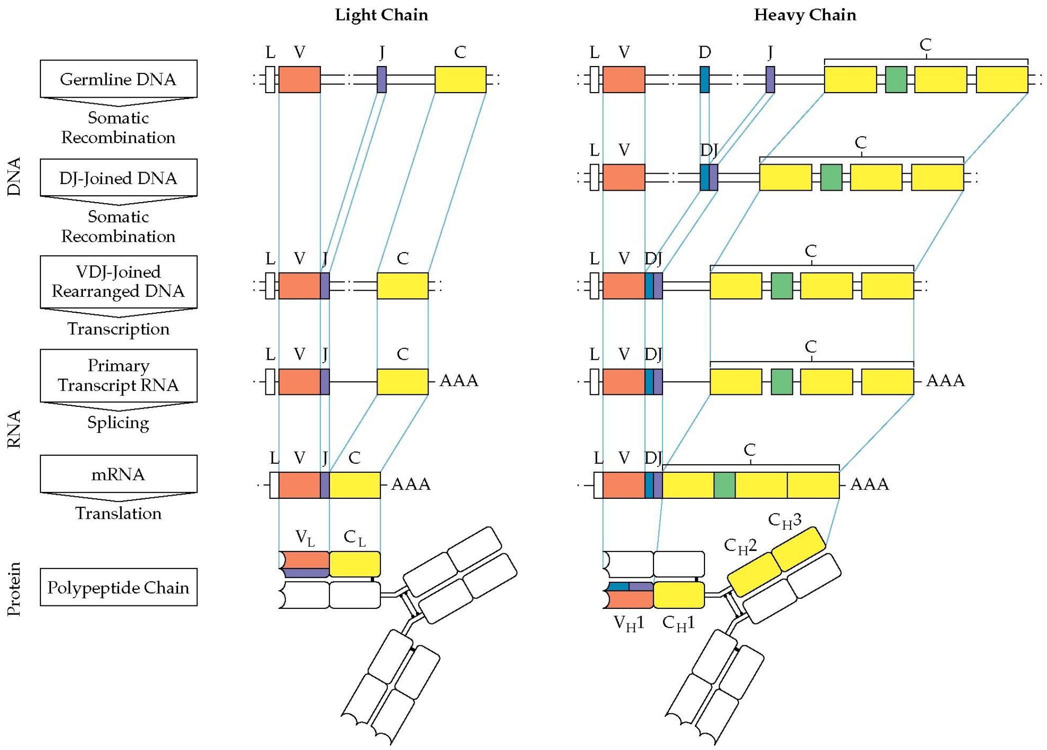


Figure 1.8: Somatic recombination in antibody heavy and light chains. From http://what-when-how.com/acp-medicine/adaptive-immunity-antigens-antibodies-and-t-cell-and-b-cell-receptors-part-1/.

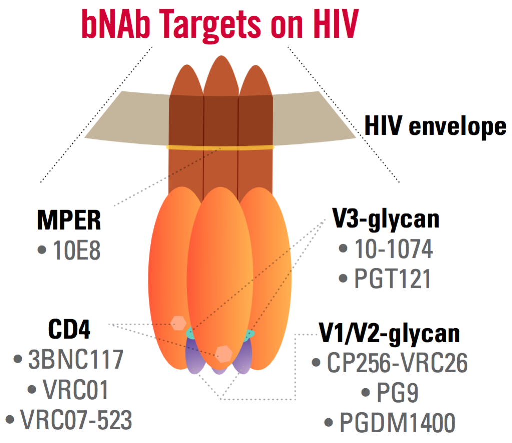


Figure 1.9: The main bNAb targets in Env. Broadly-neutralizing antibodies tend to target one of these five regions in Env. These epitopes may represent weaknesses in Env from an evolutionary standpoint that could be exploited [152]. From "HIV Vaccine Research: An update".

(bNAbs), which are able to bind and neutralize a large number of HIV strains [49]. This ability makes bNAbs an attractive target of research, since they may provide they key to vaccine design, as well as antibody-mediated therapies.

Env possesses many characteristics that make it a difficult target for antibodies, which is why few individuals develop bNAbs, and why the immune response is often inadequate to prevent the progression to AIDS. HIV's high mutation rate and fast generation time allow it to quickly evolve escape mutations, if they were not already present at low frequencies in the highly diverse population. Cleavage of gp160 into gp120 and gp41 is inefficient, causing antibodies to target the uncleaved, inactive form. Env is conformationally flexible and heavily shielded by glycans until fusion begins, at which point proximity of the host cell inhibits binding. Despite these difficulties, bNAbs do develop. Five main classes of bNAbs have been discovered, which differ in their target, as shown in Figure 1.9: the V3-high mannose patch, the V2 apex, CD4 binding site (CD4bs), the gp41 membrane proximal external region (MPER), and the gp120/gp41 interface, including the fusion peptide [65, 151].

One promising approach to developing a vaccine would be to induce the development of bNAbs in HIV-negative individuals. However, there are currently many obstacles that need to be overcome before this approach becomes clinically viable [38, 62, 81, 83, 88]. The conditions that lead to the development of bNAbs are not generally understood. Naive antibodies bind weakly to HIV, and most evolved antibody lineages that appear soon after infection are narrowly neutralizing, which means they do not neutralize a broad number of strains in a neutralization panel. In contrast, bNAbs do not appear in most cases until after years of infection, suggesting that a complex series of specific interactions with the evolving Env population are necessary to develop bNAbs. These lineages may also depend on features of the early variants of Env [61] and on the genetic background of the host. Finally, even if bNAbs can be successfully induced, they will need to be maintained at protective levels.

Deep sequencing of B cell repertoires has become a key tool for understanding the immune response [9, 39, 46, 48, 155, 158], and this approach is currently being used to understand the process of bNAb development. However, the evolution of B cell lineages cannot be fully understood except in the context of the Env population that co-evolved with them. Therefore, there is a similar need for longitudinal deep sequencing of Env populations. Most sequencing methods are insufficient for this task, because they either lack depth or they use short reads that lose long-range linkage information that could be crucial for understanding the evolution and escape. These shortcomings were recently addressed by deep, full-length sequencing of *env* with the SMRT sequencing protocol.

1.2 Full-length SMRT sequencing of *env*

Pacific Bioscience's single-molecule realtime (SMRT) sequencing is a third-generation sequencing protocol. Single-molecule methods like SMRT have a number of

Figure 1.10: The Pacific Biosciences RS II sequencer. From pacb.com.

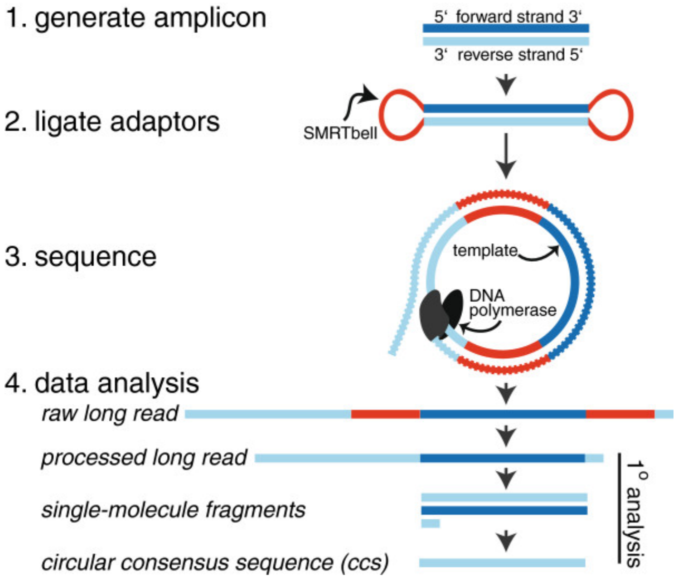


Figure 1.11: How SMRT sequencing works. From [35].

specific advantages over first- and second-generation sequencing technologies, including speed and read length [121, 124]. This platform is especially useful for viral sequencing [113].

SMRT sequencing can provides high-quality reads with sufficient length and depth to study HIV-1 *env* populations. The details of the effective sequencing depth and of the single-molecule accuracy distribution depends primarily on the amplicon length, as described below. For a comparison with other sequencing technologies, see [42]. These qualities make it the best choice currently for sequencing highly diverse amplicon populations such as HIV-1 *env*. The work described in this dissertation was done with the RS II (1.10); these benefits are presumably compounded in the newer Sequel Systems, which produce about seven times more reads than the RS II.

The entire sequencing process takes place on a silicon chip (SMRTcell) that contains a large number of microscopic holes called zero-mode waveguides (ZMWs). Each ZMW contains a single DNA polymerase, allowing a number of molecules to be sequenced in parallel. SMRTbell adapters are ligated onto the amplicon to form a circular SMRTbell molecule, which runs through the DNA polymerase repeatedly. Nucleotide-specific fluorescent tags are observed by a sensor and the fluorescence signal is used for basecalling. The entire process produces a ZMW read, which contains multiple copies of the forward and reverse amplicon sequences, interleaved with the adapter sequences. The individual copies are then combined into a higher-quality consensus sequence called a circular consensus sequence (CCS). The process is illustrated in Figure 1.11.

SMRT sequencing has a reputation for producing error-prone reads, but the error rates usually cited (typically around 12-15% [111, 113]) refer to the raw error rate of the ZMW read. Combining the forward and reverse passes into a single CCS drops the error rate considerably. The final accuracy of the CCS depends on the length of the amplicon and the length of the read. The shorter the amplicon, or the longer the read, the more

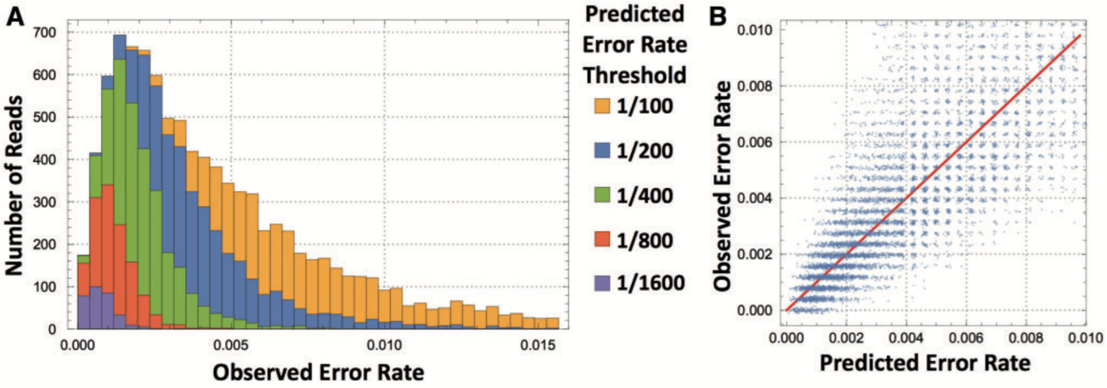


Figure 1.12: SMRT errors rates on NL4-3. (A) Distribution of error rates of full-length *env* CCS sequences, for different predicted error rate thresholds. (B) Observed error rate versus predicted error rate, showing that quality scores are well correlated with the true error rates, and therefore useful for filtering. Image from [63].

passes get combined into the final CCS, and the higher the quality. Errors tend to be mostly insertions and deletions, especially in homopolymer regions.

My collaborators at UCSD and Pacific Biosciences recently developed a SMRT sequencing protocol for full-length *env* [63]. It is now possible to sequence an entire population from a single host at multiple time points to infer how the population is evolving, for instance, in response to bNAbs or therapies that target Env. For full-length *env* sequencing, the amplicon is about 2.6kb, yielding a per-read error distribution shown in Figure 1.12 for the P5/C3 chemistry. Multiple tests confirm the accuracy of the results, including comparisons with a known NL4-3 template. Moreover, phylogenies obtained from running our newly-developed pipeline on SMRT sequences agree with Sanger sequencing, as shown in Figure 1.13.

The sequencing protocol was published in [63], which also contained an early version of the computational pipeline presented in this dissertation. The method was validated on NL4-3, as well as donors P018, K453, and H497, and results analyzed with early version of `FLEA`. We re-analyzed P018 with the current version of `FLEA`, which is described in the next chapter.

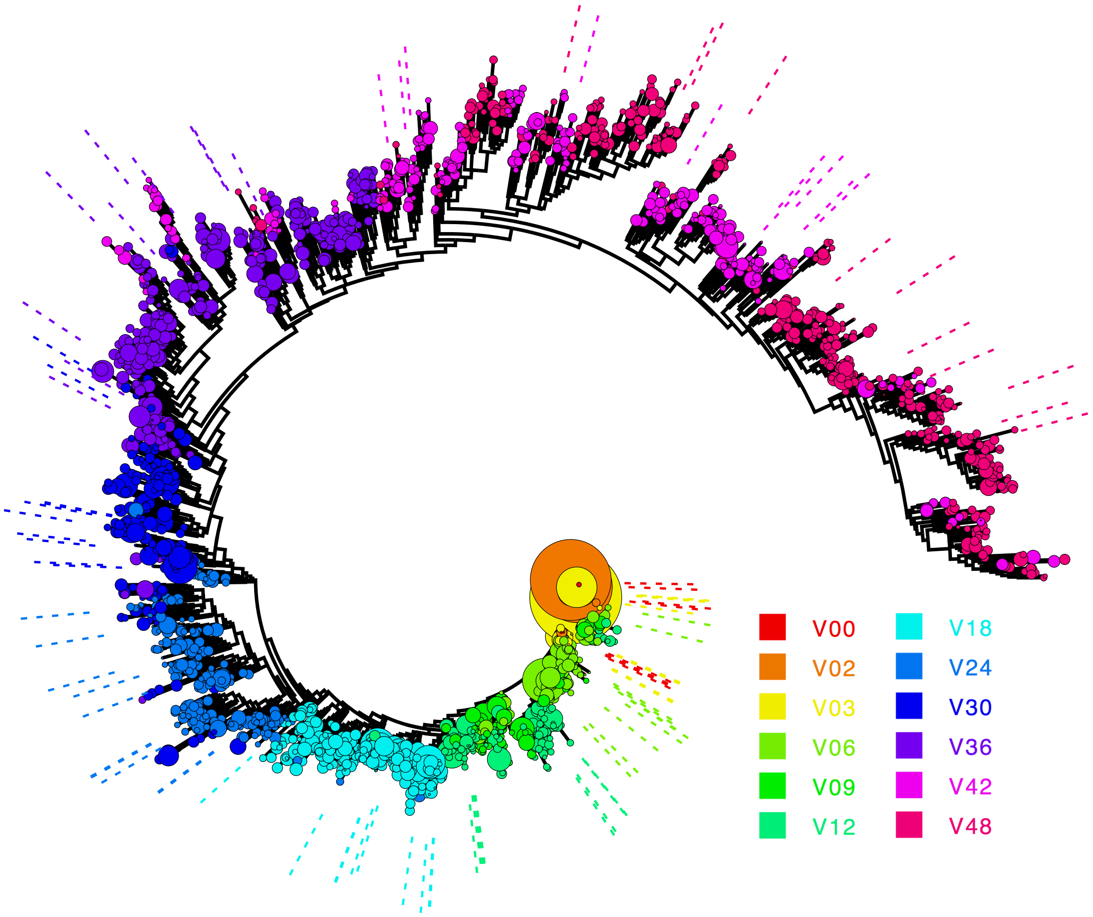


Figure 1.13: Phylogenetic tree of Env from donor PC064. Nodes are colored according to the longitudinal time point and scaled according to inferred abundance. Sequencing was done with both the SMRT sequencing protocol and Sanger sequencing, to show agreement between the two methods. Dotted lines show the locations of Sanger sequences. Image from [63].

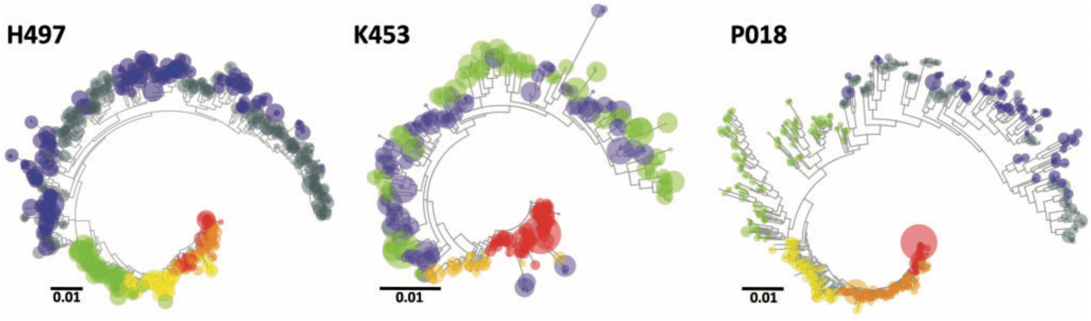


Figure 1.14: Phylogenetic trees for Env from three donors from the San Diego HIV-1 Primary Infection Research Consortium (SD-PIRC). Sequenced with SMRT sequencing and analyzed with the `FLEA` pipeline. Nodes are colored according to the longitudinal time point and scaled according to inferred abundance. All three infections show increased diversity as time goes on. Image from [63].

The sequencing depth and accuracy of this protocol allow researchers to resolve complex population structures with many closely-related minority variants and to perform phylogenetic analysis of entire viral populations in a single host. Once the sequencing technology enabled these experiments, significant computational work was required to develop the tools to analyze the data coming off the sequencer. Those computational tools and methods are the subject of this dissertation.

1.3 The rest of this dissertation

Thanks to the full-length sequencing protocol, it is now possible to obtain longitudinal full-length *env* sequences. That data can be used to aid in rational vaccine design, tracking the effect of monoclonal antibody therapies on the population, and many other kinds of population-level work, but processing the sheer number of reads it provides is a challenge. Each sequencing run is capable of producing around 50,000 reads per time point. Filtering at a 1% expected error rate and appropriate length yields about 10,000 sequences for the P5/C3 chemistry, and 15,000 for the P6/C4 chemistry. Therefore, a longitudinal experiment can easily produce over hundreds of thousands of sequences.

Moreover, the data still contains errors – for instance, even an error threshold of 1% for a 2.6 kb amplicon still yields an expected 26 errors per sequence – so even if it were possible to input all the CCSs directly into a suite of analyses, the results would be biased. For pathogens with low indel variablity, such as HCV, a common strategy is to pairwise align each read to a reference, stack them into a multiple sequence alignment, then perform further analyses[1] [146]. However, HIV *env* populations contain a large amount of indel variation, so this strategy would discard all insertions relative to the reference and break in highly variable regions.

The subject of this dissertation is a set of computational methods to solve these problems for longitudinal full-length *env* sequences. Chapter 2 describes `FLEA`, a pipeline for analysis and visualization of a large number of erroneous sequences with indel variation. Chapter 3 describes `RIFRAF`, a sequence consensus algorithm that uses quality scores and an in-frame reference sequence to infer highly-accurate consensus sequences. Chapter 4 gives an overview of the results acquired so far with `FLEA`, which has been used on samples from donors that developed lineages of broadly-neutralizing antibodies, and on participants in a phase I clinical trial of a monoclonal antibody therapy. Chapter 5 summarizes my work and proposes future directions for this research.

1.4 Acknowledgements

Chapter 1 contains, in part, material that appeared in “Rapid sequencing of complete env genes from primary HIV-1 samples”. Melissa Laird Smith, Ben Murrell, Caroline Ignacio, Elise Landais, Steven Weaver, Pham Phung, Colleen Ludka, Lance Hepler, Gemma Caballero, Tristan Pollner, Yan Guo, Douglas Richman, IAVI Protocol C Investigators & The IAVI African HIV Research Network, Pascal Poignard, Ellen

[1] https://github.com/veg/HIV-NGS

E Paxinos, Sergei L Kosakovsky Pond, Davey M Smith,*Virus Evolution* 2016. The dissertation author was an author of this paper.

Chapter 2

FLEA

Abstract

Next generation sequencing of viral populations has advanced our understanding of viral population dynamics, the development of drug resistance, and escape from host immune responses. Many applications require complete gene sequences, which can be impossible to reconstruct from short reads. HIV-1 *env*, the protein of interest for HIV vaccine studies, is exceptionally challenging for long-read sequencing and analysis due to its length, high substitution rate, and extensive indel variation. While long-read sequencing is attractive in this setting, the analysis of such data is not well handled by existing methods. To address this, we introduce `FLEA` (Full-Length Envelope Analyzer), which performs end-to-end analysis and visualization of long-read sequencing data.

`FLEA` consists of both a pipeline (optionally run on a high-performance cluster), and a client-side web application that provides interactive results. The pipeline transforms FASTQ reads into high-quality consensus sequences (HQCSs) and uses them to build a codon-aware multiple sequence alignment. The resulting alignment is then used to infer phylogenies, selection pressure, and evolutionary dynamics. The web application

provides publication-quality plots and interactive visualizations, including an annotated viral alignment browser, time series plots of evolutionary dynamics, visualizations of gene-wide selective pressures (such as dN/dS) across time and across protein structure, and a phylogenetic tree browser.

We demonstrate how `FLEA` may be used to process Pacific Biosciences HIV-1 *env* data and describe recent examples of its use. Simulations show how FLEA dramatically reduces the error rate of this sequencing platform, providing an accurate portrait of complex and variable HIV-1 *env* populations.

A public instance of `FLEA` is hosted at http://flea.datamonkey.org. The Python source code for the `FLEA` pipeline can be found at https://github.com/veg/flea-pipeline. The client-side application is available at https://github.com/veg/flea-web-app. A live demo of the P018 results can be found at http://flea.murrell.group/view/P018.

2.1 Introduction

Next generation sequencing (NGS) has become an invaluable tool for studying HIV-1 and other rapidly evolving viruses by providing direct high resolution measurements of viral genetic diversity within the host. NGS has been used to study immune escape [24, 36, 47, 70, 82, 102, 141], drug resistance [40, 50, 51, 69, 91, 137, 141], transmission bottlenecks [12, 69, 144, 147], population structure and dynamics [36, 41, 47, 59, 76, 108, 133, 157, 159], tropism dynamics [126], and multiplicity of infection [100]. It is also used in clinical virology [13, 113]. For reviews of the promises and challenges of NGS applications in virology, see [71], [145], [85], and [5].

Full-length sequences can resolve features that are difficult to assemble from short sequences [50, 116]. For instance, Pacific Biosciences SMRT sequences were able to resolve 1.5 kb *msg* isoforms from *Pneumocystis jirovecii*, but reads from a 454 instrument could not be assembled correctly [116]. For tracking evolutionary patterns in viral populations, accurately resolving these features provides a more accurate history of the population, which becomes especially relevant when epistatic interactions and linkage between mutations effect phenotypic changes in the pathogen [43, 103, 150]. For example, studies of HIV-1 *env* frequently use functional assays to measure the potency with which a given antibody or donor serum neutralizes a specific *env* strain [123], which requires knowing the full *env* sequence.

We have developed a pipeline for handling long read HIV-1 *env* sequencing data from within-host viral populations: the Full-Length Envelope Analyzer (FLEA). FLEA addresses the specific challenges posed by large volumes of such data, e.g., using the sequencing protocols we previously described in Laird *et al* [63], which also contains an overview of a prototype of FLEA. Here we describe the full pipeline and experimentally demonstrate its ability to resolve populations of closely related variants. FLEA uses state-of-the-art tools and methods at every step and can be accessed through a web browser or on a high-performance cluster. FLEA is readily extensible to other genes and systems.

FLEA has recently been used by the authors in two high-profile studies. In [14], we describe how FLEA was used to process PacBio HIV-1 *env* data from a clinical trial of monoclonal antibody 10-1074. For sequences sampled before and after therapy, FLEA reveals that prior to antibody therapy low-frequency *env* variants were present with mutations that typically confer resistance to 10-1074. Additionally, when resistance emerges, it emerges multiple times, exploiting many different resistance pathways. FLEA was also used to characterize the longitudinal *env* population that drove development of a broadly neutralizing antibodies against the apex of the *env* trimer, sampled from donor PC64 from the Protocol C primary infection cohort [65].

There exist dozens of standalone pipelines developed for analyzing HIV-1 and related sequence data, including longitudinal samples [40, 51, 70, 72]. However, it was necessary to develop a new tool due to HIV-1 *env*'s extensive natural indel variation and the high rate of indels in long PacBio reads, which are especially problematic when any spurious indel in the 2.6kb *env* amplicon corrupts the reading frame, rendering the sequence uninterpretable. With HIV-1 *env*, the common strategy of mapping reads to a reference fails because the diversity in variable regions of *env*, predominantly driven by extensive indel processes, means that these regions in sampled reads lack homology to those in any heterologous reference sequence. Instead, `FLEA` relies on a fine-grained cluster-and-consensus strategy to remove spurious indels from reads. The task is related to Liang *et al.* (2016), but, rather than distinguishing a small number of variants at 81-91% identity, we must distinguish potentially hundreds of variants that differ by only a handful of bases.

In addition to a standalone application, `FLEA` is also available as an online resource that provides interactive visualizations for all its analyses. To allow researchers to further examine and dissect their results, `FLEA` also provides access to raw data, such as aligned consensus sequences and phylogenetic trees.

2.2 Design and Implementation

2.2.1 Pipeline

The input to `FLEA` is a set of FASTQ files from the PacBio RS-II or Sequel. Each set corresponds to one time point, containing circular consensus sequence (CCS) reads, which can be obtained using the "Reads of Insert“ protocol on PacBio's SMRTportal or SMRTanalysis tools. Upon completion, the `FLEA` pipeline produces results as JSON (Javascript Object Notation) files, a standard format for machine (and human-) readable structured data. The logic of `FLEA` is implemented in `Nextflow` [27], a workflow framework for deploying parallel pipelines to clusters and clouds.

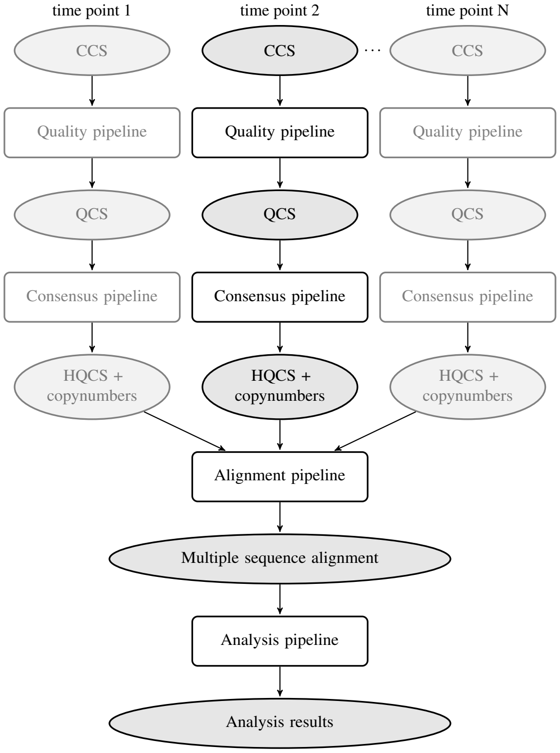


Figure 2.1: Overview of the `FLEA` pipeline, broken into conceptual sub-pipelines. The *Quality* and *Consensus* sub-pipelines process each time point separately. Duplicate steps in other time points are grayed out. CCS stands for “circular consensus sequences”; QCS for “quality-controlled sequences”, and HQCS for “high-quality consensus sequences”.

`FLEA` consists of multiple sub-pipelines, as shown in Fig. 2.1. Details of the quality and consensus pipelines are depicted in Fig. 2.2. Together, these two pipelines take error-prone CCS reads and convert them into unique high-quality consensus sequences. The alignment pipeline generates a multiple sequence alignment, which is used by multiple methods in the analysis pipeline.

