Using Deep Sequencing with a Primer ID to Resolve the Structure of Viral Populations and Reveal Pre-existing Drug Resistance Mutations in the HIV and HCV Protease Genes

Cassandra B. Jabara

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology

Chapel Hill
2012

Approved by

Corbin D. Jones, Ph.D.

Ronald Swanstrom, Ph.D.

Stanley M. Lemon, M.D.

Christina L. Burch, Ph.D.

Charles E. Mitchell, Ph.D.

Abstract

CASSANDRA B. JABARA: Using Deep Sequencing with a Primer ID to Resolve the Structure of Viral Populations and Reveal Pre-existing Drug Resistance Mutations in the HIV and HCV Protease Genes
(Under the direction of Ronald Swanstrom, Stanley M. Lemon, and Corbin D. Jones)

Human Immunodeficiency Virus (HIV) and Hepatitis C Virus (HCV) are among the most deadly chronic viral diseases affecting the human population. The rich genetic diversity produced within a host includes adaptive resistance alleles that may enable viral escape from drug selective pressures. An in-depth characterization of the intrahost population and the genetic path it takes to escape drug selection may reveal how to prevent the evolution of resistance. Resolving the fine-scale genetic structure of a viral population requires deep sampling of the genetic variation within a viral population. I developed a novel technique, Primer ID, which reproducibly (Chapter 4) captures viral diversity while correcting for PCR biases and error inherent in deep sequencing protocols (Chapter 2). Deep sequencing with a Primer ID was applied to the targeted re-sequencing of protease for two different viral genomes, HIV (Chapter 2-3) and HCV (Chapter 4). The allelic distribution of genetic variation of HIV and HCV was skewed towards low-frequency polymorphisms, some of which were resistance-associated variants (Chapters 2-4). I observed that pre-existing resistance mutations could be directly selected during a drug treatment (Chapter 2). However, the path to resistance was often complex and confounded by variance in the steady-state frequency of resistance alleles, sampling depth, and the effective population size (Chapter 3).

Once a population of HIV escaped a drug, it was observed that resistance-associated variants were added *de novo* in a step-wise manner, not brought together by recombination of pre-existing haplotypes. HCV-HIV co-infection decreased overall population diversity (Chapter 4). This difference did not correlate with a change in the overall frequency of pre-existing resistance mutations, but specific resistance alleles were enriched in either mono- or co-infected populations. Further application of deep sequencing with a Primer ID will result in a greater understanding of the population dynamics of both HIV and HCV and determine if the standing genetic variation can be used to predict if a patient will fail therapy and how a viral population responds to selective pressures. Together, improvements in predictive power will result in an enhancement of therapeutic success rate and sustained virologic response.

This work is dedicated to those who push technological boundaries to study the abysmal and enthralling intricacies of RNA viral populations evolving inside hosts

Acknowledgements

nani gigantum humeris insidentes

If I had seen it further it is by standing on y^{e} shoulders of Giants.

Newton to Hooke, 5 February 1676

I am incredibly fortunate to be mentored in an unparalleled collaborative and cross-disciplinary environment by a small handful of exceptional and passionate experts. I cherish every interaction, discussion, suggestion, critique, and most importantly, question. They have pushed me to think, invent, develop, and refine. To strive for excellence knowing there is never an endpoint, only advancing the boundary of what we know.

I would like to thank the members of my thesis committee: Corbin Jones, Ronald Swanstrom, Stanley Lemon, Christina Burch, and Charles Mitchell, for the critical evaluation of my work. I would also like to thank Jeffrey Anderson for the intellectual and experimental mentorship that went into building the foundation for these projects, and Jesse Walsh, Fengyu Hu, and Jeffrey Roach for their technical expertise. In addition I am grateful for the discussions and insight from Piotr Mieczkowski, Joseph Eron, Myron Cohen, Michael Fried, Prema Menezes, David Margolis, and colleagues in the Departments of Biology, Microbiology, and Infectious Diseases, the Center for Genome Sciences, and the Swanstrom, Jones, and Lemon labs.

This work would not be possible without the generous donation of time and resources from study subjects and clinical providers. I would like to thank Dale Kempf of Abbott Laboratories and Michael Fried of the University of North Carolina at Chapel Hill School of Medicine for making samples available. I would also like to thank the UNC Center for AIDS Research, Lineberger Comprehensive Cancer Center, The National Institute of Health, and Merck for their financial support.

Table of Contents

Chapter

List of Tables

Table

List of Figures

Figures

List of Abbreviations

cART	Combined antiretroviral therapy
cDNA	Complementary deoxyribonucleic Acid
DAA	Direct acting antiviral
dsDNA	Double stranded deoxyribonucleic Acid
dsRNA	Double stranded ribonucleic acid
Env	Envelope glycoprotein
FDA	U.S. Food and Drug Administration
FDR	False discovery rate
Gag	Group-specific antigen
HAART	Highly active antiretroviral protein
HCV	Hepatitis C virus
HIV	Human immunodeficiency virus
IDU	Intravenous drug use
INF	Interferon
LD	Linkage disequilibrium
nRTI	Nucleoside reverse transcriptase inhibitor
NS	Nonsynonymous
NS3	Non-structural protein 3
NS4A	Non-structural protein 4A
NS4B	Non-structural protein 4B
NS5A	Non-structural protein 5A

NS5B	Non-structural protein 5B
PCR	Polymerase chain reaction
PEG	Pegylated
Pol	Polymerase
Pro	Protease
RAV	Resistance associated variant
RBV	Ribavirin
RdRp	RNA-dependent RNA polymerase
RIG-I	Retinoic acid-inducing gene 1
RNA	Ribonucleic acid
RNase H	Ribonuclease H
RT	Reverse transcriptase
RTV	Ritonavir
S	Segregating sites
SGA	Single genome amplification
SNP	Single nucleotide polymorphism
ssRNA	Single stranded ribonucleic acid
TLR3	Toll-like receptor 3
vRNA	Viral ribonucleic acid
VSV	Vesicular stomatitis virus

List of Symbols

D	Tajima’s D statistic
θ	Total sequence diversity
π	Average pairwise sequence diversity
ω	Fitness
μ	Mutation rate

Chapter 1

Introduction

1.1 The Evolvability of RNA viruses

RNA viruses are the dominant causative agent of emerging infectious diseases worldwide, yet only a small number of viral families can be effectively controlled through vaccination or antiviral drugs. Of all chronic viral diseases affecting the human population, Human Immunodeficiency Virus (HIV) and Hepatitis virus are the most deadly (1). HIV currently is infecting approximately 35 million people worldwide and is associated with 2 million deaths each year. Hepatitis C Virus (HCV) is infecting approximately 200 million people, and annually there are 366,000 deaths due to cirrhosis of the liver and hepatocellular carcinoma. Despite 30 years of active research and drug development, HIV remains incurable. Similarly, around 70 percent of HCV-infected individuals have persistent replication (2-4), and most potent drugs against this virus are still in clinical trials (1, 5). The substantial worldwide morbidity and mortality caused by these two viruses necessitates a better understanding of how to effectively counteract and control their spread within the human population. The most direct means of decreasing the interhost transmission rate is to suppress or eradicate intrahost viral populations.

The human host inherently contains strong barriers to pathogen infection, replication, and transmission. Despite highly evolved innate and adaptive immunities, the majority of human hosts cannot naturally clear HIV or HCV infections and often transmit these viruses to new hosts. As a result, direct acting antivirals (DAA) are needed as a pharmacologic intervention for suppressing the viral burden and decreasing transmission risk, and DAAs are increasing in abundance, availability, and sophistication. However, there are significant pitfalls to current day drug therapies against HIV and HCV. The newly FDA licensed DAAs against HCV have low genetic barriers that a virus population can rapidly evolve to overcome, which results in viral rebound. Highly Active Antiretroviral Therapy (HAART) against HIV is a multifaceted selection regiment with a combination of drugs, most of which contain high genetic barriers. However, cessation of HAART always results in viral rebound.

1.1.1 Protease inhibitors are a major drug class used to treat HIV and HCV infections.

There are six antiretroviral drug classes, 24 drugs in total, licensed to treat HIV. Each class targets a different component of the viral life cycle, from inhibiting attachment of the virion to the host cell through preventing virion maturation post-budding. The recommended treatment for HIV infection involves dual nucleoside reverse transcriptase inhibitors (nRTI) in combination with a third class, commonly a protease inhibitor (6). The standard of care for HCV infection is immunomodulation and mutagenicity using a combination of pegylated interferon-α and ribavirin (PEG-INF/RBV). However, as of late April and mid May of 2011, two DAAs were FDA approved to treat HCV infection. Both were protease inhibitors.

Protease inhibitors are an important drug class for counteracting both HIV and HCV infections. These drugs are designed to specifically inhibit the viral protease, an enzyme critical for the production of infectious virions. For HIV, its aspartic protease is released through autocatalysis of the Gag-Pro-Pol precursor polyprotein. Protease homodimers then cleave viral polyprotein peptide bonds through hydrolysis, with the orchestrating water molecule anchored in the active site by two opposing aspartic acid residues. For an HIV virion to mature, proteolytic processing of three polyprotein precursors, Gag, Gag-Pro-Pol, and Env, is required. Without protease function, the immature HIV virion is unable to initiate a new infectious cycle.

Similarly, the HCV protease NS3 is critical for HCV infection. NS3 is a serine protease that heterodimerizes with the viral cofactor NS4A. The NH_2 terminal cleaves four sites downstream of the NS2-3 junction in the HCV polyprotein, NS3-4A, NS4A-4B, NS4B-5A, and NS5A-5B (7). For proteolytic function, a catalytic triad is required in conjunction with a tetrahedrially coordinated metal ion (8). In addition to viral polyprotein cleavage, HCV's protease also processes host signaling molecules activated by dsRNA, blocking signaling through the TRL3 and RIG-I pathways (reviewed in (5)).

Both HIV and HCV proteases are extremely attractive drug targets for suppression of viral populations due to their critical roles in the viral life cycles. Furthermore, because of their enzymatic activity, they also contain structurally conserved active sites. The protease active site is in essence a pocket in which polyprotein cleavage occurs, therefore small molecule inhibitors designed to bind within the pocket will block viral polyproteins and other target molecules from being cleaved. However, amino acid and/or structural changes within the active site can decrease or prevent inhibitor binding. If drug inhibition is not complete then a certain degree of proteolytic processing will occur. Viral production under drug selection is the primary obstacle to therapeutic success, and results in the clinical presentation of drug resistance.

1.1.2 The permissive replication of RNA viruses and a high mutation rate result in the rapid generation of genetic variation.

Genetic variation is the raw material upon which selection acts. The ability of these viral populations to escape strong selection pressures—such as small molecule inhibitors—is due to a rich landscape of genetic heterogeneity. This variation is primarily introduced through polymerase nucleotide misincorporation during genomic template copying. It is the exceptionally high rate at which nucleotide misincorporation occurs that results in the quick diversification of an RNA virus population. HIV's error-prone Reverse Transcriptase (RT) introduces mutations at an estimated rate of 2.16×10^{-5} substitutions/site/generation during transcription of viral RNA (vRNA) to double-stranded DNA (dsDNA) (9, 10). HCV's RNA-dependent RNA polymerase (RdRp), NS5B, emulates the retroviral substitution rate, adding 1×10^{-4} to 1×10^{-5} substitutions/site/generation (11). However, HCV does not have a double-stranded intermediate genome, thus RdRp is exclusively used to go from positive-sense to negative-sense to positive-sense RNA, adding errors with each template copy.

Recombination creates additional diversity by enabling allelic shuffling within the population. It has been posited that recombination in HIV occurs in up to 40% of progeny virions (12, 13). HIV is pseudodiploid; each HIV virion contains two single-stranded RNA (ssRNA) genomes non-covalently linked at the 5' dimerization initiation sequence, but only one provirus is synthesized per viral particle. RT is associated with ribonuclease H (RNase H), an enzyme that degrades RNA from RNA-DNA heteroduplexes. During minus strand synthesis, RNase H frees up the growing chain, so if polymerase encounters a nick, break, or pause due to secondary structure in the RNA template, it can jump to the other molecule in the dimer, forming a recombinant. This model, originally called forced copy choice (12) but then broadened to copy choice (14), is the predominant theory for recombination in HIV (15, 16).

Unlike HIV, HCV cannot readily form recombinants *in vivo*, though there are some rare documented cases of intersubtype and intergenotypic recombination (17-19). HCV's replicase is membrane bound, which constrains RNA templates within lipid rafts and creates a barrier to template switching (20). Moreover, early competitive exclusion may prevent distinct variants arising to appreciable frequencies. If recombination occurs, chimers retain substantial homology to the parental templates and would be difficult to detect. Finally, recombination between subtypes can be readily induced *in vitro*, but recombinants have poor fitness. Sequence divergence incompatibilities would further limit the rise of recombinant genomes within a population (21). Thus if recombination occurs in HCV, it is rare, although poor sensitivity may result in underestimates of recombination (22).

1.1.3 Viruses can rapidly evolve resistance to protease inhibitors and other direct acting antivirals.

Nucleotide misincorporation by viral polymerase has an approximate random distribution across the viral genome, though it may be biased towards non-helical secondary structure (23-27) and other mutational biases (28, 29). The viral polymerase's misincorporation rate (μ) is the frequency at which new mutations are introduced into a population. If a misincorporation event results in a nonsynonymous change, and the new amino acid does not reduce viral fitness to zero, the allele[1] may be preserved in the viral

[1] An allele is defined as a genomic change from the consensus sequence of the intrahost untreated population.

population under a therapy naïve environment. New mutations will be added at a rate μ, but the frequency at which any given allele (q) will exist is the equilibrium achieved between the rate of introduction of that allele and the intensity at which selection (s) removes it, or $q = \mu/s$ (30).

Because of the mutation/selection balance, a viral population will contain a number of low-frequency *de novo* mutations. Some minor variants will be unviable, however, while others may confer resistance. If a new allele reduces or prevents protease inhibitor binding, thus allowing for the genotype to replicate in the face of strong selection, it has the potential to be sweep under that selection, becoming a major haplotype and allowing for drug escape.

The growth and adaptation of a viral population in the presence of drug selection is observed as viral rebound with drug resistance, and it can occur rapidly due to a short viral life cycle. The time needed for an HIV virion to attach, replicate, and produce progeny that infect new cells is estimated at 2 days, with a maximum of approximately 10^{10} virions produced per day within a person (31). In contrast, HCV's replication cycle has a half life of 2.7 hours (32), and approximately 10^{12} virions are produced daily (32-34). The high production rate of new mutations introduced into the small, ~10kb viral genomes results in rapid diversification of viral populations within a host. For both viruses, it has been theoretically estimated that at any given time within a large population, every single mutation can exist (35-37).

If a drug was prescribed to an individual infected with HIV or HCV, and their intrahost population had variants containing resistance to the applied drug, viral escape and rebound may rapidly occur. Genotypic testing for drug resistance mutations is recommended before starting drug therapy (6), and this technology will reveal variants ≥20% in frequency (38, 39).

The clinical significance of pre-existing resistance alleles has not been clearly elucidated, and likely will be unique per drug and treatment regiment. Previous studies using allele-specific PCR have implicated that pre-existing variants may preclude increased susceptibility for therapy failure (40, 41). Deep sequencing studies have produced correlative (42), partial (43) or non-correlative (44) results on the impact of pre-existing resistance alleles, but none have examined the haplotypes on which they reside. Understanding if pre-existing resistance alleles can be selected under a drug can further inform therapeutic choices and circumvent suboptimal selection. Aside from cost and morbidity associated with therapy failure, the outgrowth of drug resistance also complicates subsequent therapeutic intervention, as drugs within a class commonly contain overlapping drug resistance mutations, thus further narrowing downstream drug choices for viral suppression (6).

When a patient fails therapy, a population of susceptible viruses bottlenecks under drug selective pressures, but then grows out with resistance. Virologic rebound requires genomic change that confers a phenotype allowing escape of drug selection and selection for that phenotype under a drug. In order to better understand how a viral population evolves resistance, several questions need to be addressed:

Do new resistant variants arise during selection or do they grow out from the standing genetic variation? An extraordinarily high mutation rate adds variants *de novo*, rapidly diversifying the standing genetic variation. When a population rebounds with resistance, is this due to *de novo* resistance mutations arising during the selective pressure (not a result of it), or the sweeping of pre-existing resistant variants in the standing genetic variation?

Is the evolutionary path to resistance consistent? Escaping a selective event may only require a small number of discrete mutations. *In vivo* studies have demonstrated this process to be a sequential, step-wise addition of resistance and compensatory alleles (45-47), indicating that the path to resistance may be conserved across populations. Alternative *in vivo* studies, however, indicate that the patterns are more complex (48), therefore individual populations may take unique paths to resistance.

Antagonistic epistasis has been shown to occur between beneficial alleles (49). If two adaptive alleles arise on different haplotypes, clonal interference predicts that their progression to fixation will be slower than if there was only one beneficial allele sweeping to fixation (50). Interference has been demonstrated in RNA populations such as vesicular stomatitis virus (VSV) (51), but has not been clearly demonstrated *in vivo*. The Red Queen Hypothesis, or the perpetual arms race amongst competing viral subpopulations, is perhaps better supported for RNA viruses in the face of drug selection due to a “leap-frog” effect of adaptive (or maladaptive) alleles (52, 53) versus smooth climbs to fitness peaks. Although many of these studies are from *in vitro* and *in silico* observations, they point to unique paths to adaptation.

Do deterministic or stochastic processes dictate the fate of new mutations? Whether selection or drift determines which alleles survive a population bottleneck is directly related to the effective population size (N_e). For HIV, arguments for both large populations governed by deterministic processes (54) and small populations driven by stochastic models (55-57) have been made. If stochastic forces govern a population it will take a unique path to resistance, as the fate of new mutations will be driven by drift (58). Similarly, pre-existing resistance mutations will be non-informative, as they are likely lost. If deterministic forces shape population response, pre-existing alleles likely are retained through a population bottleneck, increasing their probability of selection and expansion.

Does recombination play a large role in shaping the evolution of haplotypes? If two alleles are associated or disassociated non-randomly, their appearance on a haplotype may be indicative of a non-additive change in fitness (ω). Linkage disequilibrium (D) that is maladaptive can be broken up by recombination (59), allowing escape from the effects of Muller's ratchet and error catastrophe. For non-recombining viruses, a single allele needs to rise to fixation then gain a second beneficial allele from mutation for two beneficial alleles to arise on a single haplotype. HIV has a documented high recombination rate, but whether recombination or the *de novo* addition of new alleles during the path to resistance plays a larger role has not been clearly demonstrated.

Does suboptimal compliance increases population diversity. A rough genotypic fitness landscape has been experimentally demonstrated for the RNA virus ϕ6 (60). Environmental effects further complicate Fisher's (phenotypic) landscape. Environmental heterogeneity promotes population diversity by creating multiple fitness peaks. Disturbance, like selection, can destroy or promote diversification. Ecological succession has demonstrated maintenance of diversity through intermediate disturbances; organisms are not killed so frequently that only a few can survive at any given time yet the environment is not so stable that superior competitors can quickly displace (61). Applying disturbance ecology to intrahost selection would predict that inconsistent, intermediate drug selection would result in the emergence of multiple resistance alleles. Drug concentration heterogeneity, thus increasing a range for the path to resistance, has been demonstrated to increase the rate at which resistance is developed (62).

How does a co-infecting pathogen affect viral diversity? The persistent infection of more than one pathogen in a host is common, but the complexity of within-strain and within-host interactions make it difficult to predict how strains influence each other and shape disease outcome (63, 64). It is known that HCV-HIV co-infection increases patient morbidity and mortality (65) by causing a three fold acceleration in fibrosis, cirrhosis, and liver disease (66). The biologic effects and clinical observations of mono- versus co-infection suggest differences in intrahost viral diversity, but previous studies have failed to demonstrate a difference (67, 68), or consensus (69-72).

1.2 Sequencing Approaches for Minor Alleles

The structure of an intrahost viral population and the path it takes to resistance has large theoretical and clinical implications. To date, these questions have only been explored from techniques that have limited and/or biased sampling of the extant diversity.

1.2.1 Common methods for genotyping resistance may obscure the origins and nature of resistance alleles.

In the clinical setting, genotypic assays that resolve drug resistance mutations use Sanger-based sequencing technology. A population of viral templates is sampled from an individual, sequenced, and then reported as a single consensus sequence. The consensus sequence captures the most common alleles, and high frequency allelic variants can be detected and quantified *de novo* by assessing chromatogram peaks. However, variants below 20-25% in frequency are typically not resolved (38, 39), and accuracy is further compounded by laboratory-introduced biases (73). Furthermore, the association, or linkage, between different variable sites is lost.

There are alternative methods to resolving variants below the resolution of Sanger sequencing. Allele-specific PCR, for example, uses primers specific to drug resistance polymorphisms on individual sites within the viral genome. Due to the sensitivity of PCR, this potentiates the resolution of variants present in less than 1% in of the population. However, targets require *a priori* selection of sites and variants, and akin to Sanger sequencing, linkage is lost (74-83).

Thus, the detection of allelic variants has been constrained by either a *de novo* analysis of sequencing variation but at the cost of a low resolution, or by an *a priori* resolution of individual minor variants but at the cost of losing linkage and novel uncharacterized alleles. Furthermore, these assays are labor-intensive and do not lend themselves to high-throughput techniques. Although resistance thresholds that are clinically relevant have been poorly defined, there is evidence that low-abundance drug resistant variants are selected under drug therapy, resulting in virologic failure. Clearly defining the presence and frequency of drug resistance mutations prior to antiviral treatment that result in failure could prevent suboptimal therapies and incomplete viral suppression.

1.2.2 PCR artifacts and a high sequencing error rate bias deep sequencing resolution of viral populations.

The high throughput, thus high-resolution capabilities of next generation sequencing platforms has a great potential to be applied within a clinical setting in resolving minor drug resistance mutations *de novo*, retain linkage across a template, and provide drug resistance screening at a lower cost per sample than standard genotypic assays. However, the high error rate inherent from the sequencing chemistry in combination with laboratory-introduced biases has limited the clinical application and resolution of minor variants to levels below that of population sequencing but above what allele-specific PCR can obtain. Therefore, the utility of next generation platforms for sequencing of viral populations derived from clinical samples is restricted.

High throughput platforms require a large amount of starting material, typically 500ng-1ug of DNA. Patient-derived samples contain a limited number of viral templates, therefore PCR is an necessary first step prior to a sequencing protocol. Significant biases are introduced into the viral population by PCR. Polymerase, during its many rounds of copying, will misincorporate nucleotides. This inflates the genetic diversity of the population. When incomplete templates prime a subsequent round of synthesis, chimeric genomes, or recombinants, are produced (84, 85). PCR-mediated recombination not only disrupts linkage between sites, but also creates artifactual linkage. Templates entering at different PCR cycles will result in some genomes amplifying more than others, resulting in a skewing of allele frequencies.

Another major limitation of viral template PCR amplification prior to a deep sequencing protocol is the re-sampling of amplified templates, or PCR re-sampling. PCR reactions typically start with an unknown number of templates. When PCR efficiencies are not 100%, the random dropout of sequences results in an underestimation of diversity. Moreover, sampling of a large, amplified population does not equate with sampling of individual genomes, therefore depth is a correlate of the amount of starting material, not degree of re-sampling (86).

1.2.3 Consensus sequences constructed from amplified products derived from an individual template resolve PCR and sequencing error.

Strategies that create a consensus sequence, such as Single Genome Amplification (SGA), will call the correct base at each position (87-90). In the SGA strategy, PCR amplification is preceded by endpoint dilution titration, such that a single template is present per reaction. During amplification, PCR biases will introduce error, but the influence of diversity from recombination events and differential template amplification will be masked by overall sample homogeneity. Although misincorporation occurs, it is randomly distributed across the template, therefore the majority of reads per site will be correct. With traditional sequencing methods, minor variants due to misincorporation will be masked, resulting in the correct base call per position. While this approach is effective for eliminating the PCR and sequencing biases, it does not lend itself to high throughput techniques and has limited utility when applied to a large viral population.

High throughput sequencing can resolve PCR biases because each individual template from a PCR reaction has the potential to be sequenced. The task of extracting biological polymorphisms from PCR biases and sequencing error has been farmed out to an ever-increasing number of software and bioinformatics tools that range in utility, algorithmic complexity, and degree of auxiliary analyses (91-94). However, the unifying factor that ties together these tools is that they are all assessing error from biological polymorphisms indirectly.

Indirect, or statistical inference of biological diversity can generally resolve major variants. However, minor biological variants whose frequency nears the error threshold are greatly skewed or lost due to procedural biases. Therefore, an individual's resistance profile is fundamentally limited to variants whose frequency is well above the sequencing error and influence of PCR biases, and PCR re-sampling continues to remain uncorrected for. In order to make high throughput sequencing useful in resolving minor drug resistant variants and determining their role in virologic outcome, PCR biases, sequencing error, and PCR re-sampling all need to be directly overcome. Only after procedural error is removed can one examine a population accurately for drug resistance and adaptation to drug selective pressures.

The high evolvability of HIV and HCV is due to selection of adaptive alleles within the standing genetic variation, and understanding this process may reveal how to prevent it. My work begins to explore the path to drug resistance by finely resolving the structure of intrahost populations as it overcomes selection. Why individuals fail therapy is likely a question that will need to be re-interpreted on a per-patient basis, as my work suggests a complex interplay between selection pressure, viral diversity, and population dynamics.

Chapter 2

Accurate Sampling and Deep Sequencing of the HIV-1 Protease Gene Using a Primer ID

Reprinted with modification from: Jabara C.B., Jones C.D., Roach J., Anderson J.A., and Swanstrom R. (2011) Accurate Sampling and Deep Sequencing of the HIV-1 Protease Gene Using a Primer ID. *Proc Natl Acad Sci USA* 108:20166-71.

2.1 Abstract

Viruses can create complex genetic populations within a host, and deep sequencing technologies allow extensive sampling of these populations. Limitations of these technologies, however, potentially bias this sampling, particularly when a polymerase chain reaction (PCR) step precedes the sequencing protocol. Typically, an unknown number of templates are utilized in initiating the PCR amplification and this can lead to unrecognized sequence resampling creating apparent homogeneity; also PCR-mediated recombination can disrupt linkage, and differential amplification can skew allele frequency. Finally, misincorporation of nucleotides during PCR and errors during the sequencing protocol can inflate diversity. We have solved these problems by including a random sequence tag in the initial primer such that each template receives a unique Primer ID. After sequencing, repeated identification of a Primer ID reveals sequence resampling. These resampled sequences are then used to create an accurate consensus sequence for each template,

correcting for recombination, allelic skewing, and misincorporation/sequencing errors. The resulting population of consensus sequences directly represent the initial sampled templates. We applied this approach to the HIV-1 protease (*pro*) gene to view the distribution of sequence variation of a complex viral population within a host. We identified major and minor polymorphisms at coding and noncoding positions. In addition, we observed dynamic genetic changes within the population during intermittent drug exposure, including the emergence of multiple resistant alleles. These results provide an unprecedented view of a complex viral population in the absence of PCR resampling and artifactual error.

2.2 Introduction

High throughput sequencing allows the acquisition of large amounts of sequence data that can encompass entire genomes (95-98). With sufficient amounts of starting DNA, PCR is not needed prior to the library preparation step of the sequencing protocol. Sequencing miscalls inherent in high throughput sequencing approaches are resolved using multiple reads over a given base.

Deep sequencing can also capture the genetic diversity of viral populations (72, 99-103), including intrahost populations derived from clinical samples. This approach offers the opportunity to view population diversity and dynamics, and viral evolution in unprecedented detail. One place where the presence of minor variants is of immediate practical importance is in the detection of drug resistant variants. Standard bulk sequencing methods, however, will miss allelic variants below 20% in frequency within a population (38, 39). Alternative assays can detect less abundant variants that confer drug resistance, but require *a priori* selection of sites and variants (74-83, 104). Thus deep sequencing approaches offer the opportunity to identify minor variants associated with resistance *de novo* with the goal of understanding their role in therapy failure.

While screening for drug resistant variants is a practical application of the deep sequencing technology, this technology also addresses broader questions of sequence diversity and structure for a complex population like HIV-1. However, the relatively high sequencing error rates of these technologies artificially increase genetic diversity, which confounds the detection of natural genetic variation especially when sequencing a highly heterogeneous viral population (31, 105, 106). Moreover, the use of PCR to amplify the amount of material prior to starting the sequencing protocol adds the potential for several serious artifacts (107): first, nucleotide misincorporation by polymerase during the many rounds of amplification artificially increases sequence diversity; second, artifactual recombination during amplification occurs when premature termination products prime a subsequent round of synthesis which can obscure the linkage of two sequence polymorphisms (84, 85); third, differential amplification can skew allelic frequencies; and fourth, PCR amplification can create a significant mass of DNA from a small number of starting templates, which obscures the true sampling of the original population as these few starting templates/genomes get resampled in the PCR product, creating sequence resampling rather than the observation of independent genomes (86). Overall, these biases artificially decrease true diversity while introducing artifactual diversity and also skew allelic frequencies which can lead to incongruence between the real and observed viral populations. Most investigators use statistical tools to attempt to control for the types of sequencing errors that are associated with each sequencing platform.

To make deep sequencing useful for complex populations it is necessary to overcome PCR resampling, which is mistaken for sampling of the original population, and PCR and sequencing errors, which can be mistaken for diversity. As nucleotide misincorporation is largely random across sites, and template switching/recombination is more likely to occur in the later cycles of a PCR reaction (108), strategies that create a bulk or consensus sequence for each sampled template will call the correct base at each position. One approach to sampling highly heterogeneous populations, such as the HIV-1 *env* gene, is through endpoint dilution titration of the template prior to nested PCR, such that a single template is present in each PCR amplification (87-90). In addition to masking the misincorporations, PCR-mediated recombination produces recombinant templates identical to the parental sequence. Although highly accurate, this technique is labor-intensive, and, as population sampling is dependent on the number of templates sequenced, this methodology does not lend itself to the identification of minor variants or to understanding the structure of a complex population, nor is it easily adaptable to a high throughput approach.

We have developed a new, high throughput technique for directly resolving the genetic diversity of a viral population. This technique avoids the recording of PCR and sequencing errors that create artificial diversity, and corrects for artificial allelic skewing and PCR resampling, revealing the original genomes in the population. This is accomplished by embedding a degenerate block of nucleotides within the primer used in the first round of cDNA synthesis. This creates a random library of sequences within the primer population. As primers are individually used out of this library, each viral template is copied such that the complement (cDNA) now includes a unique sequence tag, or Primer ID. This Primer ID is carried through all of the subsequent manipulations to mark all sequences that derive from each independent templating event, and PCR resampling then becomes over-coverage for each template to create a consensus sequence of that template. Using this approach, we were able to directly remove error, correct for PCR resampling, and capture the fluctuation of minor variants in the viral population within a host. We also resolved minor drug resistant variants below 1% in frequency prior to the initiation of antiretroviral therapy, and were able to correlate these variants with the emergence of drug resistance.

2.3 Materials and methods

vRNA extraction and cDNA synthesis. Viral RNA was extracted from three plasma samples taken longitudinally from an individual infected with subtype B HIV-1 who was participating in a protease inhibitor efficacy trial (M94-247). Two samples were collected at approximately 6 months before and immediately prior to the addition of the protease inhibitor ritonavir to a failed therapy regimen (plasma viral loads of 285,360 copies of viral RNA/ml and 321,100 copies of viral RNA/ml, respectively), and one sample was collected during ritonavir therapy (at approximately two months on therapy, 349,920 copies of viral RNA/ml) but during a time of apparent intermittent compliance. For each plasma sample, vRNA was extracted from pelleted (25,000 x g for 2 hours) viral particles using the QiaAMP Viral RNA Kit (Qiagen, Valencia, CA). Approximately 10,000 copies of viral RNA from each sample were present in the cDNA synthesis reaction as previously described (87, 109, 110). The tagging primer used was, 5’-GCCTTGCCAGCACGCTCAGGCCTTGCA(BARCODE)CGNNNNNNNNTCCTGGCTTTAATTTTACTGGTACAGT-3’. The barcode represented TCA, GTA, and TAT for study days 58, 248, and 303, respectively. The 3’ end of the tagging primer targeted downstream of the protease coding domain (HXB2 2568-2594). The oligonucleotides were purchased from IDT and were purified by standard desalting.

