1-1-2011

Viral Diversity by Deep Sequencing: Approaches to Analyzing Effects of Anti-HIV Treatments

Rithun Mukherjee
University of Pennsylvania, rithun@gmail.com

Viral Diversity by Deep Sequencing: Approaches to Analyzing Effects of Anti-HIV Treatments

Abstract
HIV is a deadly virus responsible for the AIDS pandemic, which has claimed countless lives since its origins in the early 1980s. A cure for HIV is still elusive - HIV can exist as a diverse and dynamic population that adapts quickly to immune and drug pressures, making elimination of infection difficult. Advances in antiretroviral (ARV) therapy have resulted in effective control of HIV for some but not all patients. This dissertation reports case studies of the response of viral populations to selection pressures exerted by emerging anti-HIV therapies. Deep sequencing technology was used to probe viral swarms at high-resolution, which helped make clinically relevant conclusions. Further, novel computational approaches were implemented to control procedural noise and carefully interpret signal. In one study, we examine HIV integrase inhibitors (INIs), which are among the latest ARV drugs. INIs act at a pre-integration level by aborting viral integration, which would normally lead to lasting infection. Raltegravir (RAL) is the only FDA-approved INI to date. Investigating drug resistance is crucial to informing future course of ARV therapy. We describe evolving HIV swarms in patients exhibiting a switch in RAL-resistance profiles. To understand implications of RAL administration, we analyzed the pre-therapy or treatment-naÃ¯ve context for the viral populations in-depth. Our findings suggest that predominant mutations arise only in presence of RAL - in its absence, they do not constitute fit polymorphisms. For all their effectiveness, drugs have not eradicated HIV. A recent clinical case, however, involving transfer of HIV-resistant cells to an infected patient, resulted for the first time in possible cure. This emphasized the importance of gene-modification and cell-based therapies to treat HIV. One such strategy showing promise uses an antisense to target HIV. The approach has been safe although clinical efficacy has not been fully determined. In support of one such study, we deep-sequenced viral swarms in the presence of antisense-modified cells. Encouragingly, we observed minority strains harboring evidence of antisense pressure in vivo, demonstrating the potential of alternative therapy. Finally, this dissertation underscores the significance of rare signatures in HIV populations, and outlines methods to investigate them.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Genomics & Computational Biology

First Advisor
Frederic D. Bushman

Keywords
Antisense, Deep sequencing, Gene therapy, HIV, Pyronoise, Raltegravir

VIRAL DIVERSITY BY DEEP SEQUENCING: APPROACHES TO ANALYZING EFFECTS OF ANTI-HIV TREATMENTS

Rithun Mukherjee

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2011

Supervisor of Dissertation

Frederic D Bushman, Professor of Microbiology, University of Pennsylvania

Graduate Group Chairperson

Maja Bucan, Professor of Genetics, University of Pennsylvania

Dissertation Committee:

Warren J Ewens, Professor of Biology, University of Pennsylvania

James A Hoxie, Professor of Medicine, University of Pennsylvania

Hongzhe Li, Professor of Biostatistics, University of Pennsylvania

Michael D Miller, Site Lead, Infectious Disease – HIV, Merck Research Laboratories

VIRAL DIVERSITY BY DEEP SEQUENCING:
APPROACHES TO ANALYZING EFFECTS OF ANTI-HIV TREATMENTS

Rithun Mukherjee

ACKNOWLEDGMENTS

I do not have enough words to thank my advisor Rick Bushman, so I will just say that the best in this dissertation is a reflection of his mentoring. From short elevator trips to brief restroom meet-and-greets, Rick was always ready with an advice or anecdote, and I will miss that enthusiasm and intensity very much. I am also thankful for his patience and encouragement especially during challenging research adventures, which was largely instrumental in grooming my critical thinking. I will particularly treasure his careful training on the art of scientific writing, presentation and public speaking – these are gifts for a lifetime. Finally, I am grateful to him for providing the research resources and supporting my time in the lab, without which this dissertation would not have been written.

The Bushman lab was an excellent place for my graduate training. I was fortunate to have Chris Hoffmann train me in the best practices of HIV and sequencing related lab-work. Also, stimulating discussions with him over the years contributed in no small way in shaping my thoughts and research approach over the years. I was glad I shared workspace with Troy Brady. He was my first refuge for a question or discussion, sharing my excitement over results, or receiving a reality check on my research and therefore on my excitement levels. And when in need of a computer or programming related tip-of-the-day, I had the ready presence of Nirav Malani. His recommendations and quick-fix suggestions made many a day. Many a computational discussion was also enriched talking to Scott Sherrill-Mix and Kyle Bittinger. I thank Frances Male for technical assistance and Mali Skotheim and Caitlin Greig for administrative support. I also acknowledge all other lab members, past and present, for making the Bushman lab a wonderful place to work and train.

I express my gratitude to all my committee members for serving on my committee and guiding me through to a successful completion of my degree requirements. I am

very grateful to Warren Ewens, the chair of my committee, for his close attention to my academic progress and well being throughout my time at Penn. I am also fortunate to have the benefit of his teaching and instruction – the little I know of statistics is largely a product of that. For advice on deep sequencing related statistical issues, I also thank Hongzhe Li. To have access to the biological and clinical expertise of Michael Miller and James Hoxie was a boon – their emphasis on having me see the big picture behind my research has been an invaluable lesson.

I am thankful for the collaborations with Carl June, James Riley, Gwendolyn Binder-Scholl and Michael Miller. I thank Shane Jensen for help on statistical problems that are an inevitable outcome of research on HIV quasi-species. I thank Farida Shaheen, for training me to use Center for AIDS Research (CFAR) facilities, and Erik Toorens, for processing many a 454/Roche sequencing run for me at the Sequencing Center. The Genomics and Computational Biology Graduate Group has been like a family – among faculty, I thank Maja Bucan, Harold Riethman, Junhyong Kim, Li-San Wang and especially Sridhar Hannenhalli, for his guidance and computational training over an extended rotation project and beyond. It has been a pleasure to forge good friendships among students – Molly, Praveen, Greg, Perry, Serena, Najaf, Rumen, Adam, Kathleen, and many others, and it was special to know Le, my only classmate.

I thank my family, friends and teachers in India and elsewhere, whose good wishes for my academic career is a blessing. I express my gratitude to my father, for tirelessly espousing good education and inculcating the value of learning all my life. I thank my close friends, Chakresh, Harinder, Uttam, Sougata, Shanker, Anna and Achyuta. Among my teachers, I particularly thank Sudhir Kaicker, Naidu Subbarao, Prabeer Sinha, Baishnab Tripathy, Ramesh Bamezai, Shyamal Goswami and R K Kale. Finally, I acknowledge my childhood friend and wife Bhramar, whose loving support has been indispensible for my academic career. My time at Penn began with our marriage and without her companionship, this dissertation would not have been written.

ABSTRACT

VIRAL DIVERSITY BY DEEP SEQUENCING:

APPROACHES TO ANALYZING EFFECTS OF ANTI-HIV TREATMENTS

Rithun Mukherjee

Frederic D Bushman

HIV is a deadly virus responsible for the AIDS pandemic, which has claimed countless lives since its origins in the early 1980s. A cure for HIV is still elusive – HIV can exist as a diverse and dynamic population that adapts quickly to immune and drug pressures, making elimination of infection difficult. Advances in antiretroviral (ARV) therapy have resulted in effective control of HIV for some but not all patients. This dissertation reports case studies of the response of viral populations to selection pressures exerted by emerging anti-HIV therapies. Deep sequencing technology was used to probe viral swarms at high-resolution, which helped make clinically relevant conclusions. Further, novel computational approaches were implemented to control procedural noise and carefully interpret signal. In one study, we examine HIV integrase inhibitors (INIs), which are among the latest ARV drugs. INIs act at a pre-integration level by aborting viral integration, which would normally lead to lasting infection. Raltegravir (RAL) is the only FDA-approved INI to date. Investigating drug resistance is crucial to informing future

course of ARV therapy. We describe evolving HIV swarms in patients exhibiting a switch in RAL-resistance profiles. To understand implications of RAL administration, we analyzed the pre-therapy or treatment-naïve context for the viral populations in-depth. Our findings suggest that predominant mutations arise only in presence of RAL – in its absence, they do not constitute fit polymorphisms. For all their effectiveness, drugs have not eradicated HIV. A recent clinical case, however, involving transfer of HIV-resistant cells to an infected patient, resulted for the first time in possible cure. This emphasized the importance of gene-modification and cell-based therapies to treat HIV. One such strategy showing promise uses an antisense to target HIV. The approach has been safe although clinical efficacy has not been fully determined. In support of one such study, we deep-sequenced viral swarms in the presence of antisense-modified cells. Encouragingly, we observed minority strains harboring evidence of antisense pressure *in vivo,* demonstrating the potential of alternative therapy. Finally, this dissertation underscores the significance of rare signatures in HIV populations, and outlines methods to investigate them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 INTRODUCTION

Background and motivation

A goal of modern medicine is to understand and control infectious agents. The host immune system protects by targeting structures or sequences of the pathogen that are sufficiently distinct from those of the host for specific recognition. Pathogens that are highly mutable have the ability to withstand a variety of pressures – they are difficult targets for the host as they can readily escape the immune system. Examples include RNA viruses [1] and single stranded DNA viruses [2, 3], which quickly produce a wide spectrum of mutants after infection by virtue of an error-prone replication system.

HIV is among the most deadly of these viruses. It has already claimed millions of lives in a pandemic that is ongoing. The virus is fascinating in that it infects some of the very cells in the body that protect against infections. In fact HIV infects more efficiently when its target cells are in an activated state, which ironically is their state when there is an infection to be cleared. After replication in a host, HIV is not a single viral strain but a distribution of strains, each different from the other, although related in sequence. As discussed below, such entities have unique properties and have been termed *quasi-species*.

Given the challenges posed by HIV, developing therapeutic strategies has been a huge challenge to biomedical research. HIV is the topic of the studies described in the following chapters. These investigations were performed in clinical settings, in which a particular therapeutic approach was applied to either human subjects or animal models. Outcomes and aspects of efficacy of the treatment strategies are discussed. The larger goal in each case, however, was highlighting clinically relevant

observations on the nature of the HIV quasi-species. In addition to assessing treatment efficacy, this is relevant in deciding the future course of therapy.

In this dissertation, I outline ways in which we have explored aspects of HIV quasi-species under selective pressure from therapy. In each case, a high-resolution picture of circulating viral variants was obtained by high-throughput deep sequencing technology. Further, novel algorithms were implemented to reduce procedural error whereas bioinformatics and statistical techniques were devised to robustly identify meaningful signals. Beyond a collection of methods, I hope the reader also gains an appreciation of the nature of the data in studies like these and why specific computational approaches are necessary.

Multiple therapeutic strategies are available to treat HIV infection. Given the difficulties of controlling a complex quasi-species, absolute eradication of virus is still mostly hypothetical – progress is often assessed by identifying treatment-induced signatures in "pockets" of the quasi-species structure. I hope that this dissertation illustrates approaches that can be adapted to such studies. In the following sections of this introduction, I discuss relevant concepts and provide background on the clinical context of the studies that follow in chapters 2-4. In parallel, I aim to inform the reader about advances made in sequencing that makes such studies feasible. Finally, I introduce some of the bioinformatics and algorithms used in analyzing the sequencing data.

The quasi-species

Mutability

Replication of the genetic material of all species results in error at some rate. Representative mutation rates per replication cycle per genome or μ_g are shown in

Table 1.1. It can be seen that there is a difference between RNA viruses and species with DNA genomes. For the latter, μ_g is around 0.004 irrespective of genome size. Higher eukaryotes such as humans have a bulk of DNA that is non-functional – in such cases μ_g is estimated for the *effective* genome, which is the part where any change is likely to have functional consequences and where mutations are largely expected to be deleterious. DNA is more stable than RNA, and the replication machinery has robust proofreading and repair activities [4].

GENOME	SIZE (in bases)	μ_g
RNA		
Poliovirus	7.4E+03	0.8000
Influenza A	1.4E+04	1.0000
MLV	8.0E+03	0.0600
RSV	9.3E+03	0.4300
HIV-1	9.7E+03	0.2200
DNA		
Bacteriophage M13	6.4E+03	0.0046
Bacteriophage λ	4.9E+04	0.0038
Bacteriophage T4	1.7E+05	0.0040
E. coli	4.6E+06	0.0025
S. cerevisiae	1.2E+07	0.0027
C. elegans	8.0E+07	0.0040
H. sapiens	3.2E+09	0.0040

Table 1.1 Mutation rates for the replication machinery of representative genomes.
Retroviruses are indicated in green and eukaryotes in red. Number of errors per replication cycle per genome (μ_g) is contrasted with genome size. For the eukaryotes, μ_g is calculated for the functional part of the genome where errors are likely to produce effects [92].

Viruses with RNA genomes, in contrast, have higher values for μ_g (Table 1.1). For their replication, they depend on RNA dependent polymerases. These include RNA dependent RNA polymerases (RdRPs) as in poliovirus and influenza virus. Retroviruses use reverse transcriptases (RTs). Both classes of enzymes are about a million times more error-prone than the DNA polymerases that replicate DNA genomes [4]. This poses a risk that the genomic information encoded as RNA could

be lost in a few generations. This happens at a theoretical limit called the *error threshold* [5] beyond which mutations accumulate at a rate faster than they can be selected against. However RNA being more labile than DNA, the size of genomic RNA is limited (Table 1.1). This limits the number of mutations on any individual genome such that μ_g values are still high enough to produce diversity quickly but low enough to avoid surpassing the error threshold.

Complex distribution of mutants

Many viruses with RNA genomes have large burst sizes and short generation times. The rapid exponential growth and high μ_g combine to continuously generate mutants at a brisk pace. This results in a large number of non-functional genomes. The huge number of progeny produced, however, ensures that there is still a large arsenal of diverse functional forms in the viral population. Manfred Eigen in the 1980s proposed a theoretical framework to describe such dynamic populations featuring a spectrum of variants – the concept of *quasi-species* [6].

The term quasi-species defines distributions of variants of a replicating system such that the density of variants is proportional to their fitness [7, 8]. The fitness coefficients are in turn dependent on selection differentials characteristic of the environment. For a replicating system in equilibrium with its environment, the quasi-species, described in terms of variants and their densities, is a stable distribution. Existence as quasi-species helps RNA viruses survive the often rapidly changing pressures exerted by the host environment. The unit of replication and evolution is no longer the individual genome but a population of related but different genomes, posing a substantial challenge for controlling such pathogens.

Nature of the quasi-species

The typical size of a RNA genome such as for HIV-1 is of the order of 10kb. Theoretically, there can be $4^{10,000}$ different sequences for such a genome. This astronomical number represents the entire sequence space available for exploration to the HIV-1 replication system. Of all possible sequences, most do not code for meaningful genomes. Only a fraction does, and under given selection pressure that fraction defines a quasi-species.

The nature of the quasi-species is illustrated in the context of dynamics of the replication system in Figure 1.1 [7]. Consider the cubes in this panel to represent available sequence space. Assuming a viral population is centered on a wild-type genome, three situations can arise. A high-fidelity replication system results in a population that remains centered around the wild-type (Fig. 1.1a). This preserves the original genomic information but generates little diversity resulting in populations susceptible to immune eradication or antiviral agents. In contrast, an extremely error-prone system ends up exploring all available space (Fig. 1.1b). Although this yields a rich spectrum of variants, there is rapid and continual dissipation of genomic information as error rates exceed the error threshold, yielding replication incompetent genomes.

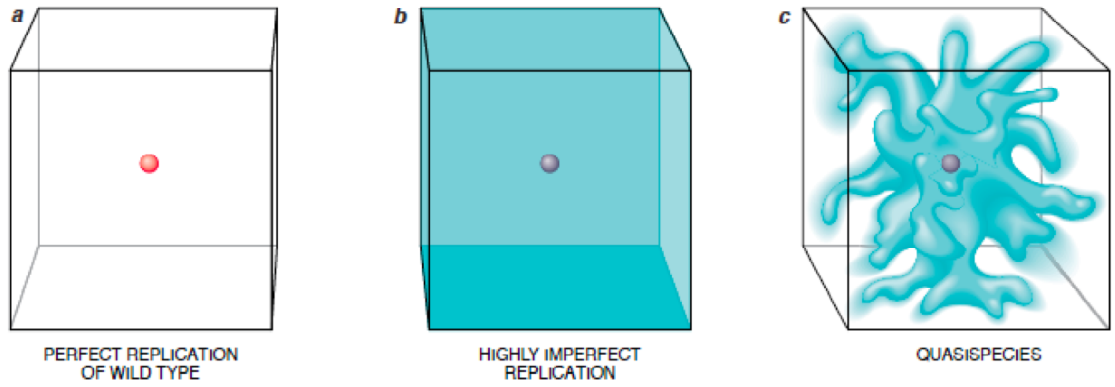


Figure 1.1 Replication system dynamics govern population distributions.

a. Perfect replication does not produce diversity. **b.** Imperfect replication generates high diversity at the cost of disintegrating genomic information **c.** Systems that balance creation of variation with preservation of information exists as diverse yet structured entities called quasi-species. Available

space for genome diversity is represented by a cube. Founder strain or population is shown as a sphere in the center of the cube. Area shaded in blue depicts space explored by replicating system over time. Permission to reuse this figure was granted by the publisher, © 1993 JSDinfographics, New York, NY.

Just below the error threshold, however, the viral population explores the sequence space more optimally [9]. First-generation mutants accumulate around the original wild-type in numbers that depend on their relative fitness. These in turn lead to newer mutants that arise especially from well-adapted existing mutants. In time, the distribution of sequences represents a cloud centered on the wild-type but asymmetrically exploring pockets of the sequence space dictated by fitness and selection considerations (Fig. 1.1c). In Eigen's words, this cloud is a quasi-species. Beyond the error threshold the cloud loses form (Fig. 1.1b) whereas much below it the cloud contracts to a single point (Fig. 1.1a). Successful viruses such as HIV-1 therefore function just below the error threshold in order to achieve high evolutionary flexibility while maintaining integrity of information.

The HIV-1 quasi-species

In HIV-1, mutations occur across all codon positions and non-coding regions. This is ideal for generating diversity and facilitates escape of the quasi-species under selection pressures. How does the virus cope with preserving essential structural motifs? The virus does indeed have conserved positions that preserve a structural backbone. These could be targets for the host immune system or therapy, except features having less functional constraints and therefore high variability conceal them. In HIV *env*, which is the viral surface protein and thus a target for both antibodies and vaccines, only about 20% of positions are constant whereas 70% are variable and 10% are hyper-variable [9]. It has been proposed that such flexibility in sequence space afforded to the quasi-species helps in outlasting the diversity generating capacity of the immune system [10]. This is a candidate explanation for

the delayed onset of AIDS, which takes place many months to years following initial HIV-1 infection.

HIV population sizes are often limited. For example, there is bottlenecking at the point of transmission leading to small founder populations at the start of infection [11]. Similarly, in a different scenario, HIV population size is restricted, such as with HAART [12, 13], or immune pressure exerted by broadly neutralizing antibodies or certain HLA alleles as in HIV elite controllers [14]. In these cases given constraints on viral proliferation and a changing environment of pressures exerted, HIV may not achieve the full requirements and flexibility of quasi-species. That is an aim for anti-HIV therapy – to restrain HIV to a localized region of the full quasi-species such that it cannot elude the therapy or immune response by mutation. Thus HIV clinical isolates will often not meet the full definition of a quasi-species, where each variant is present in proportion to its fitness. In this case, it is advisable to use other terms such as 'population' or ‘swarm’ [15].

Anti-HIV therapy

The HIV replication cycle

It is necessary to study the replication cycle of the pathogen to identify treatment opportunities. A virus like HIV-1 needs both viral and host proteins to replicate. Careful characterization of the steps involved helps in determining processes that are sufficiently distinct from those in the host, which are good choices for therapy. The HIV replication cycle outlining the various steps is illustrated in Figure 1.2.

HIV-1 infects cells that bear the molecules CD4 and either CCR5 or CXCR4 on their surface [16]. Target cells are constituents of the immune system, including CD4 T-cells and macrophages. Both CD4, which is the primary receptor, and one of the two

co-receptors, CCR5 or CXCR4, are required for viral entry and fusion with the host cell, steps that have been targeted by drugs. The transmitted founder virus is also a focus of therapy – it has been proposed that only one or a few particles typically establish initial infection [11]. The founder virus is mainly the CCR5 utilizing kind [17] whereas CXCR4 utilizing variants usually appear late in pathogenesis near the onset of AIDS [18]. Accordingly, much research has been undertaken to inhibit or reduce the CCR5 interaction.

The envelope protein of HIV-1 (Env) is a key player in the entry process [19]. The surface gp120 subunit first engages CD4. This promotes interaction of the V3 loop of gp120 with co-receptor. Subsequently, conformational changes in the trans-membrane gp41 subunit of Env result in formation of a six-helix bundle that mediates fusion of the viral and target cell membranes. This releases the viral core into the cytoplasm where it is uncoated. Next the viral RNA genome is reverse transcribed into DNA. This crucial step, which is a hallmark of retroviruses, is catalyzed by the viral enzyme reverse transcriptase (RT) [20]. RT was the target of the earliest antiretroviral drugs, and RT inhibitors continue to be a mainstay of modern HIV combination therapies.

Reverse transcription leads to the second characteristic feature of the retroviral replication cycle, which is integration of the viral DNA into the host genome. The viral enzyme involved is integrase (IN) [21]. Together with several host factors, IN associates with the viral DNA to form the pre-integration complex (PIC), which enters the nucleus. There IN catalyzes the joining of the viral and host DNA. This too has been a target of anti-HIV therapy. Upon integration, HIV-1 is stably incorporated as a part of the host genome.

The integrated viral DNA or provirus co-opts the cellular machinery to transcribe and translate all viral proteins and as well as produce genomic viral RNA [22]. The latter is packaged with the viral Gag and Gag-Pol poly-proteins at the cell surface.

Finally nascent viral particles bud off the host cell. Before they are ready to infect new cells, however, they need to undergo maturation, involving cleavage of Gag and Gag-Pol into its structural and enzymatic components by the third viral enzyme, protease (PR) [23]. Drugs have also been developed to inhibit PR, thereby preventing the formation of infectious virus.

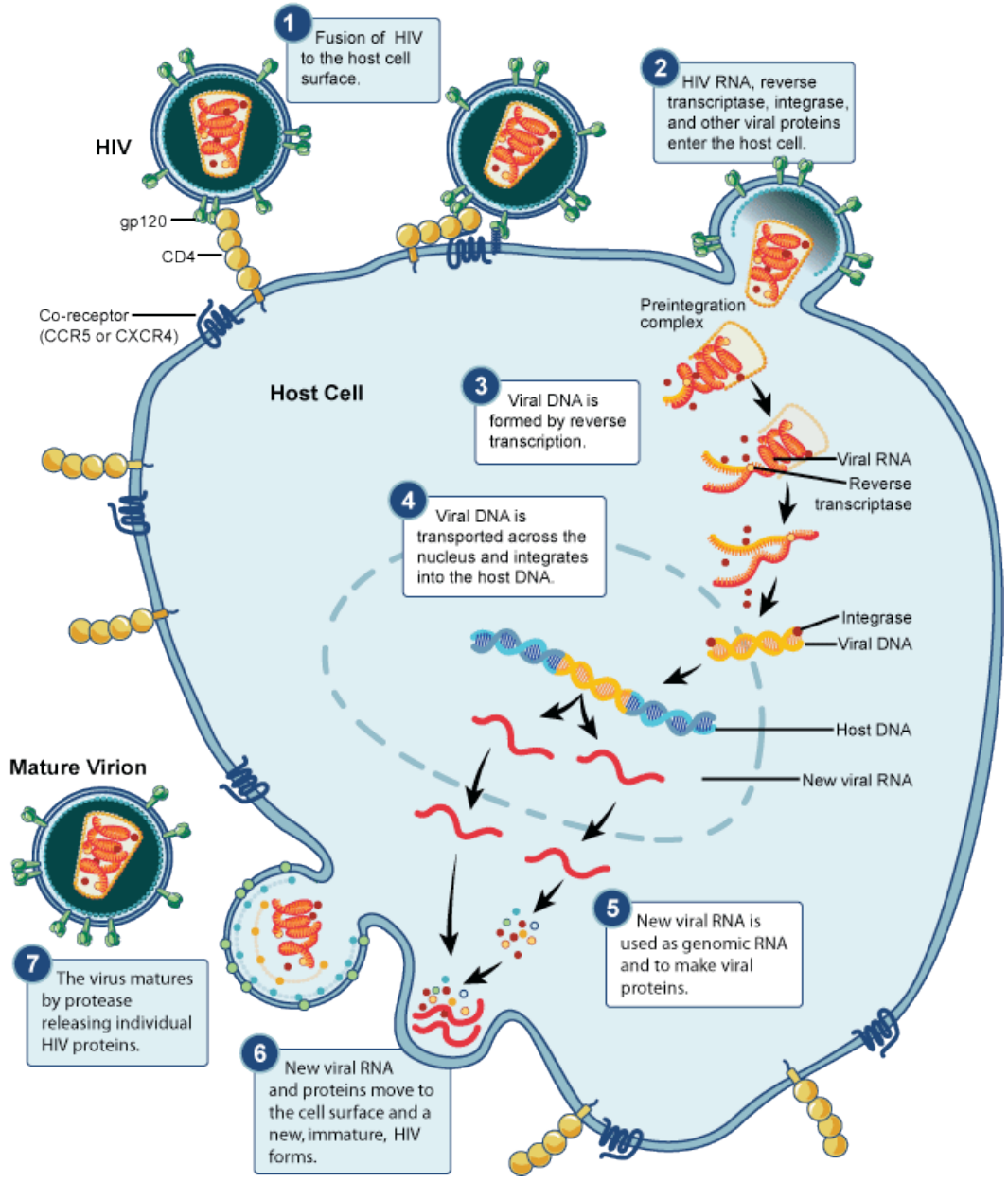


Figure 1.2 The HIV replication cycle

Courtesy: National Institute of Allergy and Infectious Diseases

Highly active anti-retroviral therapy or HAART

Developing drugs against HIV-1 has been challenging due to the complex and dynamic nature of the quasi-species. For the early part of the HIV pandemic, until the mid-1990s, only a handful of antiretroviral drugs received FDA approval. All of them were nucleoside analogs lacking a 3' hydroxyl group such that their incorporation terminated DNA synthesis [24, 25]. They were designed to interfere with reverse transcription by aborting production of viral DNA. Such compounds were called nucleoside RT inhibitors (NRTIs) and the earliest ones included zidovudine (AZT), lamivudine (3TC) and stavudine (d4T).

There is always ongoing DNA synthesis in the host cell, so any NRTIs that can be incorporated into host cell DNA have the potential for high toxicity. Also mutant forms of RT arise that overcome NRTIs by either removing them or precluding their incorporation [25]. Around the mid-1990s non-nucleoside RT inhibitors, or NNRTIs, were developed that targeted HIV-1 RT specifically instead of the process or synthesis that the enzyme catalyzes. Nevirapine and delavirdine were the earliest NNRTIs. These inhibited by binding to RT and changing the conformation of its active site [25]. NNRTIs suffered, however from low fitness costs of resistant RT, which meant that HIV-1 readily escaped from drug pressure [26].

PR was the second viral enzyme that was targeted by FDA-approved inhibitors. PR inhibitors or PIs inhibited the cleavage activity of PR [24]. As a consequence resistance mutations emerged both on PR as well as around cleavage sites of the Gag and Gag-Pol polyproteins, which are the substrates for PR [26, 27]. Nelfinavir, ritonavir and saquinavir were among the first PIs that received approval, also in the mid-90s. Thus around 1995-6, there was not only a proliferation in the number of antiretroviral drugs but also availability of drugs that had different targets and mechanisms of inhibition. This prompted the initiation of combination therapy with three or more drugs under the premise that even with its high μ_g HIV-1 would find it

difficult to escape them simultaneously [28-30]. Additionally, the cumulative reduction of ongoing viral replication by multiple drugs would be sufficient to control viral loads in patients at levels where few healthy cells would be infected [31]. Optimization of such cocktails of drugs led to establishment of well-defined HAART regimens [32-34].

Poor adherence and/or tolerance of drug regimens lead to breakthrough resistance in virus. Resistant strains emerging in individuals can also be transmitted resulting in their wider circulation in the host population. This demands continuously finding new ways of targeting the virus. Such efforts have led to drugs inhibiting the entry step. Enfuvirtide, or T20, is an FDA-approved drug that limits viral fusion – it is a peptide mimic of the six-helix bundle of gp41 that interferes with the inter-helix interactions [35]. Maraviroc, another recent drug in the market, acts by antagonizing CCR5 and is unique in targeting an essential host factor for HIV [36].

Integrase inhibitors

Success has also been achieved in targeting IN, the third enzyme of HIV-1. Researchers at Imperial College London recently described the structure of full-length retroviral IN in complex with viral and target DNA providing insights into the integration reaction [37, 38]. It was proposed that the active complex or intasome has an IN tetramer bound to viral DNA. The intasome causes bending of the target DNA, which facilitates the joining of viral and host DNA strands. The structure of the intasome is depicted in Figure 1.3 [38]. Thus far all integrase inhibitors inhibit the strand transfer reaction of IN, so they are called INSTIs [39]. The only drug that has so far received FDA approval is Raltegravir (RAL). Studies of INSTIs have focused on investigating drug efficacies and characterizing resistance. In chapter 2 of this dissertation we describe an analysis of patients undergoing RAL therapy who exhibited a switch in RAL-resistance profiles.

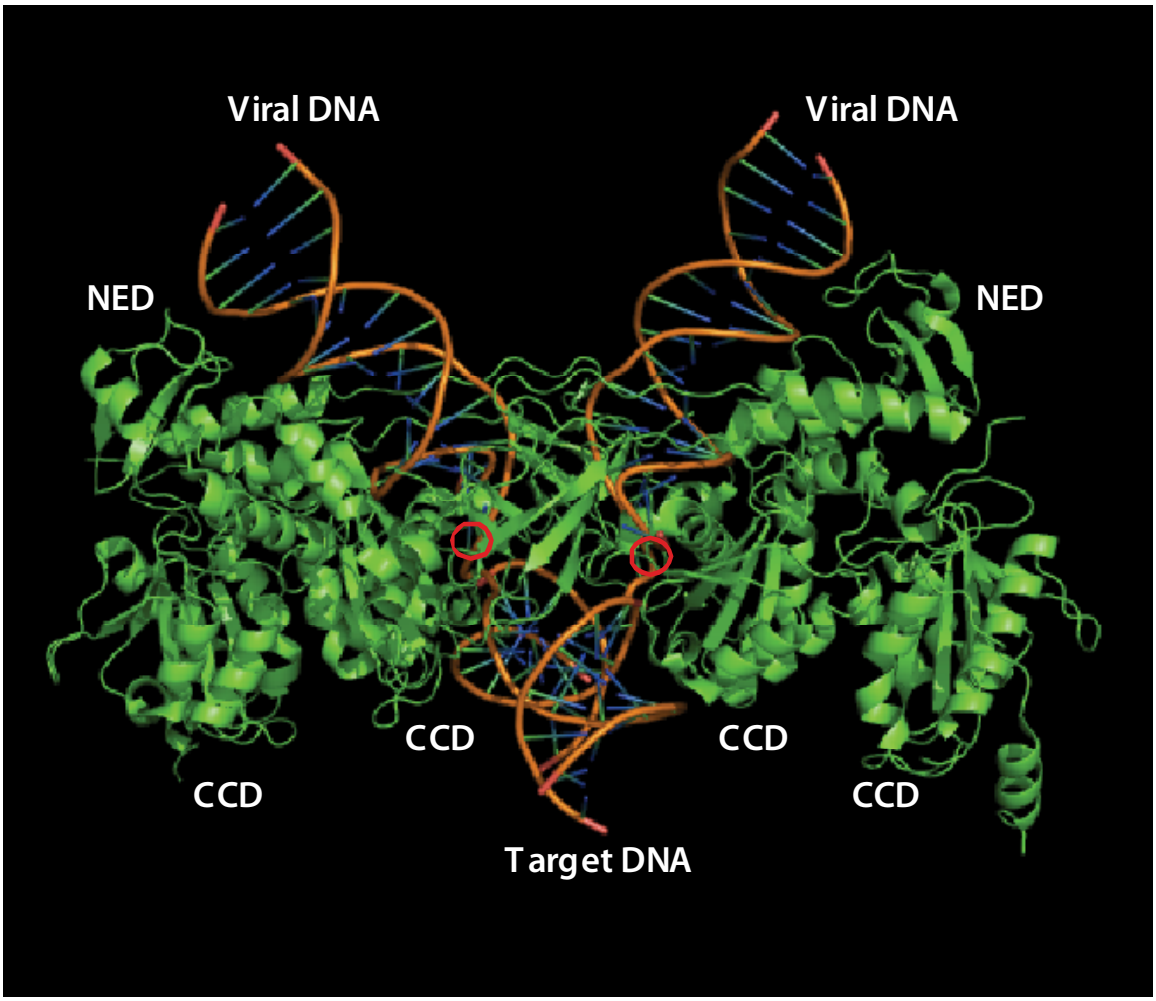


Figure 1.3 The intasome of prototype foamy virus (PFV)

The strand transfer complex consists of integrase tetramer, viral DNA and target or host DNA. The spatial arrangement of the catalytic core domains (CCD) of the four IN molecules around the viral and target DNA is shown. The active site (red circles) is part of the CCD juxtaposing on the junctions between the viral and target DNA molecules. The N-terminal extension domains (NED) are also displayed. Figure adapted from Maertens *et al* [38].