2.2.2 Quality assurance sub-pipeline

The first steps remove low quality reads and filter out common sequencing artifacts. Parameters given in these steps were chosen for full-length HIV-1 envelope sequences from the RS-II or Sequel platforms. Other reads with different properties (error

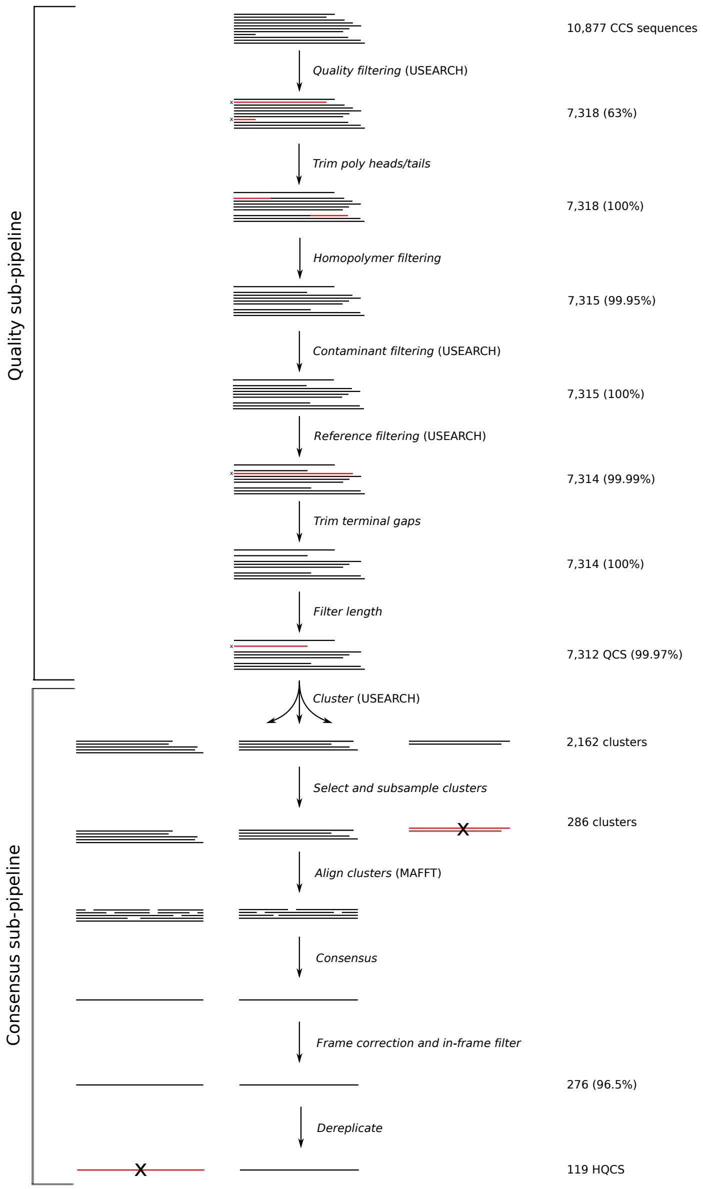


Figure 2.2: Quality and consensus sub-pipelines. These steps are repeated independently on each time point. Numbers are reported from the analysis of sequences from the first time point (V03) of donor P018, which is three months post infection. Percentages give the fraction of sequences retained after filtering. Tasks indicate whether they use third-party tools `USEARCH` or `MAFFT`.

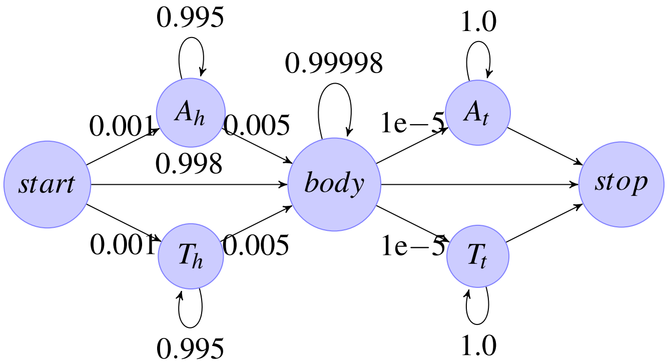


Figure 2.3: Hidden Markov model used for trimming poly-A and poly-T heads and tails. A head and tail states have a small ($p = 0.01$) probability to emit non-A bases, and similarly for T. The *body* state emits all four bases with equal probability. The *start*, and *stop* states emit nothing.

rates, error models, lengths, homopolymer distributions, etc.) likely require different parameters. All steps are run independently per time point.

1. **Filter by error rate.** The input FASTQ files contain Phred scores for each base, encoding the probabilities of incorrect base calls. `USEARCH` [30] is used to remove reads with an expected error rate greater than 1%, computed as the mean of the per-base error probabilities.

2. **Trim heads/tails.** A fraction of reads from the Laird *et al.* sequencing protocol contain poly-A or poly-T heads or tails (cause unknown), which can be hundreds of bases long and sometimes contain a small number of other bases.

   These heads and tails are trimmed with a hidden Markov model (Fig. 2.3) implemented in Pomegranate [53]. The emission probabilities of the model were fixed, and the transitions trained using Baum-Welch. The Viterbi path for each sequence is computed, and bases emitted by head and tail nodes are removed.

3. **Filter long runs.** Reads with homonucleotide runs longer than 16 bases are discarded. This length was chosen to be twice the length of the longest such run in the LANL HIV database [37].

4. **Filter contaminants and trim reads.** Sample contamination can introduce non-native sequences that interfere with subsequent analyses, so these contaminants must identified and discarded. `USEARCH` is used to compare reads to a contaminant database and a reference database using `usearch_global`. Alignments returned from querying the database are then used to trim reads to the gene boundaries. Trimming terminal insertions is vital for the accuracy of downstream tasks, such as length filtering and clustering.

   The contaminant database contains HXB2 and NL4-3 *env*, each ubiquitous in labs working with *env* sequences and a common source of sample contamination. Reads that match with $\geq 98\%$ identity are discarded. Since a 1% error rate cutoff was earlier used, this parameter conservatively ensures that these contaminants are almost certainly identified.

   The reference database contains thirty-eight sequences representing the major HIV-1 Group M subtypes from the LANL sequence database [37]. Reads with $\leq 70\%$ identity to every sequence in the reference database are discarded. This cutoff is chosen to retain reads remotely similar to HIV-1 Group M while excluding contaminants such as human or bacterial genome reads. If a sample is from SIV, or from a non group-M HIV+ donor, then more appropriate reference sequences should be added to the database.

5. **Filter by length.** By default, sequences shorter than 90% or longer than 110% of the length of the reference sequence are discarded. However, sequences with large deletions are frequently observed in HIV. These likely represent replication incompetent envelopes, and their reduced length can cause them to be dramatically oversampled due to PCR length bias. Users who want to include these species in their analyses should modify these parameters.

Reads that pass this quality assurance phase have low expected error rates and no homonucleotide runs, are within 70% identity of at least one reference sequence, are (after trimming) no more than 10% different in length than a reference sequence, and do not match the contaminant database. We refer to these sequences as quality-controlled sequences (QCS).

Consensus sub-pipeline for variant identification

Even for highly diverse populations, unique reads in a sequencing run outnumber the true unique variants, predominantly due to sequencing errors. The problem is far more significant in long reads than in short reads, precluding the use of amplicon denoising strategies used to reduce error rates in short read sequencing [31]. To accommodate this effect, the next phase of the `FLEA` pipeline clusters and combines QCS reads, attempting to infer the true variants in each time point. It also attempts to detect and correct frameshift errors.

1. **Cluster.** `USEARCH` is used with the `cluster_fast` command to generate clusters with 99% nucleotide identity. This parameter approximates the 1% error cutoff used in the error rate filtering step, so that pairwise distances of sequences in the same cluster are consistent with the sequencing error. `cluster_fast` runs in a single pass, so it is sensitive to input order. Sequences are sorted from lowest to highest quality according to expected error rate; experiment suggests that this order yields better results (see supporting information).

2. **Select and subsample clusters.** Clusters with fewer than three members are discarded, because they are too small to de-noise by majority consensus. Clusters with more than 50 members are subsampled to the top 50 with the lowest expected error rate to speed up the multiple sequence alignment step.

3. **Align and consensus.** `MAFFT` [57] is used to align each cluster. The consensus sequence of each alignment is computed.

4. **Frame correction** In-frame consensus sequences from all time points are collected into a `USEARCH` database for frame correction. `usearch_global` is then used to align each out-of-frame sequence to its top hit. The nucleotide alignment is used to correct incomplete codons: short insertions (1 or 2 base pairs) are discarded, and single deletions are replaced with the aligned base. Sequences with longer insertions or deletions are discarded. All changes are logged, so that the user can identify the sequences that have been corrected.

5. **Uniqueness** Non-unique consensus sequences are dereplicated using `usearch --fastx_uniques`.

6. **Copy numbers** The number of sequences per cluster provides an estimate of the relative abundance of that HQCS in the population. Those numbers are further augmented by adding sequences orphaned by cluster filtering and HQCS dereplication. `usearch_global` is used to assign each QCS to its nearest HQCS. The number of sequences accrued by each HQCS is interpreted as its copy number.

All of these tasks are run separately for each time point, yielding sets of unique in-frame consensus sequences. We refer to these sequences as high-quality consensus sequences (HQCS).

Alignment sub-pipeline

The HQCSs from all time points are combined into a single file, translated to protein sequences, and aligned using `MAFFT`. A Python script then transfers the gaps from each aligned protein sequence to the corresponding nucleotide sequences to produce a codon-level nucleotide multiple sequence alignment of all unique variants from all time points.

Analysis sub-pipeline

The analyses used in FLEA take as input the two outputs of the alignment phase: a codon multiple sequence alignment of all unique HQCS sequences from all time points, and their associated copy numbers. These data are used for the following analyses.

1. **Time point metrics.** HyPhy [107] scripts are used to compute evolutionary metrics (total, *dN*, and *dS* divergence and diversity) and phenotypic metrics (protein length, potential N-linked glycosylation sites, isoelectric point) for each annotated region (e.g., V1, MPER) in the amplicon for each time point.

2. **MRCA.** The most recent common ancestor is inferred by taking the copy-number-weighted codon consensus of the codon-aligned HQCSs from the earliest time point. By including gaps, the MRCA sequence is already aligned with the rest of the multiple sequence alignment. This strategy is acceptable for primary infection studies from single founders with very low early diversity.

3. **Reference coordinates.** MAFFT is used to assign HXB2 [115] coordinates to the gapped MRCA sequence, which are then transferred to the full multiple sequence alignment.

4. **Infer phylogeny.** A maximum-likelihood phylogenetic tree is inferred with FastTree2 [109, 110] under the general time reversible model.

5. **Ancestral sequence reconstruction.** HyPhy is used to infer ancestral sequences at the internal nodes of the phylogeny, using joint maximum likelihood reconstruction and the HKY85 substitution model [45].

6. **Multidimensional scaling.** TN93 [127] is used to compute a distance matrix for all HCQC sequences using the Tamura Nei 93 distance [138]. Metric multidimensional scaling [140] (implemented in `scikit-learn` [21]) is used to find a two-dimensional embedding of the sequences that approximates their pairwise distances.

7. **FUBAR.** Site-specific selection rates are inferred using `FUBAR` [93], implemented in `HyPhy`.

8. **Position-specific changes.** Entropy and Jensen-Shannon divergence are computed for each position in each time point.

The results of these analyses are provided to the user in an interactive web application, described next.

2.2.3 Web application

The `FLEA` web app is built using modern web design principles. It consists of two parts: a Javascript client-side app, written using the `Ember.js` [33] framework, and a server-side REST (REpresentational State Transfer) service for serving JSON-formatted data. There are two main benefits to using this decoupled pattern for scientific web applications. First, the client-side code only needs to be downloaded once, at the start of the session. The data are requested from the server and cached as needed. Once everything is loaded, the visualizations run entirely in the browser with no delays for page loads. Second, the REST service may be reused by other apps and third-party tools.

The web app presents the results of the `FLEA` analysis as a series of interactive visualizations. The report is organized into the following sections.

Multidimensional scaling. A two dimensional embedding of the HQCSs is visualized as a bubble plot, showing changes in population structure over time, as shown in Fig. 2.4. This visualization has been especially useful for investigating populations with super-infection, or with multiple founders, where aggressive recombination between vastly different *env* variants precludes the use of phylogenies.

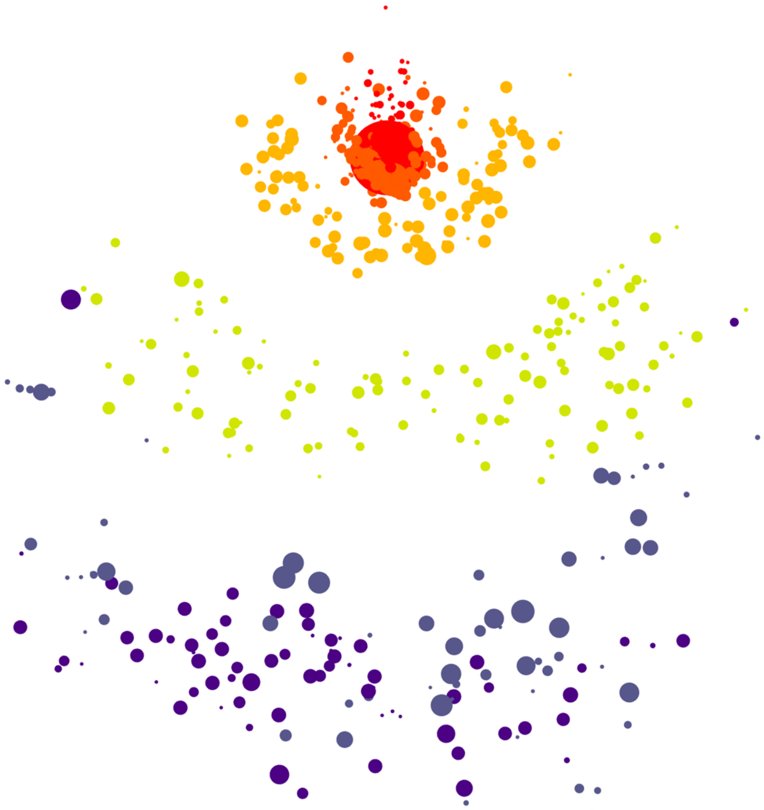
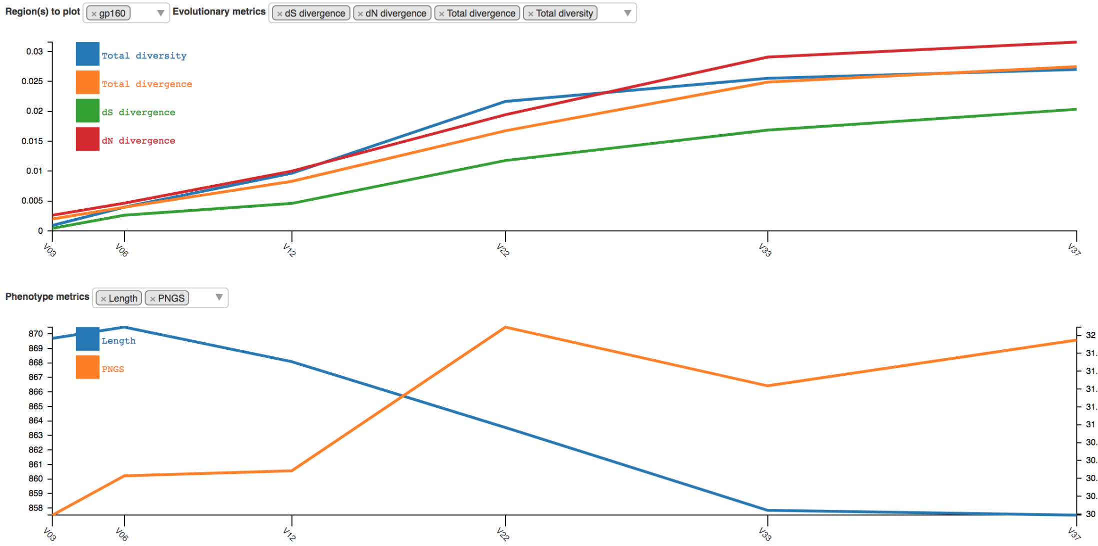


Figure 2.4: Screenshot of the multidimensional scaling plot. The embedding in two dimensions preserves pairwise evolutionary distances between HQCSs. Node area is proportional to copy number, and color corresponds to time point. The increasing genetic diversity of the population is visible as time goes on.

Evolutionary trajectory. The evolutionary trajectory viewer plots evolutionary and phenotypic metrics for each time point and multiple regions in the amplicon, giving a high-level overview of population dynamics over time. Fig. 2.5 shows the plot for the



Region(s) to plot × gp160

Evolutionary metrics × dS divergence × dN divergence × Total divergence × Total diversity

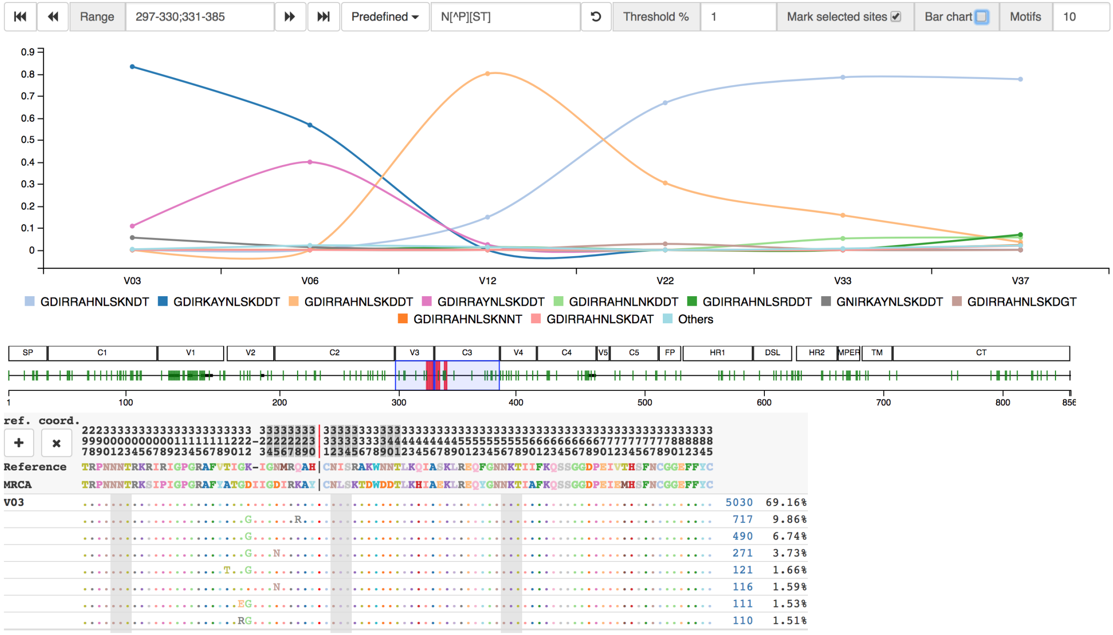
Phenotype metrics × Length × PNGS

Figure 2.5: Screenshot of the evolutionary trajectory report. Four evolutionary metrics (*dS* divergence, *dN* divergence, total divergence, and total diversity) and two phenotype metrics (length and possible N-linked glycosylation sites) are shown for gp160.

entire gp160 region of HIV-1 Env, which is generated with the `D3.js` plotting library [87].

Sequences. The multiple sequence alignment of all the HQCSs sequences is the foundation for all subsequent analyses. It is displayed in the amino acid sequences viewer, which contains a custom alignment browser and an interactive motif dynamics plot, as shown in Fig. 2.6.

Protein structure. The protein structure viewer maps evolutionary metrics to an interactive three-dimensional structure of the protein, customized from PDB ID `5FUU`, a recently resolved cryo-EM structure [68], and rendered using `pv` [77]. Missing residues are rendered as spheres which are positioned by Bézier curve interpolation. dN/dS ratios, Jensen Shannon divergence, and entropy may all be mapped to the protein structure, as shown in Fig. 2.7. The same metrics are also plotted in one dimension for each time point, as shown in Fig. 2.8. The protein visualization interacts with the sequence viewer



⏮ ⏪ Range 297-330;331-385 ⏩ ⏭ Predefined ▾ N[^P][ST] ↺ Threshold % 1 Mark selected sites ☑ Bar chart □ Motifs 10

🟦 GDIRRAHNLSKNDT 🟦 GDIRKAYNLSKDDT 🟧 GDIRRAHNLSKDDT 🟪 GDIRRAYNLSKDDT 🟩 GDIRRAHNLNKDDT 🟩 GDIRRAHNLSRDDT ⬛ GNIRKAYNLSKDDT 🟫 GDIRRAHNLSKDGT 🟧 GDIRRAHNLSKNNT 🟥 GDIRRAHNLSKDAT 🟦 Others

```
ref. coord.
          222333333333333333333333 33333333|3333333333333333333333333333333333333333333333333333333
 +   ×    99900000000000111111111222-22222223|3333333334444444444555555555566666666667777777777888888
          789012345678923456789012 34567890|1234567890123456789012345678901234567890123456789012345
Reference TRPNNNTRKRIRIGPGRAFVTIGK-IGNMRQAH|CNISRAKWNNTLKQIASKLREQFGNNKTIIFKQSSGGDPEIVTHSFNCGGEFFYC
MRCA      TRPNNNTRKSIPIGPGRAFYATGDIIGDIRKAY|CNLSKTDWDDTLKHIAEKLREQYGNNKTIAFKQSSGGDPEIEMHSFNCGGEFFYC
V03       .................................|.......................................................  5030  69.16%
          .......................G......R..|.......................................................   717   9.86%
          .......................G.........|.......................................................   490   6.74%
          .......................G...N.....|.......................................................   271   3.73%
          ....................T..G.........|.......................................................   121   1.66%
          ...........................N.....|.......................................................   116   1.59%
          ......................EG.........|.......................................................   111   1.53%
          ......................RG.........|.......................................................   110   1.51%
```

Figure 2.6: Screenshot of amino acid sequences viewer. Sequences are grouped by identity, with aggregate copy number and population percentage shown to the right. An overview of the amplicon, optionally annotated with region names, provides fast access to different locations of the alignment. Selecting columns of the alignment interactively updates the amino acid dynamics plot, showing the dynamics of the selected motif over time. In this case, the trajectory shows changes in the N332 glycan supersite. Sites inferred by FUBAR to be undergoing positive selection are selectable.

by showing alignment positions and highlighting the residues in the selected sequence motif.

Trees. The tree viewer renders a tree browser with `phylotree.js` [128], as shown in Fig. 2.9. Leaf nodes are scaled to the copy number of their sequence. The tree zoom level, layout, and coloring is interactively modifiable. Motifs selected in the sequence viewer are mapped to the tree. Ancestral nodes are colored by motif, allowing inferred changes to be tracked through the entire phylogeny.

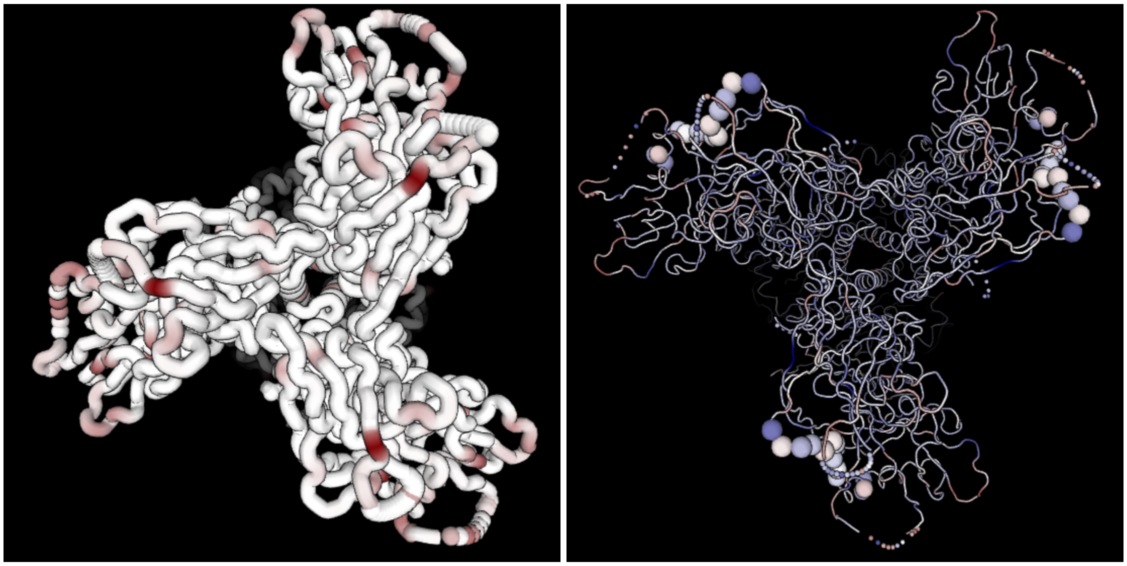


Figure 2.7: Screenshots of the interactive three-dimensional Env structure, colored according to JS divergence (left) and dN/dS values (right). Positions imputed to be undergoing more positive selection ($dN/dS > 1$) are darker red, and positions undergoing more purifying selection ($dN/dS < 1$) are darker blue. The right structure also shows motif positions highlighted in the sequence viewer.

2.3 Results

The entire pipeline was run on HIV-1 *env* reads from donor P018, which are available from the NCBI Sequence Read Archive under BioProject PRJNA320111, and were sequenced as part of [63] on the RS-II instrument, using the older generation P5/C3 PacBio sequencing chemistry. The full dataset contains 58,468 CCS reads. The reads are split across six time points, which are coded as V03, V06, V12, V22, V33, and V37, where Vx corresponds to a visit x months post infection. The number of reads per time point ranges from 7,530 in V33 to 11,806 in V06.

2.3.1 Results on simulated data

The true sequences and copy numbers are not known for the P018 data. In order to assess the accuracy of our inferred sequence population, we used the HQCSs from

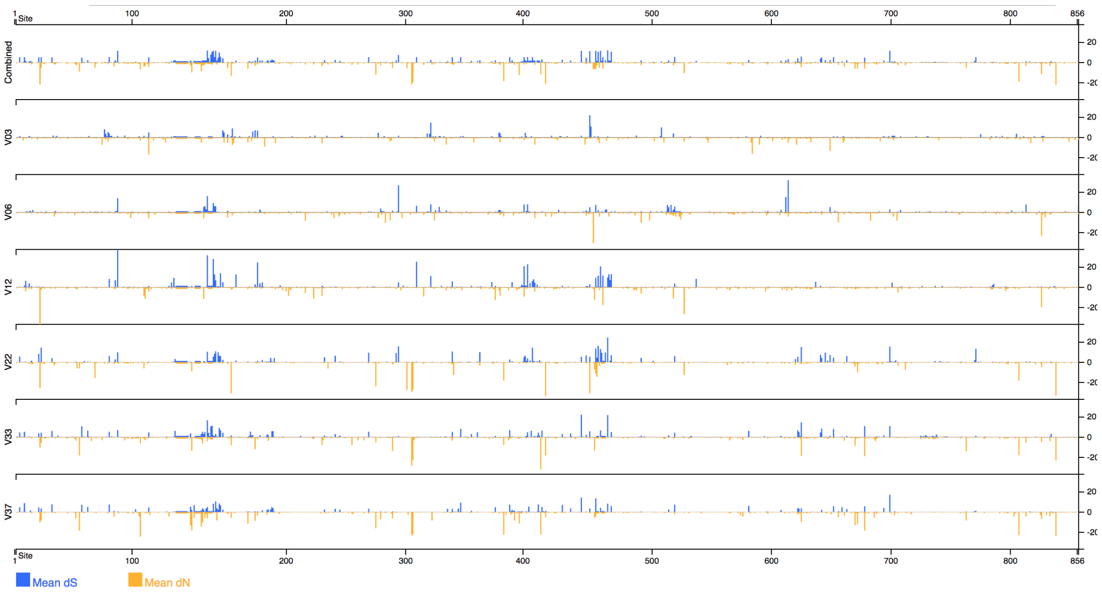


Figure 2.8: Screenshot of dN/dS values mapped to protein positions and separated by time point.

a previous `FLEA` run to simulate a gold standard dataset on which to assess the `FLEA` pipeline.

The simulation procedure starts with the HQCSs and copy numbers from the `FLEA` results on P018, then augments them with additional mutated sequences to create a gold standard set of templates. Mutated sequences were added because our clustering strategy may artificially merge similar templates. For each template, noisy reads with a SMRT-style error profile were sampled. Full details of the simulation process appear in the supporting information. These simulated reads were sent through the `FLEA` pipeline, both with and without frame correction.

The resulting QCS and HQCS sequences were compared to the ground truth using Earth Mover's Distance (EMD), using normalized copy numbers for the population weights and edit distance for the distance matrix. The fully constrained EMD has units that can be directly interpreted as the average change per nucleotide necessary to transform one sequence population into another. We also calculate two variants of EMD

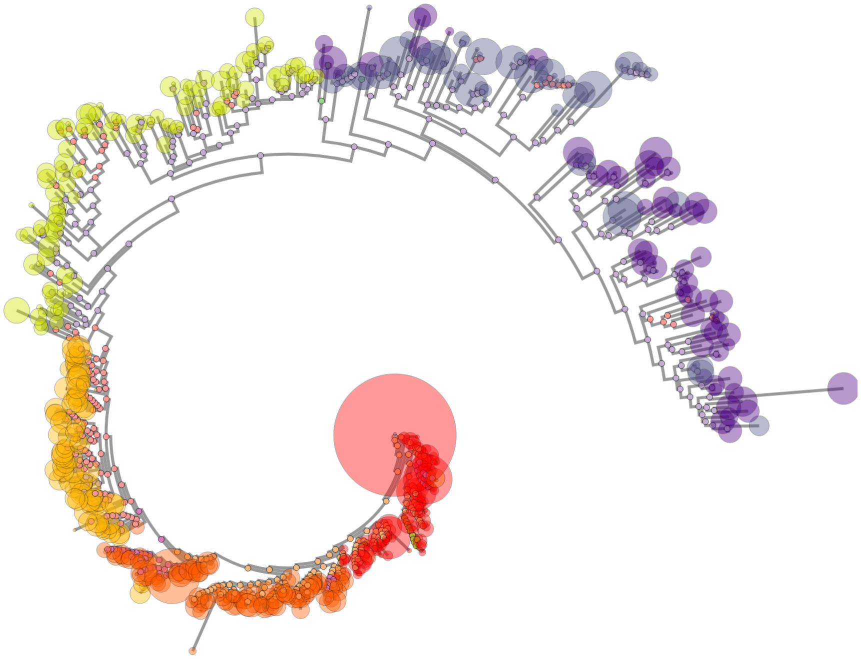


Figure 2.9: Screenshot of the phylogenetic tree viewer. Leaf node size corresponds to sequence copy number. Node color corresponds to time point. Since ancestral sequences have been inferred, ancestral nodes are colored according to the selected motif, which in this case is the N332 glycan supersite.

for further insight into how well the inferred population B estimates the sequences in the ground truth population A. EMD_{FP} removes the constraint on A, allowing any amount of flow from A to B. It is a measure of false positives because it grows when B contains extra sequences distant from any in A. Similarly, EMD_{FN} removes the constraint on B. It grows when B fails to recapitulate sequences in A, and therefore is a measure of false negatives.

