Identifying genetic signatures of vaccine-induced immune responses in HIV-1 infected MRKAd5 STEP vaccine study subjects

Shyamala Iyer

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of philosophy

University of Washington

2015

Reading Committee:

James I. Mullins, Chair

Roger E. Bumgarner

John E. Mittler

Program Authorized of Offer Degree:

Microbiology

University of Washington

University of Washington

Abstract

Identifying genetic signatures of vaccine-induced immune responses in HIV-1 infected MRKAd5 STEP vaccine study subjects

Shyamala Iyer

Chair of the Supervisory Committee:

Professor James I. Mullins

Microbiology

Massively parallel sequencing technologies have been extensively applied in HIV-1 research to study the presence of minority variants. Insight gained through these technologies includes identification of minor drug resistant variants and immune escape variants. However, given the massive amounts of data generated, processing the sequences and discerning true minor variants from sequencing artifacts is important. Additionally, errors introduced during viral template amplification and incorrect quantification of templates prior to the sequencing process can further obfuscate resolving mismatch errors within the sequences. I address these issues in this dissertation. We developed a computational algorithm, CorQ, to correct specific patterns of sequencing errors and call Single Nucleotide Polymorphisms (SNPs). When coupled with additional error correction steps, we observed a 97% reduction in insertion, and deletion sequencing errors. In addition, we observed over 98% specificity in SNP detection compared to other available error correction methods. We observed reduced SNP calling specificity when

error correction programs were tested on sequences with simulated PCR amplification mismatch errors, with the highest specificity of 70% observed with a combination CorQ algorithm, highlighting the difficulty in resolving errors generated during PCR amplification. We observed over 99% concordance in consensus variants observed in multiple HIV-1 infected subjects sequenced with traditional Sanger sequencing and pyrosequencing. The majority of SNPs that were specific to subjects' pyrosequences were present at less than 2% of the subjects viral sequence population. We observed higher accuracy in variant frequencies in positions where read coverage exceeded the number of input templates.

We have applied the developed error correction algorithm and observations from SNP variant comparisons to identifying major and minor variants observed within predicted T-cell epitope regions in HIV-1 infected study subjects enrolled in the MRKAd5 STEP vaccine trial. We observed genetic signatures of immune responses primed by the vaccine on breakthrough HIV-1 sequences. We observed greater genetic distances to the vaccine sequence in breakthrough sequences from vaccine recipients than placebo recipients and this difference was most significant within T-cell epitope regions. Additionally over time, the vaccine-primed immune responses resulted in reduced epitope diversity and decreased rates of epitope evolution over time. Combined, these results strongly support the hypothesis that the MRKAd5 vaccine resulted in T-cell mediated selection occurring post-infection. The results from our study are the first evidence of vaccine-induced anamnestic pressure influencing CTL epitope evolution and epitope diversity over time during HIV-1 infection.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

I would like to express my gratitude to my Ph.D. advisor, Dr. James I Mullins for providing me with scientific guidance through the course of my Ph.D. career. His meticulousness and attention to detail have helped me immensely in the course of writing manuscripts for publications. I sincerely thank him for his patience during the course of time it has taken me to complete the requirements for my Ph.D.

I would like to thank members from the Mullins group for their help and scientific advice. I would particularly like to thank Eleanor Casey whose contribution to the various projects in this dissertation has been invaluable.

I would like to acknowledge my Doctoral Supervisory Committee, Dr. Roger E. Bumgarner, Dr. John E. Mittler, Dr. Paul Edlefsen and Dr. Gabriele Varani for their time and guidance through my research.

Most importantly, I would like to thank my daughter who has helped me grow and change in ways I never imagined and who fills my life with joy and laughter.

Chapter 1. Challenges in the development of a global HIV-1 vaccine

Introduction

Human Immunodeficiency virus (HIV) is a member of the family Retroviridae and the genus Lentiviridae [1]. HIV is the causative agent of Acquired Immunodeficiency Syndrome (AIDS). The HIV genome is composed of 9.8Kb positive sense, single stranded RNA that is reverse transcribed to viral DNA by the enzyme reverse transcriptase [1]. Between the two types of HIV, HIV-1 and HIV-2, HIV-1 is more virulent and responsible for most of the HIV infection globally, whereas HIV-2 infections are rare outside of Africa [2].

The HIV-1 RNA genome encodes the essential retrovirus genes: *gag*, *pol* and *env* as well as the additional accessory and regulatory genes *vif*, *vpr*, *vpu*, *rev*, *tat* and *nef* (Figure 1).

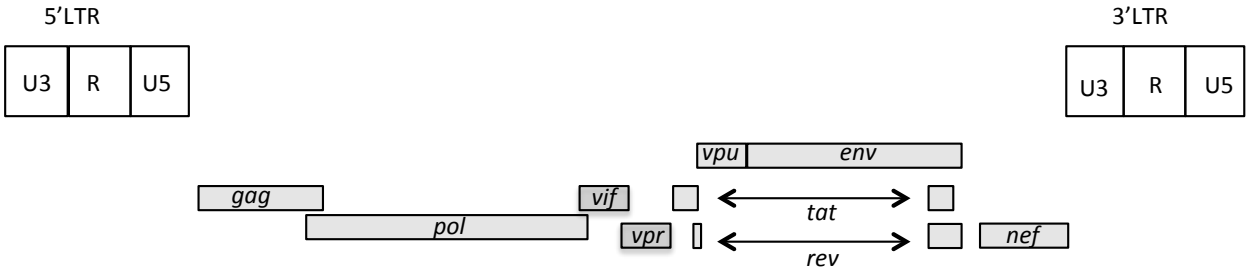

Figure 1. HIV-1 genome organization. *pol*, *gag* and *env* genes encode enzymatic and structural proteins. Regulatory gene products are encoded by *tat* and *rev* genes. The *vif*, *vpu*, *vpr* and *nef* genes encode auxiliary proteins. Long terminal repeats (LTRs) are the sites for initiation of viral RNA synthesis and necessary for proviral integration into host cell chromosomes.

Origin of diversity in HIV

HIV has several intrinsic mechanisms that ensure rapid evolution. The high number of replication cycles, error-prone reverse transcription as a result of lack of proofreading activity by reverse transcriptase, and high recombination rates between variants in an infected individual all contribute to the generation of a large population of HIV-1 variants within an individual [3-8].

The lack of proofreading activity results in a mutation rate in the range of $5.9x10^{-4}$ to $5.3x10^{-5}$ mutations per base pair in one replication cycle [9]. The HIV genome is 10^4 base pairs in length and approximately 10^{10} virions are produced per day [9], which amounts to millions of viral variants produced within an infected person in a single day. This heterogeneous population of viruses generated during replication is known as a viral quasispecies [10]. Adding to this complexity is the immune pressure from the infected individual's cellular (involving mostly T-cells) and humoral responses (antibody mediated responses) forcing selection of viral variants leading to the generation of additional viral diversity within the viral quasispecies [11-20]. The variability of HIV-1 within a host has been compared to the global variation observed with influenza A [21].

In addition to the viral quasispecies within an infected individual, there is genetic variation between the subtypes in the major group of circulating HIV-1 variants (Group M). There are thirteen distinct subtypes or clades in addition to dozens of circulating recombinant forms (CRFs), reviewed in [22]. Figure 2 below shows the geographical distribution of the HIV-1 subtypes along with the circulating recombinant forms (CRFs). Within a subtype, amino acid variation in the HIV-1 Env protein ranges between 15-20%, and between subtypes variation at the amino acid level can range up to 42% [2,22,23]. Given this diversity reflected by presence of multiple HIV-1 subtypes, circulating recombinant forms and constant viral evolution within infected individuals, it is a daunting challenge to develop a global HIV-1 vaccine that can provide protection against a diverse collection of HIV-1 sequences in multiple demographic populations. Globally, an estimated 35.3 million people are living with HIV, as recorded by the 2013 UNAIDS update [24]. The number of new HIV infections every year globally still remains

high at 2.3 million [24]. A safe and effective vaccine continues to be the best hope to combat the global HIV pandemic.

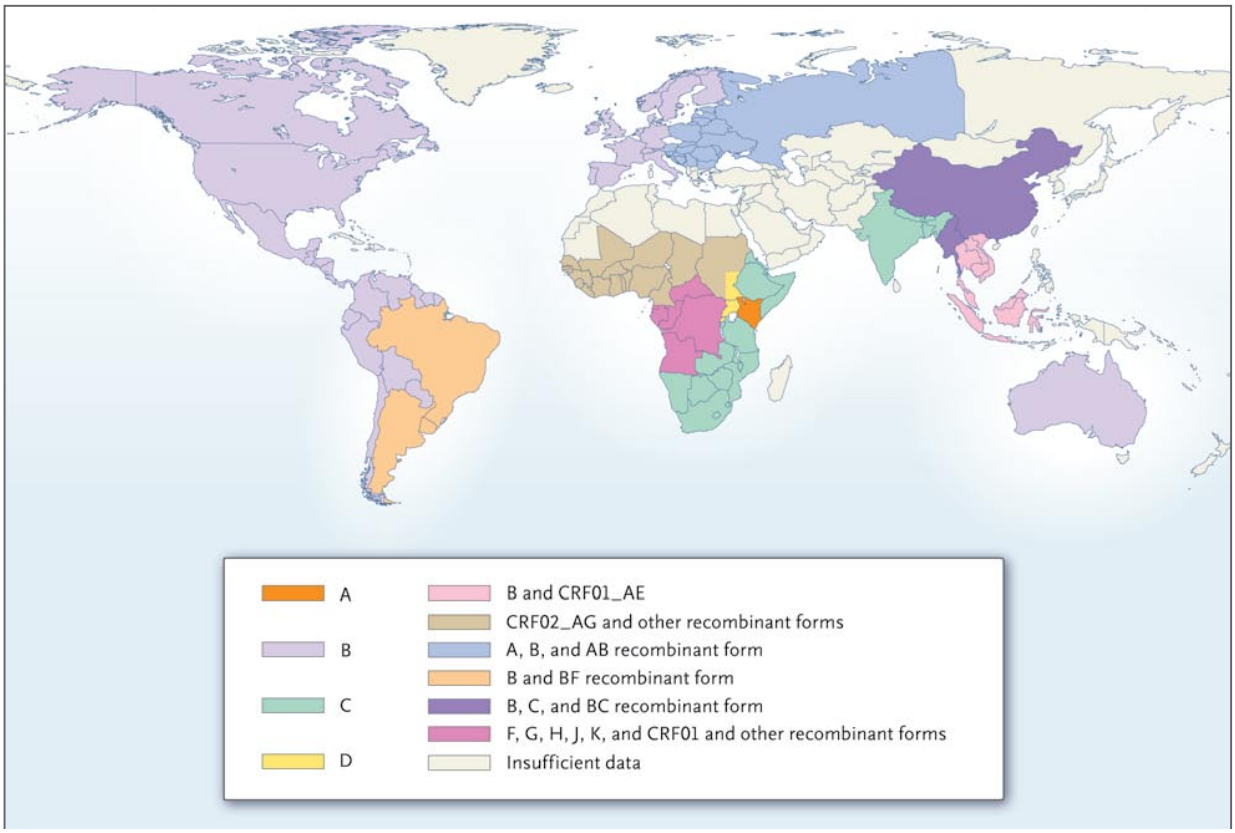

Figure 2. Global distribution of HIV-1 Subtypes and recombinant forms. This figure has been reproduced with permission from [22], Copyright Massachusetts Medical Society.

HIV-1 vaccine efficacy trials

There have been more than 200 vaccine products tested since 1987 [25] but of these only four vaccine strategies have been tested in six human efficacy trials [26], Table 1. These efficacy trials broadly follow two strategies for viral control: generate neutralizing Abs (nAbs) to prevent virus entry or generate cell mediated immune responses, or a combination of both. The rationale behind antibody (Ab) based vaccines is to target the HIV surface glycoproteins and neutralize HIV particles. The first two trials (Table 1) VAX004 [27] and VAX003 [28] evaluated a bivalent subunit of gp120 and these were trials were designed to elicit Abs to the viral Envelope protein.

These trials were not effective in generating nAbs against a diverse array of circulating HIV-1 strains and failed to prevent HIV-1 acquisition [29,30].

The Merck Ad5 (MRKAd5) STEP vaccine trial employed a different strategy by focusing on eliciting HIV-1 specific cytotoxic T lymphocytes (CTLs). There is evidence that within infected individuals HIV-1 specific CTLs are able to control viral replication [12]. Studies in individuals that are able to naturally contain HIV-1 infection and maintain a low plasma viral load have shown the importance of cellular immunity in controlling viremia. Certain human leukocyte antigen (HLA) alleles, in particular HLA-B*57 and HLA-B*27 have been correlated with the ability to control viral replication, which is consistent with the importance of T cells in HIV-1 infection control [31]. There is more evidence within nonhuman primates where it has been shown that CD8+ lymphocyte depletion correlated with an increase in viral load, and non-human primate vaccine induced CTLs have been shown to control viral replication and disease progression [32-36]. These lines of evidence led to testing a T-cell based vaccine with recombinant Gag, Pol and Nef proteins within the Adenovirus5 (Ad5) viral vector (STEP, MRKAd5). The STEP (HVTN502) and HVTN 503/Phambili trials were the first efficacy trials that evaluated an HIV vaccine designated to stimulate T-cell responses. These vaccine trials were halted before the trial completion as the first interim analysis showed the vaccine failed to prevent HIV-1 acquisition and did not reduce the viral load within infected trial subjects [37-39]. While the Ad5 vaccine in STEP induced CD8+ T-cell responses in the majority of vaccinees, the responses were weak and only directed against a small number of epitopes [38].

The RV144 vaccine comprised of a canarypox viral vector prime expressing Env, Gag and Pol (ALVAC) followed by an AIDSVAX B/E boost, which had the same protein subunit vaccine used in VAX003 [40]. The vaccine efficacy against HIV-1 infection acquisition was

31.2%, but the vaccine did not have an observable effect on CD4+T-cell counts or viral load after infection. Furthermore, the vaccine efficacy was highest, approaching 60%, within the first six months following immunization [41]. In an analysis for immune correlates, binding of immunoglobulin G (IgG) to V1/V2 (the first two variable loops regions within gp120 subunit of Env), was associated with a reduced risk of infection, and plasma levels of IgA against Env abolished the IgG protective effect [42]. Despite the inclusion of the gp120 subunit, this vaccine did not elicit broad neutralizing Abs and failed to generate measurable CD8+ T-cell responses. The most recently conducted HIV vaccine efficacy trial, HVTN505, evaluated a vaccine containing DNA prime expressing HIV-1 subtype B Gag, Pol and Nef, in addition to Env proteins from three HIV-1 subtypes A, B, and C. The rAd5 vaccine boost consisted of four rAd5 vectors expressing HIV-1 subtype B Gag-Pol fusion protein and Env glycoprotein from subtypes A, B and C. The trial was halted early in 2013 before completion when interim analysis showed no reduction in HIV-1 acquisition or reduction in viral load within infected study subjects [43].

HIV vaccine efficacy trial	Trial dates	Vaccine strategy	Main results from vaccine trial
VAX 004, VAX 003	1998-2004	Recombinant HIV Envelope protein	No vaccine efficacy, no generation of neutralizing Abs
STEP study, HVTN 502, Merck Ad5 ***(Primarily North America)*** **HVTN503/Phambili** ***(South Africa)***	2004-2009	rAd5 vector, Clade B *gag*/*pol*/*nef*	No effect on HIV-1 acquisition or reduction in viral load
RV144	2003-2010	ALVAC Canarypox vector prime (*gag*, *pol*, *env*) + recombinant gp120 protein boost (AIDSVAX B/E)	31.2% overall vaccine efficacy. No effect in post infection viral load or CD4+ counts
HVTN505	2005-2013	Prime: DNA Vaccine expressing Gag, Pol, Nef and Env + Boost: rAd5 vector expressing Gag, Pol and Env	Trial was halted in April 2013 as interim analysis showed no vaccine efficacy. The vaccine had no effect on HIV-1 acquisition or reduction in viral load

Table 1. HIV-1 vaccine efficacy trials in humans. The table summarizes the important results from major HIV-1 vaccine human efficacy trials.

Immune responses elicited with vaccination

As we see from the efficacy trials described above, HIV-1 vaccines are designed to elicit either the humoral immune response or the cellular immune responses or ideally a combination of both (Figure 3). The enormous diversity observed in Env amino acid sequences within and between infected individuals makes the design and development of nAbs a daunting task. Several lines of evidence have recently demonstrated the importance of antibodies in controlling HIV-1 and SIV infections. In an analysis for immune correlates, binding of immunoglobulin G (IgG) to V1/V2 (the first two variable loops regions within gp120 subunit of Env), was associated with a reduced risk of infection, and plasma levels of IgA against Env abolished the IgG protective effect [42]. The RV144 vaccine was shown to have increased efficacy against viruses that matched the Env immunogen in the V2 region [44]. Studies in nonhuman primates have also shown that Env-specific Abs are associated with decreased SIV infection risk when challenged with SIVmac251 and SIVsmE660 viruses [45-47]. Similarly in other non-human primate studies, neutralizing Env-specific Abs have been shown to have protect against HIV and Simian Human Immunodeficiency Virus (SHIV) challenges [48-57].

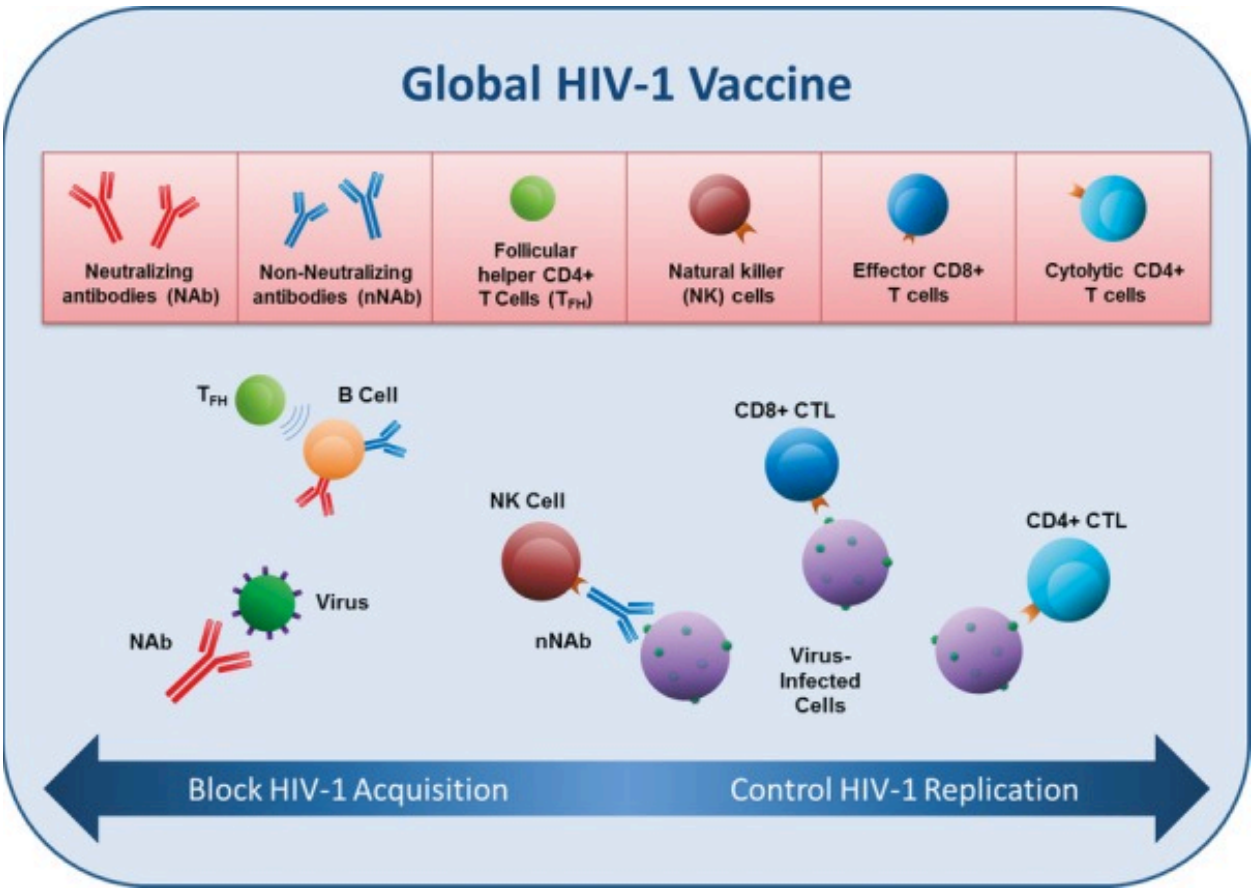

Figure 3. Immune responses targeted by a global HIV-1 vaccine. There are two broad approaches to HIV-1 vaccine design: blocking HIV-1 acquisition through generation of neutralizing Abs (NAbs) through vaccination and controlling HIV infection through non-neutralizing Abs (nNAbs) and vaccine induced HIV-1 specific T-cells (CD8+ and CD4+ cells). Figure reproduced from [58] under the terms of Creative Commons Attribution Non-Commercial No Derivatives License.

Lessons learned from immune correlates in vaccine efficacy trials highlight the importance of eliciting cellular immune responses capable of controlling viral replication (Figure 3) in addition to activating humoral immune responses to block HIV-1 acquisition [59]. A large body of literature has shown that cellular immune responses can control viral replication in HIV-1 infected humans and SIV infected rhesus monkeys through CD8+ T lymphocytes [36,60-68], through Natural Killer (NK) cells [69] and also viral control through CD4+ T lymphocytes [70-72]. Gag specific cellular immune responses in particular seem to be important for virologic

control. In non-human primates it was recently demonstrated that Gag specific CD8+ T cells are correlated with both *in vivo* and *in vitro* virologic control following SIV challenge in vaccinated monkeys [73]. Gag specific cellular immune responses have also been shown to be associated with virologic control in HIV-1 infected individuals [74-79]. Another important factor to consider is the location and phenotype of cellular immune responses elicited by vaccination. In a recent finding, Fukazawa *et*. *al*. demonstrated that the degree of protection from a live attenuated SIV vaccine strongly correlated with the magnitude of SIV specific effector T cells in lymph nodes, and the maintenance of these protective T cells was associated with the replication of the vaccine virus in the follicular helper T cells [80]. Notwithstanding the vast number of studies on non-human primates depicting efficacy of T cell responses in controlling SIV infections, T-cell based vaccines, such as the one tested in the STEP trial, designed to control HIV-1 replication through HIV-1 specific CD8+ T lymphocytes, have proven to be ineffective in preventing HIV-1 acquisition and HIV-1 viral replication in efficacy trials conducted to date. However, despite the lack of efficacy with T cells based vaccines, there has been evidence of immune selection pressure on founder strains infecting vaccinated subjects in the STEP trial, demonstrating for the first time that vaccine-induced immune responses can have an imprint on infecting strains [81].

HIV-1 vaccine immunogen design strategies

In the HIV-1 efficacy trials conducted so far, four vaccine design strategies have been evaluated, and the results from these trials have provided significant information for future vaccine development. The RV144 trial demonstrated that a safe and effective HIV-1 vaccine is possible [40,41] but many questions still remain about the number and types immunogens to be

included as part of the vaccine. HIV-1 immunogen selection and design strategies for use in vaccines can be divided into four broad categories.

1) **Subtype-specific antigens**: These vaccines include antigens to match local circulating HIV-1 strains. The idea behind these is that subtype-specific immune responses will have a higher likelihood of recognizing local strains. Such a strategy was adopted in the RV144 trial that used antigens matching the local Thai circulating subtype B strains and the circulating recombinant form CRF01_AE [40,41]. Since certain subtype-specific vaccines may be geographically limited in efficacy, these vaccines would have to be redesigned to be effective in other regions of the world.

2) **Vaccine antigens that elicit broadly neutralizing antibodies**: One of the important objective when designing a HIV-1 vaccine immunogen has been to elicit Abs that can neutralize a wide variety of HIV-1 strains and subtypes [20,26,82]. Many of the critical epitopes that are targeted by nAbs are located within the CD4+ binding site and even within sites in the V1/V2 variable loops in gp120 (HIV-1 subtype C nAbs). These epitopes are conformational and depend on the three dimensional structure of the Envelope (Env) trimer [83-87]. One of the challenges in developing an Env based HIV-1 immunogen, is the stability of recombinant Env trimers when synthesized on a large scale. Some of the techniques used to stabilize these trimers have centered around modifying the antigenicity, biochemistry and biophysics of the recombinant trimers [88]. Nevertheless, immunogenicity testing shows that recombinant trimers are only marginally better than monomers [89]. Recent studies have shown that a stable recombinant Envelope trimer can be generated with antigenic properties and glycosylation patterns closely resembling the native Envelope glycoprotein, and these recombinants have been shown to elicit improved neutralizing

Ab responses compared to monomeric gp120 in guinea pigs [90]. This trimer design strategy is proposed for phase I clinical trials in the next few years [90].

Detection of nAbs to HIV-1 in recent studies [91-93] has important implications in vaccine design. These nAbs have on average a viral neutralization breadth of around 80% and they recognize one of four sites on the viral envelope spike: a) the CD4 binding site on gp120, b) the V1/V2 loops on gp120, c) glycoproteins on the V3 loop of gp120 or d) the membrane-proximal external region of gp41 [91-93]. Notwithstanding these important discoveries, one has to keep in mind that broadly nAbs only arise in 10-30% of infected individuals after a period of 2-4 years [86,94,95]. Based on characterizing the IgGs from sera of HIV-1 individuals, the anti-HIV-1 antibodies are highly somatically mutated when compared to other IgGs cloned from the same subjects [96]. Potent broadly nAbs from HIV-1 infected individuals have undergone 2-10 times more somatic mutations than most human Abs [97]. Additionally, the accumulation of extensive somatic mutations over time within infected individuals necessitate the need to alter traditional vaccine strategies to effectively stimulate broadly nAbs via immunization.

3) **Vaccine antigens that elicit highly conserved HIV-1 specific cellular immune responses**: To address the challenge of HIV-1 diversity, one approach has been to use conserved HIV-1 regions as immunogens and elicit cellular immune responses to these conserved regions. These are usually sections within the HIV-1 genome that are highly conserved across multiple strains and subtypes, and hence the rationale is that immune responses targeted towards these regions will recognize multiple HIV-1 subtypes, and most likely, these immune responses will impose a high fitness cost for escape mutations to arise within the viral population [98,99]. One approach that has been tried is to select natural HIV-1 sequence antigens that are most conserved among

circulating HIV-1 strains. This was tested with the Merck Ad5 STEP vaccine that used an Ad5 vector expressing natural sequence antigens that were most likely to be conserved amongst circulating HIV-1 strains. Gag, Pol and Nef vaccine sequences for the STEP vaccine were selected based on phylogenetic similarity to a consensus of all HIV-1 subtype B sequences [37,100]. While there is evidence that the STEP vaccine exerted pressure on founder HIV-1 variants that established productive infection within infected trial subjects, immune responses against CTL epitopes was limited and not sufficient to prevent infection or lower viral load set-point [81].

One approach to address HIV-1 diversity in vaccine design is to omit variable segments of the HIV-1 genome and focus on only the conserved elements of the HIV-1 genome [101-104]. The HIV_{CONSV} immunogen [102] includes long fragments of conserved regions within the HIV-1 genome and when expressed by DNA and viral vector vaccines, this strategy has been shown to be immunogenic in Phase I trials [105,106]. A variation to this approach has been to design immunogens that are strictly based on conserved HIV-1 segments with mutable regions excluded completely [103]. Immune responses generated with DNA vectors expressing these highly conserved elements (CE) were compared with immunization with p^{55gag} DNA. Immunization with p^{55gag} DNA induced poor CD^{4+} mediated cellular responses whereas responses to the CE vectors were reactive across subtypes and comprised of both CD^{4+} and CD^{8+} T cells [107]. Further, DNA vaccination in macaques with the conserved elements vaccine shows increased T-cell breadth of response [108].

4) **Vaccine antigens to elicit highly diverse HIV-1 specific immune responses:** A contrasting strategy to eliciting immune responses to highly conserved HIV-1 regions is to generate diverse

immune responses against a broad array of HIV-1 sequences. The rationale for this immunogen strategy is that by generating a diverse array of immune responses against multiple regions of the HIV-1 genome (immunogen breadth) and simultaneously generating immune responses against multiple variants within epitopes (immunogen depth), we can maximize the likelihood that there will be a match to the infecting founder strain within the infected individual. The immunogens are so designed to elicit diverse immune responses including HIV-1 specific CTLs and humoral immune responses. There are two primary approaches to elicit broad HIV-1 immune responses via vaccination. One approach is to design a multivalent vaccine immunogen that represents antigens from different HIV-1 clades. A recent example of such an approach is the phase II clinical trial sponsored by the HIV Vaccine Trials Network (HVTN), HVTN 505, which tested the DNA prime/Ad5 boost vaccine expressing Env proteins from subtypes A, B and C. Unfortunately this trial was halted when preliminary analysis showed no vaccine efficacy and no reduction in HIV-1 acquisition rates [41]. The second approach to eliciting broad immune responses is the design "mosaic" immunogens [109,110]. These mosaic immunogens have been designed based on an *in silico* analysis of global HIV-1 sequences to provide maximum coverage of viral sequence diversity information [110]. Multiple studies have reported that mosaic immunogens have elicited greater breadth and depth of HIV-1 specific immune response compared to consensus or natural HIV-1 sequences in non-human primates including improved Env specific binding and neutralizing Ab responses [111,112]. A recent study in rhesus monkeys has shown that mosaic HIV-1 immunogens have elicited immune responses of greater magnitude compared to a vaccine that was constructed by excising and concatenating six highly conserved regions (including Gag, Pol and Env) from the full length mosaic antigens [113].

The overarching goal of all HIV-1 vaccine design strategies is to overcome the challenges posed by HIV-1 sequence diversity and the only way to determine which of these strategies used alone or in combination will result in protection against HIV-1 infection in humans is through large scale clinical efficacy trials.

Sequence analyses in HIV-1 vaccine efficacy trials

Sieve analysis

Understanding the characteristics of the viruses that evade a vaccine-induced host immune response to establish a productive HIV-1 infection (breakthrough virus) can provide 1) valuable insights into vaccine efficacy, 2) evidence of genetic signatures of vaccine-induced host immune responses and 3) useful information for subsequent immunogen design [114]. In the course of a HIV-1 efficacy trial, due to randomization at study entry and blinding of both investigators and study subjects to treatment arms, it is expected that vaccine and placebo recipients would be exposed to similar circulating HIV-1 strains during the duration of the efficacy trial. It can be then be assumed that if the trial was conducted with no bias in treatment assignment, that the differences between the HIV-1 sequences characterized from the two treatment groups can be attributed to vaccination [115]. Genetic imprints of vaccine-induced immune pressure on infecting HIV-1 strains in vaccine recipients can be compared against the genetic makeup of HIV-1 sequences infecting a control placebo population. Associating specific viral variants with vaccine efficacy will provide important insights to effects of vaccine-induced immune pressure on viral selection and evolution [115,116]. This approach is termed as ‘sieve analysis’ [115,116].

Sieve analysis can be used for both generating hypotheses regarding which immune characteristics are important in terms of protection offered by the vaccine, and also for confirmatory analysis in which HIV-1 genetic comparisons between vaccine and placebo groups can validate whether a particular immunological response is a correlate of protection against HIV-1 infection [114]. Sieve analysis differs from comparing correlates of risk analysis since the former includes both vaccine and placebo groups whereas in the latter analysis the vaccine subjects who subsequently became HIV-1 infected are compared with other vaccine subjects who remained uninfected during the trial period. Hence, when immune correlates are identified in a risk analysis, they do not necessarily predict vaccine efficacy [114,115]. In contrast, the advantage of sieve analysis is that any differences observed in breakthrough HIV-1 sequences between vaccine and placebo groups can be described by the vaccine status since we assume that the vaccine treatment was assigned randomly at the start of the trial [114-116].

Acquisition and post-infection sieve effects

Acquisition sieve effect describes the scenario where vaccine-induced immune responses block infection of certain HIV-1 variants that are most genetically similar to the vaccine insert (Figure 4) [115,117]. In this case, when breakthrough sequences are compared to the vaccine insert sequences, one would expect to find the sequences from the vaccine recipient more divergent from the vaccine than the placebo breakthrough sequences.

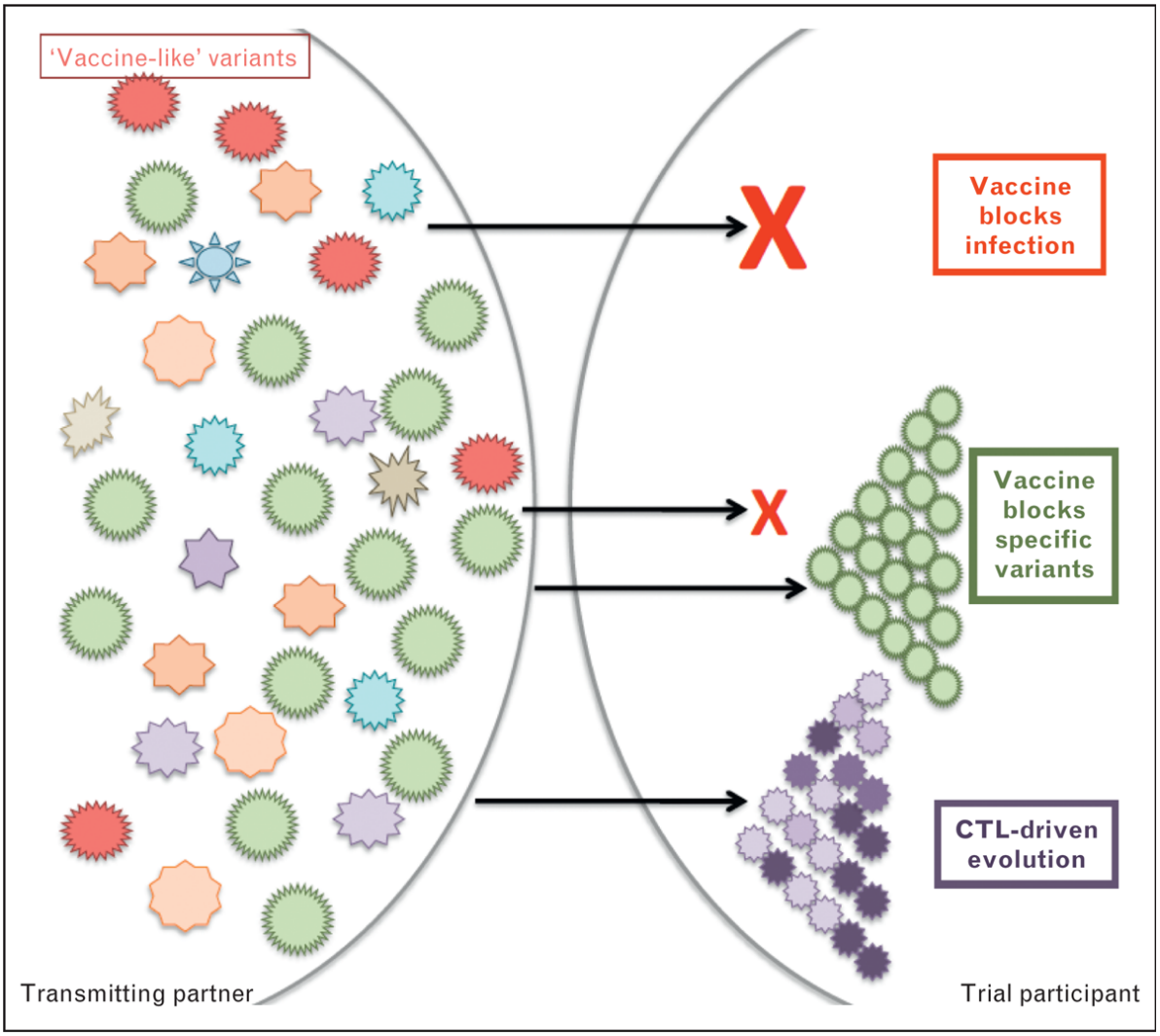

Figure 4. Schematic description of sieve acquisition and post-infection effects. There are three possible scenarios in a vaccine trial. This figure is reproduced from [114] with permission (License number: 3393210853176).