IN requires two bivalent metal ions such as magnesium (Mg^{2+}) or manganese (Mn^{2+}) at its active site to catalyze strand transfer. INSTIs inhibit IN by binding to the active site, chelating the two metal cations and displacing the terminal 3' nucleotide of the viral DNA in the PIC [37, 40]. IN counteracts this by acquiring mutations in the active site. Against RAL, the primary mutations are Y143H/R/C, Q148H/R/K and

N155H [41]. These primarily alter the native conformation of IN such as to make INSTI binding energetically unfavorable [42].

Of all major IN resistance forms, Q148H in conjunction with its accessory mutation G140S/A results in the largest reduction in RAL susceptibility [42, 43]. This configuration also exhibits the most cross-resistance to other INSTIs currently evaluated in clinical trials, namely Elvitegravir (EVG or GS9137) and Dolutegravir (S/GSK 1349572) [44]. The study in chapter 2 describes patients presenting a switch from N155H to Q148H+G140S. Investigating fine details of the genotypic spectrum of the viral populations that underlies this transition could be crucial for both understanding resistance evolution and informing therapy. Also it is important to interrogate levels of natural polymorphisms for these mutations. This would relate to their fitness in the wild-type quasi-species for IN, that is in absence of INSTI. It could also explain the propensity of their outgrowth following drug treatment and help in understanding the genetic barrier to INSTIs. Accordingly, in the context of the patients studied, pre-RAL occurrence of resistance polymorphisms is analyzed in detail in chapter 2.

The Berlin patient – curing HIV?

In 2008 an HIV-infected person presented with symptoms of acute myeloid leukemia (AML) in Berlin, Germany. He underwent chemotherapy, however AML relapsed and stem cell transplantation (SCT) from a healthy HLA-matched donor was performed. Keeping in mind the HIV status of the patient, the donor was selected to be of homozygous CCR5Δ32 genotype [45]. CCR5Δ32 is a 32 bp deletion in the CCR5 gene that is found in roughly 1-3% of the Caucasian population. It results in a truncated gene product that abrogates cell surface expression of CCR5 [46, 47]. This is restrictive for cellular entry of CCR5-utilizing HIV – crucially the transmitted virus as well as those dominating until the later stages of pathogenesis

are CCR5-tropic. The objective behind sourcing SCT from a CCR5Δ32 donor by design was repopulating the immune system with progenitor cells that would also be restrictive to CCR5-tropic virus.

The Berlin patient underwent a conditioning therapy pre-SCT intended to eradicate leukemic cells as well as suppress immune response to donor cells [45]. For the former, myelo-ablative chemotherapy and total body irradiation (TBI) was performed – in the process potentially also eliminating infected cells with actively replicating or latent virus. SCT resulted in successful engraftment of donor cells and both peripheral and mucosal CD4 T cell populations were reconstituted at levels comparable to HIV-uninfected SCT recipients as well as healthy control patients [48]. Over the long-term both HIV target cell populations – T cells and macrophages – showed a transition to the donor CCR5Δ32 genotype. In addition to the peripheral circulation, this was also true of the mucosal circulation, which is where HIV infection is first initiated. Finally, post-SCT for over 40 months now, there is no detectable virus in the patient either in the plasma or in tissues – all in the absence of HAART that was discontinued to reduce drug toxicity following SCT-conditioning. Consistent with this, researchers reported a decreasing trend in the titers of anti-HIV antibodies [48].

Although the CCRΔ32 genotype is resistant to CCR5-tropic HIV, it does not protect against CXCR4-tropic or dual tropic virus. Indeed researchers demonstrated that engrafted donor cells in the Berlin patient were prone to infection by CXCR4-tropic HIV [48]. Deep sequencing of the pre-SCT viral population had revealed a minority of CXCR4-tropic strains [45]. Thus the possibility remained of ensuing viral infection post-SCT. Undetectable virus for almost 4 years, however, suggests that post-SCT levels of circulating or latent CXCR4-tropic HIV were insufficient for reinitiating HIV infection.

The Berlin patient is a striking example of possible cure from HIV. Potentially conditioning therapy contributed to cure by expunging long-lived cells and viral reservoirs therein [48]. Nevertheless reconstituting the patient immune system with HIV-resistant cells was a significant factor, emphasizing the promise of cell therapy to control HIV. An important consideration is cure from HIV in the absence of HAART. This overcomes several challenges of long-term drug use: persistent costs, toxicities, and tolerance issues that often lead to poor adherence and breakthrough of resistance. Furthermore HAART controls HIV but never 'cures' – cessation of drugs is accompanied by breakthrough of virus. This happens due to eventual activation of latent HIV existing as provirus in long-lived cells. Such viral reservoirs are a formidable hurdle for cure and beyond the reach of HAART [49].

Cell and gene therapy targeted to HIV

Cell and gene therapy strategies are directed at long-term modification of HIV target cell populations in the host such as to reduce susceptibility to the virus in the absence of drugs. Approaches have been directed against either viral targets or cellular factors that are essential for viral replication, and either proteins or nucleic acids. Although therapeutic benefit has not been conclusively established thus far, researchers have reported survival advantage of gene-modified cells in the presence of virus. Also, trials have successfully addressed safety considerations related to genotoxicity. Transient lowering of viral loads were observed in some cases. There is hope that improvements can made with better *in vivo* engraftment, persistence and expression of gene-modified cells and their products. Certainly given the example of the Berlin patient, this is an area with much promise.

Gene therapy at the level of proteins

The first gene therapy clinical trial was conducted in the mid-1990s in the early days of HAART. HIV patients were infused with autologous (or self) CD4 T cells modified with an antiviral protein called RevM10 [50]. This is a form of HIV Rev with mutations in a conserved domain that interacts with host factors. RevM10 binds to the Rev-response element or RRE on viral transcripts but does not promote their nuclear export. Thus it interferes with Rev function and inhibits viral replication by exerting a dominant negative effect. As opposed to control cells transduced with ΔRevM10, which does not produce a protein, RevM10 transduced cells survived longer in HIV patients [50, 51]. The latter also inhibited HIV replication *in vitro* although *in vivo* there was no pronounced effect on viral load [50, 51]. Persistence of modified cells was a limitation – when a RevM10 plasmid was delivered by microgold particles, the half-life was 1-2 weeks [50]. This was improved upon to a few months with a murine retrovirus (MoMuLV) based delivery vector, which presumably led to stable integrated forms of RevM10 [51]. Trials were assessed for safety – no adverse biochemical or physiological events were reported and no replication competent MoMuLV or undesirable immune responses to RevM10 detected.

A more recent clinical trial involved the membrane anchored antiviral peptide C46 (maC46), a mimic of a gp41 heptad helix [52]. It is similar to the fusion inhibitor T-20 and inhibits HIV replication *in vitro* by blocking entry. The trial was a first in using a pre-integration step target in a gene therapy setting. This can afford protection to modified cells without the requirement of HIV integration. HIV patients that received autologous T cells modified with a retroviral vector encoding maC46 exhibited no adverse effects, insertional mutagenesis or attenuating immune responses to maC46. *In vivo* levels of modified cells correlated with the reinfusion doses and in some patients maC46-marked cells persisted for a year. Although modified cell numbers were not high enough to reduce viral load, they possibly stimulated T cell growth such that infusions led to transient increases in total CD4 cell counts.

Host proteins essential to HIV have also been targeted. Such strategies have fewer concerns with viral escape. However, they have to contend with possible toxic effects on the host, especially if targeted protein has important roles in the body. Given the success story of the Berlin patient, and the HIV-resistance afforded by the CCR5Δ32 genotype, much effort has been invested in reducing or abolishing CCR5 expression on cell surfaces. In addition CXCR4 has also been targeted. To knockdown host gene expression, investigators have adopted RNAi techniques [53]. Ribozymes, which are catalytic RNA that bind and degrade complementary target RNA, have also been used [54]. Alternative approaches have focused on knocking out host genes [55, 56]. Advances in design of recombinant endonucleases called zinc finger nucleases or ZFNs have facilitated this [57]. ZFNs can be engineered to cleave genomic DNA at a unique locus. Although off-target effects are a concern, there are several advantages. Even transient expression of ZFNs can result in a permanent disruption of a gene – this also obviates the need for using integrating vectors as delivery vehicles. Clinical trials are ongoing to study effects of T cells modified with CCR5-targeting ZFN (clinicaltrials.gov Id NCT00842634, NCT01252641).

RNA based strategies

Gene therapy at the level of nucleic acid overcomes problems with immunogenicity that a modified cell expressing an antiviral protein could experience. RNAi and ribozyme based techniques have been popular although avoiding off-target effects is challenging. Results of the largest gene therapy trial till date, involving 74 HIV-1 patients, were reported in 2009 – a ribozyme that targeted the overlapping region of HIV-1 *vpr* and *tat* was used. This antiviral payload was delivered by a MoMuLV vector and was called OZ1 [58]. The study was also the first phase II clinical trial that tested anti-HIV gene-modified hematopoetic stem cells (HSCs). Compared to T cells, *ex vivo* transduction and expansion of HSCs has been difficult and associated

genotoxicity has been documented [59-61]. Still it is attractive to modify HSCs as antiviral resistance can then be transferred on a long-term basis to descendent immune cells of both myeloid and lymphoid lineages that HIV targets.

In the OZ1 trial, participants were infused with modified or control autologous HSCs. Importantly patient groupings into treated and control cohorts was randomized and the study double-blinded. OZ1-modified cells persisted in some patients at least for 100 weeks – OZ1 RNA was also detectable till that time in a few cases indicating the potential for durable *in vivo* expression. Although viral replication was not controlled following post-infusion interruption of HAART, viral loads over the long-term were lower in the treated group. Significantly, viral loads were even lower in patients with long-lasting OZ1 expression. Also, CD4 counts were higher in treated patients. Finally no serious adverse events were reported. Integration site analysis of OZ1 did not detect proliferation of any single clone and viral sequencing did not establish any OZ1 resistance.

Yet another RNA-based gene therapy procedure employs RNA decoys to divert HIV proteins. Good targets are HIV Tat and Rev: Tat binds to a trans-acting responsive or TAR element on the 5' UTR of viral mRNAs to promote transcription, whereas Rev recognizes RRE also on viral mRNAs – other than the completely spliced forms – to promote nuclear export. A phase I clinical trial conducted in 1999 transduced autologous HSCs with an RRE decoy delivered by a MoMuLV vector [62]. Gene-modified HSCs were detectable until almost a year post-infusion in some samples. Also they survived much longer than control HSCs that contained vector but not RRE decoy. Participants did not exhibit any complications and did not present with evidence for replication competent MoMuLV.

The story of anti-HIV env antisense

A possible first therapeutic benefit in HIV patients due to gene therapy in a clinical setting was reported in 2006 by investigators at University of Pennsylvania. The study enrolled 5 HIV patients who were failing HAART and involved the use of a ~1kb long anti-HIV antisense directed to HIV *env* [63]. Autologous CD4 T cells were modified with this antiviral RNA delivered as part of a lentiviral vector called VRX496. Remarkably, 3 patients showed at least 0.5 log reductions in viral loads over a period of 3-6 months post-infusion of gene-modified cells. In one patient, this reduction continued until at least a year, when a sharp drop of ~2 log was registered. After a year CD4 T cell counts had improved in 4 of 5 patients and no adverse events were reported.

The premise of antisense treatment was that antisense RNA would form duplexes with HIV messages rendering them impotent. In addition, the duplexes would be long double-stranded RNA (dsRNA) molecules that belong to the category of pathogen-associated molecular patterns (PAMPs) – these could then be recognized and degraded by mechanisms of cellular immunity [64]. Compared to ribozyme or RNAi, a long antisense is likely to make viral escape more difficult. As a corollary multiple mutations would be needed to evade antisense pressure, which might debilitate any escape variant.

The VRX496 study was also the first clinical trial for lentiviral vectors. Use of other retroviral vectors that integrate in the promoter regions of genes had led to incidences of insertional mutagenesis [60, 61]. In contrast, lentiviruses tend to integrate within transcriptional units away from promoters – vectors based on lentiviruses were thus developed as safer alternatives. Consistent with this, long-term follow-up analysis of VRX496 integration sites in patients did not detect evidence of genotoxicity [65]. Employing a lentivirus vector like VRX496 that contains elements of HIV also confers the benefit of conditional vector mobilization. By relying on HIV Tat and Rev for transcription and expression, integrated VRX496 is only activated upon HIV infection. Also, in the presence of HIV, VRX496 RNA can

potentially co-package with HIV Env forming antiviral particles that could spread HIV resistance from cell-to-cell. There was possible short-term VRX496 mobilization in patients but importantly there was no indication of a replication-competent VRX496 [63].

Taken together, this phase I antisense-mediated anti-HIV trial was promising in many respects. Most significantly, there was suggestion of HIV control in some patients beyond that achieved by HAART. This called for a deeper investigation of clinical efficacy. Researchers had already reported that VRX antisense pressure *in vitro* produced mutant virus – these were sufficiently impaired so that they were not true escape forms [64]. It remained to be seen if similar observations could be recapitulated *in vivo*. To answer this, and to get a better understanding of the mechanism of antisense inhibition of HIV, the clinical trial was simulated in a mouse model. This is discussed in detail in chapter 3. Subsequent experiments illustrating measurable effects of antisense on HIV populations are also described. This is also the first report of viral signatures bearing evidence of anti-HIV pressure *in vivo* in a gene therapy setting.

In a second VRX496 clinical trial, the majority of patients exhibited lowering of viral loads with one patient suppressing virus for about 15 weeks in the absence of HAART (clinicaltrials.gov Id NCT00295477). Given the results presented in chapter 3, molecular effects of antisense, if any, were investigated within patient virus. Longitudinal samples corresponding to the HAART-interrupted period in patients following VRX496 infusion were studied. Data and analysis are presented in chapter 4 – antisense effects were detectable at low levels. Interestingly, effects could be observed only for the initial time-points sampled, which is also when more antisense-modified cells are present.

Advances in sequencing and bioinformatics applications for HIV research

Early developments leading to sequencing of single genomes

Until the advent of high-throughput platforms onwards of 2005, capillary-based Sanger sequencing [66] was the main tool to read DNA. The basic chemistry in this process relies on sequencing by synthesis. The sequence is read using fluorescently labeled dideoxy-nucleotides, which terminate DNA polymerization. These incorporate at every base position in a fraction of the clonal DNA population being polymerized – this helps decipher the sequence. Sanger sequencing is still the choice for many low-scale applications today, although costs per sequenced base are unattractive for large-scale projects. However read lengths up to ~1kb and errors in the order of 0.001% are still among the best [67].

Initial efforts to genotype HIV populations from clinical samples used sequencing based on Sanger chemistry or hybridization to high-density oligonucleotide arrays, such as Affymetrix GeneChips [68, 69]. Sanger sequencing of HIV populations generates a composite sequence for the bulk of the viral swarm – termed population sequencing. This carries limited information on polymorphisms. GeneChip-sequencing using probes that differ at given positions can address this deficiency. Synthesizing probes, however, relies on knowledge of sequence thereby limiting its use for *de novo* applications. Neither approach allows the sequence estimation of individual variants. Thus linkage information is also lost. Furthermore they are unsuited to quantification of minority variants that can be clinically relevant.

In 1990, researchers at University of Edinburgh, Scotland, reported amplification of single molecules of provirus isolated by limiting dilution [70, 71] . Based on this approach, John Coffin and colleagues described an assay to perform single genome

sequencing [72]. They progressively diluted HIV cDNA, obtained by reverse-transcribing viral RNA, until only about 30% of real-time PCR reactions set up for a sample yielded amplified product. At these levels, according to the Poisson distribution, sample dilutions would correspond to on average 1 viral cDNA per positive reaction. In this way, individual genomes could be amplified and studied. The approach was appropriately named single genome amplification or SGA. Multiple independent viral genomes from a given sample were analyzed by setting up appropriate number of PCR reactions. Linkage information for viral variants was preserved. Additionally, since single genomes were amplified within any given reaction, PCR errors – such as due to nucleotide mis-incorporation or inter-template recombination – were suppressed. Also a single sequence was determined per positive reaction avoiding re-sampling of the same original template, which is a problem with bulk PCR of viral populations. In the report by Coffin and colleagues, no recombination-related error was detected whereas base substitution error was 0.01%.

Importantly, SGA detected drug-resistance mutations (DRMs) missed by population sequencing. DRMs identified by SGA at levels of 10-35% in samples were detected only 1 in 4 times by population sequencing while DRMs below 10% were rarely detected [72]. Investigators have since used SGA to study HIV transmission and establish founder virus genotypes [11, 73]. SGA has also been employed to demonstrate the contribution of HIV latency to residual viremia in patients displaying HIV control by HAART [74]. For all its merits, SGA is extremely resource-intensive due to requirement for huge number of PCR reactions and individual determination of sequences in the positive reactions. Throughput in SGA studies has been in the order of hundred to a few thousand sequences with an average of less than 100 sequences per sample. Such numbers could still miss rare signatures in HIV populations important in a clinical setting, prompting the need for high-throughput single molecule sequencing approaches.

Deep sequencing and associated computational techniques

The ability to perform numerous sequencing reactions in parallel on arrays has led to the rise of high-throughput sequencing platforms [67]. This involves several steps. From a sample of interest a library is prepared. This has fragmented whole genomes or amplified products from targeted regions with oligonucleotide adaptors attached at the ends of molecules. Next, clonal copies of DNA molecules sampled from the library are generated, either by emulsion PCR or bridge PCR with the oligonucleotide adaptors [67]. Resulting clonal clusters are spatially distributed on an array and subjected simultaneously to cyclic flows of enzymes, nucleotides and other reagents. Sequencing by synthesis iteratively adds the appropriate nucleotide for each clone accompanied by a light signal. This information is captured over the entire array and across all flows. The imaging technology precisely resolves all clones on the array – in this way massively parallel sequencing determines a single sequence read per clone.

Illumina is currently the market leader among next generation sequencing (NGS) technologies [75]. Per run 100-200 Gb can be generated with yields up to 1 billion read clusters, each 100-150 bases long (for paired ends, the numbers are 2 billion and 2 X 100-150 bases respectively). Error rates are in the order of 0.1%. The SOLiD platform offered by Life Technologies is a competitor in terms of information yield per run. Sequencing is performed by ligation of oligomer probes and is mediated by a DNA ligase, rather than polymerase [76]. The same template is read multiple times beginning at positions frame-shifted with respect to each other. This scheme ensures low error rates, however read lengths are shorter at ~60 bases. Whereas Illumina uses bridge PCR, SOLiD relies on emulsion PCR to achieve clonal amplification.

The third popular NGS platform is the 454/Roche system [77]. It employs emulsion PCR followed by a sequencing process based on pyrosequencing [78]. Compared to

the Illumina or SOLiD systems, a 454 run generates only around 1 million reads and it is also 15-30 times costlier per base [79]. Nevertheless 454/Roche has its advantages such as faster run times and longer read lengths [79, 80]. It should be noted that numbers and figures quoted here for all NGS systems are current estimates (http://knowledgebank.blueseq.com/sequencing-platforms/) and are subject to changes with continuous progress in technologies.

454 pyrosequencing was introduced in 2005 as the first commercially available NGS platform [77]. Within a short time, researchers started using it for deep sequencing HIV populations [81, 82]. Longer read lengths help analyze polymorphism linkages on viral variants. Costs can be reduced by multiplexing samples by bar-coding [82, 83], which is easily done during library preparation. Also, the depth afforded by 454/Roche is adequate to sample clinical samples of HIV. For these reasons 454/Roche is a good NGS platform of choice for studying HIV diversity. All deep sequencing studies described in this dissertation have used it to gain valuable insights into existence of rare signatures in HIV swarms.

With benefits of 454/Roche comes the cost of sequencing errors [84]. Homopolymers, or a succession of identical nucleotides, are a significant source of errors and arise due to non-linearity of signal for a sequencing reaction flow cycle to number of nucleotides incorporated. Errors can inflate sample diversity and is crucial to correct interpretation of rare variants. Sophisticated algorithms have been developed for de-noising pyrosequence data [85-90]. Of these the PyroNoise method developed by Quince *et al* is preferable as it uses raw flow data, which is the source of error, to interpret base calls. Other methods use sequence data, by which stage error may already be fixed. Reeder's method utilizes flow data, however assignment of reads to solution haplotypes is based on a simplistic greedy clustering algorithm. Although this improves computational efficiency, there is a cost of mis-assigning reads. PyroNoise has already been used to study viral populations in a

report describing HCV transmission. In following chapters it has now been applied extensively in the analysis HIV populations.

The need for multiple alignments of high-throughput 454 data has been addressed with development of efficient multiple sequence alignment (MSA) tools such as Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) and MUSCLE [91]. In chapter 2, a pipeline with PyroNoise followed by MSA based on a pair-wise alignment strategy is presented. This reduces error to levels comparable to SGA. In contrast to SGA, PCR re-sampling is another challenge inherent of the 454/Roche process as library preparation involves bulk PCR amplification of samples. This issue is also addressed in Chapter 2. Finally in all chapters, studies highlight the use of appropriate deeply sequenced control data to help in interpretation of rare biological signatures. Thus this dissertation illustrates that with careful experimental design and application of computational techniques, the power of deep sequencing can be reliably harnessed to gain significant insights into the nature of dynamic and diverse HIV swarms.

References

[1]. Jenkins GM, Rambaut A, Pybus OG, Holmes EC. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol* 2002; **54**:156-165.
[2]. Shackelton LA, Parrish CR, Truyen U, Holmes EC. High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proc Natl Acad Sci U S A* 2005; **102**:379-384.
[3]. Shackelton LA, Holmes EC. Phylogenetic evidence for the rapid evolution of human B19 erythrovirus. *J Virol* 2006; **80**:3666-3669.
[4]. Steinhauer DA, Holland JJ. Rapid evolution of RNA viruses. *Annu Rev Microbiol* 1987; **41**:409-433.
[5]. Biebricher CK, Eigen M. The error threshold. *Virus Res* 2005; **107**:117-127.

[6]. Eigen M, Gardiner W, Schuster P, Winkler-Oswatitsch R. The origin of genetic information. *Sci Am* 1981; **244**:88-92, 96, et passim.
[7]. Eigen M. Viral quasispecies. *Sci Am* 1993; **269**:42-49.
[8]. Biebricher CK, Eigen M. What is a quasispecies? *Curr Top Microbiol Immunol* 2006; **299**:1-31.
[9]. Eigen M. The origin of genetic information: viruses as models. *Gene* 1993; **135**:37-47.
[10]. Nowak MA, May RM, Anderson RM. The evolutionary dynamics of HIV-1 quasispecies and the development of immunodeficiency disease. *AIDS* 1990; **4**:1095-1103.
[11]. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* 2008; **105**:7552-7557.
[12]. Havlir DV, Strain MC, Clerici M, Ignacio C, Trabattoni D, Ferrante P, et al. Productive infection maintains a dynamic steady state of residual viremia in human immunodeficiency virus type 1-infected persons treated with suppressive antiretroviral therapy for five years. *J Virol* 2003; **77**:11212-11219.
[13]. Palmer S, Wiegand AP, Maldarelli F, Bazmi H, Mican JM, Polis M, et al. New real-time reverse transcriptase-initiated PCR assay with single-copy sensitivity for human immunodeficiency virus type 1 RNA in plasma. *J Clin Microbiol* 2003; **41**:4531-4536.
[14]. O'Connell KA, Brennan TP, Bailey JR, Ray SC, Siliciano RF, Blankson JN. Control of HIV-1 in elite suppressors despite ongoing replication and evolution in plasma virus. *J Virol* 2010; **84**:7018-7028.
[15]. Eigen M. On the nature of virus quasispecies. *Trends Microbiol* 1996; **4**:216-218.
[16]. Hunter E. Viral Entry and Receptors. In: *Retroviruses.* Coffin JM, Hughes SH, Varmus HE, (editors). Plainview: Cold Spring Laboratory Press; 1997. pp.71-119.
[17]. Salazar-Gonzalez JF, Salazar MG, Keele BF, Learn GH, Giorgi EE, Li H, et al. Genetic identity, biological phenotype, and evolutionary pathways of

transmitted/founder viruses in acute and early HIV-1 infection. *J Exp Med* 2009; **206**:1273-1289.
[18]. Brumme ZL, Goodrich J, Mayer HB, Brumme CJ, Henrick BM, Wynhoven B, et al. Molecular and clinical epidemiology of CXCR4-using HIV-1 in a large population of antiretroviral-naive individuals. *J Infect Dis* 2005; **192**:466-474.
[19]. Doms RW. Unwelcome guests with master keys: how HIV enters cells and how it can be stopped. *Top HIV Med* 2004; **12**:100-103.
[20]. Telesnitsky A, Goff SP. Reverse Transcriptase and the Generation of Retroviral DNA. In: *Retroviruses.* Coffin JM, Hughes SH, Varmus HE, (editors). Plainview: Cold Spring Harbor Laboratory Press; 1997. pp.121-160.
[21]. Brown PO. Integration. In: *Retroviruses.* Coffin JM, Hughes SH, Varmus HE, (editors). Plainview: Cold Spring Harbor Laboratory Press; 1997. pp.161-203.
[22]. Rabson AB, Graves BJ. Synthesis and Processing of Viral RNA. In: *Retroviruses.* Coffin JM, Hughes SH, Varmus HE, (editors). Cold Spring Harbor: CSH Laboratory Press; 1997. pp.205-261.
[23]. Swanstrom R, Wills JW. Synthesis, Assembly, and Processing of Viral Proteins. In: *Retroviruses.* Coffin JM, Hughes SH, Varmus HE, (editors). Plainview: Cold Spring Harbor Laboratory Press; 1997. pp.263-334.
[24]. Richman DD. HIV chemotherapy. *Nature* 2001; **410**:995-1001.
[25]. Sarafianos SG, Marchand B, Das K, Himmel DM, Parniak MA, Hughes SH, et al. Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. *J Mol Biol* 2009; **385**:693-713.
[26]. Shafer RW, Schapiro JM. HIV-1 drug resistance mutations: an updated framework for the second decade of HAART. *AIDS Rev* 2008; **10**:67-84.
[27]. Cote HC, Brumme ZL, Harrigan PR. Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir. *J Virol* 2001; **75**:589-594.
[28]. Collier AC. Efficacy of combination antiretroviral therapy. *Adv Exp Med Biol* 1996; **394**:355-372.

[29]. Collier AC, Coombs RW, Schoenfield DA, Bassett RL, Timpone J, Baruch A, et al. Treatment of Human Immunodeficiency Virus Infection with Saquinavir, Zidovudine, and Zalcitabine. *N. Engl. J. Med.* 1996; **334**:1011-1017.
[30]. Collier AC, Coombs RW, Schoenfeld DA, Bassett R, Baruch A, Corey L. Combination therapy with zidovudine, didanosine and saquinavir. *Antiviral Res* 1996; **29**:99.
[31]. Shen L, Peterson S, Sedaghat AR, McMahon MA, Callender M, Zhang H, et al. Dose-response curve slope sets class-specific limits on inhibitory potential of anti-HIV drugs. *Nat Med* 2008; **14**:762-766.
[32]. Carpenter CC, Fischl MA, Hammer SM, Hirsch MS, Jacobsen DM, Katzenstein DA, et al. Antiretroviral therapy for HIV infection in 1996. Recommendations of an international panel. International AIDS Society-USA. *JAMA* 1996; **276**:146-154.
[33]. Carpenter CC, Fischl MA, Hammer SM, Hirsch MS, Jacobsen DM, Katzenstein DA, et al. Antiretroviral therapy for HIV infection in 1997. Updated recommendations of the International AIDS Society-USA panel. *JAMA* 1997; **277**:1962-1969.
[34]. Carpenter CC, Fischl MA, Hammer SM, Hirsch MS, Jacobsen DM, Katzenstein DA, et al. Antiretroviral therapy for HIV infection in 1998: updated recommendations of the International AIDS Society-USA Panel. *JAMA* 1998; **280**:78-86.
[35]. Lalezari JP, Eron JJ, Carlson M, Cohen C, DeJesus E, Arduino RC, et al. A phase II clinical study of the long-term safety and antiviral activity of enfuvirtide-based antiretroviral therapy. *AIDS* 2003; **17**:691-698.
[36]. Dorr P, Westby M, Dobbs S, Griffin P, Irvine B, Macartney M, et al. Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrob Agents Chemother* 2005; **49**:4721-4732.
[37]. Hare S, Gupta SS, Valkov E, Engelman A, Cherepanov P. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* 2010; **464**:232-236.

[38]. Maertens GN, Hare S, Cherepanov P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* 2010; **468**:326-329.
[39]. McColl DJ, Chen X. Strand transfer inhibitors of HIV-1 integrase: bringing IN a new era of antiretroviral therapy. *Antiviral Res* 2010; **85**:101-118.
[40]. Cherepanov P, Maertens GN, Hare S. Structural insights into the retroviral DNA integration apparatus. *Curr Opin Struct Biol* 2011; **21**:249-256.
[41]. Hatano H, Lampiris H, Fransen S, Gupta S, Huang W, Hoh R, et al. Evolution of integrase resistance during failure of integrase inhibitor-based antiretroviral therapy. *J Acquir Immune Defic Syndr* 2010; **54**:389-393.
[42]. Hare S, Vos AM, Clayton RF, Thuring JW, Cummings MD, Cherepanov P. Molecular mechanisms of retroviral integrase inhibition and the evolution of viral resistance. *Proc Natl Acad Sci U S A* 2010; **107**:20057-20062.
[43]. Fransen S, Gupta S, Danovich R, Hazuda D, Miller M, Witmer M, et al. Loss of raltegravir susceptibility by human immunodeficiency virus type 1 is conferred via multiple nonoverlapping genetic pathways. *J Virol* 2009; **83**:11440-11446.
[44]. Metifiot M, Marchand C, Maddali K, Pommier Y. Resistance to integrase inhibitors. *Viruses* 2010; **2**:1347-1366.
[45]. Hutter G, Nowak D, Mossner M, Ganepola S, Mussig A, Allers K, et al. Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N Engl J Med* 2009; **360**:692-698.
[46]. Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, et al. Homozygous Defect in HIV-1 Coreceptor Accounts for Resistance of Some Multiply-Exposed Individuals to HIV-1 Infection. *Cell* 1996; **86**:367-377.
[47]. Paxton WA, Liu R, Kang S, Wu L, Gingeras TR, Landau NR, et al. Reduced HIV-1 infectability of CD4+ lymphocytes from exposed-uninfected individuals: association with low expression of CCR5 and high production of beta-chemokines. *Virology* 1998; **244**:66-73.
[48]. Allers K, Hutter G, Hofmann J, Loddenkemper C, Rieger K, Thiel E, et al. Evidence for the cure of HIV infection by CCR5Delta32/Delta32 stem cell transplantation. *Blood* 2011; **117**:2791-2799.

[49]. Finzi D, Blankson J, Siliciano JD, Margolick JB, Chadwick K, Pierson T, et al. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med* 1999; **5**:512-517.
[50]. Woffendin C, Ranga U, Yang Z, Xu L, Nabel GJ. Expression of a protective gene-prolongs survival of T cells in human immunodeficiency virus-infected patients. *Proc Natl Acad Sci U S A* 1996; **93**:2889-2894.
[51]. Ranga U, Woffendin C, Verma S, Xu L, June CH, Bishop DK, et al. Enhanced T cell engraftment after retroviral delivery of an antiviral gene in HIV-infected individuals. *Proc Natl Acad Sci U S A* 1998; **95**:1201-1206.
[52]. van Lunzen J, Glaunsinger T, Stahmer I, von Baehr V, Baum C, Schilz A, et al. Transfer of autologous gene-modified T cells in HIV-infected patients with advanced immunodeficiency and drug-resistant virus. *Mol Ther* 2007; **15**:1024-1033.
[53]. Tamhane M, Akkina R. Stable gene transfer of CCR5 and CXCR4 siRNAs by sleeping beauty transposon system to confer HIV-1 resistance. *AIDS Res Ther* 2008; **5**:16.
[54]. DiGiusto DL, Krishnan A, Li L, Li H, Li S, Rao A, et al. RNA-based gene therapy for HIV with lentiviral vector-modified CD34(+) cells in patients undergoing transplantation for AIDS-related lymphoma. *Sci Transl Med* 2010; **2**:36ra43.
[55]. Wilen CB, Wang J, Tilton JC, Miller JC, Kim KA, Rebar EJ, et al. Engineering HIV-resistant human CD4+ T cells with CXCR4-specific zinc-finger nucleases. *PLoS Pathog* 2011; **7**:e1002020.
[56]. Perez EE, Wang J, Miller JC, Jouvenot Y, Kim KA, Liu O, et al. Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nat Biotechnol* 2008; **26**:808-816.
[57]. Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, et al. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* 2005; **435**:646-651.
[58]. Mitsuyasu RT, Merigan TC, Carr A, Zack JA, Winters MA, Workman C, et al. Phase 2 gene therapy trial of an anti-HIV ribozyme in autologous CD34+ cells. *Nat Med* 2009; **15**:285-292.