To see the effect of sequencing runs of different depths, the experiment was repeated for 300, 1,000, 3,000, and 10,000 reads per time point. The results, which appear in Table 2.1, show the benefit of FLEA's approach of reducing sequence errors via clustering and consensus. The QCS sequences, although they have few false negatives ($EMD_{FN} = 0.0782$) for $n = 10,000$, are dominated by false positives ($EMD_{FP} = 8.3$). However, adding the consensus sub-pipeline virtually eliminates false positives

Table 2.1: EMD metrics for various numbers of reads, averaged across all time points. "mean errors" gives the average number of errors in the reads, estimated from the simulated Phred scores.

n	mean errors	consensus type	EMD	EMD_{FP}	EMD_{FN}
300	9.63	QCS	12.3769	8.3418	2.8956
		HQCS	7.1570	0.4050	5.4271
		HQCS (corrected)	6.4752	0.3020	4.5533
1000	9.63	QCS	10.5433	8.3686	1.2551
		HQCS	2.8279	0.0610	1.1453
		HQCS (corrected)	2.7557	0.0666	1.0405
3000	9.6	QCS	9.5053	8.2837	0.3908
		HQCS	1.6432	0.0146	0.4322
		HQCS (corrected)	1.5168	0.0045	0.2925
10000	9.56	QCS	9.0734	8.3080	0.0782
		HQCS	1.0549	0.0336	0.1735
		HQCS (corrected)	1.0146	0.0073	0.1463

($EMD_{FP} = 0.0336$), at the cost of only a 2.4x increase in false negatives, for a 8.6x improvement in EMD to 1.0549. The frame correction step further improve both EMD_{FP} and EMD_{FN} because it turns false positives into true positives.

The full-length *env* sequencing protocol yields approximately 10,000 reads per run; the P018 data averaged 9,744 reads per time point. Therefore, these results with $n = 10,000$ suggest that FLEA is capable of taking a full sequencing run of CCS reads from a diverse viral population with an average of 9.56 errors per sequence and inferring HQCSs with an average of 1.01 errors per sequence, which corresponds to an average error rate of 0.038%. Moreover, these error rates are mostly caused by low-abundance sequences in both the true population and the inferred FLEA sequences. Figure 2.10 shows that FLEA perfectly recovers all sequences from all time points that account for at least 1.6% of the population.

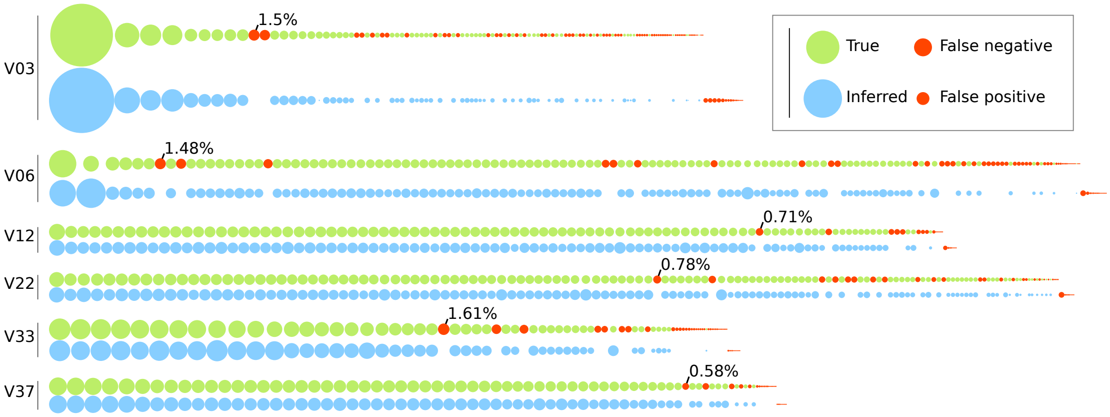


Figure 2.10: Comparison of true sequence abundances versus copy numbers inferred by FLEA for each time point of the simulated P018 data. Each node represents one sequence, with the area denoting its relative abundance in the population. The true population (top) is colored green. For each true sequence, the matching HQCS sequences appears below it in blue. Red nodes denote false negatives and positives. The most common false negative for each time point is annotated with its abundance.

2.3.2 Results on real data: donor P018

FLEA was run directly on the P018 sequences, and the results are summarized here. The full results of this run are available to view at http://flea.murrell.group/view/P018.

Fig. 2.2 shows the number of sequences from the V03 time point that make it to each stage of the quality and consensus pipelines . At three months post infection, the majority amino-acid sequence variant is shared by 52.1% of the population, and the next most common variants accounts for just 8.66%. This relative lack of diversity is consistent with early infection dynamics. By 37 months post infection there is much more diversity: the most common variant accounts for only 3.96% of the population.

Donor P018 shows signs of potential N332 glycan specificity, as shown by the motif trajectories in Fig. 2.6. The glycan supersite, centered around N332 in V3, is a common target for broadly-neutralizing antibodies [64] because they are often conserved, so mutations in these regions are associated with escape [25]. A year into sampling (V12), mutations 328R and 330H dominate, and the majority of sequences also contain

339N from 22 months (V22) onwards.

2.4 Discussion

The FLEA pipeline analyzes longitudinal full-length *env* sequences and provides visualizations of the results. Using simulations, we show that FLEA is capable of inferring accurate HIV-1 *env* consensus sequences and population frequencies. Despite each CCS read containing an average of ten errors, our approach distinguishes variants that differ by as little as one base from an amplicon with high indel variation. It uses those high-quality consensus sequences to generate a codon-aware multiple sequence alignment of all time points, estimate ancestral sequences, infer the phylogenetic tree, and perform many other population-level analyses with high accuracy. These results are presented in a visualization suite that is highly general and applicable to many related sequencing problem.

While our USEARCH-based clustering and consensus strategy for de-noising long PacBio amplicons performs well when error rates are $< 1\%$, there is a clear need for more sophisticated long-read de-noising algorithms that exploit the additional depth of lower quality reads that we currently discard. This will be especially beneficial for longer PacBio amplicons, because the CCS read quality distribution degrades with length. For example, while we can currently obtain around 15,000 CCS reads $< 1\%$ from a P6/C4 RS-II run of our 2.6kb *env* amplicon; this read count drops to $\sim 1,000$ for full-length 9kb HIV genomes.

Both the pipeline and client-side visualizations are under development, with many improvements planned, including a novel clustering algorithm that reduces false positives and a novel consensus algorithm that uses quality scores and performs frame correction. We plan to integrate epitope prediction into the FLEA pipeline and add appropriate visualizations for the case when users have IC_{50} values available for their sequences. Finally, FLEA will be expanded to support other amplicons.

2.5 Supporting Information

2.5.1 Simulation method

The simulation procedure begins with a population of copynumber-weighted HQCS sequences from a real FLEA run. The population is augmented with mutants of the input sequences, to ensure that the simulated ground truth population contains sequences that differ by only a few bases. Each HQCS has a $p = 0.2$ probability to donate 30% of its abundance to a closely-related mutant, which contains one, two, or three substitutions with equal probability. This mutated population is treated as the ground truth for all experiments.

In order to simulate sequencing at different depths, different numbers of N reads are drawn from the same ground truth population for each time point. For each value of N (300, 1,000, 3,000, and 10,000 in this paper), and for each time point, N sequences are sampled with replacement from the copynumber-weighted population. Each read is then mutated with an error model derived from true Pacific Biosciences sequences, in order to mimic the errors introduced by sequencing, especially homopolymer length errors.

To simulate a read r from template t, it is necessary to model both r itself and its Phred scores. First an error rate p is drawn from $p \sim Gamma(\alpha = 2, \theta = 0.0017)$. The length n for each run of identical bases in t (including singletons) is lengthened or shortened with equal probability to be $m = max(n \pm \varepsilon, 0)$, where $\varepsilon \sim Poisson(\lambda = p/c \cdot n^{1.5})$. c is calibration parameter chosen in these experiments to be 1.55 to match observed errors. This process introduces homopolymer length errors, which account for most of the error in Pacific Sciences reads. Then point mutations are introduced at each position with probability $p/4$ of occurring and equal probability for each nucleotide.

Finally, error probabilities are computed for each base as $P = p/4 + m^{1.5}/m$, which is the per-base mutation rate plus a homopolymer error rate. The final simulated Phred scores are obtained by adding error per-base errors $\varepsilon \sim \mathcal{N}(0, 0.1)$ in the natural log domain to these probabilities, then converting to Phred scores.

2.5.2 Sequence order for clustering

`USEARCH`'s `cluster_fast` algorithm runs in a single pass, and therefore is sensitive to the order of the input sequences. We investigated four different strategies: none (no re-ordering), shuffle (randomly shuffle the sequences), sort (sort from high to low quality, as measured by expected number of errors), and reverse sort (sort from low to high quality). Ten trials of simulated sequencing were run to generate 3,000 reads. `FLEA` was run on each dataset with all four ordering strategies.

The results clearly favor reverse sorting, as shown in Table 2.2, which does better on average across the ten trials, and in the worst case it does much better. In the worst case, other methods suffer from false negatives, as shown in Table 2.4. We hypothesize that this behavior is caused by reads from the rare templates – which have a low chance of having a high-quality representative read – loading onto the nearest high-quality template.

Table 2.2: *EMD* score statistics for different ordering strategies, summarized over ten trials.

strategy	min	median	max
none	1.161733	1.754568	4.650523
shuffle	1.042900	1.877201	15.177280
sort	1.362890	2.170702	15.585899
reverse sort	1.077585	1.495208	2.853009

Table 2.3: EMD_{FP} score statistics for different ordering strategies, summarized over ten trials.

strategy	min	median	max
none	0.0	0.012702	0.726121
shuffle	0.0	0.005160	0.634556
sort	0.0	0.019261	0.839726
reverse sort	0.0	0.005992	0.079191

Table 2.4: EMD_{FN} score statistics for different ordering strategies, summarized over ten trials.

strategy	min	median	max
none	0.109267	0.363806	4.064444
shuffle	0.104379	0.359525	13.443283
sort	0.170185	0.623503	13.672487
reverse sort	0.096213	0.233316	1.011137

2.5.3 Pipeline visualizations

Nextflow provides pipeline introspection and performance tools including tracing reports, task order graphs, and timeline visualizations (Figure 2.11).

2.6 Acknowledgements

Chapter 2, in full, is a reprint of material that has been submitted as “Full-Length Envelope Analyzer (FLEA): A tool for longitudinal analysis of viral amplicons”. Kemal Eren, Steven Weaver, Robert Ketteringham, Morné Valentyn, Melissa Laird Smith, Venkatesh Kumar, Sanjay Mohan, Sergei L Kosakovsky Pond, Ben Murrell. The dissertation author was the primary investigator and author of this paper.

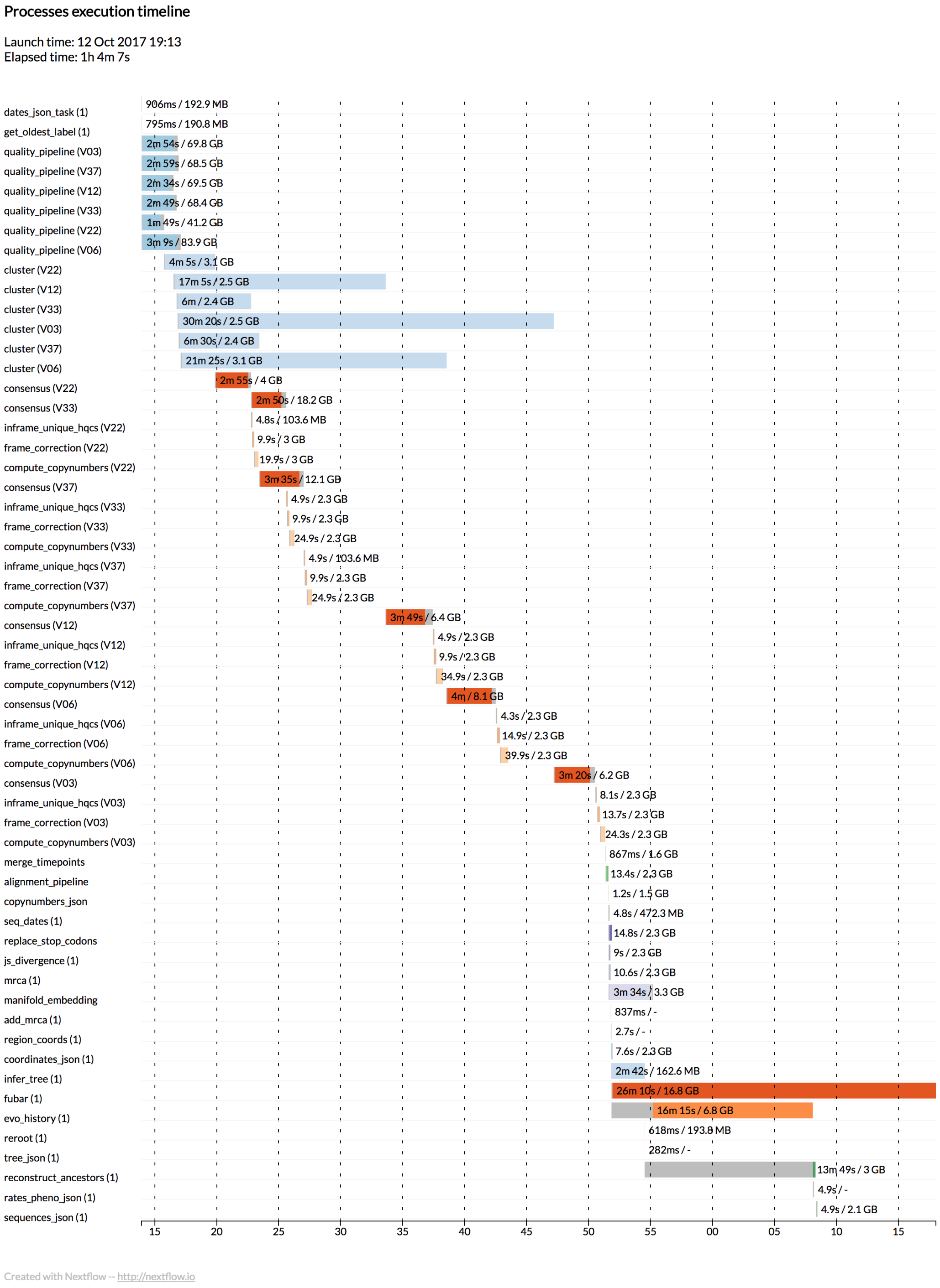


Figure 2.11: Timeline of each task in the FLEA pipeline. Tasks are annotated with time per task and max memory used. Image generated with Nextflow's `-with-timeline` option.

Chapter 3

RIFRAF

Abstract

Motivation: Protein coding genes can be studied using long-read next generation sequencing. However, high rates of indel sequencing errors are problematic, corrupting the reading frame. Even the consensus of multiple independent sequence reads retains indel errors. To solve this problem, we introduce `RIFRAF`, a sequence consensus algorithm that takes a set of error-prone reads and a reference sequence and infers an accurate in-frame consensus. `RIFRAF` uses a novel structure, analogous to a two-layer hidden Markov model: the consensus is optimized to maximize alignment scores with both the set of noisy reads and with a reference. The template-to-reads component of the model encodes the preponderance of indels, and is sensitive to the per-base quality scores, giving greater weight to more accurate bases. The reference-to-template component of the model penalizes frame-destroying indels. A local search algorithm proceeds in stages to find the best consensus sequence for both objectives.

Results: Using Pacific Biosciences SMRT sequences of NL4-3 *env*, we compare our approach to other consensus and frame correction methods. `RIFRAF` consistently finds a consensus sequence that is more accurate and in-frame, especially with small numbers of reads. It was able to perfectly reconstruct over 80% of consensus sequences from as few as three reads, whereas the best alternative required twice as many. RIFRAF is able to achieve these results and keep the consensus in-frame even with a distantly related reference sequence. Moreover, unlike other frame correction methods, RIFRAF can detect and keep true indels while removing erroneous ones.

Availability: RIFRAF is implemented in Julia, and source code is publicly available at https://github.com/MurrellGroup/Rifraf.jl.

Contact: bmurrell@ucsd.edu

3.1 Introduction

The problem of finding the consensus of a set of sequences is fundamental to bioinformatics, especially in the age of high-throughput sequencing. This paper addresses the task of reconstructing an unknown true sequence from a set of error-prone reads. Many algorithms that solve this task focus on *de-novo* or reference-guided assembly of short reads [96, 101, 104]. However, with the advent of third-generation single-molecule sequencing technologies, such as Pacific Biosciences' SMRT sequencing protocol [32], it is now possible to perform full-length sequencing of entire genes or small genomes. Here we will focus on finding the consensus of a set of *amplicon* sequences - where the sequences have the same start and end points. An example application would be targeted sequencing of an entire gene from a viral population (eg. [63]). We focus just on the consensus reconstruction problem, assuming that reads have first been grouped by genetic identity, either using primer ID barcodes [52, 131], or some form of clustering.

Consensus sequences found via multiple sequence alignment may be inaccurate when there are few reads available, or when the reads contain many errors. SMRT sequencing in particular is known to contain mostly indel errors, especially in homopolymer runs. For example, in [63], we discovered that 80% of the sequencing errors were indels. If these indels carry over into the consensus sequence, they cause frameshift errors which corrupt the reading frame, and render the amino acid sequence uninterpretable. If a reference sequence with a trusted reading frame is available, it can be exploited to inform the consensus.

Current approaches that attempt to reconstruct in-frame consensus sequences consider these problems separately. There are approaches to infer the consensus of multiple reads, and there are approaches to correct the reading frame of an already-inferred consensus sequence. Here, we solve these problems jointly, simultaneously considering evidence from the reads and the reference sequence.

One common approach to inferring consensus sequences is from multiple sequence alignments (MSAs), from which the consensus is calculated by taking the most common base in each column. A myriad of multiple alignment algorithms are available [106], any of which may be used for this task. This paper uses `MAFFT` [56, 57] as an example of this strategy when comparing alternatives. A multitude of tools, such as the `cons` command in `EMBOSS` [117], are available for computing the consensus of these alignments. Another approach is to use a partial order alignment [67] representation of the set of sequences, and find the consensus sequence using dynamic programming to extract the heaviest bundles [66]. This paper uses `poaV2`[1] for comparison. Other implementations of this approach include `pbdagcon`[2], which was released by Pacific Biosciences specifically for raw SMRT sequence reads, and `nanopolish` [74], which wraps `poaV2` for Oxford Nanopore reads. Finally, specialized consensus methods are available for specific sequencing technologies; these methods model the specific behavior

[1] https://sourceforge.net/projects/poamsa/

[2] https://github.com/PacificBiosciences/pbdagcon

of their target protocol, such as read length and error model. In this domain, Pacific Biosciences developed the `Quiver` [17] and `Arrow` algorithms[3] for building circular consensus sequences from raw ZMW reads.

Existing approaches for reading frame correction (such as `FrameBot`, which we use here as a comparator) exploit frame-aware codon alignment to a protein reference, followed by inserting or deleting bases in the target sequence [148]. Related algorithms include `FALP` and `LAST` [130], `Frame-Pro` [29], `HMMFrame` [161], and others. Another approach is hybrid sequencing, which supplements long single-molecule reads with short reads [106]. Methods such as `HGAP` [17] use hybrid sequence data to find and remove indels.

This paper introduces a new method for inferring consensus sequences of such reads: the Reference-Informed Frame-Resolving multiple-Alignment Free consensus algorithm (`RIFRAF`). `RIFRAF` considers evidence from both the reads and the reference simultaneously, allowing reads to inform the frame correction process, and is sensitive to the read quality scores to ensure that high-quality bases are more informative. These features allow `RIFRAF` to make highly accurate predictions, even for a small number of error-prone reads. Unlike other frame-correction methods, `RIFRAF` can detect true frameshift-causing indels and keep them while removing spurious indels.

3.2 Methods

`RIFRAF` addresses the following sequence consensus problem. Let t be an unknown template sequence, which is sequenced N times to generate a set of N pairs of reads and quality scores $\mathcal{R} = \{(s^i, p^i)\}_{i=1}^N$. Each read s^i is a noisy observation of t, and each p^i is a vector of error probabilities, one for each base in s^i. The ith character in read

[3] https://github.com/PacificBiosciences/GenomicConsensus

s is denoted s_i, and the substring from the ith to the jth character is denoted $s_{i...j}$. p_i is the probability that s_i is an error; an error at a base is either a substitution, an insertion, or a deletion has occurred next to it. The task is to infer a consensus sequence c that matches the unknown t. Additionally, we also consider a reference sequence r and prefer that c not contain insertions or deletions that change its reading frame relative to r. This is especially useful when the template that generated the reads in $\mathcal{R}$ had an intact reading frame, but the reads themselves have a high indel rate.

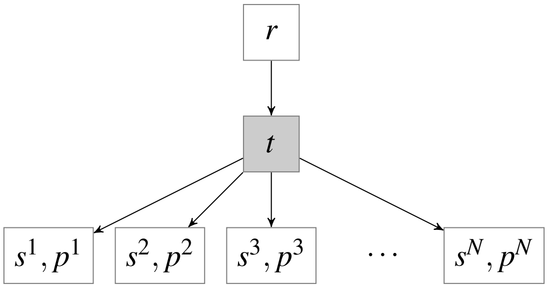


Figure 3.1: Structure of the full model. The unknown template t (grey) has the same reading frame as known reference r. The sequencing process generates error-prone reads $s^1 \ldots s^N$ with quality scores $p^1 \ldots p^N$.

The structure of the full RIFRAF model is shown in Figure 3.1. It infers the unknown template by optimizing two objectives: the quality-aware alignment to the reads, and a frame-aware alignment to the reference. The optimization procedure starts with an initial consensus sequence and proceeds in an iterative greedy manner, mutating the consensus sequence at every step to improve those objectives. RIFRAF uses a number of techniques to speed up convergence: filtering mutations, accepting multiple mutations, forward and backward alignments, banding, batching, increasing indel penalties, and multi-stage optimization.

RIFRAF is implemented in Julia [8], a high-level scientific computing language.

3.2.1 Objective 1: pairwise alignment to reads

In order to find the optimal consensus, it is necessary to assign a score to candidates. `RIFRAF` scores consensus sequences by a global pairwise alignment [98, 134] of each read s with the current values of c. Let $\mathbf{A}$ be the $|s|+1 \times |c|+1$ dynamic programming matrix for aligning c and s. Each $a_{i,j}$ is the score of aligning prefix $s_{1\ldots i}$ to prefix $c_{1\ldots j}$. $a_{0,0}$ is initialized to 0, and the last cell $a_{|s|+1,|c|+1}$ contains the score for the full alignment. The score function for c and s is defined as the full alignment score: $S(c|s) = a_{|s|+1,|c|+1}$. The overall score of consensus sequence c is the sum over the alignment scores for all reads: $S(c|\mathcal{R}) = \sum_{(s,p)\in\mathcal{R}} S(c|s)$.

The sequencing process has an error rate ρ, which by assumption can can be partitioned into $\rho = \rho_{mismatch} + \rho_{insertion} + \rho_{deletion}$. These parameters account for the different error profiles of different sequencing technologies. For instance, in SMRT sequencing, indels are more likely than substitutions. The base move scores for the alignment are derived from these error probabilities.

Typical pairwise alignment uses fixed scores for moves. However, `RIFRAF` also incorporates sequence qualities into the move scores to generate more accurate alignments. The scores for match, insertion, and deletion moves depend on the error probabilities p in the following way. Let $q = \log_{10} p$ (base 10 is used instead of the usual natural logarithm for compatibility with quality scores such as Phred scores). Let $q_{mismatch} = \log_{10} \rho_{mismatch}$, and similarly for the others. Then move scores are calculated as follows

- A diagonal move from $a_{i-1,j-1}$ to $a_{i,j}$ has score $\log_{10}(1-p_i)$ if $s_i = c_j$ (ie. a match), else $q_{mismatch} + q_i$ (ie. a mismatch).
- A vertical move (insertion relative to c) from $a_{i-1,j}$ to $a_{i,j}$ has score $q_{insertion} + q_i$.
- A horizontal move (deletion relative to c) from $a_{i,j-1}$ to $a_{i,j}$ has score $q_{deletion} + max(q_i + q_{i+1})$. If $i = 0$, the score is just $q_{deletion} + q_1$; similarly, $i = |s|$, the score

is just $q_{deletion} + q_{|s|}$.

Intuitively, the penalties for mismatches, insertions, and deletions are more severe when the consensus does not match higher quality regions of the reads. PHRED values are capped at 30 because rarer sources of error that are not sequencing errors (eg. PCR errors) may have very confident PHRED scores, and we do not wish these to be overly informative. This cap can be adjusted if these sources of error can be ruled out (for example if PCR was not used to generate the amplicon library).

The best consensus c^* under Objective 1 (pairwise alignment to reads) is the one that maximizes $S(c|\mathcal{R})$.

3.2.2 Objective 2: Frame-aware alignment to reference

To perform frame correction, RIFRAF requires a reference nucleotide sequence r, which is known to be in-frame. It models the reference sequence r as having diverged from the template t, where the differences between r and t represent evolutionary events, not sequencing error as in Objective 1. The score for the consensus-reference alignment is modified to reflect this difference. First, two new moves are allowed during alignment: codon insertion and codon deletion, each with their own penalty, as shown in Figure 3.2. Second, a new parameter t_{indel} is used as a multiplier for the non-codon insertion and deletion penalties. Together, these two modifications bias the alignment to prefer only codon indels, keeping the consensus in-frame. Because it uses nucleotide alignments, this method works may be expected to work better with more closely related reference sequences, where nucleotide similarity is preserved.

We first let RIFRAF converge to a draft template c without the reference sequence. This draft template is used to approximate the divergence between the true template and the reference, taking the edit distance normalized by the max length $d(r,c)/max(|r|,|c|)$ to obtain a per-base probability of template/reference disagreement (which is used in the

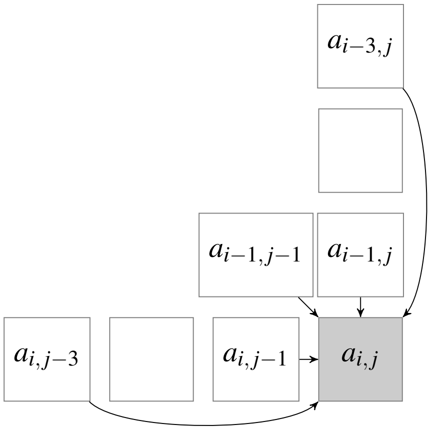


Figure 3.2: Codon moves in the reference alignment dynamic programming matrix. The goal is to favor a consensus that preserves the reading frame. Thus, in addition to the usual single match, insertion, and deletion moves, codon insertions and deletions are also allowed, with a lower penalty than single-base indels.

same manner as the per-base quality scores p in Objective 1). Reference (mis)match, indel, and codon error rates are provided as parameters, and the scores for each move are computed from error rate ρ as $\log_{10}(\rho)$, as before.

The insertion and deletion scores are multiplied by a penalty t_{indel}, which controls the influence of single insertions and deletions in the reference alignment. If t_{indel} is small, frame-destroying indels may appear in the consensus, but if it is large, the consensus will be forced into the reference reading frame, even if the unobserved template really did contain indels. As we show in Section 3.3, this penalty can be tuned to discard spurious indels while keeping true ones.

`RIFRAF` combines both objectives into a single score, allowing the reads to inform the frame correction. The score of the consensus to reference alignment is denoted $S_r(c|r)$, and the full score function is:

$$S(c|\mathcal{R},r) = S(c|\mathcal{R}) + S_r(c|r).$$

3.2.3 Optimization procedure

An exhaustive search for the optimal consensus c^* would be intractable, so `RIFRAF` uses a variant of the following greedy search algorithm, with some optimizations to speed up convergence:

1. Start with a guess c^0. `RIFRAF` chooses the read with the lowest expected number of errors.

2. For the most recent guess c^i, examine a set of candidate single mutations, such as insertions, deletions, and substitutions. Note that these candidates vary at each optimization stage. Keep all that improve the score $S(c^i|\mathcal{R}, r)$. Call the set of candidate mutations $\mathcal{C}$.

3. If $\mathcal{C}$ is empty, accept c^i and terminate. Otherwise, choose some subset of $\mathcal{C}$, apply them to c^i to obtain c^{i+1}, and iterate.

`RIFRAF` works in two stages, first optimizing just $S(c|\mathcal{R})$, and then optimizing the full $S(c|\mathcal{R}, r)$.