There are three possible scenarios of vaccine-induced immune responses influencing the population of breakthrough HIV-1 variants. If the vaccine is 100% effective in blocking productive HIV-1 infection (Figure 4, red box), there are no breakthrough sequences in the vaccinated group. In the second scenario, vaccine-induced immune responses can exclude certain HIV-1 variants from establishing infection. This is defined as an acquisition sieve effect (Figure 4, green box). The third scenario of post-infection sieve effect implies that HIV-1 breakthrough variants are driven by vaccine-induced immune pressure to accumulate more mutations or

accumulate them at a faster rate, when compared to rate of mutations driven by non-vaccine induced immune pressure [114].

Sieve analysis of breakthrough sequences from the STEP/HVTN502 trial

The STEP/HVTN502 trial, which tested an Ad5 vector with *gag/pol/nef* HIV-1 subtype B vaccine insert, was not effective in preventing HIV-1 acquisition and did not reduce viral loads in infected trial participants [37,38]. The hypothesis of whether the vaccine-induced immune responses impacted breakthrough HIV-1 sequences was tested by sieve analysis [81]. Divergence from the vaccine insert was estimated in peptide regions from Gag, Nef and Pol that were predicted, *in silico*, to be CD8+ T-cell epitopes (as these peptides would most likely bear genetic signatures of vaccine-induced immune pressure) and these distances were compared across vaccine and placebo treatment groups. Sieve analysis of 465 genomes from 65 trial subjects showed that the epitopic distances between breakthrough sequences and vaccine insert sequences were higher in subjects in the vaccine group than subjects in the placebo group [81]. This was the first time evidence of selective pressure from vaccine-induced immune responses elicited by a T-cell based HIV-1 vaccine was described.

These observed results could be explained by acquisition sieve effect where certain variants are excluded from establishing productive infection within subjects in the vaccine group. Additionally, the observed results could also be the result of post-infection sieve effects where vaccine-induced immune pressure is the driving force behind breakthrough variants accumulating more mutations, ultimately resulting in immune escape and leading to selective outgrowth of escaped variants. These scenarios are not mutually exclusive and it is important to explore these observed effects in greater detail. Doing so will help us identify key genetic signatures that are linked to vaccine-induced immune responses generated within an infected

individual. Additionally, quantifying the effect of these immune responses on HIV-1 breakthrough sequences will help in the future design of T-cell based vaccines.

In this dissertation, I will present results from a study of the breakthrough HIV-1 sequences from the HVTN502/STEP vaccine trial. As part of my dissertation I will address the following questions: 1) Can we find evidence of blocking of infection by variants similar to the vaccine insert (acquisition sieve effect)? 2) Can we find evidence for vaccine-induced immune pressure driving CTL epitope evolution? and 3) Do breakthrough sequences from study subjects followed over time show evidence for vaccine-induced anamnestic responses?

I will begin my dissertation by describing a new algorithm (CorQ, Correction by Quality), that I developed to analyze sequences from high throughput sequencing technologies used to sequence virus in subjects from the STEP vaccine trial (Chapter 2). In the chapter describing the error correction algorithm, I will also present results comparing the performance of the program CorQ to other high throughput sequence analysis programs. Following this, in Chapter 3, I will present a comprehensive comparison of Single Nucleotide Polymorphisms (SNPs) within a subset of the vaccine trial subjects that were included in both the initial study of breakthrough sequences [81] (subjects sequenced with traditional Sanger sequencing technology) and the current study (subjects sequenced high throughput sequencing technology). The purpose of this comparison is to estimate the HIV-1 variant concordance between these two vastly different sequencing technologies and highlight the differences between the two methods. Following this, I will then present evidence for vaccine-induced anamnestic pressure influencing CTL epitope diversity and evolution (Chapter 4).

CHAPTER 2. Quality score based identification and correction of pyrosequencing errors

The text in this chapter has been modified slightly from *PlosOne*, 2013, doi:

10.1371/journal.pone.0073015

Summary

Massively parallel DNA sequencing using the Roche 454/pyrosequencing platform allows in-depth probing of diverse sequence populations such as within an HIV-1 infected individual. Nonetheless, analysis of this sequence data remains challenging due to the shorter read lengths relative to that obtained by Sanger sequencing as well as errors introduced during DNA template amplification and during pyrosequencing. The ability to distinguish real variation from pyrosequencing errors with high sensitivity and specificity is crucial to interpreting sequence data. In this chapter I will present a new algorithm, CorQ (Correction through Quality), which utilizes the inherent base quality in a sequence-specific context to correct for homopolymer and non-homopolymer insertion and deletion (indel) errors. CorQ also takes uneven read mapping into account for correcting pyrosequencing miscall errors and it identifies and corrects carry forward errors. We tested the ability of CorQ to correctly call SNPs on a set of pyrosequences derived from ten viral genomes from an HIV-1 infected individual, as well as on six simulated pyrosequencing datasets generated using non-zero error rates to emulate errors introduced by PCR. We tested sensitivity and specificity of CorQ in combination with other error correction algorithms and we attained a 97% reduction in indel errors, a 98% reduction in carry forward errors, >97% specificity of SNP detection and >98% sensitivity of SNP detection. This combined procedure will permit examination of complex genetic populations with improved accuracy.

Introduction

DNA sequencing has dramatically altered the nature of biomedical research and medicine. Over the past decades there has been a dramatic reduction in cost, complexity and time required to sequence large amounts of DNA. Until the last few years, the majority of DNA sequences overwhelmingly relied on some version of the Sanger sequencing chemistry [118]. Over the past decade there has been a massive push to develop new strategies for DNA sequencing. Massively parallel sequencing (MPS) or next generation ("next-gen") sequencing technologies [119] allow for the generation of millions of sequence fragments ("sequence reads") from a single specimen. These technologies have already begun to replace Sanger sequencing for many applications, including de novo sequencing, re-sequencing and metagenomics [120,121].

Pyrosequencing chemistry

One of the earlier high throughput sequencing technologies, 454/Roche pyrosequencing, has been used extensively to study the inherent diversity of viral pathogens such as HIV-1 and HCV [122-135]. The pyrosequencing chemistry is shown in Figures 5 and 6 below. Template DNA to be sequenced is first denatured and strands are annealed to the beads conjugated with oligonucleotides complementary to the linker sequences (Figure 5). This step is carried out with very low DNA concentrations so that on average only one strand binds to each bead. Bead-bound DNA is then PCR amplified in an oil water emulsion, where each water droplet in the emulsion contains on average, a single bead [119]. Amplified DNA strands anneal to the beads resulting in beads with many copies of a homogenous PCR product. The beads are mixed with DNA bead incubation mix containing DNA polymerases and layered with enzyme beads on a Picotiter plate which contains approximately 1.6 million wells [136].

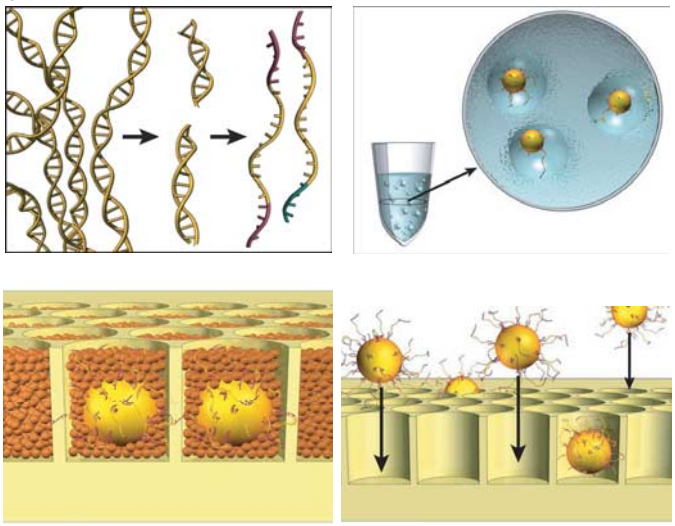

Figure 5. Pyrosequencing sample preparation. Genomic DNA is isolated, fragmented, ligated to adapters and separated into single strands. The fragments are then bound to beads and captured in droplets containing PCR-reaction mix and emulsion PCR is carried out within the droplets. The figure is adapted from [119] with permission (License number 3397230961032).

The picotiter plate is placed into the Genome Sequencer FLX Instrument (Figure 6). The fluidics system delivers the four DNA nucleotides sequentially in a fixed order across the plate. When a nucleotide complementary to the template strand is flowed across a well, the DNA polymerase extends the DNA strand by adding the nucleotide(s). Addition of one or more nucleotides generates a light signal recorded by the CCD camera in the instrument. Raw signals are background-subtracted, normalized and corrected. The normalized signal strength at each nucleotide flow within a well corresponds to the number of nucleotides that were incorporated. This correspondence is linear only up to eight consecutive nucleotides, following which signal strength deteriorates rapidly [119]. Signal strengths are stored in standard flowgram format (SFF) files for downstream analysis. A schematic of a flowgram is shown in Figure 7.

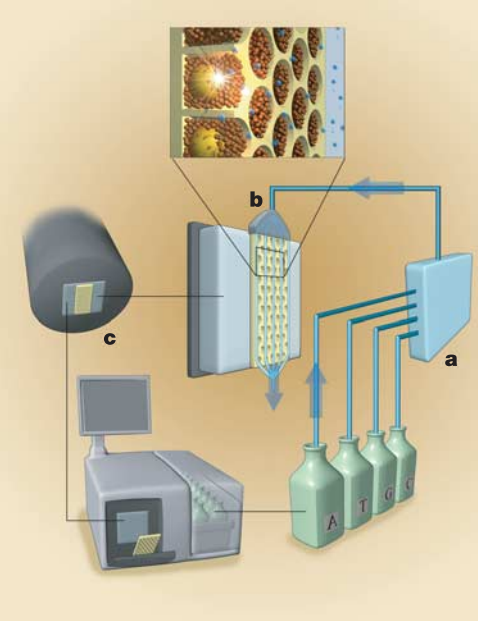

Figure 6. Pyrosequencing instrument. The sequencing instrument consists of a fluidic assembly (a), a flow chamber that includes the picotiter slide (b), a CCD camera based imaging assembly (c) and a computer that provides the user interface and instrument control. This figure has been adapted from [119] with permission (License number 3397230961032).

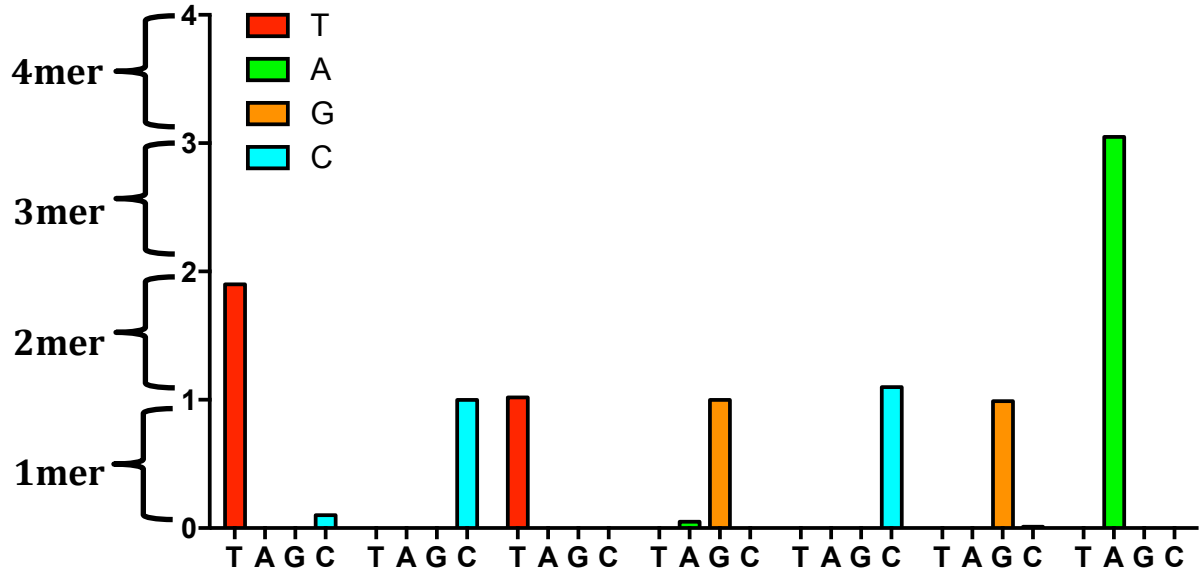

Figure 7. Graphical representation of flowgram intensities. The four colored bars represent the nucleotides during each flow (in the order TAGC). Seven consecutive flow cycles are shown in the x-axis. Signal intensities during nucleotide flow are shown in the y-axis. For certain signals, the intensities recorded by the CCD camera are such that it is difficult to determine the exact number of bases at that position. This inherent property of pyrosequencing leads to homopolymer region associated errors (over and undercalls).

The promise of MPS has to be balanced with its caveats. Each MPS platform has a much higher rate of error compared to Sanger sequencing [119,137,138]. If the sample must be PCR-amplified prior to sequencing, the errors occurring during PCR are also present in the MP sequences and can be impossible to distinguish from real variation, except in cases when using random sequence tags coupled with oversequencing to generate consensus sequences from each amplicon [139,140]. The 454/pyrosequencing platform results in uniquely high rates of overcalls and undercalls (resulting in erroneous insertions and deletions in the sequence reads) [119,138]. Carry forward errors are also unique to pyrosequencing and are caused by leftover nucleotides in a sequencing well and incomplete extension during a prior flow cycle [119].

Error rates for the GS-FLX Titanium pyrosequencing technology have been estimated on an extensive dataset of Roche Corp. Quality control DNA fragments and the sequences generated were found to have a mean error rate of 1.07%. This study found that errors were non-randomly distributed with some positions in homopolymer regions with error rates as high as 50% [138]. Additionally, 89% of the reads had some form of error with insertion errors comprising of majority, followed by deletions, mismatches and ambiguous base calls. This result suggests that instead of removing erroneous reads from downstream analysis, error correction algorithms have to be applied to the entire dataset to improve overall accuracy of the pyrosequences.

Pyrosequencing errors can be corrected at two stages: 1) at the level of light intensities (flowgrams) [141-143] and 2) at the level of correcting the machine called sequences [144-151]. Error correction algorithms in the first category include the program such as AmpliconNoise, which correct errors generated by clustering flowgrams and calculating the likelihood that each of the reads from these flowgrams was generated from a mixture of correct and incorrect sequences [142,143]. Subsequently an expectation-maximization algorithm is applied to the

clusters to determine a true sequence from each cluster. Methods from the second error correction category using a Poisson or binomial probability have traditionally assumed, incorrectly [119,138,141], that all base qualities are equal [123,144,146,149]. The latter category also includes error correction methods that rely on comparing variants to an empirical control dataset, mapping read segments to a consensus template and refining alignments locally [145]. Other approaches include taking sequences sharing common k-mers and forming multiple alignments with these reads and correcting the reads bases on a consensus sequence generated from these alignments [148]. Co-variation or phase information has also been used to distinguish between real variation and systematic error [150]. These algorithms mentioned above do not alter the flowgram intensities instead they correct pyrosequencing errors on the translated bases.

Materials and Methods

HIV-1 pyrosequences for comparing performance of error correction algorithms

Ten HIV-1 genome (from one infected individual) sequences were PCR amplified, cloned and sequenced using the Sanger sequencing method [152]. These sequences have been deposited in GenBank with accession numbers: JN024165-JN024168, JN024170-JN024173, JN024495 and JN024537. The plasmid clones were mixed in equal proportion, linearized with a restriction enzyme and used for pyrosequencing using standard protocols provided in the GS-FLX Titanium Rapid Library preparation kit (454.com/products/gs-flx-system/index.asp).

Generation of simulated pyrosequences

We generated a total of six additional simulated pyrosequencing datasets using the software program Flowsim [153]. We used two starting configurations (Table 2). The first three datasets (Set 1a-c, Table 2) were generated using a single 1500 nucleotide HIV-1 sequence as the

starting template. Three simulation runs were conducted: The first had no additional mismatch errors. The second and third had added mismatch error rates (with an equal mix of transitions and transversions) of 0.005 and 0.01 respectively, set to approximate error rates generated during template PCR amplification, and these values were selected based on previously determined DNA polymerase error rates [154]. The templates for the fourth through sixth simulated pyrosequencing datasets (Set 2a-c, Table 2) were generated from a multiple sequence alignment of 28 previously published HIV-1 sequences [152]. A 1500nt region was selected (alignment positions: 1-1522 within *gag* region) and used as input for Flowsim. Again, three simulation runs were conducted: with no additional SNP errors, and with SNP error rates of 0.005 or 0.01.

Simulated pyrosequencing set	Reads generated	Average read length (nt)
Set 1a (No additional SNP errors)	36,000	494
Set 1b (SNP error rate: 0.005)	33,000	427
Set 1c (SNP error rate: 0.01)	48,000	428
Set 2a (No additional SNP errors)	98,800	423
Set 2b (SNP error rate: 0.005)	98,880	422
Set 2c (SNP error rate: 0.01)	98,000	423

Table 2. Average number of reads and average read length for simulated pyrosequences. Two sets of simulated pyrosequences generated using Flowsim [153] is shown here**.** The first set (Set 1a, b and c) is comprised of simulated reads generated using a single 1500 nt HIV-1 sequence as the starting template. The second set (Set 2a, b and c) is comprised of simulated reads generated using a 1500nt region located within 28 HIV-1 sequences as starting templates. Simulations were done without additional SNP errors (1a, 2a) and with two different SNP error rates, 0.005 and 0.01 (1b,c and 2b,c). This table is reproduced from [155] with permission under the terms of Creative Commons Attribution License.

Error correction with AmpliconNoise

AmpliconNoise [142,143] (version 1.24) was run on flowgrams using default settings. Error correction with the AmpliconNoise suite of programs consists of two components, clustering and correcting the flowgrams with AmpliconNoise, followed by correcting PCR based errors with SeqNoise. In our preliminary evaluation we found that SeqNoise was computationally intensive, often failed on datasets larger than 20,000 reads and lacked important user definable parameters. Hence, we did not use the SeqNoise component for our subsequent analyses. The sequence and associated quality files obtained after AmpliconNoise flowgram correction were aligned with MOSAIK [156] using a sample-specific consensus sequence as reference. We adjusted the reference to query sequence mismatch parameter in MOSAIK to vary between 20 – 30%. These mismatches included both SNP and indels and allowed mapping a greater number of reads to the reference sequence, subsequently resulting in smaller loss of data.

Read filtering and chimera check

Sequences with ambiguous base calls (N) or less than 100 bases in length were removed, and we implemented an optional check to test for chimeric sequences. Chimeras are generated when sequences are amplified from a multi-template population [157] as well as naturally during HIV infection. The majority of *in-vitro*-generated chimeras arise due to incomplete primer extension during PCR [157]. To detect chimeras, we counted the number of SNP mismatches in a given read relative to the consensus sequence. In the CorQ analyses presented here we set this parameter to require between 20-40% SNP mismatches between the consensus sequence and a given read to assign a sequence as a chimera, since this mismatch rate was optimal for chimera detection amongst several methods [158]. For analyzing sequences with inherently greater diversity, we recommend varying this parameter to better distinguish a sequence variant from an artificially generated chimeric sequence.

CorQ implementation: Correcting poor quality indel and miscall errors

CorQ uses the filtered sequence alignment file to correct indel and miscall errors. First, quality values are mapped to the bases in a multiple sequence alignment, and positions with insertions and deletions in homopolymer and non-homopolymer regions are flagged (Equation 1). Two or more consecutive bases of the same kind are considered part of a homopolymer. In the flagged positions, the average base quality, Q_i, of indel bases is estimated by summing up the individual quality of all non-gap bases (A, G, T or C) at that position from all reads divided by the number of reads with non-gap bases. Similarly, the average base qualities of all non-gap bases in a non-indel position, Q_{i-1} and Q_{i+1}, adjacent to the flagged position are also estimated. For each indel occurring after a homopolymeric or non-homopolymeric sequence, the average base quality difference, $Q_{reduction,i}$, is calculated by first estimating the difference between average scores of flagged and adjacent positions: Q_{i-1}, Q_i and Q_{i+1}, Q_i followed by estimating an average of the score difference (Equation 1). This value is compared against a distribution of quality reductions across all flagged positions in the entire alignment. Flagged positions with a reduction in base quality higher than the distribution mean are flagged for correction. Both single and multi-base indels are handled in a similar manner. Additionally, indels present only in a single read are also flagged for correction. Sequences containing flagged insertions are corrected by removing the incorrectly inserted base and sequences with flagged deletions are corrected by adding the consensus base. CorQ also creates an annotation file that tracks changes made to each corrected read.

$$Q_{reduction,i} = 1/2\left[\left(\frac{\sum_{1}^{n_{i-1}} Q_{i-1}}{n_{i-1}} - \frac{\sum_{1}^{n_{indel,i}} Q_i}{n_{indel,i}}\right) + \left(\frac{\sum_{1}^{n_{i+1}} Q_{i+1}}{n_{i+1}} - \frac{\sum_{1}^{n_{indel,i}} Q_i}{n_{indel,i}}\right)\right] \quad (1)$$

Equation 1. Calculation of reduction in base quality in flagged indel positions. Q_{i-1}, Q_i and Q_{i+1} represent the quality of non-gap bases within the flagged position (i) and adjacent positions ($i-1$, $i+1$). $n_{indel,i,}$ n_{i-1} and n_{i+1} are the number of reads with non-gap bases at positions i, $i-1$ and $i+1$ respectively.

To identify and flag potential sequencing miscalls, the difference between the average base quality of consensus bases at the flagged position i, $Q_{i,consensus}$, and average base quality of SNP variant at position i, $Q_{i,SNP}$, are calculated for all positions in the alignment in which SNPs are observed relative to the consensus (Equation 2). The difference in average quality between the consensus and SNP bases is compared against a distribution of quality reductions for all flagged positions with identified SNPs and the SNP is flagged for correction if the quality difference is higher than the distribution mean. The consensus character at that position then replaces a flagged SNP. Positions with a SNP present only within a single read in the dataset are also flagged for correction.

$$Q_{reduction,i} = \frac{\sum_{1}^{n_{consensus,i}} Q_{i,consensus}}{n_{consensus,i}} - \frac{\sum_{1}^{n_{SNPi}} Q_{i,SNP}}{n_{SNPi}} \quad (2)$$

Equation 2. Calculation of reduction in base quality in flagged SNP positions. $Q_{i,consensus}$, and $Q_{i,SNP}$ represent the base quality of consensus and SNP bases respectively in the flagged position (i). $n_{consensus,i,}$ and $n_{SNP,i}$ are the number of reads with consensus and SNP bases at flagged position i, respectively.

To accommodate uneven read coverage (number of reads mapping to each base) from the two different sequencing orientations, we implemented additional checks when correcting potential sequencing miscalls. We have made read coverage difference as one of the input parameters in CorQ to allow users to set a coverage difference threshold that best captures the observed read coverage differences. SNPs that fall within regions of the designated coverage

difference are marked but not corrected, as we cannot rule out the possibility that a detected SNP is not "true" simply due to lack of adequate reads mapping to that position.

We also implemented a method within CorQ to identify and correct carry forward errors. Carry forward errors occur when insufficient flushing between the flows results in leftover nucleotides in a well, resulting in signal peaks at the wrong position during the next base incorporation [119]. The presence of homopolymers increases the likelihood of this type of error [119,137]. Carry forward events cause single base insertions usually near, but not adjacent to homopolymer regions [137]. CorQ detects this specific pattern of single base insertions occurring after runs of homopolymeric nucleotides and flags them as carry forward errors if the inserted base is not the consensus at that position, and if it is the same base type as the preceding homopolymeric stretch. The flagged inserted bases are removed from reads.

Comparison to other error correction methods

We tested the sensitivity and specificity of CorQ to identify true SNPs within a dataset created by pyrosequencing ten HIV-1 genomes that had previously been sequenced, after cloning into plasmids, by the Sanger method, as well as the set of six simulated datasets. CorQ was tested against four other pyrosequencing error correction programs: CORAL [148], Segminator II [145], QuRe [151] and V-Phaser [150] and the flowgram correction method AmpliconNoise [143] using reads mapping to the three HIV genes *gag*, *env* and *nef*. All programs were run according to the default parameters recommended by the authors. All the tested programs used Fasta and associated base quality files as their input. Reads less than 100 bases in length or reads with ambiguous bases (N) were removed prior to testing the error correction programs. We implemented CorQ on the following set of data files: a) uncorrected fasta and quality files, b) Flowgram corrected fasta and quality files (from AmpliconNoise) and c) files generated from the

quality recalibration program Pyrobayes [159]. Pyrobayes uses data likelihoods and prior distributions to determine the Bayesian posterior probability of the correct number of bases given a measured incorporation signal [159] and results in a recalibrated base quality for each called base. We used the consensus of the Sanger sequences from the 10 viral genomes as the reference for generating multiple sequence alignments in all the above comparisons. We also compared the performance of each error correction program on indel attrition. The exact count of insertions and deletions are not obtained from the output from QuRe and SegminatorII, hence these programs were not included in this comparison.

Results

Pyrosequencing of 10 HIV-1 genomes resulted in 26,620 reads mapping to *gag*, 48,927 reads mapping to *env* and 21,963 reads mapping to the *nef* genes (Figure 8). Read coverage for both sequencing orientations is shown in Figure 8 below. The uneven coverage map shown below is the result of uneven template shearing. While these coverage maps are more uneven than typical pyrosequencing runs performed by us, they highlight an important concern for algorithms calling SNPs in regions of poor read coverage and for determining the actual depth of population sampling across a genome – coverage and depth vary across the target sequences, and thus are poorly summarized by a single measure.

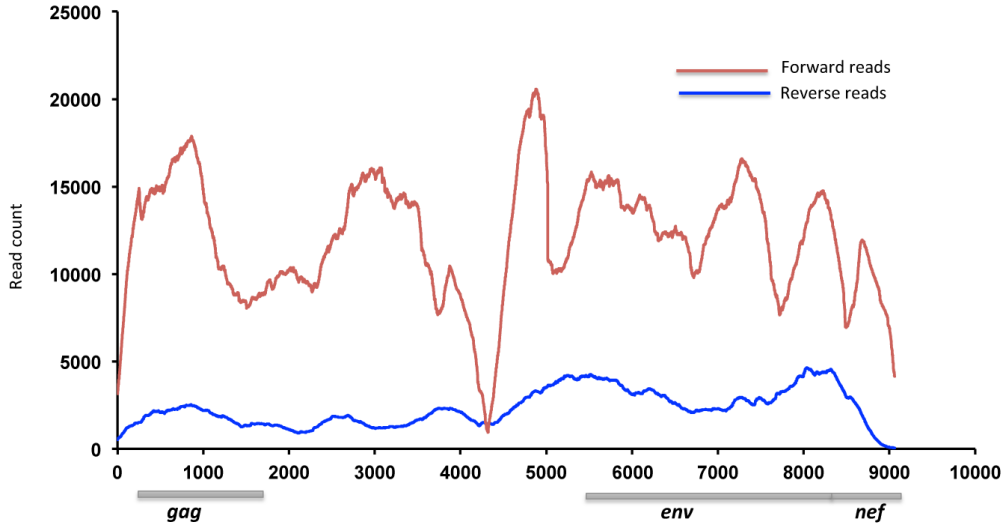

Figure 8. 454 read coverage across the HIV-1 genome. Locations of the *gag*, *env* and *nef* genes evaluated in this study are shown. A total of 26,620 reads mapped to *gag*, 48,927 to *env* and 21,963 to the *nef* gene. Reads were aligned to a sample-specific consensus using MOSAIK [156]. This figure has been adapted from [155] under the terms of Creative Commons Attribution License.

An overview of the CorQ algorithm is shown in Figure 9. Following AmpliconNoise, a reference-guided multiple sequence alignment is generated with MOSAIK. Reads less than 100 bases and reads with ambiguous bases are removed as part of the preprocessing step. Short reads are generally a result of premature stops in strand synthesis or out-of-phase strand synthesis. These out-of-phase strands show early deterioration in signal quality, leading to shorter read lengths [119,138]. Regions within the multiple sequence alignment with insertions and deletions are classified as occurring in homopolymer (a region with two or more consecutive nucleotides of the same type) or non-homopolymer regions.

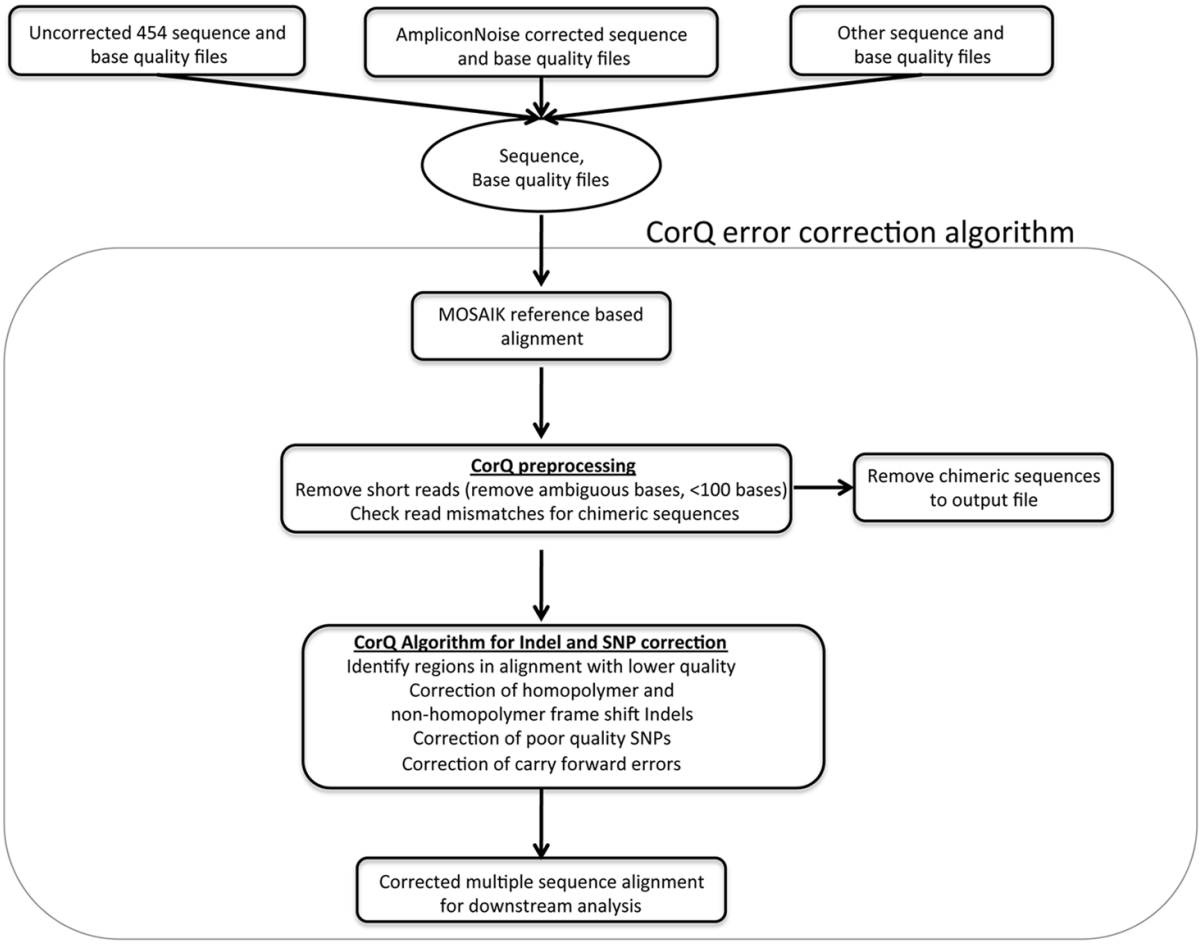

Figure 9. Overview of the CorQ 454 error correction methodology. The starting point for the CorQ algorithm is a set of sequence and base quality files. MOSAIK [156] is used for reference-based alignment. Positions with out-of-frame insertions and deletions (indels) are identified within the alignment and average base qualities are calculated for these regions (See Equation 1). SNPs are similarly identified and called (See Equation 2). This figure is adapted from [155] under the terms of Creative Commons Attribution License.

The average difference in base quality between an indel position and adjacent positions are then calculated. The rationale for this step is that in the event a base corresponds to a sequencing overcall or undercall, the quality of that base should be lower than the neighboring bases – CorQ measures this drop in base quality relative to the adjacent bases. A distribution of average base quality reductions across indel positions within the alignment is used to make error correction calls. We observed similar patterns of quality reductions across the three gene regions

(Figure 10). This bolsters our hypothesis that erroneous bases have poorer quality in the reads that contain them, and that the base quality adjacent to an erroneous base should be higher in the majority of reads. This allows CorQ to identify regions with a drop in average base quality across an alignment.