[59]. Baum C, Dullmann J, Li Z, Fehse B, Meyer J, Williams DA, et al. Side effects of retroviral gene transfer into hematopoietic stem cells. *Blood* 2003; **101**:2099-2114.
[60]. Hacein-Bey-Abina S, Garrigue A, Wang GP, Soulier J, Lim A, Morillon E, et al. Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J Clin Invest* 2008; **118**:3132-3142.
[61]. Wang GP, Berry CC, Malani N, Leboulch P, Fischer A, Hacein-Bey-Abina S, et al. Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial. *Blood* 2010; **115**:4356-4366.
[62]. Kohn DB, Bauer G, Rice CR, Rothschild JC, Carbonaro DA, Valdez P, et al. A clinical trial of retroviral-mediated transfer of a rev-responsive element decoy gene into CD34(+) cells from the bone marrow of human immunodeficiency virus-1-infected children. *Blood* 1999; **94**:368-371.
[63]. Levine BL, Humeau LM, Boyer J, MacGregor RR, Rebello T, Lu X, et al. Gene transfer in humans using a conditionally replicating lentiviral vector. *Proc Natl Acad Sci U S A* 2006; **103**:17372-17377.
[64]. Lu X, Yu Q, Binder GK, Chen Z, Slepushkina T, Rossi J, et al. Antisense-mediated inhibition of human immunodeficiency virus (HIV) replication by use of an HIV type 1-based vector results in severely attenuated mutants incapable of developing resistance. *J Virol* 2004; **78**:7079-7088.
[65]. Wang GP, Levine BL, Binder GK, Berry CC, Malani N, McGarrity G, et al. Analysis of Lentiviral Vector Integration in HIV+ Study Subjects Receiving Autologous Infusions of Gene Modified CD4+ T Cells. *Mol Ther* 2009;.
[66]. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 1977; **265**:687-695.
[67]. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008; **26**:1135-1145.
[68]. Gunthard HF, Wong JK, Ignacio CC, Havlir DV, Richman DD. Comparative performance of high-density oligonucleotide sequencing and dideoxynucleotide

sequencing of HIV type 1 pol from clinical samples. *AIDS Res Hum Retroviruses* 1998; **14**:869-876.
[69]. Gunthard HF, Wong JK, Ignacio CC, Guatelli JC, Riggs NL, Havlir DV, et al. Human immunodeficiency virus replication and genotypic resistance in blood and lymph nodes after a year of potent antiretroviral therapy. *J Virol* 1998; **72**:2422-2428.
[70]. Simmonds P, Balfe P, Peutherer JF, Ludlam CA, Bishop JO, Brown AJ. Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers. *J Virol* 1990; **64**:864-872.
[71]. Simmonds P, Balfe P, Ludlam CA, Bishop JO, Brown AJ. Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *J Virol* 1990; **64**:5840-5850.
[72]. Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, Bazmi H, et al. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol* 2005; **43**:406-413.
[73]. Salazar-Gonzalez JF, Bailes E, Pham KT, Salazar MG, Guffey MB, Keele BF, et al. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol* 2008; **82**:3952-3970.
[74]. Bailey JR, Sedaghat AR, Kieffer T, Brennan T, Lee PK, Wind-Rotolo M, et al. Residual human immunodeficiency virus type 1 viremia in some patients on antiretroviral therapy is dominated by a small number of invariant clones rarely found in circulating CD4+ T cells. *J Virol* 2006; **80**:6441-6457.
[75]. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res* 2006; **34**:e22.
[76]. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 2005; **309**:1728-1732.

[77]. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**:376-380.
[78]. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. *Genome Res* 2001; **11**:3-11.
[79]. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet* 2008; **24**:133-141.
[80]. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010; **11**:31-46.
[81]. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* 2007; **17**:1195-1201.
[82]. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, et al. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 2007; **35**:e91.
[83]. Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, et al. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2007; **2**:e197.
[84]. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007; **8**:R143.
[85]. Huse SM, Welch DM, Morrison HG, Sogin ML. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* 2010; **12**:1889-1898.
[86]. Zagordi O, Klein R, Daumer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* 2010; **38**:7400-7409.
[87]. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol* 2010; **17**:417-428.

[88]. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, et al. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 2009; **6**:639-641.
[89]. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 2011; **12**:38.
[90]. Reeder J, Knight R. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods* 2010; **7**:668-669.
[91]. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**:1792-1797.
[92]. Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of spontaneous mutation. *Genetics* 1998; **148**:1667-1686.

Chapter 2 SWITCHING BETWEEN RALTEGRAVIR RESISTANCE PATHWAYS ANALYZED BY DEEP SEQUENCING

This work has been published by:
Mukherjee R, Jensen ST, Male F, Bittinger K, Hodinka RL, Miller MD, Bushman FD.
AIDS 2011 Oct;**25**(16):1951-9.

Abstract

In this study we analyzed the pathways leading to resistance of HIV to the integrase (IN) inhibitor raltegravir (RAL). Three HIV-infected individuals exhibiting RAL resistance pathway switching were characterized using longitudinal analysis of viral samples from plasma. 454/Roche pyrosequencing was used to generate ~74,000 sequence reads from the IN coding region. Effects of error were controlled by denoising with Pyronoise, and by comparison to ~142,000 control reads from HIV_{NL4-3}. Viral lineages were modeled quantitatively using viral serial pathway analysis (vSPA). All three patients showed transitions from the N155H pathway to the Q148H/G140S pathway. Analysis with vSPA revealed complex pathways to the final genotype, probably involving both *de novo* mutation and recombination. No reads contained both the N155H and Q148H drug resistance mutations (DRMs), indicating that the double mutant is not a prominent intermediate, consistent with low fitness. To characterize possible drug resistant variants circulating prior to therapy, we sequenced ~70,000 reads from samples collected prior to initiating treatment. Although some pre-existing drug resistant variants were detected, N155H, the first major DRM present after initiating RAL therapy, was not detected. We conclude that the main DRMs are present at very low levels if at all prior to initiating therapy. We also outline general methods for deep sequence analysis of DRMs in longitudinal HIV samples.

Introduction

The high mutation rate of HIV can result in rapid development of drug resistance [1, 2]. For raltegravir (RAL), which inhibits DNA strand transfer by integrase (IN) [3-5], resistance mutations typically arise in the part of the IN gene encoding the catalytic domain [6-8]. Three codons can mutate to generate primary resistance mutations, which encode Y143R/C/H, Q148H/R/K and N155H. Each primary DRM is associated with a preferred set of accessory mutations [9, 10].

Patients who exhibited virologic failure following RAL treatment often switch from one resistance pathway to another [10, 11]. Specifically, the N155H pathway is commonly replaced by the Q148H/R/K pathway, resulting in reduced susceptibility to RAL and improved viral replication capacity [9, 10]. However, questions remain regarding the nature of the switch, such as whether resistance mutations were present before treatment, and the nature of intermediates during pathway switching.

Here we investigated three patients for whom conventional viral population genotyping identified an N155H to Q148H pathway switch. Longitudinal analysis of viral population evolution under RAL pressure was performed using 454/Roche pyrosequencing [12], which is well suited to tracking viral variants in complex populations [13-20]. We used the Pyronoise pre-clustering method for controlling error [21] and performed longitudinal analysis of serially sampled viral populations using vSPA [22] for rigorous lineage analysis. Particularly deep sequencing was also carried out on pre-treatment samples and controls to investigate the abundance of possible pre-existing resistance mutations.

Results

Participants studied

The three patients studied were failing their existing HAART regimen and were administered RAL as salvage therapy. All of them exhibited a drop in the viral load in the first few weeks of RAL treatment, followed by a rebound with development of resistance. Population genotyping with Sanger sequencing identified N155H as one of the first IN DRMs to appear, but later the Q148H mutant predominated. For each patient we studied a baseline sample, obtained just prior to RAL initiation, and several time points after initiating therapy (Tables 2.1 and 2.2).

Patient	Visit date	Viral load (copies/ml)	Month	# 454 reads			# Pyronoise OTUs		
				Total	Fwd	Rev	Total	Fwd	Rev
1	04/04/06	34300	0	293	162	131	22	13	9
	07/05/06	8500	3	450	234	216	22	4	18
	08/02/06	48100	4	305	193	112	11	9	2
2	07/11/06	40700	0	410	251	159	28	21	7
	09/05/06	3630	2	28	14	14	12	6	6
	10/03/06	16700	3	168	93	75	16	9	7
	10/31/06	11400	4	489	250	239	13	9	4
	11/28/06	31300	5	304	190	114	11	9	2
	03/20/07	66600	8	1012	562	450	23	17	6
	07/05/07	31100	12	550	307	243	16	13	3
3	05/30/06	36800	0	32	16	16	14	5	9
	08/29/06	11000	3	19	12	7	9	4	5
	10/24/06	31700	5	29	16	13	2	1	1
	12/27/06	50000	7	5	3	2	3	2	1
	05/22/07	95100	12	103	50	53	6	3	3
Controls									
RNA	-	50000	-	179	87	92	2	1	1
DNA	-	-	-	264	143	121	3	1	2

Table 2.1 Plasma samples and controls studied in the low depth first pass experiment using 454/Roche GS FLX pyrosequencing.

Sequencing was performed in both directions – forward (Fwd) and reverse (Rev).

Longitudinal samples assayed using pyrosequencing

HIV RNA was extracted from each plasma sample, reverse-transcribed, and PCR amplified to examine bases 3906 through 4288 (numbering of the HIV-1_{NL4-3} genome). This corresponds to IN codons 45-171, which span the reported major DRM sites. PCR products were purified and sequenced using 454/Roche pyrosequencing. RNA and DNA controls from HIV-1_{NL4-3} were also included in this experiment. All samples analyzed are summarized in Table 2.1.

Control of error is central to studies of the low abundance resistance mutations. Error in this study could originate from the reverse transcription step, the PCR step, or the sequence determination step. We used Pyronoise for control of sequencing error [21]. Pyronoise operates in a Bayesian framework to cluster sequence reads as light intensity values (flowgrams) prior to interpretation as base calls. The resulting clusters are termed operational taxonomic units or OTUs. The number of Pyronoise OTUs was then interpreted as the number of unique variants at each time point (Table 2.1). No single OTU occurred in more than one patient and no HIV-1_{NL4-3} control OTU was detected in any patient. Thus there was no detectable contamination between samples.

To determine the optimal parameters for use in the Pyronoise analysis, we generated a first set of parallel HIV-1_{NL4-3} controls (443 reads; Table 2.1), wherein the correct number of OTUs is known to be one (Figure 2.1a,b). Under the parameters used, the two HIV-1_{NL4-3} RNA read sets (one each for forward and reverse direction sequencing) each yielded one OTU, whereas the HIV-1_{NL4-3} DNA sets yielded one and two OTUs. Examination of the output showed that an artifact at the extreme edge of the sequence reads led to the formation of two OTUs in one DNA control set. Thus we conclude that Pyronoise processing resulted in effective though not perfect control of inflation of OTU numbers due to sequencing error. A much larger set of 141,582 HIV-1_{NL4-3} reads was generated for comparison to deep analysis of pretreatment time points and is discussed below.

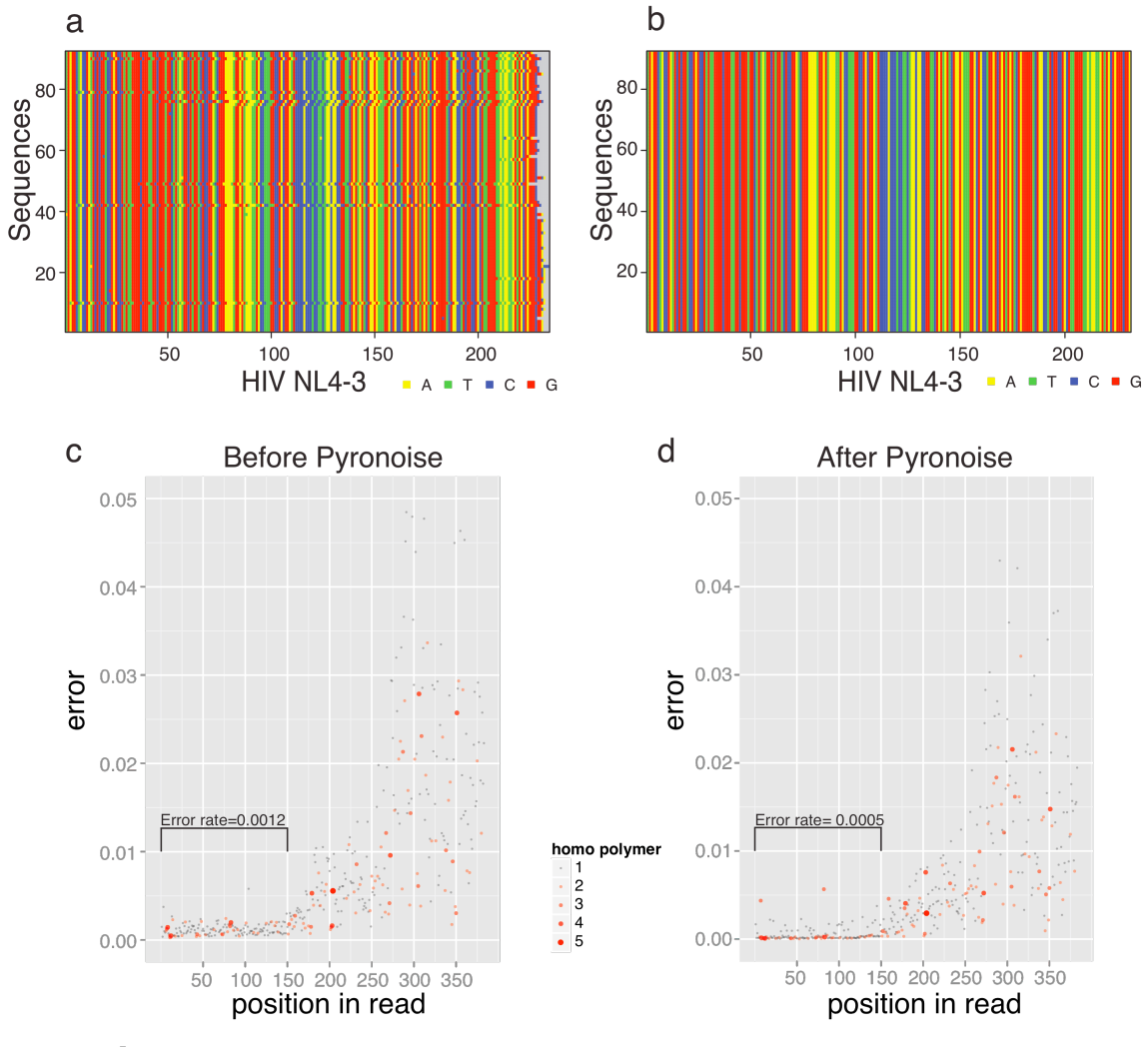


Figure 2.1 Results of denoising control data using Pyronoise.

RNA from HIV-1_{NL4-3} particles was purified, amplified by RT-PCR, and analyzed by 454/Roche pyrosequencing. After filtering for error-prone reads, sequences were denoised using Pyronoise, which pre-clusters reads based on light intensity values for each flow prior to interpretation as base calls. Alignments of a subset of reads are shown before denoising (**a**) and after denoising (**b**). The effects of denoising were also compared at each position in the HIV-1_{NL4-3} sequences studied (**c** and **d**). Shown are sequences from the forward reads. The base position is shown on the x-axis and the error rate is shown on the y-axis. Positions are assigned a homopolymer index of 1-5 (refer key) based on their location on a homopolymer. For example, a position with index 3 carries a base that occupies the 3rd place on a homopolymer with a minimum length of 3. Results are compared before denoising (**c**) and after denoising (**d**). The error rates in the high quality parts of the reads were 1.2e-3 before denoising and 5e-4 after denoising.

We estimated the error for the reverse transcription step to be 2e-4 base substitutions per nucleotide by comparing the HIV-1_{NL4-3} DNA and RNA sets. For PCR, the error rate (estimated by the manufacturer) is 1e-5 base substitutions per nucleotide per replication cycle, which results in 1.75e-4 base substitutions per nucleotide over 35 cycles of PCR. Thus pre-sequencing misincorporations add up to 3.75e-4 per nucleotide. Comparison to the measured overall rate leaves ~ 8e-4 base substitutions per nucleotide resulting from the pyrosequencing procedure. This is applicable to the highest quality portion of the reads over which such errors are uniform. In the 3' part of the reads, the error rate rises sharply, as illustrated in Figure 2.1c,d with data generated from re-sequencing of HIV_{NL4-3}. The presence of homopolymers had no influence on base substitution frequency with our alignment protocol. Our pyrosequencing error is slightly lower than that in previous studies, which used full reads and not just high quality regions [14, 15, 23].

A multiple sequence alignment (MSA) was constructed with denoised OTUs. Positions of codon polymorphisms within each patient over time are shown in Figure 2.2. The sequence reads were then translated *in silico*, and the known IN DRM sites were abstracted to create an IN resistance amino acid profile for each patient (Figure 2.3).

Longitudinal analysis of drug resistance pathways

In the pyrosequence data, the majority of early primary mutations encoded N155H in all three patients, whereas the majority of later primary mutations encoded Q148H, confirming the pathway switch inferred from population genotyping. There was no evidence of any double mutant variants harboring both 155 and 148 pathway primary mutations. All DRMs detected by population genotyping (Table 2.1) were also detected by deep sequencing. A methodological concern is that recombination during RT-PCR *in vitro* may link DRMs artifactually, but no recombinants encoding both substitutions were observed. We did not find evidence

of pre-existing IN inhibitor-related primary mutations (positions 143, 148 and 155) prior to initiating therapy in this first pass analysis.

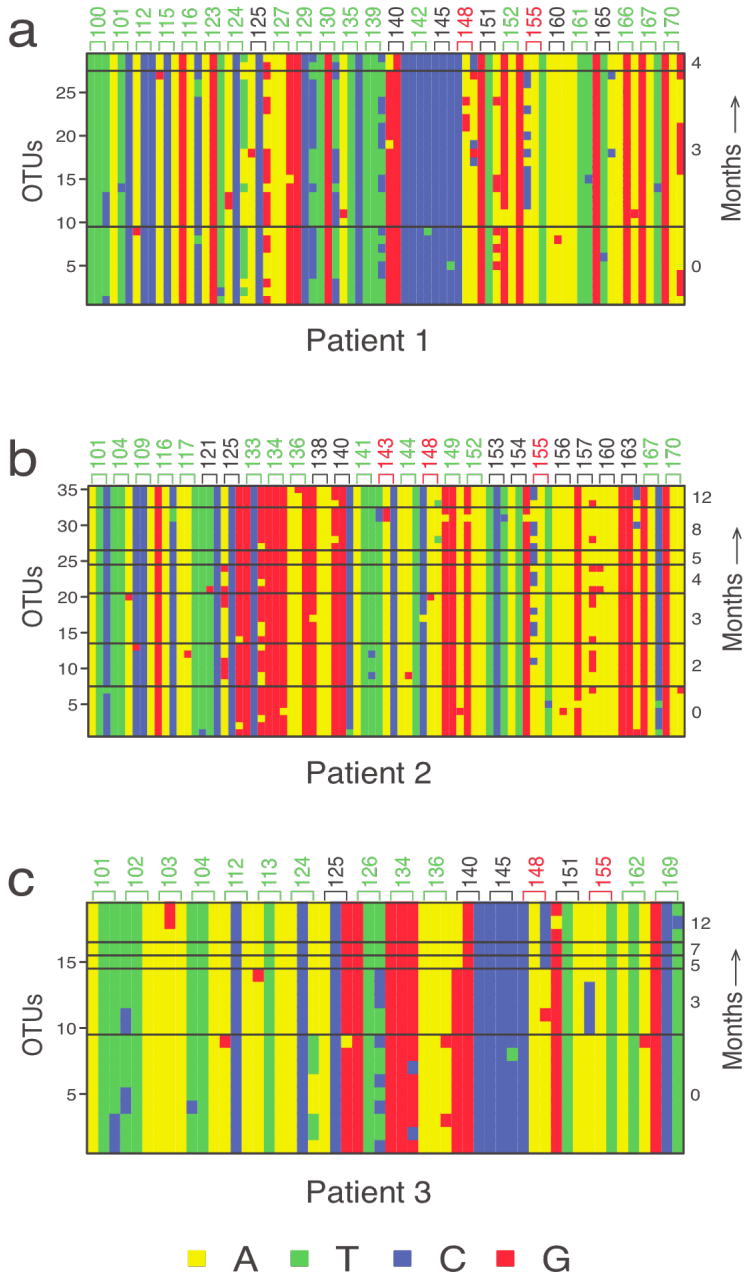


Figure 2.2 Sequence analysis of HIV populations in three patients treated with RAL and undergoing pathway switches from N155H to Q148 + G140S.

Nucleic acid sequences are shown for the three participants: (**a**) patient 1; (**b**) patient 2; (**c**) patient 3. For economy of display, only codons with polymorphisms over the time-course studied are shown. The display shows the sequence of each OTU detected without reference to their abundance. The

numbers to the left of each panel show the cumulative number of OTUs, the numbers to the right indicate the number of months since initiating therapy. A key to nucleic acid designations is shown at the bottom.

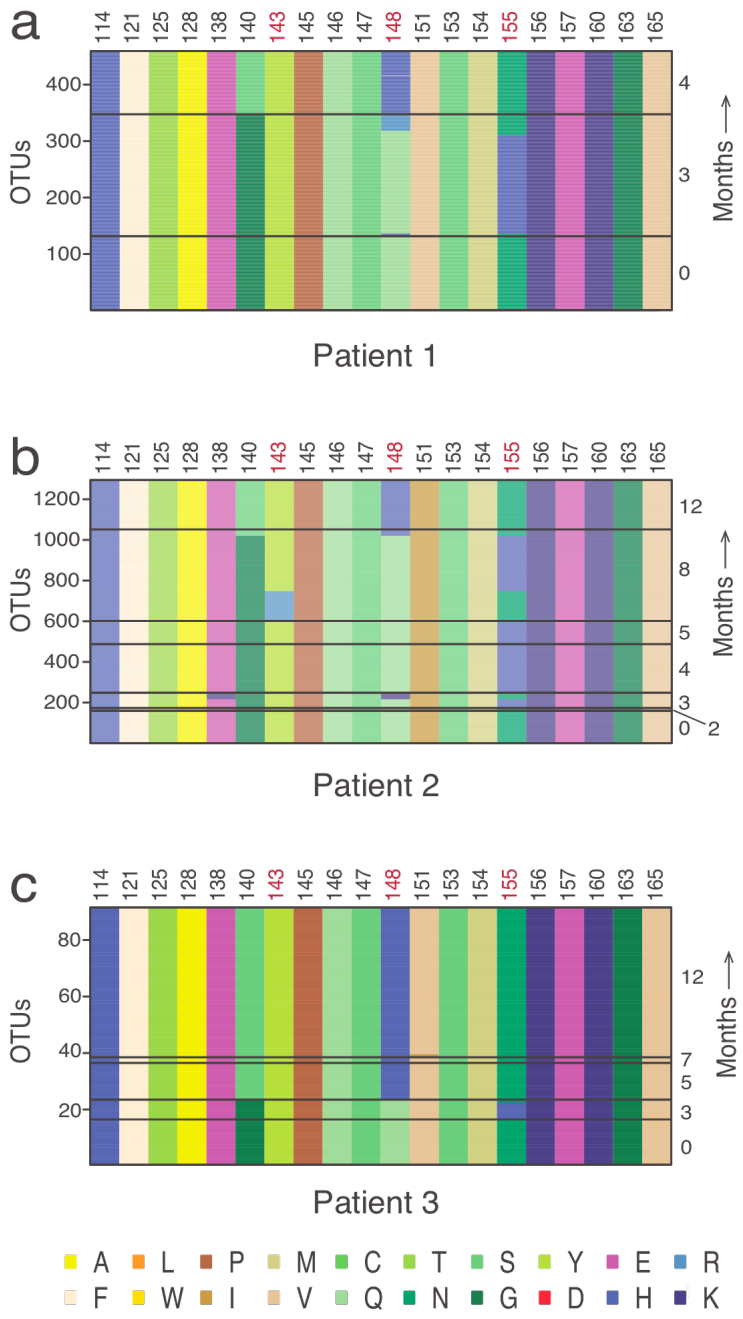


Figure 2.3 Inferred amino acid substitutions for each patient at DRM positions.

Amino acid profiles of OTUs for each subject were generated by conceptual translation: (**a**) patient 1;

(**b**) patient 2; (**c**) patient 3. For economy of display, only those codons encoding amino acid residues implicated in drug resistance are shown. IN DRM positions are listed along the x-axis. The profiles are grouped horizontally by time-point. The width of each horizontal bar indicates the total number of sequences recovered with each amino acid profile.

To track the evolution of drug-resistance lineages in a rigorous fashion, we used the vSPA algorithm [22]. For each sequence, a normalized distance vector over all other sequences is used to construct a correlation matrix. Sequences with more than a threshold correlation coefficient are clustered together based on a distribution of such matrices obtained from permuted datasets, and clusters across serial samples are linked based on average genetic distance to yield a longitudinal phylogenetic network (Figure 2.4).

In patient 1 at month 3 after initiating treatment, N155H predominated, though there were rare variants with Q148R and Q148H present (Figures 2.2a and 2.4a). Some but not all of the Q148H codons were associated with G140S (middle of Month 3 panel in Figures 2.2a and 2.4a). Even though the most common substitution at position 148 at month 3 encodes Q148R, it does not occur together with any accessory mutations at codons 138 or 140. Two separate lineages were detected at all three time-points, distinguished by polymorphisms at codons encoding amino acids 124, 125, 129, 130, and 139 (Figures 2.2a and 2.4a). Each of the collections of DRMs (N155H, Q148R, and Q148H + G140S) was found on both backgrounds. For most of the mutations, it is simplest to assume that the mutations arose once and recombined onto the different backgrounds. However, for the G140S mutation, the codon is directly adjacent to the polymorphic codon 139, so in this case recombination would need to break exactly between the two codons to generate the observed genotypes. Thus, independent mutations to generate the G140S substitution on the two backgrounds seem more likely.

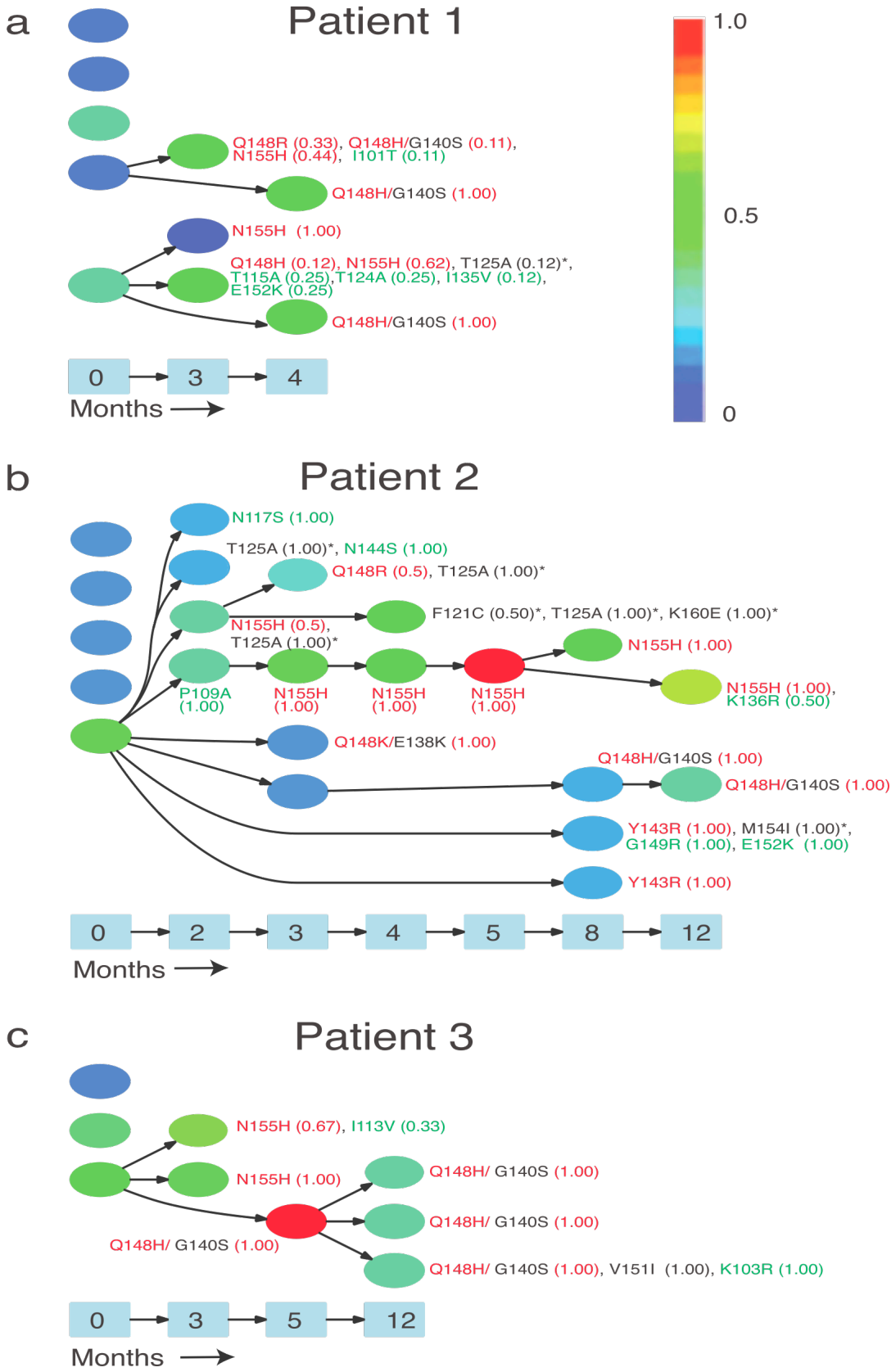

a
Patient 1
1.0
0.5
0
Q148R (0.33), Q148H/G140S (0.11),
N155H (0.44), I101T (0.11)
Q148H/G140S (1.00)
N155H (1.00)
Q148H (0.12), N155H (0.62), T125A (0.12)*,
T115A (0.25),T124A (0.25), I135V (0.12),
E152K (0.25)
Q148H/G140S (1.00)
0
3
4
Months
b
Patient 2
N117S (1.00)
T125A (1.00)*, N144S (1.00)
Q148R (0.5), T125A (1.00)*
F121C (0.50)*, T125A (1.00)*, K160E (1.00)*
N155H (0.5),
T125A (1.00)*
N155H (1.00)
N155H (1.00),
K136R (0.50)
P109A
(1.00)
N155H
(1.00)
N155H
(1.00)
N155H
(1.00)
Q148K/E138K (1.00)
Q148H/G140S (1.00)
Q148H/G140S (1.00)
Y143R (1.00), M154I (1.00)*,
G149R (1.00), E152K (1.00)
Y143R (1.00)
0
2
3
4
5
8
12
Months
c
Patient 3
N155H (0.67), I113V (0.33)
N155H (1.00)
Q148H/ G140S (1.00)
Q148H/ G140S (1.00)
Q148H/ G140S (1.00)
Q148H/ G140S (1.00), V151I (1.00), K103R (1.00)
0
3
5
12
Months

Figure 2.4 Evolutionary network of mutations following RAL treatment inferred using vSPA

(**a**) patient 1; (**b**) patient 2; (**c**) patient 3. The color of the cluster corresponds to the frequency of its membership from among sequences in that time-point (see color code). vSPA reports all mutations relative to a baseline (in this case month 0) ancestral cluster that arise at a level 2x or more in a descendent cluster. Primary mutations are in red, accessory ones in black and others in green (shown as the encoded amino acids). Mutations marked with an asterisk are not known to be associated with any primary DRM. Clusters can contain one or many OTUs. The proportion of OTUs within a cluster carrying a particular mutation is indicated in parenthesis.

Patient 2 also showed N155H switching to Q148H + G140S, but N155H was still detected at low abundance even after 8 months of therapy, and a complex collection of intermediates were detected over the period sampled. After three months of therapy, N155H, Q148K and Q148R all coexisted (Figures 2.2b and 2.4b). The Q148K + E138K combination was evident at month 3, and though this combination is reported to be a potent RAL escape variant [8] , it was not detected subsequently. By month 4, only the N155H variants were detected, whereas by month 8 Y143R, Q148H and N155H all were detected. At month 12, Q148H was the majority but N155H was still detectable, whereas Y143R was not. Patient 2 was the only subject where Y143R and N155H were detectable at later time-points. Tracking the origin of drug resistance lineages using vSPA indicated that all primary DRMs derived from a single ancestral cluster present before initiation of therapy. We also detected T97A, an accessory mutation for Y143R, and L74M and E92Q, accessory mutations for N155H, from the forward read sets (data not shown).

In Patient 3, three clusters were detected at time 0, one of which went on to give rise to the N155H drug resistant lineages by month 3. A different lineage derived from the same ancestral group emerged at month 5 and acquired the Q148H + G140S mutations, which persisted thereafter. Q148H replaced N155H after 5 months.

In depth analysis of pre-treatment time points

An important question in understanding the origin of RAL resistance is whether resistance mutations were present in the viral population prior to initiation of therapy. We thus carried out deeper sequencing analysis of the pre-treatment time-point. RT-PCR products were sequenced in both directions, yielding 69,862 sequence reads, ranging from ~18,000 to ~26,000 for each of the three participants.

Patient #	Viral load copies/ml	Pre-RT-PCR # genomes "n"	Read direction	# 454 reads "s"	Denoised # OTUs	Sampling statistics			
						Prop independent E(Y/s)		Variance $Var(Y)/s^2$	
						Simulations	Formula	Simulations	Formula
1	34300	27440	Forward	19483	3374	0.7162	0.7160	4.21E-06	5.66E-06
			Reverse	6666	1249	0.8879	0.8878	1.33E-05	1.22E-05
2	40700	32560	Forward	12824	2574	0.8267	0.8266	7.37E-06	8.00E-06
			Reverse	4815	1413	0.9301	0.9296	8.92E-06	1.20E-05
3	36800	29440	Forward	18558	2512	0.7422	0.7418	4.73E-06	6.01E-06
			Reverse	7516	1223	0.8828	0.8826	1.24E-05	1.11E-05

Table 2.2 Pre-treatment plasma samples studied in depth by 454/Roche Titanium pyrosequencing.

Viral load values for the baseline time-point for each patient are indicated. "n" gives the estimate of the number of starting templates in the RT-PCR reaction. "s" represents the sampling size – sample is drawn from a pool of amplified templates (here 35 cycles of PCR). Sampling statistics provides an estimate of numbers of starting templates actually sampled. The mean proportion of independent templates assayed ('Prop Independent') is indicated along with its variance. Estimates from both simulation studies and formulae described to assess re-sampling are tallied (see text for details of simulation and formulae).