Filtering mutations

When comparing the template to the reads, we need not consider all possible modifications to the current consensus. For example, if any candidate mutation to c does not appear in any pairwise alignment of c with a read, that mutation need not be scored. Since it has no support among any observed sequence, it is likely to hurt the alignment score. Similarly, during the frame correction stage, the model only proposes insertions or deletions that appear in the pairwise alignment to reference.

Multiple mutations

Instead of accepting only the best mutation in $\mathcal{C}$, `RIFRAF` accepts all the mutations that are separated by a certain number of positions: $n_{separate}$ (the default value is 15, i.e. five codons). The candidates are accepted in order from best to worst score. This policy allows `RIFRAF` to converge in many fewer iterations than if it only accepted one mutation per iteration. $n_{separate}$ ensures that the changes to the consensus are relatively independent of each other, and that the score of one is unlikely to be affected by the acceptance of another. After accepting mutations in $\mathcal{C}$, `RIFRAF` also compares the new score to the score that would be obtained from accepting only the single best mutation in $\mathcal{C}$, and optionally accepts that single mutation instead if it results in a better score.

Forward and backward alignments

Recomputing the full alignment matrix for each candidate mutation to c would be prohibitively expensive. For a sequence c from alphabet $\{A,C,G,T\}$, there are $4(|c|+1)$ insertions, $3|c|$ substitutions, and $|c|$ deletions to consider. Computing the alignment matrix $\mathbf{A}$ for each candidate requires $O(cs)$ operations, so each iteration of the proposed algorithm would require $O(Nc^2s)$ operations (we omit $|\cdot|$ in $O(\cdot)$ for clarity). Instead, `RIFRAF` uses forward and backward alignments to compute the new score for any single change to c by only recomputing a single column of $\mathbf{A}$ [17].

To achieve this, in addition to the prefix alignment matrix $\mathbf{A}$, where $a_{i,j}$ is the score for aligning prefix $s_{1\ldots i}$ to prefix $c_{1\ldots j}$, `RIFRAF` also computes the suffix alignment matrix $\mathbf{B}$, where $b_{i,j}$ is the score for aligning suffix $s_{i+1\ldots|s|}$ to suffix $c_{j+1\ldots|c|}$. Note that $a_{|s|,|c|} = b_{0,0}$ is the score for the full alignment. For any j, that alignment score can also be computed from columns $\mathbf{A}_{\cdot,j}$ and $\mathbf{B}_{\cdot,j}$:

$$\forall j \in [0\ldots v] : a_{|s|,|c|} = b_{0,0} = max_i(a_{i,j} + b_{i,j}) \tag{3.1}$$

Modifying c_j leaves unchanged columns $0 \dots j-1$ of **A**, and also leaves unchanged columns $j \dots v$ of **B**. Therefore, for all three types of mutations, computing the new score requires that at most only a single new column of **A** must be recalculated.

1. substitution at c_j: compute $\mathbf{A}_{\cdot,j}$; new score is $max_i(a_{i,j} + b_{i,j})$.
2. insertion after c_j: compute $\mathbf{A}_{\cdot,j+1}$; new score is $max_i(a_{i,j+1} + b_{i,j})$.
3. deletion of c_j: no new column necessary; new score is $max_i(a_{i,j-1} + b_{i,j})$.

Using the forward and backward alignments, all possible mutations to the consensus can be scored in $O(Ncs)$ operations.

During the alignment of the template and reference, additional columns must be recomputed to account for codon insertion and deletion moves.

Banding

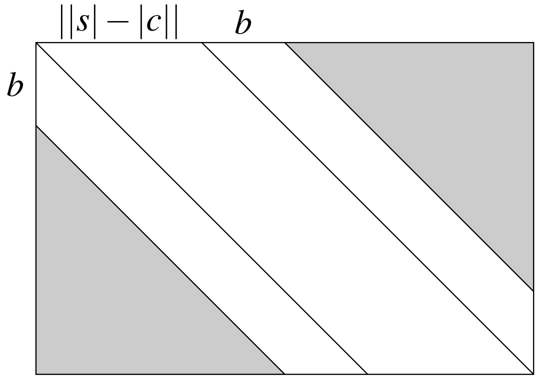


Figure 3.3: Banded alignment. Alignments must stay within the banded region of the dynamic programming matrix.

Despite the improvements from using forward and backward alignments, each iteration is still approximately quadratic in the length of the consensus, assuming $|c| \approx |s|$. Alignment banding [15, 16] further reduces the number of operations per iteration. For a given bandwidth parameter b, the maximum usable column size in **A** and **B** is

$2b + ||s| - |c|| \ll |s|$, so evaluating a possible mutation requires many fewer operations than recomputing the full column. Alignment moves are only allowed to originate inside the band, so alignment paths must stay within the band boundaries (see Figure 3.3). With banding, the time complexity per iteration becomes $O(Nc(\sqrt{s} + b))$, since $||s| - |c||$ grows like $\sqrt{|s|}$ under reasonable assumptions.

RIFRAF dynamically increases the bandwidth if the number of differences in the banded alignment is sufficiently larger than the expected number of differences implied by the read's quality scores, under the assumption that the difference between the template candidate c and the true template is much smaller than the number of sequencing errors in s. Let r be the observed number of differences between s and c, and e be the expected number of errors computed from the quality scores p. If the value of r is in the upper tail of a Poisson distribution with mean parameter e, then the bandwidth is doubled and the alignments are re-computed. α controls the size of this upper tail probability, with a default value of 0.1.

Batching

RIFRAF uses a variety of batching strategies to speed up convergence. If the number of reads is greater than a threshold k (default 5), the best k reads by error rate are fixed as the initial batch, and RIFRAF runs to convergence. This ensures that RIFRAF first converges without considering the many spurious mutations presumably present in less accurate reads. The resulting initial guess is further refined at the refinement stage, this time with a different random batch of size n_{batch} (default 20) for each iteration. Sequences are chosen for inclusion in the batch by sampling from a multinomial distribution of their error rates, parameterized by parameter ρ between 0 and 1. When $\rho = 1$, all the weight is evenly distributed among the top n_{batch} sequences. Interpolating from $\rho = 1.0$ to $\rho = 0.5$, the probabilities become proportional to the read error rates. Interpolating from $\rho = 0.5$ to $\rho = 0$, the probabilities become uniform. By default, $\rho = 0.9^i$, where i is the number of iterations since random batching activated. Like the fixed batch, this strategy speeds up convergence by initially avoiding inaccurate reads, then gradually letting them contribute to resolve uncertain bases if necessary.

Ideally, n_{batch} is small enough to make each iteration fast, but large enough that RIFRAF converges stably. RIFRAF tries to detect if n_{batch} is too small by monitoring the change in score after each iteration. If the new score is worse than the old score by more than a certain percent (10% by default), n_{batch} is increased to $2n_{batch}$, then $3n_{batch}$, etc.

Combining all of the previous optimizations, a single iteration's time complexity is reduced from $O(Nc^2s)$ to $O(n_{batch}c(\sqrt{s}+b)$.

3.2.4 Increasing indel penalties

Whenever the algorithm converges to a consensus c^i, if single indel moves were used in computing $S_r(\cdot)$, the single insertion and deletion scores are multiplied by a parameter t_{indel}, increasingly encouraging the alignment with the reference to use only codon indels, thereby keeping c in-frame. This process repeats up to m times, so the maximum multiplier is $(t_{indel})^m$. If the penalty is large enough, the consensus will always be forced into the reference's reading frame, which is the default behavior. However, some consensus sequences really are out of frame relative to the reference. The indel penalties can be tuned so that RIFRAF correctly identifies true frameshifts, with a small risk of allowing some spurious ones.

3.2.5 Multi-stage optimization

The full optimization procedure proceeds in stages, allowing RIFRAF to converge quickly by focusing on different objectives in different stages.

1. Initial stage: Do not use the reference. Propose all mutations to the consensus that appear in the pairwise alignments. Use the fixed batch, if available. If no reference was provided, stop after this step.

2. Frame correction stage: Use the full model, including the reference and reads. Propose indel candidate mutations that appear in the consensus-reference alignment. Increasingly penalize single indels in alignment of r and c. Use the fixed batch, if available.

3. Refinement stage: Propose only substitutions (no indels) to the consensus that appear in the pairwise alignments, no longer considering the reference. Use random batches, with decreasing ρ.

The initial stage quickly finds a good candidate consensus from the reads alone. The frame correction stage uses a reference to penalize indels that cause frame shift errors, correcting the reading frame of the template in a way that is maximally compatible with both the reads and the reference. Finally, the refinement stage ensures that the reference influences only the frame of c, and exerts no bias upon the nucleotides themselves. The final stage also fixes biases introduced by the fixed batch in the first two stages.

3.3 Results

We compared `RIFRAF` to two other methods: `MAFFT` [56] followed by the standard per-column consensus, and `POA` [67] with the heaviest bundle consensus algorithm [66]. All three methods were run with and without reference-guided frame correction. `RIFRAF` natively performs frame correction, but only if it is given a reference sequence. To distinguish these models in this section, we refer to the model with no reference as $\texttt{RIFRAF}_{\texttt{nr}}$, and the model with a reference as $\texttt{RIFRAF}_{\texttt{ref}}$. `FrameBot` [148] was used for correcting results from MAFFT and POA; these are referred to as MAFFT_FB and POA_FB.

A full-length sequencing run of Pacific Biosciences SMRT sequencing on *env* from HIV-1 subtype B strain NL4-3 was used for the comparison [63]. The true sequence of NL4-3 is known, so results could be compared to the ground truth. The filtered data (available on FigShare[4]) contains 27,600 reads with expected error rate 1% or better, which were further filtered and processed as follows. To make the problem more challenging and better reveal differences between methods, very high quality sequences were excluded (expected error rate $< 0.1\%$). Short fragments and long reads (often concatemers) were discarded by filtering out sequences 25 bases shorter or longer than the median of 2,597. PacBio reads come in random orientations, so reads were converted to their reverse complement, if necessary. Extra bases around the amplicon were removed by aligning to NL4-3 *env* without penalizing terminal gaps, then trimming terminal insertions. After preprocessing, 9,473 sequences remained, with a mean error rate of 0.0015 (the distribution of errors appears in Figure S1). All experiments were run for 1,000 trials on randomly sampled reads.

Choice of reference. A set of reference sequences – shown in Figure S2 – were tested to investigate how frame correction accuracy deteriorates for distantly-related references. The results are shown in Figure 3.4. Nucleotide results from MAFFT_FB and POA_FB were both equally insensitive to the choice of reference, whereas $\texttt{RIFRAF}_{\texttt{ref}}$'s results did degrade slightly. However, the reverse is true for the protein sequences, with $\texttt{RIFRAF}_{\texttt{ref}}$'s performance degrading by half an amino acid on average, and the others degrading by more than one. This difference indicates that $\texttt{RIFRAF}_{\texttt{ref}}$ not only keeps the consensus in-frame, but also makes better choices of inferring which nucleotides are truly indel errors. Finally, $\texttt{RIFRAF}_{\texttt{ref}}$ was the most accurate, regardless of choice of reference. As expected, $\texttt{RIFRAF}_{\texttt{ref}}$'s frame correction strategy works best with a closely

[4] https://doi.org/10.6084/m9.figshare.5643247

related reference, but these results show that it is capable of working even with a distant reference. Except where noted, the most distant reference, B.BR, was used for the rest of the results.

Number of sequences. Clusters of 2, 3, 4, 5, 6, 8, 10, 15, and 20 reads were randomly sampled for this experiment. The fraction of perfectly reconstructed consensus sequences per 1,000 trials appears in Figure 3.5a. For fewer than ten sequences, both versions of RIFRAF dominate the other corresponding methods. For instance, $\texttt{RIFRAF}_{\texttt{ref}}$ gets over 90% correct with access to only four reads. POA_FB does not achieve similar results until $N = 8$, and MAFFT_FB does not until between $N = 10$ and 15. Interestingly, POA's results actually degrade significantly for $n > 6$, but POA_FB continues to improve, because POA tends to include extra bases on the ends of the consensus sequence which are then removed by FrameBot. These extra bases also affect the average number of nucleotide errors (Figure 3.5b): for $N = 20$, POA averages one error per sequence, whereas all the other methods average none.

The average number of protein errors (Figure 3.6a) highlights the importance of frame correction. Frame shifts cause the translated consensus sequences to differ greatly from the true protein sequence, especially for $n < 15$. For $N = 2$, fully half of each protein sequence is wrong on average, regardless of method. Even for $N = 20$, sequences from $\texttt{RIFRAF}_{\texttt{nr}}$ and POA contain about 100 errors. On the other hand, the corrected sequences (shown in Figure 3.6b for clarity) contain nearly no errors for $n > 10$. $\texttt{RIFRAF}_{\texttt{ref}}$ again performs best here, approaching zero errors even for $N = 3$.

Interestingly, frame correction of MAFFT and POA often made the nucleotide sequences *less* accurate, whereas it improves $\texttt{RIFRAF}_{\texttt{ref}}$. This result supports the idea that RIFRAF's method of integrating frame correction into the consensus algorithm makes it more accurate by allowing all reads to inform the correction process. FrameBot, which only has access to a single consensus sequence, cannot use the extra information in the reads, and therefore cannot achieve the same accuracy.

Execution times appear in Figure 3.7. Without frame correction, all three methods are comparable for small numbers of sequences, but $\texttt{RIFRAF}_{\texttt{nr}}$ scales better, due to its batching scheme. Frame correction adds a constant factor to all three methods' execution times. $\texttt{RIFRAF}_{\texttt{ref}}$'s constant factor is larger, but, because it scales better, it overtakes the others between $N = 10$ and $N = 15$.

Sequence length. Figure 3.7 also shows execution time for varying sequence lengths. For more details on this experimental setup, see SI section 2. $\texttt{RIFRAF}_{\texttt{ref}}$ scales less well than the other methods, taking about twice as long as MAFFT_FB and POA_FB for the full-length sequences. However, it is comparable with the others at $\ell = 900$, and faster than the others for $\ell < 600$. This difference in speed is due to RIFRAF's iterative approach, which requires recomputing parts of each pairwise alignment after every iteration.

Detecting true frameshifts. In the other experiments, strict frameshift penalties were used to ensure the consensus stays in-frame. However, sometimes frameshifts are biologically plausible, such as in integrated (but non-functional) proviral Env sequences, or in the cytoplasmic tail of Env leading to a truncation, but preserving infectivity. If true frameshifts may occur in the template sequence, it may be preferable to relax this frameshift penalty. $\texttt{RIFRAF}_{\texttt{ref}}$ can be tuned to accept frameshift indels with enough support in the reads, with only a small increase in the frequency of spurious frameshift indels. To demonstrate this, single base insertions and deletions were added to NL4-3 in both homopolymer and non-homopolymer regions (details in SI section 3). t_{indel} was set to 1.05, and the max frameshift indel penalty multiplier m varied from 0 to 12. We call an in-frame sequence a "positive", so increasing m increases the false positive rate by forcing sequences with real frameshifts incorrectly into frame. To get the true positive rate, $\texttt{RIFRAF}_{\texttt{ref}}$ was also run on the unmodified sequences. Note that while we introduce only a single true indel into our "negative" cases, the analysis is always at the whole-sequence level. We are not just detecting the presence or absence of the specific indel we introduce. Thus to achieve a high true positive and low false positive rate, $\texttt{RIFRAF}_{\texttt{ref}}$ must successfully ignore spurious indels at *any* position in the "positive" cases, while successfully identifying the real indel we introduce in each "negative" case. The resulting ROC curves, which appear in Figure 3.8 for $N \in 3,5,10$, show that $\texttt{RIFRAF}_{\texttt{ref}}$ can find true indels while controlling the false positive rate, using either a closely related reference or a distant one. A useful trade-off occurs for $m = 6$, which scores close to the maximum true positive rate while keeping the false positive rate close to zero.

In agreement with the accuracy results, a more closely related reference (HXB2) improved inference for $N = 3$ for this task. As expected, real homopolymer indels in homopolymer regions are harder to discriminate than non-homopolymer indels (See SI section 3 for more detailed results).

3.4 Conclusion

RIFRAF uses quality scores and a reference sequence to infer accurate frame-corrected consensus sequences. It can often find the correct consensus, even from small numbers of reads or with a distant reference, as shown in our experimental results. RIFRAF with frame correction can be slower than taking a consensus from a multiple sequence alignment, but in experiments with real SMRT sequences it finds consensus sequences that are significantly more accurate. The benefits of using a reference to reduce frameshift errors are especially apparent when comparing translated amino acid sequences, where a single frameshift causes the entire downstream sequence to be incorrect. Finally, RIFRAF can detect and retain true frameshifts during frame correction, and, to our knowledge, is the only method capable of this.

While RIFRAF performs well with distantly related reference sequences, performance is improved when using closely related references. However, when sequencing diverse populations, we note that it is always possible to first infer a set of autologous sequences from clusters or primer ID bundles that have a large number of reads, and so should be accurate. These can then be used as references to correct the reading frame of the less-represented members of the population, providing an improved accuracy over just using a more distantly related reference. We recommend using this strategy whenever possible.

`RIFRAF` can improve the ability to resolve minority variants in sequenced populations. Its ability to find results comparable to `MAFFT` with three times fewer reads will be essential for identifying minority variants in the population with greater precision. More generally, `RIFRAF` will be useful whenever an accurate consensus sequence must be inferred from a small number of full-length sequences, especially when quality scores and a reference sequence are available.

When sequencing any population, it is often advisable to sequence a clonal representative of that population first (NL4-3 *env* here), to investigate the sequencing performance for that case. We recommend using such sequence datasets to investigate the behavior of `RIFRAF` on new genes, especially if the user seeks to detect real frameshifts. To this end, we provide a Jupyter notebook that allows one to replicate the accuracy and ROC analyses from this manuscript on any clonal amplicon dataset.

`RIFRAF` will continue to be developed along multiple lines. First, the current approach for performing frame correction needs to be faster, to keep pace with the increasing volume of available sequence data. Further work needs to be done to speed it up via optimization or algorithmic advances. Possible approaches include: re-using partial alignments, speeding up alignments with k-mer seeding, and only correcting the frame of obviously problematic regions. Another improvement would include amino acid matching penalties in the reference-to-template alignment, which would allow even more distantly related reference sequences to be used, where the nucleotide homology has been completely obliterated. Another useful feature would be to infer calibrated quality scores for the consensus sequence, in order to communicate uncertain regions to the user. Finally, RIFRAF is extensible to other systems and sequencing technologies. In particular, we plan to investigate its behavior and tune its error model for Oxford Nanopore data, and to extend the method to support amplicons containing both non-coding and coding regions, which may contain different (potentially overlapping) reading frames.

The RIFRAF source code is available at https://github.com/MurrellGroup/Rifraf.jl.

3.5 Supporting Information

3.5.1 NL4-3 and references

The distribution of estimated error rates of the NL4-3 reads appears in Figure 3.9. The phylogenetic tree of the reference database appears in Figure 3.10.

3.5.2 Length of template experiment

As a proxy for varying amplicon lengths, we sampled prefixes of varying lengths from our NL4-3 sequence dataset. This protocol is valid because SMRT sequencing does not have a positional bias (see Figure S8D in [63]), so the distribution of errors in the prefix should match the overall errors. However, these short reads are of lower overall quality than true short CCS reads, which have the benefit of more subreads in the ZMW read.

The experiment was run for 1,000 trials of clusters of size $N = 3$. The fraction of correct consensus sequences appears in Figure 3.11a. Although all methods degrade in accuracy for longer sequences, $\texttt{RIFRAF}_{\texttt{ref}}$ degrades much more slowly, getting at least

80% perfectly correct even for the full-length amplicon. For this number of sequences, $\texttt{RIFRAF}_{\texttt{ref}}$ beats even the other methods with frame correction. As expected, frame correction keeps all three methods in-frame (Figure 3.12a). The amino acid edit distance for frame correction methods alone appears in Figure 3.12b for clarity, showing that $\texttt{RIFRAF}_{\texttt{ref}}$'s frame-corrected amino acid sequences contain an average of less than one error even for the full-length amplicon, whereas the accuracy of other methods degrades from one error all the way to three errors for POA_FB and four errors for MAFFT_FB.

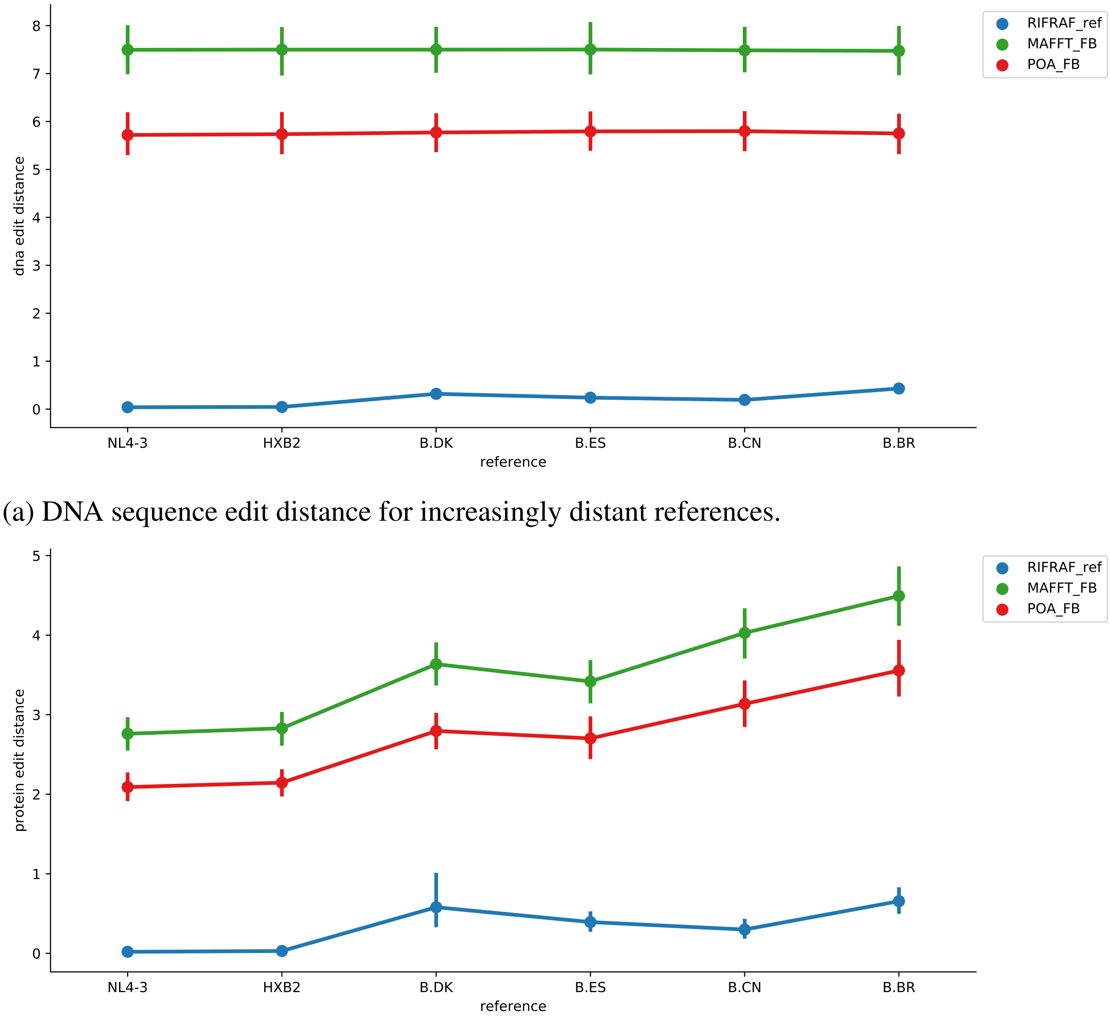
3.5.3 True indel experiments

Single indels were simulated in both homopolymer (defined as four or more identical bases in a row) and non-homopolymer regions of NL4-3 in the following manner. First, a region was sampled uniformly from all matching regions; i.e. a homopolymer region is chosen at random from all homopolymer regions. Insertions or deletions in homopolymer regions were simulated by inserting or deleting a base. Deletions to non-homopolymer regions were simulated by removing a single random base from that region. Insertions in non-homopolymer regions were simulated by choosing a non-homopolymer position and inserting a random base either before or after it.

Each read has a small probability p of not being modified to match the new template, where p is proportional to the read's estimated error rate and is calculated as the mean of the Phred scores after converting them to error probabilities. Those reads that were modified were pairwise aligned to the template and the matching base was inserted or deleted from the correct position. Quality scores for insertions were drawn from a uniform distribution between the two Phred scores on either side of the insertion.

3.6 Acknowledgements

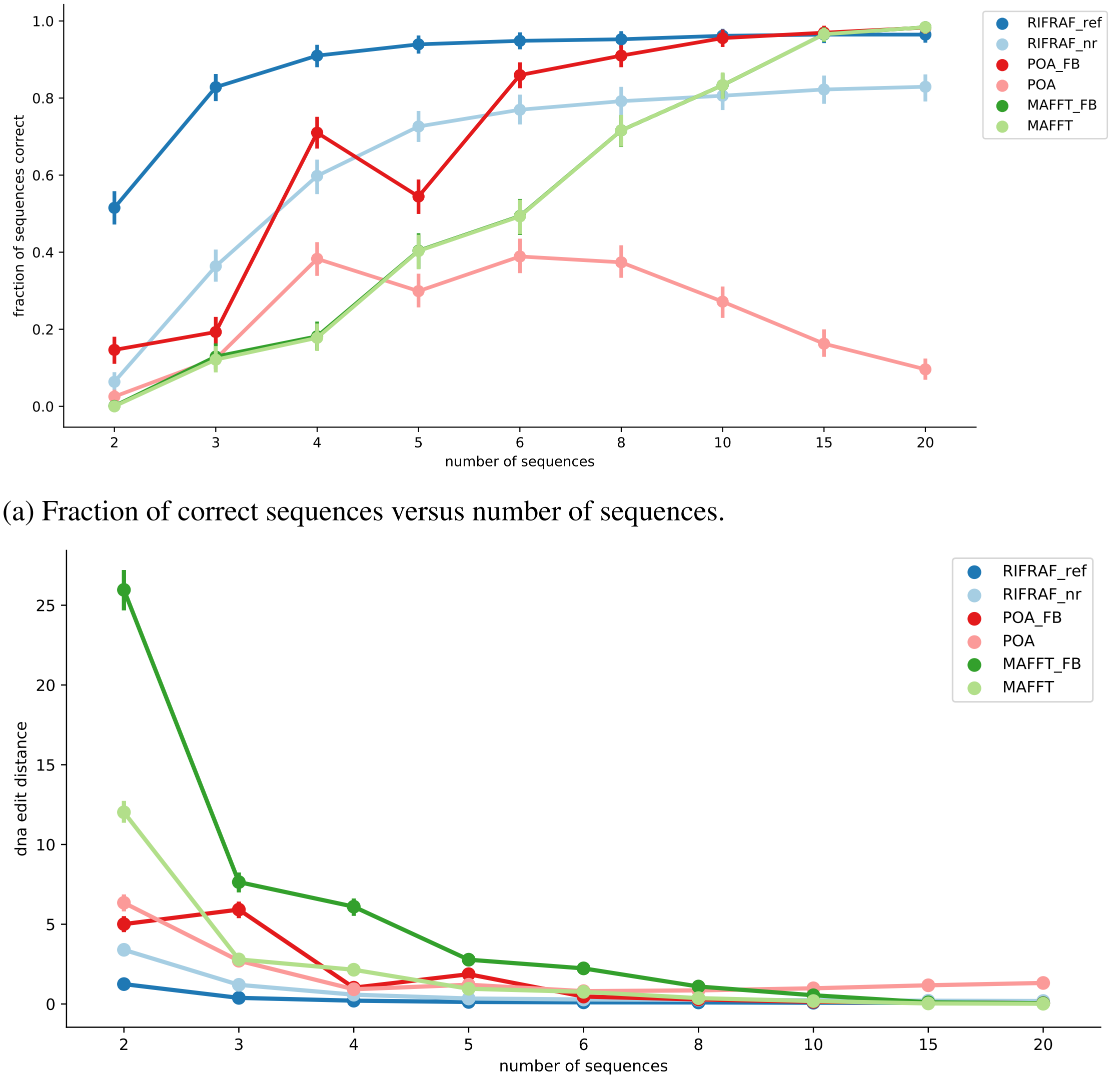
Chapter 3, in full, is a reprint of material that has been submitted as “RIFRAF: a frame-resolving consensus algorithm”. Kemal Eren, Ben Murrell. The dissertation author was the primary investigator and author of this paper.



(a) DNA sequence edit distance for increasingly distant references.

(b) Protein sequence edit distance for increasingly distant references.