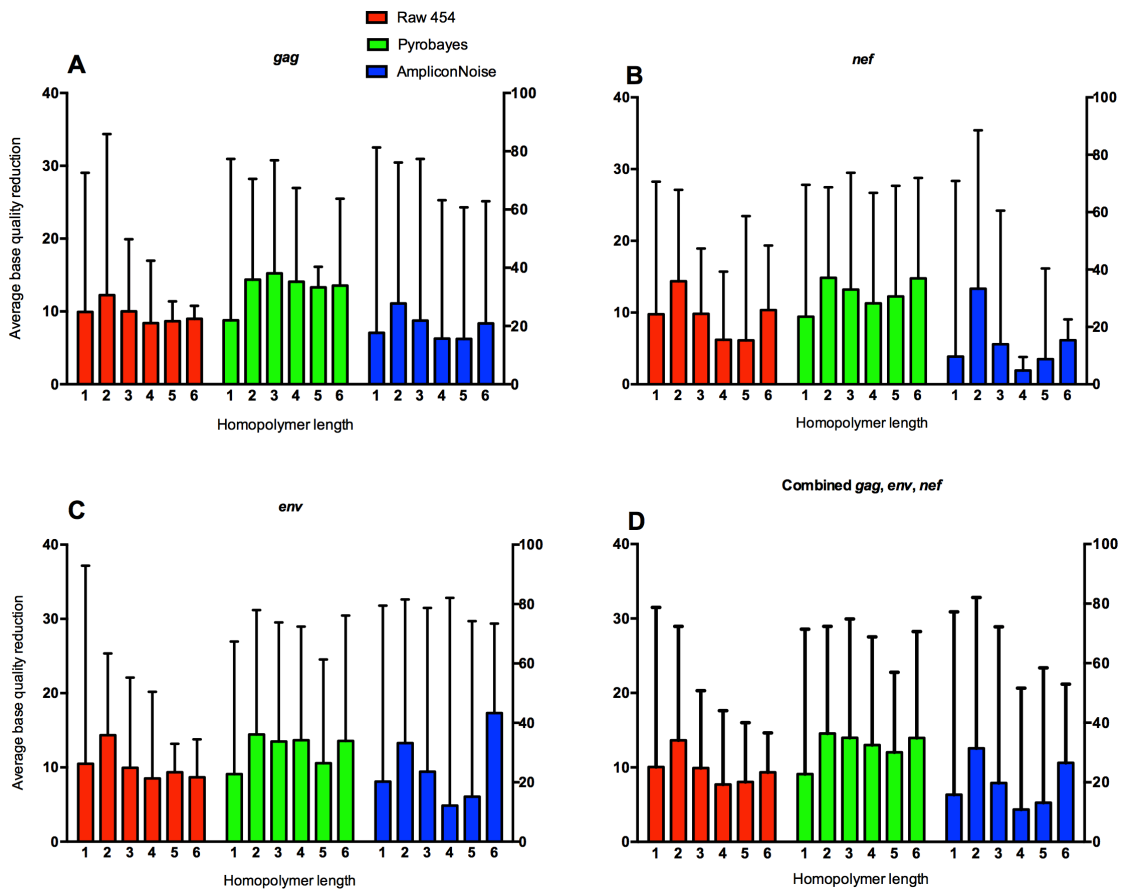

Figure 10. Average reduction in base quality for indels found in homopolymer and non-hompolymer regions. Reduction in base quality was measured as the average difference in quality between flagged positions with indels and the adjacent columns (See Equation 1). Base qualities from uncorrected sequences (raw 454), and sequences corrected with AmpliconNoise and Pyrobayes are shown for indels found in non-homopolymer regions (length of 1) and varying homopolymer lengths. Reduction in base quality is shown for indels within *gag* (A), *env* (B), *nef* (C) and the three genes combined (D). This figure is adapted from [155] under the terms of Creative Common Attribution License.

Next, we compared the ability of CorQ and previously described algorithms for their ability to flag and correct sequencing overcalls and undercalls, using true indels observed within the Sanger sequences as an indicator of effectiveness. QuRe and SegminatorII programs did not output indel counts per position and hence we omitted these programs from this comparison. Vphaser run alone or the combination of AmpliconNoise flowgram correction followed by the CorQ algorithm on Fasta and quality files reduced indel counts most effectively (95.4/96.7% reduction in *gag*, 95.3/94.7% in *nef* and 93/97% in *env*, respectively). CORAL, and Pyrobayes followed by CorQ did not result in a substantial reduction in erroneous indels (10 – 70%). Combinations of error correction methods performed better than applying a single correction method. The combination of AmpliconNoise + CorQ and AmpliconNoise + CORAL performed better than other tested methods, achieving between 95 - 97% reduction in indels. Among the individual correction methods, VPhaser performed best, reducing indels by 92 – 96%.

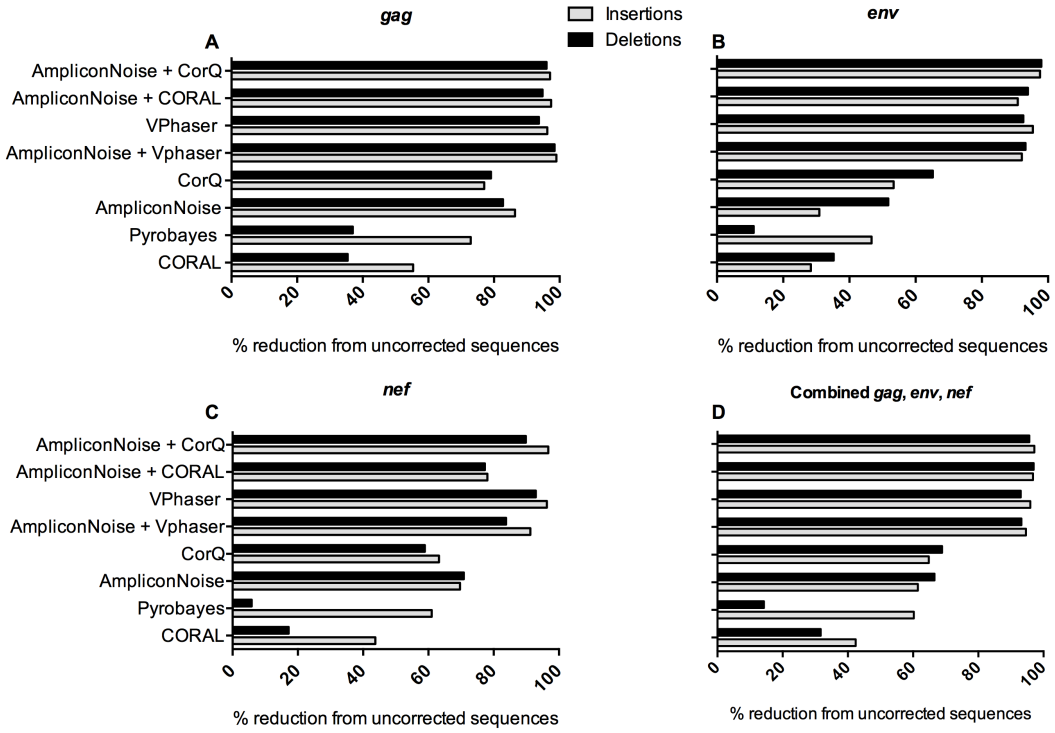

Figure 11. Attrition in indel counts after application of error correction methods. The percent reduction in number of indels within the HIV-1 ten-plasmid dataset compared to uncorrected sequences is presented. While Pyrobayes is not an error correction algorithm, but rather recalibrates quality values, the accuracy of recalibrated bases are meant to reflect overcalled and undercalled bases accurately. The % reduction in indels compared to uncorrected sequences is shown for *gag* (A), *env* (B) and *nef* (C), and all three genes combined (D). This figure is adapted from [155] under the terms of Creative Common Attribution License.

Carry forward error correction

The CorQ algorithm also corrects carry forward errors [119] near homopolymeric regions. The percent carry forward errors retained within reads after application of error correction methods is shown in Figure 12. Carry forward errors present in raw uncorrected reads are shown for comparison. In the uncorrected reads, carry forward insertion errors make up about 10-30% of the total insertions errors observed. Flowgram error correction (AmpliconNoise) corrects homopolymeric overcall insertion errors to a greater extent than carry forward insertion errors, hence, 20-30% of the insertion errors are of carry forward type after flowgram correction. Vphaser and CORAL corrected carry forward errors better than AmpliconNoise, but still retained about 10-15% of these errors. The combination of AmpliconNoise + CORAL performed only slightly better than using CORAL alone, retaining ~10% of carry forward errors. The combination of Vphaser correction followed by the carry forward correction segment of the CorQ program resulted in a further, substantial reduction in the number of carry forward errors compared to correction with Vphaser alone, retaining between 2-5% of these errors. The combination of AmpliconNoise + CorQ removed the most carry forward insertion errors, retaining only ~2%.

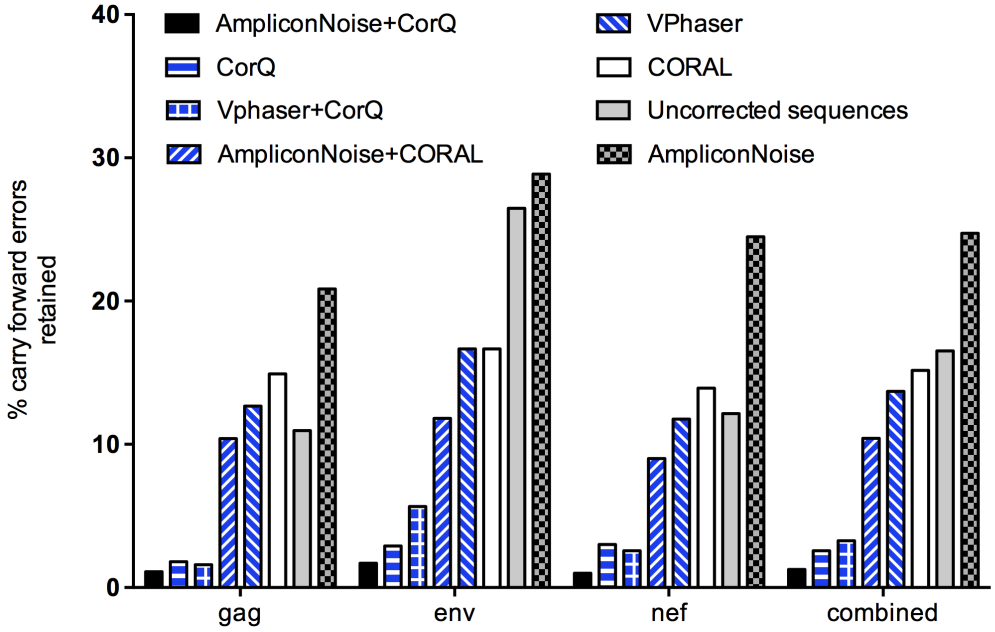

Figure 12. Carry forward errors retained after error correction. Raw uncorrected values and the percentage of carry forward errors retained after error correction with each of the algorithms listed is plotted for each of the three gene regions *gag*, *env*, *nef* and all the three genes combined. This figure is adapted from [155] under the terms of Creative Common Attribution License.

Sensitivity and specificity of pyrosequencing error correction programs

The sensitivity and specificity of SNP identification was then compared for four pyrosequencing error correcting and variant calling algorithms within the *gag*, *env* and *nef* gene regions from the 10 HIV-1 genome dataset. Since the mixture was derived from ten whole genome plasmids mixed in equal proportion, the lowest observable valid SNP would be 10%, with SNPs calls in pyrosequences validated by comparison to variants identified in Sanger sequences [152]. A total of 28 SNPs in *gag* (1500nt positions), 61 in *env* (2550nt) and 21 in *nef* (681nt) were compared. As shown in Table 3, the sensitivity of detection of variants was usually 97% or higher for most error methods, with the exception of the QuRe algorithm, which filters out regions with lower levels of coverage, and VPhaser when applied to the *nef* gene sequences. QuRe filtered out 3% of bases from correction for *gag* and *nef* but 33% of bases from correction

in *env*. These filtered regions fell within areas of poor coverage, usually at the start of the gene. V-phaser had reduced sensitivity on the *nef* dataset (61%) due to a change of valid SNPs to consensus in a region with an in-frame 18nt deletion present in 30% of the Sanger sequences. Changes to the *gapwindow* size parameter (to match the gap size observed within the sequences) as part of the Vphaser correction program did not improve *nef* sensitivity. A combination of AmpliconNoise + CORAL also showed reduced sensitivity, with values falling lower than each of these correction methods used individually. CORAL corrects errors by forming a multiple sequence alignment and generating a consensus sequence from these alignments. It is possible that the low frequency of "real" SNPs that are seen after flowgram correction are removed in CORAL when consensus sequences are generated, thus leading to a higher incidence of false negatives and reduced sensitivity. Similarly, we observed a reduced sensitivity when we combined AmpliconNoise with Vphaser, with sensitivity values falling lower than each of these correction methods used individually. The combination of AmpliconNoise + CorQ consistently resulted in higher sensitivity than the other tested error correction methods used individually or in combination.

Method	Sensitivity				Specificity			
	gag	***nef***	***env***	**combined**	***gag***	***nef***	***env***	**combined**
Uncorrected 454 reads	1	1	0.98	0.99	0.37	0.34	0.5	0.44
CorQ	1	1	0.98	0.99	0.79	0.86	0.94	0.88
AmpliconNoise	0.99	0.98	0.98	0.98	0.88	0.71	0.69	0.76
AmpliconNoise + CorQ	0.99	0.98	0.98	0.98	0.99	0.97	0.99	0.98
Pyrobayes + CorQ	0.97	1	0.98	0.98	0.78	0.7	0.78	0.77
CORAL	1	1	0.96	0.98	0.92	0.88	0.94	0.91
AmpliconNoise + CORAL	0.5	0.93	0.27	0.53	0.98	0.86	0.95	0.95
QuRe	0.96 (0.41)*	0.97 (0.61)*	0.98 (0.04)*	0.97 (0.11)*	0.97	0.92	0.99	0.96
SegminatorII	1	0.97	0.98	0.98	0.2	0.24	0.47	0.35
VPhaser	1	0.61	0.95	0.86	0.98	0.98	0.99	0.98
AmpliconNoise + VPhaser	0.54	0.25	0.41	0.38	1	0.99	1	0.99

Table 3. Comparison of CorQ against other pyrosequence error correction and SNP calling algorithms. *gag*, *env* and *nef* gene regions were used to compare the sensitivity and specificity of various algorithms. Sensitivity measures the proportion of true SNPs present in the ten HIV-1 genomes, and correctly identified by the various SNP calling programs. Specificity measures the proportion of true negatives (positions in the gene regions that are invariant) that are correctly identified by the compared programs. * Shown in parenthesis are values from QuRe when the poor coverage regions excluded from sensitivity analysis are included as false negatives. This table is adapted from [155] under the terms of Creative Common Attribution License.

With regard to specificity, the uncorrected reads had a high false positive rate (low specificity), and with the exception of SegminatorII each of the correction pipelines resulted in an increase in specificity. Repeated analyses with SegminatorII produced a high number of false positives, despite using a sample-specific consensus sequence as reference for the alignment and optimal settings recommended by the program authors. VPhaser alone, or flowgram correction (AmpliconNoise) in combination with CorQ, consistently produced the highest specificity for variant detection. Overall, combinations of error correction methods (AmpliconNoise + CORAL,

AmpliconNoise + Vphaser and AmpliconNoise + CorQ) consistently exhibited between 86 – 100% specificity.

We also performed a test to assess the effects of read coverage differences across sequencing orientations on the sensitivity and specificity of CorQ to detect and correct SNPs. We used pyrosequences mapping to the ~2500nt *env* region from the ten HIV-1 plasmid clones for this comparison and ran the combination of AmpliconNoise + CorQ with 2-fold, 5-fold, 10-fold and 20-fold coverage differences as thresholds for SNP correction (Table 4).

Coverage difference threshold	Sensitivity	Specificity
2 fold coverage	0.98	0.95
5 fold coverage	0.98	0.96
10 fold coverage	0.98	0.99
20 fold coverage	0.95	0.99

Table 4. Effect of varying coverage fold on sensitivity and specificity of SNP variant calling. AmpliconNoise + CorQ error correction was used on pyrosequences mapping to the *env* region (~2500nt) from the ten HIV-1 genome control dataset. Different fold coverage values were used as input parameters in CorQ. Sensitivity and specificity of SNP variant detection within this region is calculated for each fold coverage value. This table is adapted from [155] under the terms of Creative Common Attribution License.

As expected with a lower read coverage difference threshold (2- or 5-fold), more positions were marked to be poor coverage regions – SNPs falling within these regions are not corrected, resulting in higher false positives (reducing specificity to 95%). With higher coverage difference thresholds (20-fold), more regions with SNPs are corrected, resulting in correction of real variation present within the sequences and giving more false negatives (reduced sensitivity to 95%). We therefore used a 10-fold coverage difference (98% sensitivity and 99% specificity) with CorQ to achieve a balance between sensitivity and specificity.

Simulated pyrosequences

We tested the ability of error correction algorithms to reduce indel and substitution error rates in both homopolymeric and non-homopolymeric regions (Tables 5 and 6, respectively) on simulated pyrosequences generated with a single starting template (Sets 1a-c, Table 2). QuRe was not included in this analysis since it generates indel-removed haplotypes as the final result. SegminatorII was also excluded since it does not give indel information in the final results. The combination of AmpliconNoise + CORAL gives the highest reduction in substitution error rates for these simulated datasets. This mostly likely is a result of CORAL error correction whereby a regional consensus sequence is used to correct for low frequency variants. In the case where multiple sequencing templates are present, this correction method runs a risk of removing "true" low frequency variants (as we have shown with our sensitivity analyses), whereas in this case only a single template was used for simulation, and correction of low frequency variants was more efficient. Similar trends for indel and SNP error rate reduction was observed in homopolymeric and non-homopolymeric regions (compare Tables 5 and 6).

No simulated SNP errors	**Insertion**	**Deletion**	**Substitution**
Uncorrected	0.004	0.003	0.00032
AmpliconNoise	0.003	0.002	0.00013
CorQ	0.0035	0.002	0.00015
Pyrobayes + CorQ	0.003	0.0028	0.00026
AmpliconNoise + CorQ	0.0018	0.001	0.00003
CORAL	0.0008	0.0009	0
AmpliconNoise + CORAL	0.0003	0.0004	0
SNP error rate: 0.005			
Uncorrected	0.006	0.005	0.0025
AmpliconNoise	0.004	0.0045	0.0023
CorQ	0.005	0.0042	0.0018
Pyrobayes + CorQ	0.0056	0.0048	0.0019

AmpliconNoise + CorQ	0.002	0.0014	0.0002
CORAL	0.0008	0.0004	0.0002
AmpliconNoise + CORAL	0.0006	0.0002	0.00001
SNP error rate: 0.01			
Uncorrected	0.006	0.005	0.0045
AmpliconNoise	0.004	0.0044	0.0042
CorQ	0.0051	0.0044	0.0031
Pyrobayes + CorQ	0.0055	0.0047	0.0038
AmpliconNoise + CorQ	0.002	0.0018	0.00042
CORAL	0.0007	0.0009	0.0002
AmpliconNoise + CORAL	0.0003	0.0004	0.00002

Table 5. Comparison of insertion, deletion and substitution error rates in homopolymeric regions after error correction on simulated pyrosequences. Simulated reads were generated using Flowsim using a single 1500nt HIV-1 sequence as the starting template (Simulated datasets 1a-c). Average insertion, deletion and substitution error rates within homopolymeric regions are shown after correction with no additional SNP errors, and SNP error rates of 0.005 and 0.01. Table adapted from [155] under the terms of Creative Common Attribution License.

No simulated SNP errors	**Insertion**	**Deletion**	**Substitution**
Uncorrected	0.0008	0.0006	0.0001
AmpliconNoise	0	0.0001	0.0001
CorQ	0	0.0001	0.00001
Pyrobayes + CorQ	0.00001	0.0004	0.00009
AmpliconNoise + CorQ	0	0.00005	0.00004
CORAL	0	0.00002	0
AmpliconNoise + CORAL	0	0	0
SNP error rate: 0.005			
Uncorrected	0.0023	0.0027	0.002
AmpliconNoise	0.0021	0.0022	0.0018
CorQ	0.00098	0.001	0.0009
Pyrobayes + CorQ	0.0011	0.0019	0.0017
AmpliconNoise + CorQ	0.0001	0.0009	0.0006
CORAL	0.00007	0.0001	0.00002
AmpliconNoise + CORAL	0	0	0
SNP error rate: 0.01			
Uncorrected	0.0023	0.0027	0.0038
AmpliconNoise	0.0021	0.0022	0.0036
CorQ	0.001	0.0015	0.001

Pyrobayes + CorQ	0.0018	0.002	0.0026
AmpliconNoise + CorQ	0.0008	0.0009	0.0009
CORAL	0.00007	0.0002	0.0002
AmpliconNoise + CORAL	0.00004	0	0.00003

Table 6. Comparison of insertion, deletion and substitution error rates in non-homopolymeric regions after error correction on simulated pyrosequences. The simulated reads were generated in Flowsim using a single 1500nt HIV-1 sequence as the starting template (Simulated datasets 1a-c). Average insertion, deletion and substitution error rates within non-homopolymeric regions are shown after correction with no additional SNP errors, and SNP error rates of 0.005 and 0.01. Table adapted from [155] under the terms of Creative Common Attribution License.

Lastly we also evaluated the sensitivity and specificity of SNP identification on simulated pyrosequencing datasets. We used the three simulated datasets (Sets 2a-c, Table 2) with multiple starting templates (28 templates) for this analysis. Prior Sanger sequencing had shown a total of 145 positions with SNPs within these 28 templates [152]. We did not include SegminatorII in this comparison since our previous analysis with this program had shown that it led to lower specificity than raw uncorrected reads. Vphaser was also excluded as errors in the program led to consistently failed runs. When we compared the simulated sequences that lacked introduced SNP errors (Table 7), we observed very similar trends as observed with previous comparisons with the ten HIV-1 genome dataset (Table 3). As shown in Table 7, the sensitivity of detection was usually 95% or higher except in the combination of AmpliconNoise + CORAL that again showed a trend towards reduced sensitivity when combined. QuRe also showed reduced sensitivity when we included the poor coverage regions excluded by QuRe into our sensitivity calculations. When considering a balance between sensitivity and specificity, AmpliconNoise + CorQ performed the best amongst all the methods tested. As highlighted previously, PCR errors are harder for error correction algorithms to remove since these mutations are present within the sequencing templates. All error correction methods we tested on simulated pyrosequences with

additional SNP errors added to emulate PCR errors fared poorly for the removal of false positives with the best being AmpliconNoise + CorQ, with a specificity of 40%.

Method	Simulated pyrosequences	
	Sensitivity	**Specificity**
Uncorrected 454 reads	0.99	0.15
CorQ	0.99	0.70
AmpliconNoise	0.99	0.89
AmpliconNoise + CorQ	0.99	0.95
Pyrobayes + CorQ	0.98	0.71
CORAL	0.95	0.88
AmpliconNoise + CORAL	0.20	0.99
QuRe	0.99 (0.44)*	0.98

Table 7. Comparison of CorQ algorithm against other pyrosequence error correction and SNP calling algorithms. Simulated pyrosequences generated from 28 HIV-1 sequences as the starting template were used to compare the sensitivity and specificity of error correction algorithms. Sensitivity measures the proportion of true SNPs present within the HIV-1 templates used for simulation, and correctly identified as such by the various SNP calling programs. Specificity measures the proportion of true negatives (positions in the gene regions that are invariant) that are correctly identified as such by the compared programs. * Values from QuRe are shown when the poor coverage regions excluded from sensitivity analysis are included as false negatives (shown in parenthesis). Table adapted from [155] under the terms of Creative Common Attribution License.

Discussion

In this chapter we described a new pyrosequence error correction algorithm, CorQ that can identify and correct homopolymer and non-homopolymer indel errors, sequencing mis-incorporation errors and carry forward errors associated with homopolymeric regions. When applied to a control set of ten HIV-1 genomes (without PCR amplification), the combination of AmpliconNoise + CorQ reduced indel errors in the gene regions *gag*, *env* and *nef* by 94 to 97%. In addition to testing CorQ in combination with flowgram correction (AmpliconNoise) and base quality recalibration (Pyrobayes) programs, we also compared it to four recently published pyrosequencing variant callers, CORAL, QuRe, SegminatorII and V-Phaser. We found that

when CorQ error correction is used on flowgram-corrected fasta and quality files produced by AmpliconNoise, we get consistently higher sensitivity and specificity of SNP detection. To tease apart the contribution of CorQ and AmpliconNoise, we ran the programs separately, and found that CorQ by itself improved SNP detection specificity to a range of 79% to 94%, whereas AmpliconNoise by itself improved specificity to a range of 69% to 88%, whereas uncorrected reads had a SNP detection specificity ranging from 34% to 50%. Combining AmpliconNoise and CorQ consistently resulted in the highest combined SNP detection sensitivity and specificity amongst the error correction methods tested, with the specificity of VPhaser nearly equaling that of AmpliconNoise + CorQ. The combinations of AmpliconNoise + Vphaser and AmpliconNoise + CORAL while resulting in > 86% specificity, had poor sensitivity ranging from 25% - 93%.

The advantage of using AmpliconNoise + CorQ was most pronounced for the reduction carry forward errors. We also observed reductions in carry forward errors when we combined corrected files from Vphaser with CorQ, indicating that CorQ can be used in combination with other error correction programs to maximize the number of error free pyrosequences. We observed similar trends in sensitivity and specificity when we compared error correction methods on simulated pyrosequencing datasets. One caveat we observed in using AmpliconNoise is that it is computationally intensive, with computing time increasing exponentially on datasets over 20,000 reads, making this algorithm impractical for large datasets without extensive computational resources. Furthermore, since AmpliconNoise relies on iterative clustering, we observed that the frequencies of low-level SNPs did not correlate well with the frequencies found within uncorrected reads for sequences generated through amplicon sequencing on the Roche 454 platform (unpublished results). We therefore recommend use of AmpliconNoise for library pyrosequencing only, as described here.

CorQ takes read coverage into account when making SNP calls, particularly in regions in which there is a large discrepancy between the number of reads obtained in one sequencing orientation compared to the other. Other pyrosequencing error correction methods we tested here do not explicitly address read coverage variation across the target sequence or in different sequencing orientations. We addressed this by requiring a SNP to be present in both orientations. We also made read coverage difference threshold an input parameter for CorQ so that users can use the fold coverage that appropriately represents the data they are analyzing. We settled on a default setting of 10-fold coverage difference after initial tests showed this to achieve a good balance between SNP detection sensitivity and specificity. Thus, in regions with over a 10-fold difference in read coverage across sequencing orientations, SNPs are not corrected (by CorQ) due solely to inadequate information. While this criterion does not address all possible scenarios of read coverage across sequenced positions, we have observed that most regions with coverage discrepancies also tend to have inadequate or lack of reads in one of the sequencing orientations (unpublished observations).

While we were able to alter parameters within the CorQ program to handle both generated pyrosequences and simulated pyrosequences, altering multiple parameters simultaneously within other error correction algorithms was not always feasible. In order to ensure we were comparing the program performance with the most optimal parameters, we used multiple rounds of parameter optimization with all error correction programs. We also compared CorQ performance with other error correction programs in identifying and correcting SNP errors within simulated pyrosequences generated with varying mismatch error rates set to simulate PCR mismatch errors. As expected, none of programs evaluated were able to correct SNPs present in sequences as a result of misincorporation events occurring during PCR of the template

preparation, unless, in the case of CorQ, these SNPs also had reduced base quality. This makes identification of SNP errors as a result of PCR amplification challenging by any method as shown by our error correction tests run on simulated pyrosequences with typical PCR error rates applied.

We selected HIV-1 sequences as templates for generating additional simulated pyrosequences as this technology has become widespread in studying HIV-1 genomes. The genetic diversity of HIV-1 found within an infected individual in chronic infection is comparable to the global genetic variation seen in the influenza virus [21]. The most prominent source of HIV-1 mutation is error prone nucleic acid synthesis during replication, with rates estimated in the range of 1.4x10-5 errors per base pair, per replication cycle [160]. Viral diversity also differs in different genes and with the length of infection. The diversity of a viral population within an infected individual starts low immediately after infection but increases during the course of infection at a rate of 1% (within the *env* region) reaching up to 15% or more in long term infected individuals [161]. This extent of diversity makes pyrosequencing both a useful and challenging tool to study HIV-1. The information gleaned from pyrosequences thus has to be judged carefully for errors from both the sequencing methodology and PCR amplification.

CorQ lists frequencies of SNPs and outputs a multiple sequence alignment that can be used for downstream analysis of a variety of datasets, including microbial communities. Other error correction methods such as QuRe and V-Phaser that were tested here also generate reconstructed haplotypes that can be useful in studying microbial communities. Researchers interested in studying diverse microbial communities can use the information provided here to make decisions on selecting the right set of error correction tools. While we have tested CorQ on data derived from pyrosequencing, this algorithm is general enough to be applied to sequences

generated from other high throughput platforms that generate both sequence and associated quality files, making it a method with widespread applications in variant detection.

Author Acknowledgement

Study design: Shyamala Iyer, Dr. James. I. Mullins; Software implementation: Shyamala Iyer; Laboratory Experiments: Heather Bouzek, Eleanor Casey, Brendan Larsen; Additional software support: Wenjie Deng;

CHAPTER 3. Comparison of Major and Minor Viral SNPs Identified in Sanger and Pyrosequences in Early HIV-1 Infection

Summary

Massively parallel sequencing technologies, such as 454-pyrosequencing, allow for the identification of variants in sequence populations at lower levels than consensus sequencing and most single-template Sanger sequencing experiments, but there is little data that comprehensively compares each of these methods. In this chapter I will present results from a study comparing the single nucleotide polymorphisms (SNPs) in genetic variants observed during acute HIV-1 infection from 32 subjects that were assessed using both Sanger and 454-pyrosequencing. Pyrosequences derived from a median of 2400 viral templates per subject, and encompassing 50% of the viral genome, were compared to a median of five individually amplified full-length viral genomes sequenced using Sanger technology. There was no difference in the consensus nucleotide sequences in 27 of the subjects: among the remaining 5 subjects, disagreements were found in less than 1% of the sites evaluated (of a total of nearly 117,000 sites across all subjects). The majority of the SNPs observed only in pyrosequences were present at less than 2% of the subject's viral sequence population. Over 89% of all minor SNPs observed only in pyrosequences were found at a frequency below the detection threshold of Sanger sequencing for that subject. When only genomic positions with at least the same number of reads as the mean number of HIV templates sequenced were considered, the number of minor SNPs found only in pyrosequences was reduced by 50%, and the number of SNPs observed by both technologies was reduced by 15%. These results provide guidance regarding the design, utility and limitations of population sequencing variable template sources, and emphasize parameters

for improving the interpretation of massively parallel sequencing data to address important questions regarding target sequence evolution.

Introduction

Sanger sequencing has been widely used to study evolution of variable pathogens such as HIV, the emergence of drug resistance, and the rise of escape variants as a result of host immune pressure. Nonetheless, there are drawbacks associated with this technology. Individual template or cloned-derived sequencing is time-consuming, labor intensive, and is usually limited to tens of sequences per subject in consideration of cost. Consensus Sanger sequencing of virus populations can detect minority variants only above 10% - 25% of a heterogeneous sequence population [162] and with five individual Sanger sequences, the probability of observing a variant that represents at least 10% of the viral population is only 40% [163]. This resolution threshold is restrictive, especially when investigating minor HIV-1 variants. Massively parallel sequencing (MPS) technologies such as pyrosequencing, which involve individually amplifying and sequencing large numbers of DNA template molecules, have been applied extensively in HIV-1 research in an attempt to identify the presence of minority variants, particularly those relating to the emergence of clinically relevant HIV-1 drug resistance [122,123,125,130,131] and immune escape variants [124,126-129,132].

Four factors determine the level of minor variant resolution in MPS technologies such as 454-pyrosequencing, the related Ion Torrent system, and the Illumina platforms: a) error rate associated with the initial PCR amplification steps prior to sequencing; b) accurate quantitation of the number of amplifiable input template molecules; c) number of pyrosequences (reads) that map to the genomic position with the observed polymorphism; and d) resolution of errors that are inherent to the sequencing process.

PCR amplification can introduce errors within DNA products that can be indistinguishable from real variation after sequencing [155]. Thus, when looking for rare genetic variants, it is critical to use an enzyme with high fidelity and optimized conditions to ensure that the genetic variation found within the sequence population is representative of the virus and not an artifact of the amplification process [154].

While the term "coverage" is most often used to refer to the number of reads mapped to a genomic position, this quantity alone cannot be used to gauge the number of actual viral templates sequenced. Indeed, careful estimation of the number of amplifiable templates is rarely performed, but is essential to accurately measure population diversity [164-166]. The metric "sequencing depth", defined as the number of reads mapped to a genomic position divided by the number of estimated genome templates in the sequencing reaction was used in this study to represent template coverage. The read coverage obtained through library sequencing is often uneven [154,167], as is amplicon sequencing near the ends of the amplicons [154], hence, sequencing depth can vary greatly by nucleotide site across a sequenced region.

Errors within pyrosequences also follow distinct patterns compared to traditional Sanger sequencing. Consecutive runs of the same nucleotide (homopolymers) are particularly error-prone, resulting in inclusion of more or less bases in the read than is actually present in the DNA template. In pyrosequences generated by the GS-FLX Titanium technology, the mean homopolymer-associated error has been estimated at 1.1%, with errors showing a non-random distribution such as certain positions showing error rates as high as 50% [138]. Sanger sequencing has an estimated per-base error rate of <0.1% [168,169]. The relatively higher error rate in pyrosequences further complicates distinguishing real variants from sequencing artifacts [170-173].

Over the past few years, several error correction algorithms for identifying and correcting pyrosequencing artifacts have been described [142,143,145-151,155,174]. Despite the widespread use of pyrosequencing and associated error correction algorithms to identify minor HIV-1 variants, one should proceed with caution when asserting the biological significance of these minor variations as they approach the level of error rates. No comprehensive studies have been reported that compared the major and minor SNPs observed in pyrosequences and Sanger sequences obtained from the same source material. The studies that did compare variant populations focused heavily on the minor variants that were uniquely observed through pyrosequencing [175-177] without taking into consideration the impact of sequencing depth. Most authors reporting minor HIV-1 variants at a frequency range of 0.1% - 5% in pyrosequences do not factor in the actual number of viral templates sequenced [128,132,145], and thus have not estimated the true population frequency of those variants. Indeed, some studies have sought to have the number of templates in excess of the number of sequence reads [139,178], a protocol that further obscures the validity of individual sequences. One study that compared major and minor HIV-1 SNPs in a population of chronically HIV- infected individuals reported multiple instances of major HIV-1 variants (found in $\geq$50% of sequences) from pyrosequences that were not observed in Sanger sequences [175]. The probability of this, even with a limited number of Sanger sequences (e.g., five), is less than 20% [163]. However, as no quantitation of templates was done in that study, it is unclear whether pyrosequencing was performed on the same number of templates as were used in Sanger, or many times more, which could explain the discrepancy in the variants observed.