In designing such a study, it is desirable to avoid sequencing multiple PCR products derived from a single viral genome, as this would waste sequencing effort and potentially confuse the analysis. To minimize this potential problem, the number of viral RNA templates in each RT-PCR reaction was arranged to be greater than the number of pyrosequence reads ultimately determined (Table 2.2). For each sample, 800μl of plasma was used for RNA purification, and all was used for RT-PCR, so the

predicted numbers of templates (assuming full recovery of RNA) were 27440, 32560, and 29440 for patients 1-3 respectively. The numbers of sequences for each sample were lower, 24937, 16461, and 24943 respectively, and distributed between the forward and reverse directions (Table 2.2).

An important question centers on how many of our sequence reads corresponded to different viral genomes present in the initial pre-amplification sample, and how many were duplicates generated during PCR. We devised a general formula relating the number of viral genomes in the starting sample to the number observed in the sequence output (Methods section). For the three pre-treatment samples, we calculated 72-93% of the sequence reads corresponded to independent viral genomes from the starting RNA sample (Table 2.2).

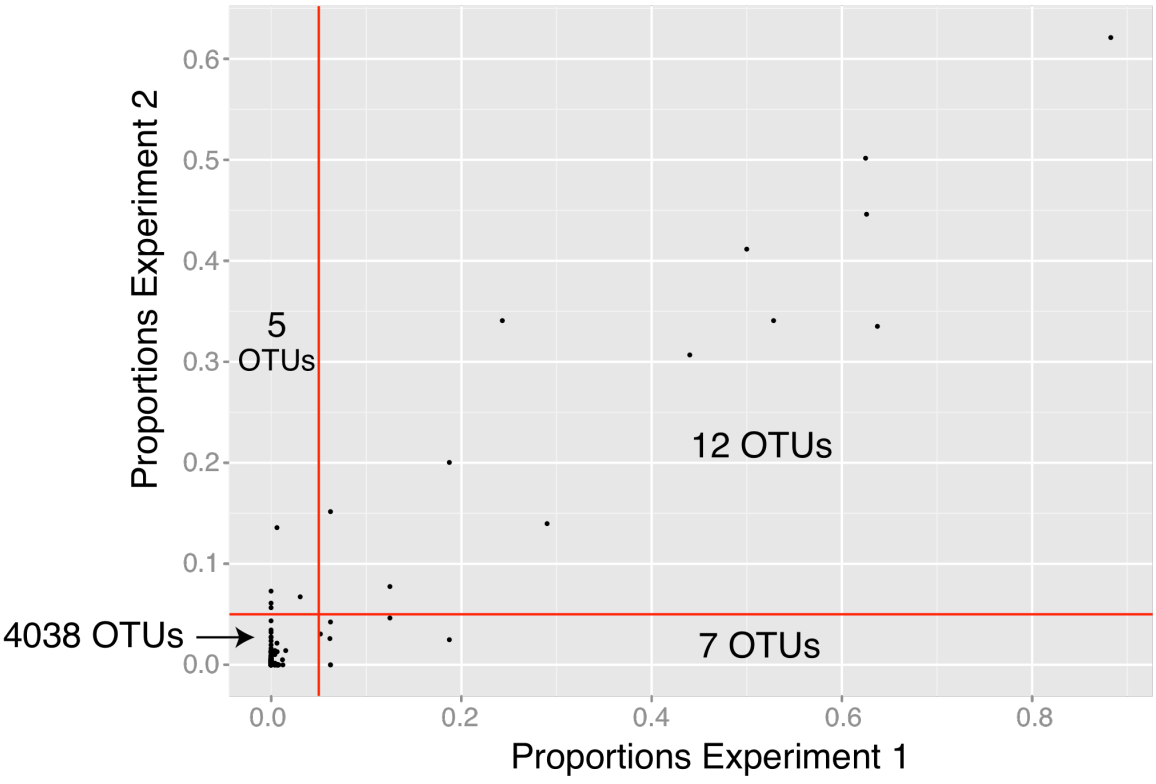


Figure 2.5 Reproducibility of OTU recoveries in duplicate analyses of the pre-treatment samples.

In this study, the pre-treatment time-points were analyzed twice, once in the low depth first pass analysis, then again in our deeper re-sequencing study, allowing a comparison of the OTUs recovered. The x- and y-axes show the relative proportion of each OTU in the two experiments. The red lines indicate the 5% abundance cut-off. To compare consistency, we asked for how many OTUs did the placement in <5% or >5% abundance cut-off groups agree (that is, in how many cases were samples in the lower left or upper right quadrants, indicating concordance). 4,050 OTUs showed concordance, or 99.7% of the total.

We next investigated the numbers of variants present in the pre-treatment samples. To assess reproducibility, the results from the first pass low coverage analysis of the pre-treatment samples were compared to the second much deeper analysis. OTU sets were reconstructed and analyzed for the high accuracy 5' part of the reads for both experiments. We plotted the proportion of reads contributed by each OTU within each participant in the two experiments (Figure 2.5). OTUs mostly fell near a line from lower left to upper right, indicating rough reproducibility between experiments. To assess concordance, we divided OTUs into those in the upper 95% of abundance or lower 5% in each experiment (Figure 2.5, indicated by red lines), and asked how many OTUs were concordant between experiments (that is present, in the lower left and upper right quadrants). We found that 4,050 out of 4,062 or 99.7% of OTUs were concordant between experiments by this measure.

454/Roche pyrosequence determination is error prone, making identification of low-level DRMs challenging. We used two strategies to distinguish low-level authentic DRMs in patient samples from spurious DRMs arising from error. In the first, we determined an additional 141,582 reads for the same IN amplicon on a homogeneous HIV_{NL4-3} template, calculated the frequency of false calls of DRMs on HIV_{NL4-3}, and then compared these frequencies to those observed in patient baseline samples. In the second strategy, we measured the error frequencies in the HIV_{NL4-3} data, and used them to model the expected frequency of spurious DRMs in the patient samples. The first approach has the advantage of modeling all aspects of the sequence acquisition process, whereas the second method allows larger numbers of

in silico generated control sequences to be compared to the patient sequences. For both, only the reverse reads were used because these had the highest quality bases over the relevant 143-155 coding positions.

(a)

Patient		Y143C	Y143H	Y143R	Q148H	Q148K	Q148R	N155H
1	Raw Reads	0.5481	0.0808	1.0000	1.0000	0.1602	0.6102	0.1602
	PyroNoised	0.9277	0.0110	1.0000	1.0000	0.1647	1.0000	1.0000
2	Raw Reads	0.4480	0.1477	1.0000	0.1937	1.0000	0.7862	1.0000
	PyroNoised	0.7480	0.0221	1.0000	0.1234	1.0000	0.1234	1.0000
3	Raw Reads	0.5135	0.0010	1.0000	0.6915	0.1780	0.4659	1.0000
	PyroNoised	0.9958	0.0001	1.0000	0.1828	1.0000	0.0259	1.0000

Significance based on Fisher's exact test

Significance maintained after Bonferroni correction for multiple testing

(b)

Patient		Y143C	Y143H	Y143R	Q148H	Q148K	Q148R	N155H
1	Prop DRM	0.0009	0.0016	0.0000	0.0000	0.0002	0.0011	0.0002
	Expt OBS	**6**	**10**	**0**	**0**	**1**	**7**	**1**
	Sim Av	3	3	0	5	3	3	3
	Sim MAX	10	12	1	16	11	10	11
	P Value	0.0488	0.0005	1.0000	1.0000	0.9280	0.0194	0.9273
2	Prop DRM	0.0011	0.0015	0.0000	0.0004	0.0000	0.0008	0.0000
	Expt OBS	**5**	**7**	**0**	**2**	**0**	**4**	**0**
	Sim Av	2	2	0	4	2	2	2
	Sim MAX	10	10	1	14	10	9	9
	P Value	0.0487	0.0030	1.0000	0.8952	1.0000	0.1355	1.0000
3	Prop DRM	0.0010	0.0023	0.0000	0.0001	0.0001	0.0012	0.0000
	Expt OBS	**7**	**17**	**0**	**1**	**1**	**9**	**0**
	Sim Av	3	3	0	6	3	3	3
	Sim MAX	11	11	1	18	11	12	11
	P Value	0.0312	0.0000	1.0000	0.9974	0.9506	0.0045	1.0000

Prop DRM	Experimentally observed proportion of DRM codons of the indicated type
Expt OBS	Experimentally observed number of DRM codons of the indicated type
Sim Av	Average #DRM substitutions due to error from 10,000 simulations
Sim MAX	Highest #DRM substitutions in any of the 10,000 simulations
	Significance based on ranking *P* values from 10,000 simulations
	Significance maintained after Bonferroni correction for multiple testing

Table 2.3 Analysis of possible DRM substitutions present in subjects prior to initiation of RAL therapy.

(a) *P* values for the comparison of DRMs at codons 143, 148, and 155 in the pre-treatment samples to error rate measured with the HIV-1_{NL4-3} sequence controls. Analysis was carried out both before and after denoising with Pyronoise. (**b)** DRMs called as significantly enriched based on simulated frequencies of false calls. Simulations were carried out on pre-denoised data using a model error rate as described in the methods section.

Table 2.3a shows the *P* values for detection of the major DRMs at amino acids 143, 148, and 155 in pre-treatment patient samples when compared to the HIV_{NL4-3} empirical control. The results are compared for both the raw sequence data and denoised data. A few substitutions show potential significance, including Y143H in all patients and Q148R in patient 3 (Table 2.3a, pink and red shading). However, after correction for multiple comparisons, only Y143H in patient 3 maintained significance (Table 2.3a, red shading).

A similar analysis was carried out comparing the pre-treatment patient sequencing data to simulations of DRM accumulation due to error (Table 2.3b). For this, the error rate of 1.2e-3 was used to generate a set of codons equal to the number sampled for each patient at each codon. A total of 10,000 such simulations were carried out, and the relationship of the observed data to simulated data assessed. Following correction for multiple comparisons, Y143H attained significance in both patient 1 and 3, but no other positions survived the test for multiple comparisons.

In all three patients, N155H, which was the first DRM to arise after RAL therapy, as well as Q148H, which was the majority primary DRM after pathway switch, were not significantly enriched in the pre-treatment samples. Thus pre-existing mutations did not contribute detectably to treatment failure.

Discussion

We analyzed the evolutionary dynamics of RAL DRMs for three patients who showed a switch from the N155H to the Q148H pathway. We describe an analytical pipeline based on Pyronoise and vSPA that may be useful for longitudinal tracking of DRMs in the future, though we note that computational feasibility becomes an issue with larger data sets. The pathways leading to the final states were complex, probably involving multiple rounds of point mutation to generate new variants and recombination to assemble variants on a single genome.

Following our initial survey, we sequenced deeply into the pre-treatment time-point to assess whether genomes containing DRMs that became abundant after treatment were detectable before treatment. Comparisons to control sequences from HIV_{NL4-3} or to results of simulations showed that most DRMs were not convincingly detectable above the error background. We did find significant enrichment for Y143H in one patient by all measures, and possible enrichment in a second patient (marginally significant relative to HIV_{NL4-3} and significant relative to simulated data). The Y143C and Q148R substitutions were marginally significant by some measures but did not survive correction for multiple comparisons. Codoner *et al*. [23] also investigated RAL resistance using 454 sequencing, and reported detection of Y143H, Y143C and Q148R prior to initiation of therapy. Our findings taken together support the idea that Y143H, and possibly Y143C and Q148R, are authentic replication-competent polymorphisms that are present in viral populations in the absence of RAL.

One of the main questions in initiating this analysis was whether the first resistance mutation to arise, N155H, was present prior to initiating therapy. We found a single read for this DRM over all patients, which is readily attributable to error. Thus, the primary DRM that appeared immediately after treatment initiation was not detectable before treatment initiation, despite sequencing to a depth of ~19,000 reads in the reverse direction, wherein the DRM positions are of highest quality. Similarly for Q148H, the major form at the last time-point, only 3 reads were

detected, a number also attributable to error. This is consistent with the idea that viruses with N155H or Q148H substitutions are considerably less fit than wild-type, and so are very low in abundance in the absence of pressure from RAL.

Materials and Methods

Deep sequencing of viral populations

Plasma samples were obtained from patients at multiple time-points and sequenced (Tables 2.1 and 2.2). RNA was purified from 200 μl plasma using the MagNA Pure LC extraction system (Roche Applied Science, Indianapolis, IN). A plasmid clone encoding HIV-1_{NL4-3} was used as a control DNA, and particles were generated by transfection and used as control viral RNA. Composite primers (Table 2.4) made of 454 sequencing adapters, barcodes and HIV$_{int}$ primers were used to amplify patient RNA and controls by *One-step* Reverse Transcriptase-PCR (Qiagen, Valencia, CA) using *RNasin* RNase inhibitor (Promega, Madison, WI) over 35 cycles as follows: 1 × (30 min at 50°C), 1 × (15 min at 95°C), 35 × (1 min at 94°C; 1 min at 60°C; 1 min at 72°C), 1× (10 min at 72°C; maintained at 4°C). Amplified products were gel-purified with the *QIAquick* Gel Extraction Kit (Qiagen, Valencia, CA) using a vacuum manifold. Purified DNA was quantified using the *Quant-iT* PicoGreen dsDNA Assay kit (Molecular Probes, Invitrogen, Eugene, OR). Pooled amplicons were pyrosequenced using the 454/Roche platform at the University of Pennsylvania.

(a)

int **Primer**	**Sequence**		
	454 Adapter	Barcode	*int* Primer
IntF3880	GCCTCCCTCGCGCCATCAG	Listed	AATAGTAGCCAGCTGTGATAAATGTC
IntR4312	GCCTTGCCAGCCCGCTCAG	in (**b**)	TGCCATTTGTACTGCTGTYTTAAG

(b)

Barcode Sequence	Patient	Month
ACACACTG	1	0
CAGTCAGT	1	3
ACGACATC	1	4
TGAGTCAC	2	0
ATCGATGC	2	2
AGACACTC	2	3
CACTACAG	2	4
CGATATGC	2	5
CGTACGAT	2	8
ATATCGCG	2	12
GACACTCA	3	0
CTACGATG	3	3
GAGTACAG	3	5
GCATATCG	3	7
GCTACGTA	3	12
ACAGACTC	Control	RNA
ACTGCTGA	Control	DNA

Table 2.4 Oligonucleotides and barcodes used in this study.
(a) Composite primers used to amplify IN codons 45-171. For the HIV_{int} primers, 'IntF' and 'IntR' refer to the forward and reverse primers respectively whereas the following numbers refer to the starting genomic coordinates on the HIV_{NL4-3} template. The composite primer sequence is made up of 454 adapter, barcode and the actual *env*AS primer (from 5' to 3'). **(b)** List of unique barcodes with their 8 bp sequence and tagged patient/time-point combination.

Three sequencing experiments were carried out. One used the 454/Roche GS FLX platform to generate relatively shallow data over all samples (4,640 total reads). The second used the 454/Roche Titanium Junior platform to acquire deep data on the three pre-treatment time points (69,862 total reads). For this, approximately 800μl of plasma per sample was ultra-centrifuged at 45,000 rpm for 75 min at 7°C to pellet virus, and the pellet was re-suspended in 100μl of phosphate buffered saline (PBS) for RNA purification using the *Illustra RNAspin* kit (GE Healthcare, Buckinghamshire, UK) with RNA carrier. RT-PCR was performed in quadruplicate and amplified products purified and concentrated using *Agencourt AMPure XP*

(Beckman Coulter, Brea, CA) prior to gel-purification. In the third experiment, the control RNA sample obtained from HIV-1_{NL4-3} was sequenced using the 454/Roche Titanium Junior platform (141,582 reads).

For the analysis below, we assume that losses during the viral isolation and cDNA synthesis steps are negligible (personal communication from manufacturer). All pyrosequence reads are available from our laboratory.

Bioinformatic analysis

We required that the first 320 flows have fewer than four low quality flows as defined by the 454/Roche criteria (note that this is not the default filtering). All pyrosequence reads were filtered for exact matches to barcodes and primers, and to remove shorter reads (<200 bases long) and reads with >2 ambiguous base calls, which are often error-prone [24]. Output sequences were assigned to each patient/time-point combination using DNA bar codes embedded in the amplification primers.

Pyrosequence reads were denoised using Pyronoise with default settings, except for the parameter 'c' set to 0.02, which pre-clusters 454/Roche flowgrams at ~95% similarity prior to interpretation as base calls [21], here termed operational taxonomic units (OTUs). Sequences were then aligned to the reference HIV-1_{NL4-3} genome (GenBank accession: M19921) using the Needleman-Wunsch pair-wise global alignment (PWA) algorithm with a match score of 5, and gap opening and extension penalties of 20 and 0.5. A multiple sequence alignment (MSA) was created by parsing the individual PWAs. In the MSA, positions that carried a reference genome base also had high read coverage (>95%), indicative of correct alignment. Positions that did not harbor a reference base, and had low read coverage (<2%) were removed, reasoning that they were alignment artifacts. The final MSA thus consisted of the 383 base positions of HIV-1_{NL4-3} from the region

encoding IN amino acids 45-171. Comparison to MSAs constructed using MUSCLE [25] for the control reads showed that our alignment method described above yielded fewer substitution errors (data not shown).

For the first sequencing experiment, reads with inferred amino acid sequences with any indeterminate amino acid in the region of high coverage were discarded. For the second, sequences were instead truncated at a point where scanning 5'-3' along the read direction yielded a pair of indeterminate amino acids, if any. For both experiments, sequences with stop codons were removed, reasoning that these were either sequencing errors or mutant genomes that would not yield replication-competent descendents. None of this filtering was performed on the control sequences in the third experiment because this set was used for error control. Of 237 OTUs in the first experiment, 213 survived and were considered for in-depth analysis (Table 2.1). In the second, of 14,908 OTUs 12,345 remained and were analyzed for pre-treatment levels of mutations (Table 2.2). Of the 12,345 OTUs 1,928 were truncated. OTUs removed or truncated were either singletons or contained relatively few reads. IN amino acid positions classified as locations of DRMs were identified using the HIV Drug Resistance Database (http://hivdb.stanford.edu/cgi-bin/INIResiNote.cgi) and [6]. The amplified region studied here had 28 such positions.

For each patient, OTUs over all time-points (first experiment) were used to generate an evolutionary network of resistance pathways by vSPA using default settings [22] except that polymorphisms accumulating in descendant lineages at a level of twice or more that found at ancestral time points were reported. We found that *vSPA* does not operate on sequence sets less than 5 in size, so for time-points with OTU numbers <5, the sequence dataset was replicated to facilitate analysis while maintaining the proportions of variants in the population.

Simulation-based framework for reporting drug-resistance mutations

For the computational simulation of accumulation of incorrect DRM calls due to error, it was necessary to first estimate the base substitution rate from control data. For this, an MSA was obtained for control HIV-1_{NL4-3} RNA reads. The MSA was constructed as previously by removing positions of low sequence coverage that are error-prone and typically correspond to indels at homopolymers (defined as any sequence with sequential identical bases). Error rates were determined for the 5' proximal bases where the sequence is most accurate, yielding an error rate of 1.2e-3 base misincorporations per position. For comparison, the error rate for the DNA control, which does not involve the RT step, was measured at 1e-3 over the same region.

To assess whether the frequencies of DRMs in the pre-treatment samples were higher than expected by chance, we used a simulation approach in addition to comparison to empirical data. For each sample we carried out 10,000 simulations per primary DRM position. For each simulation, a distribution of codons was generated of the size of the total number of codons at that position by binomially modeling an error rate of 1.2e-3 per base of the consensus codon. The numbers of resulting DRMs by *in silico* translation were then tabulated and compared to the observed frequencies of DRMs in the patient samples. *P* values were assigned by determining the proportion of simulated values that were equal or more extreme in the direction of the alternative hypothesis. Because there were seven primary DRMs queried, and three patients studied, we required that *P* values be below $0.05/21=0.0023$ to qualify as significant after Bonferroni correction for multiple comparisons. Results of the simulation-based analysis are in Table 2.3b.

Computation was carried out on a Dell Power Edge Cluster with 32 cores. Simulations and statistical tests were carried out using R.

Re-sampling statistics for numbers of starting genomes assayed in sequencing experiments after PCR amplification

Given an initial sample of *n* viral genomes, amplified by *f* PCR cycles for an amplification factor 2^f, and sequenced to yield *s* sequences, how many of the *n* original genomes will be detected among the *s* sequences? Let this quantity be defined by a random variable Y.

It can be shown that Y has a mean, $E(Y) = n\left(1 - \left(1 - \frac{1}{n}\right)^s\right)$,

and variance, $Var(Y) = E(Y) - E(Y)^2 + n(n-1)\left(1 - 2\left(1 - \frac{1}{n}\right)^s + \left(1 - \frac{2}{n}\right)^s\right)$

These formulae assume sampling with replacement, which is reasonable given the large number of amplified genomes from which we are sampling. This approximation is validated by comparisons in each case with 100 sampling simulations without replacement for 35 PCR cycles (Table 2.2; values labeled 'Simulations' and 'Formula'). Further, the equations are independent of *f* in the limit of $2^f \to \infty$ which too was validated with simulations in the ranges of *n* and *f* studied (data not shown). These equations thus provide a concrete measure of the completeness of sampling and the extent of possible over-sampling due to over-sequencing. To calculate the proportion of sequence reads corresponding to independent viral templates in the starting plasma sample, we used $\frac{Y}{s}$, which has mean $\frac{E(Y)}{s}$ and variance $\frac{Var(Y)}{s^2}$. Oversampling is then simply $1 - \frac{Y}{s}$.

Acknowledgments

We thank Michael D Miller (Merck Research Laboratories, West Point, PA) for establishing the collaboration. We are grateful to Warren J Ewens (Department of Biology, University of Pennsylvania) and Shane T Jensen (Department of Statistics,

The Wharton School, University of Pennsylvania) for advice on statistics. We also thank members of the Bushman laboratory for help and suggestions, especially Kyle Bittinger for support with Pyronoise and Frances Male and Rebecca Custers-Allen for assistance with the 454/Roche Junior sequencing workup. This work was supported by a grant from Merck Research Laboratories and NIH grant R01 AI052845.

References

[1]. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* 1995; **267**:483-486.

[2]. Coffin JM, Hughes SH, Varmus HE. *Retroviruses.* Cold Spring Harbor: Cold Spring Harbor Laboratory Press; 1997.

[3]. Espeseth AS, Felock P, Wolfe A, Witmer M, Grobler J, Anthony N, et al. HIV-1 integrase inhibitors that compete with the target DNA substrate define a unique strand transfer conformation for integrase. *Proc. Natl. Acad. Sci. U S A* 2000; **97**:11244-9.

[4]. Grobler JA, Stillmock K, Hu B, Witmer M, Felock P, Espeseth AS, et al. Diketo acid inhibitor mechanism and HIV-1 integrase: implications for metal binding in the active site of phosphotransferase enzymes. *Proc Natl Acad Sci U S A* 2002; **99**:6661-6666.

[5]. Hazuda DJ, Felock P, Witmer M, Wolfe A, Stillmock K, Grobler JA, et al. Inhibitors of Strand Transfer That Prevent Integration and Inhibit HIV-1 Replication in Cells. *Science* 2000; **287**:646-650.

[6]. Myers RE, Pillay D. Analysis of natural sequence variation and covariation in human immunodeficiency virus type 1 integrase. *J Virol* 2008; **82**:9228-9235.

[7]. Cooper DA, Steigbigel RT, Gatell JM, Rockstroh JK, Katlama C, Yeni P, et al. Subgroup and resistance analyses of raltegravir for resistant HIV-1 infection. *N Engl J Med* 2008; **359**:355-365.

[8]. Goethals O, Clayton R, Van Ginderen M, Vereycken I, Wagemans E, Geluykens P, et al. Resistance mutations in human immunodeficiency virus type 1 integrase selected with elvitegravir confer reduced susceptibility to a wide range of integrase inhibitors. *J Virol* 2008; **82**:10366-10374.
[9]. Fransen S, Gupta S, Danovich R, Hazuda D, Miller M, Witmer M, et al. Loss of raltegravir susceptibility by human immunodeficiency virus type 1 is conferred via multiple nonoverlapping genetic pathways. *J Virol* 2009; **83**:11440-11446.
[10]. Fransen S, Karmochkine M, Huang W, Weiss L, Petropoulos CJ, Charpentier C. Longitudinal analysis of raltegravir susceptibility and integrase replication capacity of human immunodeficiency virus type 1 during virologic failure. *Antimicrob Agents Chemother* 2009; **53**:4522-4524.
[11]. Charpentier C, Karmochkine M, Laureillard D, Tisserand P, Belec L, Weiss L, et al. Drug resistance profiles for the HIV integrase gene in patients failing raltegravir salvage therapy. *HIV Med* 2008; **9**:765-770.
[12]. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**:376-380.
[13]. Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N. Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction. *J Comput Biol* 2010; **17**:417-428.
[14]. Zagordi O, Klein R, Daumer M, Beerenwinkel N. Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. *Nucleic Acids Res* 2010; **38**:7400-7409.
[15]. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res* 2007; **17**:1195-1201.
[16]. Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, Gharizadeh B, et al. Viral population estimation using pyrosequencing. *PLoS Comput Biol* 2008; **4**:e1000074.
[17]. Simen BB, Simons JF, Hullsiek KH, Novak RM, Macarthur RD, Baxter JD, et al. Low-abundance drug-resistant viral variants in chronically HIV-infected,

antiretroviral treatment-naive patients significantly impact treatment outcomes. *J Infect Dis* 2009; **199**:693-701.
[18]. Le T, Chiarella J, Simen BB, Hanczaruk B, Egholm M, Landry ML, et al. Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS One* 2009; **4**:e6079.
[19]. Johnson JA, Li JF, Wei X, Lipscomb J, Irlbeck D, Craig C, et al. Minority HIV-1 drug resistance mutations are present in antiretroviral treatment-naive populations and associate with reduced treatment efficacy. *PLoS Med* 2008; **5**:e158.
[20]. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, et al. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 2007; **35**:e91.
[21]. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, et al. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 2009; **6**:639-641.
[22]. Hasegawa N, Sugiura W, Shibata J, Matsuda M, Ren F, Tanaka H. Inferring within-patient HIV-1 evolutionary dynamics under anti-HIV therapy using serial virus samples with vSPA. *BMC Bioinformatics* 2009; **10**:360.
[23]. Codoner FM, Pou C, Thielen A, Garcia F, Delgado R, Dalmau D, et al. Dynamic escape of pre-existing raltegravir-resistant HIV-1 from raltegravir selection pressure. *Antiviral Res* 2010; **88**:281-286.
[24]. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007; **8**:R143.
[25]. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**:1792-1797.

Chapter 3 HIV SEQUENCE VARIATION ASSOCIATED WITH *ENV* ANTISENSE ADOPTIVE T CELL THERAPY IN THE hNSG MOUSE MODEL

This work has been published by:
Mukherjee R, Plesa G, Sherrill-Mix S, Richardson MW, Riley JL, Bushman FD.
Mol Ther 2010 Apr;**18**(4):803-11.

Abstract

The first use of lentiviral vectors in humans involved transduction of mature T cells with an HIV-derived *env* antisense (*env*AS) vector to protect cells from HIV infection. In that study, only a minority of the patient T cell population could be gene-modified, raising the question of whether the altered cells could affect replicating HIV populations. We investigated this using humanized NOD/SCID IL-2Rγ^{null} (hNSG) mice reconstituted with ~4-11% *env*AS-modified human T cells. Mice were challenged with HIV-1_{NL4-3}, which has an *env* perfectly complementary to *env*AS, or with HIV-1_{BaL}, which has a divergent *env*. No differences were seen in viral titer between mice that received *env*AS-modified cells and control mice that did not. Using 454/Roche pyrosequencing we analyzed the mutational spectrum in HIV populations in serum--from 33 mice we recovered 84,074 total reads comprising 31,290 unique sequence variants. We found enrichment of A-to-G transitions and deletions in *env*AS-treated mice, paralleling a previous tissue culture study where most target cells contained *env*AS, even though minority of cells were *env*AS-modified here. Unexpectedly, this enrichment was only detected following challenge with HIV-1_{BaL}, where the viral genome would form an imperfect duplex with *env*AS, and not HIV-1_{NL4-3}, where a perfectly matched duplex would form.

Introduction

Highly active anti-retroviral therapy (HAART) fails to eliminate HIV completely from patients and often elicits drug resistant variants, leading to interest in additional forms of therapy. Adoptive T cell therapy using gene-modified T cells is one such approach, in which T cells are harvested from HIV-infected subjects, transduced *ex vivo* with genes that obstruct HIV replication, then re-infused back into patients. One type of anti-HIV gene encodes antisense RNA, which in several studies has been shown to inhibit HIV replication efficiently in tissue culture [1-4].

In a phase I clinical trial preceding our study, an HIV-derived lentiviral vector encoding *env* antisense (*env*AS), called VRX496, was used to treat patients that failed two or more anti-retroviral regimens [5]. The trial was a success in that no adverse events were reported in this first-in-human use of lentiviral vectors [5, 6]. There was also significant reduction in viral loads and improvement of immune function in a subset of patients, even though only a fraction of autologous T cells were modified with VRX496. This then raised the question of whether *env*AS modification of a minority of cells could influence HIV infection.

In this study, we have investigated the effect of *env*AS on HIV-1 populations in a setting where only a minority of cells was vector-modified. The vector VRX494 used here is the research counterpart of the clinically used VRX496, the only difference being that VRX494 has a GFP gene that allows convenient quantification of transduction (Figure 3.1a). Similar to the clinical vector, VRX494 is derived from HIV and directs *env*AS expression from the HIV long terminal repeat (LTR), so that following HIV infection of a vector-modified cell and expression of HIV *tat*, the anti-sense payload is expressed. In addition, *cis* signals are present in the vector to allow mobilization by HIV proteins provided in *trans* during infection, so that the vector can potentially spread from cell to cell. In a computational simulation, formation of

such defective interfering particles, combined with antisense inhibition, inhibited HIV replication effectively [7].

The anti-sense RNA encoded by VRX494 can bind the HIV RNA genome, forming double-stranded RNA. In previous work, HIV infection of VRX494-modified cells in culture, followed by recovery of challenge virus sequences, showed 1) large deletions in most HIV genomes and 2) a high frequency of A-G transitions in the *env*AS targeted region, potentially the result of A-I editing by cellular double-stranded RNA adenosine deaminase (dsRAD; a member of the Adenosine Deaminase Acting on RNA, or ADAR, family of enzymes) [4]. Such modifications may help the virus evade anti-sense pressure by reducing the complementarity of the viral genome and the *env*AS. However, most of these deleted genomes would be replication defective, and a few A-to-G changes would be unlikely to significantly destabilize pairing in the ~900 bp anti-sense RNA used. Thus, it may be more likely that the accumulation of these sequences reports the action of cellular systems acting on the double-stranded sense-antisense RNA, but that these mutations do not confer reduced sensitivity to the anti-sense [4].

Here we report an analysis of HIV-1 populations following growth in the presence of *env*AS using a humanized NOD/SCID IL-2Rγ^{null} (or hNSG) mouse model. Cohorts of mice were compared that were reconstituted with ~4-11% vector-modified T cells or with unmodified T cells to model the frequency of vector marking in the human phase I trial. We found no reduction in viral load, but we did detect a significant excess of challenge virus variants with enriched A-to-G transitions and deletions, documenting antisense pressure on HIV-1 populations even with VRX494-modification of only a minority of cells. Unexpectedly, the variants with enriched A-G transitions and deletions only accumulated significantly when the antisense was imperfectly matched to its target.

Results

A mouse model for envAS pressure on HIV-1

Previous studies measuring the effect of HIV-1 antisense therapy have used T cell populations in which the vast majority of viral target cells expressed the antisense RNA. However, in the therapeutic setting, although most of the cells that are infused express the antisense message, they are quickly diluted and persist as only a small percentage of the total [5]. To study how T cells transduced with the VRX494 *env*AS vector (Figure 3.1a) might alter viral populations in patients, we established an *in vivo* model that would mirror the published clinical trial [5], but allow infection with defined HIV challenge stocks. This was done in collaboration with the laboratory of James L Riley at University of Pennsylvania. We took advantage of the ability of NOD/SCID IL-2Rγ^{null} (NSG) mice to stably engraft human T cells [8] to model HIV-1 infection and therapy. We established 4 cohorts of 10 mice each as diagrammed in Figure 3.1b. Two control cohorts were engrafted with untransduced T cells and challenged with either of two HIV-1 derivatives, HIV_{NL4-3} or an HIV_{NL4-3} derivative engineered to express the BaL envelope (henceforth "HIV_{BaL}"). The other two cohorts were engrafted with untransduced T cells spiked with ~4-11% of VRX494-transduced T cells and challenged similarly (termed "vector-treated cohort" below).

The VRX494-transduced T cells used in this study were generated in a manner that replicated the clinical trial. Freshly isolated primary human CD4+ T cells were activated, transduced with a lentiviral vector expressing *env*AS and GFP (VRX494), expanded for 10 days, frozen, and stored in liquid nitrogen for several weeks. After thawing, GFP expression was measured by flow cytometry and these cells were found to be highly transduced (Figure 3.1c), as with the cells used for treating patients in the clinical trial [5]. Mice in the control cohorts were injected with 10 million untransduced T cells, whereas mice in the vector-treated cohorts were

injected with 9 million untransduced T cells and 1 million of VRX494-transduced cells (thus the target frequency of transduced cells was 10%).