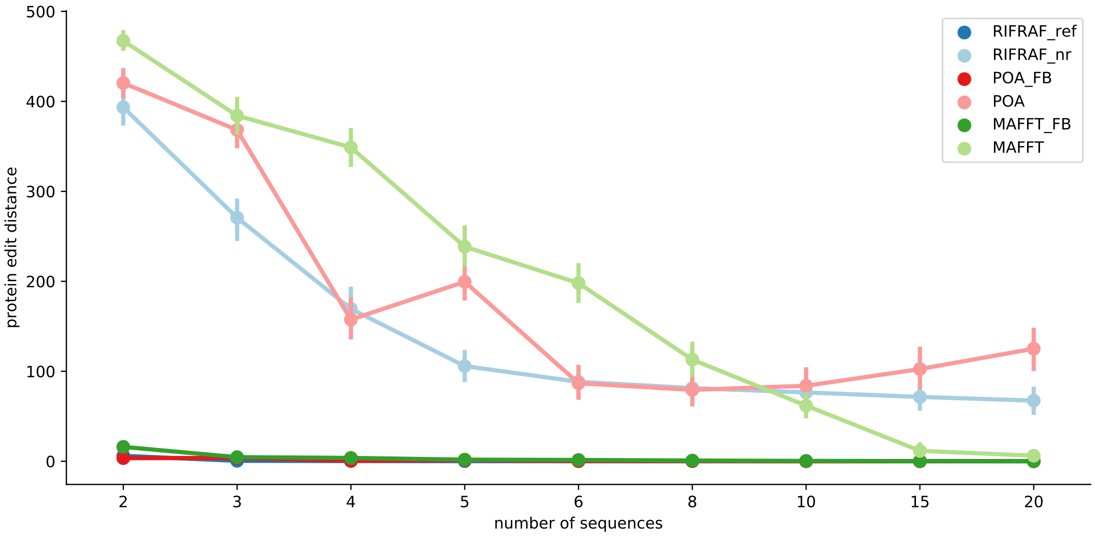
Figure 3.4: Edit distance for increasingly distant references. All methods do better with closely-related reference, but their rate of performance degradation is important because a related reference may not always be available. Run with $N = 3$ full-length reads.



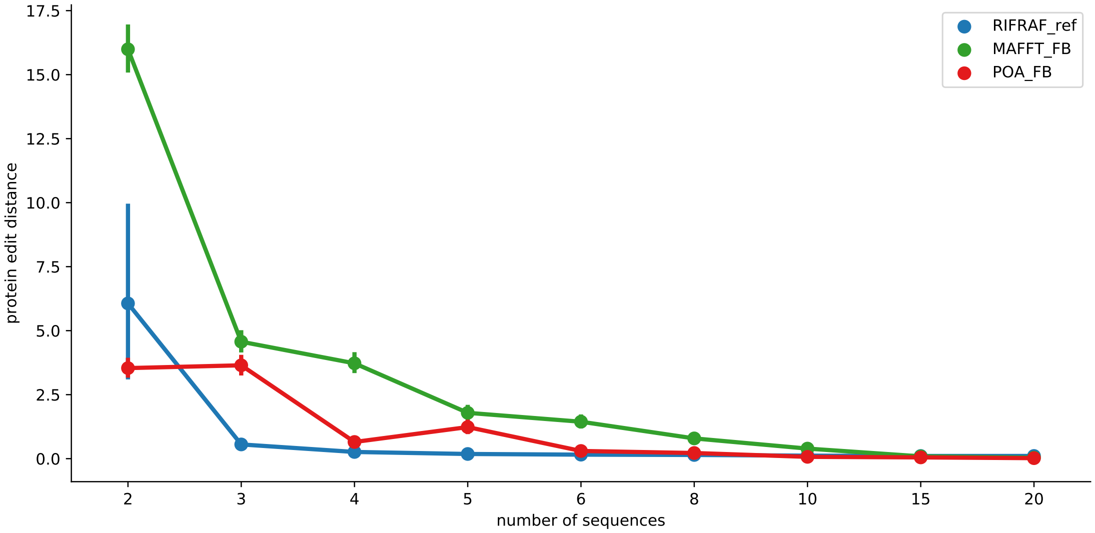
(a) Fraction of correct sequences versus number of sequences.

(b) DNA edit distance versus number of sequences.

Figure 3.5: DNA results. Fraction of correct sequences (left) and mean edit distance between the consensus and the template (right) for increasing *N*.

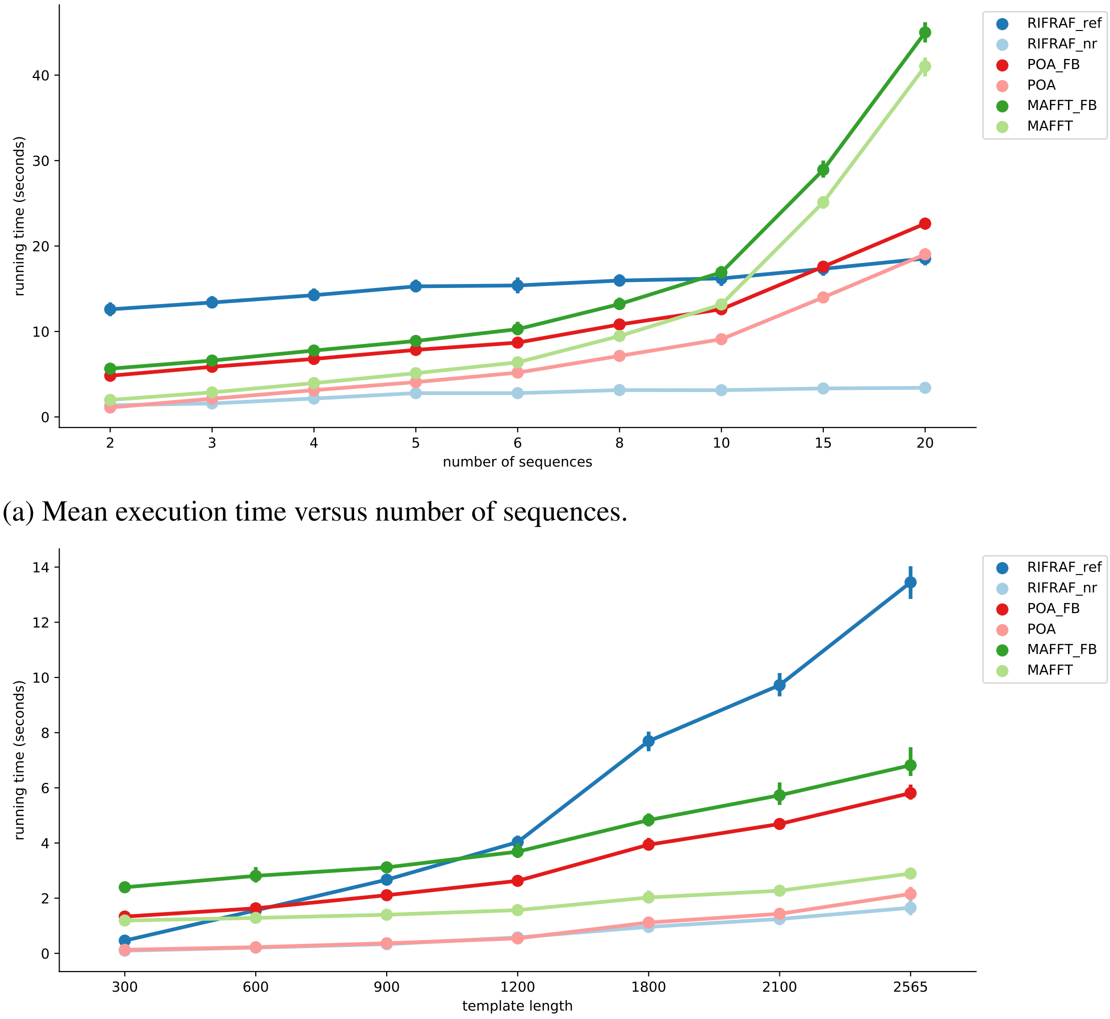


(a) Protein edit distance versus number of sequences; all methods.



(b) Protein edit distance versus number of sequences; frame correction only

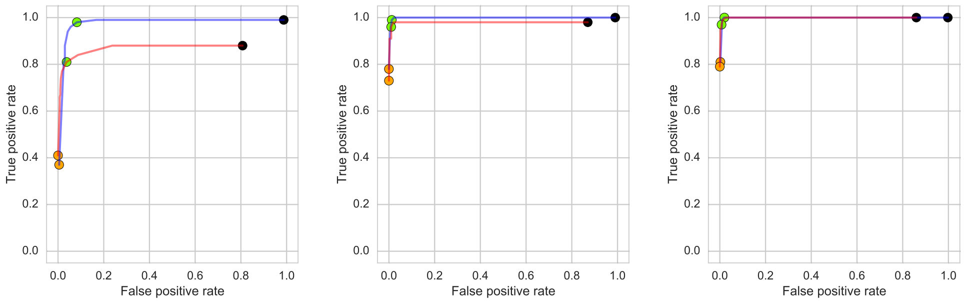
Figure 3.6: Same results as Fig. 3.5, but for the translated protein sequences. The fraction of correct sequences is not reproduced, since those figures are identical. The left figure show results for all methods. The right figure show the same data, zoomed to show the frame-corrected results. Note Y axis scale.



(a) Mean execution time versus number of sequences.

(b) Mean execution time versus template length.

Figure 3.7: Mean execution time, varying both number of sequences (left) and sequence length (right). Note that intervals on the x-axis are not linear.



(a) $N = 3$ (b) $N = 5$ (c) $N = 10$

Figure 3.8: ROC curves for true indel experiments, with max indel penalty multiplier m varying from 0 to 12. The orange point denotes results from $\texttt{RIFRAF}_{\texttt{nr}}$, while the remainder of the curve was generated by $\texttt{RIFRAF}_{\texttt{ref}}$. The green point corresponds to a max indel penalty multiplier $m = 6$. Both a related reference (HXB2, blue) and distant reference (B.BR, red) were used.

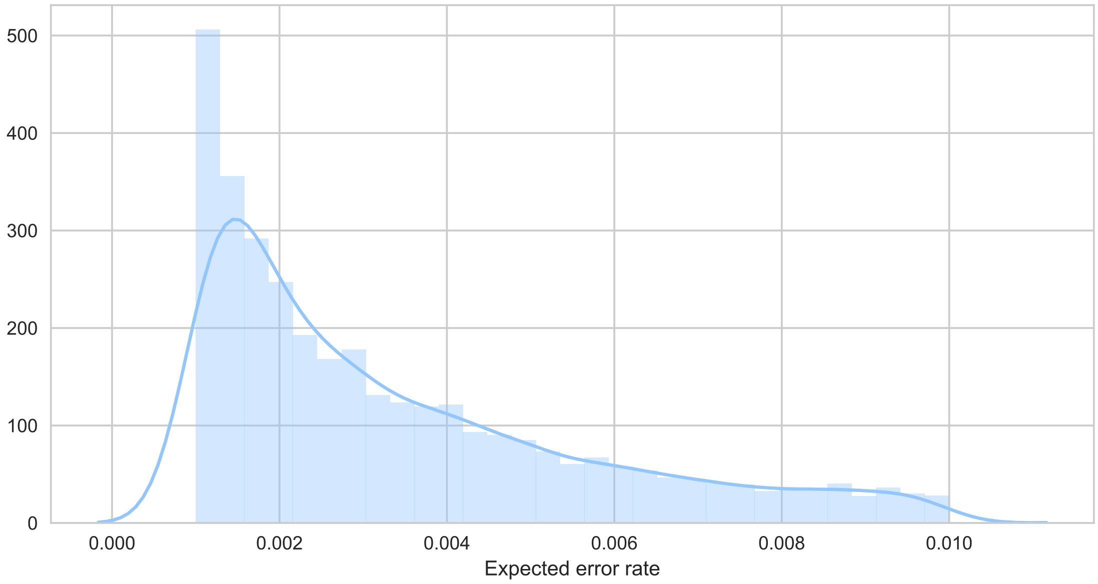


Figure 3.9: QV-derived error rates of the NL4-3 sequences used.

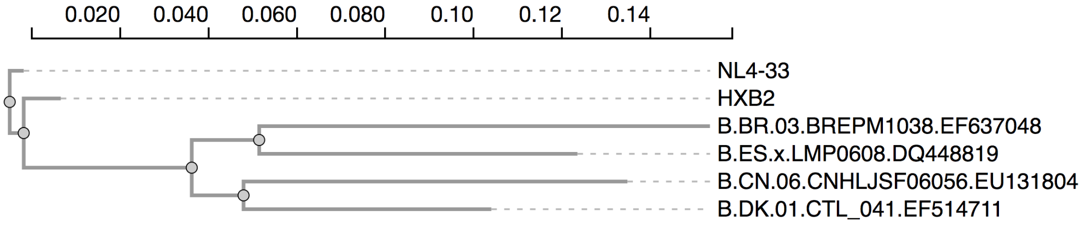
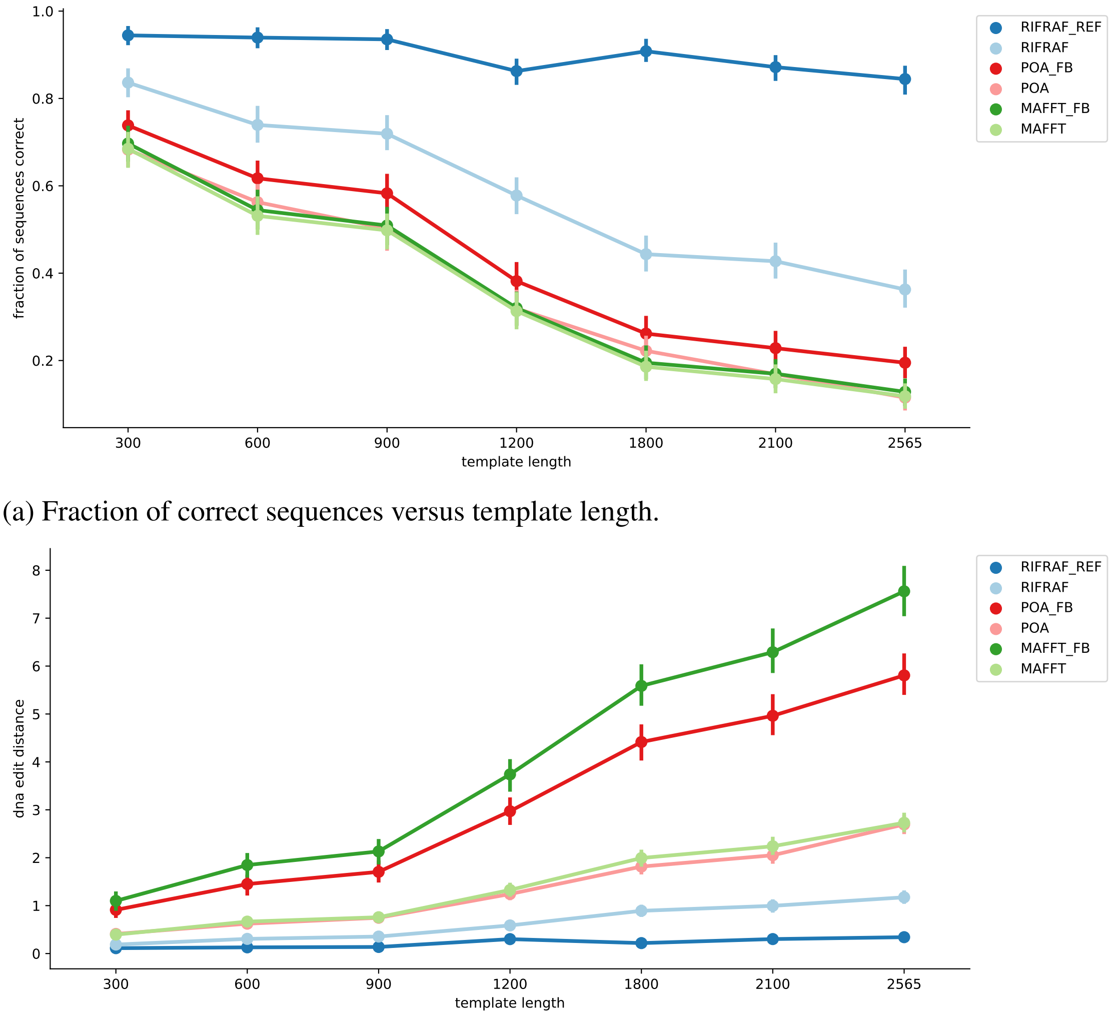


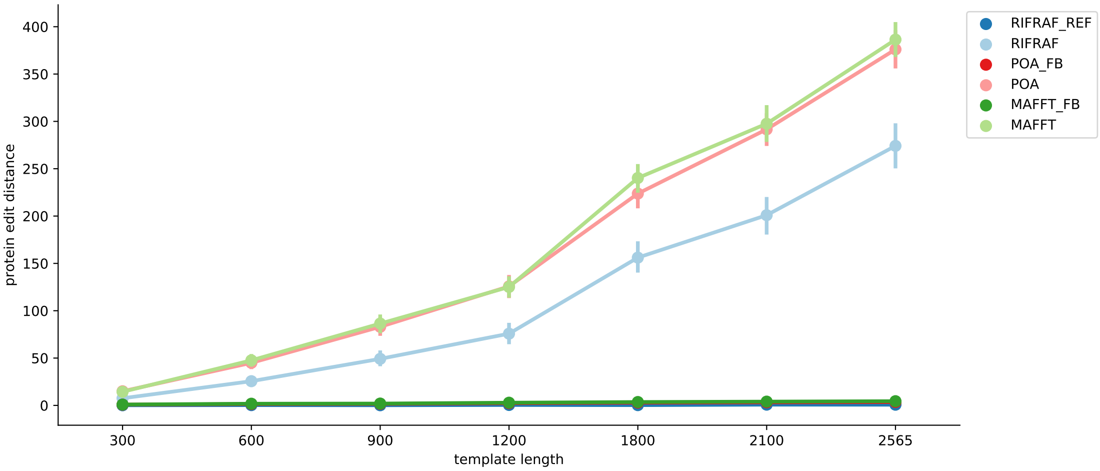
Figure 3.10: Phylogeny of references used by RIFRAF$_{ref}$. The references were aligned with MAFFT [56, 57], the phylogeny inferred by FastTree [109, 110], and visualized with PhyloTree.js (https://github.com/veg/phylotree.js).



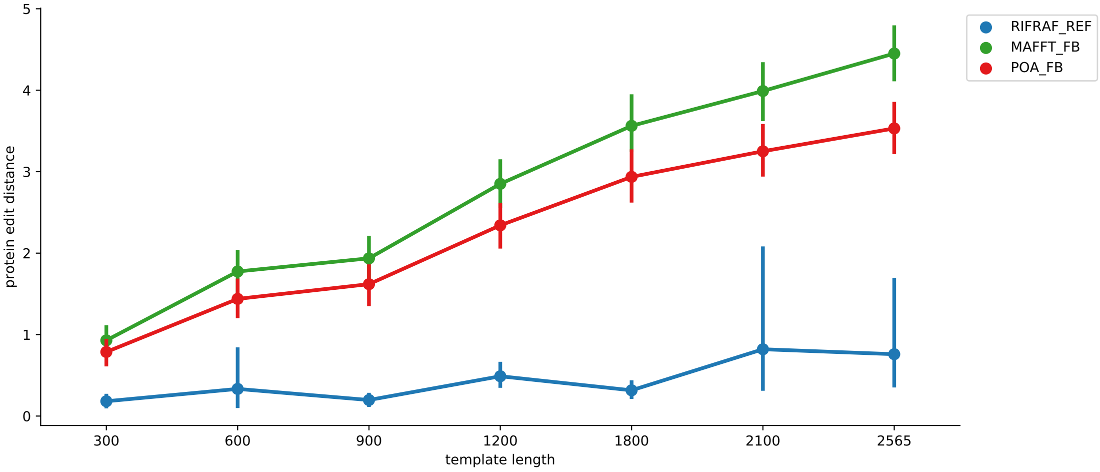
(a) Fraction of correct sequences versus template length.

(b) DNA edit distance versus template length.

Figure 3.11: Results on DNA sequences for varying sequence length. The left figure shows the fraction of correct sequences; the right figure shows mean edit distance between the consensus and the template.

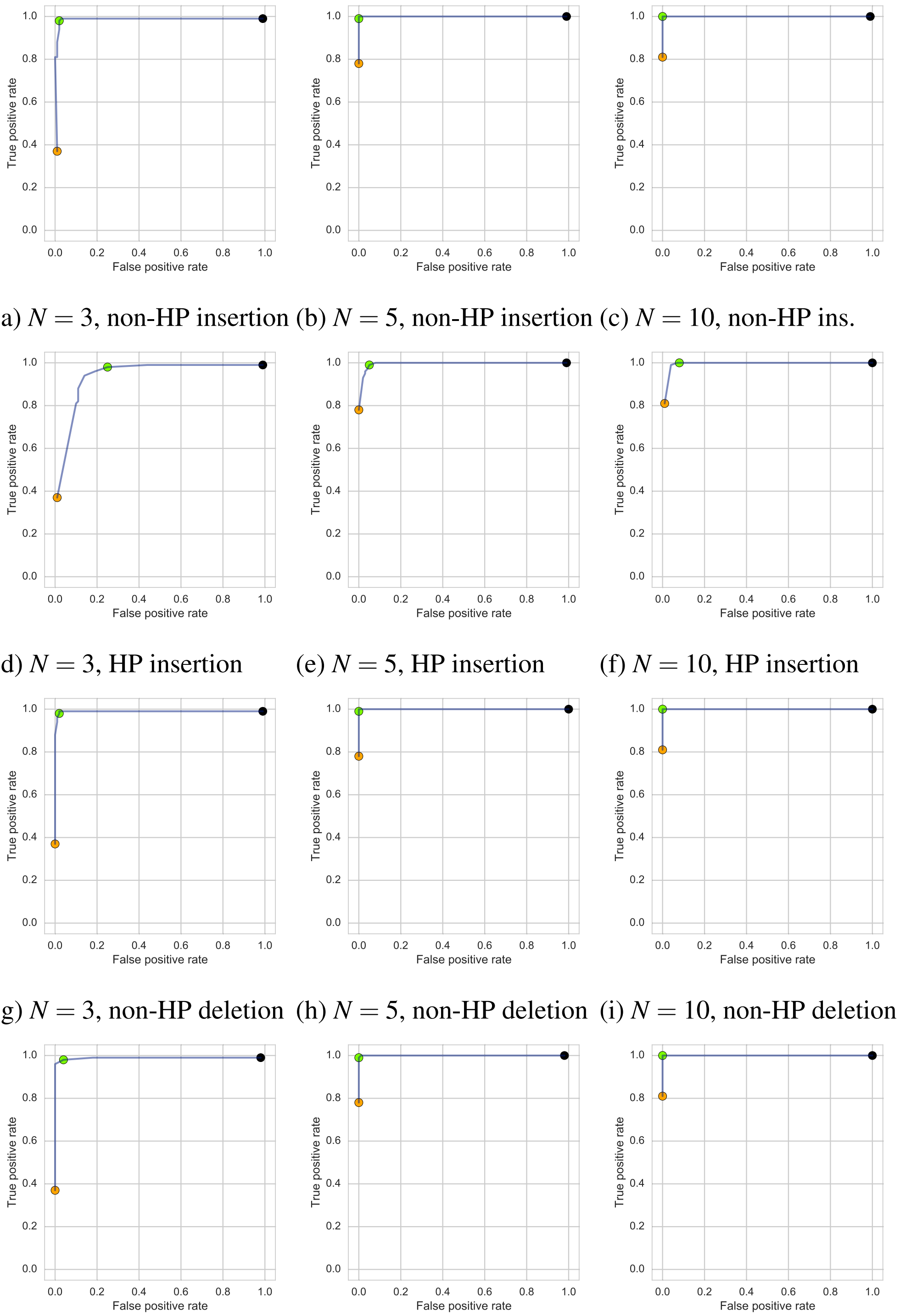


(a) Protein edit distance versus template length; all methods.



(b) Protein edit distance versus template length; frame correction only

Figure 3.12: Same results as Fig. 3.11, but for the translated protein sequences. The fraction of correct sequences is not reproduced, since those figures are identical. The left figure show results for all methods. The right figure shows the same data, zoomed in on the details of the frame-corrected results.



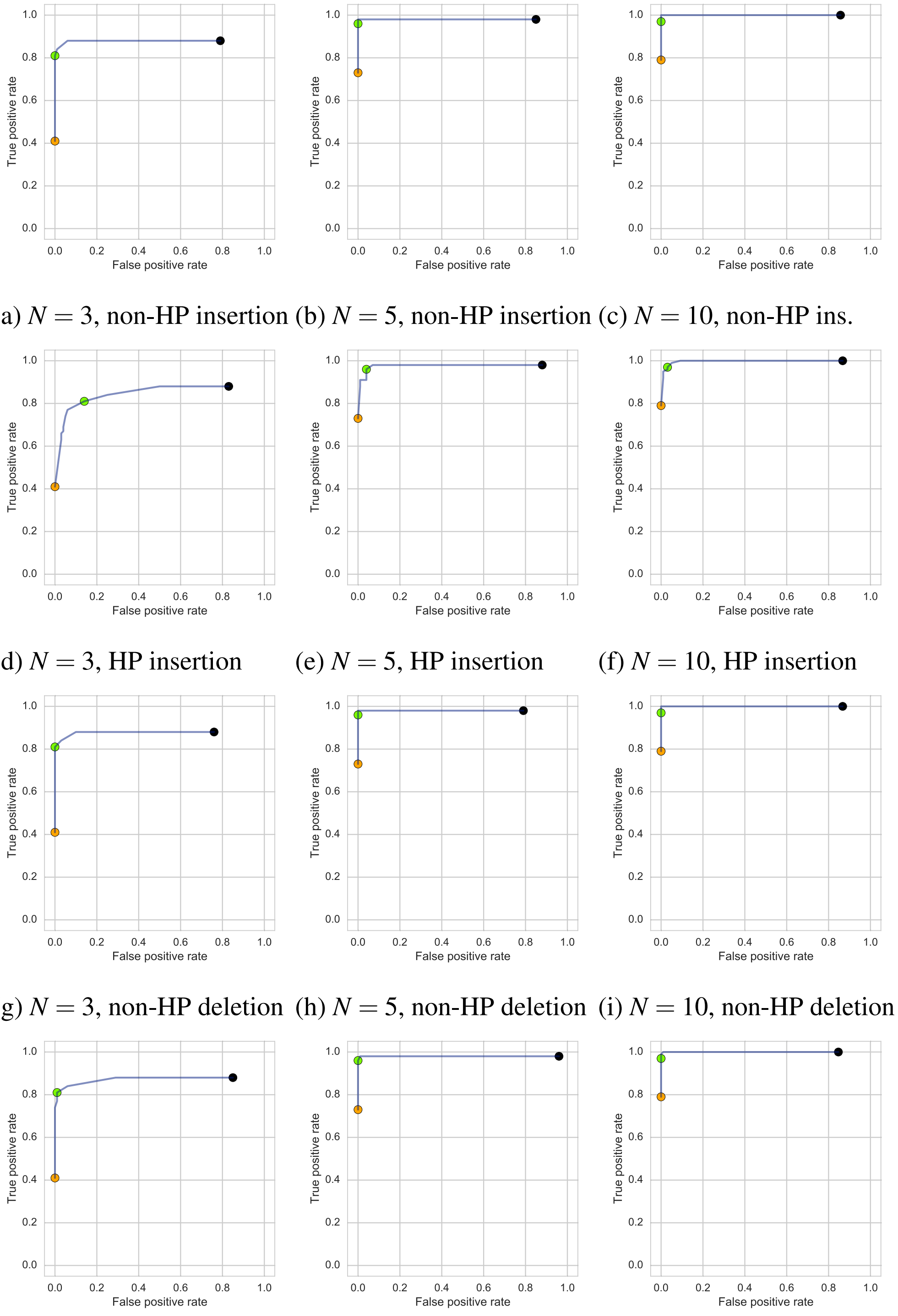
(a) $N = 3$, non-HP insertion (b) $N = 5$, non-HP insertion (c) $N = 10$, non-HP ins.

(d) $N = 3$, HP insertion (e) $N = 5$, HP insertion (f) $N = 10$, HP insertion

(g) $N = 3$, non-HP deletion (h) $N = 5$, non-HP deletion (i) $N = 10$, non-HP deletion

(j) $N = 3$, HP deletion (k) $N = 5$, HP deletion (l) $N = 10$, HP deletion

Figure 3.13: ROC curves for true indel experiments, with max indel penalty multiplier m from 0 to 12. Orange points denote runs without a reference; the rest use HXB2 as a reference. The green point corresponds to a max indel penalty multiplier $m = 6$.



(a) $N = 3$, non-HP insertion (b) $N = 5$, non-HP insertion (c) $N = 10$, non-HP ins.

(d) $N = 3$, HP insertion (e) $N = 5$, HP insertion (f) $N = 10$, HP insertion

(g) $N = 3$, non-HP deletion (h) $N = 5$, non-HP deletion (i) $N = 10$, non-HP deletion

(j) $N = 3$, HP deletion (k) $N = 5$, HP deletion (l) $N = 10$, HP deletion

Figure 3.14: The same results as in Figure 3.13, except using the more distant sequence B.BR as a reference.

Chapter 4

Biological applications

4.1 Introduction

The FLEA pipeline has already been used in multiple papers to provide new biological insights into the evolution of HIV-1. It has been used to characterize the mutations in *env* that potentially drive the development of lineages of broadly neutralizing antibodies in the Protocol C cohort. It has also been used to study evolution and escape during a phase I clinical trial of monoclonal antibody 10-1074.