In this study we comprehensively compare the minor and major SNP variations observed in Sanger sequences and pyrosequences across three HIV-1 genomic regions, *gag*, *gp120* and

nef, in 32 subjects who became infected with HIV-1 during the HVTN-502 STEP vaccine trial [81]. The concordance between SNP frequencies in both sequencing technologies, and the effect of pyrosequencing error-correction algorithms on minor variant frequencies was assessed. We also investigated whether minor SNP variants specifically observed in pyrosequences were more frequently adjacent to error-prone regions, namely homopolymers [138,179]. Finally, we assessed also the impact of sequencing depth and the number of Sanger sequences on concordance and resolution of minor variants.

Materials and Methods

Study subjects. All 32 subjects were in acute HIV-1 infection (within 1.5 months of the first PCR-positive visit) and enrolled in the STEP HIV-1 vaccine trial (Clinical Trial Identifier: NCT00095576), a double-blind phase IIb test-of-concept study of the Merck Adenovirus-5 (MRK Ad5) HIV-1 clade B vaccine with *gag*, *pol* and *nef* inserts [37,38,81]. Institutional human subjects review committees at each of the clinical sites approved the vaccine protocol prior to trial initiation, and all study participants provided written, informed consent. The trial subjects that were examined in this study included 13 placebo and 19 vaccine recipients.

Sanger sequence polymorphism analysis. The Sanger sequences used in this study were derived from single amplifiable near-full-length viral genome (NFLG) templates, and have been deposited in GenBank under accession numbers JF320002-JF320643 [81]. Sequences from each subject were quality-checked and used to generate a multiple-sequence alignment using the HIV-1 strain HXB2 as the reference sequence. A consensus sequence was then generated for each subject and used as reference to realign the sequences. The web tool InSites (http://indra.mullins.microbiol.washington.edu/DIVEIN/insites.html) was used to identify the

positions of SNPs in the aligned Sanger sequences [180]. For the comparison to pyrosequences, InSites was used to distinguish positions with SNPs present in a single Sanger sequence (private sites) and those shared by more than one sequence (phylogenetically informative sites).

Identification of founder variants. The number of variants establishing productive infection (henceforth referred as founders) for each subject was identified from Sanger sequences [81] based on phylogenetic and genetic distance analyses. Probable multiple founders were identified based on shared polymorphisms (ranging in this set between 1-4), occurring in groups of at least two sequences, that were not shared with the remaining sequences [81]. A total of six of the 32 subjects were identified as having been infected with multiple founders.

Nucleic acid extraction for pyrosequencing. Using only the blood plasma samples with the same visit date as the previously derived Sanger sequences, RNA was extracted using the Qiagen Viral RNA Mini Kit (Qiagen, Valencia, CA). cDNA was synthesized using Superscript III Reverse Transcriptase (Invitrogen, Grand Island, NY) over three 1.5kb regions corresponding to *gag*, *gp120*, and *gp41-nef*, using the first-round reverse PCR primer. The list of primers used is provided in Supplementary Table 1.

PCR amplification. PCR amplification prior to pyrosequencing was done using Advantage LA or Advantage 2 DNA Polymerase (Clontech, Mountain View, CA). The viral template input was estimated using clinical viral load measures. First round PCR was a multiplex reaction, using primers to simultaneously amplify all three non-overlapping genomic regions, *gag, gp120,* and *gp41-nef*. The second round of PCR was done separately for each gene using nested primers (Table 8). Amplified products were visualized in agarose gels or using the QiaXcel capillary electrophoresis system (Qiagen). Endpoint dilution was performed to approximate the number of

amplifiable viral copies per gene using the Quality template-estimating program [165] (http://indra.mullins.microbiol.washington.edu/quality/). Once amplifiable template numbers were determined, additional PCR reactions were performed to amplify a target of up to 5000 templates for all the subjects in this study. PCR reactions were subsequently cleaned using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) and DNA concentrations were determined spectrophotometrically. Products from all three genes from individual study participants were pooled for 454-pyrosequencing.

Library preparation and pyrosequencing. Pooled and purified PCR amplified products were quantified using the Quan-it PicoGreen dsDNA assay (Invitrogen). GS-FLX Titanium kits were used for Rapid Library Preparation and Rapid Library MID Adaptor addition (Roche, Branford, CT). 500ng of each sample was nebulized, end repaired, and ligated with 454 library adaptors and MIDs. Fragments between 600-900bp were selected for and purified using AMPure beads. Library quality was assessed using the Agilent High Sensitivity DNA Bioanalyzer kit and chip (Santa Clara, CA), and the quantity of DNA was measured using the Quan-It PicoGreen dsDNA assay. Library concentrations were calculated using the online Roche Rapid Library Quantitation calculator. Each DNA library was diluted to a working stock of $1x10^7$ molecules/µl in TE buffer. Libraries generated from multiple samples (each with distinct sequence tags) were mixed at equimolar ratios. Emulsion PCR (Roche) was performed on the combined libraries using a ratio of 2-3 DNA molecules per bead. PCR-positive beads (~10-20% of emulsion PCR products) were then selectively enriched. Four million enriched beads were loaded onto a 454 picotiter plate and pyrosequences were generated using the 454 GS FLX system.

Primer Name	Gene	Tm (°C)	Length (bp)	Sequence	HXB2 5'	HXB2 3'
StepGF_1.0_578	Gag	59.7-61.3	27	AGTAGTGTGTGCCCGTCTGTTRTGTGA	552	578
StepGF_1.1_710	Gag	65.4	24	CGACGCAGGACTCGGCTTGCTGAA	687	710
StepGF_1.2_571	Gag	65	31	TGAGTGCTTCAAGTAGTGTGTGCCCGTCTGT	541	571
StepGF_2.0_792	Gag	60.3	24	GCGGAGGCTAGAAGGAGAGAGATG	769	792
StepGF_2.1_806	Gag	62.8	24	GAGAGAGATGGGTGCGAGAGCGTC	783	806
StepGR_1.0_2511	Gag	58.4	28	TTCCAATTATGTTGACAGGTGTAGGTCC	2484	2511
StepGR_1.1_2605	Gag	59.2	18	GGGCCATCCATTCCTGGC	2588	2605
StepGR_1.2_2597	Gag	57.3	30	CATTCCTGGCTTTAATTTTACTGGTACAGT	2568	2597
StepGR_1.3_2607	Gag	61.7	22	TTGGGCCATCCATTCCTGGCTT	2586	2607
StepGR_1.4_2836	Gag	54.8	25	TGTGGTATTCCTAATTGAACTTCCC	2812	2836
StepGR_2.0_2403	Gag	56.6	27	CAATTCCCCCTATCATTTTTGGTTTCC	2377	2403
StepGR_2.1_2336	Gag	53.4	24	TGCTCCTGTATCTAATAGAGCTTC	2313	2336
StepEF_1.0_5977	Gp120	60.8	24	GSCTTAGGCATCTCCTATGGCAGG	5954	5977
StepEF_1.1_5983	Gp120	57.3	23	GCATCTCCTATGGCAGGAAGAAG	5961	5983
StepEF_2.0_6233	Gp120	59.9-62.9	28	AGAGCAGAAGACAGTGGCAATGARAGYG	6206	6233
StepEF_2.1_6228	Gp120	58.5	23	AGAGCAGAAGACAGTGGCAATGA	6206	6228
StepER_1.0_7943	Gp120	59.2	21	GATGCCCCAGACYGTGAGTTG	7923	7943
StepER_2.0_7885	Gp120	58.1	27	TTRTTYTGCTGYTGCACTATACCAGAC	7859	7885
StepER_2.1_7819	Gp120	60.4	22	GCGCCCATAGTGCTTCCTGCTG	7798	7819
StepNF_1.0_7971	Gp41-Nef	61	23	GCTCCAGGCAAGARTCYTRGCTG	7949	7971
StepNF_2.0_8001	Gp41-Nef	58	27	AAAGATACCTAMRGGATCAACAGCTCC	7975	8001
StepNR_1.0_9643	Gp41-Nef	60	26	CACTACTTGAAGCACTCAAGGCAAGC	9618	9643
StepNR_1.1_9635	Gp41-Nef	57	25	GAWGCACTCAAGGCAAGCTTTATTG	9611	9635
StepNR_2.0_9554	Gp41-Nef	57.6	24	CTARCYAGAGAGACCCAGTACAGG	9531	9554
StepNR_2.1_9613	Gp41-Nef	61.3	24	TTGAGGCTTAAGCAGTGGGTTCCC	9590	9613

Table 8. PCR primers used in this study. *gag*, *gp120* and *nef* primers begin with the letters G, E and N, respectively. Forward and reverse primers are indicated with an F or R, and first and second round primers are denoted with 1 and 2, respectively. A suffix of 0, 1 or 2 is used to denote whether that primer was the initial or alternate primer. Positions relative to the HXB2 reference sequence at the 5' (R primers) or 3' (F primers) ends are listed in the primer name.

Pyrosequence data cleaning. Pyrosequences and their associated signal intensities were processed using the error correction program CorQ [155]. Briefly, signal intensities were clustered and corrected with AmpliconNoise [142,143] for an initial improvement of insertion and deletion (indel) and SNP errors. A reference-based multiple-sequence alignment with the corrected sequences was generated for each gene using the subject consensus from the Sanger sequences [81]. Due to poor primer homology within the *gp41* region that led to reduced read coverage and sequence quality after pyrosequencing, this region was excluded in subsequent

variant analyses. Following the construction of multiple-sequence alignments, a collection of Perl programs [155] were run on the aligned sequences and associated base-quality files to identify and correct regions with poor quality in a sequence-context dependent manner. Indel errors that resulted in frameshifts were corrected. Additionally, SNPs observed in only a single read were corrected to match the consensus at that position. No further mismatch error correction was applied.

Mismatch frequency threshold. The sensitivity of minor variant detection in pyrosequencing experiments is determined in part by the PCR conditions used to generate the templates for pyrosequencing, and the number of amplifiable templates in the reaction. Subsequent to the PCR amplifications done for this study, a number of DNA polymerases and varying PCR conditions were assessed to identify differences in sensitivity and mismatch error-rates [154]. Based on initial sensitivity estimates, the DNA polymerase enzymes used in this project, Advantage LA and Advantage 2 (Clontech), were found to have high sensitivity, but mismatch error-rates as high as 1% [154]. As currently available pyrosequencing error-correction programs are not equipped to filter out mismatch errors generated during PCR amplification [155], a conservative threshold of 1% was used as the limit of detection for all SNP analyses.

Terminology. Minor SNP variants were those observed at a frequency between 1-50% in the sequences of a given subject (major variants, >50%.) A major variant difference was defined as a polymorphism at a position in which the consensus base varied between the two sequencing methods. In most cases, the differences in major variants were due to "frequency reversal" of two relatively abundant variants. SNPs observed only in pyrosequences or Sanger sequences were classified as pyrosequencing-specific (PS-SNPs) or Sanger-specific (SS-SNPs) SNPs,

respectively. Shared SNPs are nucleotide differences from the consensus that occur at the same genomic location in both the Sanger and pyrosequences in a given study subject.

Sequencing depth. The metric sequencing depth is defined as the number of reads mapped to a genomic position (read coverage) divided by the number of estimated amplifiable genomic templates in the sequencing reaction. Following PCR amplification, end point dilution is performed to estimate the number of amplifiable viral copies per gene using the quality template-estimating program [165] (http://indra.mullins.microbiol.washington.edu/quality/). Sequencing depth is used as a measure of template coverage. The read coverage is as the number of reads that map to a genomic position and is determined after a reference based multiple sequence alignment of pyrosequencing reads. Mean number of amplifiable templates and mean sequencing depth for each subject is given in Tables 9-11.

Statistical methods. Spearman's rank correlation coefficient, or Spearman's rho (**ρ**) was used to estimate correlation between SNP frequencies. Kruskal-Wallis test was performed to compare the correlation among multiple groups with Dunn's error correction for multiple comparisons. The non-parametric Mann-Whitney test was used to compare two distributions.

Results

Study subject characteristics including number of Sanger and pyrosequences generated across the three sequenced regions: *gag*, *gp120* and *nef* is given in Tables 9-11. Overall, a strong correlation in SNP frequencies was found between Sanger and pyrosequencing data sets (Figure 13). Figure 14 illustrates SNP frequencies in the 26 subjects with a single founder variant. As expected for subjects within 1.5 months post HIV-1 infection, the majority of the positions (>97%) along the three genes had no observable polymorphisms. Subjects designated as having

multiple founders had a slightly lower number of non-polymorphic sites (92 – 97%), Figure 14. Only one subject with a single founder (502-2622) had an informative SS-SNP detected within one of the gene regions (*gp120*). As expected, private SS-SNPs and PS-SNPs were more prevalent across all gene regions in the individuals with replicating multiple founders. Additionally, we observed informative SS-SNPs in subjects with multiple founders (purple regions in Figure 14).

Single founders	Median reads	Mean templates	Mean sequencing depth	Sanger sequences (SGA clones)	Samples collected (months post infection)
502-0053	27564	4461	6	10	0
502-0388	1551	517	3	5	0
502-0525	3540	771	4	5	0
502-0572	24005	3513	7	5	0
502-0717	4998	1315	3	6	0
502-0923	9011	3939	2	8	0
502-1047	20943	2684	7	5	0
502-1478	2139	527	4	5	1.3
502-1799	8809	4557	2	5	0
502-2622	2007	1971	1	5	0
502-2667	11562	1171	10	5	0
502-2794	14644	176	83	5	0
502-0524	47151	3639	13	10	0
502-0648	4149	1635	2	4	0
502-0841	30581	3182	9	4	0
502-0897	12540	3323	3	10	0
502-1046	35486	55452	1	5	0
502-1191	19236	3324	6	5	0
502-1400	7882	3905	2	5	0
502-1500	9417	1808	5	3	0
502-1897	21123	152	138	5	0
502-1926	26533	1275	21	9	0
502-2241	5427	2158	2	11	0
502-2254	6003	3372	2	6	0
502-2349	3388	3315	1	4	0
502-2437	23138	3547	6	12	0
Multiple founders					
502-2008	36090	2349	15	10	0
502-0227	26812	3513	7	11	0
502-1174	31843	3168	10	10	0
502-1399	7155	2206	3	5	0
502-1619	6129	919	7	10	0
502-2649	5918	1275	5	9	0

Table 9. The *gag* median reads, *gag* mean amplifiable templates, *gag* mean sequencing depth, *gag* SGA clones and plasma sample collection time is shown for each subject in this study. The average gene length for *gag* is 1500 bases. The first available plasma sample was sequenced.

Single founders	Treatment assignment	Median reads	Mean templates	Mean sequencing depth	Sanger sequences (SGA clones)	Samples collected (months post infection)
502-0053	Placebo	11687	1601	7	10	0
502-0388	Placebo	15920	3225	5	9	0
502-0525	Placebo	4616	616	7	5	0
502-0572	Placebo	31337	3449	9	5	0
502-0717	Placebo	4503	1487	3	2	0
502-0923	Placebo	3995	3714	1	8	0
502-1047	Placebo	19506	3834	5	5	0
502-1478	Placebo	7498	1387	5	5	1.3
502-1799	Placebo	12614	4419	3	5	0
502-2622	Placebo	2509	2375	1	5	0
502-2667	Placebo	10486	3346	3	5	0
502-2794	Placebo	12254	169	72	5	0
502-0524	Vaccine	37310	3276	11	10	0
502-0648	Vaccine	400	638	0.4	4	0
502-0841	Vaccine	39475	3162	12	4	0
502-0897	Vaccine	7593	5110	1	10	0
502-1046	Vaccine	16794	56838	0.2	5	0
502-1191	Vaccine	17074	3285	5	5	0
502-1400	Vaccine	10690	2846	4	5	0
502-1500	Vaccine	1039	854	1	4	0
502-1897	Vaccine	10298	59	173	5	0
502-1926	Vaccine	12743	638	20	8	0
502-2241	Vaccine	4451	1776	2	11	0
502-2254	Vaccine	2229	2125	1	6	0
502-2349	Vaccine	12660	1972	6	4	0
502-2437	Vaccine	2689	800	3	12	0
Multiple founders						
502-2008	Placebo	13520	2425	5	10	0
502-0227	Vaccine	22137	3837	6	11	0
502-1174	Vaccine	13378	2336	6	10	0
502-1399	Vaccine	18771	2894	6	5	0
502-1619	Vaccine	4512	788	6	10	0
502-2649	Vaccine	6794	1559	4	9	0

Table 10. The *gp120* median reads, *gp120* mean amplifiable templates, *gp120* mean sequencing depth, *gp120* SGA clones and plasma sample collection time is shown for each subject in this study. The average gene length for *gp120* is 1530 bases. The first available plasma sample was sequenced.

Single founders	Treatment assignment	Median reads	Mean templates	Mean sequencing depth	Sanger sequences (SGA clones)	Samples collected (months post infection)
502-0053	Placebo	30449	4138	7	10	0
502-0388	Placebo	8464	1567	5	9	0
502-0525	Placebo	10925	981	8	5	0
502-0572	Placebo	35670	3849	9	5	0
502-0717	Placebo	4736	1626	3	2	0
502-0923	Placebo	6926	4202	1	8	0
502-1047	Placebo	14260	2758	5	5	0
502-1478	Placebo	2673	525	5	5	1.3
502-1799	Placebo	9901	3728	3	5	0
502-2622	Placebo	3226	2769	1	5	0
502-2667	Placebo	15908	3161	5	5	0
502-2794	Placebo	19959	270	72	5	0
502-0524	Vaccine	32456	3252	10	10	0
502-0648	Vaccine	788	458	0.4	4	0
502-0841	Vaccine	49899	4135	12	4	0
502-0897	Vaccine	3987	3837	1	10	0
502-1046	Vaccine	17529	14159	0.2	5	0
502-1191	Vaccine	20441	3832	5	5	0
502-1400	Vaccine	16571	4025	4	5	0
502-1500	Vaccine	2671	1689	1	4	0
502-1897	Vaccine	39230	225	173	5	0
502-1926	Vaccine	10264	458	20	8	0
502-2241	Vaccine	5461	1906	2	11	0
502-2254	Vaccine	5281	3358	1	6	0
502-2349	Vaccine	22510	3345	6	4	0
502-2437	Vaccine	2255	661	3	12	0
Multiple founders						
502-2008	Placebo	11620	2275	5	10	0
502-0227	Vaccine	24478	4083	6	11	0
502-1174	Vaccine	23995	3679	6	10	0
502-1399	Vaccine	22216	3492	6	5	0
502-1619	Vaccine	7090	1056	6	10	0
502-2649	Vaccine	5632	1364	4	9	0

Table 11. The *nef* median reads, *nef* mean amplifiable templates, *nef* mean sequencing depth, *nef* SGA clones and plasma sample collection time is shown for each subject in this study. The average gene length for *nef* is 610 bases. The first available plasma sample was sequenced.

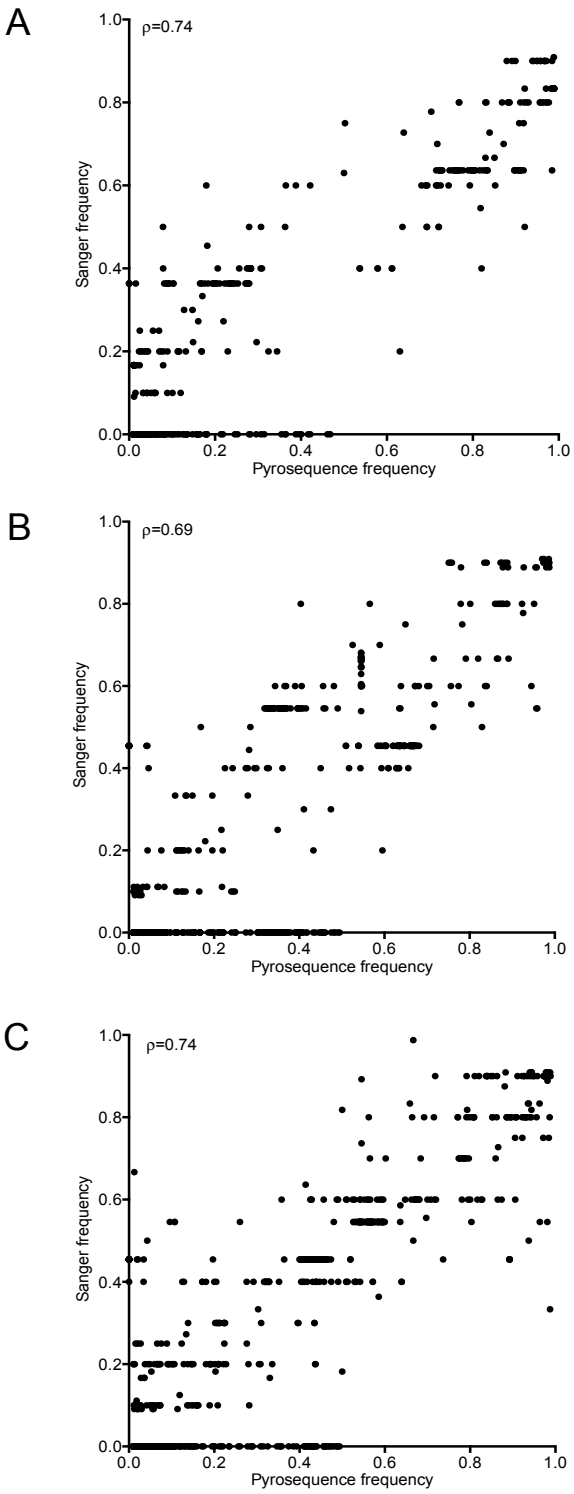

Figure 13. Correlation between SNPs observed in Sanger and pyrosequencing datasets. SNP frequencies are shown for *gag* (A), *gp120* (B), and *nef* (C). All types of SNPs evaluated (shared, Sanger-specific, and pyrosequencing-specific) from all 32 subjects are shown. Spearman's correlation coefficients are noted for each comparison.

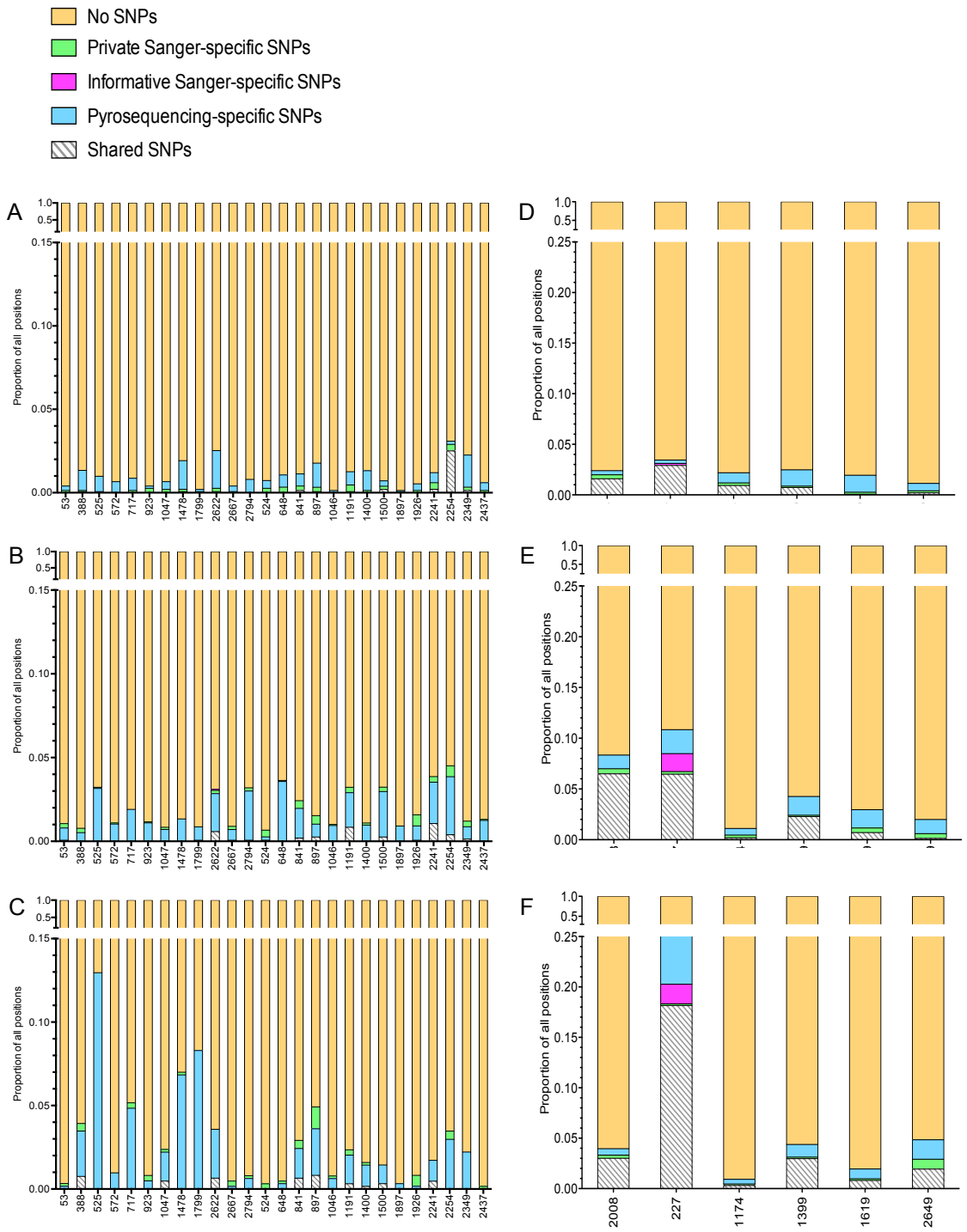

Figure 14. Proportion of positions with and without SNPs in subjects with single (A-C) and multiple (D-F) founders. The Y-axis shows the proportion of nucleotide positions in *gag* (A, D; 1500 nt), *gp120* (B, E; 1530 nt), and *nef* (C, F; 610 nt) that correspond to each category, with a linear scale and a split at 0.15 or 0.25. The X-axis indicated each subject ID (502-XXXX). The key shows the type of SNP observed.

Major variant comparisons

A consensus nucleotide sequence was generated from both pyrosequences and Sanger sequences for each subject over the 1500, 1530, and 610 nucleotide regions in *gag*, *gp120*, and *nef,* respectively. Among subjects infected with a single founder, the pyrosequence- and Sanger-derived consensus sequences were identical for 21 of the 26 subjects (80%) in *gag*, 23 (88%) in *gp120*, and 22 (84%) in *nef*, with the number of positions with consensus mismatches ranging from 1-3 (median = 1). In subjects with one or more consensus mismatches, there was an overall nucleotide identity of >99% in the consensus sequences. There were only two subjects (7%) in which a consensus base from pyrosequencing was absent in the Sanger sequences, however, both subjects had five or fewer Sanger sequences. All other instances of consensus mismatches were due to frequency reversals between shared major and relatively abundant minor variants. Among subjects infected with a single founder, there were no cases of the consensus Sanger variant being absent from the pyrosequencing dataset, and we found no evidence of consensus base discrepancies resulting from incomplete indel error correction of pyrosequences. Not surprisingly, for subjects with multiple founders, the consensus sequence concordance was lower, with all consensus base mismatches resulting from frequency reversals between the two most common variants. In addition, no relationship was found between the frequency of a variant in the pyrosequences and PCR primer sequence homology to that variant.

Minor SNP variant comparison

Figure 15A-C shows the frequency distribution of minor SNP variants in subjects with single founders. Most minor PS-SNPs (63%, 56%, 42% in *gag*, *gp120* and *nef*, respectively) represented <2% of the sequence population in subjects with single founders. Similar

distributions were found in subjects with multiple founders, although as expected, a higher fraction made up between 20-50% of the sequence population (Figure 15D-F).

To ensure that the observed PS-SNPs were not the result of artifactual mismatch errors adjacent to homopolymer regions [138,174], the sequence context of SNPs were assessed and no difference was found in the distribution of mismatches between homopolymer and non-homopolymer regions ($p=0.23$). Minor variant resolution within pyrosequences is also dependent on correctly estimating the number of amplifiable templates, as well as the number of reads mapping to each genomic position [154]. When only positions at which the sequencing depth was at least one (the number of reads was equal to, or greater than, the number of amplifiable templates used to derive products for the sequencing reaction) were considered, the number of positions with minor PS-SNPs was reduced by an average of 51%. We also quantified the number of PS-SNPs observed at a frequency below the expected Sanger sequencing threshold across all subjects and found that on average 89% of all PS-SNPs in all three gene regions were present below the detection threshold for Sanger sequencing. Of those, 81% in *gag*, 80% in *gp120*, and 60% in *nef* were present in <5% of pyrosequences.

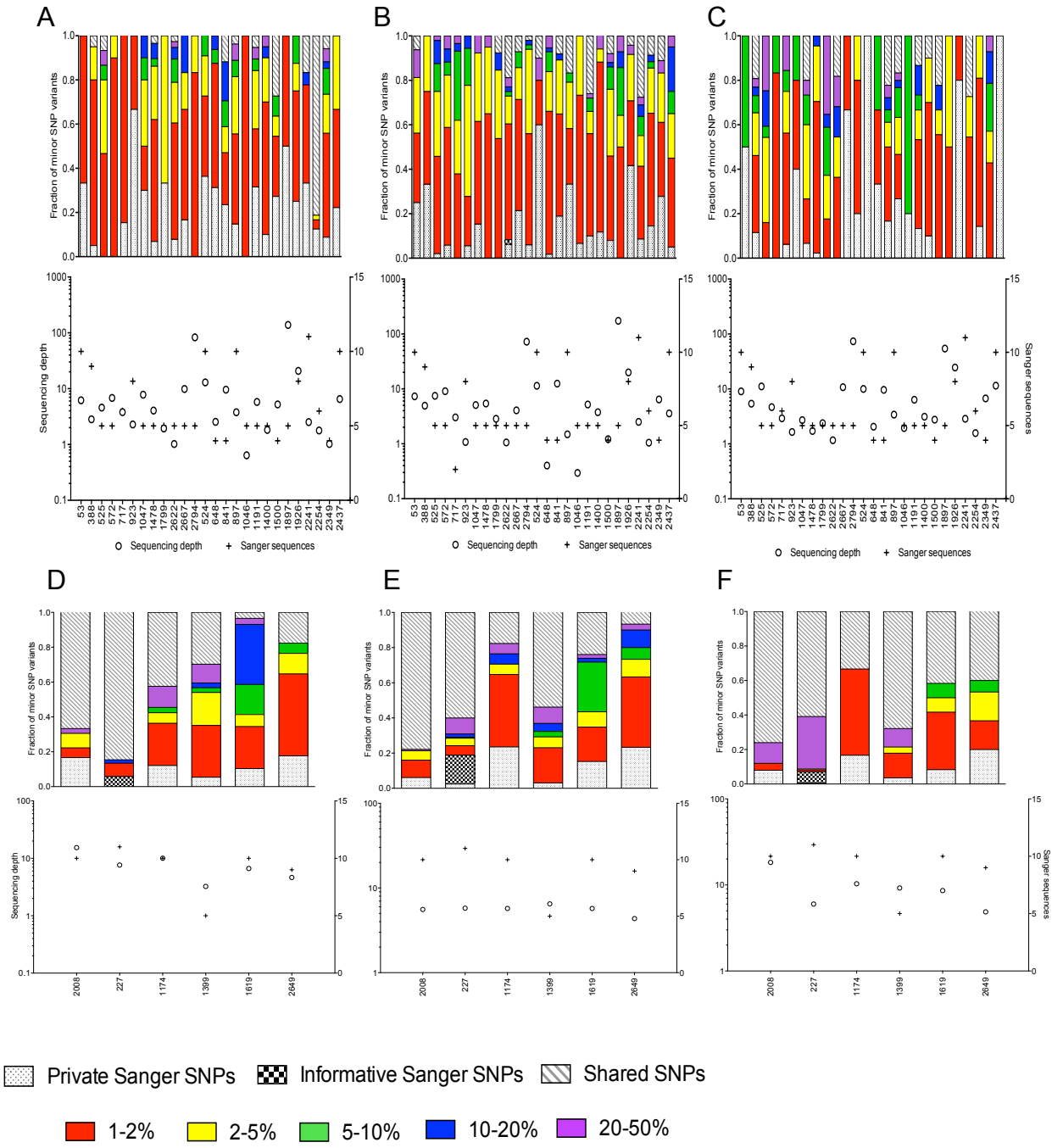

Figure 15. Frequencies of minor SNPs. Minor SNPs (frequencies between 1-50%) were compared in the 32 subjects with single and multiple founders across *gag* (A, D), *gp120* (B, E) and *nef* (C, F). The upper panels indicate for each subject the proportion of minor SNP variants in each category. The categories included shared (found in both Sanger and pyrosequences), Sanger-specific, including those that were Private (found in 1 sequence) and Informative (found in 2+ sequences), or found only in pyrosequences (with frequencies indicated by color: 1-2%, red; 2-5%, yellow; 5-10%, green; 10-20%, blue; 20-50%, purple). The lower panels show the pyrosequencing depth (o, left y-axis), defined as number of reads mapped to a position divided

by mean number of amplifiable viral templates, and the number of Sanger sequences (+, right y-axis). The X-axis lists the subject publication ID (502-XXXX)[81].

Within pyrosequences, we found phylogenetically informative Sanger SNPs well represented, with only one informative SNP missing within the pyrosequences. In this particular case there were inadequate pyrosequencing reads covering that site (sequencing depth <0.1). Among those infected with multiple founders, only one subject had informative SS-SNPs (Figure 15 (D-F). However, only one of 42 SS-SNP positions had a pyrosequencing depth of <1. Thus, their absence from the pyrosequencing data was not due to low coverage. As estimated previously for these subjects [81] a median of 2, 3 and 1 private SS-SNPs were present in *gag*, *gp120* and *nef*. Of the private SS-SNPs we observed 70% within pyrosequences, below the 1% threshold, and the remaining 30% were not observed within the pyrosequences at any frequency.

To understand the effect of sequencing depth on the resolution of minor frequency sequencing artifacts, we investigated the correlation between frequencies of PS-SNPs observed in positions as a function of sequencing depths ranging from <1 to >10 and observed no significant association in *gag* and *nef*. In *gp120* there was a significant ($p < 0.05$) increase in mean PS-SNPs frequencies comparing those with depth of < 1 to depth > 5.