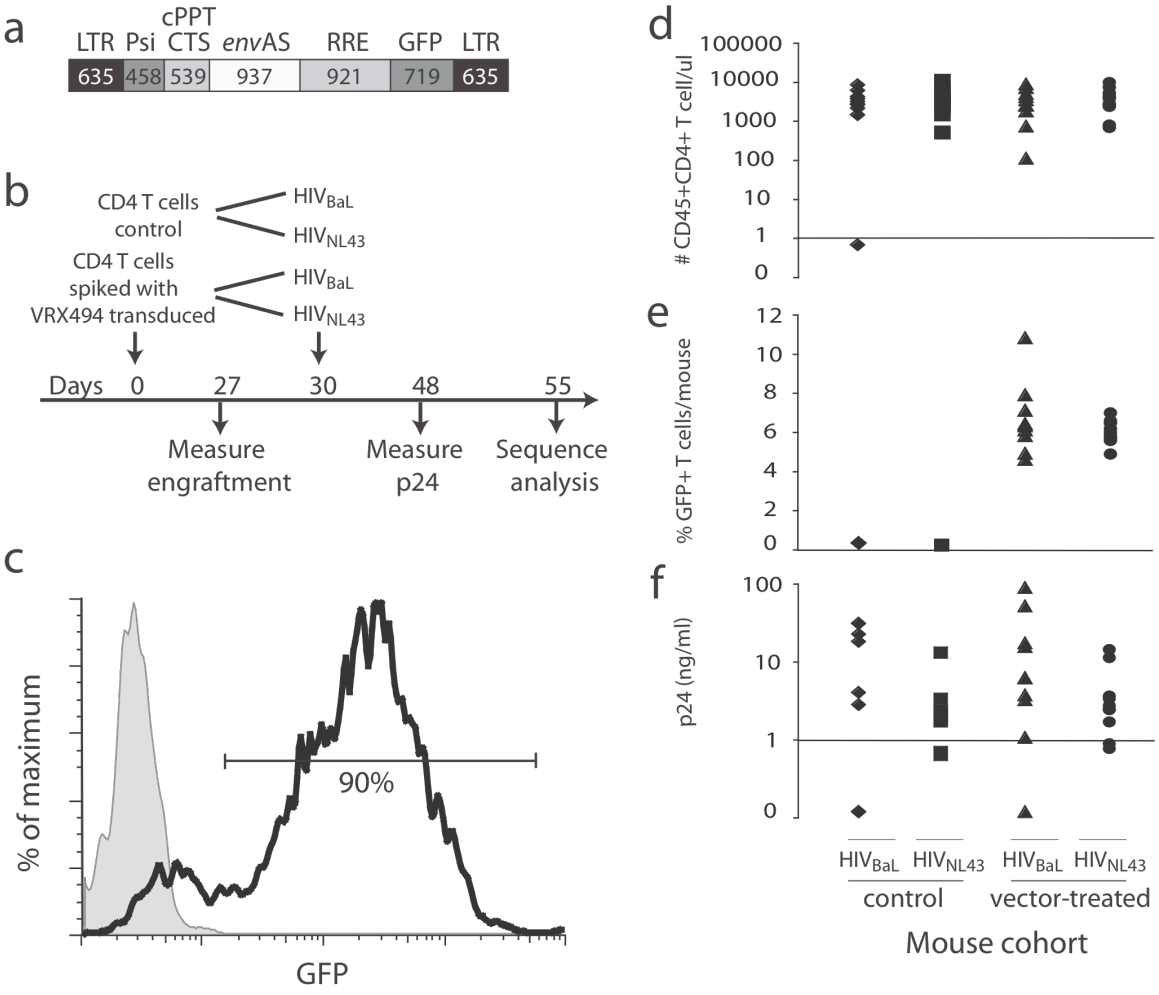


Figure 3.1 Analysis of hNSG mice with ~4-11% VRX494 *env*AS-vector modified T cells and challenged with HIV_{NL4-3} or HIV_{BaL}.

(a) The VRX494 vector. **(b)** Time line of mouse transplantation and infection. The hNSG mice were transplanted on day 0, then challenged with HIV_{NL4-3} or HIV_{BaL} on day 30. **(c)** Flow cytometry analysis of transduction of CD4 T cells with *env*AS. Cells with high transduction levels were obtained as evidenced by GFP detection. **(d)** Analysis of the levels of T cell engraftment. **(e)** Analysis of the levels of engraftment of VRX494-transduced cells, analyzed by sorting GFP-positive cells. **(f)** Analysis of the level of HIV p24 antigen in serum at 48 days after infection.

After 27 days, we measured the engraftment for each of the 40 animals. One animal failed to engraft T cells and was excluded from further analysis. For the remaining animals, similar levels of engraftment of total CD4+ T cells were seen for all cohorts (Figure 3.1d). The two cohorts that received VRX494-transduced cells showed similar distributions for the proportion of engrafted cells transduced with GFP-positive VRX494 (4-11%, Figure 3.1e).

Three days later (day 30 post-infusion of T cells) we challenged each mouse with supernatants collected from 293 T cells transfected with HIV_{BaL} or HIV_{NL4-3}. We note that this differs from the clinical trial in that the human subjects were HIV-infected before infusion of gene-modified cells, whereas here gene-modified cells were introduced into mice prior to HIV infection. After an additional 18 days (day 48 post-infusion), we measured viral replication by assaying p24 in the mouse serum (Figure 3.1f and Table 3.1a). We did not observe a significant difference in the viral titers between the vector-treated and control cohorts. This is consistent with the fact that it took ≥6 months to see substantial differences in the viral load of patients treated with the clinical *env*AS vector (VRX496). After an additional 7 days (day 55 post-infusion), some of the mice (across all four cohorts) began to display the early stages of graft versus host disease (GVHD) and the experiment was terminated. Serum obtained from the peripheral blood at the time of sacrifice was used for all subsequent analysis. Two mice died during the study, and four mice had undetectable viral loads by serum p24 assay, so a total of 33 mice were available for analysis. All work described in this section was conducted by members of the Riley lab.

Amplification and sequencing of viral quasi-species from mouse plasma

To investigate possible pressure on viral populations exerted by *env*AS, we studied the structure of viral populations using the 454/Roche pyrosequencing technology [9]. HIV RNA was extracted from serum of each mouse, reverse transcribed, and

PCR amplified according to the scheme in Figure 3.2. Three amplicons were designed covering the *env*AS targeted region of the challenge viruses – two covering each side of the *env*AS target region, and a third spanning the two. The outermost primers annealed slightly outside each edge of the *env*AS targeted region. The smaller two amplicons permit detection of A-G transitions and smaller deletions. The longer amplicon, at 1037 bases in length, is too long for use in the 454/Roche sequencing procedure, which accommodates molecules of a maximum ~500 bases in length. Thus the longer amplicon could only yield sequences with large deletions.

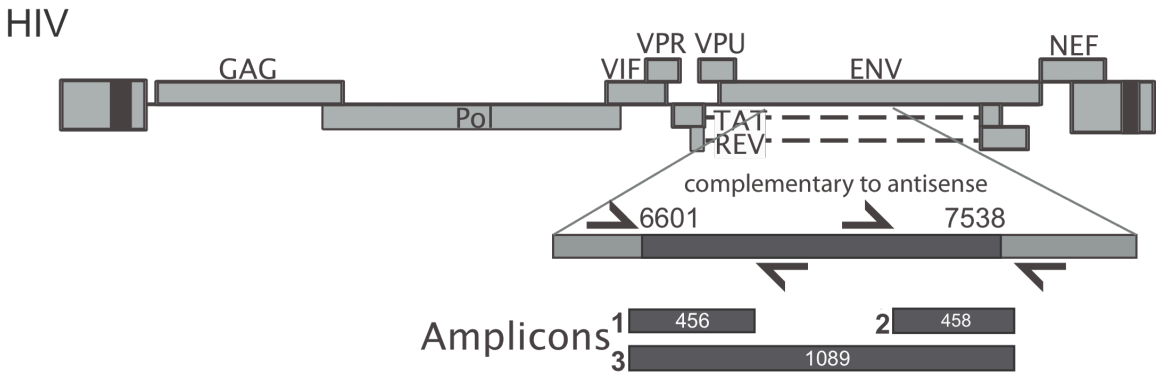


Figure 3.2 The HIV_{NL4-3} genome, showing the regions targeted by the VRX494 *env*AS, and the HIV *env* amplicons used in this study.

The numbering in the *env* gene refers to the HIV_{NL4-3} genome. The *env*AS targeted region extends from 6,601-7,538 (see blow-up). Three amplicons were designed to recover potential deletions and A-G changes. The relative positions and length of the amplicons (including 454 adapters and barcodes) are shown. Three amplicons were similarly designed in the homologous region of the HIV_{BaL} genome.

RNA from each serum sample was separately amplified and products gel-purified for each of the three amplicons. Gel regions corresponding both to full-length and shorter sequences were isolated to avoid selecting against potential deletions. To allow determination of sequences from many samples in pools, unique 8-nucleotide error-correcting barcodes [10] were incorporated into each primer to index PCR products from each mouse [6, 11-13]. Following pyrosequencing, sequence reads were assigned to the source mouse by decoding the barcode.

We analyzed a total of 84,074 sequences from 33 mice across 3 amplicons. Reads were filtered to require perfect matches to the 5' barcode and *env* primer [14], yielding 79,040 (or 94% of the total) sequences. Based on the distribution of read lengths recovered, additional filtering was performed to remove sequences with <220 bases, as short reads are reported to be error prone [15]. A total of 68,642 (or 82% of the total) sequences remained after filtering. Finally the sequence pools from each mouse were de-replicated and curated to yield a non-redundant set of 31,290 unique sequences. The distribution of these across the four study groups is summarized in Table 3.1. Subsequent analysis was carried out using the unique viral variants (rather than also considering their frequencies of isolation). Analysis involved comparison of vector-treated and control animals, so that the comparative framework minimized the influence of error arising due to mutagenesis in the PCR procedure or in pyrosequence read determination.

(a)

Control mice	**Challenge virus**	**Serum p24** (pg/ml)	**# 454 reads**	**Treated mice**	**Challenge virus**	**Serum p24** (pg/ml)	**# 454 reads**
1042	BaL	2775	151	947	BaL	1055	987
1045	BaL	22500	1248	985	BaL	89000	1517
1047	BaL	18050	1272	1037	BaL	3710	1318
1052	BaL	4008	1347	1089	BaL	15525	1701
1053	BaL	30900	899	1091	BaL	17250	1230
1085	BaL	30650	919	1095	BaL	51400	1292
1044	NL4-3	700	1312	1138	BaL	6120	151
1046	NL4-3	2200	1218	1140	BaL	6215	315
1048	NL4-3	13200	1171	946	NL4-3	2759	871
1050	NL4-3	3365	955	948	NL4-3	1700	880
1051	NL4-3	2142	531	1038	NL4-3	780	701
1055	NL4-3	1850	1489	1040	NL4-3	900	638
1086	NL4-3	642	668	1090	NL4-3	3675	888
1087	NL4-3	1735	1060	1094	NL4-3	2510	289
1092	NL4-3	2335	1188	1136	NL4-3	11450	793
1093	NL4-3	2050	1128	1137	NL4-3	2450	566
				1139	NL4-3	3380	597

(b)

Challenge	Vector-treated		Control	
	# Mice	# 454 Reads	# Mice	# 454 Reads
BaL	8	8511	6	5836
NL4-3	9	6223	10	10720

Table 3.1 Samples studied and numbers of pyrosequence reads.
(a) Samples with viral load (serum p24 assay) and number of unique pyrosequence reads. **(b)** Summary of distribution of unique pyrosequence reads across 4 study cohorts.

Analysis of nucleotide changes in VRX494-transduced mice

HIV sequences were then analyzed for enrichment for A-G transitions and deletions, which were previously found to be associated with replication in the presence of the *env*AS [4]. Each sequence was aligned to the reference HIV genome using pair-wise global alignment (PWA). A stringent gap opening penalty was selected to allow detection of deletions of up to ~900 bases. Use of a stringent gap opening penalty also allowed detection of multiple nearby mismatches, as can occur if *env*AS elicits clustered mutations. Two parsed multiple sequence alignments (MSAs), one for HIV_{BaL} and one for HIV_{NL4-3}, were constructed from the individual PWAs.

To investigate A-G transition rates, we trimmed off primer sequences, then calculated the proportion of nucleotide positions that changed from A-to-G for each sequence, yielding a value between 0 and 1. As controls, all other base change proportions per sequence were also calculated. For each base change, the distribution of proportions for the vector-treated and control groups were plotted (Figure 3.3). The boxes in the figure indicate the middle two quartiles (inter-quartile range; in most cases the intervals are so close that the box appears to be a line at 0.0). Outliers, defined as sequences with base change proportions beyond 1.5 times the inter-quartile range from the box edges, are plotted as open black circles.

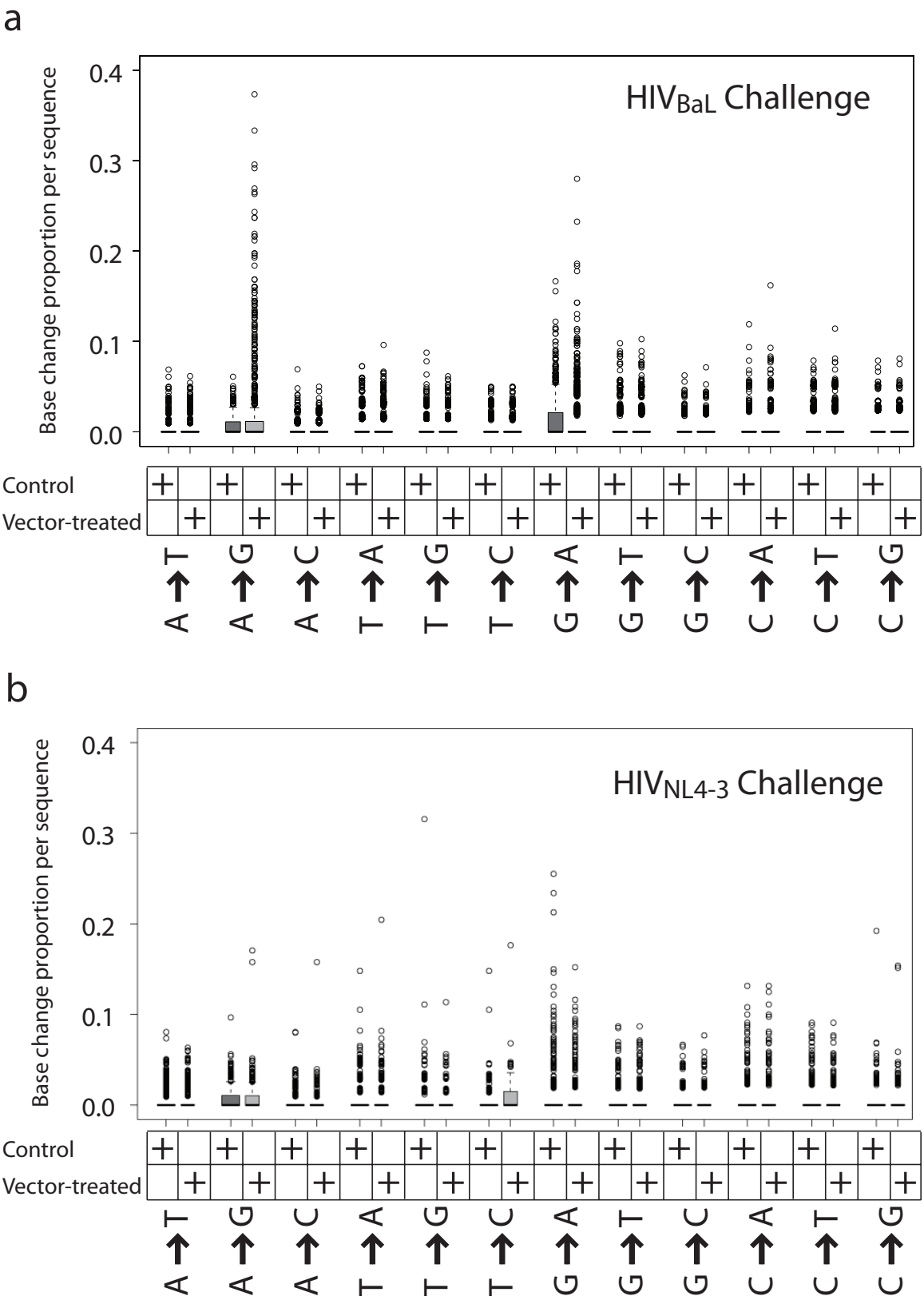

a
Base change proportion per sequence
0.4
0.3
0.2
0.1
0.0
HIV$_{BaL}$ Challenge
Control
Vector-treated
A→T
A→G
A→C
T→A
T→G
T→C
G→A
G→T
G→C
C→A
C→T
C→G
b
Base change proportion per sequence
0.4
0.3
0.2
0.1
0.0
HIV$_{NL4-3}$ Challenge
Control
Vector-treated
A→T
A→G
A→C
T→A
T→G
T→C
G→A
G→T
G→C
C→A
C→T
C→G

Figure 3.3 Box plots illustrating the types of base substitutions that accumulated during growth of HIV-1 in hNSG mice.

(a) Base substitutions that accumulate during growth of HIV_{BaL} in the vector-treated and control mice. **(b)** Base substitutions that accumulated during growth of HIV_{NL4-3} in the vector-treated and control mice. The boxes comprise all sequences with proportion changes in the middle two quartiles. Outliers, defined as sequences beyond 1.5 times the inter-quartile range, are plotted as open black circles.

As can be seen from the plots in Figure 3.3a, following challenge with HIV_{BaL}, there were substantially more sequences with high proportions of A-G transitions in the vector-treated group compared to the control group. Multiple sequences in the vector-treated group had A-G change proportions of 0.1 or more (i.e., at least 10% of A changed to G), although the control group had none. No such effect was seen for any other base change.

In contrast, we found that the A-G effect was not pronounced for HIV_{NL4-3} challenge (Figure 3.3b). The *env*AS was derived from the NL4-3 strain, so it is perfectly complementary to HIV_{NL4-3} but not to HIV_{BaL} – thus one conjecture is that mismatches in the RNA/RNA hybrid altered the processing of the duplex. We return to this point in the discussion.

We also found that viral sequences enriched in G-A transitions were present at a relatively high frequency following both HIV_{NL4-3} and HIV_{BaL}infections, though independently of whether or not *env*AS was present (Figure 3.3). High rates of G-A transitions are well documented during HIV infection, and are due to reverse transcriptase errors [16] and the C-to-T deamination activity of APOBEC family enzymes [17-19].

We next investigated whether the enrichment for A-G transitions in vector-treated samples was statistically significant. We defined informative sequences as those that had at least one A-G transition. We then calculated the probability of an A-G

transition over all informative sequences from both vector-treated and control groups, because according to our null hypothesis there is no difference between the two groups. Using the A-G transition probability, we assigned a binomial *P* value to each informative sequence based on the number of A-G events as a fraction of the total A positions examined. Thus we could define an A-G enrichment score for each informative sequence as the negative logarithm ($-\log_{10}$) of their *P* values. The lower the *P* value of a sequence, the less likely the observed frequency of A-G transitions occurring by chance, and the higher the A-G enrichment score. Sequences with statistically significant enrichment levels have *P* values of ≤0.05 (A-G enrichment score ≥1.301). Figure 3.4 depicts a representative portion of the MSA, showing only A positions in the region and comparing the most A-G enriched sequences for vector-treated and control samples following HIV_{BaL} challenge.

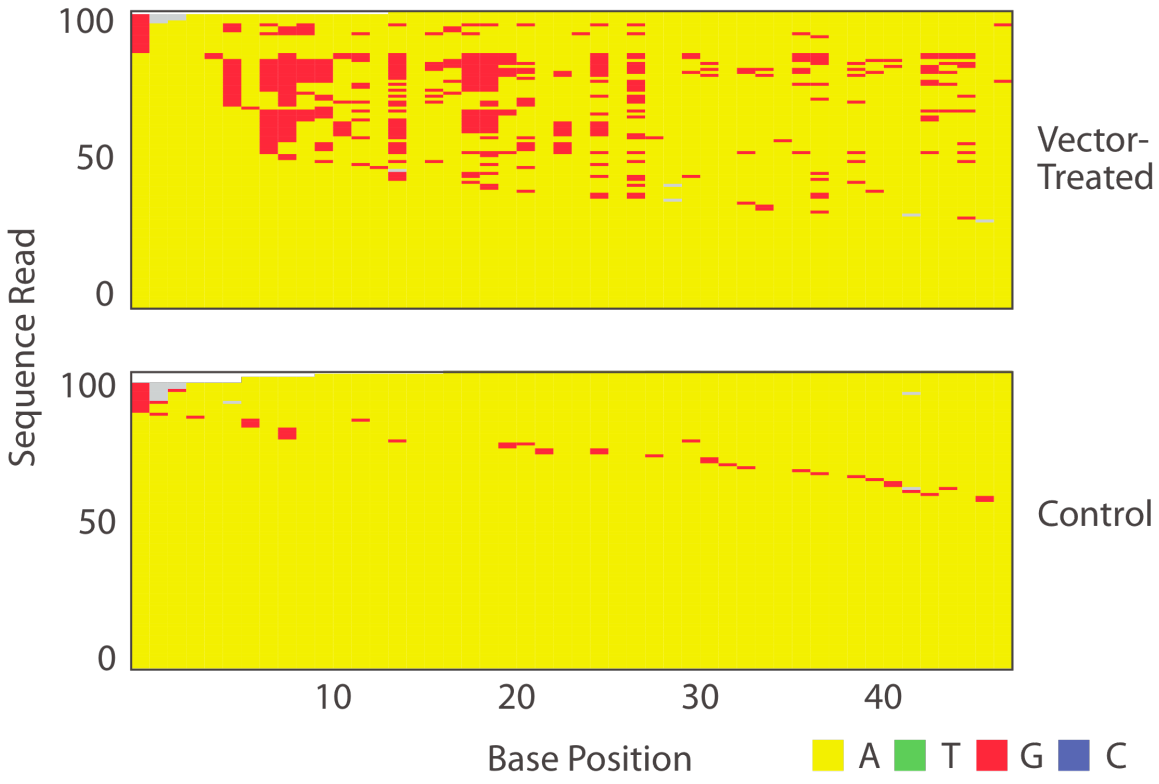


Figure 3.4 Comparison of the 100 sequences with the greatest enrichment of A-G transitions from the VRX494-treated and control mice challenged with HIV_{BaL}.

The most enriched sequences from viruses grown in vector-treated mice (top) are compared to the most enriched sequences from controls (bottom). The bases were color-coded as indicated at the bottom of the figure. Only base positions that were A (in yellow) in the starting viral stock are shown, those substituted with G are shown in red. Gray indicates sequence gaps.

As an initial statistical scan of the data, we performed a one-sided Fisher's exact test to determine whether there was a significant excess of sequences enriched in A-G transitions in the vector-treated versus the control group (pooled over all mice). For HIV_{BaL} challenge, we obtained a highly significant trend that persisted at least up to A-G enrichment scores as high as 4 (data not shown). This confirms our observations (Figures 3.3a and 3.4) there are sequences with significantly higher proportions of A-G transitions primarily in the HIV_{BaL} vector-treated group. No strongly significant trend was observed when we repeated our analysis for other base changes. A weak excess of sequences with G-A changes was seen, but this did not survive the more stringent test described below.

For the HIV_{NL4-3} challenge, the A-G effect was weak and showed no statistical significance above an enrichment score of 1.3 (data not shown). This confirms that in this case there is no preferential occurrence of sequences with high rates of A-G changes in the vector-treated group. No other types of base change showed significant enrichment following HIV_{NL4-3} challenge.

As a control for the A-G effect, we examined the frequency of A-to-G transitions outside the region targeted by *env*AS. There are 20 A sites present in the regions between the edges of the PCR primers used for sequence isolation and the *env* region complementary to *env*AS. These were compared to the 243 A sites in the *env* region targeted by *env*AS and analyzed in depth using the shorter two amplicons. Only 1 of 20 control sites (5%) was enriched in A-to-G transitions in the vector-treated compared to the control group (as determined by Fisher's exact test with Type I error of 0.05). In contrast, 70 of the 243 *env*AS targeted sites (28%) were enriched in the vector-treated group relative to the control. The difference achieved

significance (P = 0.01) by one-sided Fisher's exact test. Thus we conclude that the region of the HIV_{BaL} genome complementary to *env*AS had an elevated vector-induced frequency of A-to-G changes compared to flanking non-targeted regions of the viral genome.

Sequence features at A-to-G transitions

We investigated whether A-to-G changes were associated with any nearby sequence motifs in the HIV RNA. We aligned sites of A-to-G transitions at the affected base, and used WebLogo to scan for conserved sequence features. No strongly conserved motifs were detected, though a weak preference for a 5' A or T residue was seen (data not shown).

Nearest neighbor	% A changed to G with indicated nearest neighbor	
	Earlier Study [21]	This work
5'U	70	38
5'A	57	29
5'C	22	20
5'G	17	18
3'U	46	25
3'A	42	25
3'C	52	22
3'G	48	36

Table 3.2 Proportions of A residues converted to G with the indicated 5' and 3' nearest neighbor nucleotides.

Data from the HIV_{BaL} challenge was used for the analysis, considering only sites that were enriched at $p \leq 0.05$ for A-G transitions in the vector-treated sample.

We then analyzed the proportions of bases changed as a function of nearest neighbors and found a bias (Table 3.2). Previous studies have suggested that

dsRAD acts on A residues with 5' neighbors in order of preference U and A greater than C and G [20, 21]. We found a relatively higher proportion of modified A residues 3' of U or A compared to 3' of a G or C (Table 3.2). This suggests that dsRAD is responsible for modifying A to I in our experiments, leading to substitution with G. However, in the published study no preference for 3' neighbors was found, whereas for unknown reasons we found a preference for G residues.

We further asked whether there was a relation between the frequency of A-to-G substitutions and the extent of complementarity of the target genome with *env*AS. For this, we classified sites in the HIV_{BaL} template into those that were complementary to the corresponding antisense position and those that were not. Likewise, we classified the A sites in HIV_{BaL} as ones which underwent significant transition to G or not. We then compared the extent of sequence complementarity and fraction of sites with A-to-G transitions (using a sliding window of 25 bases). These were positively correlated with a highly significant *P* value ($P < 0.0001$) for Spearman's rank correlation coefficient. This is consistent with earlier reports of preferential RNA editing by dsRAD within perfectly matched regions of partially mismatched templates.

Comparing sequences enriched in A-G transitions among the groups of mice

A more rigorous statistical approach involves treating each mouse as a single measurement instead of analyzing all sequences pooled over a cohort of mice. For the analysis over pooled sequences described above, a single anomalous mouse might contribute sequences that dominated the behavior of the pool, particularly in a study of rare sequence variants, whereas trends reproducible over many mice are of greater interest.

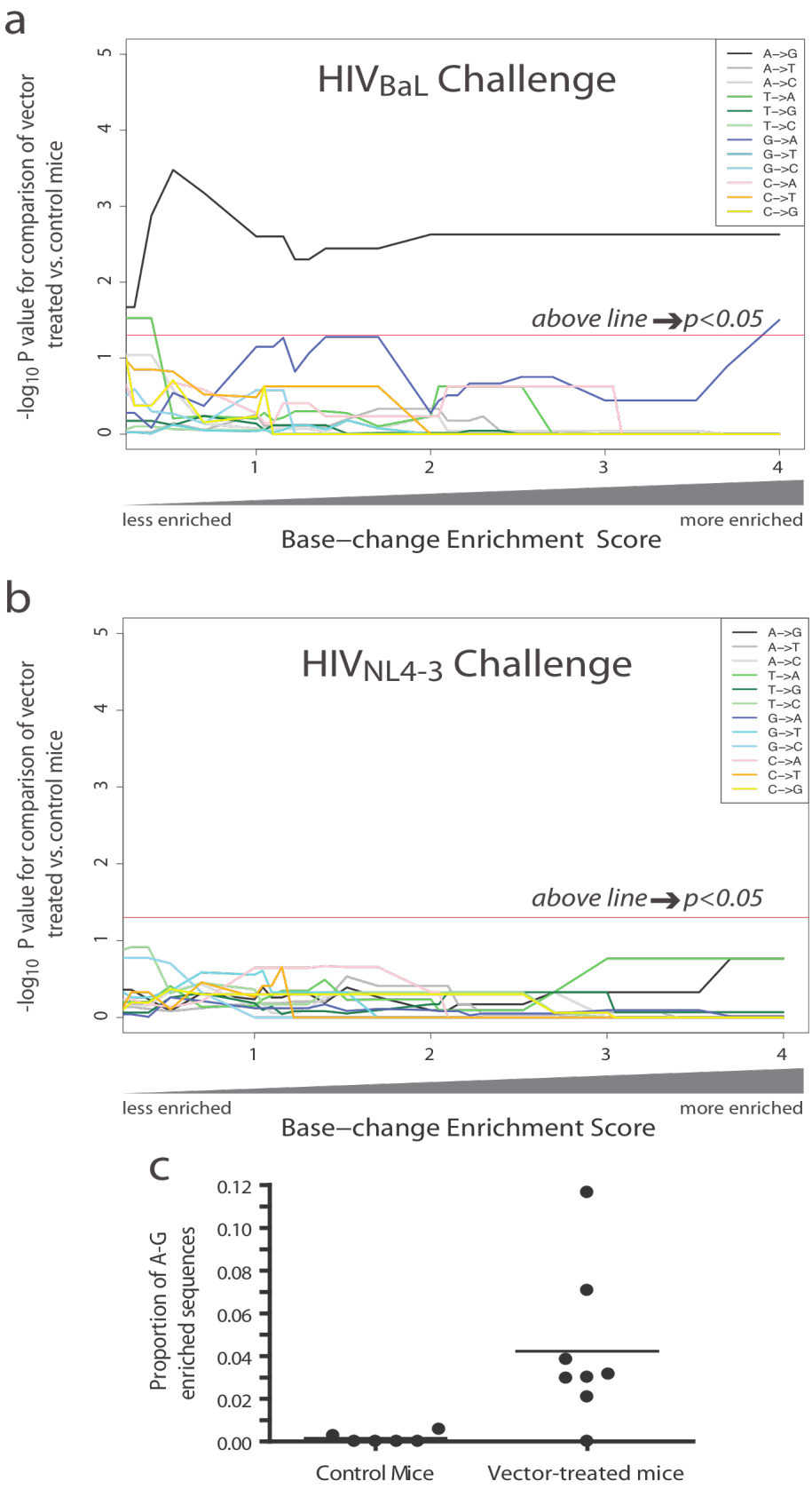

a
HIV BaL Challenge
A->G
A->T
A->C
T->A
T->G
T->C
G->A
G->T
G->C
C->A
C->T
C->G
above line → p<0.05
-log10 P value for comparison of vector treated vs. control mice
0
1
2
3
4
5
1
2
3
4
less enriched
more enriched
Base–change Enrichment Score
b
HIV NL4-3 Challenge
A->G
A->T
A->C
T->A
T->G
T->C
G->A
G->T
G->C
C->A
C->T
C->G
above line → p<0.05
-log10 P value for comparison of vector treated vs. control mice
0
1
2
3
4
5
1
2
3
4
less enriched
more enriched
Base–change Enrichment Score
c
Proportion of A-G enriched sequences
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Control Mice
Vector-treated mice

Figure 3.5 Statistical analysis of base substitution frequencies in vector-treated and control mice.

For the statistical analysis, each mouse was treated as an individual measure of proportions. **(a)** Comparison of vector-treated and control mice after HIV_{BaL} challenge. **(b)** Comparison of vector-treated and control mice after HIV_{NL4-3} challenge. In each panel, the x-axis indicates the extent of enrichment per sequence for each base substitution used in the analysis, so that at any indicated enrichment score, only sequences with at least that score were considered. Progressing from left to right indicates analysis of increasingly high levels of substitution. The y-axis indicates the $-\log_{10}$ p value from the Mann-Whitney nonparametric comparison of means (one-sided) for the excess in vector-treated cohort compared to control cohort. The analysis was carried out for each of the 12 base substitutions. The horizontal red line indicates the threshold for achieving statistical significance at *P* values ≤0.05 (above is significant). **(c)** Scatter plot showing the proportions of sequences with enrichment scores >1.3 (*P* values <0.05) of A-G transitions for each mouse in the HIV_{BaL} group.