4.2 Protocol C

The Protocol C cohort is a group of 439 HIV-positive subjects from multiple regions in sub-Saharan Africa that participated in a multi-year longitudinal study to investigate the factors leading to the development of broadly neutralizing antibodies [64]. HIV-negative individuals were monitored and enrolled in the study after they contracted HIV, ensuring that data collection started as soon as possible after the initial infection event. Blood draws were taken at regular intervals and neutralization assays were performed against a panel of heterologous virus. Consistent with other estimates of bNAb frequency, 15% of participants developed bNAbs, most after two to four years of infection. A summary of their neutralization scores over time appears in Figure 4.1. Another third of the participants developed some neutralization breadth, but not enough to be characterized as broad. Among those that did develop bNAbs, 40% of them targeted the N332 glycan supersite in V3.

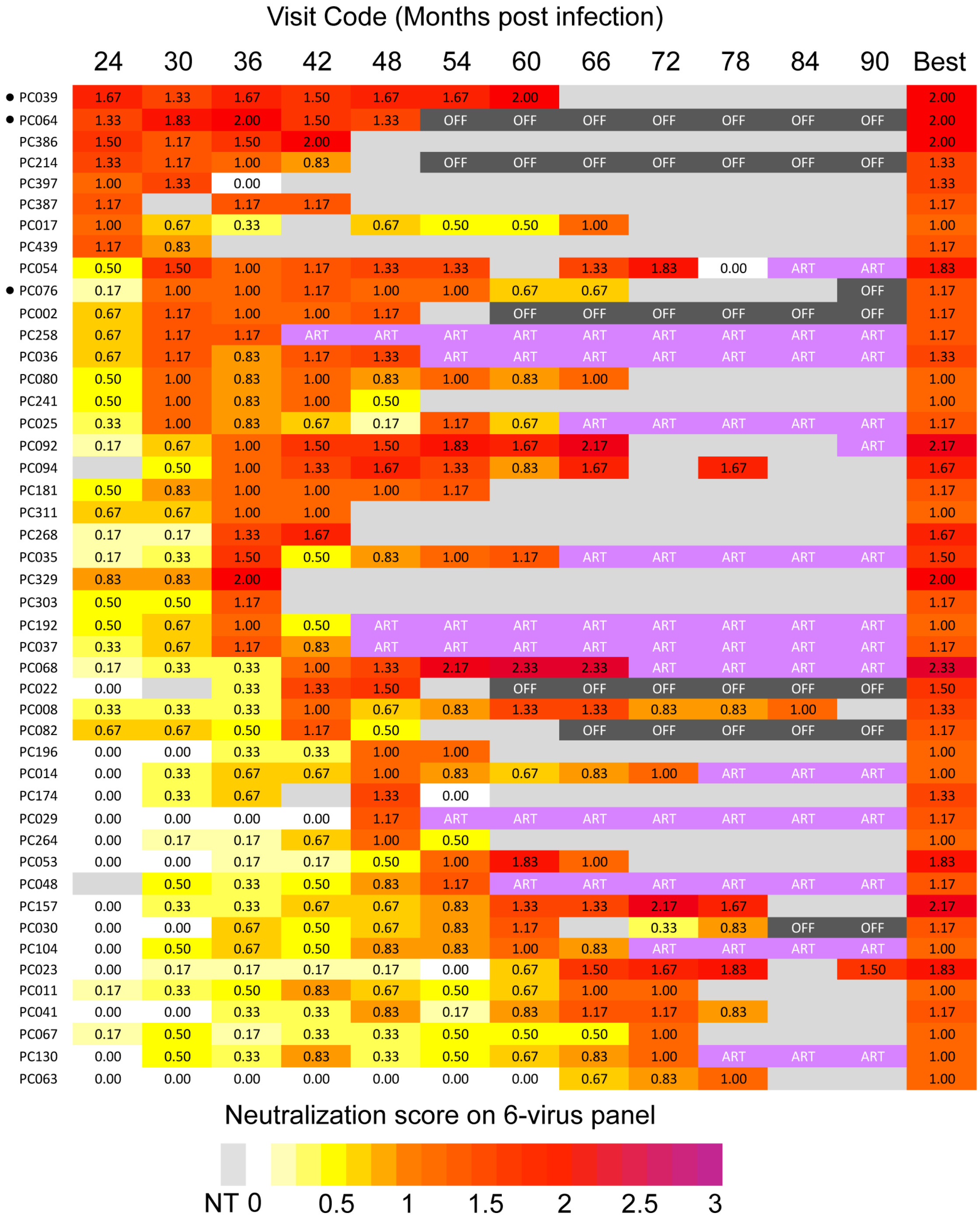
This study emphasized the need to perform in-depth followup studies in order to understand how to elicit bNAbs:

> A detailed analysis of the development of bNAb lineages in top neutralizers will help understand which specificities are most amenable to elicitation through vaccination and whether Env evolution pathways associated with specific lineages suggest particular immunogen designs or vaccine strategies. ([64])

Those in-depth analyses are currently underway. Of the top forty-six donors, PC039, PC064, and PC076 (marked in Figure 4.1) have already been studied with a combination of strategies such as neutralization assays, longitudinal sequencing of both the B-cell lineage and the Env population, structure determination of antibodies, and some Env/Ab complexes. In all three cases, `FLEA` was used to analyze the longitudinal Env sequences. This section highlights the major biological insights from each of those donors, with an emphasis on the contributions of `FLEA` to the results.

4.2.1 PC039

Donor PC039 developed antibody lineages targeting the N332 glycan supersite in variable loop 3 (V3), which was the most common target observed in the Protocol C cohort. The donor developed a maximum neutralization score of 2.0 at 60 months post infection. This donor is also notable because they were infected with two founder variants, which recombined and then evolved into two stable sub-populations, each


Visit Code (Months post infection)
Neutralization score on 6-virus panel
NT 0
0.5
1
1.5
2
2.5
3

	24	30	36	42	48	54	60	66	72	78	84	90	Best
● PC039	1.67	1.33	1.67	1.50	1.67	1.67	2.00						2.00
● PC064	1.33	1.83	2.00	1.50	1.33	OFF	OFF	OFF	OFF	OFF	OFF	OFF	2.00
PC386	1.50	1.17	1.50	2.00									2.00
PC214	1.33	1.17	1.00	0.83		OFF	OFF	OFF	OFF	OFF	OFF	OFF	1.33
PC397	1.00	1.33	0.00										1.33
PC387	1.17		1.17	1.17									1.17
PC017	1.00	0.67	0.33		0.67	0.50	0.50	1.00					1.00
PC439	1.17	0.83											1.17
PC054	0.50	1.50	1.00	1.17	1.33	1.33		1.33	1.83	0.00	ART	ART	1.83
● PC076	0.17	1.00	1.00	1.17	1.00	1.00	0.67	0.67				OFF	1.17
PC002	0.67	1.17	1.00	1.00	1.17		OFF	OFF	OFF	OFF	OFF	OFF	1.17
PC258	0.67	1.17	1.17	ART	ART	ART	ART	ART	ART	ART	ART	ART	1.17
PC036	0.67	1.17	0.83	1.17	1.33	ART	ART	ART	ART	ART	ART	ART	1.33
PC080	0.50	1.00	0.83	1.00	0.83	1.00	0.83	1.00					1.00
PC241	0.50	1.00	0.83	1.00	0.50								1.00
PC025	0.33	1.00	0.83	0.67	0.17	1.17	0.67	ART	ART	ART	ART	ART	1.17
PC092	0.17	0.67	1.00	1.50	1.50	1.83	1.67	2.17				ART	2.17
PC094		0.50	1.00	1.33	1.67	1.33	0.83	1.67		1.67			1.67
PC181	0.50	0.83	1.00	1.00	1.00	1.17							1.17
PC311	0.67	0.67	1.00	1.00									1.00
PC268	0.17	0.17	1.33	1.67									1.67
PC035	0.17	0.33	1.50	0.50	0.83	1.00	1.17	ART	ART	ART	ART	ART	1.50
PC329	0.83	0.83	2.00										2.00
PC303	0.50	0.50	1.17										1.17
PC192	0.50	0.67	1.00	0.50	ART	ART	ART	ART	ART	ART	ART	ART	1.00
PC037	0.33	0.67	1.17	0.83	ART	ART	ART	ART	ART	ART	ART	ART	1.17
PC068	0.17	0.33	0.33	1.00	1.33	2.17	2.33	2.33	ART	ART	ART	ART	2.33
PC022	0.00		0.33	1.33	1.50		OFF	OFF	OFF	OFF	OFF	OFF	1.50
PC008	0.33	0.33	0.33	1.00	0.67	0.83	1.33	1.33	0.83	0.83	1.00		1.33
PC082	0.67	0.67	0.50	1.17	0.50			OFF	OFF	OFF	OFF	OFF	1.17
PC196	0.00	0.00	0.33	0.33	1.00	1.00							1.00
PC014	0.00	0.33	0.67	0.67	1.00	0.83	0.67	0.83	1.00	ART	ART	ART	1.00
PC174	0.00	0.33	0.67		1.33	0.00							1.33
PC029	0.00	0.00	0.00	0.00	1.17	ART	ART	ART	ART	ART	ART	ART	1.17
PC264	0.00	0.17	0.17	0.67	1.00	0.50							1.00
PC053	0.00	0.00	0.17	0.17	0.50	1.00	1.83	1.00					1.83
PC048		0.50	0.33	0.50	0.83	1.17	ART	ART	ART	ART	ART	ART	1.17
PC157	0.00	0.33	0.33	0.67	0.67	0.83	1.33	1.33	2.17	1.67			2.17
PC030	0.00	0.00	0.67	0.50	0.67	0.83	1.17		0.33	0.83	OFF	OFF	1.17
PC104	0.00	0.50	0.67	0.50	0.83	0.83	1.00	0.83	ART	ART	ART	ART	1.00
PC023	0.00	0.17	0.17	0.17	0.17	0.00	0.67	1.50	1.67	1.83		1.50	1.83
PC011	0.17	0.33	0.50	0.83	0.67	0.50	0.67	1.00	1.00				1.00
PC041	0.00	0.00	0.33	0.33	0.83	0.17	0.83	1.17	1.17	0.83			1.17
PC067	0.17	0.50	0.17	0.33	0.33	0.50	0.50	0.50	1.00				1.00
PC130	0.00	0.50	0.33	0.83	0.33	0.50	0.67	0.83	1.00	ART	ART	ART	1.00
PC063	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.83	1.00			1.00

Figure 4.1: Best neutralizers from the Protocol C cohort. The 46 best neutralizers had neutralization score ≥ 1. Neutralization score shown for each month. NT: Not Tested. ART: on ART during visit. OFF: off-study during this visit. The three donors for which FLEA was used during followup studies (PC039, PC064, PC076) are marked. Image adapted from Figure 1C [64].

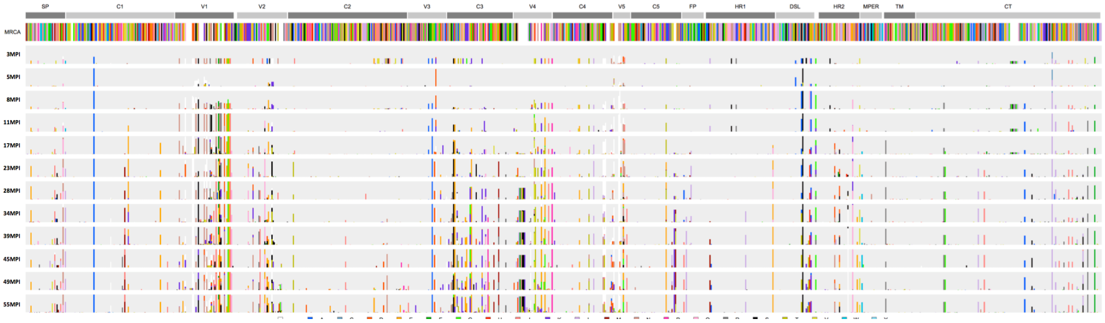


Figure 4.2: Histogram of amino acid frequencies in PC039 for each time point. Only differences from one of the parent MRCA sequences are shown. The other parent is clearly visible at 3MPI. From Figure 1 [95].

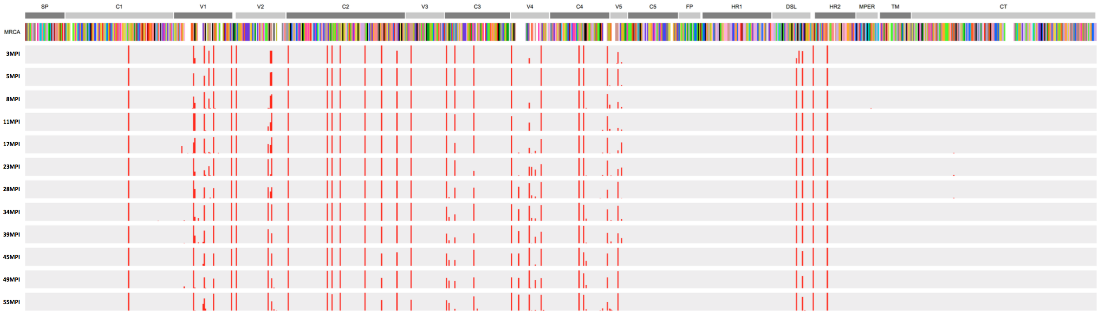


Figure 4.3: PNGS locations in PC039. Fraction of PNGS sites at each position. From Figure 2 [95].

of which separately escaped the antibody lineage. The early presence of at least two founders, followed by recombination, may possibly have contributed to the antibody lineage development.

The full-length *env* sequencing protocol was used to sequence the population at twelve time points, from three to sixty months post infection. `FLEA` was used to generate HQCSs, align them, and perform phylogentic analyses. Per-site amino acid frequences (Figure 4.2) clearly show the presence of both founder variants at 3 MPI, which also appear in the phylogenetic tree in Figure 4.5.

These multiple founders, plus the high recombination rate in HIV-1, means that alignment columns do not share a common ancestry, which is an assumption of most phylogenetic analyses. Therefore, the inferred tree does not capture the true within-host

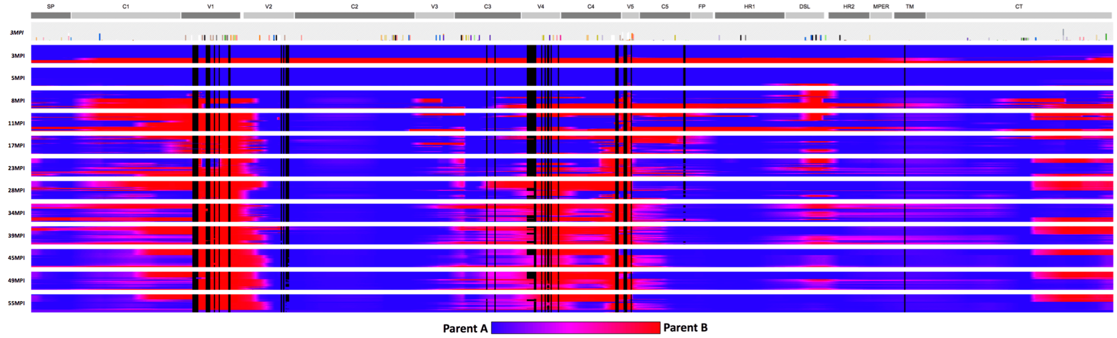


Figure 4.4: Inferred recombination in PC039. Posterior probability that each base came from parent A (blue) or parent B (red). Little recombination is visible at 3 MPI. Sequences become increasingly recombined over time. Black regions denote gaps in the alignment. From Figure 3 [95].

evolutionary history of the virus [122, 125]. To better understand these lineages, we also visualized the population in a different way by using Multidimensional Scaling (MDS) [60] to embed the sequences in two-dimensional space while preserving the pairwise distances. The embedding clearly shows the branching and independent evolution of two sub-populations of Env (Figure 4.6). A stable mixture of two sub-populations of recombinants appeared by 17 MPI, which had already escaped. Subsequently each population proceeded along different evolutionary pathways. This data was inspiration for the MDS visualization in FLEA, as shown in Figure 2.4.

A previously-described recombination model [80] was implemented and customized to infer the amount of recombination in each HQCS. The two founder variants were inferred from the earliest time point and used to build a hidden Markov model (Figure 4.7), keeping only the positions that differed in at least one parent. The model was intilized with constant transition and emission matrices, then trained using Viterbi training [114] with the constraint that the transition matrix must be symmetric with a constant diagonal. The forward and backward algorithms were used to compute the posterior probability that each base came from each parent. The training and inference process was repeated independently for each sequence. A visualization of the posterior

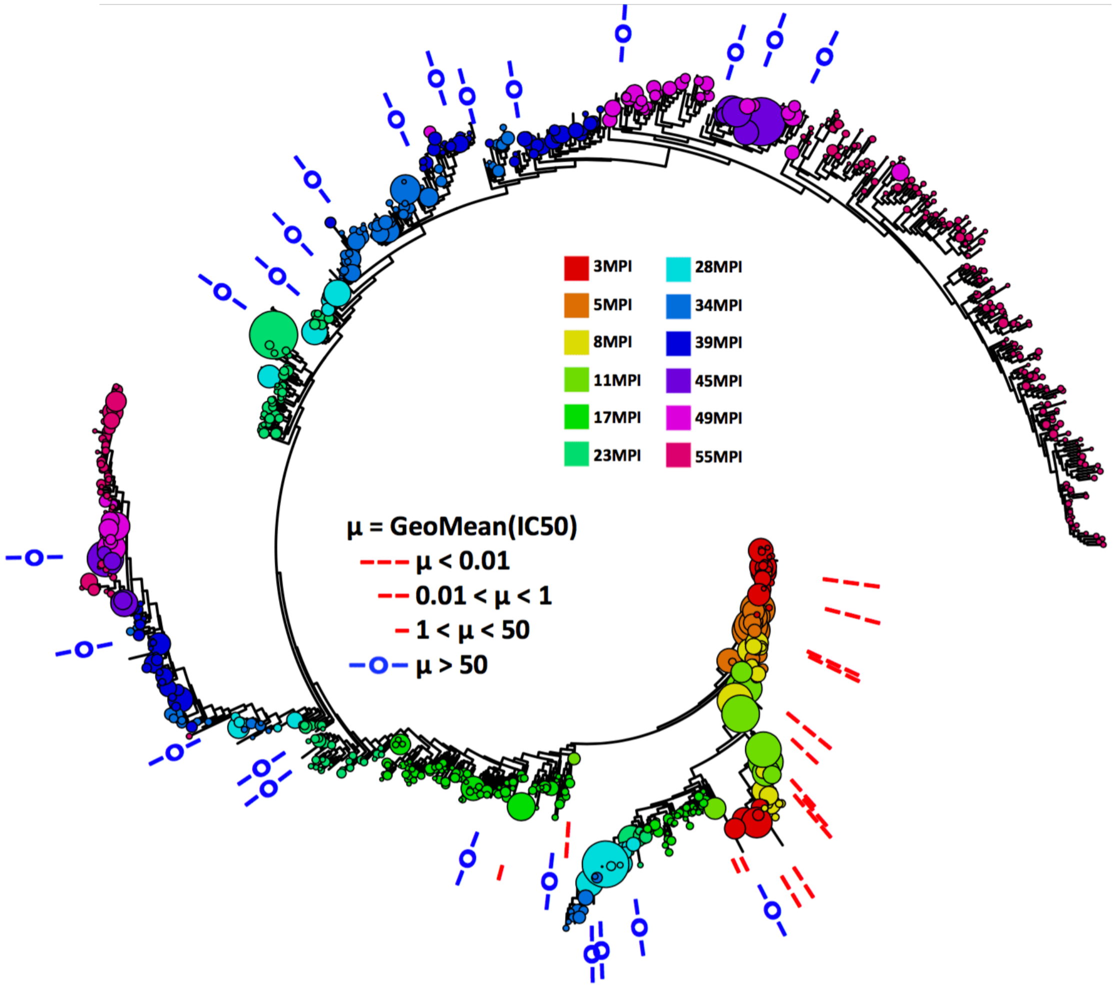


Figure 4.5: Phylogenetic tree of Env in PC039, with nodes colored according to time point. The diverging ladder structure is due to recombination and diverging sub-populations. Marked IC50 values show escape at later time points. From Figure 4A [95].

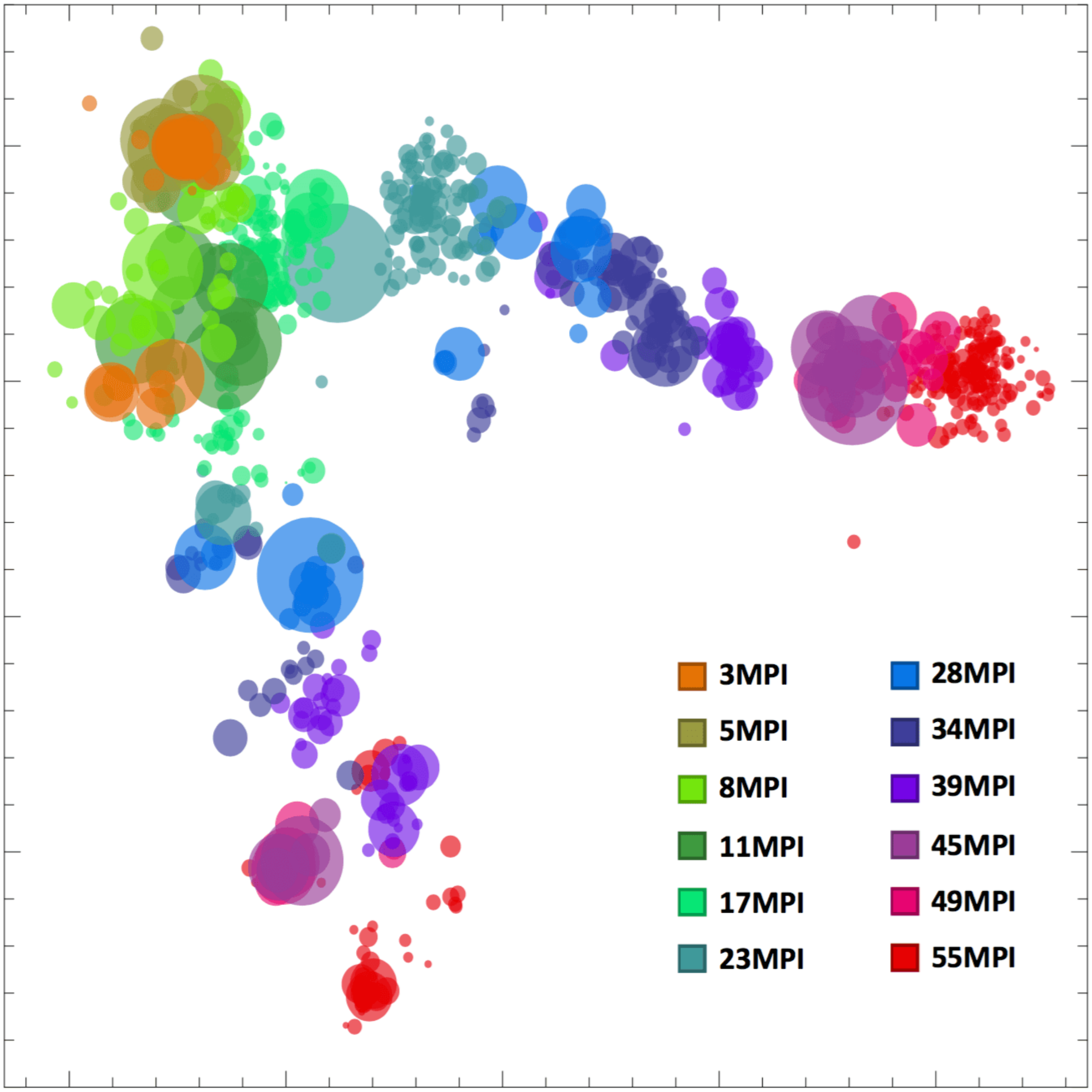


Figure 4.6: MDS embedding of Env in PC039. Multidimensional scaling (MDS) was used to generate a low-dimensional embedding of Env HQCSs sequences that preserves pairwise distances. Node color shows time since infection, and node size is proportional to inferred abundance. The evolution of two independent populations of Env is clearly visible. The populations derived from the two founder sequences are clearly visible (orange). They recombine, then split into separate lineages around 23 MPI. From Figure 4C [95].

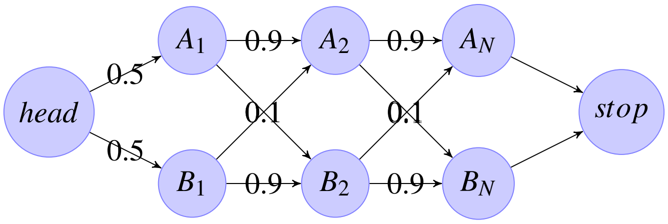


Figure 4.7: HMM for inferring recombination of two parent sequences to generate an observed child sequence. Each column in their alignment corresponds to a state from each parent. Transitioning to the other parent models a recombination event. The shown probabilities are meant to be representative, as the true probabilities differ for each observed sequence after Viterbi training.

probabilities for all HQCSs is shown in Figure 4.4.

The two Env lineages contained many genetic differences, such as the 322-323 motif, as shown in Figure 4.9, the 324-327 motif, as shown in Figure 4.8, the location of glycans (Figures 4.10 and 4.3) and the composition of the V1 loop. In particular, the deletion observed at position 322 has been functionally validated by mutagenesis and neutralization assay, and explains escape from the PC039 antibody lineage for the top arm of the MDS plot.

This work was presented by Ben Murrell in 2016 at Keystone [95]. Further work is ongoing.

4.2.2 PC076

Donor PC076, who was infected by HIV-1 subtype C, also developed a bNAb lineage targeting the N332 region, here referred to as the high-mannose patch. This donor's lineage is notable because at least one of its antibodies achieved breadth without a large number of complex changes: it had a small number of somatic hypermutations and no insertions or deletions. This result has implications for vaccine design, because it should be faster and easier to induce such an antibody, compared to one that requires an older, more complicated lineage. Presumably, the longer it took during natural infection

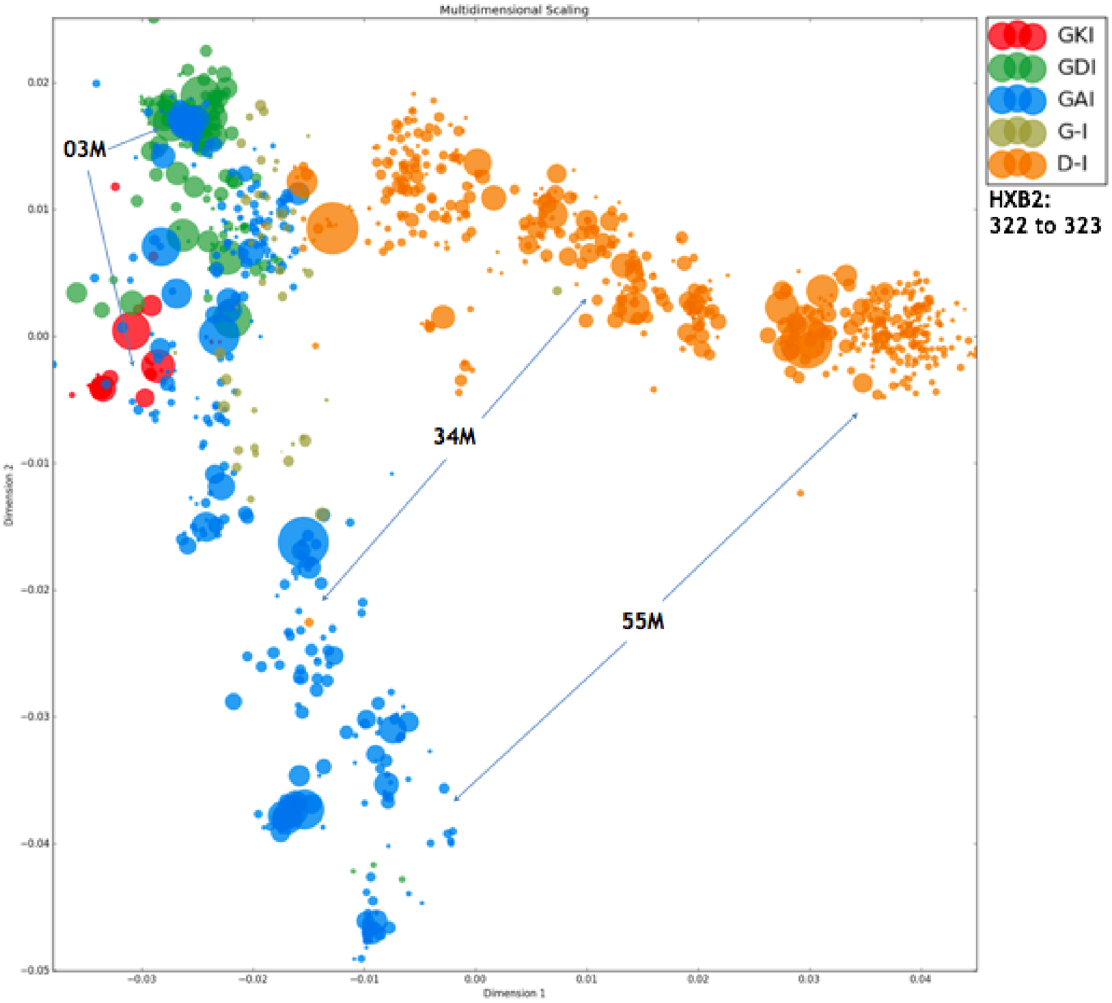


Figure 4.8: [MDS embedding showing positions 332 and 333 in PC039. From an ongoing collaboration.