Amplification of tagged sequences. The single-stranded cDNA was column purified using the PureLink PCR Purification Kit (Invitrogen, Carlsbad, CA), using Binding Buffer HC (high cut-off) and 3X wash to remove the cDNA primer. Primer removal was verified by electropherogram analysis using an Experion HighSense RNA microfluidic chip (Bio-Rad Laboratories, Hercules, CA). Samples were amplified by nested PCR, using upstream primers 5'-GAGAGACAGGCTAATTTTTTAGG-3' (HXB2 2071-2093) and 5'-ATAGACAAGGAACTGTATCC-3' (HXB2 2224-2243); the downstream primers targeted the 5' portion of the cDNA tagging primer 5-GCCTTGCCAGCACGCTCAGGC-3' then 5'-CCAGCACGCTCAGGCCTTGCA-3'. The PCR was done using Platinum *Taq* DNA Polymerase High Fidelity (Invitrogen, Carlsbad, CA). Each reaction contained 1x High Fidelity PCR Buffer, 0.2 mM dNTP mixture, 2 mM $MgCl_2$, 0.2 μM of each primer, 1.5 units of Platinum *Taq* DNA Polymerase. For the first round, the purified cDNA template was split to 2x50ul for the first round PCR, and 1ul of the purified first round product was used for nested PCR. Samples were denatured at 94^{o}C for 2 minutes, followed by 30 cycles of 94^{o}C for 15 seconds, 55^{o}C for 30 seconds, 68^{o}C for 1 minute, and a final extension at 68^{o}C for 5 minutes.

Samples were column purified after the first round of PCR using the MinElute PCR Purification Kit (Qiagen, Valencia, CA), and eluted into 30ul of buffer EB. Second round PCR product was gel purified using a 2% agarose gel and QIAquick gel extraction kit (Qiagen, Valencia, CA), with incubation of the solubilization buffer at room temperature.

DNA was quantified by Qubit fluorometer using dsDNA High Sense assay (Invitrogen, Carlsbad, CA). Product generation, quality, and primer removal for both PCR rounds was verified using an Experion DNA microfluidic chip (Bio-Rad Laboratories, Hercules, CA).

454 pyrosequencing. Tagged samples from the three time points were combined and sequenced on the 454 GS FLX platform with XLR70 Titanium sequencing chemistry as per the manufacturer's instructions (Roche, Nutley, NJ) but with under-loaded beads to minimize signal crosstalk. Sequences were processed from two independent 454 GS FLX Titanium runs (1/8th of a plate each).

Bioinformatic pipeline for raw sequence processing. A suite of programs was written to filter and parse raw 454 sequencing reads. In short, first each sequence was placed in the correct orientation as compared to a reference *pro* gene sequence. This alignment was then used to identify insertions or deletions caused by the 454 sequencing of homopolymers. When there was an insertion, the extraneous base was excised from the sequence. Deletions retained were largely resolved in the construction of the consensus sequence (see below). Second, they were evaluated for the presence of the cDNA primer 5' tail, with the encoded information (barcode and primer ID) exactly spaced. Third, individual samples were binned according to their barcodes, and then to their individual the Primer ID. Fourth, sequences were trimmed to the protease coding domain (*pro* gene). Within a barcode bin, when three sequences contained an identical Primer ID, a consensus sequence was called by majority rule. Ambiguous nucleotide designations were used when there was a tie. Sequences are available under GenBank accession numbers JN820319-JN824997.

Population analyses. A chi-squared test was used to test for significance changes in allele frequency between the two untreated time points. To control for multiple testing, collective assessment of significance was based on False Discovery Rate analysis (FDR = 0.05). Tests for linkage disequilibrium were computed by DnaSP v.5.10.01 (111). These tests were done on filtered populations devoid of sequences containing ambiguities or gaps. Tests for neutrality were computed by DnaSP and R (112) on filtered populations devoid of sequences containing ambiguities. Gaps and alleles represented by a single sequence were reverted to the consensus. Beta *P*-values were calculated against the null hypothesis that $D = 0$, assuming that D follows a beta distribution after rescaling on [0, 1] (113).

Synonymous and nonsynonymous diversity across and within populations was computed through customized bioinformatics suites. Unfiltered sequences were used in the analysis, and ambiguities, gaps, and alleles represented by a single sequence were removed from the final tabulation.

SNPs were graphically displayed through the *Highlighter* tool (www.hiv.lanl.gov).

Phylogenetic resolution of sequences. The phylogeny for the population of consensus sequences from all three time points was resolved using two alternative methods and on populations devoid of sequences containing gaps or ambiguities. When only one example of a SNP was present across all sequences, it was converted to the consensus on the assumption that it was likely generated by residual method error. First, the Neighbor-Joining tree using the Kimura translation for pairwise distance and a bootstrap of 100 iterations was constructed with QuickTree v.1.1 (114).

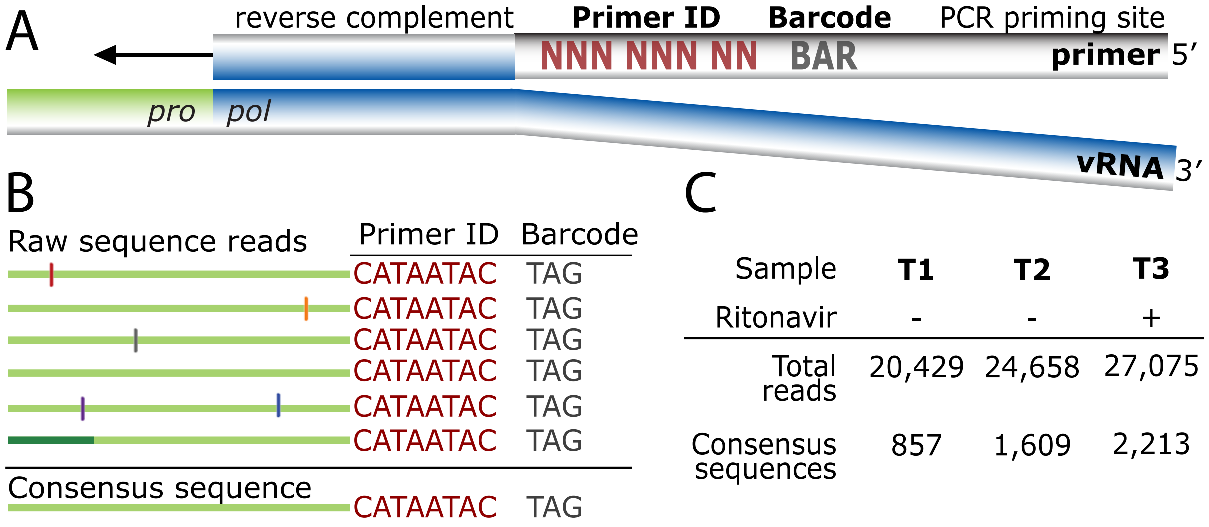
Second, Maximum likelihood phylogeny was inferred using the PHYLIP package, version 3.69 (115), and the calculated phylogeny is available upon request. The PHYLIP program *seqboot* was used to create 100 bootstraps. Resulting bootstraps were submitted to the PHYLIP program *dnamlk* for maximum likelihood inference subject to a strict molecular clock. The consensus tree of all boostrap results was constructed using the PHYLIP program *consense*.

Both phylogenetic trees were visualized by a customized modification of Figtree v.1.3.1. (116)

2.4 Results

A cDNA synthesis primer containing a Primer ID can be used to track individual viral templates.

A population of cDNA synthesis primers was designed to prime DNA synthesis downstream of the HIV-1 protease (*pro*) gene, with the primer containing two additional blocks of identifying information (Fig. 2.1*A*). The first block was a string of eight degenerate nucleotides that created 65,536 distinct sequence combinations (4^8), or Primer IDs. This region was flanked by an *a priori* selected three nucleotide barcode, creating a sample identification block so that multiple samples could be pooled together in a sequencing run (100). A designed sequence at the 5' end of the cDNA primer was used for subsequent amplification of the cDNA sequences by nested PCR.



Sample	T1	T2	T3
Ritonavir	-	-	+
Total reads	20,429	24,658	27,075
Consensus sequences	857	1,609	2,213

Fig. 2.1 Tagging viral RNA templates with a Primer ID before PCR amplification and sequencing allows for direct removal of artifactual errors and identifies resampling. (A) A primer was designed to bind downstream of the protease coding domain. In the 5′ tail of the primer, a degenerate string of eight nucleotides created a Primer ID, allowing for 65,536 unique combinations. An a priori selected three nucleotide barcode was designed for the sample ID. Finally, a heterologous string of nucleotides with low affinity to the HIV-1 genome was included in the far 5′ end for use as the priming site in the PCR amplification. (B) PCR biases and sequencing error are introduced during amplification and sequencing of viral templates. Repetitive identification of the barcode and Primer ID allow for tracking of each templating event from a single tagged cDNA. As errors are minor components within the Primer ID population, forming a consensus sequence directly removes them, and corrects for PCR resampling. (C) HIV-1 RNA templates isolated from plasma samples from two pre- and one post-intermittent ritonavir drug therapy were tagged, amplified, and deep sequenced. Tagged sequences containing full- length protease were used to create a population of

consensus sequences when at least three sequences contained an identical barcode and Primer ID.

Viral RNA was extracted from three longitudinal blood plasma samples from an individual infected with subtype B HIV-1 who was participating in a protease inhibitor efficacy trial (M94-247) (117) (Fig. 2.2). Approximately 10,000 copies of viral RNA from each sample were used in a reverse transcription reaction for cDNA synthesis and tagging using the Primer ID. The cDNA product was separated from the unused cDNA primers, then the viral sequences were amplified by nested PCR and sequenced on the 454 GS FLX Titanium. Our data were distilled from total reads of 20,429, 24,658, and 27,075 for the three time points (T1, T2, and T3, respectively). Raw sequence reads were assessed for the cDNA tagging primer and a full length *pro* gene sequence (297 nucleotides long representing 99 codons), and when three or more sequences within a sample contained an identical Primer ID, a consensus sequence was formed to represent one sequence in the population (Fig. 2.1*B*, 1*C,* S*2*).

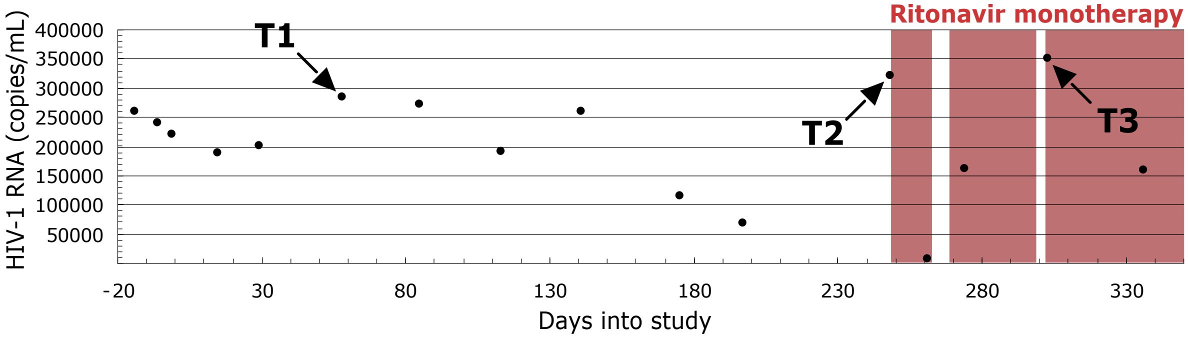


Fig. 2.2 Longitudinal sampling of blood plasma from a single individual infected with HIV-1 subtype B pre- and post- a failed ritonavir monotherapy regime. Two time-points ~6 mo apart were sampled before ritonavir therapy (T1 and T2). One time point was sampled

after failed, intermittent ritonavir monotherapy (T3). The shaded areas represent times of therapy compliance based on self-report.

With these manipulations we generated 857, 1,609, and 2,213 consensus sequences, respectively, for the three time points (Fig. 2.1*C*). The median number of reads per Primer ID was 6, ranging from 1 to 96 (Fig. 2.3). The distribution of identical Primer IDs did not form a normal distribution as would be expected if all templates were amplified equally. We saw a higher than expected number of single reads of Primer IDs; although we do not know the reason for this, such a result is consistent with different cDNA templates entering the PCR at different cycles. Since each template is individually tagged the different number of reads is an indication of allelic skewing, as noted this can be nearly 100 fold. In an analysis of a number of low abundant variants we saw a 20-fold range of representation through allelic skewing, with half of the variants up to 2-3 fold more abundant than the mean, and the other half up to 5-10 fold less abundant.

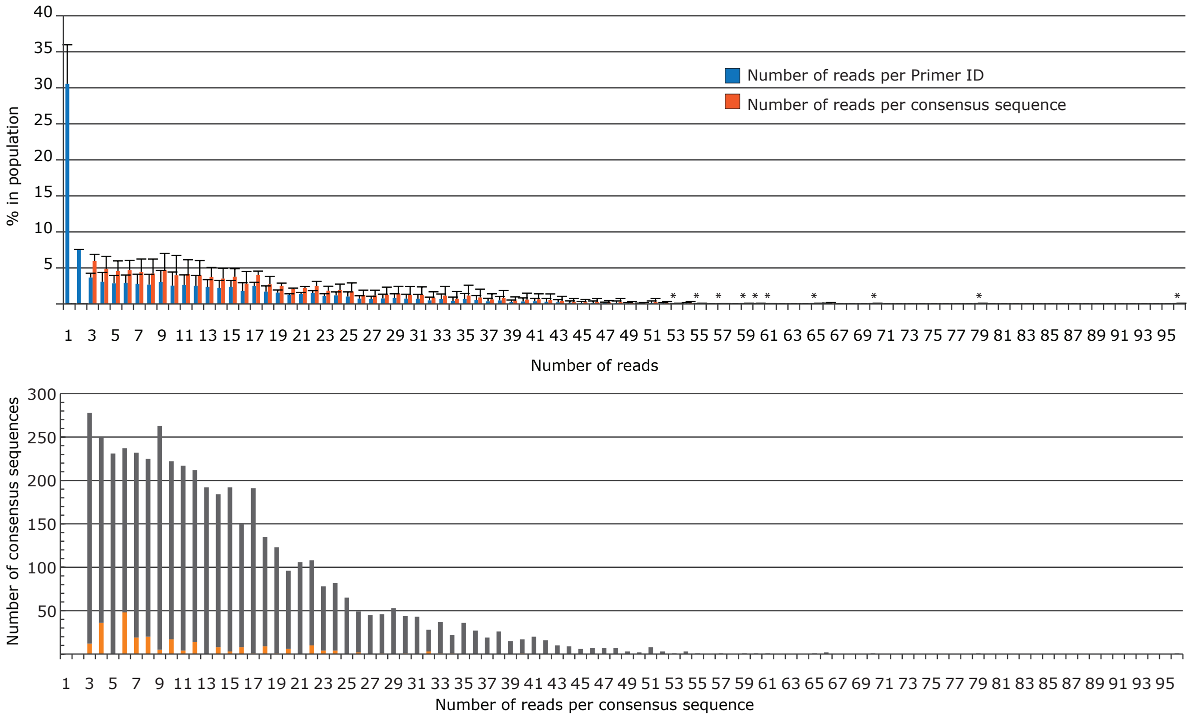


Fig. 2.3 Distribution of the number of reads per Primer ID or consensus sequence. A) Blue bars represent the distribution of resampling of the filtered sequence population immediately before consensus sequence generation. Within a single Primer ID, when three or more sequences were present, a consensus sequence was formed. The orange bars represent the distribution of the number of reads that went into each consensus sequence. The values shown represent the mean for the data from the three time points with the error bars representing the SD between the three samples. Starred bars are included to mark positions where a single sequence had high resampling occurrence. (B) Number of consensus sequences containing an ambiguity as a function of extent of resampling. All three time points were combined. Gray bars represent consensus sequences without an ambiguity, and orange bars represent consensus sequences with an ambiguity. There is a discernible pattern of an increased number of ambiguities going out to 22 reads/consensus sequence for those

consensus sequences created from an even number of reads, the result of having a tie between two different sequences at one position. However, this represents only a small fraction of the total reads (5.4%). The amino acid position with the highest ambiguity total was used per Primer ID subpopulation.

We conservatively estimate the combined in vitro error rate of the cDNA synthesis step by reverse transcriptase and the first strand synthesis by the *Taq* polymerase to be on the order of 1 mutation in 10,000 bases, or approximately one mutation per 33 *pro* gene sequences, based on an RT error rate of 1 in 22,000 nucleotides (118) and a *Taq* polymerase error rate of 1.1 in 10,000 nucleotides (119) but reduced by half since only the first round of synthesis is relevant and a misincorporations at this step gives a mixture. Later rounds of *Taq* polymerase errors should be largely lost through the creation of the consensus sequence. Thus we would expect 139 sequence misincorporations to be present in the data set of 4,679 total sequences representing T1+T2+T3, and with an excess of transitions. These would be expected to occur as 113 single copy single nucleotide polymorphisms (SNPs) and 13 SNPs that appeared twice. We observed 98 single copy SNPs in the data set with a 3 fold excess of transitions, and with three-fourths of them being coding changes, which is consistent with random mutations. We expect there to be low frequency SNPs in the viral population from rare but persistent variants that are fortuitously sampled, and from the intrinsic error rate of viral replication (the error rate during one round of viral replication would represent approximately one mutation per 150 *pro* gene sequences (105)). However, we cannot distinguish real polymorphisms from the inferred background error rate associated with the first and second rounds of in vitro DNA synthesis. Thus we have limited the analysis of population diversity to SNPs that appeared at least twice in the data set (i.e. linked to at least two separate Primer IDs), either at the same time point or at multiple time points in the overall data set (Table 2.1). We have not corrected the data set for the presumed 13 SNPs that appeared twice that are expected to be present due to error even though this represents 33% of all of the SNPs that appeared twice (13 of 39). Overall, 80% of the SNPs (i.e. any sequence change from the consensus that appeared at least once) in the total data set of 72,162 sequence reads were removed as error. Also, 60-65% of the sequence reads were revealed as resampling. Finally, allelic skewing of up to nearly 100 fold was corrected (Fig. 2.4).

Table 2.1 Frequency of nonconsensus codons per position

Consensus			Nonsynonymous							Synonymous					
AA$_{pos}$[a]	AA$_c$[b]	C$_c$[c]	C$_m$[d]	AA$_m$[e]	T1[f]	T2[g]	T3[h]	T3$_s$[i]	T3$_r$[j]	C$_m$[k]	T1[l]	T2[m]	T3[n]	T3$_s$[o]	T3$_r$[p]
4	T	ACT	GCT	A		0.06	0.05		0.09						
5	L	CTT	CCT	P	0.12		0.05	0.14							
7	Q	CAA								CAG	0.35	0.12	0.09	0.14	0.09
8	R	CGA								CGG	0.12		0.05	0.14	
9	P	CCC													
10	L	CTC	TTC	F		0.19				CTT		0.19			
11	V	GTC	ATC	I	0.23	0.25				GTT		0.12			
14	K	AAG	AGG	R		0.12				AAA	1.17	0.19	0.59	0.29	0.72
15	I	ATA	GTA	V	1.17	0.12	0.14	0.14	0.18	ATC			0.09		0.18
16	G	GGG	AGG	R		0.06	0.05		0.09	GGA	2.22	3.54	38.86	17.70	45.97
17	G	GGG	AGG	R			0.09	0.29		GGA	0.35	0.19	0.18	0.43	0.09
18	Q	CAA	GAA	E	0.23	0.12				CAG	18.55	21.75	6.46	12.81	3.53
19	L	CTA	ACA	T	0.47										
			ATA	I	19.25	19.83	20.42	19.28	24.98	TTA	0.12	0.19	0.09	0.29	
			GTA	V	3.38	5.66	46.00	25.61	52.76						
20	K	AAG	AGG	R	0.12	0.12	0.05		0.09	AAA		0.31	0.86	0.29	1.27
21	E	GAA								GAG	0.12	0.06	0.05	0.14	
22	A	GCT								GCC	0.47	0.44	0.27	0.58	0.18
										GCG	0.23				
23	L	CTA								CTG		0.19			
24	L	TTA								CTA	0.35	5.72	1.31	2.16	0.63
										TTG	12.49	0.81	0.59	1.01	0.27
25	D	GAT	GGT	G	0.12	0.12				GAC	0.23	0.93	0.05	0.14	
26	T	ACA	GCA	A		0.12									
27	G	GGA								GGG	0.12	0.06			
28	A	GCA								GCG	0.12		0.09	0.14	
29	D	GAT	AAT	N	0.12		0.05		0.09	GAC	0.23	0.19			
30	D	GAT								GAC		0.06	0.09	0.14	0.09
31	T	ACA								ACG		0.12			
32	V	GTA								GTG		0.25			
33	L	TTA	GTA	V	0.47	0.06				CTA		0.25	0.14	0.29	0.09
										TTG	0.35	0.12	0.14	0.43	
34	E	GAA	GGA	G		0.12	0.05		0.09	GAG	0.12		0.05	0.14	
			CAA				0.09								
35	E	GAA	AAA	K	0.12	0.06	0.09	0.14							
36	M	ATG	ATA	I	0.82	0.81	0.27	0.43	0.27						
37	N	AAT	AGT	S		0.19	0.05			AAC		0.06	0.05	0.14	
			GAT	D	2.33	2.30	0.95	0.86	1.27						
38	L	TTG								TTA	0.23	0.62	0.05		0.09
39	P	CCA								CCT	0.23				
40	G	GGA								GGG	0.12	0.12			
41	K	AAA	AGA	R		0.06	0.18	0.14	0.27	AAG	4.08	1.43	0.50	1.15	0.27
42	W	TGG	CGG	R	0.12	0.06									
			TAG	_	0.12		0.05		0.09						
			TGA	_			0.14		0.27						
43	K	AAA	AGA	R		0.06	0.05		0.09	AAG	0.35		0.14	0.14	0.18
44	P	CCA								CCG		0.06	0.23	0.43	0.18
45	K	AAA	AGA	R	0.12	0.12	0.05		0.09	AAG	0.58	0.99	0.41	1.29	
46	M	ATG	ATA	I		0.12	0.09	0.14	0.09						
48	G	GGA	GAA	E			0.14	0.14	0.18	GGG	0.35	0.19			
49	G	GGA	GAA	E	0.12	0.06	0.05		0.09	GGG	0.23	0.12			
50	I	ATT								ATC	0.12	0.12			
51	G	GGA								GGG	0.12	0.06			
52	G	GGT	AGT	S	0.12	0.06	0.05	0.14		GGA		0.06	0.05	0.14	
										GGC	0.12	0.31	0.09	0.14	0.09
										GGG			0.14	0.43	
53	F	TTT								TTC	0.70		0.05	0.14	
54	I	ATC	ACC	T	0.12	0.06	0.05		0.09	ATT	0.35	0.06	0.14	0.14	
55	K	AAA	AGA	R	0.12		0.05		0.09	AAG	0.12	0.06			
56	V	GTA	ATA	I	0.12		0.05	0.14		GTG		0.75	0.14	0.14	0.18
57	R	AGA	AAA	K	0.23					AGG	0.23	0.87	0.14	0.14	0.18

58	Q	CAG	TAG	_			0.05		0.09	CAA	0.93	0.50	0.23	0.29	0.27
60	D	GAT	AAT	N		0.12									
			GGT	G		0.12									
61	Q	CAA	CGA	R	0.12	0.06	0.05	0.14		CAG		0.19	0.23	0.58	
			TAA	_	0.12	0.06	0.05		0.09						
62	I	ATA	GTA	V	0.35	0.06									
63	L	CTC	CCC	P	0.12		0.41	0.58	0.36	CTT	11.32	5.41	1.27	2.88	0.45
64	I	ATA	GTA	V	1.05	0.06	0.09		0.18						
			ATG	M	0.23		0.05	0.14							
65	E	GAA	AAA	K			0.09	0.14	0.09	GAG	0.35	0.06	0.05		
66	I	ATC								ATA		0.25	0.18	0.58	
										ATT	1.98	0.19			
67	C	TGT								TGC	0.35	0.12	0.05	0.14	
68	G	GGA								GGG	0.23	0.12	0.05	0.14	
69	H	CAT	TAT	Y	0.23	0.06	0.09	0.14		CAC	0.82	0.31	0.14	0.29	0.09
70	K	AAA	CAA	Q	0.47	0.12	0.41	1.29		AAG	3.27	10.88	15.27	6.62	25.34
71	A	GCT	ACT	T		0.12	0.09								
72	I	ATA	GTA	V	0.12	0.12									
73	G	GGT								GGC	0.47	18.09	7.05	15.68	3.62
74	T	ACA								ACG	0.23	0.12			
75	V	GTA	ATA	I	0.23	0.06	0.05			GTG	1.87	0.99	0.27	0.43	0.27
			GCA	A			0.09		0.18						
76	L	TTA								CTA		0.12	0.09		0.18
										TTG	0.93	0.62	0.27	0.43	0.18
77	V	GTA	ATA	I	0.23	0.56	0.72	2.01	0.18	GTG	0.82	0.62	0.23	0.58	
			CTA	L			0.14								
78	G	GGA								GGG	1.17	1.24	0.09	0.14	
79	P	CCT								CCC	1.17	0.31	0.54	1.29	0.18
81	P	CCT								CCC	0.12	0.19			
										CCG	1.52	0.44			
82	V	GTC	ATC	I		0.06	1.27	3.60		GTA	0.35	0.31	0.05		
			CTC	L		0.06	1.08	3.45		GTT	1.05	0.75	0.41	1.01	
			GCC	A		0.12	49.89		99.91						
			TTC	F			0.14	0.43							
83	N	AAC	AGC	S	0.12		0.05		0.09	AAT	8.17	6.40	3.62	4.75	4.16
84	I	ATA	GTA	V			5.15								
85	I	ATT								ATA		0.12	0.05	0.14	
										ATC	0.12	0.12	0.05		
86	G	GGA								GGG		0.12			
										GGT	0.12	0.06			
87	R	AGA	AAA	K	0.12	0.06	0.05		0.09	AGG	0.58	0.37	0.05	0.14	
			GGA	G		0.06	0.09	0.14	0.09						
88	N	AAT								AAC	0.35	0.93			
89	L	CTA	ATA	I		0.12				CTG	1.17	0.68	1.36	1.87	1.54
										TTA	1.98	0.56	1.27	0.14	2.44
90	L	TTG	ATG	M	0.12		13.56		0.09	CTG	0.47		0.09	0.14	0.09
			TCG	S	0.12		0.05		0.09	TTA	0.47	0.19	0.14	0.43	
91	T	ACT	GCT	A		0.06	0.05		0.09	ACC	0.12	0.06	0.09	0.14	0.09
										ACG	0.12	0.12	0.77		1.54
92	Q	CAG								CAA	0.23	0.19	0.14		
93	I	ATT	CTT	L	0.12	0.06				ATC	0.23		0.09	0.14	0.09
94	G	GGT	GAT	D	0.12	0.06				GGA	0.23				
										GGC	1.28	0.25	0.50	1.29	0.18
										GGG	0.23	0.06	0.09	0.14	
95	C	TGC								TGT	0.70	0.12	0.14		0.27
96	T	ACT								ACA	0.12		0.09	0.14	0.09
										ACC	0.70	0.12	0.23	0.43	0.09
										ACG		0.06	0.05	0.14	
97	L	TTA								CTA	0.58		0.05	0.14	
										TTG	0.12	0.25	0.27	0.43	0.27
98	N	AAT								AAC	0.23	0.12	0.14		0.18
99	F	TTT	CTT	L		0.06	0.18	0.29	0.09	TTC	1.05	0.50	1.54	1.44	1.54
			GTT	V	0.23										

[a] Amino acid position, protease.
[b] Consensus amino acid in untreated population.
[c] Consensus codon in untreated population.
[d] Coding nonconsensus amino acid.
[e] Coding nonconsensus codon.
[f] Frequency of SNP in first untreated time point.
[g] Frequency of SNP in second untreated time point.
[h] Frequency of SNP in third time point, treated.
[i] Frequency of SNP in third time point, treated, susceptible population (not V82A, I84V, L90M).
[j] Frequency of SNP in third time point, treated, population containing major ritonavir resistant variant V82A.
[k] Silent nonconsensus codon.
[l] Frequency of SNP in first untreated time point.
[m] Frequency of SNP in second untreated time point.
[n] Frequency of SNP in third time point, treated.
[o] Frequency of SNP in third time point, treated, susceptible population (not V82A, I84V, L90M).
[p] Frequency of SNP in third time point, treated, population containing major ritonavir resistant variant V82A.

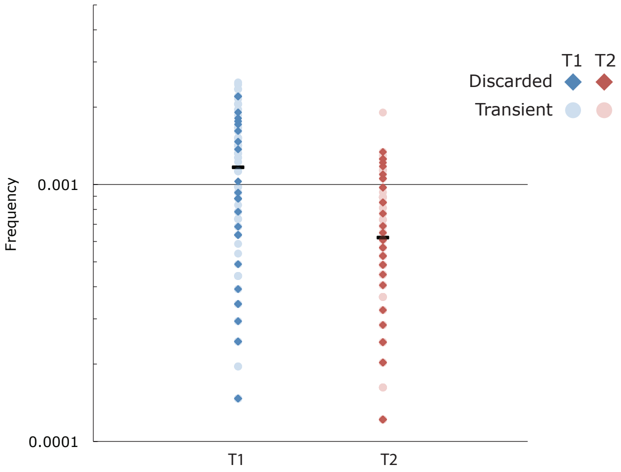


Fig. 2.4 Analysis of low abundance variants for the distribution of allelic skewing. We used discarded sequences (i.e., unique sequences represented by a single Primer ID) and transient genomes defined as having a low abundance SNP in the preconsensus population per untreated time point. Transient sequences were defined as having at least two sequences at only one of the untreated time points, or one copy at one of the untreated time points and

then again at the third time point. These sequences were used to define a set of sequences that could be compared for low frequency abundance in the total data set versus the consensus sequences. The horizontal bars represent the measured frequency of a single copy sequences in the consensus population at T1 and T2. Dark points represent discarded genomes, and light points represent transient genomes with their position indicating their abundance in the total sequence population before construction of the consensus sequences. Blue points represent sequences present at T1, red points represent sequences present at T2. These data show that allelic skewing of 2-fold upward and 10-fold downward is common before the formation of the consensus sequence.

Longitudinal sequencing of the HIV-1 protease (*pro*) gene in an untreated individual reveals dynamic changes in genetic variation.

We analyzed the sequences of the *pro* gene populations to assess allelic frequency at the two sampled time points, separated by 6 months and prior to ritonavir (117) drug selection (Fig. 2.1). The combined sequence population from the two time points (T1 and T2) before therapy consisted of 492 unique *pro* gene sequences with 155 SNPs. About 4% (i.e. 21) of these unique gene sequences were above 0.5% abundance, and these 21 unique gene sequences represented 67% of all sampled genomes, with the genome representing the overall consensus sequence comprising 21% of the total population (Fig. 2.5*A*, 2.5*B*). The relatively small number of unique gene sequences above 0.5% frequency in the population contained only 7% of the 155 detected SNPs. Thus, a large proportion of the viral population's diversity was associated with a large number of *pro* gene sequences that were present at low abundance (Fig. 2.5*A*, 2.5*C*); conversely the majority of the population consisted of a small number of SNPs. Similarly, Tajima's *D* statistic for T1 and T2 in this individual were -2.47 and -2.48, respectively (Table 2.2), indicative of a population structure that has suppressed levels of neutral mutations. This pattern is consistent with but more extreme than that observed in a prior shallow intrahost survey in which a metapopulation model was proposed to explain the pattern of Tajima's *D* statistic (55). Figure 2 shows the amino acid variability and synonymous nucleotide variability present in two or more individual genomes across the 99 codons in the *pro* gene for these samples.