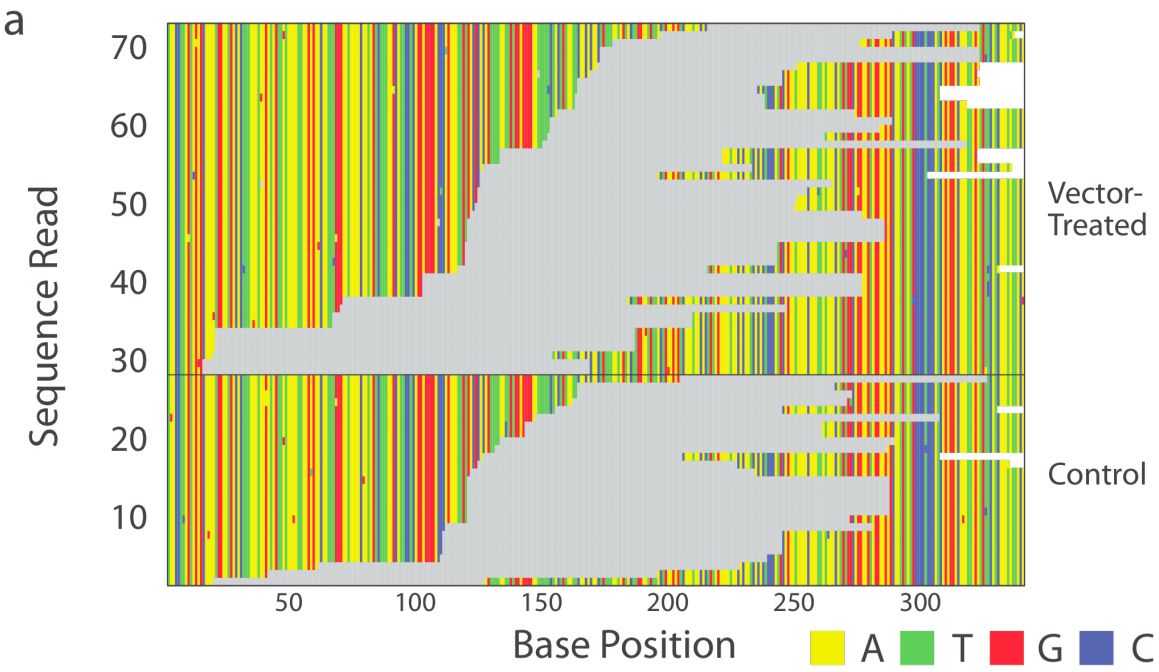
Figure 3.5a shows a comparison of the enrichment scores for base substitutions in the sequence reads, comparing vector-treated and control mice after challenge with the HIV_{BaL} virus (one sided Mann-Whitney comparison of means). The x-axis shows the enrichment score, and the y-axis shows the *P* value for the mean sequence excess comparing the vector-treated mice to control mice at the indicated enrichment score. For A-G substitutions (Figure 3.5a, black line), there is a significant excess of enriched sequences (above the horizontal red line) regardless of whether only high levels of enrichment (right side) or both low and high levels (left side) are used in the statistical analysis.

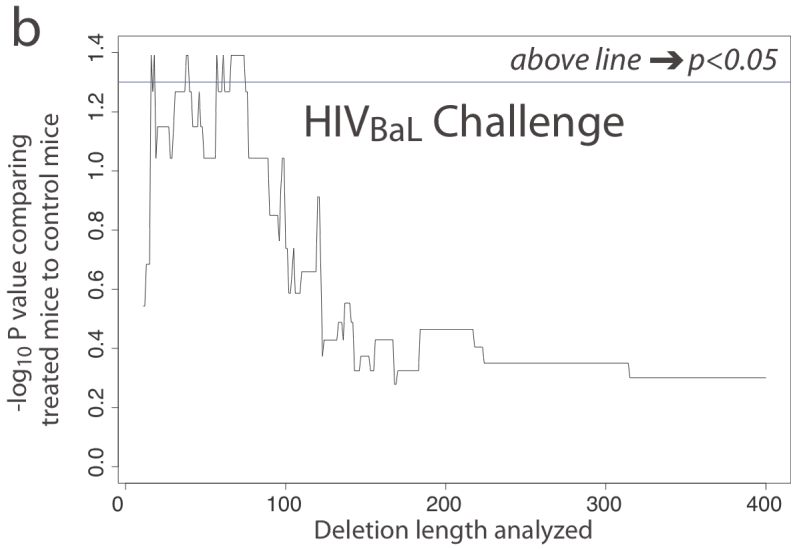
The remaining 11 transitions and transversions were also compared. None showed a consistently significant trend for sequences enriched for substitutions in the vector-treated cohort. Only T-A and G-A substitutions achieved a low level of significance at specific enrichment scores but were non-significant at most x-values and are of questionable biological importance.

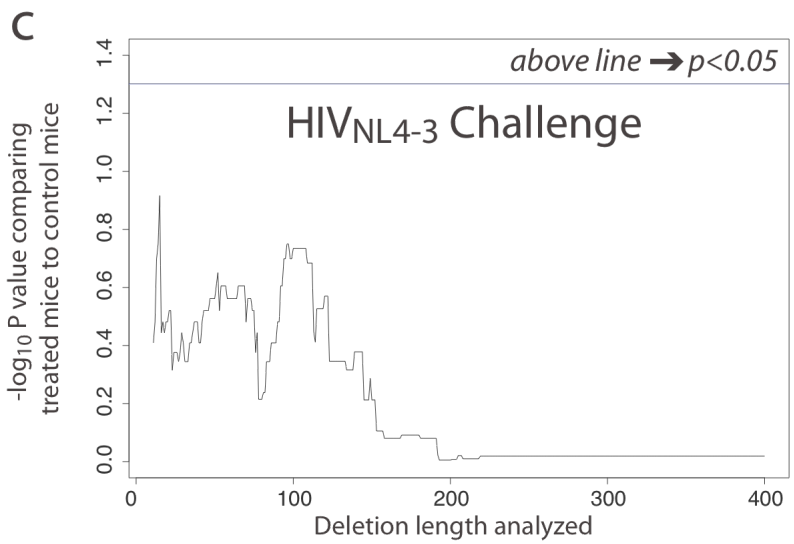
The analysis was repeated for the experiment with the HIV_{NL4-3} challenge virus (Figure 3.5b). None of the 12 substitutions achieved significance. Thus, the low level of significance seen for A-G transitions when all HIV_{NL4-3} sequences were pooled was not confirmed in the more rigorous analysis in which each mouse was treated as a single measurement.

A representation of the excess of sequences with A-G enrichment among vector-treated mice in the HIV_{BaL} challenge group is depicted in Figure 3.5c, using the enrichment level of $P < 0.05$ (score of >1.301) as an example. Most of the control mice had few or no sequences passing this A-G enrichment threshold, whereas the vector-treated mice had an average of ~4.5% passing the threshold ($P = 0.005$).

Frequency of deletions after challenge of the vector-modified cells




b
above line → p<0.05
HIVBaL Challenge
-log10 P value comparing treated mice to control mice
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
100
200
300
400
Deletion length analyzed


c
above line → p<0.05
HIVNL4-3 Challenge
-log10 P value comparing treated mice to control mice
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
100
200
300
400
Deletion length analyzed

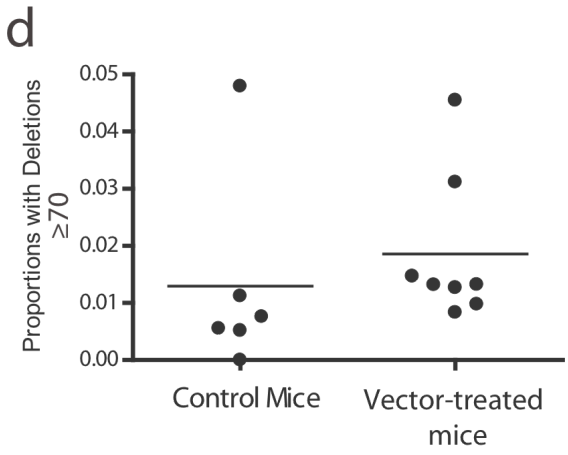

d
Proportions with Deletions ≥70
0.00
0.01
0.02
0.03
0.04
0.05
Control Mice
Vector-treated mice

Figure 3.6 Frequency of deletions in HIV-1 challenge viruses grown in the presence of vector-treated cells or controls.

(a) Illustration of the numbers and locations of deletions in HIV_{BaL} from the vector-treated and control groups corresponding to the 5' end of amplicon 2. Gray indicates sequence gaps. Deletions of ≥70 bases were plotted. **(b)** Analysis of the significance of the difference in deletion frequencies between vector-treated and control mice after HIV_{BaL} challenge. The x-axis shows the length of deletions included in the analysis, so that at any indicated value only deletions of that length or greater were included in the analysis. The y-axis shows the *P* value for the comparison of means between vector-treated and control groups calculated using the nonparametric Mann-Whitney test (one-sided). Each mouse was treated as a single measurement of proportions. **(c)** As in **(b),** but analysis of the HIV_{NL4-3} challenge group. **(d)** Comparison of proportions for deletions ≥70 bases, for the control and vector-treated mice. Each mouse is shown as a point.

Studies of antisense inhibition of HIV in tissue culture, where most cells contained the VRX494 vector, indicated that deletions also accumulated [4], so we analyzed the frequencies of deletions here. The distribution of sequences with deletions of at least 70 bases is shown in Figure 3.6a for one of the HIV_{BaL} amplicons. Inspections suggested that the number of sequences with deletions is greater in the vector-treated mice than in the controls.

We thus calculated the significance, again treating each mouse as a single measure, of the proportion of deleted viruses (Figures 3.6b,c). The x-axis indicates the length of deletion used in the analysis – progressing to the right indicates restricting the analysis to incrementally longer deletions. The y-axis shows the statistical significance of testing for the excess of sequences in vector-treated mice versus controls at the indicated deletion length. Segments of the curve above the horizontal blue line indicate statistical significance. For HIV_{BaL} challenge, some data sets with deletions in the range of ~10-100 bases showed significant enrichment in the vector-treated group, though others did not, indicating that the enrichment of deletions achieved marginal significance (Figure 3.6b). No such significant enrichment in deletions was seen with HIV_{NL4-3} challenge (Figure 3.6c). The enrichment in deletions for HIV_{BaL} sequences, using deletions of ≥70 bases for

analysis, is shown in Figure 3.6d, where each mouse is plotted individually. As can be seen, the difference in means is significant ($P = 0.04$), but the effect slight. We conclude that there is a significant though modest increase in the frequency of deletions in the HIV_{BaL} but not the HIV_{NL4-3} challenged vector-treated mice.

We next checked whether variants with deletions ≥70 bases were also enriched for A-G transitions, as both features potentially arise from *env*AS pressure. We performed Fisher's exact test to establish whether there was an excess of deletion-harboring variants among the population of sequences significantly enriched for A-G. We did not observe any significant enrichment or depletion. Thus the A-G transitions and deletions appear to accumulate independently.

Correlation between extent of T cell modification and effects of VRX494 envAS treatment

Each cohort of mice showed variation in the final extent of T cell modification by the VRX494 vector following engraftment, allowing us to investigate possible correlations with the frequency of A-to-G transitions or deletions. No significant correlation was observed between modification frequencies and enrichment of sequences with high frequencies of A-G transitions (for Spearman's rank correlation coefficient). However a modest but significant correlation ($P = 0.04$) was seen for vector modification (measured at day 48) and deletion frequency for HIV_{BaL} challenged mice. Significance was achieved when deletions of about 70 bp or larger were studied, which we know are enriched in vector-treated mice, but not with shorter lengths (data not shown). We thus conclude that a possible modest positive correlation could be seen where increased vector modification was associated with increased deletion frequency.

Discussion

Here we report modeling the effects of an HIV *env*AS vector on HIV challenge virus under conditions where cells harboring the *env*AS construct were a minor component of the cell population. This setting is of interest because studies of cells in culture, where all cells contain the *env*AS, show strong inhibition of HIV replication by the *env*AS [4], but trials to date in human subjects have achieved vector marking in only a minority of circulating T cells. Despite this, some patients showed intriguing alterations in disease parameters [5]. We thus sought to study in detail the effects on HIV populations when only a minority of cells contained the *env*AS using hNSG mice. We found a significant enrichment compared to controls for sequences enriched for A-G transitions and deletions in mice harboring the *env*AS in the experimental infections with the HIV_{BaL} challenge virus.

However, unexpectedly, we also found that A-G transitions were only enriched in samples from the HIV_{BaL} infections, and not HIV_{NL4-3} infections. The origin of the difference is unknown. According to one idea, the different co-receptor usage of the two challenge viruses might have resulted in infection of two different cell types possibly resulting in different treatment of double-stranded RNA, but at present we have no evidence in favor of this view. A more attractive model is that double-stranded RNA is processed differently depending on whether the duplexes were perfectly matched or contained mismatches. The *env*AS was derived from HIV_{NL4-3}, so RNA duplexes formed with HIV_{NL4-3} RNA would be perfectly paired, while those with HIV_{Ba} would contain a mixture of paired and mismatched regions. Favored action at duplex regions near mismatches, and the observed RNA nearest neighbor preferences, are consistent with action of the dsRAD enzyme [20, 21]. However, dsRAD is also able to act on perfectly matched duplexes, so it is unclear why no A-to-G transitions accumulated in the HIV_{NL4-3} infection.

The difference in processing could be a consequence of differential initial attack, in which only RNA duplexes containing mismatches are substrates for the duplex-processing enzyme. However, another possibility is that the perfect duplexes are attacked but degraded more quickly, so that no HIV sequences with perfect matches survive to allow detection even by deep sequencing. We presently have no basis for favoring either model.

The origins of the deletions in the HIV_{BaL} challenge virus seen here and in a previous study [4] are unclear. Not only must the RNA duplex be cleaved, it must be rejoined to form the internal deletion detected by sequencing. One candidate mechanism involves cleavage by a double-strand specific ribonuclease (RNAse III) such as Drosha, followed by copy-choice polymerization by reverse transcriptase, so that template switching from a broken RNA template to a site elsewhere on a second HIV RNA would yield the observed internal deletion. We did not observe enrichment for deletions following HIV_{NL4-3} challenge, so one model would hold that the RNAse III involved acts preferentially on mismatched dsRNA. However, we note that the same results would be obtained if products of cleavage of perfectly matched RNA were rapidly degraded. No positive correlation was seen between the deletions and A-G transitions, indicating that the enzyme systems responsible for RNA cleavage and RNA editing are likely acting independently.

It is surprising that modification of only 4-11% of cells in the mice could have caused detectable alterations in the HIV_{BaL} population, and similarly surprising to observe possible alterations in disease parameters in human subjects after gene therapy with a similar *env*AS vector [5]. However, it is possible that mobilization of the VRX494 *env*AS vector augmented the effect. The VRX494 vector contains intact LTRs, the *cis* sites needed for RNA packaging, reverse transcription, and integration, and also sites needed for response to Tat and Rev. Thus, if a cell harboring the integrated VRX494 provirus were infected by HIV, the VRX494 provirus could be transcribed, then the RNA exported from the nucleus, packaged and released.

Infection into a naïve cell, followed by reverse transcription and integration, would achieve spread of the VRX494 *env*AS vector between cells. Such conditional replication of the vector could expand the proportion of vector-containing cells and increase the likelihood of accumulating the observed mutations in the challenge viruses.

Are the A-to-G transitions and deletions viral escape mutations? This is an important question and the answer is not fully clarified here, but at present it seems unlikely that they confer escape. The large deletions in HIV envelope are sure to be *env*-minus, and a few A-to-G transitions in an ~900 bp duplex are not likely to destabilize RNA pairing significantly. More likely the mutations that accumulate in HIV_{BaL} are by-products of the action of cellular RNA-modifying enzymes that inhibit viral replication. The modified RNAs with deletions could persist in the population by complementation during co-infection with wild-type viruses, as they were found in viral particles in serum. They may also be continually arising *de novo* in the infected cell populations during ongoing viral replication, or else complementation may be efficient enough to maintain modified RNAs in the population for multiple generations.

The experimental protocol used here differed from the clinical trial in the order of gene modification versus HIV infection. In the clinical trial, gene-modified cells were re-infused into patients with pre-existing HIV infection. In our hNSG mouse model, the cells were first gene-modified and mice were infected afterward. The composition of pre-existing viral populations can affect disease course [22], and this may be modified by prior treatment history. Thus, it would be of potential interest to compare effects of infection first followed by *env*AS modification in the hNSG model.

To summarize, we observe significant effects of *env*AS pressure *in vivo* when only 4-11% of HIV-1 target CD4+ T cells are gene-modified. The effects were only

detectable with divergent *env* target sequence, suggesting action of dsRNA-modifying enzymes acting on imperfectly matched sequences. As *env*AS has been shown to control HIV-1 replication and has also been deemed safe in gene therapeutic clinical trails, this represents a promising antiviral agent.

Materials and Methods

Transduction and culture of primary human CD4 T cells

All human-related studies were done per Declaration of Helsinki Protocols. Primary human CD4+ T cells were obtained from the University of Pennsylvania CFAR Immunology Core under an institutional review board (IRB) approved protocol and stimulated with anti-CD3 and anti-CD28 antibody-coated beads as previously described [23]. The following day the activated T cells were either left alone (untransduced), or transduced with VRX494 [24], expanded and frozen. This work was done in the Riley lab.

Infection of humanized NOD/SCID IL-2Rγ^{null} (hNSG) mice

NSG mice (NOD.Cg-*Prkdc^{scid} Il2rg^{tm1Wjl}*/SzJ stock) were obtained from the Jackson Laboratory (Bar Harbor, ME) and bred and housed in the University of Pennsylvania Xenograft and Stem Core Lab. Before injection, each expanded cell population was thawed and transduced cells were mixed with untransduced T cells so that the final concentration of transduced T cells was approximately 10%. The mixed cells were washed in phosphate-buffered saline (PBS) and 10 million cells were injected intravenously into each mouse. A total of 20 animals received cells containing 10% transduced cells with VRX494, whereas 20 animals received only the untransduced cells. The HIV-1 based vector VRX494 used in this study encodes an *env*AS complementary to the HIV_{NL4-3} *env* gene. After 20 days, the number and percentage

of GFP-positive, CD4+ T cells within the NSG mouse peripheral blood was measured using TruCount beads as per the manufacturer's recommendation (BD Biosciences, San Jose, CA). One mouse in the control group showed poor engraftment and was excluded from further studies. For both control and VRX494-treated groups, the engraftment data (and additionally the engrafted proportions of VRX494-modified cells for the treated group) were used to randomize mice between the HIV_{NL4-3} and HIV_{BaL} challenge cohorts. Each group of mice was then challenged with the supernatants of 293 T cells transfected with molecular clones of HIV-1 expressing either the HIV_{NL4-3} or HIV_{BaL} *env* sequence [pNL4-3, a gift of VIRxSYS (Gaithersburg, MD) and pWT/BaL, obtained from the NIH AIDS Research & Reference Reagent Program (Germantown, MD), respectively]. HIV-1 infection of reconstituted mice was performed at Bioqual (Rockville, MD) under an approved animal protocol.

Amplification and deep sequencing of HIV quasispecies

Total RNA was extracted from mouse plasma samples using the *Illustra* RNAspin kit (GE Healthcare, Buckinghamshire, UK) with RNA carrier added to improve extraction efficiency. Composite primers (Table 3.3) made of 454 sequencing adapters, barcodes and HIV_{env} primers, 5' to 3' in that order, were used to amplify HIV-1 RNA by *One-step* RT-PCR (Qiagen, Valencia, CA) using *RNasin* RNase inhibitor (Promega, Madison, WI) and a touch-down protocol (with a total of 30 cycles of amplification) as follows:

1 x (30 min at 50°C), 1 x (15 min at 95°C), 10 x (1 min at 94°C; 1 min at 58°C-53°C, 0.5°C iterative decrement; 1 min at 72°C) (starting with a temperature of 58°C and reducing it successively by 0.5°C to reach 53°C), 20 x (1 min at 94°C; 1 min at 53°C; 1 min 15 sec at 72°C), 1 x (10 min at 72°C; maintained at 4°C). Amplified products were gel-purified separately for each mouse-amplicon combination and pooled following DNA quantitation by *Quant-iT* PicoGreen dsDNA Assay kit (Molecular Probes, Invitrogen, Eugene, OR). Pooled sequences were pyrosequenced using the 454/Roche platform at the University of Pennsylvania. A total of 84,074 raw reads

were recovered. These have been deposited with the Sequence Read Archive (SRA) of National Center for Biotechnology Information (NCBI) under submission accession SRA010586.3. The raw reads have accessions SRR033739.1-SRR033739.84074.

(a)

***env*AS Primer**	**Sequence**		
	454 Adapter	Barcode	*env*AS Primer
for NL4-3			
EnvF6553	GCCTCCCTCGCGCCATCAG	Listed	ATGGGATCAAAGCCTAAAGCC
EnvR6954	GCCTTGCCAGCCCGCTCAG	in **(b)**	CATTGTACTGTGCTGACATTTGTAC
EnvF7184	GCCTCCCTCGCGCCATCAG		GGAAATATGAGACAAGCACATTG
EnvR7587	GCCTTGCCAGCCCGCTCAG		CCATCTCTTGTTAATAGCAGCCC
for BaL			
EnvF701	GCCTCCCTCGCGCCATCAG	Listed	ATGGGATCAAAGCCTAAAGCC
EnvR1123	GCCTTGCCAGCCCGCTCAG	in **(b)**	CATTGTACTGTGCTGACATTTGAAC
EnvF1350	GCCTCCCTCGCGCCATCAG		GGAGATATAAGACAAGCACATTG
EnvR1738	GCCTTGCCAGCCCGCTCAG		CCATCTCTTGTTAATAGCAGCCC

(b)

Barcode	**Sequence**	**Mice**	**Barcode**	**Sequence**	**Mice**
bc1	CAGTCAGT	1050	bc18	TCACTGTC	985
bc2	ACACACTG	1090	bc19	TCTGTGAG	1037
bc3	ACGACATC	1044	bc20	TGAGTCAC	1089
bc4	AGACACTC	1046	bc21	ACAGACTC	1093
bc5	ATATCGCG	1048	bc22	ACTGCTGA	946
bc6	ATCGATGC	1051	bc23	AGCACTAC	948
bc7	CACTACAG	1055	bc24	AGTCGTCA	1038
bc8	CGATATGC	1086	bc25	AGTGTCAC	1040
bc9	CGTACGAT	1087	bc26	CACTGTGA	1094
bc10	CTACGATG	1092	bc27	CATCGTAG	1136
bc11	GACACTCA	1042	bc28	CGATGCTA	1137
bc12	GAGTACAG	1045	bc29	CTAGTGCA	1139
bc13	GCATATCG	1047	bc30	GAGTGACA	1091
bc14	GCTACGTA	1052	bc31	TCACGAGT	1095
bc15	GTACACGT	1053	bc32	TCGACATG	1138
bc16	TACGATCG	1085	bc33	TGAGCACT	1140
bc17	TAGCGCAT	947			

Table 3.3 Sequences of oligonucleotides and barcodes used in this study.

(a) Composite primers. There are 4 primers each for the HIV_{NL4-3} and HIV_{BaL} template corresponding to the ones depicted in Figure 3.2. For *env*AS primer, 'EnvF' and 'EnvR' refer to the forward and reverse primers respectively whereas the following numbers refer to the starting genomic coordinates on the HIV_{NL4-3} and HIV_{BaL} templates. The composite primer sequence is made up of 454 adapter, barcode and the actual *env*AS primer (from 5' to 3'). **(b)** List of unique barcodes with their 8 bp sequence and tagged mouse. Barcodes 1-10 and 21-29 were used to tag mice challenged with HIV_{NL4-3} while barcodes 11-20 and 30-33 were used for mice challenged with HIV_{BaL}.

Bioinformatics

Pyrosequence reads were barcode-decoded for assignment to source mouse samples. Sequence reads were filtered for exact match to primers. To remove sequences shorter than the read length distribution peak, that are known to have high 454 error rates, additional filtering was carried out to select only sequences ≥220 bases long. Reference genomes corresponding to the HIV_{NL4-3} and HIV_{BaL} infection stocks were obtained from NCBI's nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore/), accession IDs M19921 (Version: M19921.1, GI: 328415) and M63929 (Version: M63929.1, GI: 326765). Viral infection stocks were Sanger-sequenced and verified that they matched the respective NCBI sequences for the genomic region analyzed. Sequences were then aligned to the reference HIV-1 genome using the Needleman-Wunsch pair-wise global alignment (PWA) algorithm with a match score of 5, and gap opening and extension penalties of 20 and 0.5. A few sequences from HIV_{NL4-3} challenge group had a higher alignment score to the HIV_{BaL} reference and vice-versa; these were classified as apparent sequence crossovers and were excluded from further analysis. For remaining sequences in HIV_{NL4-3} and HIV_{BaL} group, two multiple sequence alignments (MSAs) were created, one for each group, by parsing the individual PWAs within each group. Finally, primer sequences were trimmed from each sequence because these are not part of the viral genome actually sampled.

All statistical programming and tests were performed in the R computing framework (http://www.r-project.org/). One sided statistical tests were used for analysis of the A-G substitutions and deletions because a previous report established a directional hypothesis (increased frequency after treatment) [4].

A-G analysis: For each MSA, base positions having more than 50% change from the reference sequence base over all other sequences (discounting MSA gaps) were excluded, by reasoning that the preferred base reported by pyrosequencing could likely be the correct base at these positions. For the HIV_{BaL} MSA, of the 994 positions sampled by our amplification scheme, 22 positions were excluded, of which 17 were in regions of high coverage (>1000 reads per position) corresponding to the shorter two amplicons. For HIV_{NL4-3}, all 40 positions excluded (out of 991) were in the low coverage region between the shorter two amplicons (<30 reads per position). For each base-change, overall probability of the event was obtained from the pool of all informative sequences (see Results). Using this, a binomial distribution based *P* value was assigned per sequence based on the observed proportion of base-change in a sequence. This *P* value was termed the *enrichment level* and the $-\log_{10}$ of this was called the *enrichment score*. For each base-change, excess within the vector-treated cohort, of sequences with a minimum enrichment was statistically inferred. For the pooled analysis, i.e., taking vector-treated group as a whole, this was done by Fisher's exact test. For un-pooled analysis Mann-Whitney test was done. In each case, statistical significance was tested with enrichment scores up to 4 (*P* value of 0.0001) and displayed as $-\log_{10}$ of the *P* value (Figure 3.5a,b).

Investigating Sequence Features for A-G modifications:
(a) Conserved Motifs. To research sequence motifs on the HIV-1 genome associated with A-G transitions, for informative sequences, sites with enriched changes in the vector-treated group were chosen. Sites were selected at an enrichment corresponding to *P* value ≤0.05 by Fisher's exact test. Reference genome positions 10 bases upstream and downstream of each site were aligned and WebLogo was

used for information content analysis (http://weblogo.berkeley.edu/) to detect conserved motifs.

(b) Neighbor Base Rules. All instances of XA and AX dinucleotides in the *env*AS target region were tabulated, $X \in \{A,T,G,C\}$. Then for each X, in both cases of X being 5' or 3' to A, the proportion of instances the A was a statistically preferred site for transitioning to G (as determined in (a) above with Fisher's exact test) was calculated.

(c) Genomic Distribution in Relation to Complementarity. The fraction of sites in the HIV_{BaL} template that were complementary to *env*AS was calculated across the *env*AS targeted region with a sliding window of 25 bases. Similarly, of all A sites, the fraction of A sites statistically preferred for transitioning to G (as determined in (a) above with Fisher's exact test) was calculated. The correlation between the two quantities was estimated by Spearman's rank correlation coefficient.

Deletion Analysis: Small deletions can arise by chance among viral quasi-species or during sample work up, so meaningful deletions were qualified as only those which had a length >10 bases. Additionally deletions were required to have a ≥10 base long gapless aligned region at both ends. With this constraint, the hope was to select deletions that are biological while weeding out those that could be generated as alignment artifacts through gapped matching of a few error-prone bases. Two deletion-related quantities were defined for each sequence: 1) *Total deletion length*, representing the cumulative length of all qualifying deletions, and 2) *Maximum deletion length*, representing the length of the longest qualifying deletion. Then the spectrum of total deletion lengths that can be assayed by our study ranges from 11 to ~900 bases, the upper limit being imposed by the longest amplicon. For each possible deletion length in this range, sequences with cumulative deletions of at least that size were selected and tested for excess in vector-treated mice (un-pooled analysis). Statistical significance was inferred by Mann-Whitney test and displayed as $-\log_{10}$ of the p-value (Figure 3.6b,c). Repeating analysis with maximum deletion length did not alter results significantly.

Acknowledgments

We thank James L Riley (Department of Pathology and Laboratory Medicine, University of Pennsylvania) and members of his laboratory, especially Gabriela Plesa, for their collaboration. We thank Carl H June (Department of Pathology and Laboratory Medicine, Univ. of Pennsylvania) and Gwendolyn Binder-Scholl (Adaptimmune LLC, Philadelphia, PA) for assistance in establishing this collaboration, VIRxSYS Corporation (Gaithersburg, MD, USA) for providing the lentiviral vector used in this study, Jake Yalley-Ogunro, Gary Thomas and Mark Lewis (all from Bioqual Inc., Rockville, MD) for manipulating the HIV-1 infected mice, and Gwenn Danet-Desnoyers and Anthony Secreto (both from the Stem Cell and Xenograft Core, University of Pennsylvania) for maintaining the animals prior to HIV-1 infection. We are grateful to members of the Bushman lab for help and suggestions, especially Scott Sherrill-Mix for computational support. This work was supported by NIH grants RO1AI0802020, U19AI082628, U19AI066290, Penn Genome Frontiers Institute and a grant with the Pennsylvania Department of Health. The Department of Health specifically disclaims responsibility for any analyses, interpretations, or conclusions.

References

[1]. Reyes-Darias JA, Sanchez-Luque FJ, Berzal-Herranz A. Inhibition of HIV-1 replication by RNA-based strategies. *Curr HIV Res* 2008; **6**:500-514.
[2]. Rossi JJ, June CH, Kohn DB. Genetic therapies against HIV. *Nat Biotechnol* 2007; **25**:1444-1454.
[3]. Scherer L, Rossi JJ, Weinberg MS. Progress and prospects: RNA-based therapies for treatment of HIV infection. *Gene Ther* 2007; **14**:1057-1064.

[4]. Lu X, Yu Q, Binder GK, Chen Z, Slepushkina T, Rossi J, et al. Antisense-mediated inhibition of human immunodeficiency virus (HIV) replication by use of an HIV type 1-based vector results in severely attenuated mutants incapable of developing resistance. *J Virol* 2004; **78**:7079-7088.

[5]. Levine BL, Humeau LM, Boyer J, MacGregor RR, Rebello T, Lu X, et al. Gene transfer in humans using a conditionally replicating lentiviral vector. *Proc Natl Acad Sci U S A* 2006; **103**:17372-17377.

[6]. Wang GP, Levine BL, Binder GK, Berry CC, Malani N, McGarrity G, et al. Analysis of Lentiviral Vector Integration in HIV+ Study Subjects Receiving Autologous Infusions of Gene Modified CD4+ T Cells. *Mol Ther* 2009;.

[7]. Weinberger LS, Schaffer DV, Arkin AP. Theoretical design of a gene therapy to prevent AIDS but not human immunodeficiency virus type 1 infection. *J Virol* 2003; **77**:10028-10036.

[8]. Ishikawa F, Yasukawa M, Lyons B, Yoshida S, Miyamoto T, Yoshimoto G, et al. Development of functional human blood and immune systems in NOD/SCID/IL2 receptor {gamma} chain(null) mice. *Blood* 2005; **106**:1565-1573.

[9]. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**:376-380.

[10]. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 2008; **5**:235-237.

[11]. Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, et al. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2007; **2**:e197.

[12]. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, et al. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 2007; **35**:e91.

[13]. Church G, Gilbert W. Genomic sequencing. *Proc. Natl. Acad. Sci. USA* 1984; **81**:1991-1995.

[14]. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A* 2006; **103**:12115-12120.
[15]. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007; **8**:R143.
[16]. Mansky LM, Temin HM. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol* 1995; **69**:5087-5094.
[17]. Lee YN, Malim MH, Bieniasz PD. Hypermutation of an ancient human retrovirus by APOBEC3G. *J Virol* 2008; **82**:8762-8770.
[18]. Wood N, Bhattacharya T, Keele BF, Giorgi E, Liu M, Gaschen B, et al. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC. *PLoS Pathog* 2009; **5**:e1000414.
[19]. Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* 2002; **418**:646-50.
[20]. Polson AG, Bass BL. Preferential selection of adenosines for modification by double-stranded RNA adenosine deaminase. *EMBO J* 1994; **13**:5701-5711.
[21]. Kumar M, Carmichael GG. Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc Natl Acad Sci U S A* 1997; **94**:3542-3547.
[22]. Ribeiro RM, Bonhoeffer S. Production of resistant HIV mutants during antiretroviral therapy. *Proc Natl Acad Sci U S A* 2000; **97**:7681-7686.
[23]. Riley JL, Schlienger K, Blair PJ, Carreno B, Craighead N, Kim D, et al. Modulation of susceptibility to HIV-1 infection by the cytotoxic T lymphocyte antigen 4 costimulatory molecule. *J Exp Med* 2000; **191**:1987-1997.
[24]. Humeau LM, Binder GK, Lu X, Slepushkin V, Merling R, Echeagaray P, et al. Efficient lentiviral vector-mediated control of HIV-1 replication in CD4 lymphocytes from diverse HIV+ infected patients grouped according to CD4 count and viral load. *Mol Ther* 2004; **9**:902-913.

Chapter 4 EFFECTS OF *ENV* ANTISENSE VISIBLE IN PATIENTS IN A CLINICAL TRIAL

This work is being prepared for publication.

Abstract

Antisense sequences targeting HIV genomic RNA provide a mechanism to control the virus. These inhibited HIV in tissue culture studies and mice offering an attractive choice for anti-HIV gene therapy strategies. T cells modified with an antisense directed against HIV *env* were shown to control heterogeneous viral strains *in vitro*. Subsequently, this *env* antisense or *env*AS became the first of its kind to be used in a clinical trial. Antisense was delivered to patient T cells by lentiviral vectors. This was also the first use of a lentiviral vector in a clinical setting. A few participants exhibited reduction in viral loads and improvements in immune function. To assess *env*AS pressure on evolving virus, we simulated the clinical trial *in vivo* in a mouse model. Consistent with previous work that studied *env*AS *in vitro*, we observed enrichment of molecular signatures on virus impacted by *env*AS – in the form of A-G changes and deletions. Here we investigated 8 patients participating in a follow-up phase I/II trial for evidence of such signatures in virus rebounding during structured treatment interruption (STI). We included 9 patients from an unrelated trial STI as a control group. All patients were studied longitudinally over 2-5 time-points. A total of 218,756 reads obtained by 454/Roche pyrosequencing were condensed into 127,290 viral variants following quality filtering and error control by Pyronoise. There was an indication of a higher rate for A-G transitions within the *env*AS-target region for treated patients as compared to controls. This recapitulates findings in the mouse study, although the signal is weaker. Interestingly, the A-G effect is detectable only in the initial time-point in the treated patients. We did not find a similar trend for deletions, although in a pooled analysis

over groups, there were more deletions in the treated. This analysis illustrates that *env*AS exerts pressure on circulating virus in patients. The signal is faint but this is likely a consequence of the low levels and persistence of *env*AS-modified cells in the clinical trial.