Shared SNPs Distribution and Frequency Concordance

Individuals with multiple founders, as expected, had a higher fraction of positions with shared SNPs found in both sequencing platforms (*gag*: p=0.0003, *gp120*: p=0.0009, *nef*: p<0.0001). There was also high concordance between the frequencies of both shared major and minor variants in individuals with single (Spearman's Rho ρ=0.91 for *gag*, ρ=0.81 for *gp120,* ρ=0.90 for *nef*, p<0.0001 for all three regions, Figure 16A-C) and multiple founders (ρ=0.93 for *gag*, ρ=0.91 for *gp120* and ρ=0.44 for *nef,* p<0.0001 for all three regions, Figure 16 D-F).

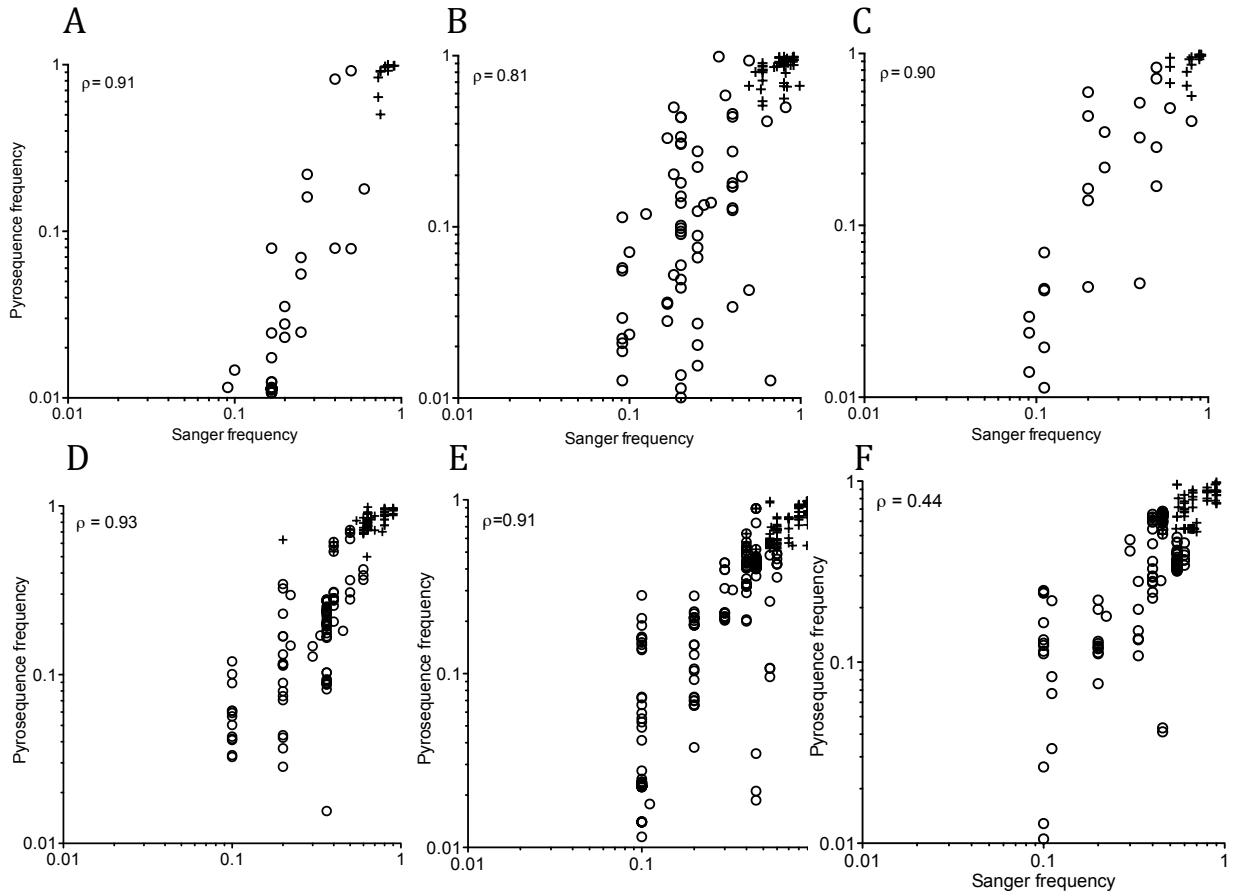

Figure 16. Comparison of frequencies of SNPs shared between Sanger and pyrosequences. Single founders: (A) *gag* (n = 54 SNPs), (B) *gp120* (n = 64), and (C) *nef* (n = 24*)*. Multiple founders: (D) *gag* (n = 103 SNPs), (E) *gp120* (n = 250), and (F) *nef* (n = 169*)*. Major (+) and minor (o) variant frequencies are plotted for subjects with a single founder, including positions at which the major and minor SNP frequencies were reversed in the two sequencing sets. Spearman correlation coefficients are shown for each comparison.

Higher sequencing depth would be expected to afford better agreement in variant frequencies between the two sequencing datasets. Among subjects with single founders, we observed no significant trend between frequency concordance and sequencing depth in *gag* and *nef* (ρ=0.07, ρ=-0.18, p=0.23). In subject *gp120* sequences, we observed reduced frequency difference with increasing sequencing depth (ρ=-0.28, p=0.01). In those with multiple founders, no trends were observed in *nef* or *gp120* (ρ=0.21 and ρ=0.08, p=0.42), however, there was a lower absolute frequency difference with increasing pyrosequencing depth in *gag* (ρ=-0.32,

p=0.0008). When sequencing depth was applied as a filter for shared SNPs below a depth of 1, the number shared SNPs was reduced by 12-19%, highlighting the issues of low read coverage affecting SNP resolution.

We also studied the effect of the number of Sanger sequences generated for a subject with the corresponding absolute difference in shared SNP frequency, for both consensus and minor variants. In subjects with a single founder variant, a significantly increased frequency concordance was observed between subjects with ≥10 Sanger sequences compared to those with ≤5 Sanger sequences in *gag* ($p<0.0001$) and for those with ≤6 vs. ≥10 Sanger sequences in *gp120* ($p<0.05$). A similar but non-significant trend was observed in *nef*. Among subjects with multiple founders, the results were similar, except that in *nef* there was a trend towards less concordance with more Sanger sequences (≤5 vs. 11, $p = 0.01$). However, a single subject with 11 Sanger sequences largely drove this result as this subject had a large proportion of positions with frequency reversals.

Discussion

Single-nucleotide polymorphisms observed in pyrosequencing data were compared to those observed in Sanger sequences in order to determine the concordance between the two technologies, and to assess the quality and utility of the information provided by the greater depth of massively parallel sequencing. Greater than 99% concordance was found between consensus bases from the two sequencing sets. Consensus differences were infrequent and generally associated with decreased numbers of Sanger sequences or frequency reversals between major and frequently observed minor variants. There was no evidence that primer

mismatches led to preferential amplification of one variant over another in subjects with multiple founders that would explain the rare consensus base differences observed.

As one strength of massively parallel sequencing technologies derives from the detection of minor variants, the distribution of minor SNPs was assessed - a majority (>53%) of the PS-SNPs observed were rare in the viral sequence population (<2%), whereas only 9% of the PS-SNPs observed in pyrosequences were found in more than 25% of the sequences. That the majority of SNPs fell below 2% of the sequence population is not surprising given the large number of reads generated through pyrosequencing. However, one has to be cautious about overstating the biological significance of these minor variants as other factors such as PCR amplification conditions, polymerase error rates, pyrosequence error correction and numbers of quantifiable viral templates all influence the threshold applied to resolve minor variants.

Only one informative SS-SNP (from an individual with a single founder) was observed at a high level (~40%) in Sanger sequences that was absent in pyrosequences. However, the pyrosequencing depth at this position was less than 0.1, which likely explains its absence in pyrosequences. Among subjects with multiple founders, one had phylogenetically informative SS-SNPs, despite the majority of positions having a pyrosequencing depth of at least one. This subject also showed the highest number of consensus base mismatches of all subjects. These differences could be the result of using different plasma vials and cDNA preparations, or an indication of the stochastic nature of quantitation of multiple viral variants. All other phylogenetically informative SNPs observed within the Sanger sequences were observed in the pyrosequencing dataset above the 1% threshold. This result lends confidence that all the consensus viral templates observed within the Sanger sequence population were also adequately sequenced by pyrosequencing. Private-site SNPs within Sanger sequences were found at a

frequency of 0.22% per nucleotide sequenced. Surprisingly, 70% of these were observed below the 1% threshold and the remainder not observed within pyrosequences at all. This suggests the possibility that private-site SNPs observed within Sanger sequences correspond to sequencing errors rather than simply reflecting low sampling depth. This result was unexpected since each Sanger-derived viral genome sequence corresponds to the consensus of reads derived from a single viral template and thus should not include PCR errors.

An average of 43 PS-SNPs per subject across all the three genes (89% of all PS-SNPs detected) were observed at frequencies below their respective Sanger sequencing thresholds. However, as the majority of these SNPs were present at a frequency of <5%, diligence must be applied to minimize external sources of error to improve the accuracy of the observed polymorphisms.

Pyrosequencing error patterns can skew minor variant distribution and frequencies [138,155,174]. However, following correction [155] the distribution of PS-SNPs adjacent to homopolymer and non-homopolymer regions showed no significant differences. Variant bias can also be introduced by PCR during viral template amplification [139,181,182]. Nonetheless, shared SNPs showed a high degree of correlation (average Spearman's ρ of 0.87) between the two sequencing methods, suggesting that the impact of PCR bias in this study was minimal, although, it remains an important consideration in the design and implementation of a PCR protocol for discerning true rare variation from sequencing artifacts.

Sequencing depths of 1 or below are not adequate to resolve low frequency sequencing artifacts from genuine low frequency variants present within viral templates. An excess of reads compared to input viral templates will help fine-tune minor frequency SNP calls. Unfortunately,

due to the uneven sequencing coverage observed with library sequencing [154,167], more than 50% of the positions with PS-SNPs observed within the current dataset were located in regions with a sequencing depth < 2 (0.71% of all sites in the current dataset). Additionally, while the number of PS-SNPs were reduced in positions with higher sequencing depth due to limited number of positions with depth >5 (0.23% of all sequenced sites in the current dataset), we did not observe significant changes in mean frequencies between PS-SNPs from positions with <5 or >5 depth in two of the three genome regions sequenced. An ideal comparison might be accomplished by analysis of pyrosequences from a genomic region with known viral templates and varying the sequencing depth to quantitate the advantage of higher sequencing depth in resolving low frequency sequencing artifacts.

The concern over higher error rates, especially from the 454 pyrosequencing and Ion Torrent platforms [183,184], necessitates the application of a conservative frequency threshold and additional filters in order to reduce or eliminate sequencing artifacts. The metric "sequencing depth" used here illustrates that increased read coverage with respect to number of amplifiable templates is associated with increased accuracy in the SNP frequencies at that position. As sequencing depth relies on read coverage and amplifiable templates, regions with poor read coverage or samples with large numbers of viral templates can decrease sequencing depth and subsequent confidence in the validity of observations of minor variants.

The results presented here provide guidance about each sequencing method's applications and limitations for assessing sequencing populations variability, and emphasize parameters critical for interpretation of massively parallel sequencing data.

Author Acknowledgement

Study design: Shyamala Iyer, Dr. James. I. Mullins; Software implementation: Shyamala Iyer; Data analysis: Shyamala Iyer, Eleanor Casey, Dr. Morgane Rolland; Laboratory experiments: Eleanor Casey, Heather Bouzek, Moon Kim, Hong Zhao, Brendan Larsen; Additional software support: Wenjie Deng

CHAPTER 4. MRKAd5 HIV-1 Vaccine-Induced Immune Selection leads to reduced T-cell Epitope Diversity and reduced rates of Epitope Evolution

Summary

In this chapter I will present results from a study that analyzed HIV-1 *gag*, *gp120* and *gp41-nef* genes from 64 volunteers enrolled in the STEP HIV-1 vaccine trial who became infected in the course of the trial. To identify genetic signatures of vaccine-induced anamnestic[1] pressure on breakthrough viruses we compared epitope distances from vaccine insert and average pairwise distances within epitope variants. We found significantly greater distances (p=0.0002) to the vaccine insert within CTL epitopes in Gag among the vaccine recipients. Amino acid residues flanking CTL epitopes within Gag sequences in vaccine recipients also had significantly greater distances (p=0.0003) to the vaccine insert. Vaccine-induced anamnestic responses could lead to increased accumulation of CTL-mediated mutations within the breakthrough founder variants. We observed a trend towards higher diversity within Gag epitope sequences in vaccine recipients infected with a single founder. Influence of vaccine-induced immune pressure on CTL epitope evolution and divergence from breakthrough founder virus was estimated in 30 subjects followed up to 20 months post infection. Gag epitope diversity, as estimated by average pairwise distances, within vaccine recipients decreased over time. Similarly, CTL epitopes within Gag in vaccine recipients had decreased rates of divergence from founder variants. No evidence for genetic signatures[2] of vaccine-induced immune selection was observed in protein regions not overlapping with CTL epitopes. The fact that we observed evidence for vaccine-induced immune selection only

[1] Renewed and rapid response by T lymphocytes and Abs on the second (or subsequent) encounter with the same Antigen

[2] Influence of host immune responses on viral sequences through selection

within predicted epitopes, and only within a protein that was a component of the vaccine strongly supports the idea that the MRKAd5 vaccine resulted in T-cell mediated selection occurring post-infection. This study provides the first evidence of vaccine-induced anamnestic pressure influencing CTL epitope evolution and epitope diversity during HIV-1 infection.

Introduction

Intrinsic viral diversity among circulating HIV-1 strains presents a major challenge for global vaccine development. The rationale for a T-cell-based vaccine, as used in the STEP/HVTN502 Phase II trial, involving the Merck Adenovirus 5 (MRKAd5) Subtype B *gag-pol-nef* vaccine [37-39], was the generation of T-cell responses that could control viral replication and attenuate HIV-1 pathogenesis. This vaccine was based on results from several studies that demonstrated a role for HIV-1 and SIV specific cytotoxic T lymphocytes (CTLs) in controlling infection [12,33,61,63,65-68,73,185]. However, translating these results into a successful T-cell based HIV-1 vaccine has proven to be difficult, partly because there remains a lack of clear understanding about the immune correlates of protection in humans [59,186-188]. This was made evident by the failure of the MRKAd5 STEP vaccine to prevent HIV-1 acquisition or reduce the viral load set point [37,38]. The MRKAd5 vaccine did induce HIV-1 specific CD8+ T-cell responses in the majority of vaccinees, but the responses were relatively weak and were only directed against a small number of epitopes [37,38].

Our group previously conducted a study on the genetic impact of vaccination on founder (breakthrough) HIV-1 sequences from STEP trial participants [81]. This work was the first to demonstrate evidence of selective pressure from vaccine-induced T-cell responses on breakthrough viruses. Viruses from vaccine recipients showed greater genetic distances to the vaccine insert sequence compared to the viruses infecting the placebo group, particularly in predicted T cell

epitope regions in Gag ($p < 0.0001$). These effects could have been due to an acquisition sieve effect, wherein vaccine-induced immune pressure excluded certain variants from establishing a productive infection. Alternatively, anamnestic pressure could result in the accumulation of more or faster acquisition of mutations (post-infection sieve effects) [81,114]. This study addresses both of these hypotheses in an attempt to better understand the extent of the genetic impact of vaccination observed in the STEP study.

In this work we analyze three potential scenarios for vaccine-induced immune responses and its impact on the infecting HIV-1 strain(s): a) Acquisition sieve analysis, evaluating whether vaccine-induced immune responses blocked certain variants genetically similar to the vaccine insert from replicating and establishing infection in the subject; b) Post-infection sieve analysis, evaluating whether vaccine-induced immune pressure lead to selection of escape mutations within CTL epitopes immediately after infection, and; c) Post-infection sieve analysis within study subjects followed up to a maximum of 20 months post-infection, to evaluate whether vaccine-induced anamnestic responses contributed to differing rates of CTL epitope evolution and escape. We amplified *gag*, *gp120* and *gp41-nef* genes from 64 STEP study participants who became infected during the course of the trial. T-cell epitope regions within Gag and Nef were observed to have the most significant results from the previous study [81]. Gp41 and Gp120 were not included as part of the vaccine insert and therefore used in this study as control regions. We used 454-Roche pyrosequencing to sequence the viral populations from subjects in acute infection as well as subjects followed longitudinally for up to 20 months post infection. The computational pipeline used to analyze the generated pyrosequences was previously published by our group [155] and described here in Chapter 2. Our analyses were focused on predicted T-cell epitope regions (9mers) as well as an extended epitope region (19mer) that included the five amino acids flanking

the predicted epitopes. Flanking regions can be critical for effective antigen processing and maintenance of epitope integrity, and escape mutations within these regions can lead to inefficient antigen processing and presentation [189,190]. As in our previous study [81], we calculated the genetic distance of founder epitope variants to the vaccine insert sequence for evidence of both an acquisition effect as well as evidence for vaccine-induced anamnestic responses. To address the post-infection sieve effects we calculated changes in pairwise diversity within predicted CTL epitope regions and estimated the fraction of epitope variants identical to the vaccine that were present in a low frequency within the subject's virus population. Additionally, we estimated the rate of genetic divergence from the founder epitope variants in infected subjects followed for up to 20 months post infection.

Results from the current analyses confirm our previously observed results [81]. As previously observed, immune pressure, recognized by genetic imprinting on the virus, was localized to predicted CTL epitope regions in Gag. We also found evidence for selective pressure within the five amino acid regions flanking predicted epitopes in Gag. Our post-infection sieve analysis showed a trend towards higher diversity within predicted CTL epitopes in vaccine recipients sequenced at the first time point post-infection, suggesting increased immune pressure within T-cell epitopes leading to accumulation of more mutations. Additionally, we found evidence for vaccine-induced anamnestic responses resulting in decreased epitope diversity and reduced rates of epitope evolution in vaccine recipients. Our results strongly support the idea that the MRKAd5 vaccine resulted in T-cell mediated selection occurring post-infection. This study is the first evidence of vaccine-induced anamnestic pressure influencing CTL epitope evolution and epitope diversity during HIV-1 infection.

Materials and Methods

The STEP/HVTN502 vaccine trial (Clinical Trial Identifier: NCT00095576) was a double blind phase-IIb test-of-concept study of the Merck Adenovirus-5 (MRKAd5) HIV-1 Subtype B vaccine with Gag, Pol and Nef inserts. Institutional human subjects review committee at each of the clinical sites approved the protocol prior to study initiation and all study participants provided written, informed consent. All subjects used in this study were enrolled in the STEP trial, and became infected with HIV-1 Subtype B virus: one subject was infected with HIV-1 Subtype C and was not included in our analyses.

Subject samples were classified as deriving from acute infection (first time point samples) if either PBMC or plasma was obtained within one month of the date of confirmed infection. Additionally, samples that were within 1.5 months of the date of confirmed infection were also designated first time point samples if these were from the first HIV-1 positive date available. PBMC and plasma samples were also sequenced from subjects that were followed for up to 20 months post-infection (longitudinal samples). Table 9 indicates the number of subjects in both the vaccine and placebo groups, as well as the number of first time point and longitudinal samples.

	Vaccine	**Placebo**
Subjects with samples collected up to 1.5 m.p.i		
First timepoint (<1.5 mpi)	32	23
Study volunteers common with Rolland et. al [81]	*19*	*16*
Subjects with samples collected up to 20 m.p.i		
Subject samples collected up to 20 m.p.i	27	18
Unique subjects with longitudinal sample data	19	11
Plasma and PBMC samples		
Plasma	32	23
PBMC	31	22

Table 12. Classification of study samples. A total of 55 subjects (32 vaccine and 23 placebo) were sampled within 1.5 months post infection (m.p.i). Number of subjects in each treatment group that were in common with the Rolland *et. al* [81] study were collected at various visits up to a maximum of 20 m.p.i. Samples that were classified as PBMC or plasma are also listed. A total of 108 subject samples were sequenced in this study (Pyrosequencing samples: 99, Sanger sequencing samples: 9)

Identification of founder variants

The number of viruses that established productive infection after multiple rounds of replication (founder variants) in each subject was identified based on phylogenetic analyses of near-full-length genomic sequences acquired by prior Sanger sequencing [81]. Subjects in which more than one founder virus established productive infection (multiple founders) were identified based on shared polymorphisms (ranging in this set between 1-4) occurring in at least two sequences that were not shared with the remaining sequences [81]. Of the 55 subjects sequenced in acute infection as part of this study, a total of nine (16%) subjects were identified as being infected with multiple founders. Additionally, subject pyrosequences were also examined for evidence of establishment of productive infection by more than viral variant.

RNA extractions

RNA from plasma was extracted using the Qiagen Viral RNA Mini Kit (Qiagen, Valenica, CA). cDNA synthesis was conducted using SuperscriptIII Reverse Transcriptase (Invitrogen, Grand Island, NY) for *gag*, *gp120*, and *gp41-nef* using gene-specific primers (*gag*: StepGR_1.0 5'-TTCCAATTATGTTGACAGGTGTAGGTCC-3', *gp120*: StepER_1.0 5'-GATGCCCCAGACYGTGAGTTG-3', *gp41-nef*: StepNR_1.0 5'-GATGCCCCAGACYGTGAGTTG-3'). A detailed list of primers used in this study is provided in Table 8 (page 57). DNA extractions from PBMC were done using the standard protocols supplied with the QIAamp DNA Blood Mini Kit and QIAcube extraction robot (Qiagen).

PCR Amplification

PCR amplification prior to pyrosequencing was done using Advantage LA or Advantage 2 DNA Polymerase (Clontech, Mountain View, CA). Template input was calculated using clinical viral load measures. The first round of PCR was a multiplex reaction, using primers to amplify all three genomic regions, *gag, gp120,* and *gp41-nef*. The second round of PCR was done separately for each gene using nested primers (Table 8, page 57). Amplified products were visualized by agarose gel electrophoresis or the QiaXcel system (Qiagen). Endpoint dilution was performed to approximate the number of amplifiable viral copies per gene using a template-estimating program, Quality [165]. Once template number was determined, additional PCR was done to amplify a target of 5,000 templates for first time points and a target of 1,000 templates for longitudinal time points. PCR reactions were subsequently cleaned using Agencourt AMPure XP Beads (Beckman Coulter, Brea, CA) and DNA concentrations were quantified. All three genes were pooled so that a total of 2µg of each sample was submitted for Roche 454 pyrosequencing. Samples that were quantified to contain less than 50 copies were sequenced using the Sanger method (Figure 17).

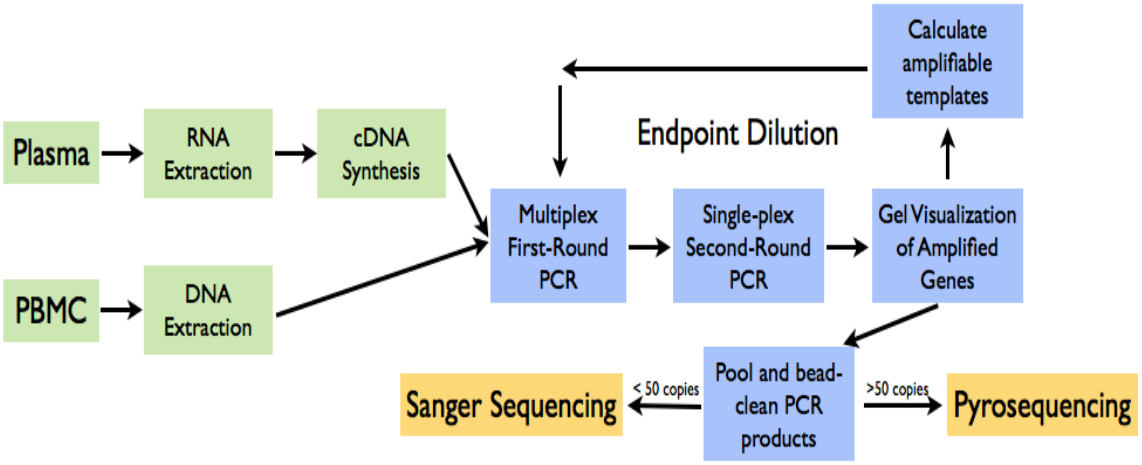

Figure 17. Overview of the experimental protocol. Samples with <50 copies were sequenced by Sanger sequencing. Samples with >50 amplifiable copies were sequenced by pyrosequencing.

For first time point samples (<1.5 months post-infection), the target number of templates for pyrosequencing was 5,000 per subject. For longitudinal samples (1.5-20 months post-infection) the target was 1,000 templates per subject.

Pyrosequencing

Pooled and purified PCR-amplified products from each study participant were quantified using Quan-it PicoGreen dsDNA assay (Invitrogen). GS-FLX Titanium kits were used for Rapid Library Preparation and Rapid Library MID Adaptor addition (Roche, Branford, CT). 500ng of each sample was nebulized to fragment the DNA, end repaired, and ligated with 454 library adaptors and sequence identifiers. Fragments between 500-900bp were selected for (based on bead concentration) and purified using AmPure beads. Library quality was assessed using the Agilent High Sensitivity DNA Bioanalyzer kit and chip (Santa Clara, CA), and the quantity of DNA was measured using a Quan-It PicoGreen dsDNA assay. The library concentration was calculated using the online Roche Rapid Library Quantitation calculator. Each DNA library (one per sample) was diluted to a working stock of $1x10^7$ molecules. Emulsion PCR (Roche) was performed on the combined libraries using a ratio of two or three DNA molecules per bead. PCR-positive beads (~10-20% of emulsion PCR products) were selectively enriched. Four million enriched beads were loaded onto the 454-picotiter plate and pyrosequences were generated using the 454 GS FLX system.

Pyrosequence data cleaning

Pyrosequences and the associated signal intensities generated were processed using the error-correction program CorQ [155] (Chapter 2). Briefly, signal intensities were clustered and corrected with AmpliconNoise [143] for an initial correction of insertion and deletion (Indel) and SNP errors. A reference-based multiple-sequence alignment was generated for each gene with

the corrected sequences using the subject consensus from prior Sanger sequencing [81]. Subsequent to multiple sequence alignments, the CorQ program outlined in in Chapter 2, [155] were run on the aligned sequences and the associated base-quality files to identify and correct regions with poor quality in a sequence-context dependent manner. Indel errors resulting in frame shifts from both homopolymer (defined as a run of two or more identical bases) and non-homopolymer regions were corrected. Additionally, SNPs observed in one read were corrected to match the consensus base at that position. No further mismatch error correction was applied. The corrected nucleotide alignment were then translated into a codon alignment. Regions within the alignment corresponding to predicted and known CTL epitopes were extracted for distance, average pairwise diversity and divergence estimates. Due to poor primer homology and read quality in *gp41*, the region was not included in any sieve analyses: the *gag*, *gp120* and *nef* genes were extracted from each alignment and processed according to our protocol.

Mismatch frequency threshold

The sensitivity of minor variant detection in pyrosequencing experiments is determined by PCR conditions, and the number of amplifiable templates input into the reaction [154,155]. Subsequent to the PCR amplification done for this study, we amplified and analyzed HIV-1 regions using a number of DNA polymerases and varying PCR conditions to identify differences in sensitivity and mismatch error-rates [154]. Based on initial sensitivity estimates, the DNA polymerase enzymes used in this project, Advantage LA and Advantage 2 (Clontech), were found to have mismatch error-rates as high as 1% [154]. As currently available pyrosequencing error-correction programs are not equipped to filter out mismatch errors generated during PCR amplification [155], we conservatively applied a threshold for the limit of detection at 1% for all

epitope and sieve analyses, as real polymorphisms or variant occurrences observed less frequently would be indistinguishable from enzyme-related errors.

Epitope Analyses

We included both known HIV-1 epitopes and potential CD8+ CTL HIV-1 epitopes predicted using NetMHC [191] for each study participant. Epitopes were identified in the MRKAd5 vaccine insert for Gag and Nef genes, and in the ConsensusB (2004) sequence for the control protein Gp120. NetMHC predicts peptide binding to 4-digit HLA alleles, and we included both the predicted strong and weak binders. Known epitopes listed at the Los Alamos National Laboratory HIV database were also included within the list of CTL epitopes. We performed sieve analyses on both epitope regions and non-epitopic regions within a subject. The non-epitopic regions within a subject are defined as the non-overlapping regions that lie outside of all predicted epitopes for that particular gene. We analyzed extended epitope regions that included up to 5 AA residues flanking the predicted epitopes.

Genetic distance of founder strain to vaccine insert

We calculated the genetic distance of the founder HIV-1 strain from the MRKAd5 vaccine insert proteins (Gag and Nef) or the ConsensusB sequence (Gp120) using an HIV-1-specific AA substitution model in the following regions [192]: 1) predicted CTL epitope regions (9mers), 2) extended CTL epitope regions that include the 5 amino acids flanking the epitope on either side (19mer), 3) only the flanking AA residues (10mers) and 4) non-epitopic regions. For each subject, the genetic distance of each individual epitope was summed up and averaged over the number of epitopes for that subject to give a single value per subject. Average subject distances are compared between treatment groups and comparisons with p values <0.05 are considered statistically significant. For trial subjects that were followed over time, a subject

average within predicted epitope regions and non-epitopic regions was estimated for all time points. The slope of genetic distance was calculated for each of these subjects by dividing the change in the genetic distance by time elapsed between sampling. Subject distances between vaccine and placebo treatment groups were compared for sieve analyses. Slope comparisons with p values < 0.05 are considered statistically significant.

Genetic distance of the most prevalent minor variant to vaccine insert

We also estimated the distance of the most prevalent minor variant to the MRKAd5 vaccine insert (Gag, Nef) and to Consensus B sequence (Gp120) for all subjects sampled within 1.5 months post infection. Variants observed between 1-50% of the population were categorized as minor variants. The distance of the most frequent minor variant in all the predicted CTL epitopes was calculated and averaged for each subject. Similarly, the average distance over non-epitopic regions was also estimated and compared between treatment groups. Treatment comparisons with p values <0.05 are considered statistically significant.

Average pairwise genetic distance

Pairwise distance is the average within subject per-epitope genetic distance (Hamming distance). We estimated pairwise distances between all reads observed above the threshold frequency of 1% in predicted CTL epitope regions, and calculated the average pairwise distance across epitopes for each subject. Similarly, average pairwise distances were also estimated for non-epitopic regions. Treatment comparisons with p values <0.05 are considered statistically significant. For subjects followed over time, a subject average within epitope specific regions and non-epitopic regions was estimated for all time points. A pairwise distance slope was calculated by dividing the change in pairwise distances with the time elapsed between sampling

and compared between treatment groups. Slope comparisons with p values <0.05 are considered statistically significant.

Frequency and number of vaccine-like variants

Across all regions in Gag and Nef with predicted or known CTL epitopes, we estimated the fraction of epitope regions with minor variants that were identical to the MRKAd5 insert (termed vaccine-like). In the case of Gp120, the fraction of epitope regions with sequences identical to ConsensusB was estimated. The fraction of identical non-epitopic regions was similarly estimated and compared between treatment groups.

Divergence

Among subjects that were followed up to 20 months post-infection, we estimated the genetic distance of epitope variants to the consensus founder epitope. A subject average for all epitope distances was calculated at each subsequent time point. A subject average was estimated for each time point across epitope or non-epitope regions. A slope of divergence was calculated for these subjects by dividing the change in genetic distance from the founder by time elapsed between sampling. Slopes were compared between treatment groups and results are considered statistically significant for p values <0.05.

Regression models for analysis of epitope evolution over time

We used two approaches to quantify changes within epitope regions over time. In the first approach, we estimated slopes of distance to insert, average pairwise distance and compared slopes between treatment groups. In the second approach, we used regression modeling to model epitope distances to the vaccine insert, average pairwise distances, and divergence from founder variants over time within epitope regions. These values were initially modeled with individual epitope curves using linear mixed models to allow for subject- and epitope- specific variation

around treatment-group curves. For all analyses, we conducted model checks of the residuals using histograms, quantile-quantile (qq) plots, scatterplots versus time, and formal tests of homoscedasticity, and we removed unsupported terms. When random effects models were not supported by the analysis, we applied fixed-effects linear regression. For the analysis of genetic distance, we found success in modeling the consensus epitopes, rather than individual variants. Epitope consensus and pairwise distances measured at time points greater than 10 months post-infection were considered influential and analyses were repeated after removal of these influential points and other data points considered outliers; both p-values are reported. When checks of the final model rejected homoschedasticity[3], we applied the White-Huber ("sandwich")-correction procedure to the results. For comparison of Gag and Nef, their genetic distance values were pooled for additional analysis, with terms included in the model to allow for gene-specific intercepts and slopes. For the measurements of genetic distances of consensus epitopes to the vaccine insert, linear mixed-effects regression models were used to predict an overall slope and intercept for the placebo-recipient subjects' epitope distances to the vaccine insert and to predict treatment effects on both the slope and intercept. This mixed-effects model combines a fixed-effect model of the per-treatment-group slope with a random-effects model for the intercept: individual per-epitope-and-subject intercepts are modeled as normally distributed around the treatment group intercept.

Measurements of divergence were computed for every observed epitope variant relative to the corresponding epitope in the estimated founder sequence, which is here taken to be the epitope's subject-specific consensus computed at the first time available point. In our selected analysis we used fixed-effects linear models with fixed (at zero) intercepts to evaluate the

[3] This describes a situation in which the error term (the noise or random disturbance between independent variables and dependent variables) is the same across all values.

treatment effect on the rate of the per-epitope average divergence from the founder. We also evaluated the subset of the data consisting only of subjects whose infections are established by a single variant (rather than multiple viral variants). Measurements of pairwise diversity were computed for each (subject, epitope) pair as the mean pairwise genetic distance (Hamming distance) among epitope variants, properly weighted by the number of reads per epitope region. For the final analysis we used fixed-effects linear regression models, allowing a non-zero intercept term.

Results

Study Subjects and Characteristics

We sequenced *gag*, *gp41-nef* and *gp120* regions from 23 and 32 unique placebo and vaccine recipients, respectively, using the Roche 454 GS FLX Titanium platform. In addition, proviral DNA and plasma from four subjects (three placebo- and one vaccine-recipient) with less than 50 amplifiable templates were sequenced using the Sanger method. A breakdown of subjects with first time point and longitudinal time point sequences, as well as the number of plasma and PBMC samples, is shown in Table 9. We obtained a median number of 153,000 and 114,000 pyrosequences respectively for first time point and longitudinal samples (Table 10).