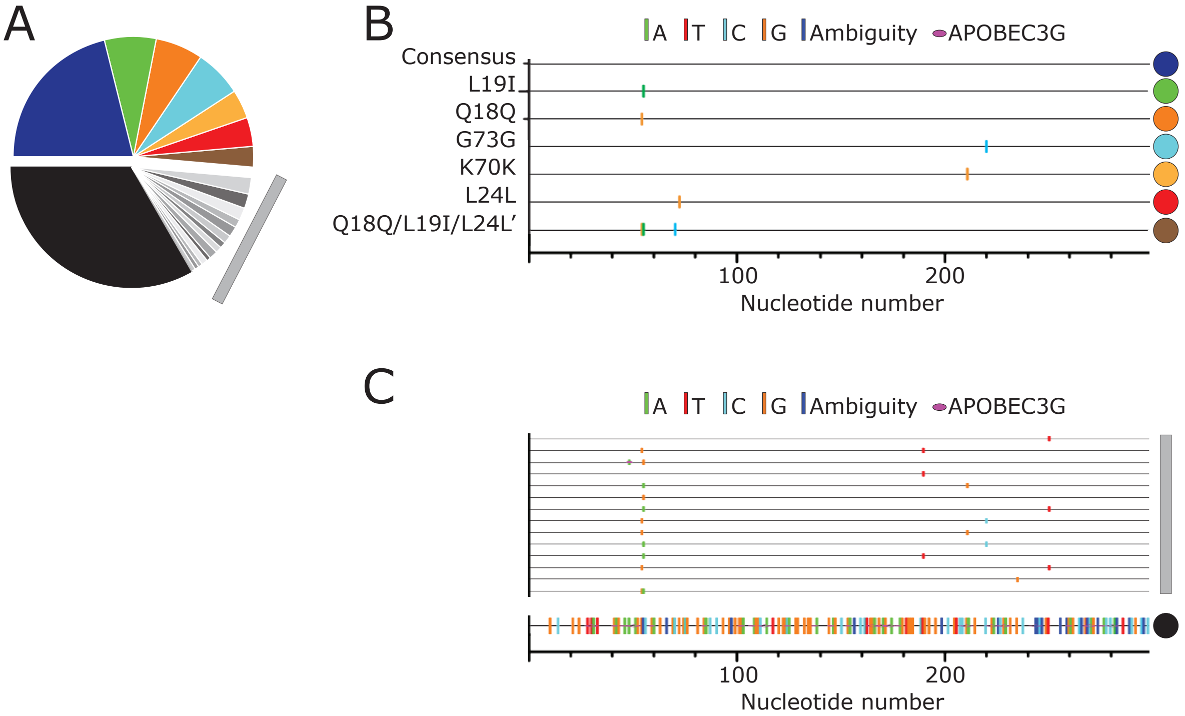


Fig. 2.5 Major and minor allelic variants in the untreated populations. (A) Frequency of major (colored) and minor (grayscale) unique pro gene sequences. Gray colors represent pro gene sequences present between 2.5 and 0.5% in frequency. Black represents the sum of all pro gene sequences individually present at <0.5%. (B) SNP distribution of the most abundant

pro gene sequences (>2.5%), with the colored dots on the right indicating the corresponding sequences identified in the pie chart (A). (C) The gray bar corresponds to SNP distribution of variants present between 2.5 and 0.5%, the same sequences indicated in panel A with the gray bar. The line at the bottom indicated by the black circle represents the sum of all variants <0.5% in frequency for the sequences shown in black in the pie chart (A).

Table 2.2 Summary of nucleotide variation in sampled time points

	T1	T2	T3	$T3_s$	$T3_r$
Number of sequences	810	1449	1925	594	972
Number of polymorphic (segregating) sites	130	145	144	95	80
Total Number of mutations	148	167	159	102	83
Average number nt differences, k	2.84	2.25	3.39	3.49	2.04
Nucleotide diversity, π	0.00955	0.00826	0.01141	0.01177	0.00687
Theta (per sequence)	20.33	21.30	19.61	14.67	11.20
Theta (per site)	0.068	0.072	0.066	0.049	0.038
Tajima's *D*	-2.46864	-2.47953	-2.28563	-2.19101	-2.26483

Synonymous Variability: There were 57 codons (with 63 variants/SNPs) that contained synonymous diversity that appeared in both pre-therapy time points, and 30 codons (with 31 variants) that appeared in only one time point. Taken together, 75 of the 99 codons contained some level of synonymous diversity (Fig. 2.2, Table 2.1). Of the 63 variants that were present in both untreated time points, 92% were transitions. Of the 31 variants that appeared in only one of the time points, 71% were transitions, representing a significantly smaller fraction of transitions than among the synonymous variants that appeared at both time points (p = 0.012;

Fisher's exact test). This suggests that synonymous transversions are selected against over time.

Nonsynonymous Variability: There were 26 codons (28 variants) that contained coding variability that appeared in both pre-therapy time points, and an additional 28 codons (33 variants) with nonsynonymous changes found in only one of the time points. Taken together, 49 of the 99 codons contained some level of nonsynonymous diversity (Fig. 2.2, Table 2.1). For the 28 nonsynonymous variants detected at both time points, 22 were transitions, and these mostly represented conservative amino acid changes. In the case of synonymous mutations two-thirds of the variants were present at both time points, while in the case of nonsynonymous mutations less than half were present at both time points (p = 0.012; Fisher's exact test). This observation suggests that at this level of sequence sampling we are able to see a difference in stability within the population in comparing synonymous and nonsynonymous substitutions.

Genetic Fluctuation: We compared the stability of minor SNPs present at both T1 and T2. A total of 14 of the 91 SNPs (synonymous and nonsynonymous that appeared at both time points) had significant changes in abundance between the two time points (Chi-square Test with a false discovery rate of 0.05). Of the 14 SNPs with significant changes in abundance, 11 had a decrease in the abundance, with an average decrease around 7.5 fold. There were 3 SNPs that had a significant increase in abundance, all of which were synonymous, ranging from a 4 to 47-fold increase. While a majority of SNPs that changed in abundance had a decrease in the frequency between T1 and T2, on a population level there was not a large change in diversity between the two time points (T1 p = 0.0079, T2 p = 0.0082 [Table 2.2]). However, the trend of increased abundance at the three sites may be driven by selection of cryptic epitopes in an alternative reading frame (see Discussion).

Significance of Rare Variants: We observed two extremes in terms of biological relevance in the untreated population among variants detected as at least two independent sequences across the three time points. At one extreme was the detection of nonviable genomes in the form of a coding variant at position 25, which mutates the active site of the protease, and the detection of a termination codon at position 61 (Table 2.1). At the other extreme was the detection of the L90M and V82A variants (at time points 1 and 2, respectively) that became the major resistance populations after ritonavir therapy was initiated (see below, Fig. 2.6); in addition, V82I and V82L were detected at T2. We found two more examples of primary resistance mutations at low abundance, K20R at all three time points and M46I at two time points, but these did not grow out in the presence of ritonavir (Fig. 2.6, Table 2.1). Similarly, fitness compensatory mutations were also detected at low abundance (L10F, M36I, L63P, A71T, and V77I), all below 1%, and only L63P increased (modestly) in abundance after exposure to ritonavir. More generally, of the 28 substitutions most closely associated with protease inhibitor drug resistance (120, 121), we found 10 such variants, half of which were detected at both pretherapy time points (Table 2.1).

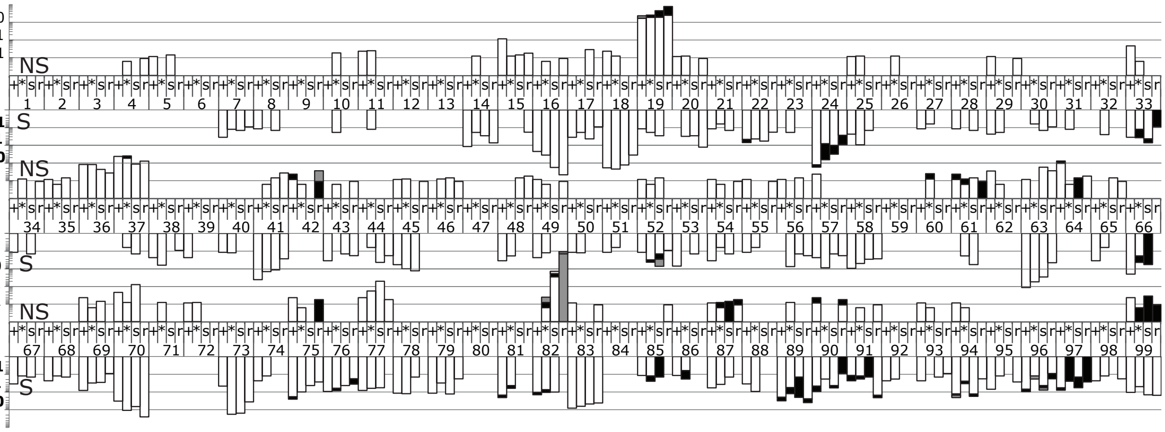


Fig. 2.6 Frequency of codon variation across all 99 positions in protease over three time points. Within a codon position, the first two bars represent untreated time points 1 and 2, respectively. Bars 3 and 4 are the third time point split based on the presence or absence of the resistance mutations to ritonavir. Bar 3 is the population of susceptible genotypes (defined as not V82A, I84V, or L90M), and bar 4 is the major resistant variant, V82A, population. Upward facing bars are nonsynonymous changes (scale in regular typeface), and downward facing bars are synonymous changes (scale in bolded typeface). Within a codon position, different shading represents different SNPs.

Assessment of linkage disequilibrium within the HIV-1 *pro* gene population

We measured LD for the sequences in the T1 and T2 populations. We identified very few examples of LD at these two time points using the Fisher's Exact Test with a Bonferroni correction. Of the 103 polymorphic sites in T1, only 3 pairs were in significant LD. Similarly, in T2 with 118 polymorphic sites, only 4 pairs displayed significant LD. A positive D (i.e. linkage) was found for 6 of the 7 pairs in the untreated populations, with one pair associating at a lower than expected frequency. Overall, LD did not appear to play a significant role in defining the *pro* gene population in this late stage individual, with only a single pair of SNPs showing linkage in both of the time points.

Detection of multiple drug resistant alleles after exposure to selection by a protease inhibitor.

The third plasma sample we examined from this subject was from a time point (T3) after the initiation of therapy with the protease inhibitor ritonavir. It is apparent from the cyclical pattern of viral load and self-report that this person had incomplete adherence to the drug regimen (Fig. 2.1). Thus we expected selective pressure from the drug to disrupt the viral population but not to select for the more homogeneous populations that are associated with virologic failure solely due to the appearance of drug resistance. The choice of this sample allowed us to look at the evolution of resistance and the persistence of polymorphisms in both the resistant and nonresistant portions of the population. Over two-thirds of the sequences from T3 carried a resistance mutation, with approximately 50% of the sequences carrying the V82A allele, the most common resistance mutation associated with resistance to ritonavir (122).

There were two divergent paths for population diversity at the third time point. For the large V82A-containing population there was a general trend of decreased diversity ($p = 0.0059$), consistent with the expected bottleneck associated with fixing a drug resistance mutation. In contrast, the diversity in the co-existing drug sensitive population was higher than the drug resistant population and comparable to the earlier time points ($p = 0.0088$) (Table 2.2).

While V82A is the most common resistance mutation associated with ritonavir resistance, the I84V allele and L90M allele can also be selected and in combination with V82A can confer a higher level of resistance (46). We detected all three of these distinct drug resistance alleles in the T3 sequence population, collectively representing 69% of the total T3 population: V82A (50% of the population), I84V (5%), and L90M (14%). These three resistance mutations appeared on different genomes, with only a single example of a sequence with two of these resistance mutations (V82A/L90M). In total there were 136 unique sequences carrying the V82A mutation (all with the GCC Ala codon), 29 unique sequences carrying the I84V mutation (all with the GTA Val codon), and 36 unique sequences carrying the L90M mutation.

There were also small groups of *pro* gene sequences in T3 that appear to be the result of selection by ritonavir. Two other substitutions at position 82, V82I and V82L, were detected at a low level at T2 and also seen at T3 but now representing 1.3% and 1.1% of the population. V82F was also detected as 0.14% of the population at T3. Finally, the compensatory mutation L63P was detected at T1 and modestly expanded at T3, with half of the sequences in the V82A background (Table 2.1).

An important issue is the number of times each of the resistance mutations evolved in the presence of drug selection. The data are consistent with the major V82A variant (42% of the V82 sequences) growing out from the pre-existing variant detected at T2. For the 6 genomic variants of V82A that each accounted for greater than 2.5% of the V82A population, all were on the background of the consensus except for the three different polymorphisms at positions 19 and 70 (Fig. 2.7*B*). In total, these represented approximately 71% of the V82A population and presumably arose via recombination with the founding sequence (Fig. 2.7*A*). The remaining 29% of the V82A-containing genomes vary in relative abundance from 2.3-0.1%, including over 100 unique sequences that each appeared once but to a large extent represent the variation seen at T1 and T2 added on to the predominant V82A genotypes.

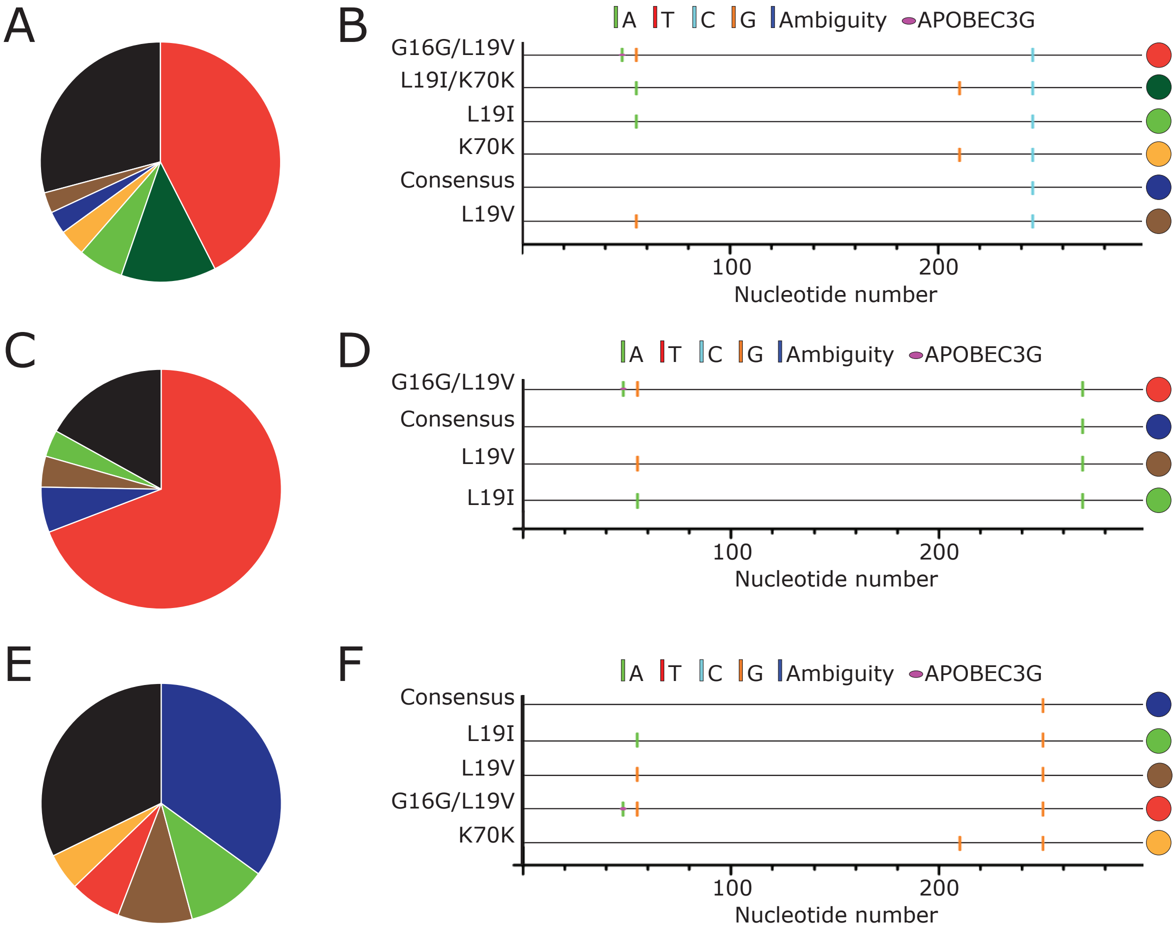


Fig. 2.7 Major and minor unique pro gene sequences in the major resistant populations V82A, L90M, and I84V. (A) Frequency of different unique pro gene sequences carrying the V82A mutation at high frequency (colored >2.5%) and low frequency (<2.5%, black and with the abundance pooled). (B) Highlighter plot showing the sequence changes from the consensus sequence for the major (>2.5%) pro gene variants carrying the V82A mutation.

The V82A substitution is indicated by the nucleotide change at position 245 shown in light blue. (C) Frequency of different unique pro gene sequences carrying the L90M mutation at high frequency (colored >2.5%) and low frequency (<2.5%, black and with the abundance pooled). (D) Highlighter plot showing the sequence changes from the consensus sequence for the major (>2.5%) pro gene variants carrying the L90M mutation. The L90M substitution is indicated by the nucleotide change at position 268 shown in green. (E) Frequency of different unique pro gene sequences carrying the I84V mutation at high frequency (colored >2.5%) and low frequency (<2.5%, black and with the abundance pooled). (F) Highlighter plot showing the sequence changes from the consensus sequence for the major (>2.5%) pro gene variants carrying the I84V mutation. The I84V substitution is indicated by the nucleotide change at position 250 shown in orange.

The composition of the I84V and L90M populations were similar to the V82A population. In each case there was a predominant population defined by a 5' polymorphism: the major L90M lineage (69% of the L90M sequences) was on the G16G/L19V background (Fig. 2.7*C*, Fig. 2.7*D*) while the major I84V lineage (35% of the I84V sequences) was on the consensus sequence background for the 5' polymorphisms (G16/L19) (Fig. 2.7*E*, Fig. 2.7*F*). The next three most abundant I84V lineages, representing 28% of the I84V sequences, differed from the most abundant sequence by other 5' polymorphisms (Fig. 2.7*F*). Similarly, the next three most abundant L90M lineages, representing 14% of the L90M sequences, differed from the most abundant L90M sequence by 5' polymorphisms (Fig. 2.7*D*). With the exception of the 5' polymorphisms and the resistance mutations, all eight of these lineages were in the consensus sequence background. The remaining sequences are accounted for by the low level variability added onto these major lineages.

As noted above, the major V82A lineage was detected at T2 (as a single genome), and this population was likely clonally amplified to form the large proportion of the drug resistant population seen at T3 (Fig. 2.7). L90M was also detected on the same *pro* gene background in the therapy-naïve environment at T1, and was likely also clonally amplified to form the large proportion of the L90M sequences (Fig. 2.8, 2.7*D*). In contrast, V82I and V82L were detected in the pre-therapy time points on background sequences that did not become the predominant sequence when these mutation modestly expanded at T3, although these two populations have complex mixtures of the 5’ polymorphisms which may indicate low level persistence and recombination during the period of drug exposure. Finally, I84V and V82F were not detected in either pre-therapy population (Table 2.1).

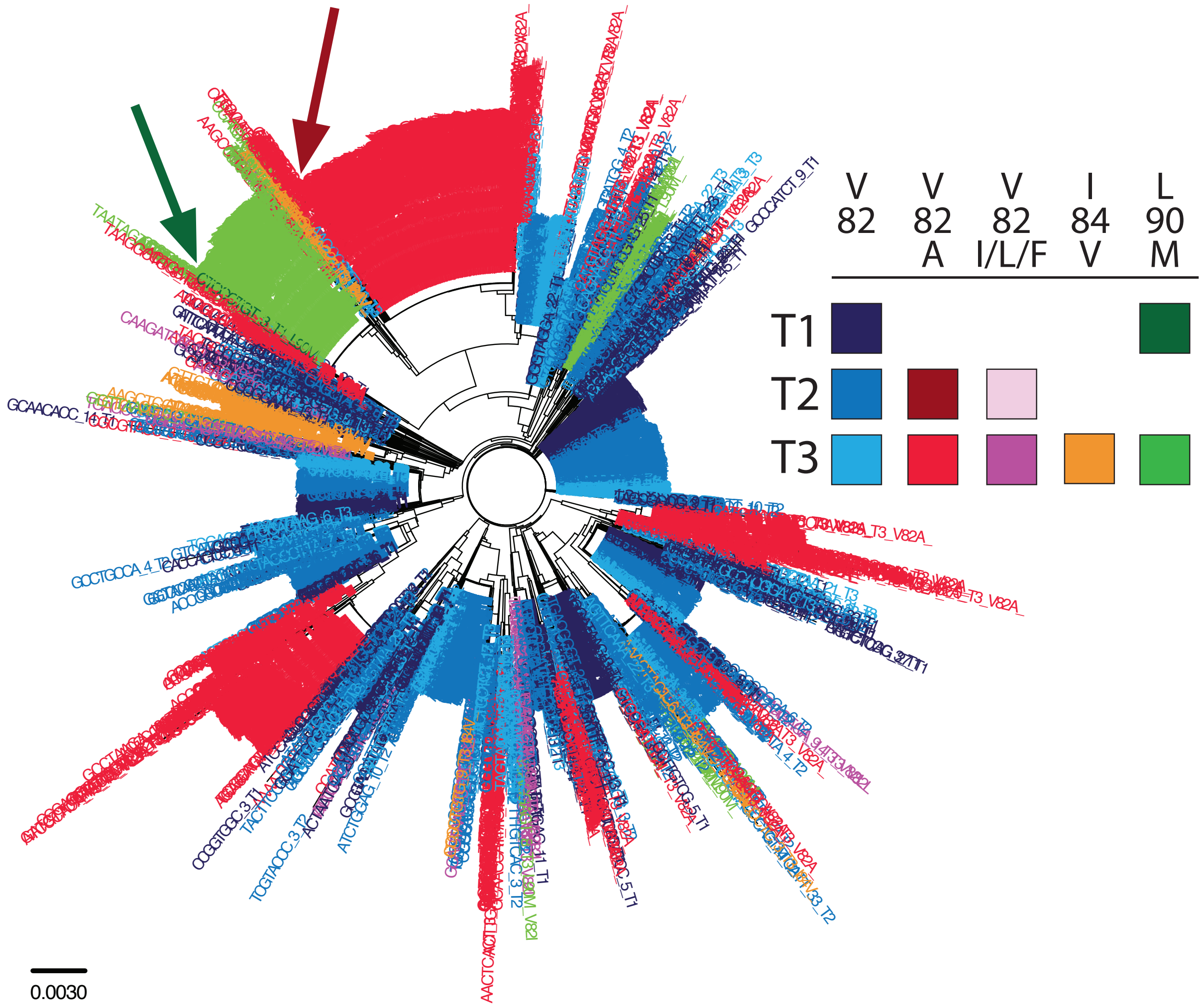


Fig. 2.8 Phylogenetic representation of protease population derived from deep sequencing with a Primer ID. A Neighbor-Joining tree was constructed from sequences derived from all three time points and colored based on susceptibility to ritonavir. Blue colored taxa represent susceptible variants (defined as not V82A/I/L/F, I84V, or L90M). Red colored taxa represent variants containing the major ritonavir resistant variant, V82A. Pink colored taxa represent the minor resistant variants V82I/L/F. Green and orange colored taxa represent the minor resistant alleles L90M and I84V, respectively. Within a color, color brightness is correlated with sample time. Dark green and red arrows point to pre-RTV low-abundance sequences that clonally amplified to their respective clades.

2.5 Discussion

Complex viral populations can form within a host (123-125). High throughput sequencing technologies allow for extensive sampling of these populations (81, 95-97, 99, 126). However, these technologies are severely limited when a PCR amplification precedes the sequencing protocol, as each sequence read has the potential to be reported as an independent observation without properly controlling for PCR resampling, PCR-mediated recombination, allelic skewing, PCR-introduced misincorporations, and sequencing errors. When working with pathogenic agents in clinical samples, the number of pathogen genomes in the sample is limited, and the use of PCR can obscure the quality of the sampling by creating a large amount of DNA from a relatively small number of starting templates. This can create artificial homogeneity, inflate estimates of segregating genetic variation, skew the distribution of alleles in the population, and introduce artificial diversity.

We have developed a novel strategy that allows each sampled template to be tagged with a unique ID by a primer that has a degenerate sequence tag incorporated during the primer oligonucleotide synthesis. This tag can then be followed through the PCR and the deep sequencing protocol to identify sequencing over-coverage (resampling) of the individual viral templates. Because the Primer ID allows for the identification of over-coverage, this can then be used to create a consensus sequence for each template, avoiding both PCR-related errors and sequencing errors. In addition, the number of different Primer IDs reflects the number of templates that were actually sampled. This allows a realistic assessment of the depth of population sampling and makes it possible to apply a more rigorous analysis of minor variants by correcting the allelic skewing during the PCR.

We tested the Primer ID approach by sequencing the HIV-1 protease coding domain at three time points in a subject who was intermittently exposed to a protease inhibitor between the second and third time points. A key feature of our approach is the removal of fortuitous errors and accounting for resampling which results in a dramatic reshaping of the original data set of 72,162 reads. There are other approaches that rely on statistical modeling that have been developed to deal with the problem of high sequencing error rates associated with deep sequencing technologies (91, 92, 127). The use of the Primer ID to create consensus sequences resulted in the removal of 80% of the unique sequence polymorphisms (defined as a change in the consensus without regard to frequency of appearance) in the data set. Similarly, allelic skewing was dramatic among the sampled sequences, in most cases ranging from 2-15 fold but going up to nearly 100 fold. While the Primer ID reveals such skewing and helps correct it, this is clearly a poorly controlled feature of PCR amplifications that can dramatically affect the observed abundance of complex populations, especially the minor variants. Allelic skewing may still persist if the cDNA primer or the upstream PCR primer binds differentially among the templates, or if cDNAs enter the PCR amplification in later rounds and are discarded because they do not result in at least three reads to allow a consensus sequence to be formed. Also, residual misincorporation errors by RT and in the first round of PCR synthesis still limit the interpretation of mutations that occur in the range of 0.01 to 0.1%. This problem is not overcome with larger numbers of sequences. Given the low diversity in these samples we removed all substitutions that appeared once since their number approximated the expected number of residual sequence errors, and this resulted in a sensitivity of detection in the range of 0.1% for SNPs that appeared above the frequency of the residual sequence error rate.

Using the Primer ID approach we were able to describe a number of features of the protease sequence population, however our results are from a single individual and therefore cannot be generalized. First, a pooled analysis of two time points six months apart showed that the variants present at greater than 0.5% in abundance made up two-thirds of the total population but represented only 4% of unique genome sequences and contained only 7% of the total unique sequence polymorphisms. About 60% of the diversity was stable over both time points, with synonymous SNPs maintained at a significantly higher proportion in the two time points than nonsynonymous SNPs. Only 18% of the total diversity represented nonsynonymous SNPs that were present at both time points. However, our ability to assess persistence of these sequences is limited by the depth of sampling, although we feel we are approaching the practical limit of sampling with this technology as we observed nonviable substitutions and estimate that most of the SNPs that appeared once were the result of remaining method error. We found no pattern of conserved linkage among these SNPs, consistent with high levels of recombination across the population.

While the overall measurement of diversity (p) was similar between the first two time points we noted that the biggest changes in SNP abundance between the two time points were in three synonymous codon positions (L24L, K70K, and G73G). These dynamic increases made these SNPs part of a larger group of SNPs that accounted for 51% of the total sequences that were otherwise identical to the consensus sequence (Q18Q, L19I, L24L, K70K, G73G, and Q18Q/L19I/L24L'). These SNPs also overlapped the major SNPs that defined subgroups of the resistant variants (L19I; L19V; G16G/L19V). We considered the possibility that there was a unifying feature of these SNPs. We found such a feature in that all of these SNPs, both coding and noncoding, result in changes in two relatively large alternative open reading frames that lie at the 5’ and 3’ ends of the *pro* gene. Alternative reading frames have been suggested to generate cryptic CTL epitopes (128-130). In this scenario, these abundant SNPs would represent various escape mutants. Such selective pressures could explain the dynamic behavior of several of these SNPs between the first two time points.

After intermittent exposure to the protease inhibitor ritonavir, we were able to identify six independent lineages of drug resistance mutations. With the intermittent exposure in this particular subject it was possible to see the major V82A lineage most often seen with ritonavir resistance, but also significant populations of I84V and L90M. We also saw minor populations of V82I, V82L, and V82F. This mixed population of resistant lineages likely represents the early stages of the evolution of resistance, a conclusion supported by the minor appearance of the L63P compensatory mutation and the complete absence of I54V, which is an often seen compensatory mutation for V82A. We saw few examples of genomes with multiple resistance mutations, although these would be expected after more extensive selection (48, 131). We and others have previously examined viral sequences that have been collected in large databases. Typically these sequences represent the single predominant sequence within an individual, and the use of these sequences allows for assessment of inter-person diversity. In the future it will be an interesting exercise to compare the conclusions reached by examining viral diversity within a person compared to viral diversity between people; however more intra-person diversity needs to be measured at this level of detail to allow comparison of inter- versus intra-person diversity.

The presence of pre-existing drug-resistant variants and their role in therapy failure is of great interest, and accurate, deep sampling of a viral population can add significantly to our understanding of this question. We were able to detect several examples of drug resistance mutations but only at a very low level. Our ability to reliably detect these mutations is limited to those that appear at a frequency of 0.1-0.2%, limited in part by the low overall diversity in the population. We were able to see examples of mutations that are typically seen only in the presence of drug selection. However, the detection was usually as one genome at two time points or two genomes at one time point. This was also the level of detection of active site mutations in the protease and of termination codons, which must represent either transient viral genomes or residual misincorporation errors. In two cases we were able to observe the resistance mutation (V82A and L90M) at pre-therapy time points linked to the same polymorphisms that were present on the variant that grew out during drug exposure. Thus while it is likely that we are detecting relevant pre-existing drug resistant variants, these are at the limit of detection and if they are maintained at a steady state level it is well under 0.5% abundance.

Most protocols of high throughput sequencing technologies still require an initial quantity of DNA that necessitates an upfront PCR step for many applications. The use of a Primer ID will help clarify the sequencing products in any strategy that uses an initial PCR step with its attendant error rate, recombination, and resampling. We believe a strategy that allows an initial tagging of individual templates prior to PCR and subsequent sequence analysis will be essential for understanding the true complexity and diversity of genetically dynamic populations.