Introduction

Survival rates for HIV-infected patients have improved dramatically with the advent of HAART or combination anti-retroviral drug therapy. HAART suppresses virus replication, however, it has failed to cure patients of HIV completely. Latent reservoirs of HIV [1-3] persist in the patients undergoing HAART, contributing to residual levels of viremia [4-6]. Treated patients still suffer from reduced life expectancy and increased risk of complications such as cardiovascular, hepatic or renal diseases, and cancer [7]. Further, HAART often does not restore CD4 T cells to normal pre-infection levels [8]. Drug-related toxicities and persistent immune deficiencies likely contribute to premature ageing-related deficiencies. To develop alternatives, research on anti-HIV gene and cell therapy has been conducted for many years. The recent case of the Berlin patient, who reportedly is cured of HIV following chemotherapy and transplantation of HIV-resistant cells [9, 10], holds promise for such approaches.

Inhibition of HIV *in vitro* by antisense oligonucleotides was among the first anti-HIV strategies reported in the beginning years of the AIDS pandemic [11, 12]. By the early 1990s researchers had described HIV suppression *in vitro* by T cells modified with antisense that targeted HIV *gag* [13, 14] or *tat, env* and *rev* in conjunction [15]. The first antisense-based gene therapy clinical trial was conducted in 2006 at the University of Pennsylvania [16]. This was also the first time a lentiviral vector was used in the clinic, to modify autologous human CD4 T cells before their reinfusion into HIV infected patients who were failing their HAART regimens. The outline of

this procedure is illustrated in Figure 4.1. The vector was called VRX496 and its payload was an antisense targeting HIV *env*, which had been shown to control HIV replication *in vitro* [17, 18]. The results of this phase I trial was encouraging – there were no safety concerns and therapeutic benefits in terms of lowering of viral load and increase in CD4 cell counts were observed in some participants. This motivated the design of a subsequent clinical trial to study if VRX496 could suppress HIV in the absence of HAART.

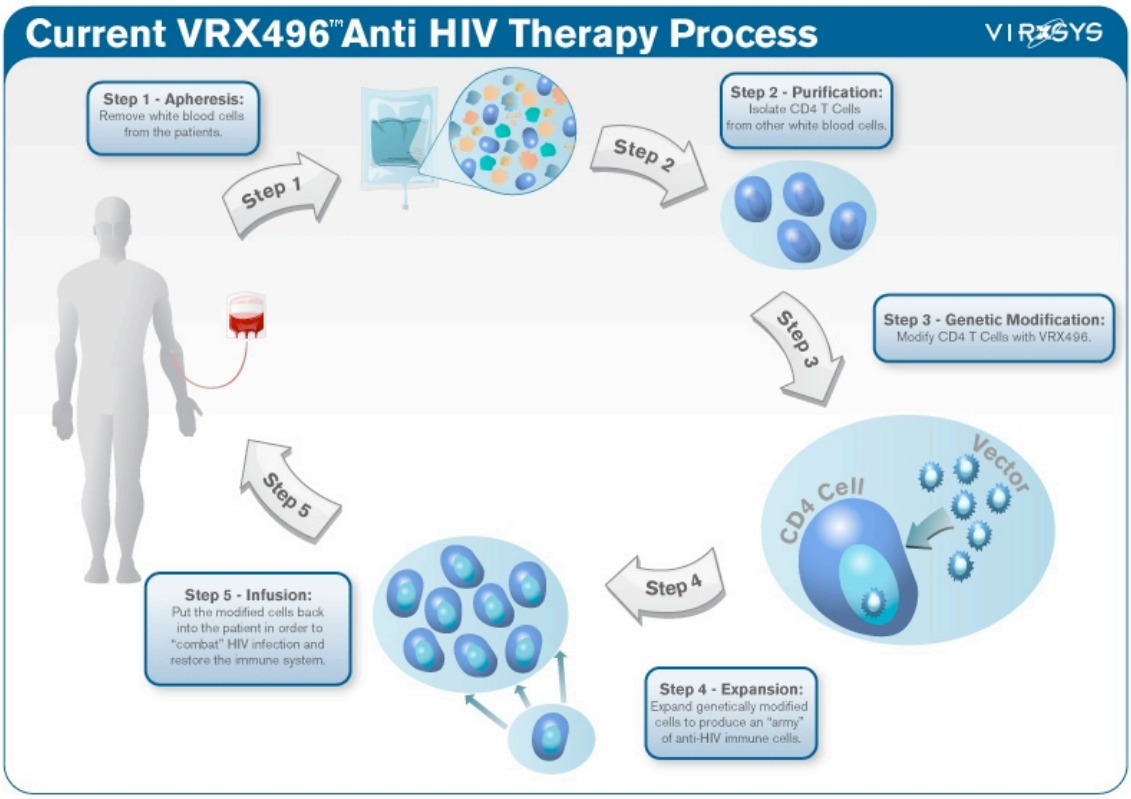


Figure 4.1 Steps in the anti-HIV antisense-based gene therapy of HIV infected patients with VRX496 vector. Permission to reuse this figure was granted by the publisher, © 2009 VIRxSYS Corporation, Gaithersburg, MD.

Accordingly a phase I/II trial was set up also at University of Pennsylvania (clinicaltrials.gov Id NCT00295477). This enrolled HIV-positive patients who were on HAART with no detectable virus (<50 copies/ml). As earlier [16], clinicians infused participants with 10 billion VRX496-modified autologous CD4 T cells – either 3 or 6 such doses were administered. This was followed by a structured

treatment interruption (STI) wherein HAART was suspended. Of 13 patients who underwent STI, 8 could be considered for evaluation of study endpoints [19]. The primary efficacy parameters involved post-STI measurements of time to viral recrudescence, viral load and CD4 counts in relation to that at baseline, persistence of modified cells, and safety.

The study is ongoing but important observations have already been made [19]. VRX496-modified cells persisted long-term in participants. Although levels of modified cells reduced over time, they were detectable in all patients for 30 weeks post-infusion to beyond a year in some. No serious adverse events have been reported and CD4 counts have remained stable post-STI compared to baseline. Interestingly two patients exhibited delayed viral recrudescence – viral breakthrough was observed respectively at 1 and 4 months post-STI. Return of viremia following treatment interruption of HAART normally occurs within 2 months [20-22] so viral suppression in the absence of HAART for 4 months was unusual. Further, the viral load set point following recrudescence was below the pre-HAART baseline value except in one patient. This amounted to a significant increase in the numbers of patients experiencing such a lowering of viral loads post-STI compared to observations made before in a similar setting of HAART interruption [23].

Given these promising results, it was important to investigate if effects of antisense on the post-STI evolving viral populations in the patients could be measured. Molecular effects of antisense on virus such as A-to-G mutations and deletions in the antisense-target region have been documented previously in cell culture experiments [17]. These findings were recapitulated *in vivo* in a mouse model of HIV [24] and is presented in Chapter 3 of this dissertation. This study modeled the clinical trial in that modified T cells were present as a minority of cells, similar to levels in patients following reinfusion of these cells. In such a scenario, antisense-induced viral sequence signatures *in vivo*, if present, are likely to occur in a minority

of HIV variants [24]. Therefore deep sequencing was used to analyze virus from patient plasma samples, the results of which are detailed in this chapter.

Results

Description of clinical samples

All 8 patients evaluated in the phase I/II VRX496 clinical trial [19] were analyzed longitudinally (Table 4.1) (referred to as the VRX group). Patient 218 was the only participant to exhibit a rise in viral set point post-STI relative to baseline. To investigate antisense pressure, we considered only those post-STI time-points after viral recrudescence and before patients restarted on HAART. Successive time points assayed were separated by 4-8 weeks in an effort to capture possible viral signatures at any point during STI.

For each patient we included the earliest time-point post-recrudescence available for analysis, because antisense pressure might be most detectable at the point of viral breakthrough. Virus was detectable in most patients by week 4 post-STI [19]. The earliest sample analyzed corresponded to week 6 (Table 4.1). Patient 201 was an exception as the earliest time-point available was for week 14. For patient 215, recrudescence was relatively delayed at 6 weeks but the earliest sample dated to 18 weeks. Patient 252 was of particular interest having suppressed virus at undetectable levels for 4 months post-STI and had a sample for week 18, which was when virus rebounded. All patient time-points studied, together with corresponding viral load and CD4 counts, are listed in Table 4.1.

We included as controls HIV-infected patients who participated in a STI trial at the Wistar Institute, Philadelphia [20] designed to evaluate effects of multiple intermittent treatment interruptions (TI). Participants were randomized into two

cohorts. One group experienced three TIs of fixed term with intervening periods of HAART resumption. Throughout this duration the other group underwent continuous HAART. Both groups were then subject to an open-ended TI (OE-TI) before participants went back on HAART.

VRX patients

ID	Visit date	Viral load (copies/ml)	CD4 (cells/μl)	Time post-TI (week)	VRX cells per 10^6 PBMCs	Pyronoise OTUs
201	5/31/07	29520	730	14	1500	2663
201	6/28/07	28174	636	18	600	1815
201	8/23/07	18684	711	26	600	3760
201	10/18/07	33467	687	34	300	2719
201	12/13/07	16165	600	42	100	2756
203	8/9/07	39524	440	6	NA	4023
203	8/28/07	26925	462	10	300	2562
203	10/22/07	18463	463	18	200	243
203	12/20/07	8049	542	26	100	2199
203	2/14/08	26563	465	34	100*	2214
203	4/8/08	30982	462	42	100*	NP
204	11/5/07	26627	318	6	NA	1494
204	12/3/07	5993	381	10	300	2151
204	2/4/08	22108	301	18	100*	2337
204	3/31/08	20650	304	26	0	3901
204	5/22/08	16940	356	34	0	2900
215	6/11/09	3360	388	18	100	2958
215	8/6/09	4522	947	26	100	NP
215	10/13/09	3952	627	34	100	1744
215	12/16/09	3619	773	42	100	2148
218	6/18/09	97623	617	6	NA	3380
218	7/22/09	15497	627	10	100	2252
218	9/16/09	28061	704	18	100	249
218	11/11/09	24588	426	26	0	641
218	1/6/10	19508	519	34	NA	NP
250	5/12/09	414365	424	6	NA	6180
250	6/9/09	101458	350	10	300	3972
251	5/12/09	77545	470	6	NA	984
251	6/9/09	10518	450	10	500	NP
251	8/6/09	36921	400	18	300	580
252	10/6/09	5052	1070	18	1100	696
252	12/8/09	365196	935	26	500	4591

Matched controls					
ID	Visit date	Viral load (copies/ml)	CD4 (cells/μl)	Time post-TI (week)	Pyronoise OTUs
8	2/14/02	55812	352	4	2268
8	3/14/02	164904	228	8	1723
8	4/11/02	140203	234	12	4167
14	10/4/01	18961	858	4	1855
14	11/8/01	21246	574	9	2629
14	11/29/01	15197	694	12	1958
22	10/9/02	44673	893	4	1279
22	11/6/02	9943	953	8	1447
22	12/4/02	147500	966	12	1991
35	5/28/03	8621	464	8	1996
35	6/25/03	5291	393	12	1557
35	7/30/03	8921	318	17	2389
41	5/30/02	124747	661	8	1415
41	6/27/02	24995	467	12	1786
41	7/25/02	15464	533	16	2920
43	7/17/02	14424	496	14	1476
43	8/15/02	17247	409	18	3637
43	11/6/02	17531	488	30	2417
44	8/29/02	23519	341	4	2012
44	9/25/02	63255	362	8	2576
44	10/16/02	61417	279	11	3281
53	2/6/03	81957	531	4	2070
53	3/6/03	6572	701	8	1010
53	4/3/03	6152	630	12	1441
54	4/24/03	7860	747	4	3143
54	5/22/03	18581	644	8	2318
54	6/19/03	25870	636	12	2417

Table 4.1 Patient time-points studied.

Sample characteristics are indicated along with Pyronoise OTU numbers recovered by denoising raw 454/Roche pyrosequence reads obtained for a given sample. TI: Treatment interruption; NA: Not available; NP: Not pyrosequenced due to insufficient DNA; 100*: Below limit of detection

For matched controls, longitudinal samples were obtained from nine patients from the OE-TI period. Similar to VRX496 patients, time-points selected for analysis were at least 4 weeks apart with the earliest being 4 or 8 weeks following OE-TI – except for patient 43, in which case the first available time-point was for week 14. Details for control samples including viral loads and CD4 counts are mentioned in Table 4.1.

Patients 35 and 53 belonged to the group with continuous HAART before OE-TI, whereas the rest were from the group with three TIs before OE-TI.

Viral RNA amplification and pyrosequencing

RNA was extracted from plasma samples for each patient/time-point combination listed in Table 4.1. HIV RNA was reverse-transcribed and amplified following which PCR products were purified and pyrosequenced. The amplification was targeted to the region of *env* targeted by antisense (*env*AS) and involved a modification of the scheme described earlier [24]. To take advantage of longer read lengths offered by the 454/Roche Titanium platform, the length of the two shorter *env*AS amplicons was increased. This also allowed inclusion of more positions outside the antisense-target region for comparison to targeted nucleotides. On the basis of earlier observations that antisense-mediated deletions were small (<100 bases) [24], we anticipated that the two new *env*AS amplicons would be sufficient to detect antisense signatures, if any.

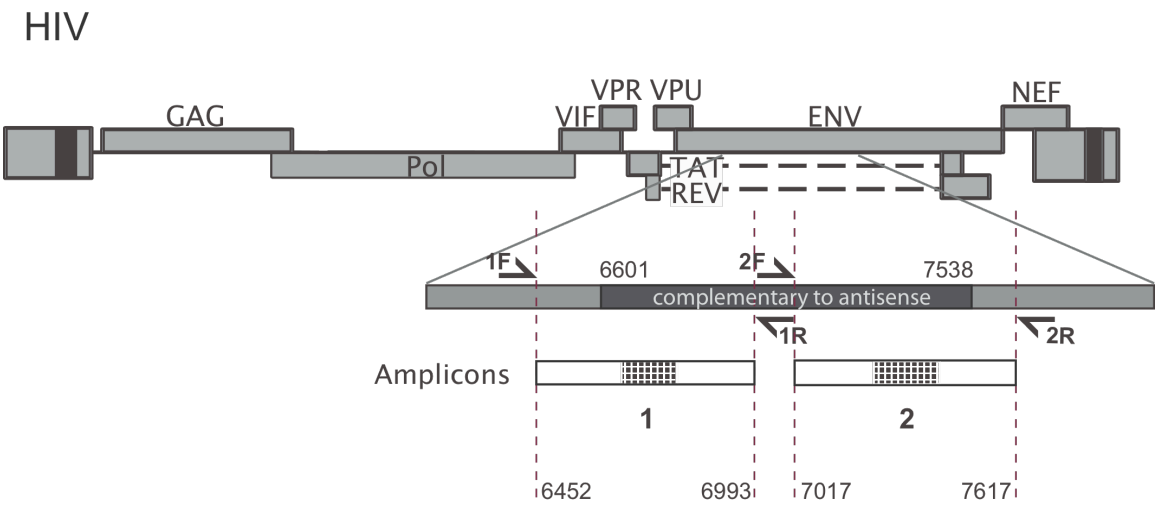


Figure 4.2 Representation of the amplification scheme in the context of the HIV_{NL4-3} genome. The blow-up shows the *env*AS target region and the amplicons designed for this study. Both

amplicons include adjacent non-target bases. Amplicons were sequenced from both directions. The first 5' 200 high quality bases for pyrosequence reads correspond to the white parts at either end of each amplicon. All numbering pertains to HIV_{NL4-3} genome positions.

The current study required amplifying virus from different patients potentially with heterogeneous HIV sequence backgrounds. Under the assumption that patients harbored HIV subtype B, which is the predominant form in the USA, the new *env*AS primers were redesigned to sites of high conservation for HIV subtype B. This improved amplification efficiency for the patient samples as evidenced by gel electrophoresis. The coordinates of the *env*AS amplicons used in this study are illustrated in Figure 4.2. For amplification, *env*AS primers were barcoded to enable multiplex pyrosequencing of samples [25, 26]. The 454/Roche microtiter sequencing plate can be physically separated into quadrants, which enabled recycling of barcodes 3-4 times over the sample set – barcode-primer combinations that performed best during testing were used more often. Complete primer and barcode sequences are cataloged in Table 4.2.

Both *env*AS amplicons were pyrosequenced in two directions generating reads originating with either the forward or reverse primers in each case. After filtering for quality, a total of 218,756 pyrosequence reads were obtained over all patient/time-point and primer combinations. To remove sequencing errors, and cluster reads into unique viral variants – termed operational taxonomic units (OTUs) – the denoising algorithm Pyronoise was used [27] according to specifications in Chapter 2. This condensed the data into 127,290 OTUs over which further analysis was carried out. OTUs are enumerated according to sample in Table 4.1. A few samples were not pooled for the pyrosequencing reaction due to undetectable DNA following amplification.

(a)

*env*AS primer	Coordinates From	To	Sequence 454 Adapter	HIV template specific part
1F	6432	6451	GCCTCCCTCGCGCCATCAG	CACATGCCTGTGTACCCACA
1R	7017	6994	GCCTTGCCAGCCCGCTCAG	TCTTCTGCTAGACTGCCATTTAAC
2F	6993	7016	GCCTCCCTCGCGCCATCAG	TGTTAAATGGCAGTCTAGCAGAAG
2R	7639	7618	GCCTTGCCAGCCCGCTCAG	CATATCTCCTCCTCCAGGTCTG

(b)

Sample ID	Visit date	Barcode Amplicon 1	Amplicon 2
201	6/1/11	TCGACATG	TCGACATG
201	6/29/11	TCGACATG	TCGACATG
201	8/24/11	TCGACATG	TCGACATG
201	10/19/11	CAGTCAGT	CGTACGAT
201	12/14/11	TCGACATG	TCGACATG
203	8/10/11	CAGTCAGT	CACTGTGA
203	8/29/11	GCATATCG	GCATATCG
203	10/23/11	GCATATCG	GCATATCG
203	12/21/11	CAGTCAGT	CACTGTGA
203	2/15/12	GCATATCG	GCATATCG
203	4/9/12	GCATATCG	GCATATCG
204	11/6/11	ACACACTG	TGAGCACT
204	12/4/11	ACGACATC	CACTGTGA
204	2/5/12	ACACACTG	TGAGCACT
204	4/1/12	ACACACTG	TGAGCACT
204	5/23/12	ACACACTG	TGAGCACT
215	6/12/13	CATCGTAG	CATCGTAG
215	8/7/13	CATCGTAG	CATCGTAG
215	10/14/13	CATCGTAG	CATCGTAG
215	12/17/13	CATCGTAG	CATCGTAG
218	6/19/13	ATATCGCG	CGTACGAT
218	7/23/13	GCTACGTA	GCTACGTA
218	9/17/13	GCTACGTA	GCTACGTA
218	11/12/13	GCTACGTA	GCTACGTA
218	1/7/14	GCTACGTA	GCTACGTA
250	5/13/13	ACGACATC	GAGTACAG
250	6/10/13	ACGACATC	GAGTACAG
251	5/13/13	ATCGATGC	CGTACGAT
251	6/10/13	ATCGATGC	CGATGCTA
251	8/7/13	ATCGATGC	CGTACGAT
252	10/7/13	CGATATGC	CTAGTGCA
252	12/9/13	CGATATGC	TGAGTCAC

Sample ID	Visit date	Barcode	
		Amplicon 1	Amplicon 2
8	2/15/06	CACTACAG	CAGTCAGT
8	3/15/06	CACTACAG	AGTCGTCA
8	4/12/06	CACTACAG	CAGTCAGT
14	10/5/05	CGTACGAT	GTACACGT
14	11/9/05	CGTACGAT	GTACACGT
14	11/30/05	CGTACGAT	GTACACGT
22	10/10/06	TAGCGCAT	CAGTCAGT
22	11/7/06	TAGCGCAT	GAGTGACA
22	12/5/06	TAGCGCAT	CAGTCAGT
35	5/29/07	GAGTGACA	GAGTGACA
35	6/26/07	GAGTGACA	GAGTGACA
35	7/31/07	GAGTGACA	GAGTGACA
41	5/31/06	TCACTGTC	CTACGATG
41	6/28/06	TCACTGTC	CTACGATG
41	7/26/06	TCACTGTC	CTACGATG
43	7/18/06	TGAGTCAC	TCTGTGAG
43	8/16/06	TGAGTCAC	TCTGTGAG
43	11/7/06	TGAGTCAC	TCTGTGAG
44	8/30/06	CACTGTGA	TCTGTGAG
44	9/26/06	CACTGTGA	AGTCGTCA
44	10/17/06	CACTGTGA	AGTCGTCA
53	2/7/07	CGATGCTA	AGTCGTCA
53	3/7/07	CGATGCTA	TCACGAGT
53	4/4/07	CGATGCTA	TCACGAGT
54	4/25/07	CTAGTGCA	TCACGAGT
54	5/23/07	CTAGTGCA	TCACGAGT
54	6/20/07	CTAGTGCA	GTACACGT

Table 4.2 Oligonucleotide sequences used for amplification.

(a) Complete sequences for primers including the 454 adapters, barcodes and HIV-specific parts are indicated. Primer coordinates are based on numbering for the HIV-1_{NL4-3} genome. Barcodes are placed between 454 adapter and HIV template-specific sequence. **(b)** Barcodes assigned to patient time-point combinations are listed for each sample.

Framework to estimate A-G error rates and deletions

To determine enrichment for A-G transitions and deletions within the *env*AS target region, if any, it was important to carefully interpret base changes in the OTUs. For this, two things were necessary: 1) a reasonable multiple sequence alignment (MSA)

wherein all OTUs would be lined up base-by-base, and 2) an appropriate reference sequence against which to measure base substitutions and deletions.

MSA for OTUs corresponding to each *env*AS primer was generated on the same principles as described in earlier chapters (also see Methods). One important difference was that to account for patient viral background heterogeneity in the *env*AS region, individual MSAs were first constructed for each patient. Also for the same reason, the majority OTU identified within a given patient was used to guide creation of these MSAs, rather than any standard reference. Next, all MSAs were brought in the context of HIV-1_{NL4-3}, which is the reference genome on which *env*AS is based. Subsequently, the *env*AS sequence was used to demarcate non-target and target bases for each read for each patient. A consensus was generated for each patient and used as the reference to help interpret changes in HIV sequences within a given patient. This was a reasonable approach to detect *env*AS-mediated effects, which were expected to be in a minority and thus potentially different from the consensus.

Preliminary assessment of A-G enrichment

As a first approach, an overall consensus sequence was created for each patient and for each *env*AS primer set from OTUs sampled over all time-points. The consensus was created two ways – by considering the frequency of isolation (or "weight") of the OTUs, or not. We call this the weighted or un-weighted consensus respectively. Accordingly two modes of analysis were carried out: 1) weighted, at the level of total reads, by replicating each OTU as many times as its weight, and 2) un-weighted, at the level of OTUs, treating each OTU equally as if they had the same weight. The "de-replication" step condensing reads into unique viral variants corresponds to denoising and clustering of reads by Pyronoise.

Since we expect *env*AS effects to be in the minority, our anticipation was that these would show up better in the un-weighted analysis. For a given OTU, any base that was different from the consensus base at that position was characterized to be a base change. Proceeding as in chapter 3, base-change proportions per OTU were plotted for both VRX and control for all possible nucleotide substitutions. This was first done for the *env*AS-target region. Unlike that following HIV-1_{BaL} challenge in *env*AS-treated mice [24], we did not find outliers that were more enriched for A-G changes in the VRX patients compared to control patients (data not shown). G-A, which arises due to APOBEC action on HIV-1, was the highest overall as well as in terms of outliers in both groups – this recapitulates earlier findings [24] and validates the approach of using the consensus to estimate nucleotide changes.

When pooled over groups, we found the proportion of OTUs enriched for A-G was higher in control patients at levels requiring a minimum of 5% or 10% change. Only at a level needing 15% or more A positions to change to G, relatively more OTUs were recorded in the VRX group. Closer inspection revealed that patient 203 alone contributed approximately 95% of the numbers in the VRX group at this level. Thus, although there was some enrichment of OTUs with high A-G rates in the VRX group, effects were not equally strong in all patients. This is in contrast to that observed in the mouse study for the treated cohort challenged with HIV-1_{BaL} [24]. To improve sensitivity, we restricted analysis for base changes to informative OTUs, which harbor at least one change of the given type (as in chapter 3). This did not lead to any further enrichment of OTUs in the VRX cohort at the different levels for A-G changes examined. A weighted analysis also did not enrich for A-G changes in VRX patients (data not shown).

Alternative approaches to evaluate A-G changes

We considered the possibility that over time immune responses or residual drug effects in both groups of patients could lead to outgrowth of mutations at some sites

at a high rate. These could add significant noise to the data while attempting to interpret the potentially more subtle changes due to *env*AS, making detection of such minority signatures difficult. Indeed, repeated base substitutions were observed longitudinally at a few sites while inspecting the MSAs. To account for this, we modified our analysis. A consensus was determined for each time-point, instead of a single consensus over all, and MSA positions where the majority base changed between time-points were discarded. As determined by Fisher's exact test (one-sided), this showed significant enrichment of OTUs in VRX patients at A-G change levels of 10% and 15% ($P < 10^{-12}$ and $P < 10^{-14}$ respectively) but not at 5% (data not shown). Statistical test was based on pooling of OTUs over all patients within each group and results were similar whether we considered all sequences or just informative ones. As earlier, patient 203 alone contributed a vast majority of the OTUs with high rates of A-G. As before, performing a weighted analysis did not improve signal detection and on the contrary, the enrichment observed at the 10% level was lost. For all further analysis, only the un-weighted approach was considered.

The earlier time-points sampled in VRX patients have the highest counts for VRX496-modified cells [19] (also see Table 4.2). Modified cell levels fall over time in all patients. Also, once there is viral breakthrough, the expectation is that the fittest forms will gradually dominate the HIV quasi-species with time under given conditions. For these reasons, it is reasonable to expect *env*AS effects to be present and most readily detectable only toward the beginning of the post-STI period at the point of recrudescence. We therefore repeated all analysis by considering only the first time-point for all patients. To estimate changes, in addition to an overall consensus like earlier, we also considered a "founder" consensus built from OTUs detected in the first time-point.

As a consequence of custom filtering during the pyrosequence data acquisition step, error increases in the 3' bases of reads (chapter 2). This could confound estimates of

true base changes, especially those generated in low frequencies. To account for this, we duplicated all analyses for the *env*AS-target region as before except considering only the first 100, and then the first 50, 5' positions for the MSAs for each primer set. This restricts analysis to increasingly high quality regions of the reads (Figure 4.2; white area within amplicons).

When considering high quality bases and constraining analysis to the first time-point and using founder consensus for interpreting changes, we found enrichment of OTUs in the VRX group by Fisher's exact test at all A-G change levels examined ($P < 10^{-5}$ at 15%, $P < 10^{-67}$ at 10%, and $P < 10^{-41}$ at 5%). This was not the case with other base changes. Importantly, we found a significant increase in numbers of OTUs with A-G among the VRX patients as estimated by Mann-Whitney test. (P = 0.114 at 10% and P = 0.046 at 5%). Enrichment of OTUs in VRX patients at the 5% level as determined by Mann-Whitney test for all base changes is given in Table 4.3a. Apart from evidence of the A-G effect in the initial time-point, we also recorded significant excess of OTUs with G-C substitutions for the third time-point. The G-C enrichment, however, did not survive the correction with non-target region substitution rates as described below.

(a)

Within Antisense

1	To A	To T	To G	To C
From A		0.760	0.046	0.405
From T	0.977		0.240	0.629
From G	0.697	0.212		0.240
From C	0.519	0.593	0.407	

3	To A	To T	To G	To C
From A		0.523	0.697	0.703
From T	0.772		0.864	0.772
From G	0.836	0.344		0.025
From C	0.994	0.568	0.736	

2	To A	To T	To G	To C
From A		0.568	0.568	0.303
From T	0.568		0.967	0.612
From G	0.836	0.164		0.612
From C	0.772	0.656	0.736	

All	To A	To T	To G	To C
From A		0.593	0.336	0.371
From T	0.970		0.882	0.593
From G	0.697	0.407		0.100
From C	0.815	0.444	0.629	

(b)

Rate Differentials

1	To A	To T	To G	To C
From A		0.556	0.057	0.084
From T	0.271		0.729	0.336
From G	0.084	0.729		0.629
From C	0.444	0.729	0.118	

3	To A	To T	To G	To C
From A		0.344	0.264	0.388
From T	0.344		0.772	0.523
From G	0.164	0.998		0.388
From C	0.836	0.344	0.432	

2	To A	To T	To G	To C
From A		0.736	0.432	0.344
From T	0.612		0.943	0.568
From G	0.523	0.388		0.697
From C	0.836	0.982	0.112	

All	To A	To T	To G	To C
From A		0.371	0.212	0.185
From T	0.556		0.882	0.444
From G	0.303	0.697		0.444
From C	0.729	0.556	0.185	

Table 4.3 *P* values for enrichment of base changes in *env*AS target region among VRX patients. **(a)** Enrichment in VRX patients of OTUs containing at least 5% base changes within the *env*AS target region. **(b)** Significant excess of rate differentials in VRX patients. For each nucleotide substitution, the difference in rates between target and non-target parts was determined per patient. In both **a.** and **b.** Mann-Whitney *P* values are listed. Analysis was performed separately for time-points 1-3 and for all time-points pooled. Time-points are indicated in a shaded box to the upper left of each panel. Significant *P* values, or that closest to significance as in **(b)**, are highlighted by white lettering.

Base change comparisons between envAS target and non-target regions

Given the above trend for A-G transitions, we investigated if normalizing against changes in the non-target region could reinforce this signal. Earlier work had reported preferential *env*AS-mediated A-G transitions in the *env*AS-target as opposed to the adjoining non-target region [17, 24]. For the current study we designed our amplicons to include 240 adjacent non-target bases (with a variation of a few bases across patients) – 150 to the left and 90 to the right of the 937 base long *env*AS-target region (Figure 4.2). This made it feasible to interrogate whether comparisons of nucleotide substitution rates within patients between target and non-target parts might bolster the A-G signal in the VRX group.

We repeated the analysis described in previous sections for the non-target regions. Although A-G rates varied across patients, there was no evidence of any difference between the control and VRX groups by Mann-Whitney test. We proceeded to calculate differences in target and non-target region base change rates. First, we identified appropriate nucleotide positions for this analysis such that non-target and target bases would be matched for quality. This was relevant as the A-G effect best showed up when restricting analysis to higher quality bases in the target region. The custom filtering performed during sequence data recovery from the 454/Roche sequencer yields high quality bases for the first ~175-200 5' positions in the direction of sequencing with errors increasing further 3' (chapter 2). For each *env*AS primer set, we selected the first 200 5' positions – indicated by the white parts inside the amplicon boxes in Figure 4.2.

Over all 4 *env*AS primer sets, we thus considered 800 nucleotide positions. These included all 240 non-target positions amplified and 560 target positions. We compared to the founder consensus to determine base changes. Base change rates were calculated separately for target and non-target parts for each patient. By taking the difference in rates between the two parts, we estimated relative base changes in the target region. This was done over all possible base changes.

Our hypothesis was that A-G rate excess would be more in the VRX group, so we performed a one-sided Mann-Whitney test to check for higher A-G rate differentials among VRX patients, which yielded $P = 0.057$ for the first time-points (Table 4.3b). This trend for A-G was not detectable for other time-points or when all time-points were pooled. Notably, no other base change recorded a higher significance across any time-points (Table 4.3b). The distribution of rate differentials for the first time-point for all base changes is shown in Figure 4.3.

We repeated our analysis by considering the first 250 5' positions for each *env*AS primer set. This gave a total of 1000 nucleotide positions – 240 of them non-target

and 760 target. Within the target parts, Fisher's test revealed OTU enrichment at different levels of A-G turnover for the initial time-point as before. Also like previously, Mann-Whitney test indicated significance across all patients at low levels of A-G transitions (1-2%). Testing for rate differentials between target and non-target parts preserved the trend for A-G in VRX patients for the initial time-point (P = 0.069; Mann-Whitney test – data not shown). Again, this effect was not visible for any other time-point. This reproduces observations recorded earlier in this section with a different selection for high quality bases. Thus we find an indication of possible *env*AS-induced A-G transitions in this study.