	Median number of reads	***P-value***	***Mean amplifiable viral templates***	***P-value***	***Median Sequencing depth***	***P-value***
First time points	153,000	0.47				
gag			2,780	0.55	6.3x	0.42
gp41-nef			2,130	0.43	6.6x	0.84
gp120			2,820	0.68	5.7x	0.36
Longitudinal time-points	114,000	0.53				
gag			650	0.40	7.9x	0.87
gp41-nef			740	0.98	8.2x	0.83
gp120			790	0.96	6.0x	0.39

Table 13. Number of viral templates, pyrosequences, and depth of sequencing. The median number of unfiltered pyrosequencing reads obtained for all three genes in first time-point and longitudinal samples for all study subjects is listed. The mean amplifiable viral templates submitted for sequencing as calculated by PCR endpoint dilution is listed for all three sequenced regions. The sequencing depth was calculated by dividing the median number of reads obtained for each subject by the number of templates sequenced. Differences in median number of reads were compared between treatment groups for subjects sequenced as part of first time point and longitudinal time points, p values in column 3. Similarly, mean amplifiable templates and median sequencing depths for each of the gene regions sequenced were compared between treatment groups, p values in columns 5 and 7. P-values were calculated by Mann-Whitney t-test.

We sequenced a mean of 2780, 2130 and 2820 viral templates for *gag*, *gp41-nef* and *gp120,* respectively, for first time point samples. For longitudinal samples, we sequenced a mean of 650, 740 and 790 viral templates for *gag*, *gp41-nef* and gp120, respectively. There was no significant difference in the number of templates sequenced per subject across treatment groups for first time points or longitudinal time points, Table 10. We obtained a median sequencing depth (number of reads mapping to a position divided by the number of sequenced templates, of 6.3x, 6.6x and 5.7x in *gag*, *gp41-nef* and *gp120,* respectively, for first time points. The median sequencing depth for longitudinal time points was 7.9x, 8.2x and 6.0x for *gag*, *gp41-nef* and *gp120,* respectively. There was no significant difference between the sequencing depth across treatment groups for the first or longitudinal time point samples (Table 10).

We predicted a total of 873, 339 and 2499 potential CTL epitopes using NetMHC [191]in Gag, Nef, and Gp120 respectively (Table 11). The number of predicted unique CTL epitopes was lower, 192, 65 and 551 in Gag, Nef and Gp120, respectively. There was no significant difference in the distribution of predicted epitopes per subject across the treatment groups (Table 11).

	Treatment	Total predicted epitopes (NetMHC)	Unique predicted epitopes (NetMHC)	Mean number of predicted epitopes per subject
Gag	Vaccine	576	102	13
	Placebo	297	90	15
Nef	Vaccine	216	36	6
	Placebo	123	29	6
Gp120	Vaccine	1602	281	39
	Placebo	897	270	42

Table 14. NetMHC predicted CTL epitopes. The number of total and unique CTL epitopes predicted by netMHC is listed for vaccine and placebo treatment groups. Distributions of number of predicted epitopes per subject were compared by Mann-Whitney test (Gag, p=0.23; Nef, p=0.78; Gp120, p=0.80).

We also compared post-infection viral loads for all subjects and found no difference between vaccine and placebo treatment groups (p=0.32). No differences in post-infection viral loads were found between vaccine and placebo groups for subjects that were followed over time (p=0.38).

Distances of consensus epitope to vaccine insert in subjects sampled at the first time point post infection

We calculated the genetic distance of the consensus peptide sequence within predicted CTL epitope regions to the MRKAd5 HIV-1 vaccine insert and found significantly greater distances among vaccinees in Gag (p=0.0002, Figure 18A), as previously observed [81]. In both studies this effect was confined to Gag, with no significant difference in distance to vaccine insert observed within predicted epitope regions in Nef and Gp120 (p=0.49 and p=0.63 respectively, Figure 18B, C). When we measured consensus distances to the vaccine insert in non-epitopic regions, we found no significant difference between treatment groups in any of the three genes (Gag, p=0.59; Nef, p=0.22; Gp120 p=0.76) (Figure 18 D-F).

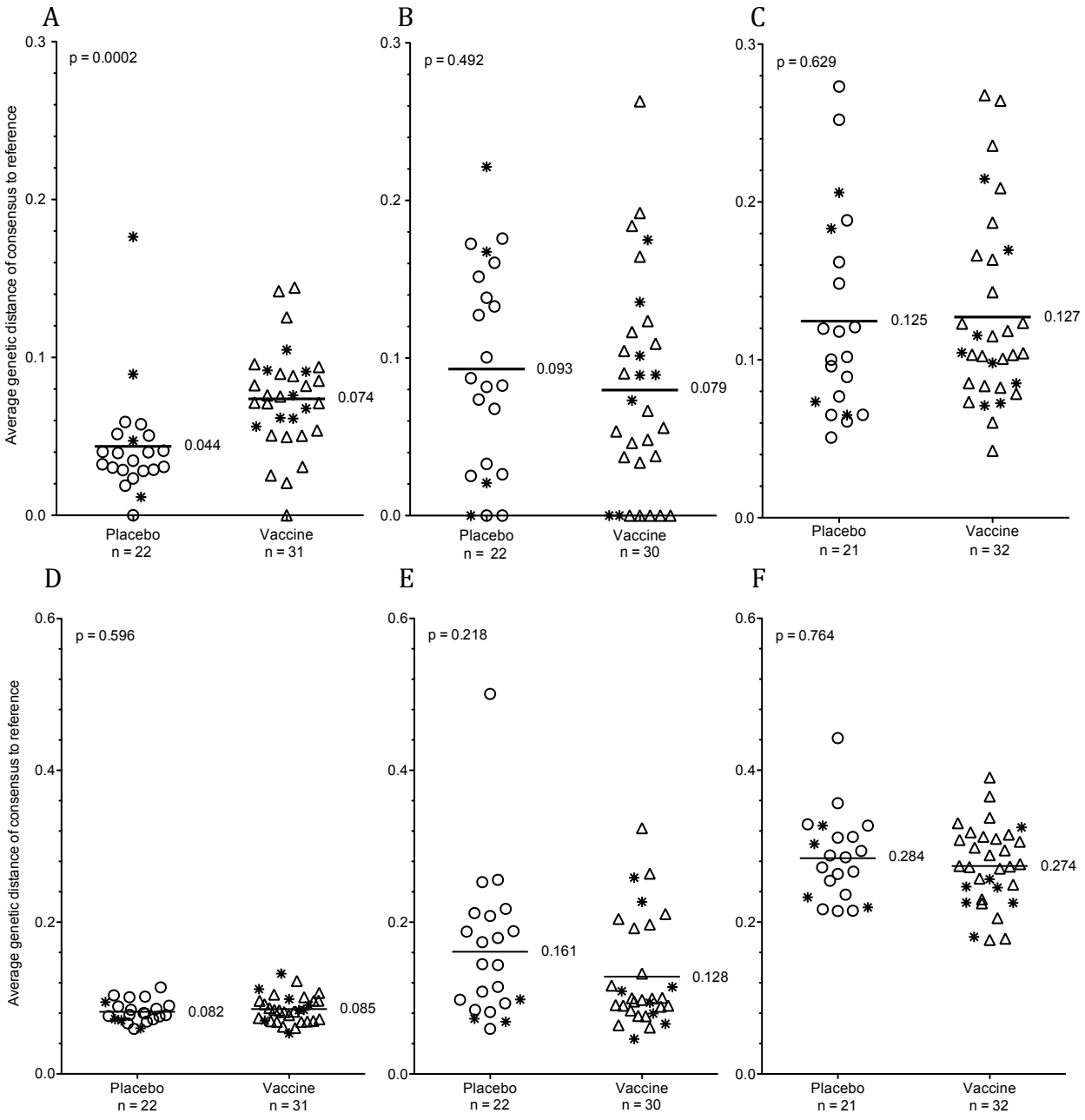

Figure 18. Genetic distances of consensus epitopes to reference sequences in predicted CTL epitopic (A-C) and non-epitopic (D-F) 9mers at first time point post-infection. Genetic distance was estimated using PhyML to the MRKAd5 vaccine insert within predicted CTL epitopes for Gag (A, D) and Nef (B, E). For the control protein Gp120, distance was estimated to HIV-1 CON_B04 (C, F). A subject-specific average across predicted epitopes was calculated and compared between the vaccine and placebo group. Subjects with multiple replicating founder variants are marked with asterisks*. CTL epitopes were predicted using NetMHC and only subjects with predicted or known CTL epitopes are shown. A Mann-Whitney t-test is used to calculate p-values.

Distance of the most prevalent minor variant (at 1% - 50%) to the MRKAd5 vaccine insert within predicted CTL epitopes was estimated and no differences were observed across treatment groups in Gag, Nef and Gp120 (p=0.57, p=1.0, p=0.09 respectively, Figure 19 A-C). A trend of higher mean distances from the vaccine insert within the vaccine group compared to the placebo group (0.25 vs. 0.19) within Gp120 was observed. No significant difference between treatment groups was observed within non-epitopic regions (Gag, p=0.74; Nef, p=0.31; Gp120 p=0.81, Figure 19 D-F). In subjects with multiple replicating founder variants, with the sequences obtained from library pyrosequencing, it is often not possible within CTL epitope regions to assign minor variants to corresponding founder variants, especially when the differences between founder variants is only confined to small number of base differences. We estimated distances after removing the multiple founder subjects and observed similar results, with no difference in distances to vaccine insert observed across treatment groups in either epitopic or non-epitopic regions (predicted epitopes: Gag, p=0.21; Nef, p=0.76; Gp120 p=0.08, Non-epitopic regions: Gag, p=0.96; Nef, p=0.11; Gp120 p=0.50).

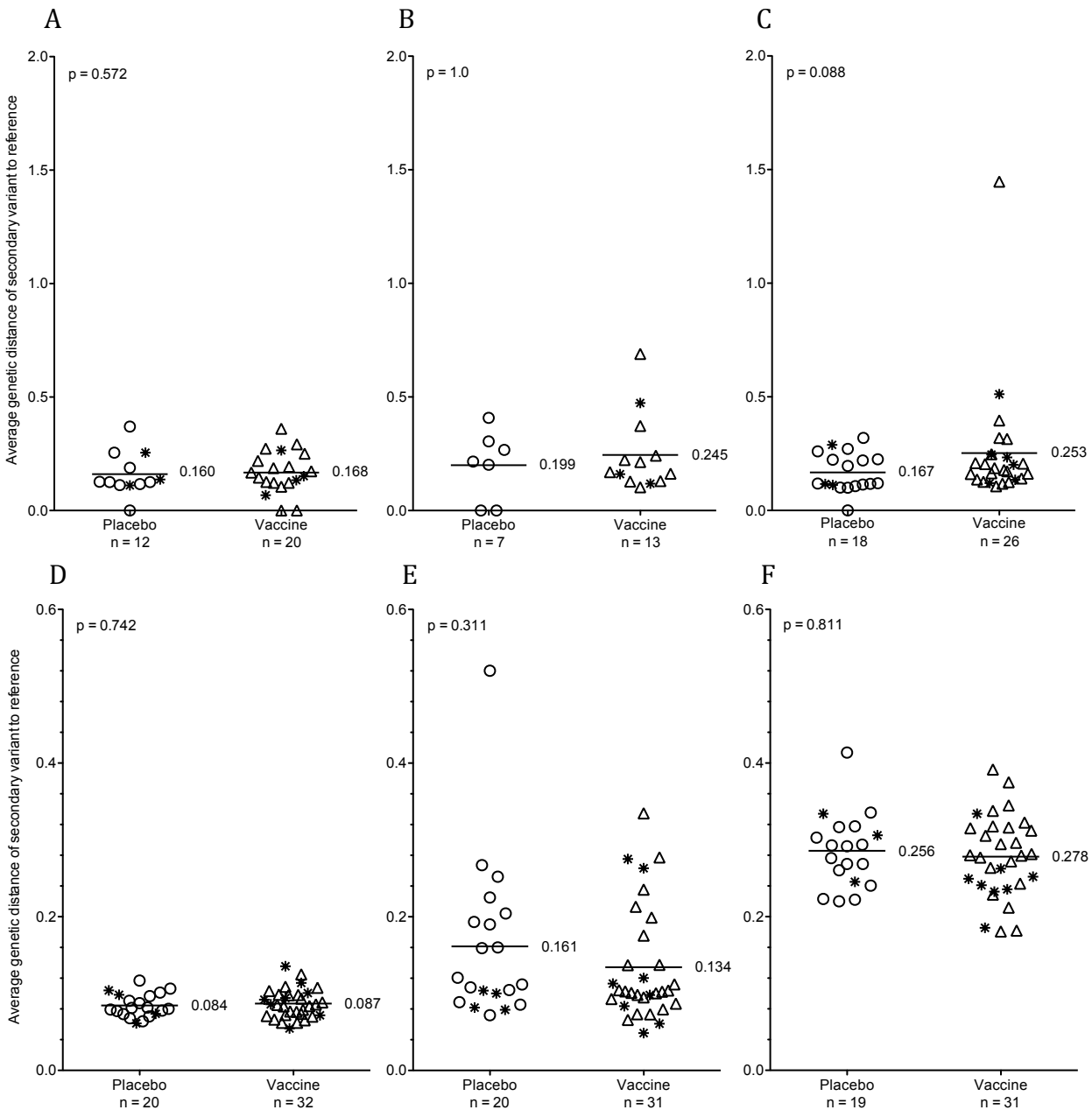

Figure 19. Genetic distance of second-most-frequent variant to reference sequence in predicted CTL epitope and non-epitopic 9mers at first time point post-infection. Genetic distance (estimated with PhyML) of the second-most-frequent variant to the vaccine insert was calculated within predicted CTL epitopes for Gag (A) and Nef (B). For the control protein Gp120, distance to HIV-1 ConsensusB was calculated (C). A subject-specific average across all predicted epitopes was calculated and compared between the vaccine and placebo group. In some cases, predicted epitopes did not have a secondary variant present above the 1% cutoff: these epitopes were not included. Subjects with multiple founder variants are marked with asterisks*. CTL epitopes were predicted using NetMHC. A Mann-Whitney t-test is used to estimate the p-values.

Distances to vaccine insert within epitope flanking regions were estimated and similar results were observed, with vaccine recipients having significantly greater divergence in Gag CTL epitopes (p=0.0003, Figure 20A). No significant differences between treatment groups were observed in Nef and Gp120 (p=0.88 and p=0.43 respectively, Figure 20B, C). No significant differences within 19mer non-epitopic regions were observed in Gag, Nef and Gp120 (Figure 20 D-F) (Gag, p = 0.49; Nef, p = 0.07; Gp120 p = 0.40). Consensus distances to vaccine insert were also estimated for the flanking 5 amino acid regions only (excluding the predicted epitope in the center) and the distances within the vaccine group remained higher in Gag (p=0.006) (Figure 21A) with no significant difference in Nef and Gp120 (p=0.35, p=0.69 respectively, Figure 21B, C).

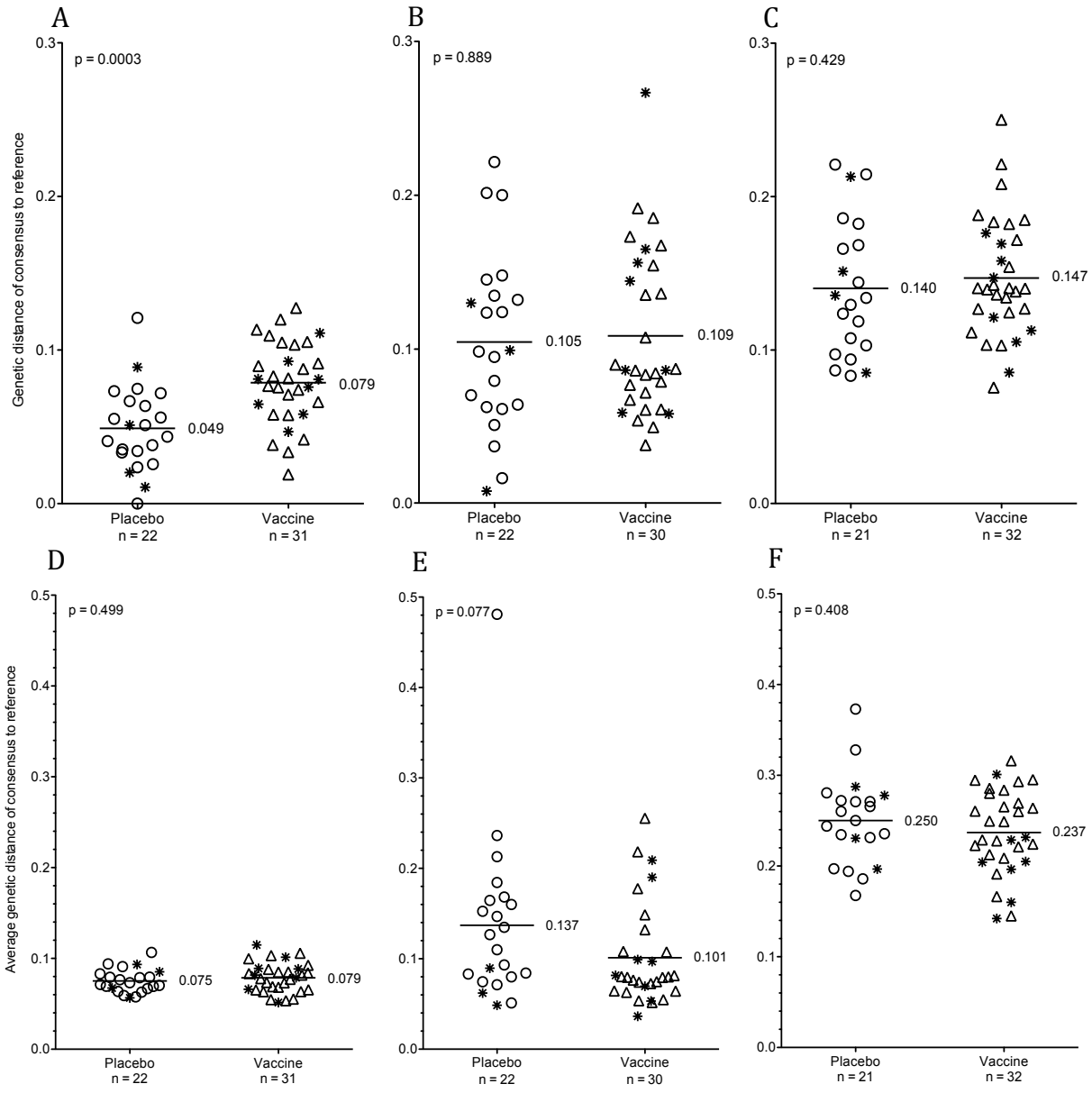

Figure 20. Genetic distance of extended CTL epitopes (19mers) to reference sequence in epitopic (A-C) and non-epitopic regions (D-F) at first time point post-infection. Genetic distance (estimated with PhyML) to the MRKAd5 vaccine insert was calculated in the 19mer regions surrounding predicted CTL epitopes (9mer epitope + 5AA flanks) in Gag (A) and Nef (B). For the control protein Gp120, distance to HIV-1 ConsensusB was calculated (C). A subject-specific average was calculated across all predicted epitopes and compared between the vaccine and placebo groups. Subjects with multiple founder variants are marked with asterisks*. CTL epitope 9mers were predicted using NetMHC. Only subjects with predicted epitopes are shown. A Mann-Whitney t-test is used to estimate the p-values.

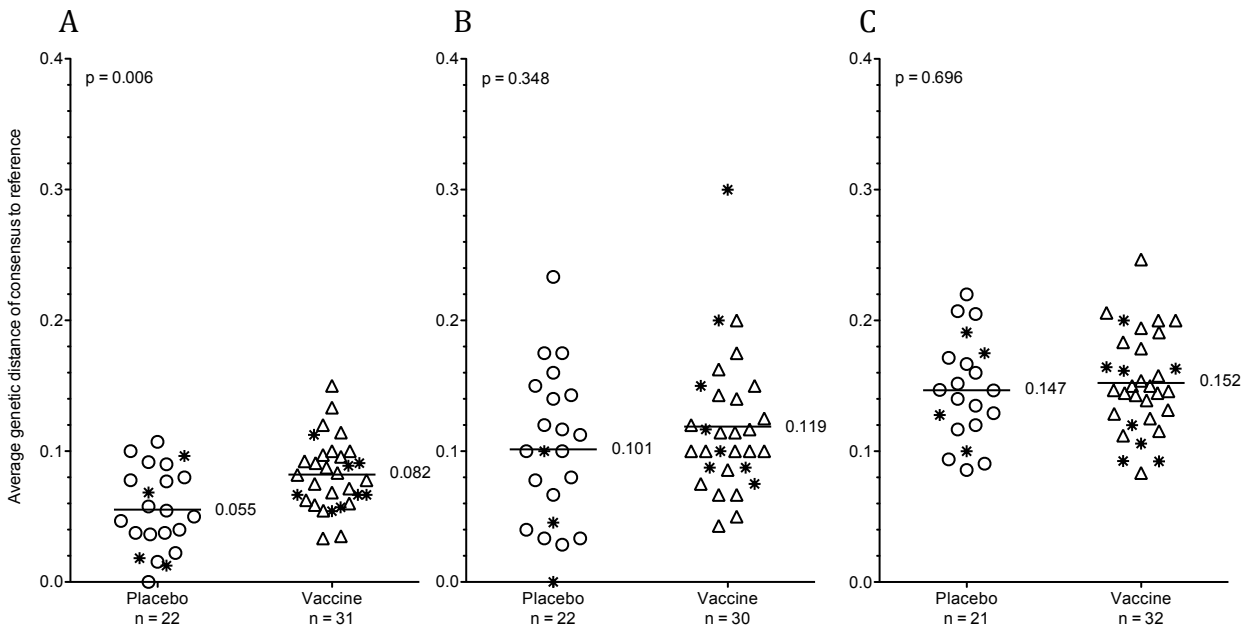

Figure 21. Average genetic distance of predicted epitope flanking regions to reference sequences. Hamming genetic distance to the MRKAd5 vaccine insert in the two 5mer regions flanking predicted CTL epitopes for Gag (A) and Nef (B). For the control protein Gp120, distance was estimated to HIV-1 CON_B04 (C). A subject-specific average across all flanking regions was calculated and compared between the vaccine and placebo group. Subjects with multiple replicating founder variants are marked with asterisks*. A Mann-Whitney t-test is used to calculate p-values.

Average pairwise distances in subjects sampled at the first time point post-infection

One of the hypotheses we tested, as part of the post-infection sieve analysis was whether vaccine-induced anamnestic responses led to increased mutations within CTL epitopes soon after infection. We estimated average pairwise hamming distances of all epitope variants (between 1% - 50% frequency) across predicted CTL epitope regions. We found no significant difference in average pairwise distances within predicted CTL epitope regions (Gag (p=0.37), Nef (p=0.80) or Gp120 (p=0.76)) (Figure 22 A, B, C respectively) or non-epitopic regions (Gag (p=0.2), Nef (p=0.28), Gp120 (p=0.55)).

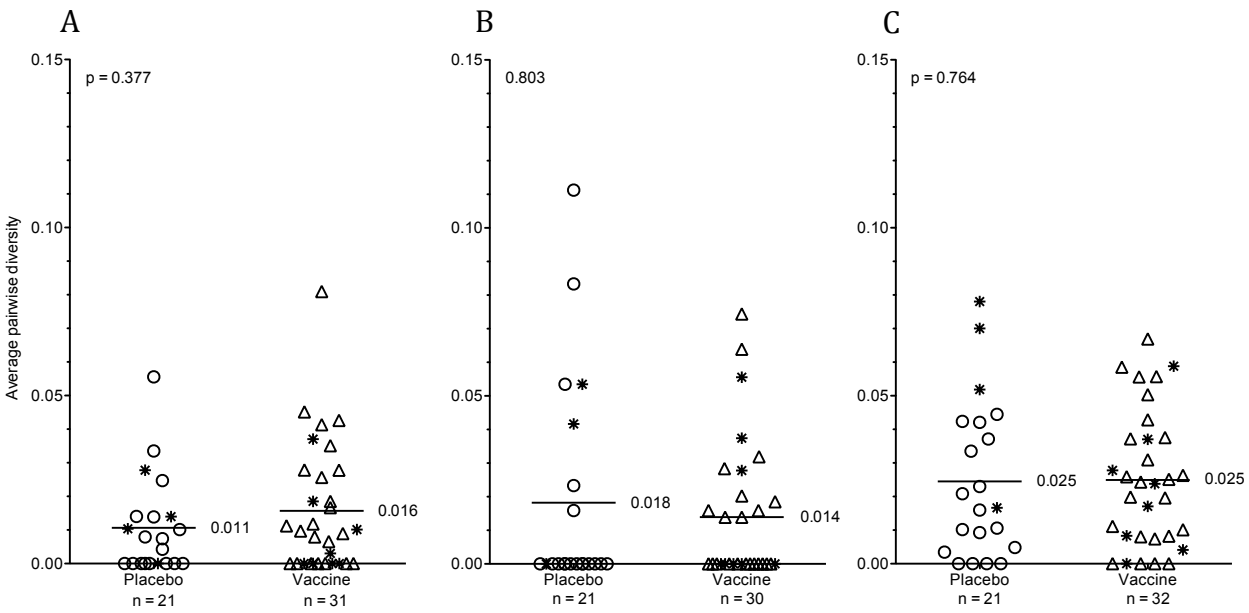

Figure 22. Average pairwise distances within predicted CTL epitope 9mers at the first time point post-infection. The pairwise distance was calculated based on hamming distances for all peptides found >1% frequency in predicted epitopes. A subject-specific average across all predicted epitopes was calculated and compared between the vaccine and placebo groups in Gag (A), Nef (B) and Gp120 (C). Subjects with multiple founder variants are marked with asterisks*. CTL epitopes were predicted using NetMHC. Only subjects with predicted or known CTL epitopes are shown. A Mann-Whitney t-test is used to estimate the p-values.

Average pairwise distances were estimated within extended epitope regions and no differences between treatment groups were observed in Gag, Nef and Gp120 (Gag, p=0.56; Nef, p=0.82; Gp120 p=0.89). Average pairwise distance estimates can be skewed by multiple founder variants, and to accommodate this, we repeated these analyses by removing subjects infected with multiple replicating founders. Again, we found similar results, with no significant difference in pairwise distances across treatment groups in Nef or the control protein Gp120 within predicted CTL epitope regions (Nef: p=0.91, Gp120: p=0.55), but observed a trend towards higher epitope diversity within Gag CTL in the vaccine group (p=0.07), Figure23. In contrast, no significant differences between treatment groups were found in non-epitopic regions

(Gag: p=0.31, Nef: p=0.19, Gp120: p=0.58) when only single founders were included to estimate average pairwise distances.

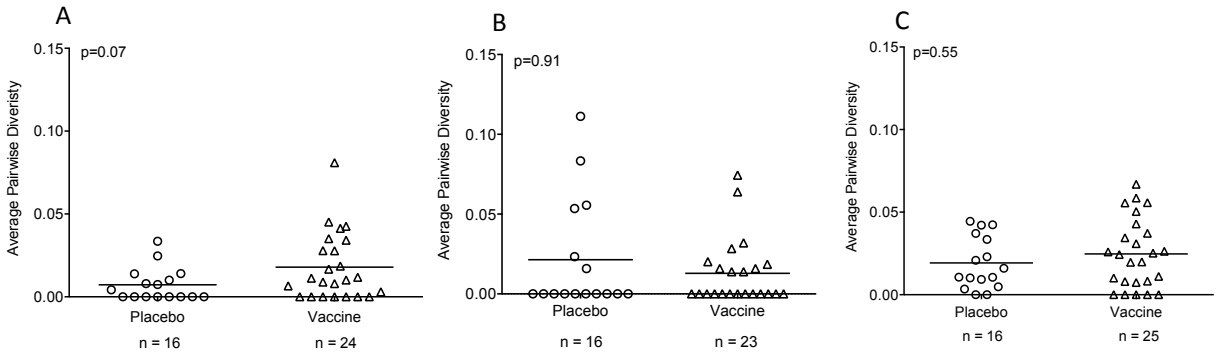

Figure 23. Average pairwise distances within predicted CTL epitope 9mers at the first time point post-infection for subjects infected with single founder variants. The pairwise distance was calculated based on hamming distances for all peptides found >1% frequency in predicted epitopes. A subject-specific average across all predicted epitopes was calculated and compared between the vaccine and placebo groups in Gag (A), Nef (B) and Gp120 (C). Subjects with multiple founder variants are not included in this distance estimates. CTL epitopes were predicted using NetMHC. Only subjects with predicted or known CTL epitopes are shown. A Mann-Whitney t-test is used to estimate the p-values.

Differences in the fraction of minority peptide variants similar to the vaccine insert

Another possible consequence of vaccine-induced immune pressure is rapid selection for variants divergent from the vaccine insert. In this case, we would expect to observe differences in the minor variant populations across treatment groups, particularly, differences in variants genetically similar to the vaccine insert. We found no differences in numbers and frequency of minority variants across all overlapping peptides (9mer AA) in Gag and Nef between vaccine and placebo groups in both predicted CTL epitope regions (Figure 24 A-B) and non-epitopic regions (Figure 24, C-D). Similarly, we found no differences when subjects infected with multiple founders were excluded: predicted epitope regions (Gag: p=0.81, Nef: p=0.43), non-epitopic regions (Gag: p=0.44, Nef p=0.94).

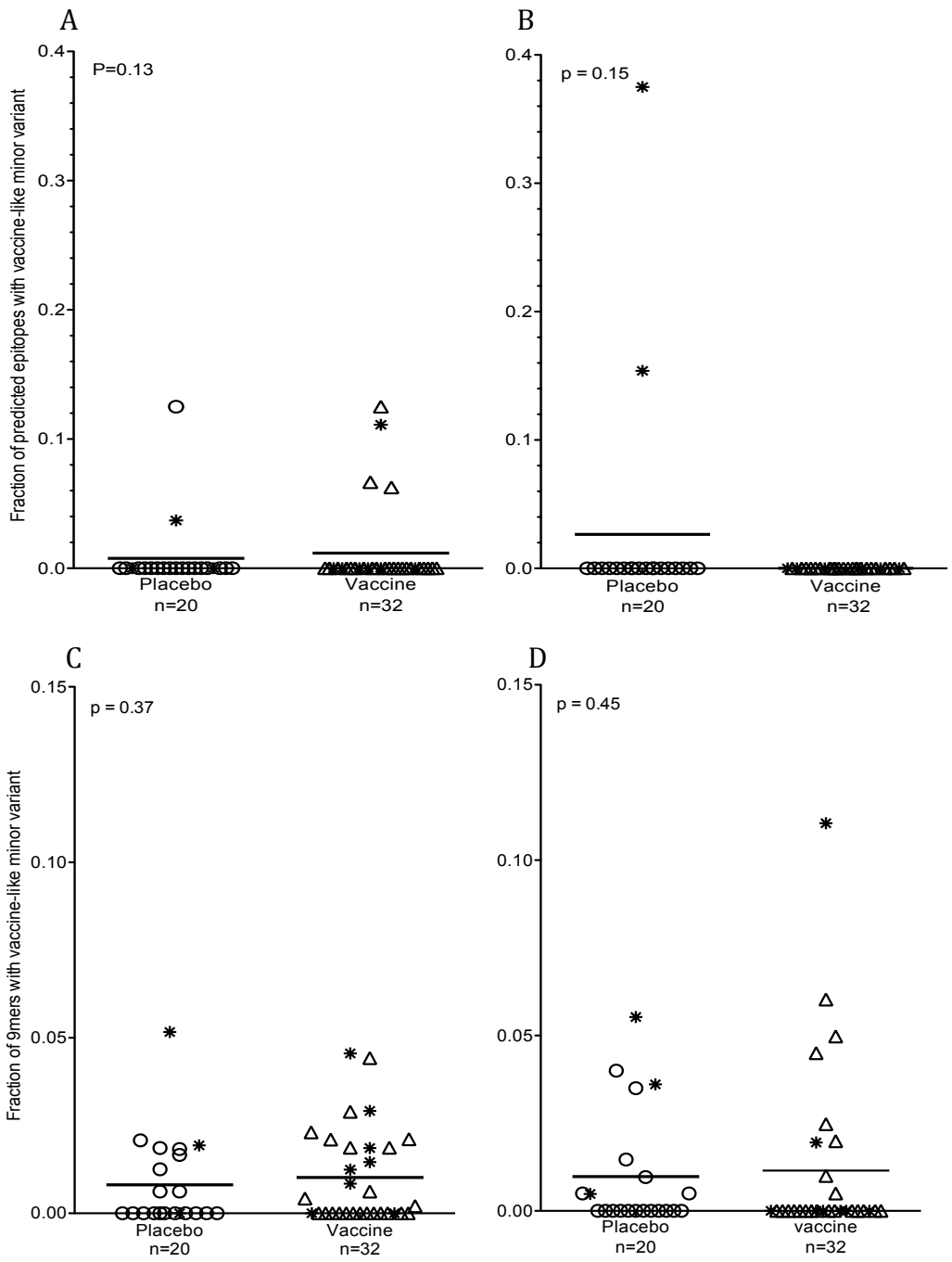

Figure 24. Fraction of CTL predicted epitopes (A,B) and non-epitopic 9mers (C,D) within minority (<50%) variants that match the vaccine insert peptide. Differences in number of minor variants identical to the vaccine insert are shown for Gag (A), and Nef (B). Subjects with multiple founder variants are marked with asterisks*. Epitopes were predicted using NetMHC. Only subjects with known or predicted epitopes were included in analysis. A Mann-Whitney test is used to estimate the p-values.

Impact of vaccine-induced anamnestic responses on distance to vaccine insert, genetic diversity and divergence from founder variant in subjects followed over time

Vaccine-induced anamnestic responses can impact epitope evolution rates, and differences in evolution rates across treatment groups were measured by estimating changes in founder distance to vaccine insert over time (distance slopes) Distance slopes within CTL epitopes and non-epitopic regions were compared across treatment groups (Figure 25 A, B). In the control protein Gp120, founder distances to ConsensusB sequences were compared (Figure 25 C). While the y-intercepts of genetic distances in Gag are higher within the vaccine group, the distance to vaccine insert do not significantly differ between treatment groups over time. Similarly, we observed no significant difference in distance to vaccine inserts over time in non-epitopic regions (Gag: p=0.42, Nef: p=0.34, Gp120: p=0.77). We also modeled genetic distance to vaccine insert using mixed-effects linear regression models to estimate an overall slope and intercept for the epitopes and treatment differences across vaccine and placebo groups. We found a significant treatment effect on the initial genetic distance (y-intercept) of Gag and Nef epitopes combined to the vaccine insert (p=0.0404), with vaccine recipients having greater initial genetic distance. The models also showed a non-significant trend towards treatment effect on the rate at which genetic distance to vaccine insert changes over time (p=0.054), with vaccine recipients trending to faster rates of increase of Genetic distance over time. These treatment effects did not significantly differ between Gag and Nef (p=0.46 intercept, p=0.23 slope). This analysis found no evidence of a treatment effect on the intercept (p=0.4244) or on the slope (p=0.5820) of genetic distance to ConsensusB in Gp120.