Chapter 3

Deep sequencing with a Primer ID reveals the dynamic paths of HIV-1 resistance during drug failure

Cassandra B. Jabara[a,b,c], Ronald Swanstrom[b,c,d], and Corbin D. Jones[a,e]

[a]Department of Biology, [b]Lineberger Comprehensive Cancer Center, [c]UNC Center for AIDS Research, [d]Department of Biochemistry and Biophysics, [e]Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, US

3.1 Abstract

The intrahost population dynamics of viruses such as Human Immunodeficiency Virus Type 1 (HIV-1) is poorly understood. A population of HIV-1 may be large, heterogeneous, and spatially and temporarily variant, but it is unclear if HIV-1 has a large or small effective population size. The effective population size (N_e) can affect the path to antiviral drug resistance. If N_e is large, pre-existing resistance alleles may be present, and the evolutionary forces acting on them deterministic. If N_e is small, pre-existing alleles may not be present and the evolution of resistance could be strongly influenced by stochastic events. To characterize the intrahost population dynamics of HIV and to finely resolve viral

diversity, we used deep sequencing technology with our previously developed method, Primer IDs. We longitudinally sampled 10 chronically infected individuals who failed drug therapy. Although viral load indicates a large census population, traditional calculations to estimate the effective population size gave values that ranged from 620-1,129. However, we noted an excess of low-frequency polymorphisms, which suggests that the population is not in mutation-drift equilibrium, downwardly biasing estimates of N_e. The nucleotide diversity in the therapy naïve populations differed among subjects, and this directly correlated with the percentage of major haplotypes within a population but not sampling depth. Therefore, there may be significant differences in the selection pressures on the viral populations among subjects. We also measured population variation after exposure to a single drug (the protease inhibitor ritonavir). The diversity of the emergent, resistant population was greater than the pre-existing diversity, although N_e and the number of polymorphic sites tended to be reduced, supporting that resistance alleles on multiple haplotypes were segregating at intermediate frequencies within the populations after drug selection. While we did observe pre-existing resistance alleles in the pre-drug therapy populations, we did not observe these alleles on the background haplotypes that grew out during treatment, suggesting that the emerging resistant haplotype was not sampled prior to treatments or that the resistance mutants arose during the selection. However, the fact that multiple haplotypes grew out with resistance mutations, and these mutations were not on the dominant haplotypes in the population, indicates that the resistance mutations that did grow out pre-existed in the population. The path to higher-level resistance within each individual was unique, but involved common major resistance alleles. Selection coefficients for the resistance haplotype were at minimum 0.01-0.04. In some populations after exposure to drug, a susceptible subpopulation persisted despite the emergence and evolution of increasingly resistant haplotypes, suggesting intermediate drug exposure in these subjects. One subject with low levels of drug in plasma had a resistance mutation grow out that confers a lower level of resistance but with less of a fitness cost as would be predicted from deterministic evolution. In sum, the population dynamics of HIV-1 are clearly complex, maintaining a small number of abundant haplotypes and a large number of minor haplotypes. The consistency with which resistance evolved suggests that N_e is much larger than calculated, with multiple resistance alleles appearing under selective pressure and with the mutations that appear determined by the level of selective pressure. These results are consistent with deterministic evolution occurring within a large effective population.

3.2 Introduction

Human Immunodeficiency Virus Type 1 (HIV-1) can create a large, diverse population within a host, but the evolutionary forces that shape the population have been highly debated. The effective population size (N_e) is the idealized number of virions required to create a population experiencing the same amount of genetic drift as the census population (132). The N_e of HIV-1 has been used to derive the frequency of pre-existing drug resistance mutations (133-137) and how recombination affects the evolution of resistance (138). Estimates of N_e and population structure ultimately determine whether deterministic or stochastic forces drive the evolution of a viral population.

The census population of HIV-1 is large and presumably diverse. The time needed for an HIV-1 virion to attach, replicate, and produce progeny that infect new cells is estimated at

2 days, with a maximum of approximately 10^{10} virions produced per day within a person (31). The viral DNA polymerase, Reverse Transcriptase (RT), is error-prone, introducing mutations at an estimated rate of 2.16×10^{-5} substitutions/site/generation during transcription of viral RNA (vRNA) to double-stranded DNA (dsDNA) (9, 10). Due to a high mutation rate, small genome size, and large census population, it has been hypothesized that every single mutation can exist (36, 37). If the effective population size is greater than the inverse of the mutation rate (~50k virions), the fate of alleles in this population is expected to be dominated by Darwinian forces such as purifying selection. However, much smaller estimates for the N_e of HIV-1 have been argued, shifting evolutionary forces to the stochastic effects of genetic drift (56, 57, 139, 140).

Most analyses of N_e assume Wright-Fisher mutation-drift equilibrium. Populations of HIV-1, however, may not be at this equilibrium. Natural selection, population structure, changes in population size, and and/or unequal reproductive ratios among progeny virions can distort estimates of N_e. For HIV-1, assumptions of neutrality have yielded very small estimates of N_e (56, 57, 139-142), whereas adding selection has produced very large estimates of N_e (54). However, the majority of studies assuming neutrality estimated genetic diversity from *env* gene sequences, a nonsynonymous biased immune target. Furthermore, the sequences were derived using methods that only sample the most common alleles, likely leading to underestimates of census population diversity (56, 57, 139-142).

Individuals can fail antiviral drug therapy. The role and clinical significance of pre-existing resistance alleles in therapy failure is not fully understood. N_e has been used to estimate the frequency of pre-existing resistance alleles and the probability that they will contribute to resistance. However, as N_e estimates greatly vary and may be biased by sampling, this measure is unreliable. Alternatively, studies have tried biological tools to correlate the presence of pre-existing resistance alleles and virologic rebound. Allele-specific PCR approaches suggest that pre-existing variants may prelude increased susceptibility to therapy failure (40, 41), but resolution is limited by *a priori* screening of unlinked alleles. Deep sequencing studies have produced correlative (42), partial (43), and non-correlative (44) results on the impact of pre-existing resistance alleles, but have been limited by sampling bias and error.

Because of previous limitations in estimating N_e and characterizing viral diversity, we sought to determine the population dynamics of HIV-1 sampled longitudinally and as the population went through a selective event. We used deep sequencing paired with a novel high-resolution sequencing technique we previously developed, Primer IDs (143), to capture and identify individual viral genomes. For the 10 subjects sampled, both susceptible and resistant populations had small N_e but the viral population structure violated the assumptions of Wright-Fisher equilibrium. All populations had the majority of genetic diversity contributed by many low frequency haplotypes indicating that a high number of alleles are at the mutation/selection balance, diversity is quickly recovered after drug selection, recombination distributed the resistance allele among the low abundant diverse portion of the population, and/or that a large number of resistance alleles on multiple haplotypes grew out. We did not observe pre-existing resistant haplotypes that were directly selected under the drug, although this could be due to an inadequate level of sampling. The consistency with which resistance evolved on diverse background haplotypes suggests that N_e is much larger than estimated by traditional calculations.

3.3 Materials and methods

Tagging, amplification, and sequencing of vRNA. Longitudinal samples from 10 chronically infected individuals (infected with HIV-1 subtype B) who were enrolled in the protease inhibitor efficacy trial M94-247, were chosen for analysis based on the presence of high viral load and virologic failure after the addition of the protease inhibitor ritonavir as monotherapy (Fig. 3.1). Between 1 and 4 samples were examined per individual over the study time course. Approximately 10,000 vRNA templates went into a cDNA synthesis reaction using a primer that annealed downstream of the protease coding domain, with the primer containing at its 5' end a barcode, a degenerate Primer ID cassette, and a PCR primer binding site, as previously described (143). Tagged cDNA molecules were amplified by nested PCR. Amplicons were quantified using the Qubit platform (Invitrogen, Carlsbad, CA), pooled in equal molar amounts, then 454 adaptors were added using the Lib-L Rapid Library protocol (Roche, Nutley, NJ). The libraries were sequenced on the 454 GS FLX+ platform with XLR70 Titanium sequencing chemistry as per the manufacturer's instructions (Roche, Nutley, NJ) but with under-loaded beads to minimize signal crosstalk.

Bioinformatic pipeline and analysis of viral populations. A suite of programs was written to filter and parse raw 454 sequencing reads. In short, raw sequencing reads were first binned by sample based on the barcode and then binned for each cDNA molecule by the Primer ID. Within a sample, when 3 or more reads were present with an identical Primer ID, a consensus sequence was constructed. The Primer ID technique and methodology is further detailed in Jabara et al. (143).

All statistical and diversity analyses were done on filtered populations of the consensus sequences and devoid of sequences containing ambiguities; gaps were resolved by alignment to the consensus sequence of the protease coding domain. Tajima's *D* test and sliding window analysis of π were computed by DnaSP v.5.10.01 (111). Sliding window analysis of π had a window length of 100 and a step size of 10 bases. Whole gene π, diversity, and haplotypes across and within populations were computed through customized bioinformatics suites. SNPs were graphically displayed through the *Highlighter* tool (www.hiv.lanl.gov).

Database population analysis. Viral protease sequences from therapy-naïve subjects infected with subtype B HIV-1 were downloaded from the Stanford HIV-1 database and aligned to the same reading frame. Polymorphisms that appeared once in the data set and mixtures were excluded. All full-length sequences were first screened for the absence of Class III alleles (primary drug resistance alleles), which would indicate prior drug exposure. The filtered population was then parsed based on the presence or absence of non-consensus polymorphisms at Class II positions (compensatory mutations). Sequence alignments, filtering, and diversity analysis was completed using customized bioinformatics suites.

3.4 Results

Protease nucleotide diversity in therapy naïve populations is different between people and related to population structure.

To resolve the genetic diversity and response to drug selection of *in vivo* HIV-1 populations, 10 chronically infected subjects who failed drug therapy were chosen for retrospective longitudinal sampling and deep sequencing. For each subject, viral RNA was isolated from serum samples representing between one and four time points taken pre- and/or post-drug exposure. For each sample, approximately 10,000 copies of viral RNA were used in the cDNA synthesis reaction, with the cDNA primer tagged with a barcode and Primer ID as previously described (143). Tagged cDNA was amplified by nested PCR and sequenced using the 454 FLX+ platform with Titanium chemistry.

The Primer ID component of the cDNA primer was used to correct for PCR biases, sequencing error, and PCR re-sampling, and this allowed us to estimate a median sampling depth of 0.2%. A depth ≤0.1% was reached for at least two time points for nine of the subjects (Table 3.1, Fig. 3.1). In order to characterize the selective pressure on the viral population, drug concentrations of ritonavir (RTV) were measured for time points after the initiation of therapy. We found that drug levels varied greatly between patients and time points. This is potentially due to differences in metabolism and, in some cases, likely intermediate adherence. For either reason this would result in suboptimal drug exposure and intermediate or cyclical selective pressure.

Table 3.1. Summary of clinical metrics and sequencing resolution.

PID[a]	Sample[b]	VL[c]	RTV[d]	RTV (ng/mL)[e]	Reads[f]	P-IDs[g]	Consensus[h]
1008	T1	13,060	+	N/A	7,553	2016	299
1032	T1	15,380	+	N/A	15,091	782	338
	T2	68,360	+	15,117	11,049	541	233
	T3	216,640	+	0	754	291	60
	T4	273,960	+	1,864	13,495	744	208
1036	T1	236,040	-	N/A	52,912	1763	954
	T2	36,840	+	5,669	47,754	1945	949
	T3	45,780	+	5,611	8,970	1202	857
	T4	99,000	+	4,667	11,057	643	349
1047	T1	279,000	-	N/A	2,984	831	391
	T2	198,320	+	0	2,520	820	219
	T3	33,040	+	66	5,594	1077	584
	T4	202,240	+	770	3,174	1028	298
1051	T1	501,600	-	N/A	62,614	3114	1765
	T2	233,020	+	983	43,037	2130	770
	T3	351,880	+	6,566	12,163	1082	666
	T4	330,480	+	5,339	5,144	1266	673
1079	T1	199,520	-	N/A	16,386	827	161
	T2	63,920	+	5,723	967	192	44
	T4	242,360	+	6,610	17,531	4183	601
1113	T1	276,100	-	N/A	15,492	1203	548
	T2	133,240	+	1,773	15,020	1154	631
	T3	227,080	+	742	15,490	1719	837
	T4	343,720	+	5,540	23,259	2200	1112
1118	T1	205,600	-	N/A	7,207	981	136
	T2	57,360	+	19,830	27,605	1180	696
1127	T1	139,080	-	N/A	19,820	2515	1387
	T2	38,720	+	1,639	11,136	711	353
	T3	55,640	+	1,160	8,211	1174	471
	T4	181,160	+	767	5,477	624	238
1157	T1	380,200	-	N/A	10,372	1518	816
	T2	349,200	+	3,988	8,366	1028	126
	T4	243,440	+	1,056	1,732	747	77

[a] Patient identification.

[b] Time point of sample.

[c] Viral load (copies/mL).

[d] Therapy naïve (-) or ritonavir experienced (+).

[e] RTV concentration in blood plasma (ng/mL).

[f] Number of pre-consensus reads containing full length protease, Primer ID, and barcoding information.

[g]Number of individual Primer IDs in population of pre-consensus reads.

[h]Number of consensus sequences constructed within a sample when 3 or more reads contained identical Primer IDs.

N/A = not applicable

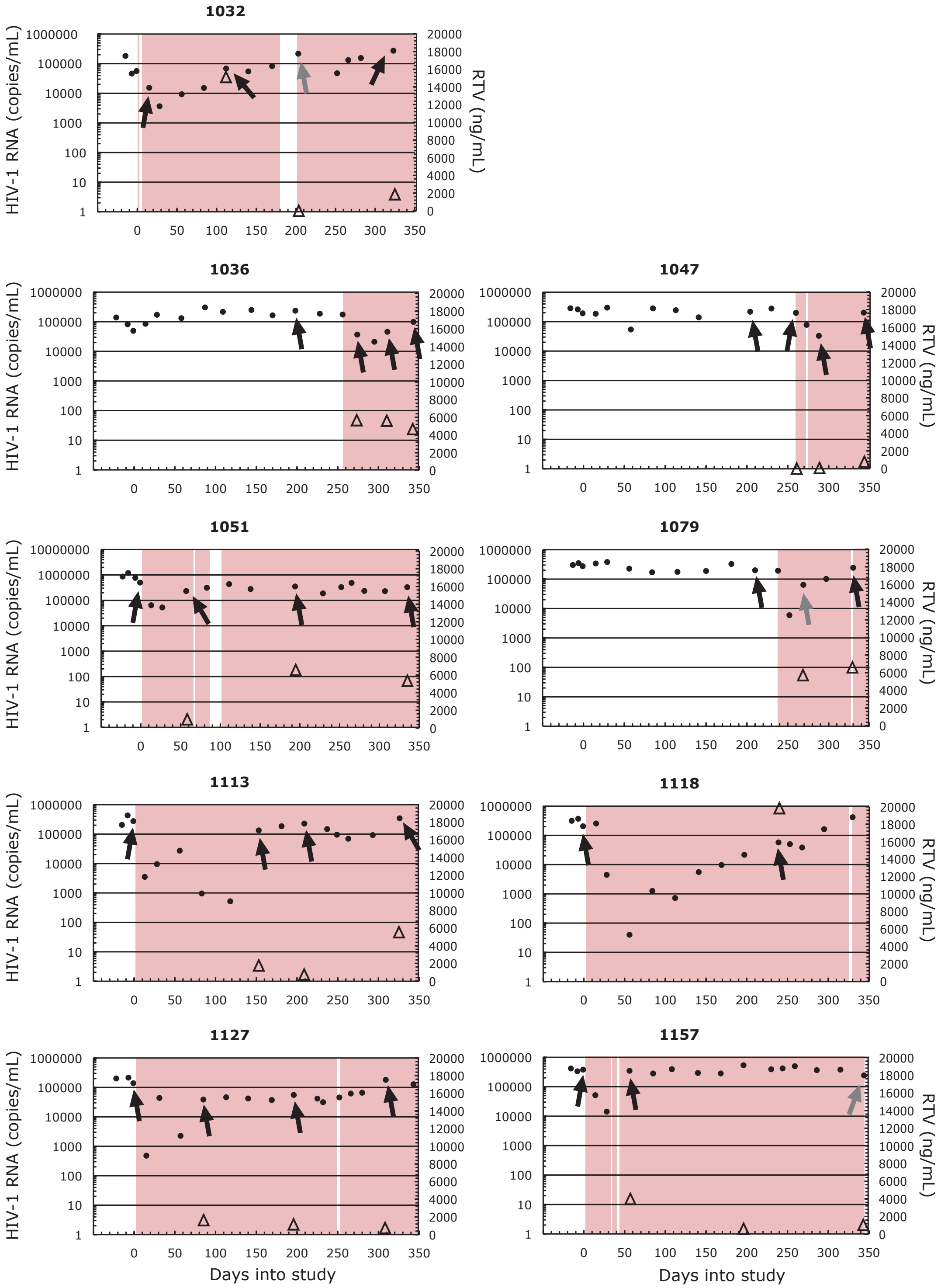


Fig. 3.1 Longitudinal sampling of blood plasma from 10 chronically infected individuals with HIV-1 subtype B that failed ritonavir monotherapy. Shaded areas represent times of RTV therapy compliance based on self-report. Black circles are viral load (copies/mL), and black open triangles are RTV drug concentrations (ng/mL). Arrows indicate time point sequenced, and shade correlates to the resolved population depth (black: ≤1%, gray: >1%).

We assessed the standing genetic variation prior to the drug selection pressure with a sliding window analysis of nucleotide diversity (π) across protease. Figure 3.2 shows reduced diversity across the active site for all individuals, as expected for a region where catalytic activity occurs. We found the mean intra-patient diversity was different between individuals and correlated with the frequency of minor haplotypes in the population. For example, the most diverse populations, found in subjects 1036 and 1079, had minor (≤1%) haplotypes making up 80% of the diversity of their populations. Conversely, haplotypes ≤1% for subject 1127 constituted only 34% of the population, resulting in lower π across protease (Fig. 3.2). This analysis points to a key feature of the HIV-1 population: a small number of abundant haplotypes, and a large number of minor haplotypes, with the proportion of these two types of sequences varying between subjects. We do not know what causes the inter-subject variation; pre-therapy diversity was not due to differences in sampling depth (Table 1), and all individuals had CD4+ counts ranging from 5-44 cells/mm^3 with no correlation between CD4 count and differences in population structure.

Subjects 1036, 1047, 1051, 1079, 1113, and 1157 were therapy naïve for all antiviral drugs. Subjects 1118 and 1127 were taking an additional antiretroviral drug, the nucleoside reverse transcriptase inhibitor (NRTI) ddC. Selection by ddC would target reverse transcriptase, thus proximal to the protease coding domain. Subject 1127 has the least amount of viral genetic diversity, possibly correlating to NRTI selection. However, π for subject 1118 is intermediate in comparison to the other subjects. It will be important to determine if an NRTI resistance mutation became fixed in either of these subjects. However, it seems unlikely that selective pressure by this weak NRTI would be responsible for shaping the population of the proximal protease coding domain. Alternative explanations for differences in π could be intrapopulation competition, or time since the most recent population bottleneck.

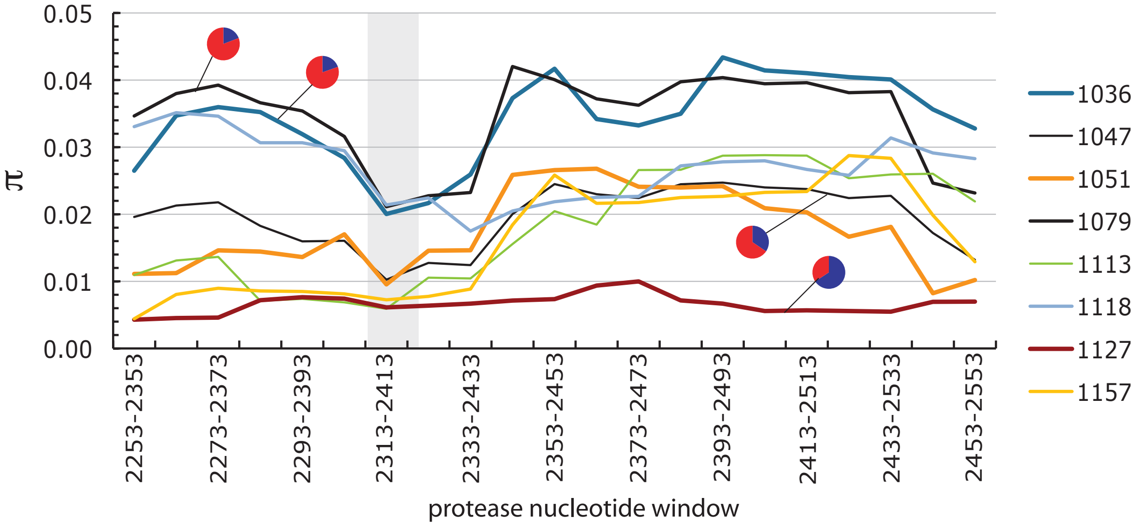


Fig. 3.2 Sliding window illustration of the nucleotide diversity (π) of protease in pre-therapy populations indicate spatial heterogeneity. Nucleotide diversity varies among subjects and the catalytic region typically harbors the least variation. Pie charts depict the percentage of population diversity that is made up of haplotypes ≥1.5% (blue) or <1.5% (red) in frequency. Gray shaded area corresponds to the active site of protease.

The majority of Class III drug resistance mutations pre-existed at frequencies above the expected PCR/sequencing error rate but below the minimum frequency needed being confidently sampled, likely confounding estimates of their frequency.

Primer ID enables in-depth resolution of population diversity, including minor variants that confer drug resistance. However, residual technical errors remaining with this method can confound whether a polymorphism is biological or artifactual based on the depth of sequencing. The frequency of error introduced during RT and the first round of PCR is expected to be 1 SNP per 33 protease genomes (1:10,000), and random misincorporation will produce a ratio of nonsynonymous to synonymous of approximately 2:1. For the pre-therapy time points, we observed approximately four-fold more single SNPs than expected given our estimate of the residual error rate of the method. Overall, the SNPs that appeared once have a mean ratio of nonsynonymous to synonymous of between 1.2 and 2. The excess of single SNPs over the expected error, and with a modest bias toward synonymous mutations, shows that our observed single SNPs are above the error rate but are not convincingly beyond an unselected distribution of nonsynonymous to synonymous mutations to be devoid of technical error.

Sampling depth will also dictate the probability a particular allele is sampled. The 8 pre-therapy viral populations were sequenced to a depth that would identify alleles corresponding to 0.13% of the population and ranging from 0.6-0.06%. Based on the Poisson distribution, the minimum frequency a single allele had to be present in order to be definitively sampled (3/N) ranged from 0.17-1.86%. Furthermore, when a low frequency allele is sampled twice—i.e. by two unique Primer IDs—there is a greater chance that the allele is biological and not technical error (the chance a randomly introduced error will mutate the same site with the same nucleotide is extremely low). Typically, a polymorphism needed to be at a frequency between 0.34-3.72% to be sampled twice (6/N).

The observed major RTV resistance variants V82A, I84V, and L90M, were not above the frequency threshold where two alleles on identical haplotype background were sampled, although a subset of them were found on more than one haplotype (Table 3.2). V82A was found in three of eight individuals at a frequency ~0.2%. I84V was found in half of the individuals and at a maximum on 0.74%. However, all of these variants were below the expected frequency at which an allele could be present and sampled with certainty. Thus, the failure to observe these alleles in other pre-treatment populations is not evidence for absence of these alleles in these populations. Instead, there is a possibility that due to sampling depth alone these and many alleles are missed.

Table 3.2. Frequency of Class II or Class III protease inhibitor resistance associated variants in therapy naïve populations.

				Patient ID[d]							
Con[a]	Class[b]	Mut[c]	Codon[d]	1036	1047	1051	1079	1113	1118	1127	1157
L10	II	F	TTC	*0.10*	61.64	*0.06*		0.55			
		I	ATC	2.41	9.97	2.78	37.89	*0.18*	1.47	99.64	12.87
		R									
		V	GTC				1.24			*0.07*	*0.12*
		C									
K20	III	M	ATG	4.93	4.35	26.46	1.24		*0.74*	*0.07*	
		R	AGG	0.52	*0.26*	1.30	*0.62*			0.36	
		I									
		T									
		V									
D30	III	N									
V32	III	I	ATA				1.24				
M36	II	I	ATA	19.60	12.28		17.39	0.73	2.94	0.94	3.92
		L	TTA								*0.12*
		V	GTG		*0.26*		1.24				0.25
		V	GTA			*0.06*					
M46	III	I	ATA	0.63		0.85	*0.62*	0.73	*0.74*	0.65	
		I	ATT							*0.07*	
		L	TTG							0.50	
G48	III	V	GTG	*0.10*							
I54	III	V	GTC	*0.10*		*0.06*					
		L									
		M									
		T	ACC				*0.62*		*0.74*	0.14	
		A									
		S									
I62	II	V	GTA	20.44	6.91		24.22	0.36	7.35	0.58	
A71	II	V	GTT		*0.26*	*0.06*			*0.74*	*0.07*	5.27
		I									
		T	ACT			*0.06*	*0.62*	0.36		0.14	
		L									
V77	II	I	ATA	2.10	4.09	0.45	14.29			0.22	2.33
V82	III	A	GCC	0.21		0.11				*0.07*	
		F									
		T									
		I	ATC			0.23			*0.74*		

		S									
I84	III	V	GTA	0.21			*0.62*		*0.74*		*0.12*
N88	III	S	AGT	*0.10*	0.51	0.11				*0.07*	
		D	GAT	*0.10*	*0.26*	*0.06*	*0.62*			*0.07*	
L90	III	M	ATG				1.86				
I93	II	L	CTT	*0.10*	0.51		1.86	1.28	44.85	*0.07*	
		M									

Italicized digits represent single observations.
[a]Consensus amino acid of resistance associated variant.
[b]Drug resistance mutation class[2]**.** Class II residues are defined as having ≥5% non-consensus variability across untreated individuals in the database and increase in non-consensus diversity with PI exposure. Class III residues are defined as having ≤5% non-consensus variability across untreated individuals in the database and increase in non-consensus diversity with PI exposure.
[c]Nonsynonymous change of consensus associated with resistance.
[c]Codon call if variant is observed.
[d]Patient identification (PID).

Class III resistance mutations are typically at low frequencies in pre-therapy environment and then increase in frequency with protease inhibitor selection (144). For each pre-therapy population, at least one Class III resistance mutation was detected (Table 3.2). Residues 32, 46, 82, 84, and 90 were found on 52/6158 haplotypes (0.08%). There were 45 total observations of a Class III allele, 71% of which were observed below the frequency a single allele would be reliably sampled. Only 18% of codons harbored resistance alleles at or above the frequency at which least two alleles would likely be sampled (specifically, residues 20 and 46).

Interpatient Class II/Compensatory Mutations in Therapy Naïve Viral Populations are associated with greater diversity in the protease.

Class II/compensatory mutations (residues 10, 36, 63, 71, 77, 93) are variable prior to therapy, but increase in frequency with drug exposure (144). Because these mutations are polymorphic in a therapy naïve environment and increase under drug selection, they may make small contributions to the fitness of emerging resistant protease. Prior work suggests that these alleles should be abundant within the pre-treatment population (144). Sequences were obtained from the Stanford database representing population/consensus sequences from many subjects naïve to protease inhibitors. These sequences were filtered to remove any haplotypes that contained major Class III resistance mutations, indicating prior drug exposure. The sequences were then split into two groups, those that contained variation at Class II positions and those that did not. By parsing sequences into these two groups, nonconsensus polymorphisms can be assessed separately from sequences that had the diversity at these positions conserved.

Non-consensus diversity ≥ 0.1% was plotted per amino acid position (Fig. 3.2). There were 9 polymorphisms in 7 amino acid positions where a substitution was present at ~2x or greater frequency in sequences containing class II mutations versus conserved at those residues (non-class II) and whose lower bound in frequency for non-class II polymorphisms was ~0.4% (representing at least 9 independent observations). Those positions were 12IP, 16E, 37YH, 45R, 69K, 70T, and 89M. The positions that changed the most (3.5-4x) between class II and non-class II sequences were 12I (1.64 vs. 0.39, 4.2x), 98M (1.87 vs. 0.48, 3.9x), and 16E (4.32 vs. 1.17, 3.7x). 12P, 37H, 45R, and 70T changed ~2.5x (3.4 vs. 1.26, 1.71 vs. 0.65, 1.44 vs. 0.57, 2.48 vs. 1, respectively). 37Y had a 2.4x difference (0.94 vs. 0.39), and 69K a 2x difference (1.66 vs. 0.83). Sequences that contained consensus diversity at the Class II positions contained ~2-3x greater non-consensus polymorphisms at positions 12N, 18L, 39Q, 70R, and 92K. This pattern of linked polymorphisms may suggest compensatory effects among these mutations.

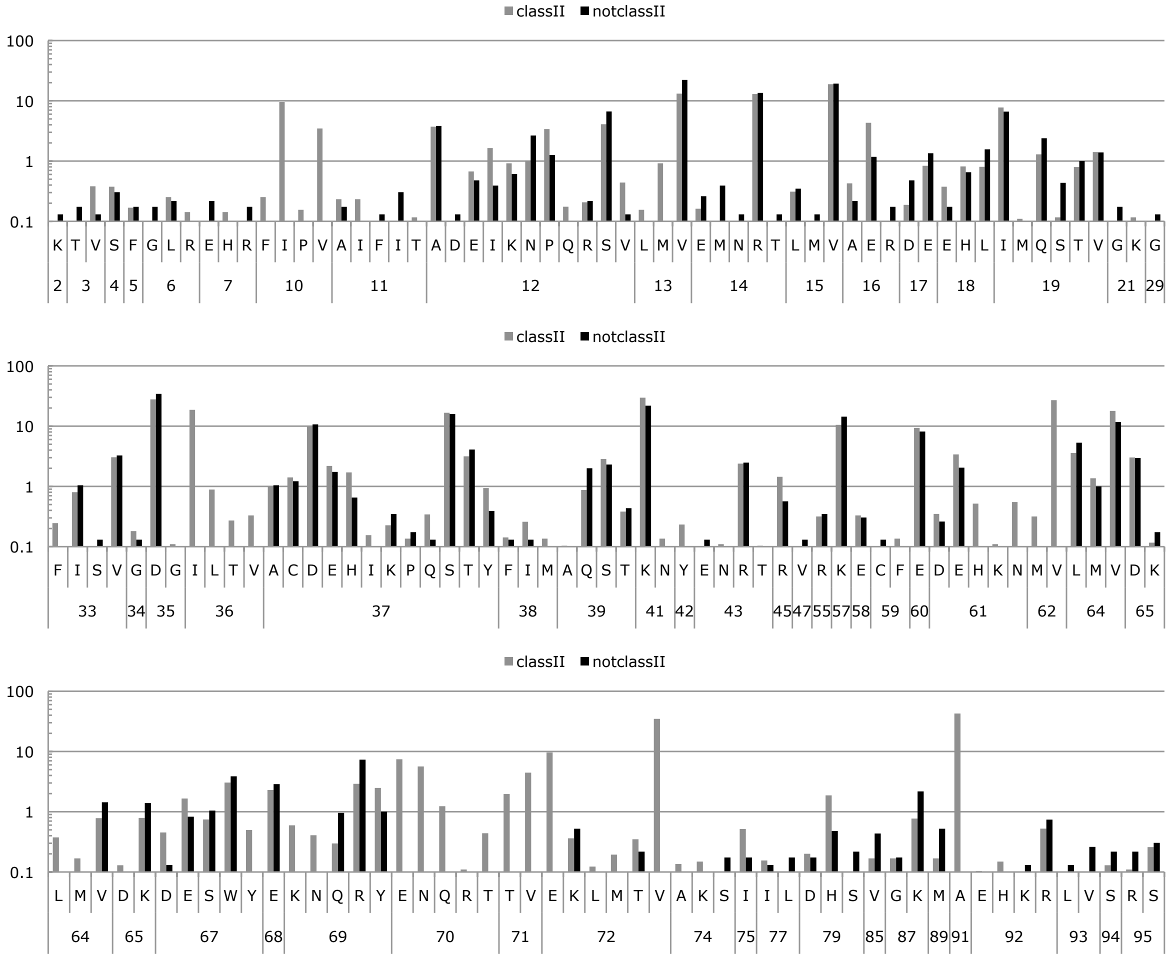


Fig. 3.3. Therapy-naïve non-consensus interpatient diversity ≥0.1% of protease in the Stanford database. Sequences derived from untreated subjects in the Stanford database were separated by the presence (gray bars) or absence (black bars) of Class II resistance associated alleles. The non-consensus diversity of these populations was then plotted per residue and polymorphism.