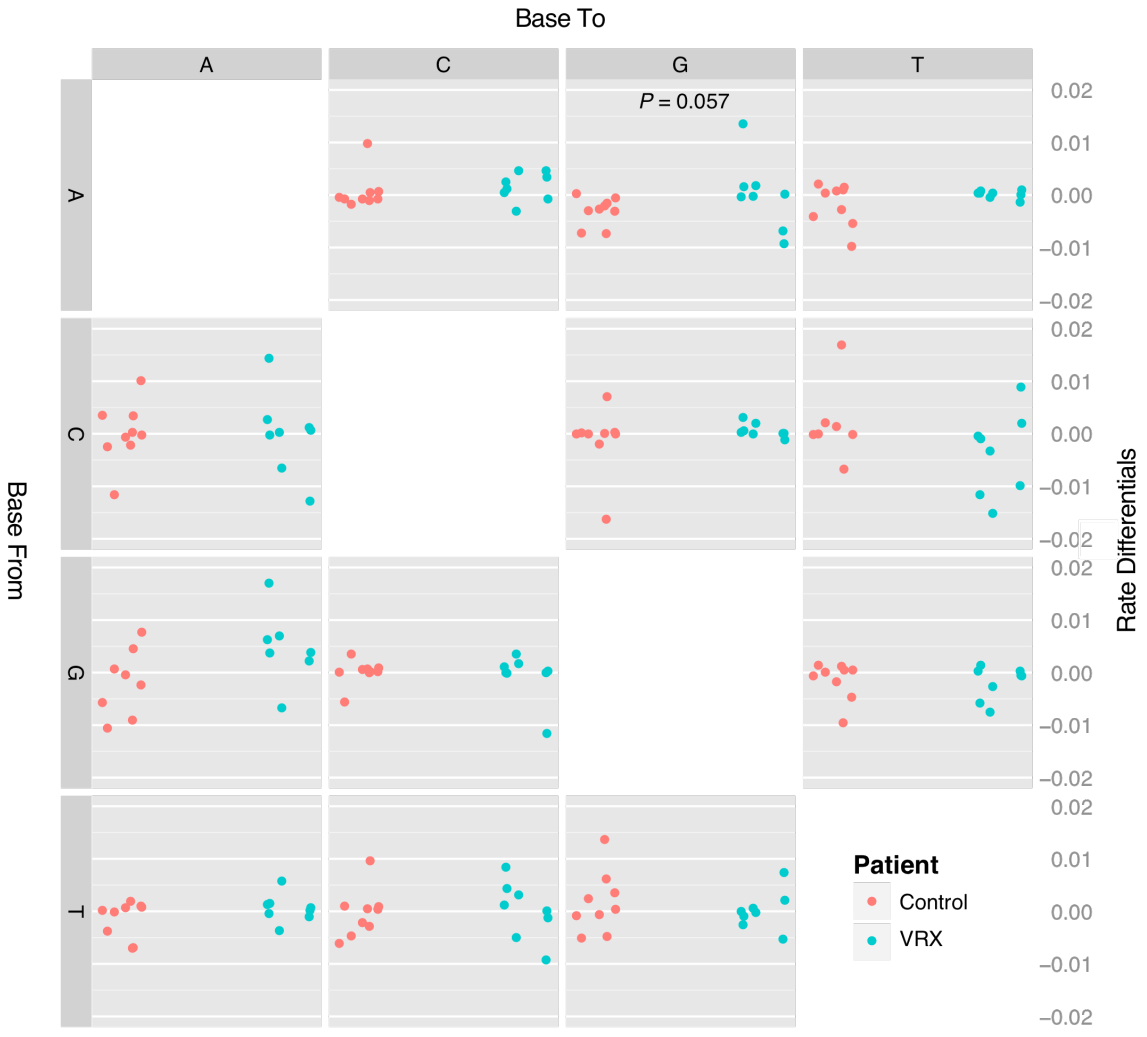


Figure 4.3 Differences in rates between *env*AS target and non-target parts for each base change per patient.

All possible base changes are shown. For A-G changes, there is a near significant trend for higher rate differentials among VRX patients compared to controls (see the panel for Base From 'A' and Base To 'G'). Two outliers each for G-A and C-T panels have been removed for visualization purposes.

Distribution of deletions

We next analyzed if *env*AS also led to higher numbers of target region deletions in viral populations among VRX patients. On similar lines as the A-G analysis, we identified deletions with the help of founder consensus and using criteria for deletions as defined in the mouse study (Chapter 3). The distribution of variants with deletions is shown in Figure 4.4. We did detect enrichment for deletion-bearing sequences in the VRX group ($P < 10^{-6}$; Fisher's exact test, one-sided), however all VRX patients did not support this ($P = 0.44$; Mann-Whitney test, one-sided). Also, in contrast to the A-G effect, there was no excess of deletions in VRX group when only the initial time-points were evaluated by Fisher's test.

Researchers studying *env*AS pressure in tissue culture reported high proportion of variants containing deletions in the target region while analyzing breakthrough virus [17]. We considered the possibility that unlike A-G effects, deletions might not be a minority signature. In this case, deletions in breakthrough variants could be reflected in founder consensus making it less useful in interpreting missing bases. Therefore all analysis was repeated using an overall consensus. This yielded significance over OTUs from all time-points pooled as well as for the first time-point separately ($P < 10^{-64}$ and $P < 10^{-21}$ respectively; Fisher's test). Still, Mann-Whitney test failed to support enrichment for deletions over all VRX patients in the per subject analysis. Performing a weighted analysis either with founder or overall consensus did not change findings. We thus conclude that evidence for deletions was not as robust as for A-G transitions, and was likely anecdotal in a few VRX patients, primarily VRX201.

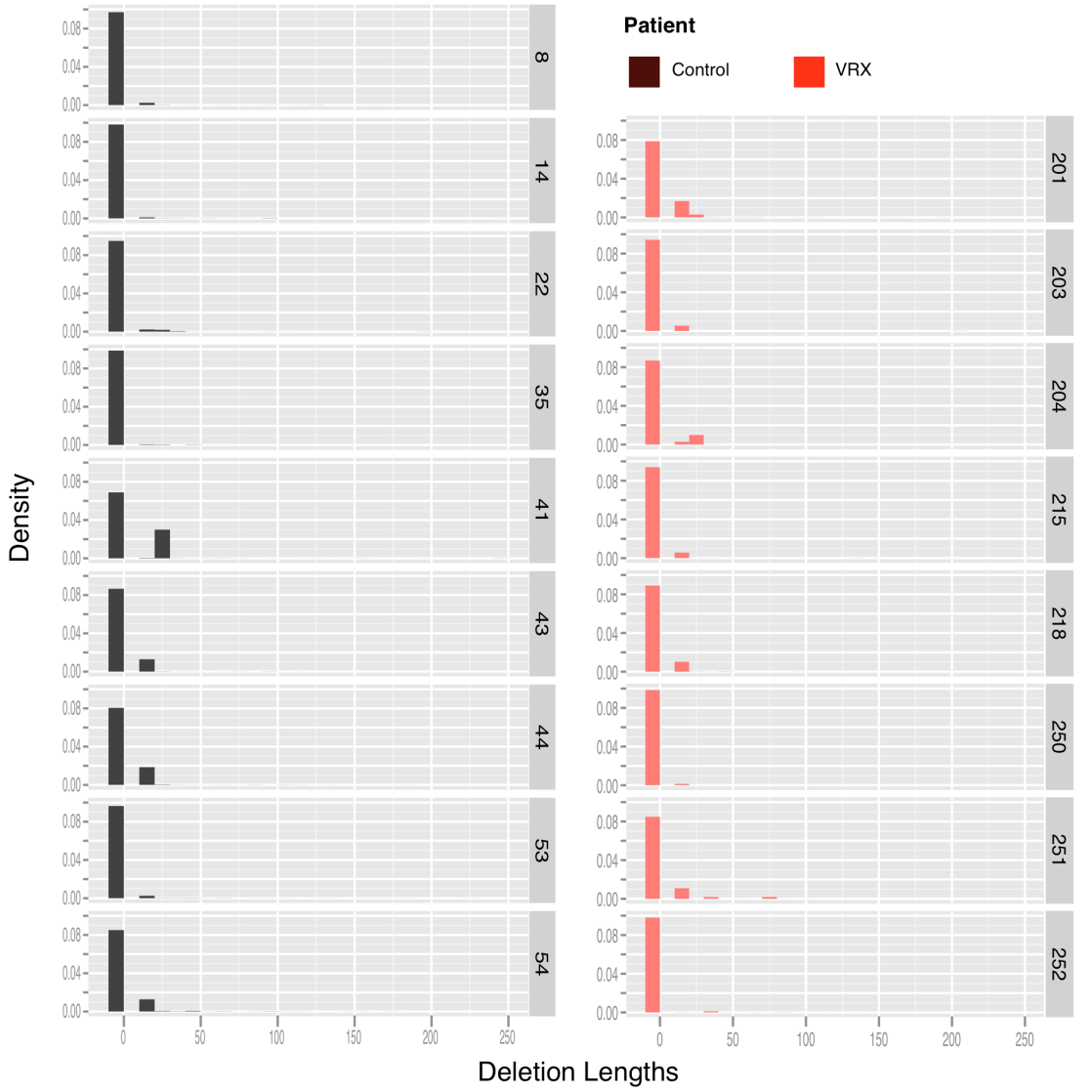


Figure 4.4 Histogram of OTUs binned by total deletion lengths.

Relative proportions, indicated by density along the y-axis, are plotted for a bin size of 10. Thus proportion OTUs in each bar corresponds to density*10. Tall bars to the left of 0 represent OTUs lacking deletions. Distributions shown are over all time-points.

Discussion

In previous work with a mouse model, we documented evidence of *env*AS pressure on circulating virus *in vivo* [24]. Understandably, these signatures were rare given that *env*AS-modified cells were in a minority. Nevertheless this highlighted the potential of *env*AS as an antiviral payload for gene therapy approaches. In the current study, we investigated patients receiving *env*AS-modified cells for possible recapitulation of observations from the mouse model. We found evidence for modestly higher A-G transitions in the VRX patients although the same was not true for deletions. Still, our findings suggest that *env*AS did indeed exert some pressure on evolving virus in the subjects studied.

The A-G effect in VRX patients was detectable only upon analysis of high quality nucleotide positions in the MSAs rather than all available bases across all reads. Yet, there are quite a few indicators that we observed an authentic *env*AS-induced signature. First, the A-G effect was only apparent for the first time-point. This tracks with declining proportions for *env*AS-modified cells in patients over time. Once virus breaks through, it is likely that any *env*AS effect would be progressively washed out as fewer virus interacts with dwindling numbers of modified cells. This would make *env*AS signatures harder to detect at later time-points.

A-G effects were most perceptible when interpreted against the founder consensus. This is consistent with the expectation that *env*AS pressure would be discernible at the level of breakthrough virus rather than in later forms. The founder consensus best represents the recrudescent HIV background – this is also the viral template that is potentially free of base changes due to immune or residual drug pressure that could confound A-G measurements. Another point to note is that although effects were weak, our reporting of the trend was based on stringent criteria. The A-G signal was replicated over the whole VRX group and not a consequence of one or a minority of patients exhibiting enhanced effects. Furthermore, in comparison to the mouse study where we used control animals, here we included the non-target

region in addition to control patients to assess the signal. Thus we can say our findings are robust.

Why was the A-G signal not as strong as previously observed in the mouse model? In the mice, modified cells numbered 4-10% of all engrafted T cells. In the clinical trial, effective levels of *env*AS-modified cells around the time of STI was less than 1% in all patients. For the initial time-points sampled in this study, corresponding numbers were even lower with some patients having no more than 0.05% [19] (also see Table 4.2). In addition, this analysis tabulated *env*AS effects across unrelated patients with heterogeneous viral backgrounds – magnitude of *env*AS pressure exerted is likely to vary with template making detection difficult. This is in contrast to the mouse study where all animals were infected with clonal HIV-1 populations resulting in straightforward measurements for base changes.

Given these limitations, it is actually surprising that we recorded signs of *env*AS action reproducibly over many VRX patients. Presumably with more power by way of more enrolled participants in the trial, we could have detected a statistically robust signal. As concerns deletions, we failed to record convincing evidence in support of *env*AS activity. Deletions have a higher fitness cost and were barely enriched in the presence of *env*AS even in the mouse model. Also, *env*AS levels were lower in the clinical trial. Together, this could explain why it may have been difficult to identify *env*AS-induced deletions. Nonetheless findings reported here are significant as for the first time in humans we show that an anti-HIV gene therapy strategy actually impacts circulating virus. With improved persistence of modified cells, it may be realistic to expect effective control of virus by antisense.

Materials and Methods

Amplicon design

HIV-1_{NL4-3}, which is a standard subtype B strain, was used as a reference genome. The sequence for HIV-1_{NL4-3} was obtained from the nucleotide database at National Center for Biotechnology Information (accession ID M19921 at http://www.ncbi.nlm.nih.gov/nuccore/). A multiple sequence alignment (MSA) for 454 HIV subtype B *env* sequences was downloaded from the HIV database at Los Alamos National Labortatory (http://www.hiv.lanl.gov/) and positions with high conservation determined. To this MSA were aligned sequences for the VRX496 *env* antisense (*env*AS) payload and HIV-1_{NL4-3} reference. This helped mark coordinates of the *env*AS amplicons designed earlier [24] on the MSA. Possible primer landing sites were identified within a region 200 bases upstream and downstream respectively of previously designed *env*AS forward and reverse primers. Sites with preferably 90% or more conservation over all positions (with a minimum of 70%) were examined for their potential to serve as primers with the help of Primer3 program (http://primer3.sourceforge.net/).

Sample preparation for pyrosequencing

Library preparation for pyrosequencing was along similar lines as in previous chapters. For each patient sample, 500μl of plasma was added to 500μl of phosphate buffered saline (PBS) to prepare 1ml aliquots for ultracentrifugation to pellet virus, under conditions specified in chapter 2. Next, 900μl of supernatant was carefully withdrawn and pellet resuspended in the remaining 100μl. PBS was added to make up volume to 100μl if there was any shortfall. RNA was purified from this 100μl volume of concentrated virus as previously. HIV-1 RNA was amplified by RT-PCR using composite primers with 454 adapters and barcodes (Table 4.2). RT-PCR cycling was identical to that in chapter 3 methods, except that following the touchdown steps, 25 (instead of 20) cycles were performed at the annealing

temperature of 53°C for a total 35 cycles of amplification. RT-PCR products were diluted 1:1 with water for a total volume of 75μl for each sample, from which DNA was purified and concentrated using Agencourt AMPure XP (Beckman Coulter). DNA was then quantified, pooled and pyrosequenced as before.

Pyrosequence data processing

Sequencing was performed on the 454/Roche Titanium platform at University of Pennsylvania. For pyrosequencing data acquisition, to increase the throughput of reads, a custom filtering was performed instead of the default one – as previously stated in chapter 2. According to steps outlined there, reads were filtered for quality, assigned to source samples and denoised using Pyronoise to remove pyrosequencing errors [27]. Multiple sequence alignments (MSAs) were generated from pair-wise global alignments (PWAs) of Pyronoise solution OTUs with a reference (chapter 2). Also as previously, MSA positions that were likely artifactual were discarded based on low read coverage (<5%). For each patient, four MSAs was generated, one for each *env*AS primer, over all time-points. To improve alignment quality, the predominant sequence isolated from a patient was used, instead of a pre-selected reference, to guide the PWAs of all denoised OTUs for that patient. Additionally for quality purposes, only the first 500 bases for reads were considered for alignment.

A-G and deletion analysis

From MSA, patient consensus over all time-pts was constructed. Next a modified consensus (consensus-Mod) was constructed by removing gaps and considering no more than the 5' 350 bases. An MSA was generated with consensus-Mod sequences for all patients with all *env*AS primers and pNL4-3 reference. The reference was used to delineate anti-sense target and non-target portions of consensus-Mod

sequences. This information was in turn used to divide the patient MSAs into target and non-target parts.

Within-patient consensus and founder sequences, weighted and un-weighted, were used to estimate base changes – the proportion of A positions in the consensus that were G in a given OTU was used as an estimate for A-G transition rate for that OTU. For each OTU, all possible base substitution rates were similarly determined. Consensus sequences were also used to identify deletions in OTUs according to qualifying criteria defined in chapter 3.

For assessing enrichment of OTUs in VRX patients harboring a certain level or more of A-G changes or deletions, one-sided statistical tests were performed following previous studies [17, 24] which laid the basis for a directional null hypothesis. Fisher's exact test was used to determine excess in VRX group pooled over all patients whereas Mann-Whitney test was used to assess replication of enrichment over all VRX patients (more stringent un-pooled analysis). In all cases, data from control helped evaluate enrichment in VRX patients. All statistical programming and tests were performed in the R computing framework (http://www.r-project.org/).

For paired analysis of base changes between the non-target and target regions, high quality target positions were selected to match with non-target positions that were all within high quality parts of pyrosequence reads. For each possible base change, substitution rates were calculated separately for the target and non-target region per patient. The difference of rates between target and non-target parts was then determined.

Acknowledgments

We thank Carl H June (Department of Pathology and Laboratory Medicine, University of Pennsylvania) and Gwendolyn Binder-Scholl (Adaptimmune LLC,

Philadelphia, PA) for their collaboration and Luis J Montaner (Wistar Institute, Phialdelphia, PA) for providing control patient samples. We are grateful to members of the Bushman laboratory for discussions and help, especially Frances Male for technical assistance with sample preparation for pyrosequencing and Kyle Bittinger for support with Pyronoise. This work was supported by NIH grant U19 AI 082628.

References

[1]. Bukrinsky MI, Stanwick TL, Dempsey MP, Stevenson M. Quiescent T lymphocytes as an inducible virus reservoir in HIV-1 infection. *Science* 1991; **254**:423-427.

[2]. Blankson JN, Persaud D, Siliciano RF. The challenge of viral reserviors in HIV-1 infection. *Annu. Rev. Med.* 2002; **53**:557-593.

[3]. Carter CC, Onafuwa-Nuga A, McNamara LA, Riddell J,4th, Bixby D, Savona MR, et al. HIV-1 infects multipotent progenitor cells causing cell death and establishing latent cellular reservoirs. *Nat Med* 2010; **16**:446-451.

[4]. Finzi D, Hermankova M, Pierson T, Carruth LM, Buck C, Chaisson RE, et al. Identification of a Reservoir for HIV-1 in Patients on Highly Active Antiretroviral Therapy. *Science* 1997; **278**:1295-1300.

[5]. Finzi D, Blankson J, Siliciano JD, Margolick JB, Chadwick K, Pierson T, et al. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med* 1999; **5**:512-517.

[6]. Bailey JR, Sedaghat AR, Kieffer T, Brennan T, Lee PK, Wind-Rotolo M, et al. Residual human immunodeficiency virus type 1 viremia in some patients on antiretroviral therapy is dominated by a small number of invariant clones rarely found in circulating CD4+ T cells. *J Virol* 2006; **80**:6441-6457.

[7]. Deeks SG, Phillips AN. HIV infection, antiretroviral treatment, ageing, and non-AIDS related morbidity. *BMJ* 2009; **338**:a3172.

[8]. Kelley CF, Kitchen CM, Hunt PW, Rodriguez B, Hecht FM, Kitahata M, et al. Incomplete peripheral CD4+ cell count restoration in HIV-infected patients receiving long-term antiretroviral treatment. *Clin Infect Dis* 2009; **48**:787-794.
[9]. Hutter G, Nowak D, Mossner M, Ganepola S, Mussig A, Allers K, et al. Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *N Engl J Med* 2009; **360**:692-698.
[10]. Allers K, Hutter G, Hofmann J, Loddenkemper C, Rieger K, Thiel E, et al. Evidence for the cure of HIV infection by CCR5Delta32/Delta32 stem cell transplantation. *Blood* 2011; **117**:2791-2799.
[11]. Zamecnik PC, Goodchild J, Taguchi Y, Sarin PS. Inhibition of replication and expression of human T-cell lymphotropic virus type III in cultured cells by exogenous synthetic oligonucleotides complementary to viral RNA. *Proc Natl Acad Sci U S A* 1986; **83**:4143-4146.
[12]. Goodchild J, Agrawal S, Civeira MP, Sarin PS, Sun D, Zamecnik PC. Inhibition of human immunodeficiency virus replication by antisense oligodeoxynucleotides. *Proc Natl Acad Sci U S A* 1988; **85**:5507-5511.
[13]. Rhodes A, James W. Inhibition of human immunodeficiency virus replication in cell culture by endogenously synthesized antisense RNA. *J Gen Virol* 1990; **71 (Pt 9)**:1965-1974.
[14]. Sczakiel G, Pawlita M. Inhibition of human immunodeficiency virus type 1 replication in human T cells stably expressing antisense RNA. *J Virol* 1991; **65**:468-472.
[15]. Gyotoku J, el-Farrash MA, Fujimoto S, Germeraad WT, Watanabe Y, Teshigawara K, et al. Inhibition of human immunodeficiency virus replication in a human T cell line by antisense RNA expressed in the cell. *Virus Genes* 1991; **5**:189-202.
[16]. Levine BL, Humeau LM, Boyer J, MacGregor RR, Rebello T, Lu X, et al. Gene transfer in humans using a conditionally replicating lentiviral vector. *Proc Natl Acad Sci U S A* 2006; **103**:17372-17377.

[17]. Lu X, Yu Q, Binder GK, Chen Z, Slepushkina T, Rossi J, et al. Antisense-mediated inhibition of human immunodeficiency virus (HIV) replication by use of an HIV type 1-based vector results in severely attenuated mutants incapable of developing resistance. *J Virol* 2004; **78**:7079-7088.
[18]. Humeau LM, Binder GK, Lu X, Slepushkin V, Merling R, Echeagaray P, et al. Efficient lentiviral vector-mediated control of HIV-1 replication in CD4 lymphocytes from diverse HIV+ infected patients grouped according to CD4 count and viral load. *Mol Ther* 2004; **9**:902-913.
[19]. Tebas P, Stein D, Zifchak L, Seda A, Binder GK, Aberra F, et al. Prolonged Control of Viremia After Transfer of Autologous CD4 T Cells Genetically Modified with a Lentiviral Vector Expressing Long Antisense to HIV env (VRX496). *CROI* 2010; **17th Conference on Retroviruses and Opportunistic Infections**.
[20]. Papasavvas E, Kostman JR, Mounzer K, Grant RM, Gross R, Gallo C, et al. Randomized, controlled trial of therapy interruption in chronic HIV-1 infection. *PLoS Med* 2004; **1**:e64.
[21]. Touloumi G, Pantazis N, Antoniou A, Stirnadel HA, Walker SA, Porter K, et al. Highly active antiretroviral therapy interruption: predictors and virological and immunologic consequences. *J Acquir Immune Defic Syndr* 2006; **42**:554-561.
[22]. Firnhaber C, Azzoni L, Foulkes AS, Gross R, Yin X, Van Amsterdam D, et al. Randomized Trial of Time-Limited Interruptions of Protease Inhibitor-Based Antiretroviral Therapy (ART) vs. Continuous Therapy for HIV-1 Infection. *PLoS One* 2011; **6**:e21450.
[23]. Tebas P, Henry K, Mondy K, Deeks S, Valdez H, Cohen C, et al. Effect of prolonged discontinuation of successful antiretroviral therapy on CD4+ T cell decline in human immunodeficiency virus-infected patients: implications for intermittent therapeutic strategies. *J Infect Dis* 2002; **186**:851-854.
[24]. Mukherjee R, Plesa G, Sherrill-Mix S, Richardson MW, Riley JL, Bushman FD. HIV Sequence Variation Associated With env Antisense Adoptive T-cell Therapy in the hNSG Mouse Model. *Mol Ther* 2010;.

[25]. Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, et al. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2007; **2**:e197.
[26]. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, et al. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. *Nucleic Acids Res* 2007; **35**:e91.
[27]. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, et al. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 2009; **6**:639-641.

Chapter 5 CONCLUSIONS AND FUTURE DIRECTIONS

In this dissertation, I describe studies undertaken to characterize evolving virus under different selection pressures. Two lines of therapies were investigated – 1) treatment with the integrase inhibitor Raltegravir, and 2) an alternative approach involving gene therapy with an anti-HIV antisense targeting the *env* gene of the virus, which to date is the only one of its kind to be used in clinical trials. In each case, the emerging technology of deep sequencing (454/Roche) was used to examine minor signatures in the HIV populations that are indicative of how the viral quasi-species responded to treatment.

The power of deep sequencing needs to be applied prudently as it comes with the cost of sequencing errors. This dissertation outlines the development and implementation of computational and statistical approaches that rigorously inspects effects of error – this helped in stringent interpretation of rare patterns among viral variants, allowing the deduction of important clinical conclusions. This was the case whether in the estimation of low-level pre-treatment resistance mutations in the Raltegravir study or in the identification of subtle effects of *env* antisense in breakthrough viral populations. In the remainder of this section, I describe possible future experiments for each of the lines of study presented in the preceding chapters discussing possible technical improvements in parallel. I end with some observations on the nature of HIV quasi-species and implications for cure.

The Raltegravir resistance pathway switch

The drug Raltegravir (RAL) has increased treatment options for HIV patients as it inhibits HIV integrase (IN), which is a different viral target than that of other anti-HIV drugs which have existed since the mid-1990s. However, resistance is a

problem, as with other drugs. The IN genotype Q148H is more resistant to RAL than others such as N155H and Y143R; Q148H also confers increased cross-resistance to other IN-inhibitors under development. For these reasons emergence of Q148H is a major challenge to treatment. In Phase II trials with RAL [1], it was observed that frequently the resistance profile of viruses in patients switched from the N155 pathway to the Q148 one. On occasions, a switch into Q148H was also recorded starting from a Y143 mutation.

Findings in Chapter 2 in patients with an N155-Q148 switch suggest that neither mutation existed detectably prior to RAL treatment. To follow-up, more patients could be studied together with deeper sequencing approaches, such as Illumina/Solexa sequencing, or more sensitive detection methods for known polymorphisms like allele-specific PCR [2, 3]. It would also be important to find other viral signatures, if any, predicting the switch. The earlier Phase II study involved patients who started out exhibiting N155H but did not shift to Q148 [1] – including them as controls for a comparison of pre-treatment quasi-species structure could reveal such predictor patterns. In experiments described in Chapter 2, one patient did have residual levels of N155H long after turnover of the resistance profile to Q148H. Interestingly, this patient also showed delayed emergence of Q148H. One could ask if in this case the Q148H mutant was less potent (or alternatively, the N155H more fit).

A complex collection of intermediate genotypes was also identified, especially in the patient with residual long-term N155H (see Chapter 2). Some of the transient IN mutations observed were not known to be associated with major RAL resistance pathways. Future experiments could help analyze if any of these are determinants of the switch. How deep did we search in the viral populations? Chapter 2 resolves this question statistically. Better still, a unique barcode can be assigned to each viral template in the starting sample that is sequenced [4, 5]. This is made possible by generating high numbers of unique barcodes by randomizing a few nucleotides.

Such accurate measurements of the number of viral variants sampled would help map the boundaries and proportions of the components of the HIV quasi-species precisely. It remains to be seen if this detailed information could be used to inform the future course of RAL, and indeed integrase inhibitor treatment.

Story of the anti-HIV *env* antisense

Drugs can control but can't cure HIV. Of alternative strategies, gene and cell therapy approaches have gained prominence with the case of the Berlin patient providing an example of possible cure [6]. In this regard, development of T cells modified with antisense as HIV-resistant cells has also shown promise with the use of antisense targeting HIV *env* entering clinical trials. Although anti-HIV gene therapy has been shown to be safe in different settings, it has been difficult to demonstrate efficacy in terms of detectable pressure exerted on the HIV. This is largely due to patients exhibiting modest levels and low persistence of gene-modified cells. Given encouraging clinical trends with *env* antisense (*env*AS), it was important to quantify possible *env*AS effects on virus.

Once again, plumbing the depths of viral populations evolving in the presence of *env*AS with 454/Roche pyrosequencing helped in identifying the genetic pressure exerted. The effects were subtle, as one would expect in the presence of only low proportions of *env*AS-modified cells. Evidence was first gathered from a mouse model simulation of the clinical trial, the results of which are presented in Chapter 3. As a sequel, Chapter 4 describes experiments performed with patient samples. Encouragingly findings of the mouse study were recapitulated in humans. A-G nucleotide changes, which are the major *env*AS-induced signature, were detectable with high-resolution deep-sequenced data but only after judicious handling of noise generated during the experimental process.

This is the first time that in a gene therapy clinical setting we have possible evidence of pressure applied on the HIV virus. Why were effects so faint? For one, *env*AS-modified cells do not persist long in patients. This is a challenge for this field. Of note, one study has reported long-term persistence of stable levels of gene-modified cells in HIV infected patients [7]. Mechanisms underlying such durability can be harnessed to potentiate enduring levels of antisense in patients. Dosing schemes for modified cells could also be optimized with the goal of improving persistence.

Following from conclusions in the mouse study (Chapter 3), it is equally possible that visibility of *env*AS pressure could be a function of adenosine deaminases acting on RNA (or ADAR enzymes), which act on the sense-antisense duplex resulting in A-I edits that manifest as the A-G changes [8]. ADAR effects on perfectly matched double-stranded RNA templates has been extensively studied and even quantified [9]. However action of ADAR on mismatched duplexes, as would arise if circulating virus were non-complementary to *env*AS, is less understood. One study by researchers at the Stockholm University, Sweden, has proposed patterns of A-I editing in this context [10]. Along these lines, to help explain the mechanism of *env*AS effects, it could be useful to study ADAR *in vitro* with templates with different degrees and quality of mismatch, and possibly alter *env*AS to maximize ADAR recruitment.

The other concern is effectiveness of *env*AS itself on heterogeneous viral templates. HIV infection in the US is dominated by subtype B. In this context, *env*AS design was based on HIV_{NL4-3}, which is a prototype HIV-1 subtype B strain. When used therapeutically, as in the already HIV infected patients enrolled in the clinical trial, *env*AS has to act against differing circulating viral backgrounds across patients. Even if all patients had subtype B infection, one would expect considerable variation between patient viral populations. This is true especially of the *env*AS target region, which is not only part of *env* – the most diverse HIV gene – but also encompasses 3 of the 7 hyper-variable loops of *env*. Thus, it is likely that *env*AS effects were further

diluted by the target template diversity. It would be interesting to examine if effects were more pronounced in patients who had baseline consensus virus sequence that was more complementary to *env*AS. Along these lines, future experiments could evaluate antisense directed against regions of the HIV genome that are more conserved across subtypes.

Back to the quasi-species

In the studies presented in this dissertation, viral populations were analyzed over the course of applied therapeutic pressure. In such situations, evolving viral swarms are often in a state of flux, trying to equilibrate with sudden modifications in selection forces in a way that best ensures survival of the population. During the process of such adjustment, however, HIV populations are in a state of instability – it does not conform to definitions of a quasi-species, which alters the quantitative models one might apply to HIV populations.

Thus, for example, until such point when resistance profiles to ensuing drug treatment have stabilized, researchers have fewer clues as to the effectiveness and future course of therapy. Nevertheless studies in this dissertation provide instances of minority signatures that are consistent over viral populations from different individuals sharing a common treatment or exhibiting similar outcomes to therapy. In the study investigating switch in RAL resistance profiles, all patients harbored low-level Y143H at the pre-treatment baseline. Similarly, all patients undergoing antisense therapy supported evidence for *env*AS effects at a point close to viral breakthrough when antisense levels were possibly the highest. Baseline or breakthrough viruses are possibly stable populations that are yet to react to ensuing therapy pressure. These may still exhibit virtues of a quasi-species and thus patterns present in them could be stable indicators of future response such as treatment failure, as in the RAL study, or possible efficacy, as in the *env*AS trial.

The development of high fidelity and highly processive polymerases can help minimize effects of PCR-related template recombination and base mis-incorporations that often confound deep sequencing studies [11]. Use of primer ID approaches, wherein genomes in the pre-PCR starting sample are tagged with unique barcodes, permit derivation of consensus sequences for re-sampled genomes that also eliminate PCR errors [4, 5]. In parallel, recent developments in de-noising algorithms are promising a more effective accounting of error from various sources such as PCR and sequencing [12]. With these improvements in both laboratory and computational techniques, we can hope to glean even finer details on signatures inherent of the HIV quasi-species. An important goal is to establish predictors for HIV response that would suggest bottlenecking of HIV distributions as it adapts its quasi-species to therapy – bottlenecking to a narrow region in sequence space where it is easily contained. I hope studies and methods outlined in this dissertation would help in this research.

References

[1]. Miller MD, Danovich RM, Ke Y, Witmer M, Zhao J, Harvey C, et al. Longitudinal Analysis of Resistance to the HIV-1 Integrase Inhibitor Raltegravir: Results from P005 a Phase 2 Study in Treatment Experienced Patients. *International HIV Drug Resistance Workshop* 2008; **17th Meeting**.

[2]. Palmer S, Boltz V, Martinson N, Maldarelli F, Gray G, McIntyre J, et al. 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* 2006; **103**:7094-7099.

[3]. Boltz VF, Zheng Y, Lockman S, Hong F, Halvas EK, McIntyre J, et al. Role of low-frequency HIV-1 variants in failure of nevirapine-containing antiviral therapy in

women previously exposed to single-dose nevirapine. *Proc Natl Acad Sci U S A* 2011; **108**:9202-9207.

[4]. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* 2011; **108**:9530-9535.

[5]. Jabara CB, Jones CD, Anderson JA, Swanstrom R. Accurate Sampling and Deep Sequencing HIV-1 Protease Using Primer ID. *CROI* 2011; **18th Conference on Retroviruses and Opportunistic Infections**.

[6]. Allers K, Hutter G, Hofmann J, Loddenkemper C, Rieger K, Thiel E, et al. Evidence for the cure of HIV infection by CCR5Delta32/Delta32 stem cell transplantation. *Blood* 2011; **117**:2791-2799.

[7]. Walker RE, Bechtel CM, Natarajan V, Baseler M, Hege KM, Metcalf JA, et al. Long-term in vivo survival of receptor-modified syngeneic T cells in patients with human immunodeficiency virus infection. *Blood* 2000; **96**:467-474.

[8]. Wulff BE, Sakurai M, Nishikura K. Elucidating the inosinome: global approaches to adenosine-to-inosine RNA editing. *Nat Rev Genet* 2011; **12**:81-85.

[9]. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. *Nat Commun* 2011; **2**:319.

[10]. Enstero M, Daniel C, Wahlstedt H, Major F, Ohman M. Recognition and coupling of A-to-I edited sites are determined by the tertiary structure of the RNA. *Nucleic Acids Res* 2009; **37**:6916-6926.

[11]. Lahr DJ, Katz LA. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques* 2009; **47**:857-866.

[12]. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 2011; **12**:38.