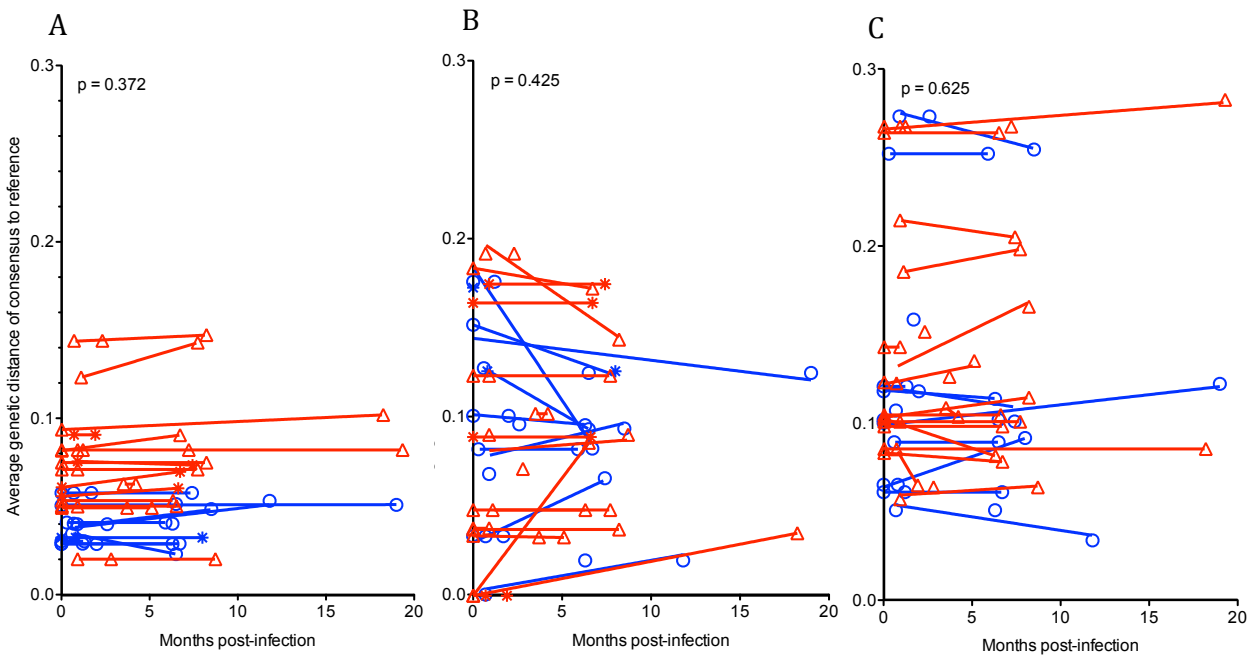

Figure 25. Genetic distance of consensus to reference sequence in predicted CTL 9mer epitopes across longitudinal samples. Genetic distance (estimated by PhyML) to the MRKAd5 vaccine insert was calculated in predicted CTL epitope 9mers for Gag (A) and Nef (B). Distances to HIV-1 ConsensusB were estimated for Gp120 (C). A subject-specific average distance across all predicted epitopes was calculated for each time point. Subjects with multiple founder variants are marked with asterisks*. CTL epitopes were predicted using NetMHC. Vaccine = red triangles; placebo = blue circles. Solid lines indicate the linear regression calculated for a particular patient. The slopes of the linear regressions were compared between the vaccine and placebo groups using a simple comparison, and a Mann-Whitney t-test was used to estimate p-values.

As described earlier, vaccine-induced anamnestic responses pressure did not result in significant differences in viral diversity across treatment groups in subjects classified to be in acute infection. To investigate whether the anamnestic responses influence viral diversity of the founder variants over time, we compared changes in epitope diversity over time (diversity slope) across treatment groups. CTL epitope diversity in Gag within the vaccine groups showed a trend towards reduced diversity over time within epitope variants in the vaccine subjects (p=0.05) (Figure 26A). The change in epitope diversity was not the result of changes in estimated viral loads; no correlation was observed between changes viral load and pairwise distance estimates

(Spearman's rho: -0.18). No significant differences in pairwise genetic distances over time were observed across treatment groups within predicted epitope regions in Nef and the control protein Gp120 (Figure 26 B-C) and within non-epitopic regions (Gag: p=0.45, Nef: p=0.76, Gp120: p=0.89). Changes in epitope diversity were also modeled with linear regression models. In Gag, overall, an effect of the vaccine treatment assignment on the slope but not the intercept was observed. Vaccine recipients showed less rapid increases in average pairwise distance (p=0.01). There was no difference in the pairwise distance intercept between vaccine and placebo groups (p=0.23). Similar results were observed when outliers and influential points (subjects sampled 10 months post infection) were removed from the analysis, with a slower increase in pairwise distances observed in the vaccine group within the Gag region (p=0.04) and no intercept difference observed between treatment groups (p=0.35). In Nef, we observed no significant vaccine treatment effect (intercept p=0.77, slope p=0.27). The models showed no evidence of a treatment effect on the slope (p=0.48) and intercept (p=0.42) in Gp120 between vaccine and placebo recipients.

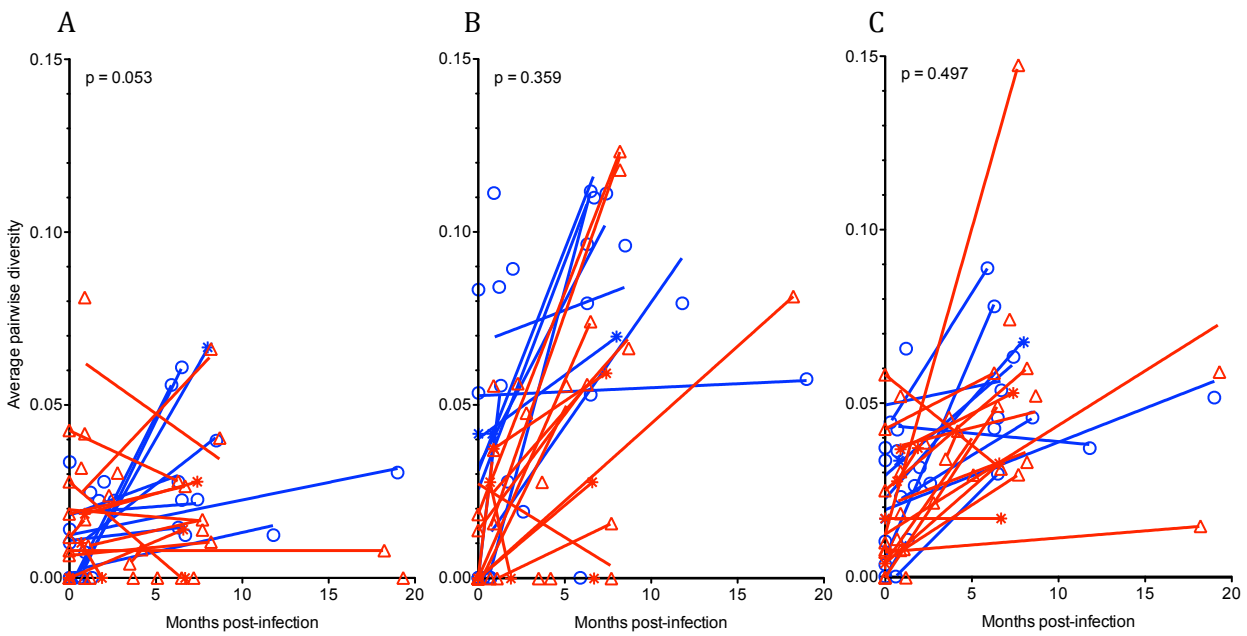

Figure 26. Average pairwise distance in predicted CTL 9mer epitopes across longitudinal samples. The pairwise distances were calculated based on hamming distances for all peptides found >1% frequency in predicted epitopes for Gag (A), Nef (B) and Gp120 (C). A subject-specific average across all predicted epitopes was calculated for each time point. Subjects with multiple founder variants are marked with asterisks*. CTL epitopes were predicted using NetMHC. Vaccine = red triangles; placebo = blue circles. Solid lines indicate the linear regression calculated for a particular patient. The slopes of the linear regressions were compared between the vaccine and placebo groups, and a Mann-Whitney t-test was used to estimate the p-values.

Vaccine induced anamnestic responses could also result in emergence and rapid selection of epitope escape variants that would replace the founder as the major variant. Distance of the consensus variant over time to the founder was estimated as a measure of epitope divergence. Distances to the founder within CTL epitope and non-epitope regions were compared across treatment groups and we observed a trend of lower divergence within CTL epitope regions in Gag in vaccine group (p=0.08), but no significant difference within Nef and Gp120 (Figure 27 B-C). No significant differences across treatment groups were observed within non-epitopic regions (Gag: p=0.34, Nef: p=0.88, Gp120: p=0.24). We also conducted regression analysis to evaluate the change in epitope divergence over time. In the combined Gag and Nef epitope

analysis, the model showed only a trend towards reduced divergence rates among vaccine recipients (p=0.07 when considering only single-founder subjects). When removing influential data points (data collected 10 months post infection) and outliers, the analysis showed significant evidence for lower divergence rates among vaccine recipients (p=0.002 when considering only single-founder subjects). There was no significant treatment effect on the slope in Gp120 (p=0.22).

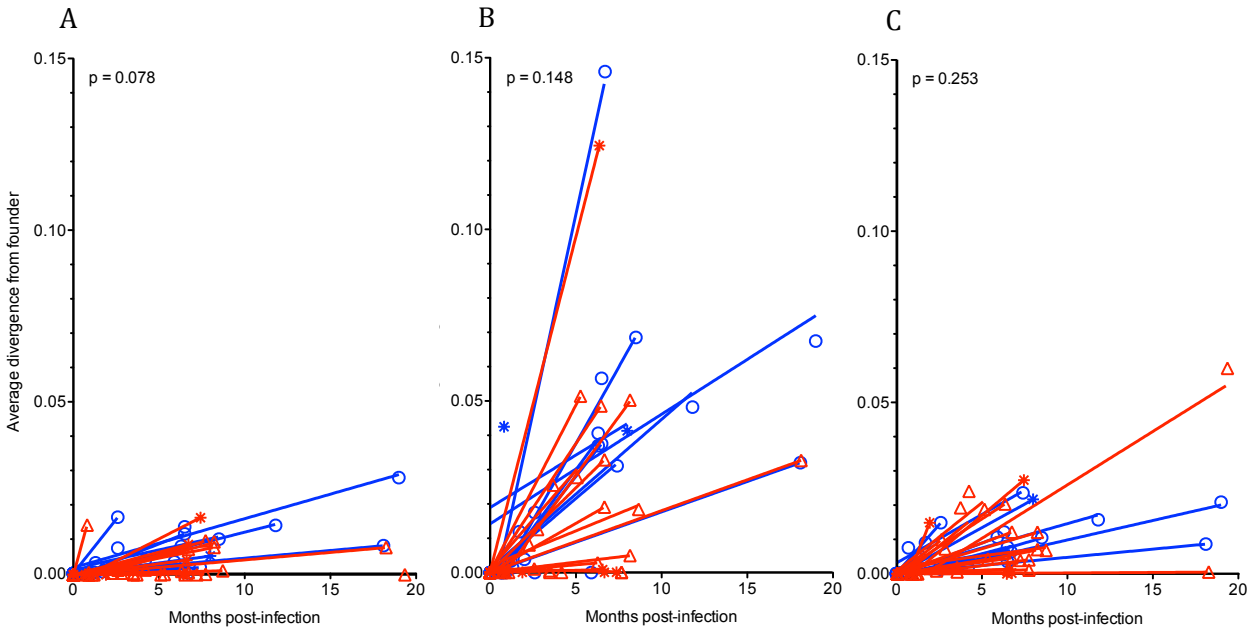

Figure 27. Consensus peptide divergence from founder variants in predicted CTL epitope 9mers over time. Genetic distance (estimated by PhyML) to the founder variant in predicted CTL epitope 9mers was estimated for Gag (A), Nef (B), and Gp120 (C). A subject-specific average distance across all predicted epitopes was calculated for each timepoint. Subjects with multiple founder variants are marked with asterisks*. CTL epitopes were predicted using NetMHC. Vaccine = red triangles; placebo = blue circles. Solid lines indicate the linear regression calculated for a particular patient. The slopes of the linear regressions were compared between the vaccine and placebo groups, and a Mann-Whitney t-test was used to estimate the p-values.

Prior circumcision status and prior Ad5 seropositivity status do not correlate with distance to vaccine insert

Assessments of correlates of HIV-1 infection risk in the STEP trial identified a statistically increased risk of HIV-1 infection in vaccine recipients who were uncircumcised and with Ad5 seropositivity at baseline, relative to controls in the first 18 months of infection [193], with the increased risk waning with time from vaccination. To investigate whether prior Ad5 seropositivity or circumcision status influenced the founder variant, we compared genetic distance of predicted CTL epitopes to vaccine insert within vaccine recipients sampled during acute infection. We found no difference in the distance to vaccine insert between Ad5 seropositive (Ad5 Ab titers >18, the lower limit of detection assay) and Ad5 seronegative (<18 Ad5 Ab titers) vaccine recipients in Gag (Ad5 >=18 mean distance: 0.08, Ad5 <18 mean distance: 0.06, p=0.10) and Nef (Ad5 >=18 mean distance: 0.07, Ad5 <18 mean distance: 0.09, p=0.28). Similarly, we found no significant difference in distances to vaccine insert between circumcised and uncircumcised vaccine recipients in Gag (circumcised mean distance: 0.08, uncircumcised mean distance: 0.07, p=0.08) and Nef (circumcised mean distance: 0.08, uncircumcised mean distance: 0.1, p=0.51). A summary schematic outlining the results is shown below.

A

Subjects sampled at the first time point post infection	Gag (included in the MRKAd5 insert)	Nef (included in the MRKAd5 insert)	Gp120 (Not included in vaccine)
Distance to vaccine insert (or ConsensusB) within breakthrough HIV-1 variants in CTL epitopes	*Higher distance in epitope and flanking regions*	No difference between vaccine and placebo	No difference between vaccine and placebo
Average pairwise distances in CTL epitope variants	No difference between vaccine and placebo	No difference between vaccine and placebo	No difference between vaccine and placebo

B

Subjects sampled up to 20 months post infection	Gag (included in the MRKAd5 insert)	Nef (included in the MRKAd5 insert)	Gp120 (Not included in vaccine)
Change in ***distance*** to vaccine insert (or ConsensusB) within breakthrough HIV-1 variants in CTL epitopes over time	No differences in between vaccine and placebo groups	No differences between vaccine and placebo groups	No differences between vaccine and placebo groups
Change in average pairwise distances in CTL epitope variants over time (***diversity***)	*Significantly slower increase in genetic diversity over time in the vaccine group*	No difference between vaccine and placebo groups	No difference between vaccine and placebo groups
Change in epitope variant distance to founder HIV-1 variant (***divergence***) over time	*Significantly reduced divergence from founder in the vaccine group*	No difference between vaccine and placebo groups	No difference between vaccine and placebo groups

Table 15. Summary of vaccine-induced differences in genetic distance, diversity and divergence within breakthrough HIV-1 variants in CTL epitope regions. A. Differences observed within subjects sampled at the first time point post infection. B. Differences observed within subjects sampled up to a maximum of 20 months post infection.

Discussion

The goal was to study the impact of vaccine induced immune responses on breakthrough HIV-1 sequences from subjects sampled at the earliest time point post-infection and also discern

the effects of vaccine-induced immune responses on breakthrough HIV-1 variants in a subset of subjects that were followed up to a maximum of 20 months post infection. This is a follow-up to our previous study demonstrating the impact of MRKAd5 vaccine on breakthrough viral populations: viruses infecting vaccinees had greater genetic distance to vaccine insert compared to the viruses infecting placebo recipients [81]. This difference was significant within the CTL epitope regions in Gag, which was part of the vaccine insert. One hypothesis to explain this involves the exclusion of certain HIV-1 variants from establishing infection within vaccine recipients. Our earlier study [81] found no direct evidence of selective exclusion: vaccinees were more likely to be infected but single and multiple founders were equally likely to be found within both vaccine and placebo treatment groups. Additionally, earlier phylogenetic analyses showed no clustering of founder viruses according to vaccine and placebo status [81].

Sieve analysis from our current study with independently generated pyrosequences confirms previous findings: genetic distances to vaccine insert within consensus CTL epitopes in Gag are significantly higher among vaccinees. We extended the previous results by comparing vaccine impact on epitope flanking regions. Distances to vaccine insert were significantly higher in Gag within the vaccine group when flanking residues were included. Similarly, higher distances to vaccine insert were observed in Gag in the vaccine recipients when only the CTL epitope flanking residue distances were compared, thus highlighting the importance of these residues in antigen processing and presentation [189,190]. No differences between treatment groups were observed when distances to vaccine insert within regions not overlapping with CTL epitopes were compared. These results suggest the possibility of selective pressure from vaccine induced anamnestic responses contributing to early epitope escape. Anamnestic responses could also result in rapid emergence of multiple divergent variants within vaccinees. We compared

average pairwise distances within CTL epitope regions in Gag and Nef across vaccine and placebo groups and found no differences. Interestingly, when we removed subjects with multiple founder variants and compared average pairwise distances within CTL epitopes, we found a trend (not significant at $p < 0.05$) towards higher diversity in Gag CTL epitopes in the vaccine group. This could signify that the immune responses primed by the vaccine select for accelerated CTL escape within breakthrough founder variants. CTL epitope regions in the control protein Gp120 showed no difference in pairwise distances across treatment groups. Higher CTL epitope diversity and rapid CTL escape as a result of anamnestic responses within the vaccine group could also result in difference in the fraction of minor epitope variants that match the vaccine insert, nonetheless, we found no differences between vaccine and placebo groups in the fraction of minor epitope variants matching the vaccine insert in Gag or Nef.

Early HIV-1 specific CD8+ T cell responses are critical for initial control of viral replication [12,36,194] and there is a strong association between the rate of disease progression and different human leukocyte antigens (HLA) class I alleles [195,196]. During natural course of HIV-1 infection, CTLs control but do not eliminate viremia [197]. While HIV-1 specific CD8+ T cell responses were generated with the MRKAd5 vaccine [37,38,198], the immune responses did not have any effect on viral load within infected study subjects. Furthermore, the vaccine-induced T cell immune responses did not adequately predict distance of breakthrough HIV-1 sequences to vaccine insert within reactive CTL epitopes [199] and the pre-infection vaccine-induced immune responses wane over time [200]. Ideally, to map genetic signatures of vaccine-induced anamnestic responses, we would have used autologous immune responses from HIV-1 infected subjects followed over time, but due to limiting number of subjects from this study with mapped immune responses (7 with Gag and 8 with Nef responses, with a median of 1 immune

response per subject), we used HIV-1 sequences from both vaccine and placebo subjects that were followed over time to identify genetic signatures of vaccine-induced anamnestic responses.

We found no difference in epitope distance to vaccine insert as a function of how long the subject was infected across treatment groups in both the simple analysis comparing the slopes between treatment groups and also from results based on regression modeling. While a simple comparison of diversity slopes between treatment groups did not show significant differences between treatment groups, results based on regression modeling showed a significantly slower rate of diversity within Gag epitopes in vaccine recipients. This observation of reduction in diversity was unexpected and the opposite of our initial hypothesis of vaccine-induced anamnestic pressure influencing increased epitope variation within the vaccine group. A scenario in which prior vaccination resulting in the generation of HIV-1 specific CTLs capable of targeting a fraction of epitope variants would result in limited variation within the epitope regions following HIV-1 infection. A similar reduction in mean pairwise variability was observed with a therapeutic vaccine study in which HIV-1 Nef was delivered by a recombinant vaccinia Ankara vector to chronically infected HIV-1 subjects, and differences in HIV-1 sequences were assessed before and after vaccination [201]. In the current study, epitope variants within Gag in the vaccinees also showed a non-significant trend toward decreased divergence from founder HIV-1 variants over time when slopes were compared between treatment groups. Results from regression modeling showed a significantly reduced rate of divergence in the combined Gag and Nef epitopes within vaccine subjects. Decreased divergence from founder was found to be significant only when outlying subject epitope values and subject samples beyond 10 months post infection were omitted from the regression models. Again, these results were the opposite of what was expected: vaccine-induced anamnestic responses influencing

rapid selection of escape variants. Limited epitope diversity as a result of vaccine-induced CTL responses against a subset of epitope variants could potentially limit the number of escape variants thus limiting divergence from founder variant. Differences between results from a simple comparison of regression slopes and mixed effects regression models could be due to the following 1) inclusion of all epitope values in the regression models compared to a subject average for slope comparisons between treatment groups, 2) combination of Gag and Nef epitopes in regression models compared to individual gene comparisons for the regression slopes and 3) exclusion of outlying epitopes and subject samples 10 months post infection in the regression models compared to inclusion of all values when comparing regression slopes between treatment groups. A comparison of regression slopes across all non-epitopic regions showed no difference between vaccine and placebo groups in distance to insert, variant diversity and divergence from founder over time within Gag, Nef and Gp120 epitopes.

The mechanisms by which vaccine-induced CTL pressure would reduce epitope diversity and divergence over time is not clear from these data. This reduction is epitope diversity however, did not correspond to reduction in subject viral load in the vaccine group and thus the reasons for lack of clinical efficacy of the MRKAd5 vaccine still remains to be elucidated. Recent success with a CMV-based vaccine that elicited broad and even T-cell responses with a mean of 34 epitopes in Gag [202] in the SIV model, highlights the potential importance of breadth in vaccine-induced immune responses, making viral escape more unlikely. In contrast, the MRKAd5 HIV-1 vaccine only elicited a narrow range of CTL responses, most likely leading to rapid generation of escape mutations. A difference of a single amino acid in an epitope (~10% variance) has been shown to eliminate between 30-50% of T-cell recognition [203]. Boosting immune function in early stages of infection, thereby slowing disease progression, is critical to

the success of T-cell based vaccines. However, results from the MRKAd5 vaccine trial and the modestly effective RV144 trial, which was based on a canarypox vector in a prime-boost combination with AIDSVAX B/E, showed no reduction in set-point viral load [37,38,42]. Hansen *et al*. have described promising results in rhesus macaques where the vaccine induced a rapid expansion of effector memory T cells which limited the early stages of SIV replication [68,202,204] and restricted viral loads to undetectable levels in 54% of the animals in the study.

Results from sieve analyses presented in this work is based on CTL epitopes predicted by NetMHC [191], nevertheless it is important to keep in mind that epitope predictors do not capture all the possible peptides that would be presented by the immune system. There has been limited comparison of different epitope prediction methods [205,206], especially studies comparing the overlap between immune responses measured with INF-γ-ELISpot assays with predicted epitopes.

We used short peptide sequences between 9 -19 amino acids in the sieve analysis instead of full-length sequences, as this would have limited the number of full-length sequences available for subsequent sieve analyses. Due to the random shearing of the template DNA during library preparation in the pyrosequencing protocol, the reads generated are staggered across the sequenced region, necessitating the use of shorter peptide fragments for sieve analysis. Multiple rounds of PCR amplification were necessary to increase viral template numbers prior to pyrosequencing, and this can lead to accumulation of mismatch errors within viral templates that would be indistinguishable from real viral sequence variation after pyrosequencing [154,155]. Sequencing errors accumulated during pyrosequencing also influence the total mismatch error rate within viral sequences. As these mismatch errors can affect minor variant resolution, we

used a frequency threshold of 1%, based on a comparison study of multiple fidelity and sensitivity of multiple DNA polymerases [154], prior to sieve analyses.

The results presented in this study are the first to describe evidence for vaccine-induced anamnestic responses impacting genetic diversity and rates of epitope divergence over time within CTL epitope regions in the vaccine group. Although the vaccine-induced selection pressure did not affect post-infection viral load, it is crucial to explore alternate immunogen design strategies in future trials that lead to the generation of immunodominant CD8+ T cell epitope responses and generation of immune responses towards highly conserved regions in the acute phase of infection as these early responses are the most likely to impact viral load set point and subsequent disease progression.

Author acknowledgment

Study design: Shyamala Iyer, Dr. Paul. T. Edlefsen, Dr. Morgane Rolland, Dr. James. I. Mullins; Software implementation: Shyamala Iyer; Data analysis: Shyamala Iyer, Eleanor Casey, Dr. Paul. T. Edlefsen, Craig. A. Margaret, Ted Holzman, Michael Flanigan; Laboratory experiments: Eleanor Casey, Moon Kim, Dylan Westfall, Heather Bouzek, Kim Wong, Hong Zhao, Juila Stoddard-Tepe, Brendan Larsen; Additional software support: Wenjie Deng

CHAPTER 5. Concluding Remarks

As part of my thesis, I developed a pyrosequencing error correction algorithm, CorQ, which automates the various steps involved in processing large numbers of sequences. The CorQ suite of programs was tested and compared against nine pyrosequencing error-correction programs (Table 3, Chapter 2) on sequenced HIV-1 genomes and simulated HIV-1 pyrosequences. CorQ used in combination with other error correction programs reduced pyrosequencing-specific errors by 97%. The performance of CorQ was maximized when combined with the pyrosequencing signal intensity error correction program AmpliconNoise [142,143]. Signal intensity clustering and correction by AmpliconNoise is computationally intensive and has been tested extensively only on data from 454-pyrosequencing, limiting its applicability to correct sequence data generated from Illumina [207] sequencing. While AmpliconNoise error correction algorithms can be used with data from the IonTorrent [208,209] sequencer, testing and optimization have to be performed before the error correction algorithms specific for correcting 454-pyrosequences can be applied to other technologies. The performance of CorQ can be improved by incorporating sequence clustering algorithms in the CorQ analysis pipeline. Sequence clustering algorithms have been widely applied to improve resolution of Operational Taxonomical Units within microbial communities [210-214] and improve the sensitivity of minor variant resolution within viral genomes [145,148,150]. Additionally, incorporating base quality information during sequence clustering might improve the sensitivity and specificity of minor SNPs observed within the sequences.

Performance comparison between multiple pyrosequence error correction programs also highlighted the challenges in identifying and correcting base misincorporation events occurring

during PCR amplification of viral templates. This is an important consideration to keep in mind when the study goal is to identify rare genetic variants, in which case it becomes critical to use enzymes with high fidelity and optimize the amplification conditions to ensure that the genetic variation found within the sequence population is representative of the virus and minimize artifacts derived from the amplification process. Another important feature to consider when designing high-throughput sequencing experiments includes considering not only the number of reads mapping to a genomic location, but also considering the number of viral templates that are input into the sequencing reaction. As highlighted earlier, estimating the amplifiable templates in a sequencing reaction is often not considered when setting frequency thresholds for minor variant detection [128,132,145]. In the study comparing concordance of minor SNP variants between two sequencing technologies (Chapter 3), we used the metric sequencing depth to filter positions with insufficient template coverage, and found that over 50% of the positions with minor SNPs observed within pyrosequences fell in regions with fewer reads than input viral templates. Given that several studies have reported immune escape HIV-1 variants at a level of 0.1% without factoring the number of viral templates in the sequencing reaction [128,132,145], including studies where the number of templates is several fold in excess of the number of sequence reads [139,178], one should proceed with caution when asserting the significance and true population frequencies of these minor HIV-1 variants.

In our study comparing SNPs observed through Sanger and Pyrosequencing, we did not find a clear relationship between increasing sequencing depth with reduction in frequency of minor SNP variants observed within pyrosequences. An ideal comparison to test the effects of increasing sequence reads with respect to viral templates would be to sequence a genomic region with known numbers of viral templates at varying sequencing depths to quantify the advantage

of higher template coverage or sequencing depth in resolving low frequency sequencing artifacts from real minor variants present within the sequence population.

Identifying minor HIV-1 variants has important implications in sieve analyses of viral sequences from infected subjects participating in the MRKAd5 STEP vaccine efficacy trial. Sieve analysis from our study confirmed and expanded previous findings: vaccine recipients showed higher divergence from the vaccine insert within CTL epitopes predicted in Gag. Sieve effects were not observed in peptide regions not overlapping with predicted CTL epitopes and within proteins not included in the vaccine insert. Importantly, we observed evidence for vaccine-induced anamnestic pressure within CTL epitope regions when breakthrough HIV-1 sequences were followed over time. Unexpectedly, pressure from immune responses primed by the vaccine within CTL epitope regions resulted in reduced epitope diversity and rate of epitope evolution. The mechanisms by which vaccine-induced CTL pressure would result in reduced diversity is not entirely clear from these data.

Nonetheless, the reasons for lack of clinical efficacy of the MRKAd5 are still not clear. One of the concerns has to do with the use of the Adenovirus vaccine vector (Ad5) containing HIV-1 clade B *gag*, *pol* and *nef* gene inserts. While this vaccine vector elicited frequent INF-γ ELISpot immune responses (in 77% vaccinees overall), it only generated a limited breadth of antigen-specific responses [38]. While an initial analysis found a vaccine-related enhancement of infection among a subgroup of Ad5 seropositive and uncircumcised subjects [215,216], extended follow-up analyses [193] demonstrated that this risk of acquisition decreased over time. However, the relationship between preexisting Ad5 seropositivity and lack of efficacy of the MRKAd5 vaccine is still unclear. A test-of-concept study using a DNA-rAd5 prime-boost regimen among circumcised subjects lacking preexisting anti-vector neutralizing Abs was halted

when an interim analysis showed that the vaccine regimen did not prevent HIV-1 infection or reduce viral load among vaccine recipients [43]. Recent success with a Cytomegalovirus (CMV) derived vector encoding SIV genes demonstrating the suppression of viral load to undetectable levels after repeated challenges with SIV highlights the importance of broad effector memory T cell responses in clearing SIV reservoirs within latently infected cells [202,204]. Further, this vaccine vector elicited broad and even T-cell responses with higher number of epitopes than that induced by the MRKAd5 HIV-1 vaccine. Adapting this success of a CMV vaccine vector to eradicating HIV-1 infections from latently infected reservoirs has been the recent focus of efforts.

The other area of focus has been to design better immunogens to elicit strong CTL responses against conserved HIV-1 segments. One such strategy, HIV_{CONSV} immunogen includes long fragments of conserved regions within the HIV-1 genome [102]. HIV_{CONSV} refocuses T-cell responses to subdominant epitopes that can provide better viral control. In preclinical studies with non-human primates, this immunogen has been shown to induce polyfunctional T-cells with increased epitope breadth [105,106]. Another strategy is to only include conserved regions in the immunogen [103]. Immune responses generated with DNA vectors expressing these highly conserved elements (CE) were compared with immunization with p^{55gag} DNA. Immunization with p^{55gag} DNA induced poor CD^{4+} mediated cellular responses whereas responses to the CE vectors were reactive across multiple HIV-1 subtypes and comprised of both CD^{4+} and CD^{8+} T cells [107]. Further, DNA vaccination in macaques with the conserved elements vaccine shows increased T-cell breadth of response [108].

Ultimately, for a successful T-cell based vaccine, the correlates of protection, which would define the specificity, breadth, and functional activity of T-cells have to be well defined.

The success of a CMV vector vaccine in clearance of SIV will be an important step toward identifying correlates of protection that can be then applied towards designing a potent T-cell based HIV vaccine that can stimulate CTL responses that are a) both broad and specific for conserved epitopes, b) able to suppress HIV-1 replication *in vitro* and c) sustained for several months following vaccination.

References

1. David M. Knipe PF (2013) Fields Virology: Lippincott Williams and Wilkins.
2. Hemelaar J (2012) The origin and diversity of the HIV-1 pandemic. Trends Mol Med 18: 182-192.
3. Allen G Rodrigo GHL (2001) Computational and Evolutionary Analysis of HIV molecular Sequences: Springer.
4. Dykes C, Balakrishnan M, Planelles V, Zhu Y, Bambara RA, et al. (2004) Identification of a preferred region for recombination and mutation in HIV-1 gag. Virology 326: 262-279.
5. Galli A, Kearney M, Nikolaitchik OA, Yu S, Chin MP, et al. (2010) Patterns of Human Immunodeficiency Virus type 1 recombination ex vivo provide evidence for coadaptation of distant sites, resulting in purifying selection for intersubtype recombinants during replication. J Virol 84: 7651-7661.
6. Jetzt AE, Yu H, Klarmann GJ, Ron Y, Preston BD, et al. (2000) High rate of recombination throughout the human immunodeficiency virus type 1 genome. J Virol 74: 1234-1240.
7. Robertson DL, Sharp PM, McCutchan FE, Hahn BH (1995) Recombination in HIV-1. Nature 374: 124-126.
8. Zhuang J, Jetzt AE, Sun G, Yu H, Klarmann G, et al. (2002) Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. J Virol 76: 11273-11282.
9. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD (1996) HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science 271: 1582-1586.
10. Briones C, Domingo E (2008) Minority report: hidden memory genomes in HIV-1 quasispecies and possible clinical implications. AIDS Rev 10: 93-109.
11. Phillips RE, Rowland-Jones S, Nixon DF, Gotch FM, Edwards JP, et al. (1991) Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition. Nature 354: 453-459.
12. Koup RA (1994) Virus escape from CTL recognition. J Exp Med 180: 779-782.
13. Havlir DV, Richman DD (1996) Viral dynamics of HIV: implications for drug development and therapeutic strategies. Ann Intern Med 124: 984-994.
14. Goulder PJ, Phillips RE, Colbert RA, McAdam S, Ogg G, et al. (1997) Late escape from an immunodominant cytotoxic T-lymphocyte response associated with progression to AIDS. Nat Med 3: 212-217.
15. Price DA, Goulder PJ, Klenerman P, Sewell AK, Easterbrook PJ, et al. (1997) Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. Proc Natl Acad Sci U S A 94: 1890-1895.
16. Richman DD, Wrin T, Little SJ, Petropoulos CJ (2003) Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc Natl Acad Sci U S A 100: 4144-4149.
17. Wei X, Decker JM, Wang S, Hui H, Kappes JC, et al. (2003) Antibody neutralization and escape by HIV-1. Nature 422: 307-312.
18. Frost SD, Wrin T, Smith DM, Kosakovsky Pond SL, Liu Y, et al. (2005) Neutralizing antibody responses drive the evolution of human immunodeficiency virus type 1 envelope during recent HIV infection. Proc Natl Acad Sci U S A 102: 18514-18519.