Diversity of the viral population within a subject is less than the diversity seen between subjects.

There were 20 residues in the database sequences (which represent the consensus/most abundant sequence in each subject) that harbored non-consensus alleles at a moderate frequency (≥5%) between subjects. These residues were compared to their frequency in the intrahost populations, and were found to be present but at a much lower frequency (Fig 3.4). Furthermore, for 18 of the 20 positions this difference was significant (Z-test).

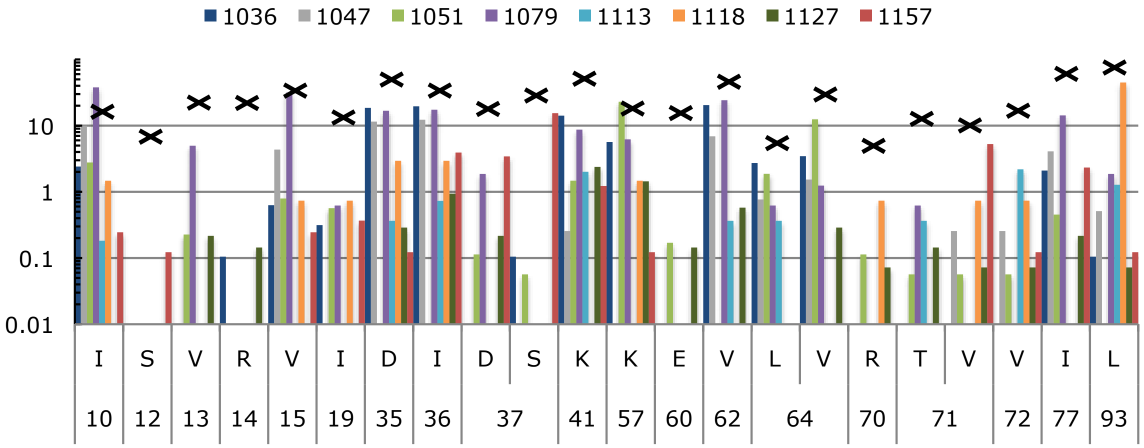


Fig. 3.4 Intrapopulation diversity contains major variants found across individuals in a database population but at significantly lower frequencies. Major (≥5%) interpatient non-consensus diversity in the database was compared to their frequency within individuals. Each colored bar represents an individual subjects, whereas black Xs indicate the weighted non-consensus mean frequency in the database population.

Plasma drug levels for the sampled populations were highly variable and did not correlate with the path to resistance.

To estimate the strength of drug selection on the treatment-experienced population, plasma levels of the protease inhibitor were measured. Across the entire cohort, drug levels ranged from below the limit of detection up to 19,830ng/mL (Fig. 3.1, Table 3.1).

All subjects experienced virologic failure and we interpret the relationship between drug exposure and resistance mutations as falling into three patterns. For subjects 1036, 1051, 1113, and 1127 higher levels of ritonavir were detected in the plasma samples and the resistance mutation V82A became largely fixed in the population, the commonly seen mutation associated with ritonavir resistance. In contrast, subjects 1079, 118, and 1157 had comparable levels of ritonavir in the plasma samples but had little or no drug resistant variants in their viral population; we interpret this to indicate poor adherence. Finally, subject 1047 had low levels of drug in the plasma but largely fixed the resistance mutation I84V with lower levels of V82A in the population; we interpret this to indicate that the I84V mutant was more fit than the V82A mutant yet provided sufficient resistance to this low level of drug exposure. The presence or absence of drug resistance mutations was not related to the pre-therapy diversity (π) in the population.

Population genetic estimates of the effective population size suggest that the effective population size of HIV-1 is much smaller than the census size.

The effective population size can inform whether evolutionary forces on a population will be deterministic or stochastic. N_e can be calculated using θ, as $\theta = 4N_e\mu$ for a diploid population, and $\theta = 2N_e\mu$ for a haploid population. HIV-1 is pseudodiploid, contains 2 copies of each allele, and can recombine, therefore the former equation was used to calculate N_e. θ is a measure of the expected nucleotide diversity of a population at mutation-drift equilibrium (58). Viral load is a surrogate for census population size. We estimated the number of segregating sites (positions in protease with a SNP), π (a measure of the number of polymorphic alleles and their frequency), θ ($4N_e\mu$), Tajima's *D* (a measure of deviation from mutation-selection equilibrium) (58), and the effective population size under the assumption of mutation-drift equilibrium. Using the equation $\theta = 4N_e\mu$, the effective population size was calculated to range from 620-1,129 virions (Table 3.3). These estimates are well within the range of previous inter-population data, but below the highest suggested values for N_e (56, 57, 139-142). There is a weak positive correlation between N_e and viral load (VL), meaning the difference between the census population size of HIV-1 (VL) and N_e across individuals is not uniform (Fig 3.5). This discrepancy suggests that the population dynamics of HIV-1–e.g. frequent population bottlenecks, high variance in offspring number, repeated selection—are reducing the effective number of reproductive virons dramatically. As a result, natural selection would not act as efficiently and a relatively small number of resistance alleles may have sufficient selective advantage to contribute to drug resistance. If HIV-1 is treated as a haploid entity, then the calculated N_e would be twice as large, but still small relative to the census population.

Table 3.3. Measures of population size and variation in the therapy naïve population.

	1036	1047	1051	1079	1113	1118	1127	1157
RNA copies/mL	236,040	279,000	501,600	199,520	276,100	205,600	139,080	380,200
S	174.00	119.00	172.00	103.00	133.00	78.00	178.00	141.00
π	0.03	0.02	0.02	0.03	0.02	0.03	0.01	0.01
θ	0.10	0.08	0.09	0.07	0.07	0.05	0.09	0.08
D	-1.90	-2.31	-2.33	-1.82	-2.24	-1.54	-2.66	-2.37
N_e	1129.28	876.39	1055.90	865.74	852.08	619.56	1086.81	917.36

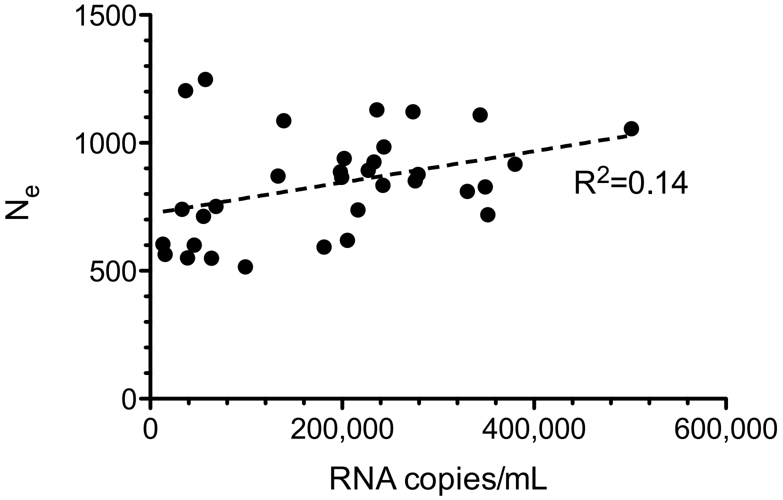


Figure 3.5 A weak positive relationship between HIV-1 census population size, as inferred from viral load, and the effective population size (N_e). N_e was inferred assuming pseudo-diploidy, a mutation rate of 2.16 x 10^{-5}, and θ estimated from the number of segregating sites.

The Tajima's *D* statistic of therapy naïve- and therapy-experienced populations does not change, suggesting the viral populations are not at mutation-drift equilibrium.

Tajima's *D* statistic measures the deviation between the expected distribution of allele frequencies at mutation-drift equilibrium and the observed distribution (113). Negative values of *D* can indicate recent selective sweeps, population bottlenecks, and population expansion; positive values can indicate balancing selection, migration, and population subdivision. We observed strongly negative Tajima's D in almost all samples (Table 3.3). Due to suboptimal selection in some subjects, the therapy-experienced population was further divided by susceptible and resistant variants. Tajima's *D* was not significantly different between pre-therapy, therapy-experienced susceptible, and therapy-experienced resistant populations, and most had a significantly negative *D* value (Fig. 3.6). The one positive *D* value was from a susceptible population that contained a small number of haplotypes comprising the majority of the variation. Because a strong negative *D* indicates an excess of low frequency polymorphisms, and because this value did not dramatically change with drug exposure, this suggests to us that multiple haplotypes with resistance mutations grew out with selection and that these haplotypes also displayed additional low abundance variants.

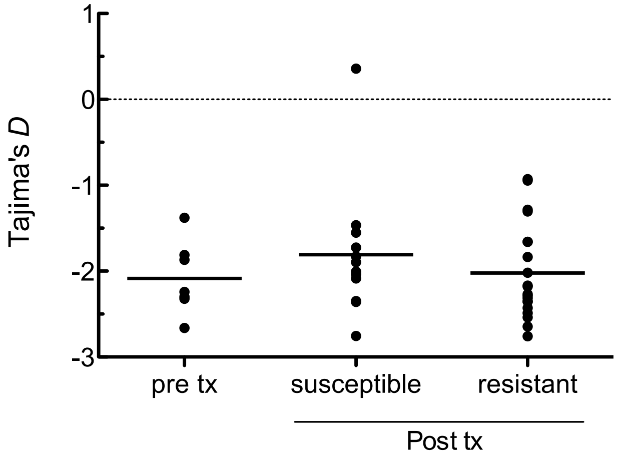


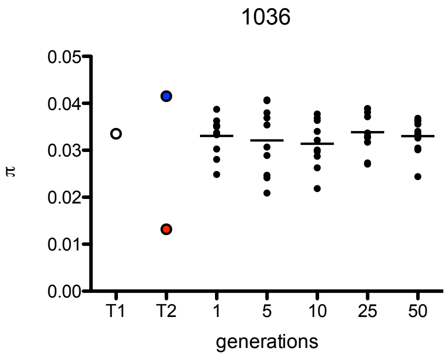
Fig. 3.6 Population diversity and structure across the entire cohort does not significantly change with drug exposure. Tajima's *D* statistic was calculated for pre- and post-therapy samples containing at least 100 sequences. Within a therapy-experienced population, susceptible and resistant subpopulations were analyzed separately.

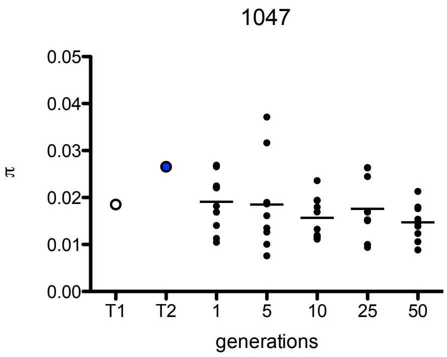
The diversity of the emerging resistant population can be greater than the pre-existing diversity, indicating the outgrowth of multiple resistance alleles.

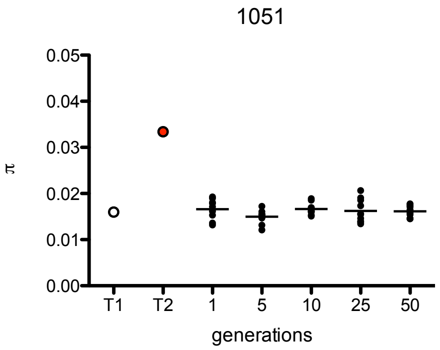
The significantly negative Tajima's *D* statistics indicates that the majority of populations were not at equilibrium. Because of this, calculations of N_e based on θ will be likely downwardly biased because $\theta = 4N_e\mu$ is only true for populations at equilibrium. Regardless, demographic complexities of intrahost populations of HIV-1 suggest that it would be difficult to predict the fate of resistance alleles from the pre-therapy data. Alternatively, one can observe how biological diversity changes under selection. The rebound population could indirectly reveal estimates of N_e based on *de novo* outgrowth of resistant alleles.

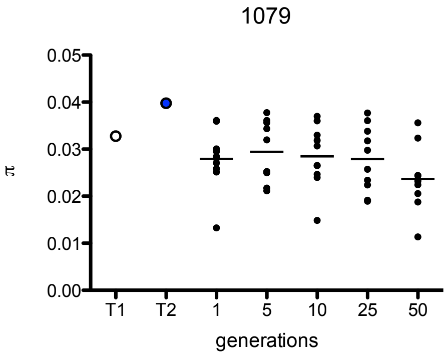
The nucleotide diversity metric π was longitudinally calculated for susceptible and resistant populations for the pre- and first post-therapy time points. Sequences were considered resistant if they contained one of the major protease inhibitor residues V82AIFTS, I84V, or L90M. To curtail sample size bias, π was only calculated for populations containing ≥40 sequences.

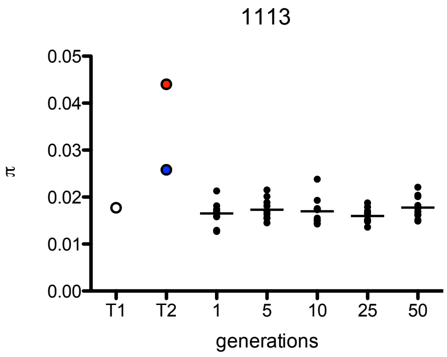
Interestingly, subjects 1051, 1113, 1127, and 1157 all had emerging resistant populations with a much greater π than the pre-existing population (Fig. 3.7), but often with fewer segregating sites (*S*). This result counteracts the simple model of drug selection resulting in the outgrowth of a predominant variant as these π and *S* values suggest multiple haplotypes with a resistant allele at intermediate frequency and a slight drop in low frequency alleles. In these subjects, the majority of variants contained resistant alleles. Furthermore, ≥ ~50% of the variation was made up of low frequency unique alleles <2% in frequency. Subject 1036 did not follow this trend because only a single major (9%) haplotype conferred resistance, and the absence of resistant alleles likely resulted in a lower π value for the emerging subpopulation in comparison to susceptible variant.

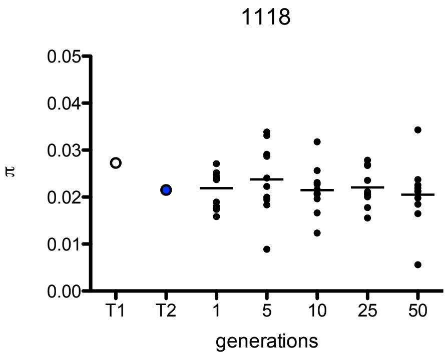

1036
π
0.00
0.01
0.02
0.03
0.04
0.05
T1
T2
1
5
10
25
50
generations

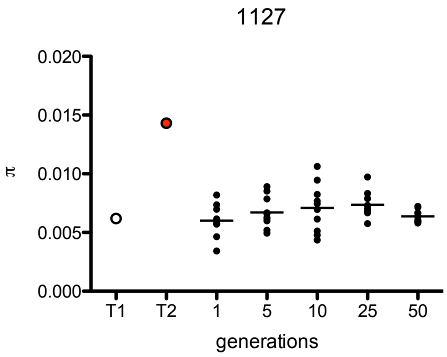

1047
π
0.00
0.01
0.02
0.03
0.04
0.05
T1
T2
1
5
10
25
50
generations


1051
π
0.00
0.01
0.02
0.03
0.04
0.05
T1
T2
1
5
10
25
50
generations


1079
π
0.00
0.01
0.02
0.03
0.04
0.05
T1
T2
1
5
10
25
50
generations


1113
π
0.00
0.01
0.02
0.03
0.04
0.05
T1
T2
1
5
10
25
50
generations


1118
π
0.00
0.01
0.02
0.03
0.04
0.05
T1
T2
1
5
10
25
50
generations


1127
π
0.000
0.005
0.010
0.015
0.020
T1
T2
1
5
10
25
50
generations

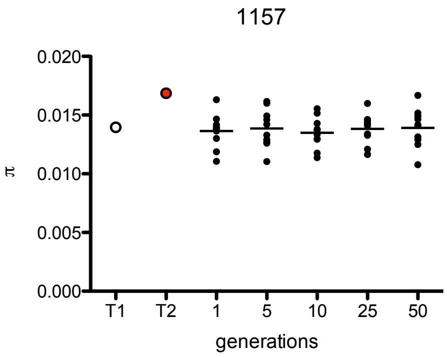

1157
π
0.000
0.005
0.010
0.015
0.020
T1
T2
1
5
10
25
50
generations

Fig. 3.7. Report of longitudinal π values from biological and simulated outgrowth populations. For each biological population, π was calculated for the pre-therapy time point (open circles), resistant (red circles) or susceptible (blue circles) populations when at least 40 sequences were present. Closed black circles represent π from 10 replicate sampling and outgrowth of the pre-therapy population that underwent a bottleneck of 90%. The bottlenecked population was allowed to grow out 1, 5, 10, 25, and 50 generations (approximately 1.5-150 days) with a mutation rate of $2.16x10^{-5}$.

The path to higher levels of resistance is unique in each subject but involves common resistance alleles.

The fate of emerging resistance alleles under selection has not been well characterized *in vivo*. To address the path to resistance, we examined the genetic diversity of longitudinally sampled populations under drug selection. Our data show that the pre-existing diversity is composed of very low frequency alleles in a large census population. Because of the limits of our sampling, determining whether haplotypes in these therapy naïve populations grew out during viral rebound populations is difficult to address. However, by looking at patterns of SNP variation on resistance haplotypes, we can infer which haplotypes likely were pre-existing and see if deterministic forces shape the evolution of resistance.

V82A, I84V, and L90M are well-characterized, major polymorphisms that confer resistance to the RTV protease inhibitor. In all populations that rebounded with resistance, at least one of these variants was present (Fig. 3.8). V82A was the most common allele enabling drug escape, emerging as single dominant resistance mutation for 1036, 1118, 1051, and 1157. V82A was sampled as pre-therapy variants in 3/8 individuals, all of which had this allele emerge but on different haplotypes than was detected as pre-existing. V82TFS are alternative nonsynonymous substitutions at position 82 that confer a low-level of resistance. These were only sampled in therapy-experienced populations, and none of these alleles grew out. I84V was sampled in 4/8 pre-therapy populations, none of which grew out in these individuals. All patients had I84V sampled in at least two time points, but this variant only grew out in patients 1047 and 1127. It was not detected as a pre-existing variant for these two individuals.

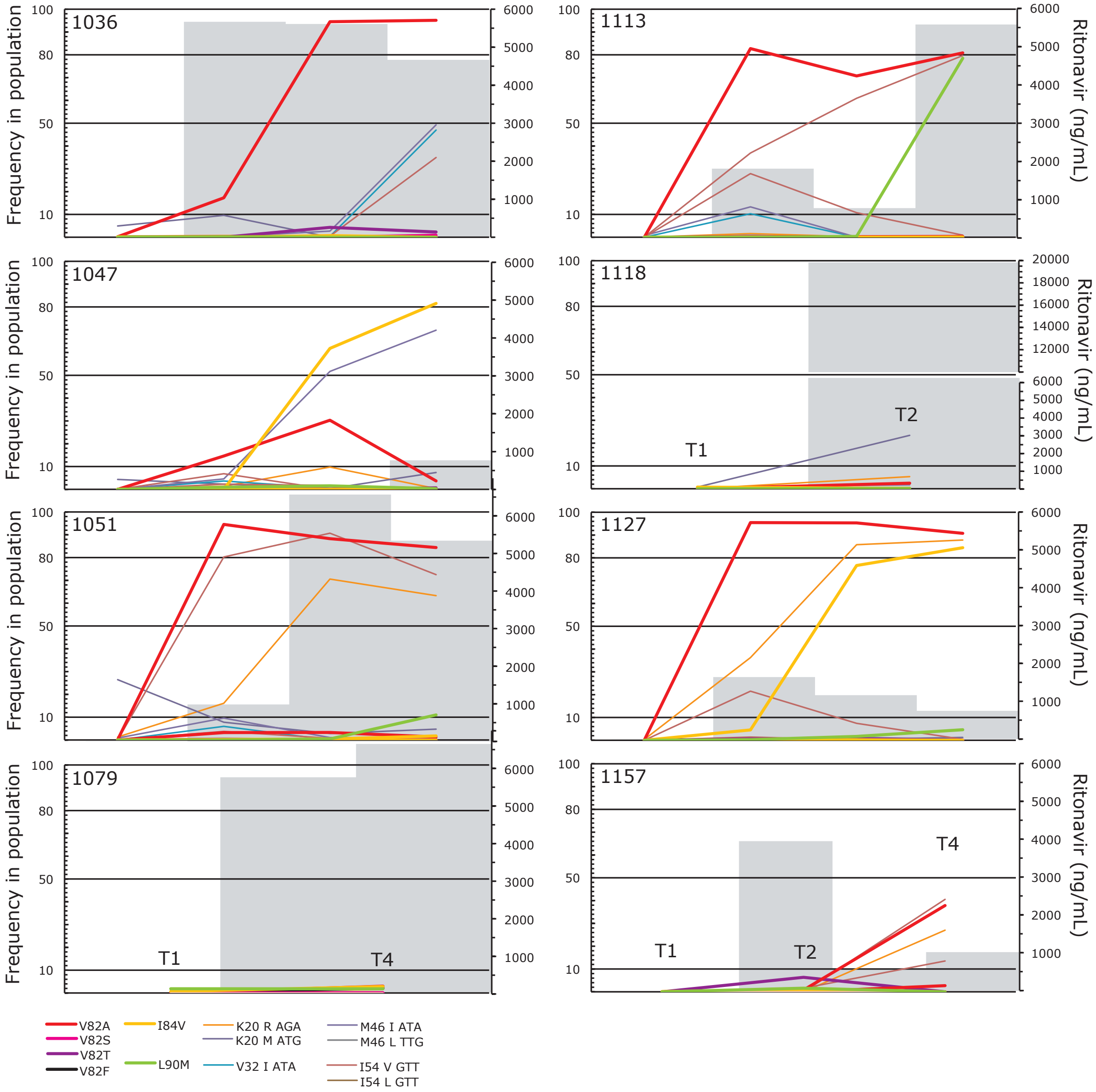
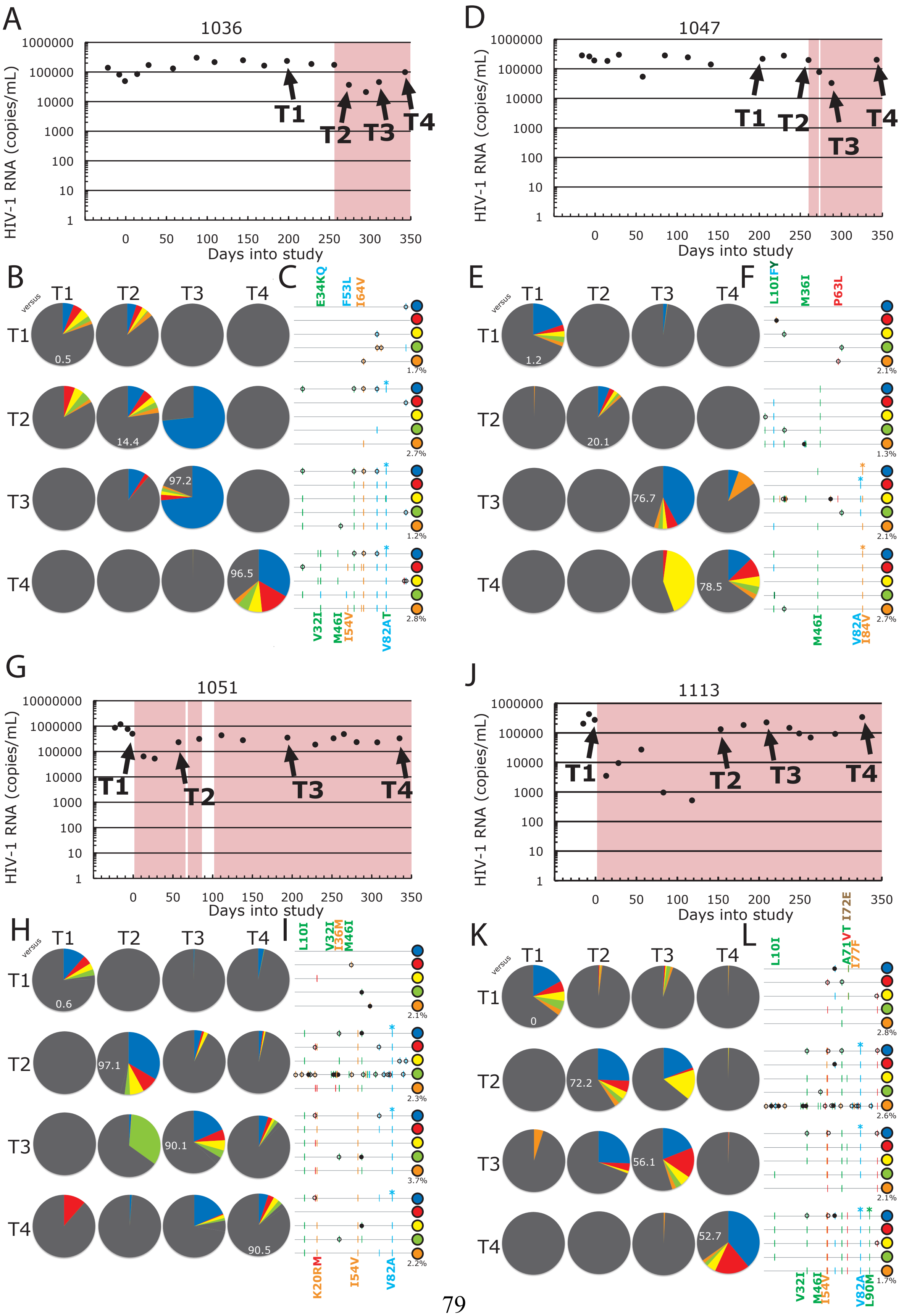


Fig. 3.8 Intrapopulation frequency of resistance associated alleles during drug selection. Major RTV resistance associated mutations V82ASTF, I84V, and L90M (thick lines) and Class III mutations (thin lines) were plotted by frequency over time. Time is illustrated from left to right, T1, T2, T3, and T4, respectively, unless otherwise noted. Gray bars (right y axis) denote the concentration of RTV in that individual sample.

It is also known that continuous, strong selection of a protease inhibitor results in the accumulation of additional resistance mutations in a step-wise manner (45). For subjects 1113 and 1127, V82A was linked to one other major resistance variant, L90M and I84V, respectively. Prior to adding either L90M or I84V, there were minor resistance mutations on the same haplotype of V82A such as K20R. The addition of a second major resistance allele further added an additional allele that was only seen when both major resistance alleles were present. For V82A, the gain of L90M also included L10I and A71V on the same major haplotypes. The addition of I84V was joined by L63P. Due to the strong level of resistance conferred by two major resistance alleles, these additional variants likely have compensatory roles.

For some subjects it was clear that intermittent drug exposure, either due to compliance or metabolic issues, allowed for the persistence of susceptible variants. This may have also limited the accumulation of additional major resistance alleles on the dominant haplotypes as the fitness cost of these alleles during time of reduced drug exposure may have prevented them from achieving high frequency. However, the overall trend supported the addition of new resistance mutations on haplotypes emerging in the rebound population (e.g. 1127, Fig. 3.9).


A
1036
HIV-1 RNA (copies/mL)
1000000
100000
10000
1000
100
10
1
0
50
100
150
200
250
300
350
Days into study
T1
T2
T3
T4
B
versus
T1
T2
T3
T4
0.5
14.4
97.2
96.5
C
E34KQ
F53L
I64V
1.7%
2.7%
1.2%
2.8%
V32I
M46I
I54V
V82AT
D
1047
HIV-1 RNA (copies/mL)
Days into study
E
78.5
1.2
20.1
76.7
F
L10IFY
M36I
P63L
2.1%
1.3%
2.1%
2.7%
M46I
V82A
I84V
G
1051
10000000
0.6
97.1
90.1
90.5
H
I
L10I
V32I
I36M
M46I
2.1%
2.3%
3.7%
2.2%
K20RM
I54V
V82A
J
1113
I72E
A71VT
I77F
K
0
72.2
56.1
52.7
L
L10I
2.8%
2.6%
2.1%
1.7%
V32I
M46I
I54V
V82A
L90M

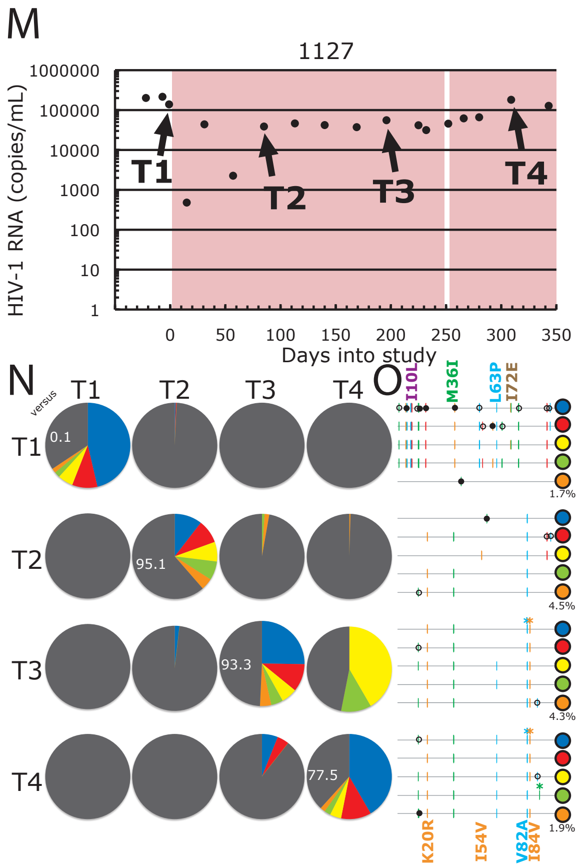


Fig. 3.9 Escape from ritonavir monotherapy is unique per patient and involves dramatic restructuring of haplotypes over time. The emergence and evolution of drug resistance is detailed for patients (A-C) 1036, (D-F) 1047, (G-I) 1051, (J-L) 1113, and (M-O) 1127. (A, D, G, J, M) For each individual, time points sampled (arrows) are plotted by study day and ritonavir exposure (shaded). (B, E, H, K, N) For each time point within an individual, the 5 most dominant haplotypes are represented in a pie chart by frequency for the other 3 time points. White inset numbers are the frequencies of non-dominant haplotypes that also have a major resistance allele at positions 82, 84, and/or 90. Color of the pie chart slice correlates to a given variant within a row. (C, F, I, L, O) Polymorphisms on each haplotype are detailed by a highlighter plot. Dominant variants that are not class III are listed above the highlighter plot above the variant and colored by the highlighter tick mark. Class III variants

are listed below the highlighter plot. Variants not described are designed by black circles as either coding (filled), or silent (open). Asterisks above the highlighter plots for each time point mark the location and color of lines that are major ritonavir resistant variants.

Because HIV-1 has a high capacity for both recombination and mutation, it is debated which of these mechanisms drives the step-wise addition of new resistance mutations on emerging haplotypes. If recombination is the driving force, the resistant recombinanat haplotype should be derived from parental haplotypes at a relatively high frequency as these haplotypes are more likely to have high levels of co-infection. (Similarly, if low frequency resistant haplotypes recombined, these increased resistance haplotypes would be more likely to be destroyed by subsequent recombination with common, but less fit haplotypes.) Under a stepwise new mutation model, one expects the most resistant haplotype to be derived from one of the common haplotypes of the previous sample.

As previously discussed, the subject 1113 had multiple major resistance mutations grow out on the same haplotype. For the final time point, V82A and L90M were linked on all of the top 5 haplotypes (Fig. 3.9*L*). In addition to V82A and L90M were also class II mutations A71V (5/5 haplotypes), L10I (top 4 haplotypes) and class III mutations I54V (5/5). The top 5 haplotypes for the prior sampled population (T3) had V82A without L90M, suggesting that L90M was added later with drug selection. Furthermore, L90M was not on any haplotypes in T3 at a high enough frequency and contained the allelic combination that would produce the variant emerging in T4 by recombination.