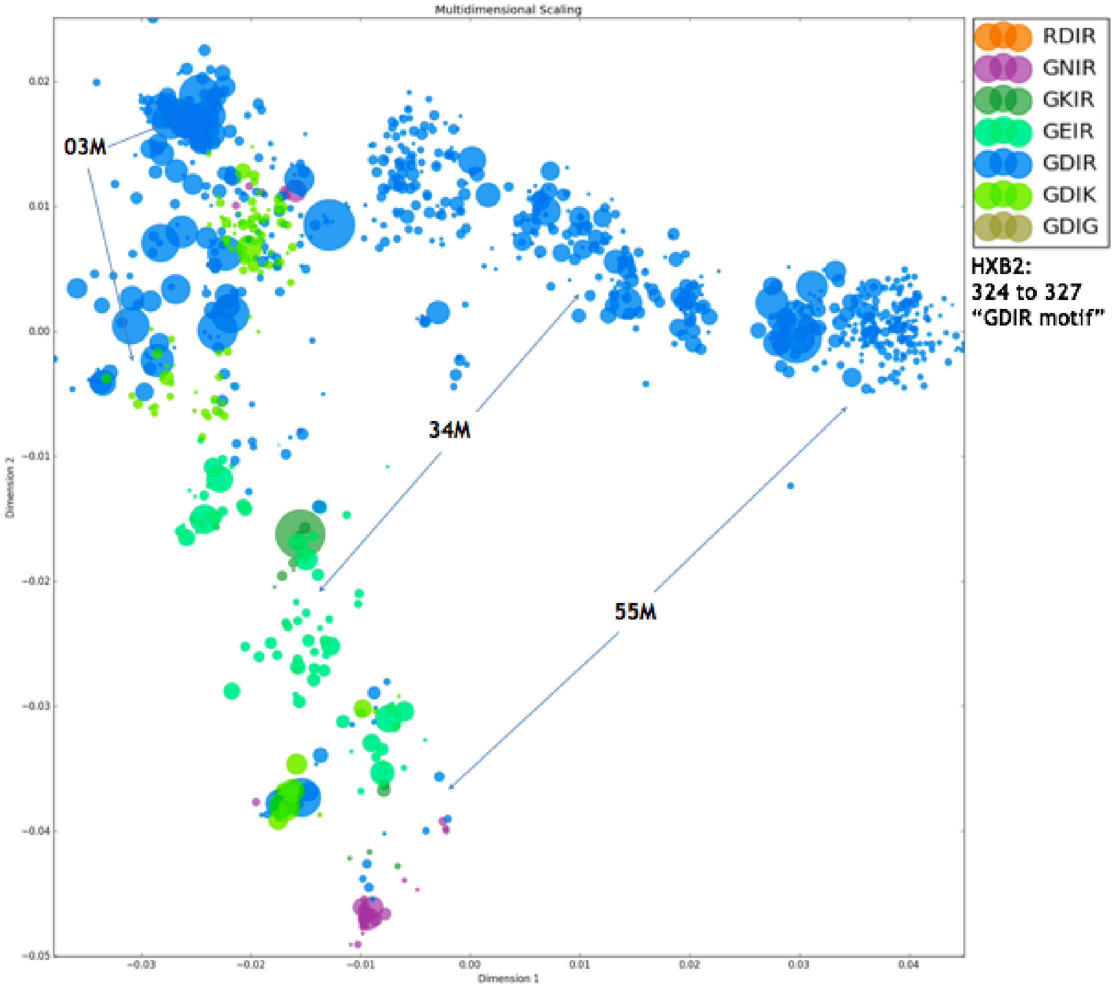


Figure 4.9: MDS embedding showing positions 324-327 "GDIR" motif in PC039. From an ongoing collaboration.

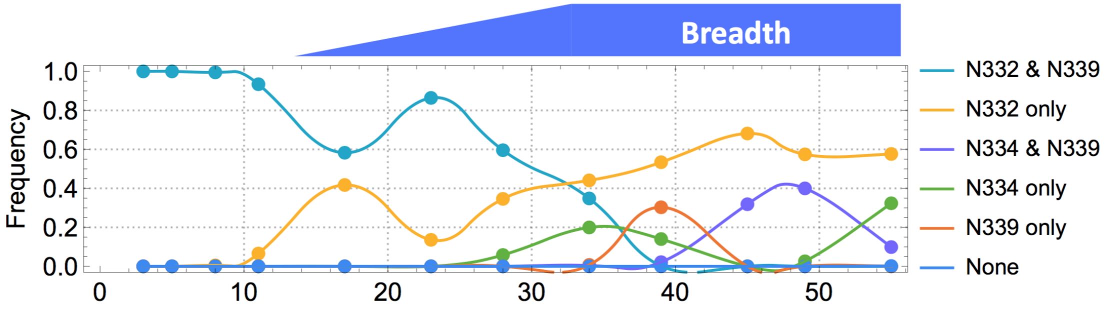


Figure 4.10: Glycan dynamics in PC039, and corresponding development of breadth. Both lineages had almost entirely escaped by 17 MPI. From [95].

for a bNAb to develop, the more rounds of sequential vaccination would be required, if it is possible at all.

The investigation proceeded as follows:

- *Confirm a broadly neutralizing response and find its target on Env.* Neutralization assays detected a broad neutralizing response at 33 MPI and N332 activity. They further confirmed that the response is N332 dependent by checking that a point mutation which eliminates the glycan (N332A) is a neutralization knockout.

- *Find the responsible mAbs.* Twelve potent monoclonal antibodies (mAbs) were isolated from the blood samples and sequenced via single-cell sequencing. Individual antibodies were tested against the neutralization panel of pseudoviruses, and the three most potent were further investigated on another panel.

- *Study Env population to find the variants that triggered the Ab lineage, mutations that drove the lineage evolution, and eventual escape mutations.* 76 autologous full-length *env* sequences were cloned and sequenced via Sanger sequencing. The capacity of each of the 12 mAbs (from the previous step) to neutralize these suggested that the lineage was triggered by virus that emerged between 5 and 10 months. The Env sequences from those time points were checked for common mutations. Mutations were confirmed by introducing them to the most common virus and again checking for neutralization.

- *Track mAb lineages that lead to the potent mAbs.* Deep sequencing was used to track the mAb lineages leading to the bNAbs studied in step #2. Multiple early lineage arms developed in parallel.

- *Get Fab structure.* In order to better understand binding and possible mechanism of neutralization, the structure of the Fab was resolved.

An early version of FLEA was used to align and analyze the Env sequences. Because sequences came from Sanger sequencing, not SMRT sequencing, the quality and consensus sub-pipelines were not necessary. Instead, the Sanger sequences were treated like HQCSs and fed directly into the alignment and analysis sub-pipelines. The results of those analyses helped to identify residues in Env that contributed to the evolution of the lineage. Those mutations could then be verified via mutagenesis.

This work was published in [75].

4.2.3 PC064

Donor PC064, who was infected by HIV-1 subtype A, developed a bNAb lineage targeting the V2 apex epitope. Fewer donors in the Protocol C cohort developed antibodies targeting V2, which is consistent with the estimated rate of 10-25% among those that do develop bNAbs [49]. Like the lineage from PC076, this lineage contained a low amount of somatic hypermutation and no indels, making it a realistic candidate for a vaccine.

Full-length longitudinal SMRT sequencing of *env* was performed, and the results were analyzed with FLEA. The inferred tree appears in Figure 4.11.

The HQCSs reconstructed by FLEA revealed key insights into the evolution of Env that could lead to a method to elicit V2-targeting antibody lineages. They showed the V2 variants that drove the development of breadth and the eventual escape trajectory, which occurred via mutations at positions 166, 167 and 169. The pattern of these mutations over time suggests that the mAb lineage evolved towards greater breadth as a response to successive escape mutations at these positions. The HQCSs also show that the Env population continued to evolve after full escape. These further changes are possibly restoring fitness lost during the escape.

This work was published in [65].

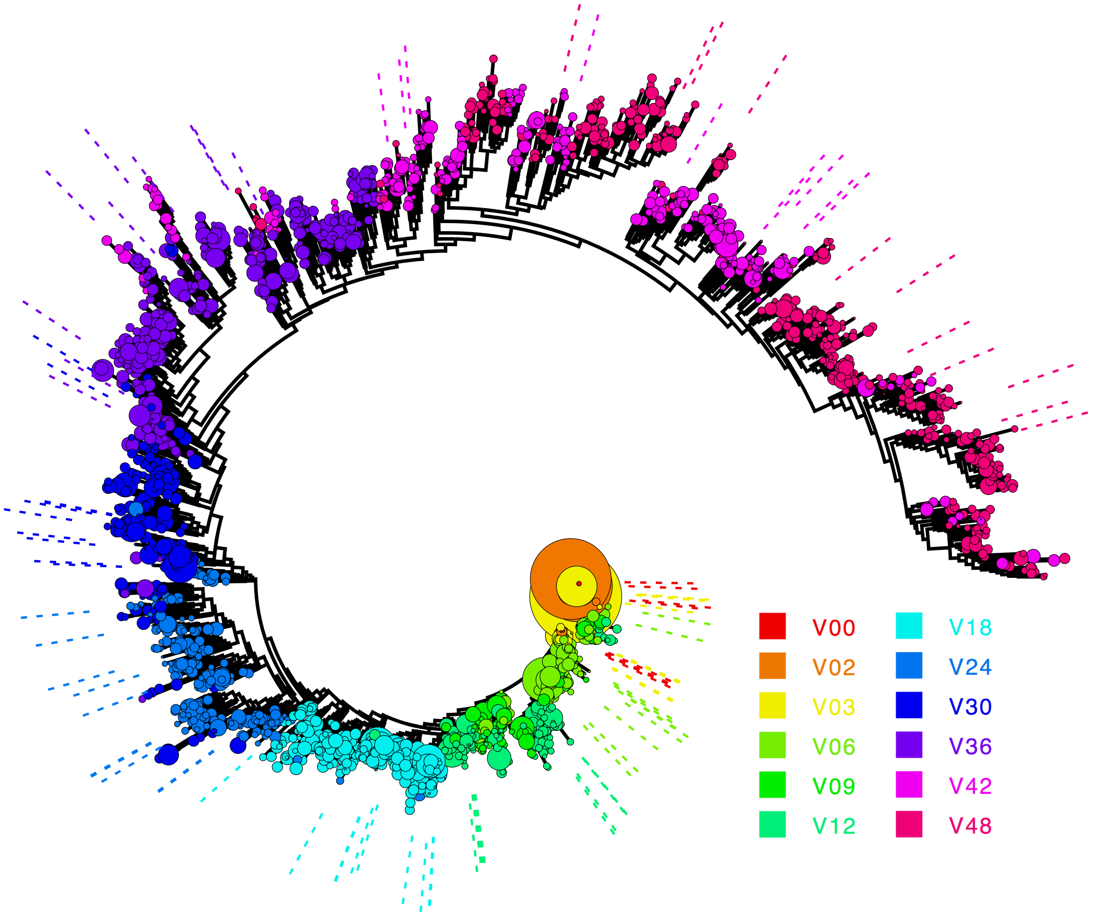


Figure 4.11: Phylogenetic tree of Env in PC064. Node size is proportional to abundances, and colors correspond to time point. Dotted lines show location of Sanger sequences, which agree with the HQCSs inferred from the SMRT sequencing reads. From Figure 4[63].

4.3 Phase I clinical trial of antibody 10-1074

FLEA was also used as part of a recent study describing a phase I clinical trial of a broadly-neutralizing monoclonal antibody [14]. As an alternative to eliciting antibodies through vaccination, another approach is to manufacture monoclonal antibodies originally isolated from donor studies (such as those described above), and directly inject them as an antibody therapy. They confer the same neutralization benefits, both in preventing infection and in controlling viremia during infection, so this approach is being considered as an alternative or supplement to both vaccines and ART. Unlike native antibodies, however, passively introduced antibodies get depleted and would have to be replenished.

This antibody used in this clinical trial, 10-1074, was isolated from an African donor described in [90]. Like the lineage in PC076 [75], this antibody also targets the V3 epitope centered on N332. 10-1074 is one of the more potent and broad monoclonal antibodies found so far. It neutralized 60.5% of pseudovirus in a large panel of neutralization assays and 77.7% of isolates from HIV-1 infected individuals from the US and Germany. In this study, 33 individuals (14 uninfected, 16 infected and off ART, 3 infected and on ART) received a single dose of varying concentration to assess its medical safety and its efficacy as a therapy for controlling viremia.

The therapy did successfully reduce viremia for a short time (approximately ten days) before the viral population evolved escape mutations and viremia rebounded. A variety of sequencing protocols were used to study this escape. Single genome sequencing (SGS) was used to sequence Env before and after treatment. It confirmed sensitivity to antibody before treatment and showed that most sequences escaped via N332 or S334 mutations that removed a PNGS. However, its depth was too low to look for minority variants: a total of 1,111 Env sequences were acquired for 15 subjects, for an average of 37 sequences per subject per time point. Primer ID based deep sequencing (PIDS) of

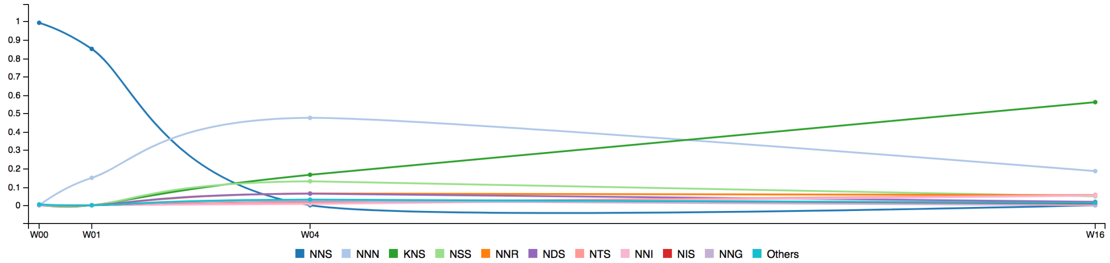


Figure 4.12: Escape in donor 1HD1. Multiple loss of glycan mutations appeared by week four, then 325K took over by week sixteen. This suggests that the final escape mutant was present at low frequencies and took longer to take over the population. Unpublished data from ongoing work with Ben Murrell.

the V3 region was also done for fifteen subjects, which could identify minority variants, since it was powered to detect mutations at 1.0% frequency, with a range of 0.5% to 2.4%). However, by focusing on only the V3 region, PIDS missed the other changes occurring in the rest of *env*. Finally, full-length SMRT sequencing of *env* in three subjects was performed, and analyzed with `FLEA`. Note that the amino acid frequencies inferred by `FLEA` agreed with the SGS and PIDS frequencies, which independently confirms `FLEA`'s accuracy.

The phylogenies inferred by `FLEA` showed that two of the three subjects already had Env populations with multiple escape variants at low frequencies prior to initiation of therapy. The results also showed that all three viral populations had escaped by the fourth week of the trial via mutations at the same few positions in V3: positions 324-327 (the ”GDIR” motif) and the PNGS at 332-334. New, unpublished longitudinal data from donor 1HD1 appears in Figure 4.12.

The aligned full-length sequences from `FLEA` also suggested that the escaped variants were still vulnerable to bNAbs targeting other epitopes. This vulnerability provides support for the idea that, like HAART, vaccines and monoclonal antibody therapies will need to prevent escape by targeting multiple Env epitopes simultaneously.

This work was published in [14].

4.4 Acknowledgements

Chapter 4 contains, in part, material that appeared in the poster “Full-length *env* deep sequencing in a donor with broadly neutralizing N332 antibodies”. Ben Murrell, Kemal Eren, Lorena S Ver, Nancy Choi, Elise Landais, Pascal Poignard, Sergei L Kosakovsky Pond, and Davey Smith, *Keystone Symposia on Molecular and Cellular Biology* 2016. The dissertation author was an author of this poster.

Chapter 5

Conclusion

During my doctoral program I developed new tools and algorithms to investigate the evolution of diverse populations of HIV-1 *env* using longitudinal SMRT sequencing. The first is `FLEA`, a pipeline for aligning, analyzing, and visualizing Env sequences. `FLEA` has already been used to study donors that developed broadly-neutralizing antibodies and to study the effects of a monoclonal antibody in a phase 1 clinical trial. The second is `RIFRAF`, a consensus algorithm that takes advantage of quality scores and that keeps the consensus sequence in frame. In addition to these projects, I also contributed to BUSTED, a new method for identifying episodic positive selection [94]

Both `FLEA` and `RIFRAF` are undergoing active development. The process of integrating `RIFRAF` into the pipeline and using the modified pipeline for new projects is currently underway. Just like `RIFRAF` made `FLEA` more accurate by replacing an off-the-shelf consensus algorithm for consensus sequences with one designed specifically for the problem at hand, improvements for other parts of the pipeline, such as clustering, are currently being developed. Moreover, although it has only been used for HIV-1 Env, `FLEA` is being adapted to other proteins in HIV-1 and even to amplicons from other pathogens. Its reference databases, error models, and parameters would need to be updated appropriately,

One outstanding question for FLEA, and more generally for the analysis of any recombining virus, is how to deal with such recombination. HIV-1 recombines at a high rate, which adds an extra layer of difficulty to the problem of reconstructing its evolutionary history and invalidates the assumptions underpinning methods such as phylogeny reconstruction. Methods exist for identifying recombination [11, 79], but currently there are no good ways to accurately reconstruct phylogenies of recombinant sequences. We plan to first address this issue by inferring recombination in every time point using the HMM model mentioned in the previous chapter and visualizing the results.

Finally, FLEA is being used to analyze breakthrough infections in two vaccination studies, involving SIV and SHIV (SIV with an HIV Env) challenge of immunized macaques. Preliminary data from the SIV challenge suggest a sieve effect, where low-frequency pre-existing escape mutants are crossing the transmission barrier and seeding the infection. This shows that FLEA is useful not just to study primary infection, but also for translational vaccine research.

Bibliography

[1] Luis M Agosto, Pradeep D Uchil, and Walther Mothes. HIV cell-to-cell transmission: Effects on pathogenesis and antiretroviral therapy. In *Trends in Microbiology*, volume 23, pages 289–295. 2015.

[2] Aikaterini Alexaki, Yujie Liu, and Brian Wigdahl. Cellular Reservoirs of HIV-1 and their Role in Viral Persistence. *Current HIV Research*, 6(5):388–400, sep 2008.

[3] Suzanna Attia, Matthias Egger, Monika Müller, Marcel Zwahlen, and Nicola Low. Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis. *Aids*, 23(11):1397–1404, 2009.

[4] Françoise Barré-Sinoussi, Anna Laura Ross, and Jean-François Delfraissy. Past, present and future: 30 years of HIV research. *Nature reviews. Microbiology*, 11 (12):877, 2013.

[5] Niko Beerenwinkel and Osvaldo Zagordi. Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology*, 1(5):413–418, 2011.

[6] Edward A Berger, Philip M Murphy, and Joshua M Farber. Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annual review of immunology*, 17(1):657–700, 1999.

[7] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic acids research*, 28 (1):235–242, 2000.

[8] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[9] Scott D. Boyd and James E. Crowe Jr. Deep sequencing and human antibody repertoire analysis. *Current Opinion in Immunology*, 8(5):583–592, 2016.

[10] Christian Brander, Nicole Frahm, and Bruce D Walker. The challenges of host and viral diversity in HIV vaccine design. *Current opinion in immunology*, 18(4): 430–437, 2006.

[11] Trevor C. Bruen, Hervé Philippe, and David Bryant. A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172(4):2665–2681, 2006.

[12] Rowena A Bull, Fabio Luciani, Kerensa McElroy, Silvana Gaudieri, Son T Pham, Abha Chopra, Barbara Cameron, Lisa Maher, Gregory J Dore, Peter A White, and Andrew R Lloyd. Sequential bottlenecks drive viral evolution in early acute hepatitis c virus infection. *PLoS Pathogens*, 7(9), 2011.

[13] M R Capobianchi, E Giombini, and G Rozera. Next-generation sequencing technology in clinical virology. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 19(1):15–22, 2013.

[14] Marina Caskey, Till Schoofs, Henning Gruell, Allison Settler, Theodora Karagounis, Edward F Kreider, Ben Murrell, Nico Pfeifer, Lilian Nogueira, Thiago Y Oliveira, Gerald H Learn, Yehuda Z Cohen, Clara Lehmann, Daniel Gillor, Irina Shimeliovich, Cecilia Unson-O'Brien, Daniela Weiland, Alexander Robles, Tim Kümmerle, Christoph Wyen, Rebeka Levin, Maggi Witmer-Pack, Kemal Eren, Caroline Ignacio, Szilard Kiss, Anthony P West, Hugo Mouquet, Barry S Zingman, Roy M Gulick, Tibor Keler, Pamela J Bjorkman, Michael S Seaman, Beatrice H Hahn, Gerd Fätkenheuer, Sarah J Schlesinger, Michel C Nussenzweig, and Florian Klein. Antibody 10-1074 suppresses viremia in HIV-1-infected individuals. *Nature Medicine*, 23(2):185–191, 2017.

[15] Kun Mao Chao, William R. Pearson, and Webb Miller. Aligning two sequences within a specified diagonal band. *Bioinformatics*, 8(5):481–487, 1992.

[16] Kun Mao Chao, Ross C. Hardison, and Webb Miller. Constrained sequence alignment. *Bulletin of Mathematical Biology*, 55(3):503–524, 1993.

[17] Chen-shan Chin, David H Alexander, Patrick Marks, Aaron A Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E Eichler, Stephen W Turner, and Jonas Korlach. Nonhybrid , finished microbial genome assemblies from long-read SMRT sequencing data. 10(6), 2013.

[18] José M. Cuevas, Ron Geller, Raquel Garijo, José López-Aldeguer, and Rafael Sanjuán. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biology*, 13(9): 1–19, 2015.

[19] Nathan W Cummins and Andrew D Badley. Making sense of how HIV kills infected CD4 T cells: implications for HIV cure. *Molecular and cellular therapies*, 2:20, 2014.

[20] Michael J. Dapp, Richard H. Heineman, and Louis M. Mansky. Interrelationship between HIV-1 Fitness and Mutation Rate. *Journal of Molecular Biology*, 425(1): 41–53, jan 2013.

[21] David Cournapeau. scikit-learn. Software download. URL https://scikit-learn.org.

[22] Mark M Davis, Kathryn Calame, Philip W Early, Donna L Livant, Rolf Joho, Irving L Weissman, and Leroy Hood. An immunoglobulin heavy-chain gene is formed by at least two recombinational events. *Nature*, 283(5749):733–739, 1980.

[23] Steven G. Deeks, Brigitte Autran, Ben Berkhout, Monsef Benkirane, Scott Cairns, Nicolas Chomont, Tae-Wook Chun, Melissa Churchill, Michele Di Mascio, Christine Katlama, Alain Lafeuillade, Alan Landay, Michael Lederman, Sharon R. Lewin, Frank Maldarelli, David Margolis, Martin Markowitz, Javier Martinez-Picado, James I. Mullins, John Mellors, Santiago Moreno, Una O'Doherty, Sarah Palmer, Marie-Capucine Penicaud, Matija Peterlin, Guido Poli, Jean-Pierre Routy, Christine Rouzioux, Guido Silvestri, Mario Stevenson, Amalio Telenti, Carine Van Lint, Eric Verdin, Ann Woolfrey, John Zaia, and Françoise Barré-Sinoussi. Towards an HIV cure: a global scientific strategy. *Nature Reviews Immunology*, 12 (8):607–614, jul 2012.

[24] Orlando DeLeon, Hagit Hodis, Yunxia OMalley, Jacklyn Johnson, Hamid Salimi, Yinjie Zhai, Elizabeth Winter, Claire Remec, Noah Eichelberger, Brandon Van Cleave, et al. Accurate predictions of population-level changes in sequence and structural properties of HIV-1 Env using a volatility-controlled diffusion model. *PLoS biology*, 15(4), 2017.

[25] Suprit Deshpande, Shilpa Patil, Rajesh Kumar, Tandile Hermanus, Kailapuri G Murugavel, Aylur K Srikrishnan, Suniti Solomon, Lynn Morris, and Jayanta Bhattacharya. HIV-1 clade C escapes broadly neutralizing autologous antibodies with N332 glycan specificity by distinct mechanisms. *Retrovirology*, 13(1):60, 2016.

[26] William S. DeWitt, Paul Lindau, Thomas M. Snyder, Anna M. Sherwood, Marissa Vignali, Christopher S. Carlson, Philip D. Greenberg, Natalie Duerkopp, Ryan O. Emerson, and Harlan S. Robins. A public database of memory and naive B-cell receptor sequences. *PLoS ONE*, 11(8):1–18, 2016.

[27] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017.

[28] Deborah Donnell, Jared M Baeten, James Kiarie, K Thomas, Wendy Stevens, Craig R Cohen, James Mcintyre, Jairam R Lingappa, and Connie Celum. Heterosexual HIV-1 transmission after initiation of antiretroviral therapy: A prospective cohort analysis. *Lancet*, 375(9731):2092–98, 2010.

[29] Nan Du and Yanni Sun. Improve homology search sensitivity of PacBio data by correcting frameshifts. In *Bioinformatics*, volume 32, pages i529–i537, 2016.

[30] Robert C Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, 2010.

[31] Robert C Edgar and Henrik Flyvbjerg. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, 31(21), 2015.

[32] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. DeWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138, 2009.

[33] Ember Core Team. Ember.js. Software download. URL https://emberjs.com/.

[34] N. R. Faria, A. Rambaut, M. A. Suchard, G. Baele, T. Bedford, M. J. Ward, A. J. Tatem, J. D. Sousa, N. Arinaminpathy, J. Pepin, D. Posada, M. Peeters, O. G. Pybus, and P. Lemey. The early spread and epidemic ignition of HIV-1 in human populations. *Science*, 346(6205):56–61, 2014.

[35] Erin B Fichot and R Sean Norman. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome*, 1(1):10, 2013.

[36] Will Fischer, Vitaly V Ganusov, Elena E Giorgi, Peter T Hraber, Brandon F Keele, Thomas Leitner, Cliff S Han, Cheryl D Gleasner, Lance Green, Chien Chi Lo, Ambarish Nag, Timothy C. Wallstrom, Shuyi Wang, Andrew J. McMichael, Barton F Haynes, Beatrice H Hahn, Alan S. Perelson, Persephone Borrow, George M Shaw, Tanmoy Bhattacharya, and Bette T Korber. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS ONE*, 5(8), 2010.

[37] Brian Thomas Foley, Thomas Kenneth Leitner, Cristian Apetrei, Beatrice Hahn, Ilene Mizrachi, James Mullins, Andrew Rambaut, Steven Wolinsky, and Bette Tina Marie Korber. HIV Sequence Compendium 2017. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2017.

[38] Yvonne Geiß and Ursula Dietrich. Catch Me If You Can The Race Between HIV and Neutralizing Antibodies. *AIDS Reviews*, pages 107–113, 2015.

[39] George Georgiou, Gregory C Ippolito, John Beausang, Christian E Busse, Hedda Wardemann, and Stephen R Quake. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology*, 32(2):158–168, 2014.

[40] Sara Gianella, Wayne Delport, Mary E Pacold, Jason A Young, Jun Yong Choi, Susan J Little, Douglas D Richman, Sergei L Kosakovsky Pond, and Davey M Smith. Detection of minority resistance during early HIV-1 infection: natural variation and spurious detection rather than transmission and evolution of multiple viral variants. *J Virol*, 85(16):8359–67, Aug 2011.

[41] Sara Gianella, Sergei L. Kosakovsky Pond, Michelli F. Oliveira, Konrad Scheffler, Matt C. Strain, Antonio De la Torre, Scott Letendre, Davey M. Smith, and Ronald J. Ellis. Compartmentalized HIV rebound in the central nervous system after interruption of antiretroviral therapy. *Virus Evolution*, 2(2), 2016.

[42] Sara Goodwin, John D. McPherson, and W. Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, 2016.

[43] Aditi Gupta and Christoph Adami. Strong selection significantly increases epistatic interactions in the long-term evolution of a protein. *PLoS genetics*, 12(3), 2016.

[44] Hillel Haim, Ignacio Salas, and Joseph Sodroski. Proteolytic Processing of the Human Immunodeficiency Virus Envelope Glycoprotein Precursor Decreases Conformational Flexibility. *Journal of virology*, 87(3):1884–1889, 2012.

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

[46] Linling He, Devin Sok, Parisa Azadnia, Jessica Hsueh, Elise Landais, Melissa Simek, Wayne C. Koff, Pascal Poignard, Dennis R. Burton, and Jiang Zhu. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. *Scientific Reports*, 4 (1):6778, 2015.