Comparing the pre-therapy population (T1) to the emerging population after drug exposure (T2-4), major haplotypes were absent or at extremely low levels for 4/5 individuals

(Fig. 3.9*E, H, K, N*). The haplotypes and frequency for patient 1036 (Fig. 3.9*B*) were largely preserved between the two time points, though T2 was sampled at a time of population contraction. Therefore, for each time point across the 5 individuals, the evolution of resistance involved a step-wise accumulation of mutations. These mutations were likely added *de novo*, not by recombination. Variants would have to be at a high enough frequency and with a complementary set of alleles that would produce recombinants identical to the haplotypes that emerged. In these individuals, recombination does not appear to play a dominant role in the emergence of resistance.

3.5 Discussion

HIV-1 population size and structure, as well as evolutionary forces that shape it, remain poorly understood and often debated (54, 56, 57, 139-142). At the crux of this debate are estimates of the effective population size (N_e), which can determine how strong selection must be for deterministic forces versus stochastic forces to shape the fate of new alleles and the population response to selection (58). Historically, quantifying N_e in HIV-1 has largely been limited to mathematical modeling, and many of these models infer *in vivo* diversity based on sequence data that only reveals major haplotypes within a population (54, 56, 57, 139-142). In contrast, we use next generation deep sequencing technology with a Primer ID to directly resolve the genetic diversity of *in vivo* HIV-1 populations (143) and then sampled these populations through a drug selection pressure to characterize the path to resistance and the forces that shape it.

Although HIV-1 appears to have a large census population size, N_e is comparatively small. We uncovered a large number of low frequency unique haplotypes that would typically be missed, but found that intermediate frequency variants were often lacking. As indicated by strongly negative Tajima's *D,* these minor variants were in excess of what would be expected under the Wright-Fisher mutation-drift model, indicating population not at equilibrium. Therefore, estimates of N_e based on θ (a measure of total population diversity) are likely downwardly biased, as $\theta = 4N_e\mu$ only at equilibrium (113). Typically, population bottlenecks, rapid population expansion, and selective sweeps can result in a negative *D* (113). The pre-therapy population and post therapy populations both had strongly negative *D*, which suggests that the demographic force shaping diversity is acting throughout the infection. Similarly, the selection for resistance alleles imposed by drug treatment did not strongly reduce *D* further. For the majority of emerging resistant populations, π was significantly higher than the diversity of the pre-therapy population. Since the structure of the population did not change, the higher π value for the emerging resistant population indicates that resistance emerges on novel alleles that have more SNPs between them in comparison to susceptible alleles in the pre-therapy population. This may be partially explained by an increase in frequency, but not fixation, of multiple resistance haplotypes within a population, despite a drop in the number of segregating sites. This pattern may suggest that there is appreciable clonal interference among emerging resistance haplotypes (51).

The apparent small N_e and skewed allele frequency distributions did not prevent the evolution of multiple resistant viral haplotypes—all subjects ultimately failed therapy and Class III (major) drug resistance alleles were segregating in all subjects. However, these same alleles are at low frequency prior to drug selection, consistent with their deleterious effects on fitness. Thus, the selective benefit of these alleles during drug treatment must also overcome their normally deleterious effects. The fitness cost of individual Class III polymorphisms in the therapy naïve environment can be calculated by assuming pre-therapy alleles are at mutation-selection balance and viral mutation rate (μ) 2.16x10^{-5}. The cost (s_c) for all Class III resistance alleles in the pre-therapy environment ranged from 0.1-3.1% (Table 3.4). N88SD, V82A, G48V, I54V, and M46I all had s_c ~2-3%. However, it should be noted that V32I, M46I ATT codon, M46L, and L90M were only sampled in a single individual therefore the estimate of s_c at best approximate for those polymorphisms.

Table 3.4. Average pre-therapy fitness values of Class III resistance mutations.

Allele[a]	Codon[b]	s_c[c]	S_{t10}[d]
K20M	ATG	0.61	1.72
K20R	AGG	0.47	1.48
V32I	ATA	0.17	1.33
M46I	ATA	0.31	1.43
M46I	ATT	3.09	4.01
M46L	TTG	0.43	1.35
G48V	GTG	2.16	3.05
I54V	GTC	2.88	3.80
I54T	ACC	0.73	1.96
V82A	GCC	2.03	2.94
V82I	ATC	0.62	1.90
I84V	GTA	0.87	2.05
N88S	AGT	1.91	2.88
N88D	GAT	2.00	3.01
L90M	ATG	0.12	1.27

[a] Class III resistance allele
[b] Resistance allele codon
[c] % disadvantage
[d] Average total % fitness when $s_t = 10/N_e$

For selection to dominate the fate of an advantageous allele with selective benefit s_b, $N_e s_b >> 1$ (58). Thus, a minimum estimate of s_b is the inverse of the effective population size (N_e). Therefore, for s to be beneficial, the minimum frequency it has to reach ($1/N_e$) is 0.09-0.16%. However, a more realistic and conservative estimate of this frequency (e.g. $10/N_e$, so that $N_e s_b >> 1$) is 0.9-1.6%. The total selective benefit (s_t) of a resistance allele during drug treatment is therefore $s_c + s_b = w$. The average fitness for each Class III resistance mutation ranged from 1.2-4.0% (Table 3.4). This is still likely an underestimate as an allele with a 1% advantage would require 5-6 years to sweep to fixation (58) and most resistance alleles swept in less than 200 days (~100 generations) (Fig. 3.8-9).

Each patient had a unique path to failing RTV monotherapy, and the dominant resistant haplotypes sampled during the course of failure were not identical to pre-therapy haplotypes nor were they always sampled in lateral time points (Fig. 3.9). The evolution of RTV failure followed a step-wise accumulation of mutations on common haplotypes (Fig. 3.9). Major variants that emerged within the population were likely formed by a *de novo* addition of alleles versus recombination events, as evidenced by an absence of major haplotypes in prior time points that could have been in a high enough frequency to recombine to form the emerging haplotype.

Why individuals fail therapy is likely a question that will need to be re-interpreted on a per-patient basis. Our sub-cohort of 10 individuals all had very similar clinical factors. They were chronically infected with HIV-1 subtype B, had extremely low CD4 counts, high viral loads, and no previous protease inhibitor exposure. It is remarkable that when given the same exact drug, each individual failed in completely different ways. Our work suggests a complex interplay between selection pressure, viral diversity, and population dynamics. It also suggests that the pre-therapy population may not be able to determine virologic outcome, or increased longitudinal sampling and depth is needed prior to drug exposure to better ascertain pre-existing resistance-associated variants. Our work illustrates how slight deviations in drug selection pressure shapes the emergence of resistance and dominant haplotypes. We also have evidence that the *de novo* addition of resistance-associated alleles on common haplotypes has a greater role in what dominant variants emerge versus recombination between haplotypes.

Chapter 4

Ultra-high resolution Primer ID sequencing reveals higher viral diversity and different pre-existing protease inhibitor resistance-associated mutations in HCV mono-infected versus HCV-HIV co-infected individuals

Cassandra B. Jabara[a,b,c,d], Fengyu Hu[d,e], Corbin D. Jones[a,f], Ronald Swanstrom[b,c,f], and Stanley M. Lemon[d,e]

[a]Department of Biology, [b]Lineberger Comprehensive Cancer Center, [c]UNC Center for AIDS Research, [d]UNC Liver Center, [e]Division of Infectious Diseases, [f]Carolina Center for Genome Sciences, [f]Department of Biochemistry and Biophysics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, US

4.1 Abstract

Interferon-sparing, all direct-acting antiviral (DAA) combination therapies are demonstrating increasing efficacy in treatment of chronic hepatitis C. Pre-existing resistance-associated variants (RAVs) may influence virologic outcome. Deep sequencing technology has the potential to provide novel insight into the frequency of such variants, but is confounded by high procedural biases and error. We previously described a Primer ID sequencing strategy that directly detects and eliminates many potential sources of error in deep sequencing, allowing for accurate ultra-high resolution analysis of viral populations. We applied this method to detect very low frequency (≤0.1%) variants in the HCV NS3 protease in HCV-treatment naïve individuals, and to determine whether genetic diversity and

the frequency of pre-existing RAVs differs between mono-infected and HCV-HIV co-infected individuals. In a cohort of 15 HCV mono-infected and 13 HCV-HIV co-infected subjects well-controlled on HARRT, Primer ID sequencing revealed that all populations had an excess of low frequency polymorphisms. We observed significantly less genetic diversity and possibly qualitatively dissimilar pre-existing RAVs in co-infected vs. mono-infected subjects. These differences may presage variation in virologic responses and possibly altered patterns of resistance to DAAs in mono- versus co-infected patients.

4.2 Introduction

Hepatitis C Virus (HCV) infects approximately 200 million people, and causes ~366 thousand deaths annually due to liver cirrhosis and hepatocellular carcinoma. The majority of contemporary transmissions are through blood-to-blood contact by way of intravenous drug use (IDU). Around 70 percent of individuals infected with HCV develop persistent infection (2-4), requiring antiviral drug intervention to suppress the intrahost viral population. Until recently, the standard of care for treatment of chronic HCV infection was a combination of pegylated interferon-α and ribavirin (PEG-INF/RBV) (145, 146); this therapy, poorly tolerated by many patients, has limited efficacy. Furthermore, only 42-46% of genotype 1 infected patients achieve a sustained virologic response (SVR) on PEG-INF/RBV (145, 146). Due to the poor efficacy of PEG-INF/RBV and development of compounds that directly target hepatitis C viral proteins, treatment is moving away from PEG-INF/RBV to all Direct Acting Antiviral (DAA) approaches. All DAA therapies offer several advantages over PEG-INF/RBV: a higher percentage of HCV infected individuals are eligible for treatment, DAAs are better tolerated, DAAs counteract the virus through direct targeting, and there is an increased breath of options and combinations, allowing for a more personalized approach and re-treatment of individuals experiencing virologic breakthrough. Disadvantages of DAAs are that they are expensive, patients can experience virologic breakthrough (albeit commonly delayed), and DAAs contain low genetic barriers.

DAA development has focused on the NS3/4A serine protease due to the multifaceted and critical functions it has in viral production and persistence. The NH_2-terminal third of NS3 contains proteolytic activity; NS3 heterodimerizes with co-factor NS4A, and this complex cleaves four sites downstream of the NS2-3 junction in the HCV polyprotein (7), allowing for the release of nonstructural proteins. Furthermore, NS3/4A cleaves host signaling molecules activated by dsRNA, blocking the TRL3 and RIG-I pathways (reviewed in (5)). In addition to proteolysis, the COOH-terminal of NS3 contains NTPase activity and a RNA helicase domain. Both are critical for replication (147). Although these two domains fold separately from each other, the COOH-terminal residues are found positioned within the shallow active site (147-149), and have been used as guidelines for the development of small molecule inhibitors (150, 151).

Due to interconnectedness of proteolytic and helicase activities, protease inhibitors can effectively suppress the viral population through multiple mechanisms, and viral diversity that decreases or prevents DAA binding likely has large negative fitness effects. However, because HCV can produce large, heterogeneous populations in a host, resistance associated variants (RAVs) are commonly present in an untreated environment. HCV's replication cycle has a half life of 2.7 hrs (range 1.5-4.6 hrs) (32), and approximately 10^{12} virions are produced per day (32-34). HCV's RNA-dependent RNA polymerase (RdRp) NS5B adds $1x10^{-4}$ to $1x10^{-5}$ substitutions/site/generation (11), and is biased in G:U and U:G mismatches (29). The quick life cycle and high production rate of new mutations introduced into the small, ~10kb viral genomes has led to estimates that at any given time within a large population, every single mutation can pre-exist (35), but the frequency at which an allele will exist is the balance between the rate at which it is added (mutation) and removed (selection).

Deeply characterizing the standing genetic variation within an *in vivo* population can detect pre-existing resistance associated variants. However, their steady-state frequencies and whether they can be directly selected under a drug is unknown. Furthermore, intrahost viral diversity can be influenced by a co-infecting pathogen, potentiating differences in pre-existing RAVs and DAA response. The HCV infected IDU cohort overlaps with individuals infected with the blood-borne Human Immunodeficiency Virus (HIV). HCV-HIV co-infection is predominantly with HCV genotype 1, (152) and co-infection is common in countries that have a high IDU rate and disease prevalence of both pathogens. For example, HIV-HCV co-infection in Russia and the Ukraine has been reported at 70% (153).

The biologic effects and clinical observations of mono- versus co-infection suggest dissimilarity of intrahost viral diversity. HCV-HIV co-infection increases patient morbidity and mortality (65) by causing a three fold acceleration in fibrosis, cirrhosis, and liver disease (66). HIV has direct effects on liver injury; the virus can fuse to hepatic stellate cells and hepatocytes by way of CCR5 and CXCR4 co-receptors, inducing apoptotic and pro-inflammatory pathways (154, 155). Despite this, previous studies have failed to demonstrate a difference (67, 68), or consensus (69-72) on whether HCV-HIV co-infection influences viral diversity, though these observations are likely confounded by variance in sampling depth, cohort size, methodology, and genomic regions examined .

The standing genetic variation of *in vivo* HCV populations is poorly characterized, leading to biased estimates of RAVs and ignorance of their fate after selection. Because co-infection may influence HCV diversity and differences in RAV frequency, we applied a novel, high-resolution deep sequencing approach to directly resolving the genetic variation of NS3 in 15 HCV populations mono-infected and 13 populations co-infected with HIV. Deep sequencing with Primer IDs directly corrected for PCR biases and error inherent in deep sequencing protocol (143) and captured thousands of genomes per population.

Because Primer IDs can form secondary structure and interactions with structured RNA genomes may result in non-random tagging, we first compared replicate tagging and sequencing reactions for 4 subjects, 2 mono-infected and 2 co-infected. We found that throughput was enhanced by using patient-specific primers and decreased PCR cycling conditions. Increasing cDNA synthesis temperatures resulted in a decrease in depth but an increase in the number of sequences recovered per Primer ID for mono-infected but not co-infected subjects for reasons that remain unclear. Independent re-constructions of replicate populations revealed a strong correlation of individual haplotype frequencies, further supporting random tagging.

For the 15 mono-infected and 13 co-infected populations, average nucleotide diversity (π) was higher in mono-infected individuals. A sliding window analysis of π demonstrated variability between subjects and residues, including active site residues. Each population contained an excess of low frequency polymorphisms, indicating population disequilibrium. Pre-existing RAVs were detected in 13/13 co-infected and 13/15 mono-infected individuals. When comparing the presence and frequency of pre-existing resistance mutations between mono-infected and co-infected individuals, RAVS were more frequent at

NS3 A.A. 36 and 168 in co-infected individuals, and at 41, 55, 107, 109, and 170 in mono-infected subjects. Repeated resampling analyses suggested these differences sort with HIV co-infection. These preliminary findings indicate that DAA response between the two groups may be different as well, and support alternative treatment approaches based on co-infection status.

It is currently unknown whether RAVs in the standing genetic variation can be directly selected under a DAA. If pre-existing diversity can predict drug response, not only would studies like this increase the probability of obtaining SVR by informing drug choice, but also advocate for alternative therapeutic approaches in mono-infected and co-infected populations.

4.3 Materials and Methods

vRNA extraction and cDNA synthesis. Human blood sera samples provided by the University of North Carolina Hospitals AIDS Clinic and UNC Liver Center for analysis were HCV genotype 1a positive, had a viral load ≥ $5x10^5$ IU/mL, and CD4+ count >350 cells/mm^3. Co-infected subjects were on HAART and had undetectable HIV viral loads. HCV genotyping was confirmed by RT-PCR and nested-PCR with primers targeting core NS5B proteins.

For each sample, vRNA was extracted using the QIAamp Viral RNA Kit (Qiagen, Valencia, CA). Approximately 10,000 copies of viral RNA from each sample were present in the cDNA synthesis reaction as previously described (87, 109, 110). The tagging primer used was, 5'-

ACCTTGCAAGCACGCTCTGGCCTTGAANNNNNNNNNCT(BARCODE)GAACACCGG GGACCTCATGGTTGTCTC -3'. The barcode represented individual *a priori* selected sample IDs. The 3' end of the tagging primer targeted downstream of H77 amino acid 175 (H77 3971-3945) and was customized per patient based on population sequencing. Oligonucleotides were purchased from IDT and were purified by standard desalting.

Amplification of tagged sequences. The single-stranded cDNA was column purified using the PureLink PCR Purification Kit (Invitrogen, Carlsbad, CA), using Binding Buffer HC (high cut-off) 3X wash to remove the cDNA primer. Primer removal was verified by electropherogram analysis using an Experion HighSense RNA microfluidic chip (Bio-Rad Laboratories, Hercules, CA). Samples were amplified by nested PCR, using upstream primers 5'-TAYTGCTYGGRCCRGCYGA-3' (H77 3370-3388) and 5'-AGTGGAGGGTGAGGTCCAGAT-3' (H77 3505-3523). The second round upstream PCR primer was customized per patient based on population sequencing. The downstream primers targeted the 5' portion of the cDNA tagging primer 5-ACCTTGCAAGCACGCTCTGGC-3' then 5'-CAAGCACGCTCTGGCCTTGAA-3'. The PCR reaction used PrimerSTAR^{TM} HS kit (TaKaRa, Japan). Each reaction contained 1x PrimerSTAR^{TM} HS Buffer Premix. For the first round of PCR, purified cDNA was split into to two 50ul reaction system. Each reaction was amplified 20 cycles of 98°C for 10 seconds, 68°C for 45 seconds. After the first round PCR, 1ul of the combined first round PCR reactions went into second round PCR. Second round PCR was amplified for 20 cycles of 98°C for 10 seconds, 68°C for 45 seconds, and gel purified using a 2% agarose gel and MinElute gel extraction kit (Qiagen, Valencia, CA), with incubation of the solubilization buffer at room temperature. DNA was quantified by Qubit fluorometer using dsDNA High Sense assay (Invitrogen, Carlsbad, CA). Product generation, quality, and primer removal was verified using a Bioanalyzer DNA microfluidic chip (Agilent, Santa Clara, CA).

454 pyrosequencing. Amplicons were quantified by picogreen and equimolarly pooled into individual libraries containing a median of 4 samples each. Rapid library adaptors were added to the amplicon populations by a blunt end ligation reaction, and this library was diluted, clonally amplified onto beads using emPCR, and sequenced on the 454 GS Junior with Titanium sequencing chemistry (Roche, Nutley, NJ). After a sequencing run, raw reads were initially processed through the native amplicon pipeline using default settings (Roche, Nutley, NJ).

Bioinformatic pipeline for raw sequence processing. A suite of programs was written to filter and parse raw 454 sequencing reads as previously described (143). In short, each full-length read was evaluated for a barcode and Primer ID in the cDNA primer 5' tail. When three sequences contained an identical Primer ID within a sample, a consensus sequence was constructed using ClustalW followed by MUSCLE (156, 157) then called by majority rule. Ambiguous nucleotide designations were used when there was a tie.

Intrahost population analysis. All statistical and diversity analyses were completed on populations devoid of sequences containing ambiguities and whose gaps were resolved by the consensus. Tajima's test and sliding window analysis of π were computed by DnaSP v.5.10.01 (111). The sliding window analysis measured π in windows of 100 nts and a step size of 10 bases. Population diversity and haplotypes were computed through customized bioinformatics suites.

RAV resampling simulation. A random draw simulation was written to conduct a label-blind analysis of RAV presence while controlling for sampling depth. In short, intrahost populations were split into two groups based on co-infection status. Within a group, individual populations were randomly chosen and shallowly sampled to a uniform depth. The sampled sequences making up the synthetic population were then examined for the presence and frequency of RAVs. This process was repeated for both groups, and 100 times within a group. Finally, the difference in RAV frequency between the two groups was logged. To create a label-blind control, all populations were pooled together prior to random sampling but otherwise treated identically.

4.4 Results

Comparison of two independent cDNA tagging reactions demonstrates similar Primer ID re-sampling distributions.

Primer IDs are a string of degenerate sequences embedded in the cDNA synthesis primer that are used to label individual viral genomes. Due to the degenerate block, some Primer ID combinations may form secondary structure and preferentially target the highly structured HCV genome, leading to tagging and re-sampling bias. A tagging bias may distort the final population resolution, as reads within a Primer ID population may be representative of the entire population. A re-sampling bias will result in a decrease of throughput, as many reads originate from a single Primer ID.

To increase the fidelity of primer-template binding, we designed the cDNA synthesis target site and upstream primers to match the intrahost consensus diversity. Furthermore, we limited the number of PCR cycles to decrease the re-sampling for all Primer IDs. Finally, because secondary structure may interfere with cDNA synthesis, we wanted to verify that higher reaction temperatures did not affect population resolution.

Approximately 10,000 copies of vRNA from 2 mono-infected (M1, M2) and 2 co-infected (C1, C2) subjects were independently tagged with cDNA primers targeting NS3 and containing a Primer ID. The amplified fragment encompassed residues 36-170 of NS3, critical positions that can confer DAA resistance. Amplicon populations were sequenced on the 454 Junior platform. The second run used standard cDNA synthesis conditions, whereas the first run had slightly higher temperatures. The two pyrosequencing runs yielded 116,211 and 113,932 raw reads, respectively, with 99.8% and 99.0% of sequences above 495bps. There were 1,492 to 3,589 consensus sequences resolved per sample, resulting in population depths between 0.067% - 0.028%. Because sampling a low-abundance allele is based on the Poisson distribution, the depth of sampling that would sample an allele with certainty (3/n) was 0.2%-0.08% (Table 4.1).

Table 4.1. Population sequencing throughput and depth for two replicate 454 Junior runs

		Run 1	Run 2
	Raw reads	116,211	113,932
	% above 495bp	98.81	99.05
Pre-consensus reads			
	M1	29,766	15,413
	M2	20,619	18,098
	C1	14,461	20,370
	C2	37,864	48,410
Primer IDs			
	M1	3,736	4,536
	M2	2,492	3,582
	C1	8,111	11,071
	C2	3,663	4,872
Consensus sequences			
	M1	2,798	2,445
	M2	2,086	2,571
	C1	1,492	2,150
	C2	2,947	3,589
Maximum Primer ID re-sampling			
	M1	58	22
	M2	37	26
	C1	19	27
	C2	77	71

Slightly higher cDNA synthesis temperatures (Run 1) resulted in more sequences being built per Primer ID and a decreased number of consensus sequences for the mono-infected subjects (Table 4.1). The change in temperature did not affect the distribution for co-infected patients for reasons that remain unclear, though this observation is on a very a small sample size and may not have biological implications (Fig. 4.1).

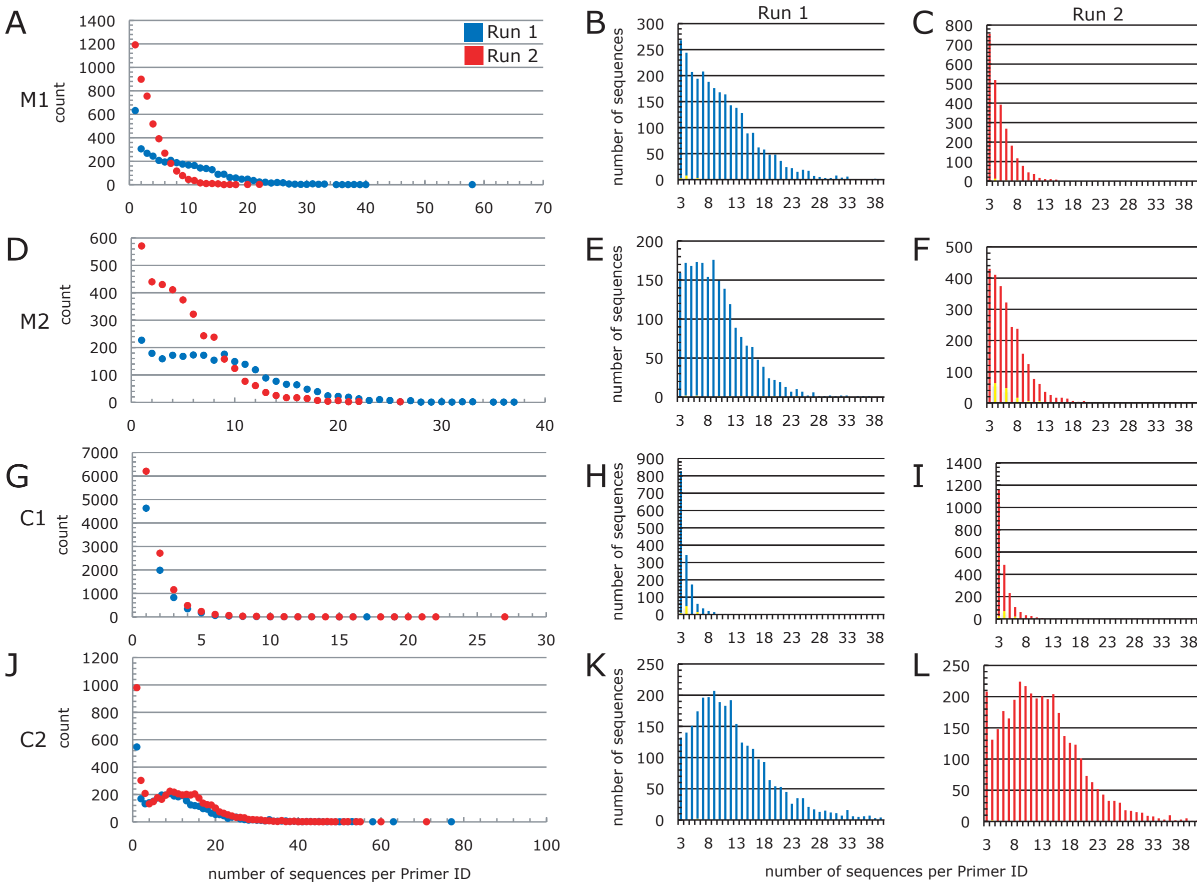


Fig. 4.1 cDNA synthesis temperature affected Primer ID re-sampling for mono- but not co-infected subjects. Distribution of the number of reads per consensus sequence in mono-infected patients M1 (A-C) and M2 (D-F), blue bars, and co-infected patients C1 (G-I) and C2 (J-L), red bars. For mono-infected patients M1 and M2, there were was a difference in the number of sequences that built each consensus. Less reads per Primer ID resulted in more ambiguous calls (F, H, I).

There was an absence of Primer ID oversampling but some evidence of resampling bias.

Although cDNA synthesis conditions resulted in a difference in the number of sequences that built each Primer ID in the mono-infected populations, none of the Primer ID tags for any of the populations were oversampled (Table 4.2). The maximum number of times a Primer ID was observed was between 27, and 77, indicating that a single Primer ID combination did not have an increased fitness in tagging and/or amplification.

Table 4.2. Replicate re-sampling of Primer IDs indicate absence of a re-sampling bias

	Run 1	count	Run 2	count
M1	CACCGTAT	58	TATCCGGC	22
	TTCGTCGA	40	CTACGACT	22
	CCCGCCCC	39	ACCGAACA	22
	CGATTGCC	38	CCCGCCCC	20
	ACGTGTTC	38	CCCATATA	20
	ATCACATA	37	GAGATTCA	18
	ACCAGTAC	36	GCGCCGCG	17
	CCGCCACC	35	CACCCCGC	17
M2	AACATCCC	37	CGGGTGAA	26
	ATGTGCTA	36	AATGCAAT	26
	AGATGATA	35	TGGCGAAA	22
	TTAACATA	33	GTCGATAT	22
	ACATGCTA	33	ATAACTAT	22
	CATACTCA	32	GTGAGCAC	21
	AACATGAC	32	ACATAGAA	21
C1	CCCCGCCC	19	CCCGCCCC	27
	CTCCCCAC	17	CCCCTCCC	22
	CCGCCCAC	16	CCCCCGCC	21
	CCCCTCCC	16	CCCCCACC	20
	CCCCCGCC	16	CCCTCCCC	19
	CCCTCCCC	15	CCCCACCC	18
C2	CCTAGTCT	77	AATAACAA	71
	AATACCAA	63	GCCGCGAG	60
	GACACAGG	58	TATTGTTA	55
	GTTGAATC	54	ACCAATAT	54
	CCTACATG	54	TAGCGTAC	52
	TTCATTAA	53	ATCGGAGG	52
	ATTCAAGA	53	ATTGATAA	51

For M1, M2, and C2, The Primer IDs resampled the most within a patient were largely different across replicates, indicating that tagging is random. The most prevalent Primer IDs from C1 were C-enriched, indicating that the degenerate region in the tagging primer was not optimally randomized during oligonucleotide synthesis or Cs were better at tagging and/or amplifying the population. However, tagging and sequencing increased numbers of populations (discussed below) have not reproduced this observation.

There is a strong reproducibility of the frequency of identical haplotypes resolved across independent tagging and sequencing runs.

The distribution of Primer IDs support unbiased tagging and amplification of viral sequences. To see if the distribution of alleles and their frequencies was reproducible, the genetic variation for all haplotypes present ≥ 1% in the first run was compared to their frequency in the second run. There was a strong correlation (R^2=0.81) of haplotype frequency across replicates (Fig. 4.2). Furthermore, this correlation likely would have been stronger had identical cDNA synthesis methods been used. Regardless, these data clearly show that the genetic variation of HCV populations can be reproducible resolved using deep sequencing with Primer IDs.

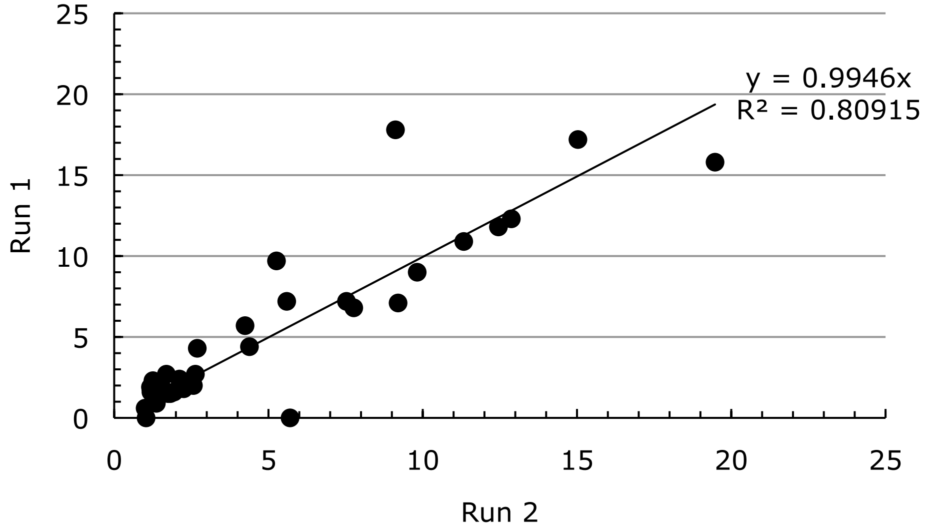


Fig. 4.2. Comparison of the frequency of haplotypes ≥ 1% between Run 1 and Run 2 demonstrates tagging and sequencing reproducibility. For each haplotype found ≥ 1% in Run 1, the frequency was compared to Run 2, and the two frequencies were plotted against each other. A strong linear correlation demonstrates that Primer ID tagging, amplification, and sequencing were highly reproducible.

As the frequency of RAVs near our limit of sensitivity, their resolution is confounded by the sampling probability and early introduced error.

Although haplotypes ≥1% can be sampled reproducibly, minor variants whose frequency nears the population depth may be missed by chance. Similarly, minor variants that appear as single occurrences cannot be differentiated from technical error, confounding whether a polymorphism is biological or artifactual. Slightly less than half of the pre-existing RAVs were sampled as single occurrences within a population (Table 3). However, half of these residues were recorded in the replicate run, supporting the polymorphism as biological, not error. To determine whether a RAV is biological without performing replicates, an alternate assay such as allele-specific PCR could be implemented in parallel. Furthermore, depth of sampling and/or repeated sampling will increase the probability that a low-frequency allele will be revealed.

Table 4.3. Frequency of resistance associated alleles across independent tagging and sequencing runs.

pt	RAV	codon	Run 1	Run 2
mono-infected				
M1	V36M	ATG	0.072	
	T54A	GCT		0.041
	V55A	GCC		0.082
	R109K	AAG		0.041
	R155K	AAG		0.041
M2	V36M	ATG	0.048	0.039
	V55A	GCC	0.048	0.156
	Q80K	AAA	98.802	98.950
	Q80K	AAG	1.103	0.933
	Q80K	AGA	0.096	0.117
	R109K	AAG		0.039
	R155K	AAG		0.039
	V170T	ACC	0.096	0.039
co-infected				
C1	V36A	GCG	0.134	0.047
	V36L	TTG	0.067	
	T54S	TCT	0.067	
	T54S	GCT		0.047
	V55A	GCC	0.067	0.047
	Q80K	AAA	99.196	99.581
	Q80K	AAG	0.670	0.279
	R155K	AAG	0.134	
	V170T	ACC	0.067	0.093
C2	V36M	ATG	0.068	0.028
	V36A	GCG		0.111
	F43I	ATC	0.034	
	F43S	TCC		0.028
	T54A	GCT	0.068	0.028
	Q80K	AAA	99.525	99.387
	Q80K	AAG	0.339	0.502
	Q80R	AGA	0.136	0.028
	R109K	AAG	0.068	
	A156V	GTC		0.028

454 Junior run throughput and patient resolution for non-replicates

The replicate sampling and resolution of 2 mono-infected and 2 co-infected individuals supported that Primer IDs could be used to accurately resolve viral diversity of HCV and that RAVs commonly pre-exist in the standing genetic variation. In order to better characterize whether a co-infecting pathogen influences diversity, and whether there was a difference in pre-existing RAVs, 24 additional patients, 13 mono-infected (samples 1-3,6-15), and 11 co-infected (samples 3-13) were sequenced (Table 4.4). The diversity of NS3 for monoinfected patients 1-3 was resolved on the GS FLX with Titanium sequencing chemistry, but due to sequencing length constraints, only contained NS3 residues 36-138. The replicate Junior runs previously described were combined per patient, and added to the cohort (M1-M2, C1-C2). Mean sampling depth for the entire cohort was 0.14% (0.11% median) (Table 4.4)

Table 4.4 Population sampling depth for mono-infected and co-infected subjects

	Sample	Barcode	Primer IDs	Consensus sequences	Depth
Mono-infected					
	1	TAT	9,106	2090	0.05
	2	TGT	2,113	703	0.14
	3	TAG	1,030	400	0.25
	M1	ACG	8,272	5,243	0.02
	M2	AGC	6,074	4,657	0.02
	6	GCAG	2,791	556	0.18
	7	GAG	307	231	0.43
	8	CAC	2,475	947	0.11
	9	CGCG	911	283	0.35
	10	GACT	4,900	2637	0.04
	11	CGT	4,362	2622	0.04
	12	TGTC	2,397	1079	0.09
	13	CATA	2,453	707	0.14
	14	AGAT	2,041	1001	0.10
	15	ACAG	567	385	0.26
Co-infected					
	C1	TAT	19,182	3,642	0.03
	C2	TTC	8,535	6,536	0.02
	3	GTA	2,418	1,053	0.09
	4	CTG	3,966	778	0.13
	5	ACGA	3,768	885	0.11
	6	ATAC	2,507	1765	0.06
	7	TCAT	965	760	0.13
	8	GTGT	852	569	0.18
	9	CTAT	1,132	893	0.11
	10	GCTA	1,343	774	0.13
	11	GATC	3,547	1829	0.05
	12	TTC	2,564	1196	0.08
	13	ATGC	2,562	834	0.12

Sliding window analysis of nucleotide diversity (π) across NS3 amino acids 36-175 show a greater diversity in the mono-infected population.

The standing genetic variation of each population revealed rich allelic diversity, including low frequency alleles. However, in order to assess whether the diversity is significantly different between the mono-infected and co-infected population, a sliding window analysis of the average pairwise diversity (π) was measured. Sliding window π reveals localized polymorphism levels, and demonstrated that the inter-population values of π across individual residues of NS3 did not identify specific regions of diversity or conservation (Fig. 4.3*A*). Similarly, the mean diversity across individuals varied greatly and was not position dependent. The lack of regional polymorphic conservation or diversity indicates that there isn't a strong selective across these positions, including active site residues. This is supported biologically, as NS3 does not form a highly structured pocket (8). In contrast, diversity is conserved across the highly structured HIV-1 protease's active site (Fig. 4.3*D*).

The differences in morbidity and mortality between HCV mono-infected and HCV-HIV co-infected individuals suggested dissimilarity in viral diversity, and we found that average diversity across NS3 was significantly greater in the mono-infected subjects than the co-infected subjects (Student t's test, p<0.0001) (Fig. 4.3*BC*).

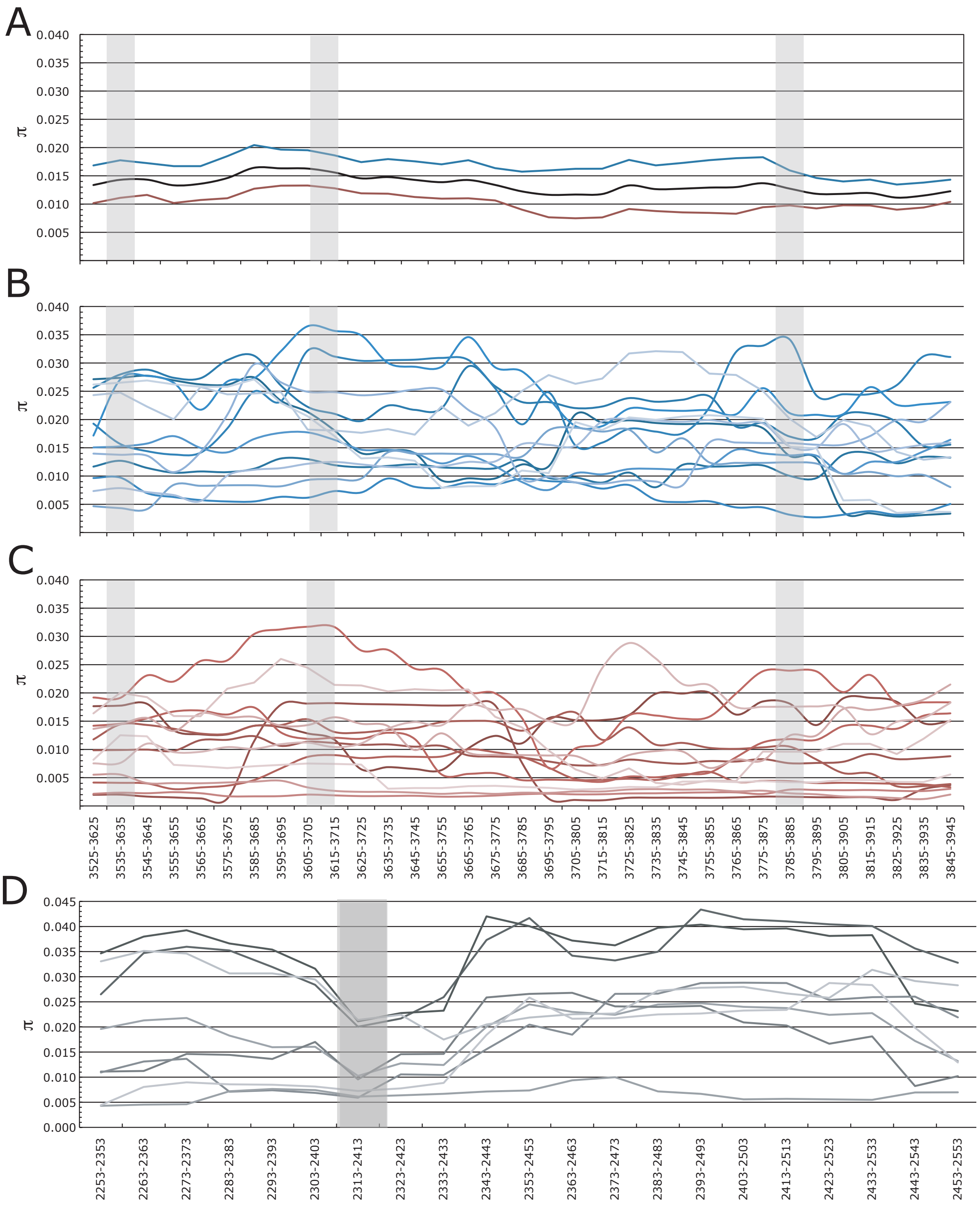


Fig. 4.3 Sliding window analysis of nucleotide diversity (π) across NS3 amino acids 36-175 and the HIV protease coding domain. A) Average sliding window π across NS3 in all

individuals (black), mono-infected (blue), and co-infected (red). B) Sliding window across NS3 in individual mono-infected individuals. C) Sliding window across NS3 in co-infected individuals. D) Sliding window across HIV-1 protease coding domain in therapy naïve individuals. Shaded areas indicate active site residues.

Nucleotide diversity (π) and Tajima's D were significantly different between mono-infected and co-infected populations.

Whole gene pairwise nucleotide diversity (π) was measured for the mono-infected and co-infected populations. The total π in the mono-infected subjects had a significantly greater diversity than the co-infected subjects (Student's t-test, p=0.01), as expected from the sliding window analysis. π provides a measure of diversity, but the Tajima's *D* statistic measures the deviation between the expected distribution of allele frequencies at mutation-drift equilibrium and the observed distribution. Negative values of *D* can indicate recent selective sweeps, population bottlenecks, and population expansion whereas positive values can indicate balancing selection, migration, and population subdivision.

For all populations, Tajima's *D* was ≤ -2.0 indicating an excess of low frequency polymorphisms. Tajima's *D* was also significantly lower in the co-infected population (Student's t-test, p=0.02). The lower D values in co-infected populations suggest that co-infection enhances the evolutionary forces skewing allelic variation in HCV populations. For instance, increased number of selective sweeps could shift this equilibrium.

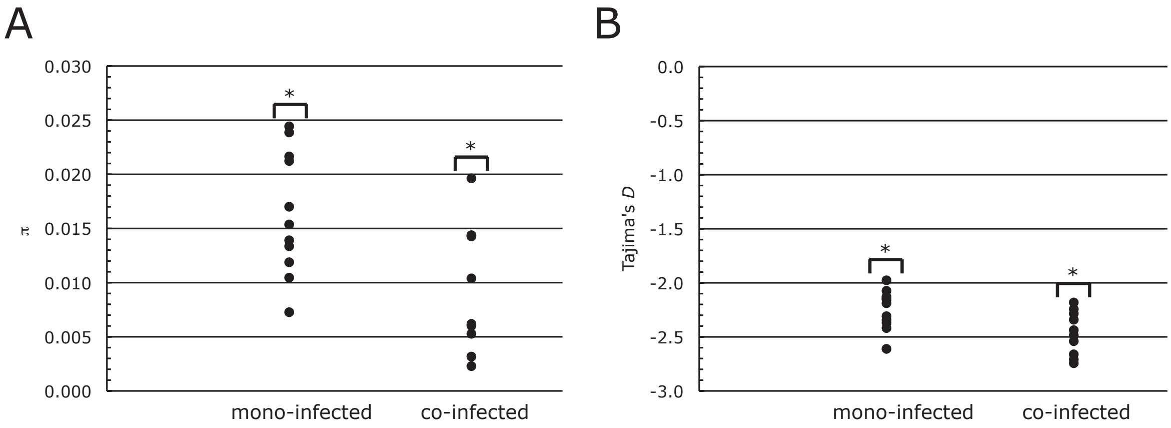


Fig. 4.4 Measures of the average and total population diversity are significantly different between mono-infected and co-infected subjects and indicates population disequilibria. For all mono-infected and co-infected populations, A) pairwise nucleotide diversity and B) Tajima's *D* statistics are reported; asterisks indicate that this difference is significant.

Table 4.5. Frequency of NS3 RAVs in mono-infected and co-infected populations.

		Mono-infected															
			1	2	3	M1	M2	6	7	8	9	10	11	12	13	14	15
V36	A	GCG	0.05														
	G																
	L	TTG															
	M	ATG				0.04	0.04										
Q41	R	CGA						0.18		0.11							
F43	C																
	S	TCC									2.12	0.04					
	I	ATC		0.14													
	V																
T54	A	GCC	0.14					0.18						0.09			
	A	GCT				0.02											
	S	TCT			6.75												
V55	A	GCC	0.05		99.75	0.04	0.11			0.21			0.15			0.10	
	A	GCT			0.25												
Q80	K	AAA		98.00			98.88	0.36		96.62	98.59	97.91	98.97	97.96		99.00	99.48
	K	AAG		1.44			1.01			3.06	1.41	1.86	0.99	1.76		0.80	0.52
	H																
	R	AGA					0.11							0.19			
	R	AGG										0.08					
	R	CGA	0.05														
	R	CGG															
V107	I	ATT						0.18		0.11							0.26
	I	ATC				0.29	0.11					0.04	0.04	0.19		0.10	
R109	K	AAG				0.02	0.02									0.10	
S138	T	ACC															
R155	G	GGG						0.18		0.11							
	I																
	K	AAG				0.02	0.02						0.04				
	L																
	M	ATG															
	S																
	T																
	Q																
	E																
	N																
A156	I																
	S	TCC															
	T																
	V	GTC															
	D	GAC											0.04				
	G																
D168	A																
	E	GAG						0.18									
	E	GAA															
	G																
	H																
	N	AAC															
	V																
	I																
	T																
	Y																
V170	A																
	T	ACC					0.06					0.04	0.04	0.19			0.26

			Co-infected												
			C1	C2	3	4	5	6	7	8	9	10	11	12	13
V36	A	GCG	0.08	0.06		0.13	0.11		0.13		0.22		0.05		
	G														
	L	TTG	0.03												
	M	ATG		0.05									0.11		
Q41	R	CGA											0.05		
F43	C														
	S	TCC		0.02											
	I	ATC		0.02											
	V														
T54	A	GCC													
	A	GCT	0.03	0.05	0.09										
	S	TCT	0.03					0.06							
V55	A	GCC	0.05			0.13									
	A	GCT													
Q80	K	AAA	99.42	99.45		0.13		0.06		99.65			99.40		
	K	AAG	0.44	0.43								0.13	0.44		
	H														
	R	AGA		0.08											
	R	AGG								0.35					
	R	CGA			0.09	0.13	0.23								0.12
	R	CGG										0.13			
V107	I	ATT													0.12
	I	ATC		0.08			0.11		0.26			0.13	0.11		
R109	K	AAG		0.03											
S138	T	ACC			0.09										
R155	G	GGG			2.09				0.13						
	I														
	K	AAG	0.05		0.09		0.11				0.11				0.12
	L														
	M	ATG													0.12
	S														
	T														
	Q														
	E														
	N														
A156	I														
	S	TCC			1.04										
	T														
	V	GTC		0.02										0.08	
	D	GAC													
	G														
D168	A														
	E	GAG													
	E	GAA				0.13									
	G														
	H														
	N	AAC			0.09										0.12
	V														
	I														
	T														
	Y														
V170	A														
	T	ACC	0.08		0.19		0.11								

Some RAVs are significantly more enriched in only the mono-infected or co-infected population.

An excess of low frequency polymorphisms was demonstrated for all populations, however this allelic structure was more pronounced in the co-infected populations. RAVs were observed as low frequency variants in mono-infected and co-infected subjects (Table 4.5), but each subject was sampled to a different depth (Table 4.4) making comparisons between the two groups difficult.

To help determine whether a particular RAV was different across all mono-infected or co-infected individuals within this cohort, a random sampling simulation was conducted. In short, a sub-cohort for either mono-infected or co-infected individuals was randomly generated, and each individual within that sub-cohort was randomly and shallowly sampled to a uniform depth. The sequences from the superficial sampling across individuals within the sub-cohort were then examined for the presence of individual RAVs. This process was repeated 100 times, and the distribution of individual RAVs for mono-infected populations were compared to the distribution of RAVs from co-infected populations. This was compared to a background distribution obtained from comparing the distribution of RAVs from randomly generated populations using identical methods to the mono-infected and co-infected sampling but sampling all individuals together (label blind). Statistical significance was determined between mono-infected and co-infected individuals versus the background sampling using Student's t-test. To try to control for a single individual affecting significance, two individuals with the highest RAV frequencies, one mono-infected and one co-infected, were removed and the simulation was re-run. RAVs were reported only if they continued to be significantly for both simulations.

There were 7 RAVs that were significantly different in frequency between mono-infected and co-infected individuals when compared to label-blind sampling of the entire cohort. NS3 residues 41, 55, 107, 109, and 170 had a higher presence of RAVs in mono-infected versus co-infected individuals, and NS3 residues 36 and 168 had a higher presence of RAVs in co-infected versus mono-infected individuals.

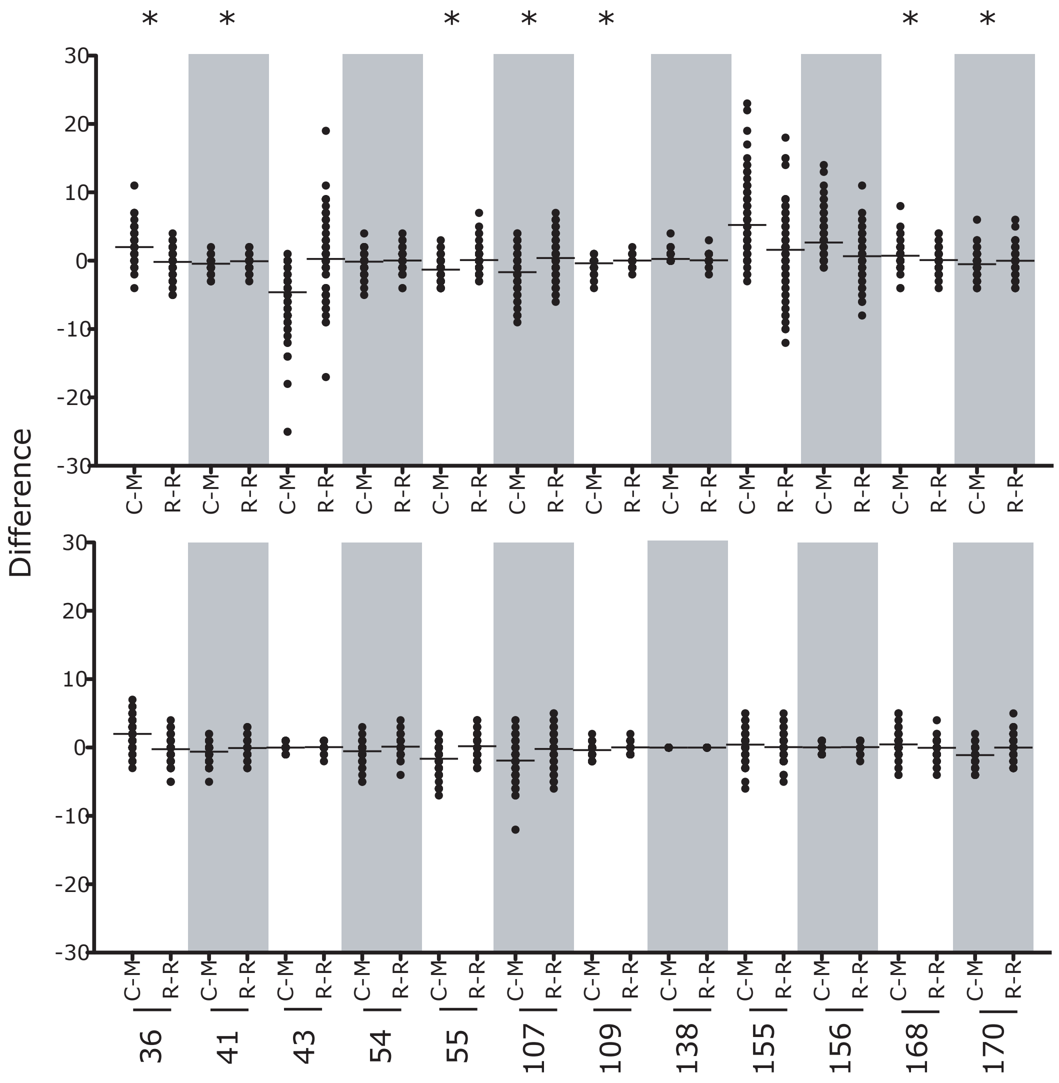


Fig. 4.5 RAVs 36, 41, 55, 107, 109, 168, and 170 have significantly different frequencies in either mono-infected or co-infected individuals. Plotted is the difference in RAV counts between either co-infected versus mono-infected (C-M) subpopulations or random versus random (R-R). Top graph includes all individuals, whereas bottom graph has one individual from each group removed that had the highest RAV frequency. Starred RAVs had significantly different frequencies in co-infected versus mono-infected populations in comparison to the background sampling and retained significance with outlier individuals removed. RAVs 41, 55, 107, 109, and 170 had a higher frequency in mono-infected individuals, and RAVs 36 and 168 were more frequently found in co-infected individuals.

4.5 Discussion

Co-infection is associated with a higher patient morbidity and mortality (65-66), and these clinical markers potentially translate to differences in viral diversity. Because DAAs target key viral enzymes, alleles that decrease or prevent drug binding may permit viral replication during drug selection. It is currently unclear what the steady-state frequency of RAVs are in therapy naïve environments, and whether this variation can be directly selected under a DAA.

We applied an ultra-high resolution deep sequencing approach to examine pre-existing viral diversity across NS3 (residues 36-170) in 15 mono-infected and 13 co-infected individuals. Using Primer IDs, we were able to achieve a mean sequencing depth of 0.1% across the cohort (Table 4.4). We found that viral diversity was higher in mono-infected over co-infected individuals, and this difference was preserved across the entire sequenced length of NS3 (Fig. 4.3-4). We also observed that there were not specific regions of conservation or diversity when comparing across individuals congruent with x-ray crystallographic constructions of NS3 that demonstrate a lack of a clearly defined deep pocket for substrate binding (Fig. 4.3) (8).

Although mono-infected individuals had a significantly higher diversity than co-infected individuals, this did not equate to a significantly higher presence of pre-existing RAVs (Table 4.5). However, when comparing individual RAV positions between the 13 mono-infected individuals that have sequencing encompassing residues 36-170 and the 12 co-infected individuals, there were 7 positions that had a significantly higher frequency in either one or the other group (Fig. 4.5). RAVs 36 and 168 were more likely to be found in co-infected individuals, whereas RAVs 41, 55, 107, 109, and 170 were more likely to be found in mono-infected individuals.

A biological explanation for these differences is not obvious. In this cohort, co-infected individuals are on HAART, and chance HIV drug interactions with HCV may exert a selection pressure that translates to viral diversity. Although all individuals had CD4+ counts >350 cells/mm^3 and co-infected individuals had undetectable HIV viral loads, chronic inflammation due to HIV infection may cause differences in immune function (reviewed in (158)). It may be plausible that HIV infection is still ongoing in a compartment such as the liver (154, 155), but otherwise undetectable in the sera. As HIV can promote liver injury (154, 155), the two viral strains may be in close proximity to one another and driving changes in diversity.

Before RAV differences between mono-infected and co-infected individuals can be assigned with confidence, larger numbers of subjects need to be accurately sequenced to comparable depths. If pre-existing variants can be directly selected under a drug, the biological differences between mono-infected and co-infected individuals may presage variation in virologic responses and possibly altered patterns of resistance to DAAs.

Chapter 5

Concluding remarks

Historically there has been an absence of high-resolution tools that could accurately capture the diversity of a viral population. The structure, dynamics, and evolution of viruses were elucidated through macroscopic snapshots and *in silico* modeling. In this thesis I have presented a new approach that allows significantly greater resolution of the complexity of a viral population, and started to examine basic but critical questions such as what does a population look like and how does a population evolve drug resistance.

5.1 What we can learn from ultra-high resolution of viral populations.

If a viral population were like a city, the 30,000 ft airplane view is the equivalent of Sanger-based sequencing. This single summation would allow one to distinguish the difference between Chicago and Manhattan, for example. Clonal sequencing would start to reveal major buildings and features, such as Willis Tower, Navy Pier, and other large landmarks. Deep sequencing with a Primer ID is like walking the streets of Chicago. A rich array of detail is revealed from this ground level view, and what these details can provide about the nature of the city beyond what can be achieved from a clonal-sequencing viewpoint is something that has yet to be fully realized. Similarly, not every detail will inform, and may even sidetrack from understanding the larger network. Differentiating which details matter will be a major challenge going forward in this increasingly intricate resolution.

RNA viruses dominate emerging infectious diseases, and HIV and hepatitis viruses are the most deadly chronic viral diseases. Clinical diagnostics of an infected individual can inform what the subtype of the infecting pathogen is and whether or what resistance alleles exist. At present, these platforms rely on Sanger-based sequencing technology to assess potential resistance-conferring positions on a population level. Next generation sequencing techniques may be integrated into a clinical diagnostic setting to shift not only to a detailed resolution of allelic variants, but also retain linkage across positions.

As a field, we currently have a poor understanding of the steady-state frequencies of minor variants, and how they change before and after a selective event. The standing genetic variation constantly changes due to an erroneous viral polymerase. Due to transitional biases, we know that certain resistance alleles are more likely to be created *de novo* compared to others, but their associated fitness effects keeps them at a mutation-selection balance dictated by the sum of the fitness cost on the residing haplotype.

For all populations sequenced, there was an excess of low frequency polymorphisms over that which would be expected from a population at equilibrium (Table 2.2, 3.3, Fig. 3.6, 4.4). This haplotype structure was universal across HCV and HIV. For HIV populations exposed to a drug selection pressure, pre-existing, rebound resistant, and rebound susceptible populations also contained this structure (Fig. 3.6). This observation further emphasizes the ability of a population to quickly regain diversity after a selective event.

We also wanted to examine if a co-infecting pathogen affects population diversity. From a small cohort of HCV mono-infected and HCV-HIV co-infected individuals, we observed that there is higher diversity in mono-infected individuals (Fig. 4.3-4). This diversity did not correlate to an increase in the overall presence of RAVs in mono-infected individuals (Table 4.5), but there was a statistically significant difference in the pre-existence of handful of alleles between these two groups (Fig. 4.5). Further sequencing work is needed to determine whether this difference holds, and if it does, for what alleles. Furthermore, a biological or clinical link would also be needed. However, if our preliminary observations are biologically sound, these differences may be reflected in DAA response.

Aside from ascertaining what resistant variants at what frequency are more likely to pre-exist, it is currently unknown whether they can be directly selected under a drug selective pressure. We have shown from a single HIV-infected individual that clonally amplified resistant alleles were identical to pre-existing variants (Fig. 2.8). Subsequent studies have indicated that the emerging resistant haplotype was not sampled prior to treatments or that the resistance mutants arose during the selection (Fig. 3.9). However, the fact that multiple haplotypes grew out with resistance mutations, and these mutations were not on the dominant haplotypes in the population, indicates that the resistance mutations that did grow out pre-existed in the population. Further work is need to determine if there is a relationship between resistance in the standing genetic variation and what grows out under a selective event.

We observed that the path to resistance can be highly variable and unique per individual. For a handful of HIV infected individuals longitudinally followed, dominant haplotypes that emerged over time were largely different than the preceding haplotypes (Fig. 3.9). Furthermore, they could not be explained by recombination across prior haplotypes, implying that the path to resistance is by *de novo* mutations. We observed that measuring the selective pressure at the time of sampling was inconsistent to population composition (Fig.

3.1, Table 3.3, 3.9). For example, low drug levels were measured for populations containing multiple resistance alleles linked on a common haplotype, and high drug levels were observed for populations containing a large percentage of susceptible variants. This observation has several clinically significant implications. The incongruence between drug levels and population composition could be due to a lag in population response. Furthermore, it could illustrate intermittent adherence issues, as excessive drug levels could be due to dosing directly prior to sampling. Variable drug levels over time could be linked to the emergence of multiple resistance alleles on distinct haplotypes.

If pre-existing resistance alleles are not consistently informative towards how a population will respond to a drug selective pressure, the role of minor variants in predicting population response may be better suited in looking at resistance decay. After removal of a selective agent from a population that is resistant to it, resistance alleles, once having a fitness advantage, will decrease in frequency if that advantage is lost. The rate at which they decay may be dependent on the particular alleles present and the haplotype on which they reside, but decay could be informative in determining whether and which drugs may achieve a sustained virologic response in re-treatment.

Deep sequencing with a Primer ID resulted in the first accurate ultra-high resolution of a single gene within a viral population (Fig. 2.8, Table 2.1). The challenge going forward is determining which details revealed from this technique are clinically significant. We are increasingly integrating personalized therapeutics with clinical choices. Personalized therapeutics has not entered infectious disease diagnostics aside from assaying major drug resistance mutations by Sanger-based population sequencing. This is in stark contrast to achievements in antiviral drug development. For example, all-DAA approaches are increasingly being used for the treatment of HCV, however this has been without comparable achievements in resistance diagnostics. As a field, we are assessing pre-existing alleles using a technique invented in 1974 and that misses allelic variants present at less than 25% abundance. Alternative, more sensitive approaches have not been adopted into the clinical setting mostly due to the high labor and low throughput.

5.2 The idealized future of next generation sequencing.

Future sequencing platforms may make Primer IDs unnecessary for the deep sequencing of viral populations, however several major technological obstacles would have to be overcome. First, the initial sample concentration could permit the sequencing of low copy samples. Second, sequencing would have to be done off of native vRNA. Third, sequencing length would have to encompass the entire genome. Fourth, the error profile would have to be as close to zero as possible. Fifth, throughput would have to be high enough to resolve minor variants within a population at a depth determined to contain variants that may inform population resolution in a cost and time effective manner.

Not mandating a high sample concentration prior to a sequencing protocol would allow one to skip preceding PCR and all of the amplification biases, recombination, and re-sampling that is introduced from this erroneous technique. Currently, 500ng of dsDNA go into library preparation for the 454 platform. Even if 454 had the capacity to sequence entire genomes, the initial input of starting material would always necessitate preceding PCR of clinical samples. High throughput platforms ideally would be dynamic enough to handle all disease states; clinical samples that contain low or undetectable viral loads could be sequenced through samples of viologic failure.

Allowing the sequencing of vRNA directly would avoid having input material that has to be molecularly adapted to the particular platform through cDNA synthesis and PCR, steps that are now needed since the addition of adaptors to the ends of the DNA sequencing target typically require a concentration of DNA larger than that found in clinical samples and also requires that the starting material is double-stranded. PacBio can sequence DNA-RNA heteroduplex material. Even if the platform could accurately sequence whole viral genomes and one could add their bell adaptors to small amounts of starting material, cDNA synthesis would still be required, however error between complementary bases introduced during cDNA synthesis could be differentiated based on sense.

A needed improvement in sequencing technologies is increasing sequencing length, ideally to where the reads are long enough to encompass the entire viral genome. Internal genes of interest could be reached along with the preservation of linkage across multiple genes. The linkage of distal genes is particularly critical for resistance surveillance of viral populations simultaneously counteracted by antiviral drugs that have different gene targets. Oxford Nanopore is the only platform that will be available the near future that appears to have the capacity to consistently achieve whole genome viral sequencing but at a high error rate (discussed below). However, there are a number of other disadvantages that trump the long sequencing length when applied to viral populations that preclude Nanopore's immediate adaptation to vRNA.