[47] Matthew R Henn, Christian L Boutwell, Patrick Charlebois, Niall J Lennon, Karen A Power, Alexander R Macalalad, Aaron M Berlin, Christine M Malboeuf, Elizabeth M Ryan, Sante Gnerre, Michael C Zody, Rachel L Erlich, Lisa M Green, Andrew Berical, Yaoyu Wang, Monica Casali, Hendrik Streeck, Allyson K Bloom, Tim Dudek, Damien Tully, Ruchi Newman, Karen L Axten, Adrianne D Gladden, Laura Battis, Michael Kemper, Qiandong Zeng, Terrance P Shea, Sharvari Gujja, Carmen Zedlack, Olivier Gasser, Christian Brander, Christoph Hess, Huldrych F. Günthard, Zabrina L. Brumme, Chanson J Brumme, Suzane Bazner, Jenna Rychert, Jake P Tinsley, Ken H. Mayer, Eric Rosenberg, Florencia Pereyra, Joshua Z Levin, Sarah K Young, Heiko Jessen, Marcus Altfeld, Bruce W Birren, Bruce D Walker, and Todd M Allen. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. *PLoS Pathogens*, 8(3), 2012.

[48] Dongni Hou, Cuicui Chen, Eric John Seely, Shujing Chen, and Yuanlin Song. High-throughput sequencing-based immune repertoire study during infectious disease. *Frontiers in Immunology*, 7(AUG):1–11, 2016.

[49] P Hraber, M S Seaman, R T Bailer, J R Mascola, D C Montefiori, and B T Korber. Prevalence of broadly neutralizing antibody responses during chronic HIV-1 infection. *Aids*, 28(2):163–169, 2014.

[50] Da Wei Huang, Castle Raley, Min Kang Jiang, Xin Zheng, Dun Liang, M Tauseef Rehman, Helene C Highbarger, Xiaoli Jiao, Brad Sherman, Liang Ma, Xiaofeng Chen, Thomas Skelly, Jennifer Troyer, Robert Stephens, Tomozumi Imamichi, Alice Pau, Richard A Lempicki, Bao Tran, Dwight Nissley, H Clifford Lane, and Robin L Dewar. Towards Better Precision Medicine: PacBio Single-Molecule Long Reads Resolve the Interpretation of HIV Drug Resistant Mutation Profiles at Explicit Quasispecies (Haplotype) Level. *Journal of data mining in genomics & proteomics*, 7(1):1–30, 2016.

[51] Michael Huber, Karin J Metzner, Fabienne D Geissberger, Cyril Shah, Christine Leemann, Thomas Klimkait, Jürg Böni, Alexandra Trkola, and Osvaldo Zagordi. MinVar: A rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing. *Journal of Virological Methods*, 240:7–13, 2017.

[52] Cassandra B Jabara, Corbin D Jones, Jeffrey Roach, Jeffrey A Anderson, and Ronald Swanstrom. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences*, 108 (50):20166–20171, 2011.

[53] Jacob Schreiber. Pomegranate. Software download. URL https://github.com/jmschrei/pomegranate.

[54] Anjali Joshi, Erin B Punke, Melina Sedano, Bethany Beauchamp, Rima Patel, Cassady Hossenlopp, Ogechika K Alozie, Jayanta Gupta, Debabrata Mukherjee, and Himanshu Garg. CCR5 promoter activity correlates with HIV disease progression by regulating CCR5 cell surface expression and CD4 T cell apoptosis. *Scientific Reports*, 7, 2017.

[55] Salim S Abdool Karim and Quarraisha Abdool Karim. Antiretroviral prophylaxis: A defining moment in HIV control. *The Lancet*, 378(9809):2011–2014, 2011.

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

[57] Kazutaka Katoh, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30(14):3059–3066, 2002.

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

[59] Cornell Kortenhoeven, Fourie Joubert, A Bastos, and Celia Abolnik. Virus genome dynamics under different propagation pressures: reconstruction of whole genome haplotypes of west nile viruses from NGS data. *BMC Genomics*, 16(1):118, 2015.

[60] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[61] Peter D. Kwong and John R. Mascola. Human Antibodies that Neutralize HIV-1: Identification, Structures, and B Cell Ontogenies. *Immunity*, 37(3):412–425, 2012.

[62] Peter D. Kwong, John R. Mascola, and Gary J. Nabel. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. *Nature Reviews Immunology*, 13(9):693–701, 2013.

[63] Melissa Laird Smith, Ben Murrell, Kemal Eren, Caroline Ignacio, Elise Landais, Steven Weaver, Pham Phung, Colleen Ludka, Lance Hepler, Gemma Caballero, Tristan Pollner, Yan Guo, Douglas Richman, Pascal Poignard, Ellen E. Paxinos, Sergei L. Kosakovsky Pond, and Davey M. Smith. Rapid Sequencing of Complete env Genes from Primary HIV-1 Samples. *Virus Evolution*, 2(2), 2016.

[64] Elise Landais, Xiayu Huang, Colin Havenar-Daughton, Ben Murrell, Matt A. Price, Lalinda Wickramasinghe, Alejandra Ramos, Charoan B. Bian, Melissa Simek, Susan Allen, Etienne Karita, William Kilembe, Shabir Lakhi, Mubiana Inambao, Anatoli Kamali, Eduard J. Sanders, Omu Anzala, Vinodh Edward, Linda Gail Bekker, Jianming Tang, Jill Gilmour, Sergei L. Kosakovsky-Pond, Pham Phung, Terri Wrin, Shane Crotty, Adam Godzik, and Pascal Poignard. Broadly Neutralizing Antibody Responses in a Large Longitudinal Sub-Saharan HIV Primary Infection Cohort. *PLoS Pathogens*, 12(1):1–22, 2016.

[65] Elise Landais, Ben Murrell, Bryan Briney, Sasha Murrell, Kimmo Rantalainen, Alejandra Ramos, Lalinda Wickramasinghe, Melissa Laird Smith, Kemal Eren, Zachary Berndsen, Natalia De Val, Mengyu Wu, Audrey Cappelletti, Yolanda Lie, Terri Wrin, Paul Algate, Etienne Karita, B Andrew, Ian A Wilson, Dennis R Burton, Davey Smith, L Sergei, Pascal Poignard, Computational Biology, Vaccine Immunology, La Jolla, Biomedical Informatics, San Diego, Monogram Biosciences, Monogram Biosciences, Theraclone Sciences, and Project San Francisco. HIV Envelope Glycoform Heterogeneity and Localized Diversity Govern the Initiation and Maturation of a V2 Apex Broadly Neutralizing Antibody Lineage. *Immunity*, 2017.

[66] Christopher Lee. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics*, 19(8):999–1008, 2003.

[67] Christopher Lee, Catherine Grasso, and Mark F. Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, 2002.

[68] Jeong Hyun Lee, Gabriel Ozorowski, and Andrew B Ward. Cryo-EM structure of a native, fully glycosylated, cleaved HIV-1 envelope trimer. *Science*, 351(6277): 1043–1048, 2016.

[69] Niall Lennon, Aaron M Berlin, Matthew R Henn, Christian L Boutwell, Patrick Charlebois, Niall J Lennon, Karen A Power, Alexander R Macalalad, Aaron M Berlin, Christine M Malboeuf, Elizabeth M Ryan, Sante Gnerre, Michael C Zody, Rachel L Erlich, Lisa M Green, Andrew Berical, Yaoyu Wang, Monica Casali, Hendrik Streeck, Allyson K Bloom, Tim Dudek, Damien Tully, Ruchi Newman, Karen L Axten, Adrianne D Gladden, Laura Battis, Michael Kemper, Qiandong Zeng, Terrance P Shea, Zabrina L Brumme, Chanson J Brumme, Suzane Bazner, Jenna Rychert, Jake P Tinsley, H Ken, Bruce W Birren, Bruce D Walker, and Todd M Allen. Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition ... Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. (June), 2017.

[70] Preston Leung, Rowena Bull, Andrew Lloyd, and Fabio Luciani. A bioinformatics pipeline for the analyses of viral escape dynamics and host immune responses during an infection. *BioMed research international*, 2014:264519, jan 2014.

[71] Preston Leung, Auda A Eltahla, Andrew R Lloyd, Rowena A Bull, and Fabio Luciani. Understanding the complex evolution of rapidly mutating viruses with deep sequencing: beyond the analysis of viral diversity. *Virus research*, 2016.

[72] Ma Liang, Castle Raley, Xin Zheng, Geetha Kutty, Emile Gogineni, Brad T Sherman, Qiang Sun, Xiongfong Chen, Thomas Skelly, Kristine Jones, Robert Stephens, Bin Zhou, William Lau, Calvin Johnson, Tomozumi Imamichi, Minkang Jiang, Robin Dewar, Richard A Lempicki, Bao Tran, Joseph A Kovacs, and Da Wei Huang. Distinguishing highly similar gene isoforms with a clustering-based bioinformatics analysis of PacBio single-molecule long reads. *BioData Mining*, 9 (1):13, 2016.

[73] Susan J. Little, Sergei L Kosakovsky Pond, Christy M. Anderson, Jason A. Young, Joel O. Wertheim, Sanjay R. Mehta, Susanne May, and Davey M. Smith. Using HIV networks to inform real time prevention interventions. *PLoS ONE*, 9(6):1–8, 2014.

[74] Nicholas J Loman, Joshua Quick, and Jared T Simpson. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, 12(8):733–735, 2015.

[75] Daniel T. MacLeod, Nancy M. Choi, Bryan Briney, Fernando Garces, Lorena S. Ver, Elise Landais, Ben Murrell, Terri Wrin, William Kilembe, Chi-Hui Liang, Alejandra Ramos, Chaoran B. Bian, Lalinda Wickramasinghe, Leopold Kong, Kemal Eren, Chung-Yi Wu, Chi-Huey Wong, Sergei L. Kosakovsky Pond, Ian A. Wilson, Dennis R. Burton, Pascal Poignard, Matt A. Price, Jill Gilmour, Pat Fast, Anatoli Kamali, Eduard J. Sanders, Omu Anzala, Susan Allen, Eric Hunter, Etienne Karita, William Kilembe, Shabir Lakhi, Mubiana Inambao, Vinodh Edward, and Linda-Gail Bekker. Early Antibody Lineage Diversification and Independent Limb Maturation Lead to Broad HIV-1 Neutralization Targeting the Env High-Mannose Patch. *Immunity*, 44(5):1215–1226, may 2016.

[76] Serghei Mangul, Nicholas C. Wu, Nicholas Mancuso, Alex Zelikovsky, Ren Sun, and Eleazar Eskin. Accurate viral population assembly from ultra-deep sequencing data. *Bioinformatics*, 30(12):329–337, 2014.

[77] Marco Biasini. pv. Software download. URL http://biasmv.github.io/pv/.

[78] Alyssa R Martin and Robert F Siliciano. Progress toward HIV eradication: case reports, current efforts, and the challenges associated with cure. *Annual review of medicine*, 67:215–228, 2016.

[79] Darren P. Martin, Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1(1):1–5, 2015.

[80] Darren P Martin, Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1(1), 2015.

[81] John R Mascola and Barton F Haynes. HIV-1 neutralizing antibodies: understanding nature's pathways. *Immunological reviews*, 254(1):225–244, 2013.

[82] Rosemary M McCloskey, Richard H Liang, P Richard Harrigan, Zabrina L Brumme, and Art F Y Poon. An evaluation of phylogenetic methods for reconstructing transmitted HIV variants using longitudinal clonal HIV sequence data. *Journal of virology*, 88(11):6181–94, 2014.

[83] Laura E. McCoy and Dennis R. Burton. Identification and specificity of broadly neutralizing antibodies against HIV. *Immunological Reviews*, 275(1):11–20, 2017.

[84] David H McDermott, Peter A Zimmerman, Florence Guignard, Cynthia A Kleeberger, Susan F Leitman, Philip M Murphy, Multicenter AIDS Cohort Study

(MACS, et al. CCR5 promoter polymorphism and HIV-1 disease progression. *The Lancet*, 352(9131):866–870, 1998.

[85] Kerensa McElroy, Torsten Thomas, and Fabio Luciani. Deep sequencing of evolving pathogen populations: applications, errors, and bioinformatic solutions. *Microbial Informatics and Experimentation*, 4(1):1, 2014.

[86] Jan Medlock, Abhishek Pandey, Alyssa S Parpia, Amber Tang, Laura A Skrip, and Alison P Galvani. Effectiveness of UNAIDS targets and HIV vaccination across 127 countries. *Proceedings of the National Academy of Sciences*, 114(15): 4017–4022, apr 2017.

[87] Mike Bostock, Jason Davies, Jeffrey Heer, Vadim Ogievetsky, and community. D3.js. Software download. URL http://d3js.org/.

[88] Penny L. Moore, Carolyn Williamson, and Lynn Morris. Virological features associated with the development of broadly neutralizing antibodies to HIV-1. *Trends in Microbiology*, 23(4):204–211, 2015.

[89] H. Morbach, E. M. Eichhorn, J. G. Liese, and H. J. Girschick. Reference values for B cell subpopulations from infancy to adulthood. *Clinical and Experimental Immunology*, 162(2):271–279, 2010.

[90] Hugo Mouquet, Louise Scharf, Zelda Euler, Yan Liu, Caroline Eden, Johannes F Scheid, A. Halper-Stromberg, P. N. P. Gnanapragasam, D. I. R. Spencer, M. S. Seaman, H. Schuitemaker, T. Feizi, M. C. Nussenzweig, and P. J. Bjorkman. Complex-type N-glycan recognition by potent broadly neutralizing HIV antibodies. *Proceedings of the National Academy of Sciences*, 109(47):E3268–E3277, nov 2012.

[91] Rithun Mukherjee, Shane T. Jensen, Frances Male, Kyle Bittinger, Richard L. Hodinka, Michael D. Miller, and Frederic D. Bushman. Switching between raltegravir resistance pathways analyzed by deep sequencing. *Aids*, 25(16):1951–1959, 2011.

[92] Kenneth Murphy and Casey Weaver. *Janeway's immunobiology*. Garland Science, 2016.

[93] Ben Murrell, Sasha Moola, Amandla Mabona, Thomas Weighill, Daniel Sheward, Sergei L Kosakovsky Pond, and Konrad Scheffler. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Molecular biology and evolution*, page mst030, 2013.

[94] Ben Murrell, Steven Weaver, Martin D Smith, Joel O Wertheim, Sasha Murrell, Anthony Aylward, Kemal Eren, Tristan Pollner, Darren P Martin, Davey M Smith, et al. Gene-wide identification of episodic selection. *Molecular biology and evolution*, 32(5):1365–1371, 2015.

[95] Ben Murrell, Kemal Eren, Lorena S Ver, Nancy Choi, Elise Landais, Pascal Poignard, Sergei Kosakovsky Pond, and Davey Smith. Full-length *env* deep sequencing in a donor with broadly neutralizing N332 antibodies. In *Keystone Symposia on Molecular and Cellular Biology*, 2016.

[96] Niranjan Nagarajan and Mihai Pop. Sequence assembly demystified. *Nature Reviews Genetics*, 14(3):157–167, 2013.

[97] Fumiyo Nakagawa, Margaret May, and Andrew Phillips. Life expectancy living with HIV. *Current Opinion in Infectious Diseases*, 26(1):17–25, 2013.

[98] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similiarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[99] Oluwafemi O Oguntibeju. Quality of life of people living with HIV and AIDS and antiretroviral therapy. *HIV/AIDS (Auckland, N.Z.)*, 4:117–24, 2012.

[100] Mary E Pacold, Sergei L Kosakovsky Pond, Gabriel A Wagner, Wayne Delport, Daniel L Bourque, Douglas D Richman, Susan J Little, and Davey M Smith. Clinical, virologic, and immunologic correlates of HIV-1 intraclade B dual infection among men who have sex with men. *AIDS*, 26(2):157–65, Jan 2012.

[101] Sankar K. Pal, Sanghamitra Bandyopadhyay, and Shubhra Sankar Ray. Evolutionary computation in bioinformatics: A review. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 36(5):601–615, 2006.

[102] Aridaman Pandit and Rob J de Boer. Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants. *Retrovirology*, 11:56, 2014.

[103] Mariona Parera, Nuria Perez-Alvarez, Bonaventura Clotet, and Miguel Angel Martínez. Epistasis among deleterious mutations in the HIV-1 protease. *Journal of molecular biology*, 392(2), 2009.

[104] Konrad Paszkiewicz and David J. Studholme. De novo assembly of short sequence reads. *Briefings in Bioinformatics*, 11(5):457–472, 2010.

[105] Martine Peeters, Matthieu Jung, and Ahidjo Ayouba. The origin and molecular epidemiology of HIV. *Expert review of anti-infective therapy*, 11(9):885–896, 2013.

[106] Pervez, M Babar, Asif Nadeem, M Aslam, A Awan, Naeem Aslam, Tanveer Hussain, Nasir Naveed, Salman Qadri, Usman Waheed, and Muhammad Shoaib. Evaluating the Accuracy and Efficiency of Multiple Sequence Alignment Methods. *Evolutionary Bioinformatics*, page 205, dec 2014.

[107] Sergei L Kosakovsky Pond and Spencer V Muse. HyPhy: hypothesis testing using phylogenies. In *Statistical methods in molecular evolution*, pages 125–181. Springer, 2005.

[108] Art F Y Poon, Luke C Swenson, Evelien M Bunnik, Diana Edo-Matas, Hanneke Schuitemaker, Angélique B. van 't Wout, and P. Richard Harrigan. Reconstructing the Dynamics of HIV Evolution within Hosts from Serial Deep Sequence Data. *PLoS Computational Biology*, 8(11), 2012.

[109] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26(7):1641–1650, 2009.

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

[111] Michael Quail, Miriam E Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, 13(1):341, 2012.

[112] Thomas C Quinn. HIV epidemiology and the effects of antiviral therapy on long-term consequences. *AIDS*, 22(Suppl 3):S7–S12, sep 2008.

[113] Miguel E. Quiñones-Mateu, Santiago Avila, Gustavo Reyes-Teran, and Miguel A Martinez. Deep sequencing: Becoming a critical tool in clinical virology. *Journal of Clinical Virology*, 61(1):9–19, sep 2014.

[114] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

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

[116] Anthony Rhoads and Kin Fai Au. PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, 13(5):278–289, 2015.

[117] Peter Rice, Ian Longden, and Alan Bleasby. EMBOSS: the European molecular biology open software suite, 2000.

[118] D. D. Richman, D. M. Margolis, M. Delaney, W. C. Greene, D. Hazuda, and R. J. Pomerantz. The Challenge of Finding a Cure for HIV Infection. *Science*, 323 (5919):1304–1307, 2009.

[119] D. D. Richman, D. M. Margolis, M. Delaney, W. C. Greene, D. Hazuda, and R. J. Pomerantz. The Challenge of Finding a Cure for HIV Infection. *Science*, 323 (5919):1304–1307, 2009.

[120] Brittany Rife and Marco Salemi. On the early dynamics and spread of HIV-1. *Trends in microbiology*, 23(1):3–4, 2015.

[121] Richard J Roberts, Mauricio O Carneiro, and Michael C Schatz. The advantages of SMRT sequencing. *Genome Biology*, 14(6):405, 2013.

[122] Derek Ruths and Luay Nakhleh. Recombination and phylogeny: effects and detection. *International Journal of Bioinformatics Research and Applications*, 1 (2):202–212, 2005.

[123] Marcella Sarzotti-Kelsoe, Robert T Bailer, Ellen Turk, Chen-li Lin, Miroslawa Bilska, Kelli M Greene, Hongmei Gao, Christopher A Todd, Daniel A Ozaki, Michael S Seaman, et al. Optimization and validation of the TZM-bl assay for standardized assessments of neutralizing antibodies against HIV-1. *Journal of immunological methods*, 409, 2014.

[124] Eric E. Schadt, Steve Turner, and Andrew Kasarskis. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2):227–240, 2010.

[125] Mikkel H. Schierup and Jotun Hein. Consequences of recombination on traditional phylogenetic analysis. *Genetics*, 156(2):879–891, 2000.

[126] Mariano M Sede, Franco a Moretti, Natalia L Laufer, Leandro R Jones, and Jorge F Quarleri. HIV-1 tropism dynamics and phylogenetic analysis from longitudinal ultra-deep sequencing data of CCR5- and CXCR4-using variants. *PloS one*, 9(7): e102857, jan 2014.

[127] Sergei L Kosakovsky Pond. TN93. Software download, . URL https://github.com/veg/tn93.

[128] Sergei L Kosakovsky Pond. phylotree.js. Software download, . URL https://github.com/veg/phylotree.js.

[129] George M Shaw and Eric Hunter. HIV Transmission. *Cold Spring Harbor Perspectives in Medicine*, 2(11):a006965–a006965, nov 2012.

[130] Sergey L. Sheetlin, Yonil Park, Martin C. Frith, and John L. Spouge. Frameshift alignment: Statistics and post-genomic applications. *Bioinformatics*, 30(24): 3575–3582, 2014.

[131] Daniel J Sheward, Ben Murrell, and Carolyn Williamson. Degenerate Primer IDs and the birthday problem. *Proceedings of the National Academy of Sciences*, 109 (21):E1330–E1330, 2012.

[132] Robert F. Siliciano and Warner C. Greene. HIV latency. *Cold Spring Harbor Perspectives in Medicine*, 1(1):1–19, 2011.

[133] Pavel Skums, Nicholas Mancuso, Alexander Artyomenko, Bassam Tork, Ion Mandoiu, Yury Khudyakov, and Alex Zelikovsky. Reconstruction of viral population structure from next-generation sequencing data using multicommodity flows. *BMC bioinformatics*, 14 Suppl 9(Suppl 9):S2, 2013.

[134] Temple F. Smith and Michael S. Waterman. Comparison of biosequences. *Advances in Applied Mathematics*, 2(4):482–489, 1981.

[135] Christoph D Spinner, Christoph Boesecke, Alexander Zink, Heiko Jessen, Hans-Jürgen Stellbrink, Jürgen Kurt Rockstroh, and Stefan Esser. HIV pre-exposure prophylaxis (PrEP): a review of current knowledge of oral systemic HIV PrEP in humans. *Infection*, 44(2):151–158, 2016.

[136] Jonathan D. Steckbeck, Anne Sophie Kuhlmann, and Ronald C. Montelaro. C-terminal tail of human immunodeficiency virus gp41: Functionally rich and structurally enigmatic. *Journal of General Virology*, 94(PART11):1–19, 2013.

[137] Evguenia S Svarovskaia, Ross Martin, John G McHutchison, Michael D Miller, and Hongmei Mo. Abundant drug-resistant NS3 mutants detected by deep sequencing in hepatitis C virus-infected patients undergoing NS3 protease inhibitor monotherapy. *Journal of clinical microbiology*, 50(10):3267–74, oct 2012.

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

[139] S Teeraananchai, SJ Kerr, J Amin, K Ruxrungtham, and MG Law. Life expectancy of HIV-positive people after starting combination antiretroviral therapy: a meta-analysis. *HIV medicine*, 18(4):256–266, 2017.

[140] Warren S Torgerson. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4), 1952.

[141] Athe M N Tsibris, Bette Korber, Ramy Arnaout, Carsten Russ, Chien Chi Lo, Thomas Leitner, Brian Gaschen, James Theiler, Roger Paredes, Zhaohui Su, Michael D Hughes, Roy M. Gulick, Wayne Greaves, Eoin Coakley, Charles Flexner, Chad Nusbaum, and Daniel R Kuritzkes. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS ONE*, 4(5):1–12, 2009.

[142] UNAIDS. UNAIDS Data 2017. Technical report, 2017.

[143] Fran Van Heuverswyn and Martine Peeters. The origins of HIV and implications for the global epidemic. *Current infectious disease reports*, 9(4):338–346, 2007.

[144] Andrew Varble, Randy A Albrecht, Simone Backes, Marshall Crumiller, Nicole M Bouvier, David Sachs, Adolfo García-Sastre, and Benjamin R. Tenoever. Influenza a virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host and Microbe*, 16(5):691–700, 2014.

[145] Antony T Vincent, Nicolas Derome, Brian Boyle, Alexander I. Culley, and Steve J. Charette. Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. In *Journal of Microbiological Methods*, volume 138, pages 60–71. Elsevier B.V., 2017.

[146] Gabriel A Wagner, Mary E Pacold, Sergei L Kosakovsky Pond, Gemma Caballero, Antoine Chaillon, Abby E Rudolph, Sheldon R Morris, Susan J Little, Douglas D Richman, and Davey M Smith. Incidence and prevalence of intrasubtype HIV-1 dual infection in at-risk men in the United States. *The Journal of infectious diseases*, 209(7):1032–1038, 2013.

[147] Gary P Wang, Scott a Sherrill-Mix, Kyong-Mi Chang, Chris Quince, and Frederic D Bushman. Hepatitis C virus transmission bottlenecks analyzed by deep sequencing. *Journal of virology*, 84(12):6218–28, jun 2010.

[148] Qiong Wang, J. F. Quensen, Jordan A Fish, T. Kwon Lee, Yanni Sun, James M. Tiedje, and James R. Cole. Ecological Patterns of nifH Genes in Four Terrestrial Climatic Zones Explored with Targeted Metagenomics Using FrameBot, a New Informatics Tool. *mBio*, 4(5):e00592–13–e00592–13, sep 2013.

[149] Corey T Watson, Karyn M Steinberg, John Huddleston, Rene L Warren, Maika Malig, Jacqueline Schein, A Jeremy Willsey, Jeffrey B Joy, Jamie K Scott, Tina A Graves, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *The American Journal of Human Genetics*, 92(4):530–546, 2013.

[150] Daniel M Weinreich. High-throughput identification of genetic interactions in HIV-1. *Nature genetics*, 43(5), 2011.

[151] Constantinos Kurt Wibmer, Penny L. Moore, and Lynn Morris. HIV broadly neutralizing antibody targets. *Current Opinion in HIV and AIDS*, 10(3):135–143, may 2015.

[152] Constantinos Kurt Wibmer, Penny L Moore, and Lynn Morris. HIV broadly neutralizing antibody targets. *Current opinion in HIV and AIDS*, 10(3):135, 2015.

[153] C. B. Wilen, J. C. Tilton, and R. W. Doms. HIV: Cell Binding and Entry. *Cold Spring Harbor Perspectives in Medicine*, 2(8):a006866–a006866, aug 2012.

[154] Joseph K. Wong and Steven A. Yukl. Tissue reservoirs of HIV. *Current Opinion in HIV and AIDS*, 11(4):362–370, jul 2016.

[155] Daniel J Woodsworth, Mauro Castellarin, and Robert A Holt. Sequence analysis of T-cell repertoires in health and disease. *Genome Medicine*, 5(10):98, 2013.

[156] M. Worobey, P. Telfer, S. Souquiere, M. Hunter, C. A. Coleman, M. J. Metzger, P. Reed, M. Makuwa, G. Hearn, S. Honarvar, P. Roques, C. Apetrei, M. Kazanji, and P. A. Marx. Island Biogeography Reveals the Deep History of SIV. *Science*, 329(5998):1487–1487, 2010.

[157] X. Wu, Tongqing Zhou, Jiang Zhu, B. Zhang, Ivelin Georgiev, Charlene Wang, Xuejun Chen, Nancy S Longo, Mark Louder, Krisha McKee, Sijy O'Dell, Stephen Perfetto, Stephen D Schmidt, Wei Shi, Lan Wu, Y. Yang, Z.-Y. Yang, Z. Yang, Zhenhai Zhang, Mattia Bonsignori, John A Crump, Saidi H Kapiga, Noel E Sam, Barton F Haynes, Melissa Simek, Dennis R Burton, Wayne C Koff, Nicole A Doria-Rose, Mark Connors, James C Mullikin, Gary J Nabel, Mario Roederer, Lawrence Shapiro, Peter D Kwong, and John R Mascola. Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing. *Science*, 333(6049):1593–1602, 2011.

[158] Gur Yaari and Steven H. Kleinstein. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine*, 7(1):121, 2015.

[159] Li Yin, Li Liu, Yijun Sun, Wei Hou, Amanda C Lowe, Brent P Gardner, Marco Salemi, Wilton B Williams, William G Farmerie, John W Sleasman, and Maureen M Goodenow. High-resolution deep sequencing reveals biodiversity, population structure, and persistence of HIV-1 quasispecies within host ecosystems. *Retrovirology*, 9(1):108, 2012.

[160] Ingrid Young and Lisa McDaid. How acceptable are antiretrovirals for the prevention of sexually transmitted HIV?: A review of research on the acceptability of oral pre-exposure prophylaxis and treatment as prevention. *AIDS and Behavior*, 18(2):195–216, 2014.

[161] Y Zhang and Y Sun. HMM-FRAME: accurate protein domain classification for metagenomic sequences containing frameshift errors. *Bmc Bioinformatics*, 12: 198, 2011.

[162] Peter A Zimmerman, Alicia Buckler-White, Ghalib Alkhatib, Todd Spalding, Joseph Kubofcik, Christophe Combadiere, Drew Weissman, Oren Cohen, Andrea Rubbert, Gordon Lam, et al. Inherited resistance to HIV-1 conferred by an inactivating mutation in CC chemokine receptor 5: studies in populations with contrasting clinical phenotypes, defined racial background, and quantified risk. *Molecular Medicine*, 3(1):23, 1997.