The largest technological hurdle that needs to be overcome is decreasing the error to negligible values. This is particularly problematic with third-generation real-time sequencing, where high in/del rates dominant the error profile compared to the high misincorporation/misread error of the second-generation approaches. Although the partitioning of generations is subjective, NextGen platforms can be divided into two different concepts. The earlier machines rely on a controlled exposure of nucleotides to a sequencing target and the report of a signal that indicates what had been incorporated, allowing for a base call. 454 and Illumina rely on light, whereas Ion Torrent gauges pH change. Newer machines rely on real-time reporting. PacBio employs a tethered polymerases and a laser to capture the identity of the fluorescently labeled nucleotide as it is incorporated. Nanopore's pores sense the chemical composition of individual nucleotides as a strand of genomic material is threaded through it. Both PacBio and Nanopore have the ability to sequence a vast amount of genomic material in a very small amount of time due to the real-time reporting. However, they are both prone to very high levels of random in/del calls. For example, due to the non-metronome nature of polymerase nucleotide incorporation, small deviations in synthesis tempo cannot be easily assigned as a true deletion or pause during PacBio base calling. Similarly, a nucleotide captured by the laser but that does not get incorporated will be erroneously called as an insertion.

When sequencing highly heterogeneous viral populations, randomly distributed in/del errors are much more problematic to resolve than misincorporation and homopolymeric errors. In/dels change the reading frame, and the identification of them likely requires indirect inference to what the true sequence should be. For platforms like PacBio, whose in/del rate approaches 15-20%, error alone in the Primer ID and barcoding region coupled with an extremely low throughput does not allow for the two techniques to be easily merged. The circular consensus sequencing of shorter input material allows a higher coverage per base, but at the trade-off of decreased length. Although this technique in principle should be able to decrease the in/del rate resolved below that of 15-20%, the current bioinformatics pipeline native to the machine cannot accurately create circular consensus sequences that would allow for subsequent Primer ID consensus formation. Even if the bioinformatics were corrected, the decreased length required for circular consensus sequencing in addition to the inherent low-throughput of PacBio does not easily translate this platform to either viral sequencing or targeted viral re-sequencing. Although Nanopore could theoretically sequence vRNA, sequencing is single-pass, thus the in/del profile would remain uncorrected much like PacBio unless multiple passes could be achieved on the same molecule or if a proofreading mechanism were incorporated into the pore.

When sequencing a viral population, depth, thus throughput will be dictated by the frequency of minor variants found to have biological significance. For example, the frequency of a minor allele of interest in the standing genetic variation is determined by the mutation selection balance. Sequencing depth will be dictated by that equilibrium frequency and the probability of sampling it. Once dominant minor variants may still be present after the removal of a selection pressure, and their frequency may influence downstream population response. For example, if a population escapes a drug selection pressure with resistance, and the drug is removed, a critical frequency may be needed for rebound if the same selective pressure is re-applied.

If vRNA could be directly and accurately sequenced, 3,000 individual reads would be needed to resolve variants present at 0.1%. The number of sequencing reactions per sample would have to be multiplied out by the negative or failed rate inherent in the system to determine the initial throughput needed. Ideally, multiple samples could be sequenced in parallel. If not physically partitioned, molecular labeling of vRNA would be required, involving a high fidelity technique that would not introduce error such as single-stranded ligation.

Primer IDs provide a direct means of correcting for PCR biases, PCR re-sampling, and sequencing error when preceding PCR is required prior to a sequencing protocol (143). There are a number of technological advances required to remove PCR prior to sequencing. Even if native vRNA could be sequenced from clinical samples, if the sequencing error rate is higher than that from early-introduced residual error from PCR, PCR may still be required to identify and remove it. Next-generation platforms need to either report multiple reads from a single molecule, or encapsulate a proofreading ability to make base calling as high fidelity as possible.

Applying Primer IDs to clinical diagnostics.

Critical advances of high throughput sequencing platforms are needed to make Primer IDs unnecessary (detailed above). If one were to integrate deep sequencing with Primer IDs for clinical diagnostics, several requirements would need to be met. The next generation platform used would have to have a low in/del error rate profile and ideally lack of a homopolymer miscall bias. Although many homopolymers do not interfere with resistance conferring positions, some do, such as HIV's RT resistance K65R and K103N. Currently, only Illumina's sequencing technology can reliably call bases across homopolymeric regions, which gives this platform a very low in/del rate, though sequencing length would still require targeted re-sequencing and the loss of linkage across distal genes.

Two critical conditions must be met for Primer IDs to accurately be used for resolving viral diversity. First, they need to be under-sampled. Second, they have to randomly tag. The Primer ID was intentionally designed to be 8 nucleotides in length. 4^8 produced 65,536 combinations, whereas a 7mer would only produce 16,384 combinations. By *in silico* random-draw sampling simulations, a 8mer allowed for 10k templates of input material to be tagged with only ~5% of the population having unique templates tagged by the same random combination assuming 100% tagging efficiency. Under-sampling 4^7 would have resulted in too little input material to be useful. A 9mer would allow for an even greater number of combinations, and degenerate nucleotides much longer than this have been applied in eukaryotic systems for molecular tagging.

The reason why the Primer ID is 8 nucleotides in length is to control for the randomness of tagging. vRNA can be highly structured. Adding a degenerate region to a tagging primer can also create structure within the primer. If particular Primer IDs gain a fitness advantage in tagging due to secondary structure effects, that combination becomes over-represented and results in a net loss of complexity in the Primer ID sequence library in addition to fewer consensus sequences representing the different starting templates.

For Primer IDs to be used in clinical diagnostics, reproducibility is critical, and directly correlates to randomness of tagging. For 2 mono-infected and 2 co-infected individuals, the same HCV positive serum samples were tagged, amplified, and sequenced in replicate. We demonstrated that not a single Primer ID was oversampled, and different Primer IDs were dominant across the different runs (Table 4.2). These two observations indicate that not a single Primer ID had a greater fitness in tagging. Slightly higher cDNA synthesis temperatures resulted in more reads being built per Primer ID in mono-infected, but not co-infected individuals (Fig. 4.1*A,D,G,J*). More reads per Primer ID also correlated with a decrease in ambiguous calls (Fig. 4.1). In comparing all haplotypes present ≥1% in frequency between the first and second replicate, there was a correlation coefficient of 0.8 (Fig. 4.2). This indicates that independent tagging, amplification, and sequencing of a given sample is reproducible.

HCV RAVs can be extremely low in frequency, emulating the expected error rate. When RAVs were observed in one run but not the other (Table 4.3), this could be due to the error or sampling. To determine if a RAV is biological when it is only represented by a single occurrence, one could always follow-up with allele specific PCR.

To integrate Primer IDs into clinical diagnostics, randomness and reproducibility are critical. Unlike eukaryotic systems, vRNA can be highly structured, and degenerate regions within a tagging primer can fold to create non-random interactions with the viral genome. In our work, we employed several techniques to encourage random interaction to a large degree of success. To use Primer IDs in clinical diagnostics, tagging would have to be further optimized, particularly for highly structured viral genomes. Although sample-specific primers encourage high fidelity targeting and hot cDNA synthesis reactions decrease structure, linearization of vRNA and tagging primers would likely be needed to truly remove the potential for secondary structure targeting. Only after consistent demonstrations of reproducibility and randomness across genes, genomes, and individuals can Primer IDs be used to reveal minor variants within a clinical diagnostics setting.

Literature Cited

1. Margeridon-Thermet S & Shafer RW (2010) Comparison of the Mechanisms of Drug Resistance among HIV, Hepatitis B, and Hepatitis C. *Viruses* 2(12):2696-2739.

2. Kenny-Walsh E (1999) Clinical outcomes after hepatitis C infection from contaminated anti-D immune globulin. Irish Hepatology Research Group. *N Engl J Med* 340(16):1228-1233.

3. Seeff LB*, et al.* (2001) Long-term mortality and morbidity of transfusion-associated non-A, non-B, and type C hepatitis: A National Heart, Lung, and Blood Institute collaborative study. *Hepatology* 33(2):455-463.

4. Thomas DL*, et al.* (2000) The natural history of hepatitis C virus infection: host, viral, and environmental factors. *JAMA* 284(4):450-456.

5. Lemon SM (2010) Induction and evasion of innate antiviral responses by hepatitis C virus. *J Biol Chem* 285(30):22741-22747.

6. Thompson MA*, et al.* (2010) Antiretroviral treatment of adult HIV infection: 2010 recommendations of the International AIDS Society-USA panel. *JAMA* 304(3):321-333.

7. Bartenschlager R (1999) The NS3/4A proteinase of the hepatitis C virus: unravelling structure and function of an unusual enzyme and a prime target for antiviral therapy. *J Viral Hepat* 6(3):165-181.

8. Kim JL*, et al.* (1996) Crystal structure of the hepatitis C virus NS3 protease domain complexed with a synthetic NS4A cofactor peptide. *Cell* 87(2):343-355.

9. Abram ME, Ferris AL, Shao W, Alvord WG, & Hughes SH (2010) Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. *J Virol* 84(19):9864-9878.

10. Mansky LM (1996) Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Res Hum Retroviruses* 12(4):307-314.

11. Steinhauer DA, Domingo E, & Holland JJ (1992) Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene* 122(2):281-288.

12. Coffin JM (1979) Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses. *J Gen Virol* 42(1):1-26.

13. Rhodes T, Wargo H, & Hu WS (2003) High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication. *J Virol* 77(20):11193-11200.

14. Xu H & Boeke JD (1987) High-frequency deletion between homologous sequences during retrotransposition of Ty elements in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 84(23):8553-8557.

15. Neher RA & Leitner T (2010) Recombination rate and selection strength in HIV intra-patient evolution. *PLoS Comput Biol* 6(1):e1000660.

16. Shriner D, Rodrigo AG, Nickle DC, & Mullins JI (2004) Pervasive genomic recombination of HIV-1 in vivo. *Genetics* 167(4):1573-1583.

17. Colina R*, et al.* (2004) Evidence of intratypic recombination in natural populations of hepatitis C virus. *J Gen Virol* 85(Pt 1):31-37.

18. Moreau I*, et al.* (2006) Serendipitous identification of natural intergenotypic recombinants of hepatitis C in Ireland. *Virol J* 3:95.

19. Kalinina O, Norder H, Mukomolov S, & Magnius LO (2002) A natural intergenotypic recombinant of hepatitis C virus identified in St. Petersburg. *J Virol* 76(8):4034-4043.

20. Gao L, Aizaki H, He JW, & Lai MM (2004) Interactions between viral nonstructural proteins and host protein hVAP-33 mediate the formation of hepatitis C virus RNA replication complex on lipid raft. *J Virol* 78(7):3480-3488.

21. Gates AT, Sarisky RT, & Gu B (2004) Sequence requirements for the development of a chimeric HCV replicon system. *Virus Res* 100(2):213-222.

22. Simmonds P (2004) Genetic diversity and evolution of hepatitis C virus--15 years on. *J Gen Virol* 85(Pt 11):3173-3188.

23. Le SY, Chen JH, Braun MJ, Gonda MA, & Maizel JV (1988) Stability of RNA stem-loop structure and distribution of non-random structure in the human immunodeficiency virus (HIV-I). *Nucleic Acids Res* 16(11):5153-5168.

24. Le SY, Chen JH, Chatterjee D, & Maizel JV (1989) Sequence divergence and open regions of RNA secondary structures in the envelope regions of the 17 human immunodeficiency virus isolates. *Nucleic Acids Res* 17(8):3275-3288.

25. Schinazi RF, Lloyd RM, Jr., Ramanathan CS, & Taylor EW (1994) Antiviral drug resistance mutations in human immunodeficiency virus type 1 reverse transcriptase

occur in specific RNA structural regions. *Antimicrob Agents Chemother* 38(2):268-274.

26. Yoshida K, Nakamura M, & Ohno T (1997) Mutations of the HIV type 1 V3 loop under selection pressure with neutralizing monoclonal antibody NM-01. *AIDS Res Hum Retroviruses* 13(15):1283-1290.

27. Simmonds P & Smith DB (1999) Structural constraints on RNA virus evolution. *J Virol* 73(7):5787-5794.

28. Knies JL*, et al.* (2008) Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. *Mol Biol Evol* 25(8):1778-1787.

29. Powdrill MH*, et al.* (2011) Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance. *Proc Natl Acad Sci U S A* 108(51):20509-20513.

30. Haldane JBS (1927) A mathematical theory of natural and artificial selection, Part V: Selection and mutation. *Proceedings of the Cambridge Philosophical Society* 23:838-844.

31. Perelson AS, Neumann AU, Markowitz M, Leonard JM, & Ho DD (1996) HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* 271(5255):1582-1586.

32. Neumann AU*, et al.* (1998) Hepatitis C viral dynamics in vivo and the antiviral efficacy of interferon-alpha therapy. *Science* 282(5386):103-107.

33. Ramratnam B*, et al.* (1999) Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *Lancet* 354(9192):1782-1785.

34. Herrmann E, Neumann AU, Schmidt JM, & Zeuzem S (2000) Hepatitis C virus kinetics. *Antivir Ther* 5(2):85-90.

35. Rong L, Dahari H, Ribeiro RM, & Perelson AS (2010) Rapid emergence of protease inhibitor resistance in hepatitis C virus. *Sci Transl Med* 2(30):30ra32.

36. Coffin JM (1995) HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* 267(5197):483-489.

37. Ho DD (1996) Viral counts count in HIV infection. *Science* 272(5265):1124-1125.

38. Van Laethem K*, et al.* (1999) Phenotypic assays and sequencing are less sensitive than point mutation assays for detection of resistance in mixed HIV-1 genotypic populations. *J Acquir Immune Defic Syndr* 22(2):107-118.

39. Gunthard HF, Wong JK, Ignacio CC, Havlir DV, & Richman DD (1998) Comparative performance of high-density oligonucleotide sequencing and dideoxynucleotide sequencing of HIV type 1 pol from clinical samples. *AIDS Res Hum Retroviruses* 14(10):869-876.

40. Kuritzkes DR*, et al.* (2008) Preexisting resistance to nonnucleoside reverse-transcriptase inhibitors predicts virologic failure of an efavirenz-based regimen in treatment-naive HIV-1-infected subjects. *J Infect Dis* 197(6):867-870.

41. Paredes R*, et al.* (2010) Pre-existing minority drug-resistant HIV-1 variants, adherence, and risk of antiretroviral treatment failure. *J Infect Dis* 201(5):662-671.

42. Lehman DA*, et al.* (2012) Low-frequency nevirapine resistance at multiple sites may predict treatment failure in infants on nevirapine-based treatment. *J Acquir Immune Defic Syndr* 60(3):225-233.

43. Codoner FM*, et al.* (2011) Added value of deep sequencing relative to population sequencing in heavily pre-treated HIV-1-infected subjects. *PLoS One* 6(5):e19461.

44. Messiaen P*, et al.* (2012) Ultra-deep sequencing of HIV-1 reverse transcriptase before start of an NNRTI-based regimen in treatment-naive patients. *Virology* 426(1):7-11.

45. Molla A*, et al.* (1996) Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nat Med* 2(7):760-766.

46. Eastman PS*, et al.* (1998) Genotypic changes in human immunodeficiency virus type 1 associated with loss of suppression of plasma viral RNA levels in subjects treated with ritonavir (Norvir) monotherapy. *J Virol* 72(6):5154-5164.

47. Zhang YM*, et al.* (1997) Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites. *J Virol* 71(9):6662-6670.

48. Resch W, Parkin N, Watkins T, Harris J, & Swanstrom R (2005) Evolution of human immunodeficiency virus type 1 protease genotypes and phenotypes in vivo under selective pressure of the protease inhibitor ritonavir. *J Virol* 79(16):10638-10649.

49. Burch CL & Chao L (2004) Epistasis and its relationship to canalization in the RNA virus phi 6. *Genetics* 167(2):559-567.

50. Gerrish PJ & Lenski RE (1998) The fate of competing beneficial mutations in an asexual population. *Genetica* 102-103(1-6):127-144.

51. Miralles R, Gerrish PJ, Moya A, & Elena SF (1999) Clonal interference and the evolution of RNA viruses. *Science* 285(5434):1745-1747.

52. Clarke DK*, et al.* (1994) The red queen reigns in the kingdom of RNA viruses. *Proc Natl Acad Sci U S A* 91(11):4821-4824.

53. Quer J*, et al.* (1996) Reproducible nonlinear population dynamics and critical points during replicative competitions of RNA virus quasispecies. *J Mol Biol* 264(3):465-471.

54. Rouzine IM & Coffin JM (1999) Linkage disequilibrium test implies a large effective population number for HIV in vivo. *Proc Natl Acad Sci U S A* 96(19):10758-10763.

55. Shriner D, Liu Y, Nickle DC, & Mullins JI (2006) Evolution of intrahost HIV-1 genetic diversity during chronic infection. *Evolution* 60(6):1165-1176.

56. Shriner D*, et al.* (2004) Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. *Genetics* 166(3):1155-1164.

57. Brown AJ (1997) Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. *Proc Natl Acad Sci U S A* 94(5):1862-1865.

58. Kimura M (1983) *The neutral theory of molecular evolution* (Cambridge University Press, Cambridge).

59. Otto SP & Lenormand T (2002) Resolving the paradox of sex and recombination. *Nat Rev Genet* 3(4):252-261.

60. Burch CL & Chao L (1999) Evolution by small steps and rugged landscapes in the RNA virus phi6. *Genetics* 151(3):921-927.

61. Connell JH (1978) Diversity in tropical rain forests and coral reefs. *Science* 199(4335):1302-1310.

62. Kepler TB & Perelson AS (1998) Drug concentration heterogeneity facilitates the evolution of drug resistance. *Proc Natl Acad Sci U S A* 95(20):11514-11519.

63. Christensen NO, Nansen P, Fagbemi BO, & Monrad J (1987) Heterologous antagonistic and synergistic interactions between helminths and between helminths and protozoans in concurrent experimental infection of mammalian hosts. *Parasitol Res* 73(5):387-410.

64. Johnson PT & Hoverman JT (2012) Parasite diversity and coinfection determine pathogen infection success and host fitness. *Proc Natl Acad Sci U S A* 109(23):9006-9011.

65. Joshi D, O'Grady J, Dieterich D, Gazzard B, & Agarwal K (2011) Increasing burden of liver disease in patients with HIV infection. *Lancet* 377(9772):1198-1209.

66. Thein HH, Yi Q, Dore GJ, & Krahn MD (2008) Natural history of hepatitis C virus infection in HIV-infected individuals and the impact of HIV in the era of highly active antiretroviral therapy: a meta-analysis. *AIDS* 22(15):1979-1991.

67. Winters MA, Chary A, Eison R, Asmuth D, & Holodniy M (2010) Impact of highly active antiretroviral therapy on hepatitis C virus protease quasispecies diversity in HIV co-infected patients. *J Med Virol* 82(5):791-798.

68. Netski DM, Mao Q, Ray SC, & Klein RS (2008) Genetic divergence of hepatitis C virus: the role of HIV-related immunosuppression. *J Acquir Immune Defic Syndr* 49(2):136-141.

69. Lopez-Labrador FX*, et al.* (2007) Trends for genetic variation of Hepatitis C Virus quasispecies in Human Immunodeficiency virus-1 coinfected patients. *Virus Res* 130(1-2):285-291.

70. Sherman KE, Andreatta C, O'Brien J, Gutierrez A, & Harris R (1996) Hepatitis C in human immunodeficiency virus-coinfected patients: increased variability in the hypervariable envelope coding domain. *Hepatology* 23(4):688-694.

71. Shuhart MC*, et al.* (2006) HIV infection and antiretroviral therapy: effect on hepatitis C virus quasispecies variability. *J Infect Dis* 193(9):1211-1218.

72. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, & Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* 17(8):1195-1201.

73. Hirsch MS*, et al.* (2008) Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. *Clin Infect Dis* 47(2):266-285.

74. Palmer S*, et al.* (2006) Persistence of nevirapine-resistant HIV-1 in women after single-dose nevirapine therapy for prevention of maternal-to-fetal HIV-1 transmission. *Proc Natl Acad Sci U S A* 103(18):7094-7099.

75. Flys TS*, et al.* (2006) Quantitative analysis of HIV-1 variants with the K103N resistance mutation after single-dose nevirapine in women with HIV-1 subtypes A, C, and D. *J Acquir Immune Defic Syndr* 42(5):610-613.

76. Cai F*, et al.* (2007) Detection of minor drug-resistant populations by parallel allele-specific sequencing. *Nat Methods* 4(2):123-125.

77. Beck IA*, et al.* (2008) Optimization of the oligonucleotide ligation assay, a rapid and inexpensive test for detection of HIV-1 drug resistance mutations, for non-North American variants. *J Acquir Immune Defic Syndr* 48(4):418-427.

78. Johnson JA*, et al.* (2007) Simple PCR assays improve the sensitivity of HIV-1 subtype B drug resistance testing and allow linking of resistance mutations. *PLoS One* 2(7):e638.

79. Johnson JA*, et al.* (2008) Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy. *PLoS Med* 5(7):e158.

80. Metzner KJ*, et al.* (2003) Emergence of minor populations of human immunodeficiency virus type 1 carrying the M184V and L90M mutations in subjects undergoing structured treatment interruptions. *J Infect Dis* 188(10):1433-1443.

81. Metzner KJ*, et al.* (2009) Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naive and -adherent patients. *Clin Infect Dis* 48(2):239-247.

82. Paredes R, Marconi VC, Campbell TB, & Kuritzkes DR (2007) Systematic evaluation of allele-specific real-time PCR for the detection of minor HIV-1 variants with pol and env resistance mutations. *J Virol Methods* 146(1-2):136-146.

83. Li JZ*, et al.* (2011) Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. *JAMA* 305(13):1327-1335.

84. Meyerhans A, Vartanian JP, & Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Res* 18(7):1687-1691.

85. Yang YL, Wang G, Dorman K, & Kaplan AH (1996) Long polymerase chain reaction amplification of heterogeneous HIV type 1 templates produces recombination at a relatively high frequency. *AIDS Res Hum Retroviruses* 12(4):303-306.

86. Liu SL*, et al.* (1996) HIV quasispecies and resampling. *Science* 273(5274):415-416.

87. Salazar-Gonzalez JF*, et al.* (2008) Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol* 82(8):3952-3970.

88. Palmer S*, et al.* (2005) Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* 43(1):406-413.

89. Simmonds P, Balfe P, Ludlam CA, Bishop JO, & Brown AJ (1990) Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *J Virol* 64(12):5840-5850.

90. Edmonson PF & Mullins JI (1992) Efficient amplification of HIV half-genomes from tissue DNA. *Nucleic Acids Res* 20(18):4933.

91. Zagordi O, Geyrhofer L, Roth V, & Beerenwinkel N (2010) Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol* 17(3):417-428.

92. Zagordi O, Klein R, Daumer M, & Beerenwinkel N (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* 38(21):7400-7409.

93. Beerenwinkel N & Zagordi O (2011) Ultra-deep sequencing for the analysis of viral populations. *Curr Opin Virol* 1(5):413-418.

94. Zagordi O, Bhattacharya A, Eriksson N, & Beerenwinkel N (2011) ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12:119.

95. Margulies M*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437(7057):376-380.

96. Eid J*, et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910):133-138.

97. Bentley DR*, et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456(7218):53-59.

98. Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11(1):31-46.

99. Fischer W*, et al.* (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One* 5(8):e12303.

100. Hoffmann C*, et al.* (2007) DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 35(13):e91.

101. Bushman FD*, et al.* (2008) Massively parallel pyrosequencing in HIV research. *AIDS* 22(12):1411-1415.

102. Varghese V*, et al.* (2009) Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. *J Acquir Immune Defic Syndr* 52(3):309-315.

103. Mitsuya Y*, et al.* (2008) Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations. *J Virol* 82(21):10747-10755.

104. Halvas EK*, et al.* (2006) Blinded, multicenter comparison of methods to detect a drug-resistant mutant of human immunodeficiency virus type 1 at low frequency. *J Clin Microbiol* 44(7):2612-2614.

105. Mansky LM & Temin HM (1995) Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 69(8):5087-5094.

106. Hughes JP & Totten P (2003) Estimating the accuracy of polymerase chain reaction-based tests using endpoint dilution. *Biometrics* 59(3):505-511.

107. Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J Biosci Bioeng* 96(4):317-323.

108. Judo MS, Wedel AB, & Wilson C (1998) Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Res* 26(7):1819-1825.

109. Abrahams MR*, et al.* (2009) Quantitating the multiplicity of infection with human immunodeficiency virus type 1 subtype C reveals a non-poisson distribution of transmitted variants. *J Virol* 83(8):3556-3567.

110. Keele BF*, et al.* (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* 105(21):7552-7557.

111. Librado P & Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25(11):1451-1452.

112. Paradis E, Claude J, & Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2):289-290.

113. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585-595.

114. Howe K, Bateman A, & Durbin R (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 18(11):1546-1547.

115. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.69. (Department of Genome Sciences, University of Washington, Seattle, WA).

116. Drummond AJ & Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.

117. Cameron DW*, et al.* (1998) Randomised placebo-controlled trial of ritonavir in advanced HIV-1 disease. The Advanced HIV Disease Ritonavir Study Group. *Lancet* 351(9102):543-549.

118. Potter J, Zheng W, & Lee J (2003) Thermal stability and cDNA synthesis capability of SuperScript™ III reverse transcriptase. in *Focus* (Invitrogen Corporation, Carlsbad, CA), p 27.

119. Barnes WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene* 112(1):29-35.

120. Johnson VA*, et al.* (2005) Update of the drug resistance mutations in HIV-1: Fall 2005. *Top HIV Med* 13(4):125-131.

121. Shafer RW, Jung DR, & Betts BJ (2000) Human immunodeficiency virus type 1 reverse transcriptase and protease mutation search engine for queries. *Nat Med* 6(11):1290-1292.

122. Carrillo A*, et al.* (1998) In vitro selection and characterization of human immunodeficiency virus type 1 variants with increased resistance to ABT-378, a novel protease inhibitor. *J Virol* 72(9):7532-7541.

123. Drake JW & Holland JJ (1999) Mutation rates among RNA viruses. *Proc Natl Acad Sci U S A* 96(24):13910-13913.

124. Duffy S, Shackelton LA, & Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* 9(4):267-276.

125. Onafuwa-Nuga A & Telesnitsky A (2009) The remarkable frequency of human immunodeficiency virus type 1 genetic recombination. *Microbiol Mol Biol Rev* 73(3):451-480, Table of Contents.

126. Shafer RW (2009) Low-abundance drug-resistant HIV-1 variants: finding significance in an era of abundant diagnostic and therapeutic options. *J Infect Dis* 199(5):610-612.

127. Eriksson N*, et al.* (2008) Viral population estimation using pyrosequencing. *PLoS Comput Biol* 4(4):e1000074.

128. Cardinaud S*, et al.* (2004) Identification of cryptic MHC I-restricted epitopes encoded by HIV-1 alternative reading frames. *J Exp Med* 199(8):1053-1063.

129. Bansal A*, et al.* (2010) CD8 T cell response and evolutionary pressure to HIV-1 cryptic epitopes derived from antisense transcription. *J Exp Med* 207(1):51-59.

130. Berger CT*, et al.* (2010) Viral adaptation to immune selection pressure by HLA class I-restricted CTL responses targeting epitopes in HIV frameshift sequences. *J Exp Med* 207(1):61-75.

131. Hance AJ*, et al.* (2001) Changes in human immunodeficiency virus type 1 populations after treatment interruption in patients failing antiretroviral therapy. *J Virol* 75(14):6410-6417.

132. Wright S (1931) Evolution in Mendelian Populations. *Genetics* 16(2):97-159.

133. Rouzine IM, Rodrigo A, & Coffin JM (2001) Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. *Microbiol Mol Biol Rev* 65(1):151-185.

134. Bonhoeffer S, May RM, Shaw GM, & Nowak MA (1997) Virus dynamics and drug therapy. *Proc Natl Acad Sci U S A* 94(13):6971-6976.

135. Bonhoeffer S & Nowak MA (1997) Pre-existence and emergence of drug resistance in HIV-1 infection. *Proc Biol Sci* 264(1382):631-637.

136. Ribeiro RM & Bonhoeffer S (2000) Production of resistant HIV mutants during antiretroviral therapy. *Proc Natl Acad Sci U S A* 97(14):7681-7686.

137. Ribeiro RM, Bonhoeffer S, & Nowak MA (1998) The frequency of resistant mutant virus before antiviral therapy. *AIDS* 12(5):461-465.

138. Althaus CL & Bonhoeffer S (2005) Stochastic interplay between mutation and recombination during the acquisition of drug resistance mutations in human immunodeficiency virus type 1. *J Virol* 79(21):13572-13578.

139. Seo TK, Thorne JL, Hasegawa M, & Kishino H (2002) Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. *Genetics* 160(4):1283-1293.

140. Achaz G*, et al.* (2004) A robust measure of HIV-1 population turnover within chronically infected individuals. *Mol Biol Evol* 21(10):1902-1912.

141. Nijhuis M*, et al.* (1998) Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. *Proc Natl Acad Sci U S A* 95(24):14441-14446.

142. Rodrigo AG*, et al.* (1999) Coalescent estimates of HIV-1 generation time in vivo. *Proc Natl Acad Sci U S A* 96(5):2187-2191.

143. Jabara CB, Jones CD, Roach J, Anderson JA, & Swanstrom R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc Natl Acad Sci U S A* 108(50):20166-20171.

144. Hoffman NG, Schiffer CA, & Swanstrom R (2003) Covariation of amino acid positions in HIV-1 protease. *Virology* 314(2):536-548.

145. Fried MW*, et al.* (2002) Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection. *N Engl J Med* 347(13):975-982.

146. Manns MP*, et al.* (2001) Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for initial treatment of chronic hepatitis C: a randomised trial. *Lancet* 358(9286):958-965.

147. Kolykhalov AA, Mihalik K, Feinstone SM, & Rice CM (2000) Hepatitis C virus-encoded enzymatic activities and conserved RNA elements in the 3' nontranslated region are essential for virus replication in vivo. *J Virol* 74(4):2046-2051.

148. Dumont S*, et al.* (2006) RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. *Nature* 439(7072):105-108.

149. Yao N, Reichert P, Taremi SS, Prosise WW, & Weber PC (1999) Molecular views of viral polyprotein processing revealed by the crystal structure of the hepatitis C virus bifunctional protease-helicase. *Structure* 7(11):1353-1363.

150. Lamarre D*, et al.* (2003) An NS3 protease inhibitor with antiviral effects in humans infected with hepatitis C virus. *Nature* 426(6963):186-189.

151. Steinkuhler C*, et al.* (1998) Product inhibition of the hepatitis C virus NS3 protease. *Biochemistry* 37(25):8899-8905.

152. Medrano J*, et al.* (2011) Hepatitis C virus (HCV) treatment uptake and changes in the prevalence of HCV genotypes in HIV/HCV-coinfected patients. *J Viral Hepat* 18(5):325-330.

153. Soriano V*, et al.* (2008) Spontaneous viral clearance, viral load, and genotype distribution of hepatitis C virus (HCV) in HIV-infected patients with anti-HCV antibodies in Europe. *J Infect Dis* 198(9):1337-1344.

154. Tuyama AC*, et al.* (2010) Human immunodeficiency virus (HIV)-1 infects human hepatic stellate cells and promotes collagen I and monocyte chemoattractant protein-1 expression: implications for the pathogenesis of HIV/hepatitis C virus-induced liver fibrosis. *Hepatology* 52(2):612-622.

155. Thibault S, Fromentin R, Tardif MR, & Tremblay MJ (2009) TLR2 and TLR4 triggering exerts contrasting effects with regard to HIV-1 infection of human dendritic cells and subsequent virus transfer to CD4+ T cells. *Retrovirology* 6:42.

156. Larkin MA*, et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947-2948.

157. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792-1797.

158. Hunt PW (2012) HIV and inflammation: mechanisms and consequences. *Curr HIV/AIDS Rep* 9(2):139-147.