19. Telenti A (2005) Adaptation, co-evolution, and human susceptibility to HIV-1 infection. Infect Genet Evol 5: 327-334.
20. Burton DR, Poignard P, Stanfield RL, Wilson IA (2012) Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses. Science 337: 183-186.
21. Korber B, Gaschen B, Yusim K, Thakallapally R, Kesmir C, et al. (2001) Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull 58: 19-42.
22. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM (2008) The challenge of HIV-1 subtype diversity. N Engl J Med 358: 1590-1602.
23. Ndung'u T, Weiss RA (2012) On HIV diversity. AIDS 26: 1255-1260.
24. Report UG (2013) UNAIDS report on the global AIDS epidemic 2013.
25. Report I (2013) IAVI Report, The publication on AIDS Vaccine Research.
26. Saunders KO, Rudicell RS, Nabel GJ (2012) The design and evaluation of HIV-1 vaccines. AIDS 26: 1293-1302.
27. Flynn NM, Forthal DN, Harro CD, Judson FN, Mayer KH, et al. (2005) Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. J Infect Dis 191: 654-665.
28. Pitisuttithum P, Gilbert P, Gurwith M, Heyward W, Martin M, et al. (2006) Randomized, double-blind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV-1 vaccine among injection drug users in Bangkok, Thailand. J Infect Dis 194: 1661-1671.
29. Gilbert P, Wang M, Wrin T, Petropoulos C, Gurwith M, et al. (2010) Magnitude and breadth of a nonprotective neutralizing antibody response in an efficacy trial of a candidate HIV-1 gp120 vaccine. J Infect Dis 202: 595-605.
30. Montefiori DC, Karnasuta C, Huang Y, Ahmed H, Gilbert P, et al. (2012) Magnitude and breadth of the neutralizing antibody response in the RV144 and Vax003 HIV-1 vaccine efficacy trials. J Infect Dis 206: 431-441.
31. Carrington M, O'Brien SJ (2003) The influence of HLA genotype on AIDS. Annu Rev Med 54: 535-551.
32. Duerr A, Wasserheit JN, Corey L (2006) HIV vaccines: new frontiers in vaccine development. Clin Infect Dis 43: 500-511.
33. Johnston MI, Fauci AS (2007) An HIV vaccine--evolving concepts. N Engl J Med 356: 2073-2081.
34. Letvin NL, Schmitz JE, Jordan HL, Seth A, Hirsch VM, et al. (1999) Cytotoxic T lymphocytes specific for the simian immunodeficiency virus. Immunol Rev 170: 127-134.
35. Shiver JW, Fu TM, Chen L, Casimiro DR, Davies ME, et al. (2002) Replication-incompetent adenoviral vaccine vector elicits effective anti-immunodeficiency-virus immunity. Nature 415: 331-335.
36. Schmitz JE, Kuroda MJ, Santra S, Sasseville VG, Simon MA, et al. (1999) Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. Science 283: 857-860.
37. Buchbinder SP, Mehrotra DV, Duerr A, Fitzgerald DW, Mogg R, et al. (2008) Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. Lancet 372: 1881-1893.

38. McElrath MJ, De Rosa SC, Moodie Z, Dubey S, Kierstead L, et al. (2008) HIV-1 vaccine-induced immunity in the test-of-concept Step Study: a case-cohort analysis. Lancet 372: 1894-1905.
39. Gray GE, Allen M, Moodie Z, Churchyard G, Bekker LG, et al. (2011) Safety and efficacy of the HVTN 503/Phambili study of a clade-B-based HIV-1 vaccine in South Africa: a double-blind, randomised, placebo-controlled test-of-concept phase 2b study. Lancet Infect Dis 11: 507-515.
40. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, et al. (2009) Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med 361: 2209-2220.
41. Rerks-Ngarm S, Paris RM, Chunsutthiwat S, Premsri N, Namwat C, et al. (2013) Extended evaluation of the virologic, immunologic, and clinical course of volunteers who acquired HIV-1 infection in a phase III vaccine trial of ALVAC-HIV and AIDSVAX B/E. J Infect Dis 207: 1195-1205.
42. Haynes BF, Gilbert PB, McElrath MJ, Zolla-Pazner S, Tomaras GD, et al. (2012) Immune-correlates analysis of an HIV-1 vaccine efficacy trial. N Engl J Med 366: 1275-1286.
43. Hammer SM, Sobieszczyk ME, Janes H, Karuna ST, Mulligan MJ, et al. (2013) Efficacy trial of a DNA/rAd5 HIV-1 preventive vaccine. N Engl J Med 369: 2083-2092.
44. Rolland M, Edlefsen PT, Larsen BB, Tovanabutra S, Sanders-Buell E, et al. (2012) Increased HIV-1 vaccine efficacy against viruses with genetic signatures in Env V2. Nature 490: 417-420.
45. Barouch DH, Liu J, Li H, Maxfield LF, Abbink P, et al. (2012) Vaccine protection against acquisition of neutralization-resistant SIV challenges in rhesus monkeys. Nature 482: 89-93.
46. Letvin NL, Rao SS, Montefiori DC, Seaman MS, Sun Y, et al. (2011) Immune and Genetic Correlates of Vaccine Protection Against Mucosal Infection by SIV in Monkeys. Sci Transl Med 3: 81ra36.
47. Lai L, Kwa S, Kozlowski PA, Montefiori DC, Ferrari G, et al. (2011) Prevention of infection by a granulocyte-macrophage colony-stimulating factor co-expressing DNA/modified vaccinia Ankara simian immunodeficiency virus vaccine. J Infect Dis 204: 164-173.
48. Baba TW, Liska V, Hofmann-Lehmann R, Vlasak J, Xu W, et al. (2000) Human neutralizing monoclonal antibodies of the IgG1 subtype protect against mucosal simian-human immunodeficiency virus infection. Nat Med 6: 200-206.
49. Shibata R, Igarashi T, Haigwood N, Buckler-White A, Ogert R, et al. (1999) Neutralizing antibody directed against the HIV-1 envelope glycoprotein can completely block HIV-1/SIV chimeric virus infections of macaque monkeys. Nat Med 5: 204-210.
50. Mascola JR, Lewis MG, Stiegler G, Harris D, VanCott TC, et al. (1999) Protection of Macaques against pathogenic simian/human immunodeficiency virus 89.6PD by passive transfer of neutralizing antibodies. J Virol 73: 4009-4018.
51. Hessell AJ, Rakasz EG, Poignard P, Hangartner L, Landucci G, et al. (2009) Broadly neutralizing human anti-HIV antibody 2G12 is effective in protection against mucosal SHIV challenge even at low serum neutralizing titers. PLoS Pathog 5: e1000433.
52. Hessell AJ, Rakasz EG, Tehrani DM, Huber M, Weisgrau KL, et al. (2010) Broadly neutralizing monoclonal antibodies 2F5 and 4E10 directed against the human immunodeficiency virus type 1 gp41 membrane-proximal external region protect against

mucosal challenge by simian-human immunodeficiency virus SHIVBa-L. J Virol 84: 1302-1313.
53. Parren PW, Marx PA, Hessell AJ, Luckay A, Harouse J, et al. (2001) Antibody protects macaques against vaginal challenge with a pathogenic R5 simian/human immunodeficiency virus at serum levels giving complete neutralization in vitro. J Virol 75: 8340-8347.
54. Hessell AJ, Poignard P, Hunter M, Hangartner L, Tehrani DM, et al. (2009) Effective, low-titer antibody protection against low-dose repeated mucosal SHIV challenge in macaques. Nat Med 15: 951-954.
55. Burton DR, Hessell AJ, Keele BF, Klasse PJ, Ketas TA, et al. (2011) Limited or no protection by weakly or nonneutralizing antibodies against vaginal SHIV challenge of macaques compared with a strongly neutralizing antibody. Proc Natl Acad Sci U S A 108: 11181-11186.
56. Veazey RS, Shattock RJ, Pope M, Kirijan JC, Jones J, et al. (2003) Prevention of virus transmission to macaque monkeys by a vaginally applied monoclonal antibody to HIV-1 gp120. Nat Med 9: 343-346.
57. Moldt B, Rakasz EG, Schultz N, Chan-Hui PY, Swiderek K, et al. (2012) Highly potent HIV-specific antibody neutralization in vitro translates into effective protection against mucosal SHIV challenge in vivo. Proc Natl Acad Sci U S A 109: 18921-18925.
58. Stephenson KE, Barouch DH (2013) A global approach to HIV-1 vaccine development. Immunol Rev 254: 295-304.
59. Koup RA, Douek DC (2011) Vaccine design for CD8 T lymphocyte responses. Cold Spring Harb Perspect Med 1: a007252.
60. McDermott AB, Koup RA (2012) CD8(+) T cells in preventing HIV infection and disease. AIDS 26: 1281-1292.
61. Deeks SG, Walker BD (2007) Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy. Immunity 27: 406-416.
62. Blackbourn DJ, Mackewicz CE, Barker E, Hunt TK, Herndier B, et al. (1996) Suppression of HIV replication by lymphoid tissue CD8+ cells correlates with the clinical state of HIV-infected individuals. Proc Natl Acad Sci U S A 93: 13125-13130.
63. Frahm N, Kiepiela P, Adams S, Linde CH, Hewitt HS, et al. (2006) Control of human immunodeficiency virus replication by cytotoxic T lymphocytes targeting subdominant epitopes. Nat Immunol 7: 173-178.
64. Hersperger AR, Migueles SA, Betts MR, Connors M (2011) Qualitative features of the HIV-specific CD8+ T-cell response associated with immunologic control. Curr Opin HIV AIDS 6: 169-173.
65. Hersperger AR, Pereyra F, Nason M, Demers K, Sheth P, et al. (2010) Perforin expression directly ex vivo by HIV-specific CD8 T-cells is a correlate of HIV elite control. PLoS Pathog 6: e1000917.
66. Saez-Cirion A, Lacabaratz C, Lambotte O, Versmisse P, Urrutia A, et al. (2007) HIV controllers exhibit potent CD8 T cell capacity to suppress HIV infection ex vivo and peculiar cytotoxic T lymphocyte activation phenotype. Proc Natl Acad Sci U S A 104: 6776-6781.
67. Streeck H, Jolin JS, Qi Y, Yassine-Diab B, Johnson RC, et al. (2009) Human immunodeficiency virus type 1-specific CD8+ T-cell responses during primary infection

are major determinants of the viral set point and loss of CD4+ T cells. J Virol 83: 7641-7648.
68. Hansen SG, Ford JC, Lewis MS, Ventura AB, Hughes CM, et al. (2011) Profound early control of highly pathogenic SIV by an effector memory T-cell vaccine. Nature 473: 523-527.
69. Jost S, Altfeld M (2012) Evasion from NK cell-mediated immune responses by HIV-1. Microbes Infect 14: 904-915.
70. Porichis F, Kaufmann DE (2011) HIV-specific CD4 T cells and immune control of viral replication. Curr Opin HIV AIDS 6: 174-180.
71. Ranasinghe S, Flanders M, Cutler S, Soghoian DZ, Ghebremichael M, et al. (2012) HIV-specific CD4 T cell responses to different viral proteins have discordant associations with viral load and clinical outcome. J Virol 86: 277-283.
72. Soghoian DZ, Jessen H, Flanders M, Sierra-Davidson K, Cutler S, et al. (2012) HIV-specific cytolytic CD4 T cell responses during acute HIV infection predict disease outcome. Sci Transl Med 4: 123ra125.
73. Stephenson KE, Li H, Walker BD, Michael NL, Barouch DH (2012) Gag-specific cellular immunity determines in vitro viral inhibition and in vivo virologic control following simian immunodeficiency virus challenges of vaccinated rhesus monkeys. J Virol 86: 9583-9589.
74. Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, et al. (2011) Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc Natl Acad Sci U S A 108: 11530-11535.
75. Edwards BH, Bansal A, Sabbaj S, Bakari J, Mulligan MJ, et al. (2002) Magnitude of functional CD8+ T-cell responses to the gag protein of human immunodeficiency virus type 1 correlates inversely with viral load in plasma. J Virol 76: 2298-2305.
76. Julg B, Williams KL, Reddy S, Bishop K, Qi Y, et al. (2010) Enhanced anti-HIV functional activity associated with Gag-specific CD8 T-cell responses. J Virol 84: 5540-5549.
77. Kiepiela P, Ngumbela K, Thobakgale C, Ramduth D, Honeyborne I, et al. (2007) CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. Nat Med 13: 46-53.
78. Streeck H, Lichterfeld M, Alter G, Meier A, Teigen N, et al. (2007) Recognition of a defined region within p24 gag by CD8+ T cells during primary human immunodeficiency virus type 1 infection in individuals expressing protective HLA class I alleles. J Virol 81: 7725-7731.
79. Zuniga R, Lucchetti A, Galvan P, Sanchez S, Sanchez C, et al. (2006) Relative dominance of Gag p24-specific cytotoxic T lymphocytes is associated with human immunodeficiency virus control. J Virol 80: 3122-3125.
80. Fukazawa Y, Park H, Cameron MJ, Lefebvre F, Lum R, et al. (2012) Lymph node T cell responses predict the efficacy of live attenuated SIV vaccines. Nat Med 18: 1673-1681.
81. Rolland M, Tovanabutra S, deCamp AC, Frahm N, Gilbert PB, et al. (2011) Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. Nat Med 17: 366-371.
82. Stamatatos L (2012) HIV vaccine design: the neutralizing antibody conundrum. Curr Opin Immunol 24: 316-323.

83. Gorny MK, Stamatatos L, Volsky B, Revesz K, Williams C, et al. (2005) Identification of a new quaternary neutralizing epitope on human immunodeficiency virus type 1 virus particles. J Virol 79: 5232-5237.
84. Pantophlet R, Burton DR (2006) GP120: target for neutralizing HIV-1 antibodies. Annu Rev Immunol 24: 739-769.
85. McElrath MJ, Haynes BF (2010) Induction of immunity to human immunodeficiency virus type-1 by vaccination. Immunity 33: 542-554.
86. Moore PL, Gray ES, Sheward D, Madiga M, Ranchobe N, et al. (2011) Potent and broad neutralization of HIV-1 subtype C by plasma antibodies targeting a quaternary epitope including residues in the V2 loop. J Virol 85: 3128-3141.
87. Moore PL, Gray ES, Wibmer CK, Bhiman JN, Nonyane M, et al. (2012) Evolution of an HIV glycan-dependent broadly neutralizing antibody epitope through immune escape. Nat Med 18: 1688-1692.
88. Phogat S, Wyatt R (2007) Rational modifications of HIV-1 envelope glycoproteins for immunogen design. Curr Pharm Des 13: 213-227.
89. Beddows S, Franti M, Dey AK, Kirschner M, Iyer SP, et al. (2007) A comparative immunogenicity study in rabbits of disulfide-stabilized, proteolytically cleaved, soluble trimeric human immunodeficiency virus type 1 gp140, trimeric cleavage-defective gp140 and monomeric gp120. Virology 360: 329-340.
90. Kovacs JM, Nkolola JP, Peng H, Cheung A, Perry J, et al. (2012) HIV-1 envelope trimer elicits more potent neutralizing antibody responses than monomeric gp120. Proc Natl Acad Sci U S A 109: 12111-12116.
91. Walker LM, Phogat SK, Chan-Hui PY, Wagner D, Phung P, et al. (2009) Broad and potent neutralizing antibodies from an African donor reveal a new HIV-1 vaccine target. Science 326: 285-289.
92. Wu X, Yang ZY, Li Y, Hogerkorp CM, Schief WR, et al. (2010) Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. Science 329: 856-861.
93. Huang J, Ofek G, Laub L, Louder MK, Doria-Rose NA, et al. (2012) Broad and potent neutralization of HIV-1 by a gp41-specific human antibody. Nature 491: 406-412.
94. Mikell I, Sather DN, Kalams SA, Altfeld M, Alter G, et al. (2011) Characteristics of the earliest cross-neutralizing antibody response to HIV-1. PLoS Pathog 7: e1001251.
95. Gray ES, Madiga MC, Hermanus T, Moore PL, Wibmer CK, et al. (2011) The neutralization breadth of HIV-1 develops incrementally over four years and is associated with CD4+ T cell decline and high viral load during acute infection. J Virol 85: 4828-4840.
96. Klein F, Diskin R, Scheid JF, Gaebler C, Mouquet H, et al. (2013) Somatic mutations of the immunoglobulin framework are generally required for broad and potent HIV-1 neutralization. Cell 153: 126-138.
97. Corti D, Langedijk JP, Hinz A, Seaman MS, Vanzetta F, et al. (2010) Analysis of memory B cell responses and isolation of novel monoclonal antibodies with neutralizing breadth from HIV-1-infected individuals. PLoS One 5: e8805.
98. Korber BT, Letvin NL, Haynes BF (2009) T-cell vaccine strategies for human immunodeficiency virus, the virus with a thousand faces. J Virol 83: 8300-8314.
99. Wang YE, Li B, Carlson JM, Streeck H, Gladden AD, et al. (2009) Protective HLA class I alleles that restrict acute-phase CD8+ T-cell responses are associated with viral escape

mutations located in highly conserved regions of human immunodeficiency virus type 1. J Virol 83: 1845-1855.
100. Priddy FH, Brown D, Kublin J, Monahan K, Wright DP, et al. (2008) Safety and immunogenicity of a replication-incompetent adenovirus type 5 HIV-1 clade B gag/pol/nef vaccine in healthy adults. Clin Infect Dis 46: 1769-1781.
101. Altfeld M, Allen TM (2006) Hitting HIV where it hurts: an alternative approach to HIV vaccine design. Trends Immunol 27: 504-510.
102. Letourneau S, Im EJ, Mashishi T, Brereton C, Bridgeman A, et al. (2007) Design and pre-clinical evaluation of a universal HIV-1 vaccine. PLoS One 2: e984.
103. Rolland M, Nickle DC, Mullins JI (2007) HIV-1 group M conserved elements vaccine. PLoS Pathog 3: e157.
104. Yang OO (2009) Candidate vaccine sequences to represent intra- and inter-clade HIV-1 variation. PLoS One 4: e7388.
105. Rosario M, Borthwick N, Stewart-Jones GB, Mbewe-Mvula A, Bridgeman A, et al. (2012) Prime-boost regimens with adjuvanted synthetic long peptides elicit T cells and antibodies to conserved regions of HIV-1 in macaques. AIDS 26: 275-284.
106. Rosario M, Bridgeman A, Quakkelaar ED, Quigley MF, Hill BJ, et al. (2010) Long peptides induce polyfunctional T cells against conserved regions of HIV-1 with superior breadth to single-gene vaccines in macaques. Eur J Immunol 40: 1973-1984.
107. Kulkarni V, Rosati M, Valentin A, Ganneru B, Singh AK, et al. (2013) HIV-1 p24(gag) derived conserved element DNA vaccine increases the breadth of immune response in mice. PLoS One 8: e60245.
108. Kulkarni V, Valentin A, Rosati M, Alicea C, Singh AK, et al. (2014) Altered response hierarchy and increased T-cell breadth upon HIV-1 conserved element DNA vaccination in macaques. PLoS One 9: e86254.
109. Thurmond J, Yoon H, Kuiken C, Yusim K, Perkins S, et al. (2008) Web-based design and evaluation of T-cell vaccine candidates. Bioinformatics 24: 1639-1640.
110. Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, et al. (2007) Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nat Med 13: 100-106.
111. Barouch DH, O'Brien KL, Simmons NL, King SL, Abbink P, et al. (2010) Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. Nat Med 16: 319-323.
112. Santra S, Liao HX, Zhang R, Muldoon M, Watson S, et al. (2010) Mosaic vaccines elicit CD8+ T lymphocyte responses that confer enhanced immune coverage of diverse HIV strains in monkeys. Nat Med 16: 324-328.
113. Stephenson KE, SanMiguel A, Simmons NL, Smith K, Lewis MG, et al. (2012) Full-length HIV-1 immunogens induce greater magnitude and comparable breadth of T lymphocyte responses to conserved HIV-1 regions compared with conserved-region-only HIV-1 immunogens in rhesus monkeys. J Virol 86: 11434-11440.
114. Edlefsen PT, Gilbert PB, Rolland M (2013) Sieve analysis in HIV-1 vaccine efficacy trials. Curr Opin HIV AIDS 8: 432-436.
115. Gilbert PB, Self SG, Ashby MA (1998) Statistical methods for assessing differential vaccine protection against human immunodeficiency virus types. Biometrics 54: 799-814.

116. Gilbert P, Self S, Rao M, Naficy A, Clemens J (2001) Sieve analysis: methods for assessing from vaccine trial data how vaccine efficacy varies with genotypic and phenotypic pathogen variation. J Clin Epidemiol 54: 68-85.
117. Berman PW (1998) Development of bivalent rgp120 vaccines to prevent HIV type 1 infection. AIDS Res Hum Retroviruses 14 Suppl 3: S277-289.
118. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265: 687-695.
119. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.
120. Petrosino JF, Highlander S, Luna RA, Gibbs RA, Versalovic J (2009) Metagenomic pyrosequencing and microbial identification. Clin Chem 55: 856-866.
121. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, et al. (2012) High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol 10: 599-606.
122. Palmer S, Boltz V, Maldarelli F, Kearney M, Halvas EK, et al. (2006) Selection and persistence of non-nucleoside reverse transcriptase inhibitor-resistant HIV-1 in patients starting and stopping non-nucleoside therapy. AIDS 20: 701-710.
123. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res 17: 1195-1201.
124. Bimber BN, Burwitz BJ, O'Connor S, Detmer A, Gostick E, et al. (2009) Ultradeep pyrosequencing detects complex patterns of CD8+ T-lymphocyte escape in simian immunodeficiency virus-infected macaques. J Virol 83: 8247-8253.
125. Varghese V, Shahriar R, Rhee SY, Liu T, Simen BB, et al. (2009) Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors. J Acquir Immune Defic Syndr 52: 309-315.
126. Bimber BN, Dudley DM, Lauck M, Becker EA, Chin EN, et al. (2010) Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing. J Virol 84: 12087-12092.
127. Cooper CJ, Metch B, Dragavon J, Coombs RW, Baden LR (2010) Vaccine-induced HIV seropositivity/reactivity in noninfected HIV vaccine recipients. JAMA 304: 275-283.
128. Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, et al. (2010) Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. PLoS One 5: e12303.
129. Love TM, Thurston SW, Keefer MC, Dewhurst S, Lee HY (2010) Mathematical modeling of ultradeep sequencing data reveals that acute CD8+ T-lymphocyte responses exert strong selective pressure in simian immunodeficiency virus-infected macaques but still fail to clear founder epitope sequences. J Virol 84: 5802-5814.
130. Li JZ, Paredes R, Ribaudo HJ, Svarovskaia ES, Metzner KJ, et al. (2011) Low-frequency HIV-1 drug resistance mutations and risk of NNRTI-based antiretroviral treatment failure: a systematic review and pooled analysis. JAMA 305: 1327-1335.
131. Liu J, Miller MD, Danovich RM, Vandergrift N, Cai F, et al. (2011) Analysis of low-frequency mutations associated with drug resistance to raltegravir before antiretroviral treatment. Antimicrob Agents Chemother 55: 1114-1119.

132. Henn MR, Boutwell CL, Charlebois P, Lennon NJ, Power KA, et al. (2012) Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog 8: e1002529.
133. Gregori J, Esteban JI, Cubero M, Garcia-Cehic D, Perales C, et al. (2013) Ultra-deep pyrosequencing (UDPS) data treatment to study amplicon HCV minor variants. PLoS One 8: e83361.
134. Palmer BA, Dimitrova Z, Skums P, Crosbie O, Kenny-Walsh E, et al. (2014) Analysis of the evolution and structure of a complex intrahost viral population in chronic hepatitis C virus mapped by ultradeep pyrosequencing. J Virol 88: 13709-13721.
135. Park CW, Cho MC, Hwang K, Ko SY, Oh HB, et al. (2014) Comparison of quasispecies diversity of HCV between chronic hepatitis c and hepatocellular carcinoma by Ultradeep pyrosequencing. Biomed Res Int 2014: 853076.
136. Leamon JH, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, et al. (2003) A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. Electrophoresis 24: 3769-3777.
137. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8: R143.
138. Gilles A, Meglecz E, Pech N, Ferreira S, Malausa T, et al. (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. BMC Genomics 12: 245.
139. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A 108: 20166-20171.
140. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, et al. (2012) Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 109: 14508-14513.
141. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 18: 763-770.
142. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, et al. (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nat Methods 6: 639-641.
143. Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12: 38.
144. Eriksson N, Pachter L, Mitsuya Y, Rhee SY, Wang C, et al. (2008) Viral population estimation using pyrosequencing. PLoS Comput Biol 4: e1000074.
145. Archer J, Rambaut A, Taillon BE, Harrigan PR, Lewis M, et al. (2010) The evolutionary analysis of emerging low frequency HIV-1 CXCR4 using variants through time--an ultra-deep approach. PLoS Comput Biol 6: e1001022.
146. Beerenwinkel N, Zagordi O (2011) Ultra-deep sequencing for the analysis of viral populations. Curr Opin Virol 1: 413-418.
147. Prosperi MC, Prosperi L, Bruselles A, Abbate I, Rozera G, et al. (2011) Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. BMC Bioinformatics 12: 5.
148. Salmela L, Schroder J (2011) Correcting errors in short reads by multiple alignments. Bioinformatics 27: 1455-1461.
149. Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N (2011) ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics 12: 119.

150. Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, et al. (2012) Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. PLoS Comput Biol 8: e1002417.
151. Prosperi MC, Salemi M (2012) QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. Bioinformatics 28: 132-133.
152. Herbeck JT, Rolland M, Liu Y, McLaughlin S, McNevin J, et al. (2011) Demographic processes affect HIV-1 evolution in primary infection before the onset of selective processes. J Virol 85: 7523-7534.
153. Balzer S, Malde K, Lanzen A, Sharma A, Jonassen I (2010) Characteristics of 454 pyrosequencing data--enabling realistic simulation with flowsim. Bioinformatics 26: i420-425.
154. Larsen BB, Chen L, Maust BS, Kim M, Zhao H, et al. (2013) Improved detection of rare HIV-1 variants using 454 pyrosequencing. PLoS One 8: e76502.
155. Iyer S, Bouzek H, Deng W, Larsen B, Casey E, et al. (2013) Quality score based identification and correction of pyrosequencing errors. PLoS One 8: e73015.
156. Lee WP, Stromberg MP, Ward A, Stewart C, Garrison EP, et al. (2014) MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. PLoS One 9: e90581.
157. Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. Nucleic Acids Res 18: 1687-1691.
158. Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, et al. (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res 21: 494-504.
159. Quinlan AR, Stewart DA, Stromberg MP, Marth GT (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nat Methods 5: 179-181.
160. Abram ME, Ferris AL, Shao W, Alvord WG, Hughes SH (2010) Nature, position, and frequency of mutations made in a single cycle of HIV-1 replication. J Virol 84: 9864-9878.
161. Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, et al. (1999) Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 73: 10489-10502.
162. Leitner T, Halapi E, Scarlatti G, Rossi P, Albert J, et al. (1993) Analysis of heterogeneous viral populations by direct DNA sequencing. Biotechniques 15: 120-127.
163. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, et al. (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A 105: 7552-7557.
164. Liu SL, Rodrigo AG, Shankarappa R, Learn GH, Hsu L, et al. (1996) HIV quasispecies and resampling. Science 273: 415-416.
165. Rodrigo AG, Goracke PC, Rowhanian K, Mullins JI (1997) Quantitation of target molecules from polymerase chain reaction-based limiting dilution assays. AIDS Res Hum Retroviruses 13: 737-742.
166. Mallona I, Weiss J, Egea-Cortines M (2011) pcrEfficiency: a Web tool for PCR amplification efficiency prediction. BMC Bioinformatics 12: 404.
167. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D (2011) Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS One 6: e28240.

168. Li Y, Chen W, Liu EY, Zhou YH (2013) Single Nucleotide Polymorphism (SNP) Detection and Genotype Calling from Massively Parallel Sequencing (MPS) Data. Stat Biosci 5: 3-25.
169. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26: 1135-1145.
170. Bakker MG, Tu ZJ, Bradeen JM, Kinkel LL (2012) Implications of pyrosequencing error correction for biological data interpretation. PLoS One 7: e44357.
171. Becker EA, Burns CM, Leon EJ, Rajabojan S, Friedman R, et al. (2012) Experimental analysis of sources of error in evolutionary studies based on Roche/454 pyrosequencing of viral genomes. Genome Biol Evol 4: 457-465.
172. Gianella S, Delport W, Pacold ME, Young JA, Choi JY, et al. (2011) Detection of minority resistance during early HIV-1 infection: natural variation and spurious detection rather than transmission and evolution of multiple viral variants. J Virol 85: 8359-8367.
173. Varghese V, Wang E, Babrzadeh F, Bachmann MH, Shahriar R, et al. (2010) Nucleic acid template and the risk of a PCR-Induced HIV-1 drug resistance mutation. PLoS One 5: e10992.
174. Deng W, Maust BS, Westfall DH, Chen L, Zhao H, et al. (2013) Indel and Carryforward Correction (ICC): a new analysis approach for processing 454 pyrosequencing data. Bioinformatics 29: 2402-2409.
175. Liang B, Luo M, Scott-Herridge J, Semeniuk C, Mendoza M, et al. (2011) A comparison of parallel pyrosequencing and sanger clone-based sequencing and its impact on the characterization of the genetic diversity of HIV-1. PLoS One 6: e26745.
176. Recordon-Pinson P, Papuchon J, Reigadas S, Deshpande A, Fleury H (2012) K65R in subtype C HIV-1 isolates from patients failing on a first-line regimen including d4T or AZT: comparison of Sanger and UDP sequencing data. PLoS One 7: e36549.
177. Ji H, Masse N, Tyler S, Liang B, Li Y, et al. (2010) HIV drug resistance surveillance using pooled pyrosequencing. PLoS One 5: e9263.
178. Carlson JM, Schaefer M, Monaco DC, Batorsky R, Claiborne DT, et al. (2014) HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. Science 345: 1254031.
179. Shao W, Boltz VF, Spindler JE, Kearney MF, Maldarelli F, et al. (2013) Analysis of 454 sequencing error rate, error sources, and artifact recombination for detection of Low-frequency drug resistance mutations in HIV-1 DNA. Retrovirology 10: 18.
180. Deng W, Maust BS, Nickle DC, Learn GH, Liu Y, et al. (2010) DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites. Biotechniques 48: 405-408.
181. Gonzalez JM, Portillo MC, Belda-Ferre P, Mira A (2012) Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. PLoS One 7: e29973.
182. Pinto AJ, Raskin L (2012) PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. PLoS One 7: e43093.
183. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434-439.
184. Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol 9: e1003031.

185. Sacha JB, Chung C, Rakasz EG, Spencer SP, Jonas AK, et al. (2007) Gag-specific CD8+ T lymphocytes recognize infected cells before AIDS-virus integration and viral protein expression. J Immunol 178: 2746-2754.
186. Pantaleo G, Koup RA (2004) Correlates of immune protection in HIV-1 infection: what we know, what we don't know, what we should know. Nat Med 10: 806-810.
187. Plotkin SA (2008) Vaccines: correlates of vaccine-induced immunity. Clin Infect Dis 47: 401-409.
188. Plotkin SA (2010) Correlates of protection induced by vaccination. Clin Vaccine Immunol 17: 1055-1065.
189. Milicic A, Price DA, Zimbwa P, Booth BL, Brown HL, et al. (2005) CD8+ T cell epitope-flanking mutations disrupt proteasomal processing of HIV-1 Nef. J Immunol 175: 4618-4626.
190. Le Gall S, Stamegna P, Walker BD (2007) Portable flanking sequences modulate CTL epitope processing. J Clin Invest 117: 3563-3575.
191. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, et al. (2008) NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res 36: W509-512.
192. Nickle DC, Heath L, Jensen MA, Gilbert PB, Mullins JI, et al. (2007) HIV-specific probabilistic models of protein evolution. PLoS One 2: e503.
193. Duerr A, Huang Y, Buchbinder S, Coombs RW, Sanchez J, et al. (2012) Extended follow-up confirms early vaccine-enhanced risk of HIV acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus HIV vaccine (Step Study). J Infect Dis 206: 258-266.
194. Borrow P, Lewicki H, Hahn BH, Shaw GM, Oldstone MB (1994) Virus-specific CD8+ cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. J Virol 68: 6103-6110.
195. Kawashima Y, Pfafferott K, Frater J, Matthews P, Payne R, et al. (2009) Adaptation of HIV-1 to human leukocyte antigen class I. Nature 458: 641-645.
196. Streeck H, Nixon DF (2010) T cell immunity in acute HIV-1 infection. J Infect Dis 202 Suppl 2: S302-308.
197. McMichael AJ, Borrow P, Tomaras GD, Goonetilleke N, Haynes BF (2010) The immune response during acute HIV-1 infection: clues for vaccine development. Nat Rev Immunol 10: 11-23.
198. Fitzgerald DW, Janes H, Robertson M, Coombs R, Frank I, et al. (2011) An Ad5-vectored HIV-1 vaccine elicits cell-mediated immunity but does not affect disease progression in HIV-1-infected male subjects: results from a randomized placebo-controlled trial (the Step study). J Infect Dis 203: 765-772.
199. Janes H, Frahm N, DeCamp A, Rolland M, Gabriel E, et al. (2012) MRKAd5 HIV-1 Gag/Pol/Nef vaccine-induced T-cell responses inadequately predict distance of breakthrough HIV-1 sequences to the vaccine or viral load. PLoS One 7: e43396.
200. Janes H, Friedrich DP, Krambrink A, Smith RJ, Kallas EG, et al. (2013) Vaccine-induced gag-specific T cells are associated with reduced viremia after HIV-1 infection. J Infect Dis 208: 1231-1239.
201. Hoffmann D, Seebach J, Cosma A, Goebel FD, Strimmer K, et al. (2008) Therapeutic vaccination reduces HIV sequence variability. FASEB J 22: 437-444.

202. Hansen SG, Sacha JB, Hughes CM, Ford JC, Burwitz BJ, et al. (2013) Cytomegalovirus vectors violate CD8+ T cell epitope recognition paradigms. Science 340: 1237874.
203. Lee JK, Stewart-Jones G, Dong T, Harlos K, Di Gleria K, et al. (2004) T cell cross-reactivity and conformational changes during TCR engagement. J Exp Med 200: 1455-1466.
204. Hansen SG, Piatak M, Jr., Ventura AB, Hughes CM, Gilbride RM, et al. (2013) Immune clearance of highly pathogenic SIV infection. Nature 502: 100-104.
205. Lafuente EM, Reche PA (2009) Prediction of MHC-peptide binding: a systematic and comprehensive overview. Curr Pharm Des 15: 3209-3220.
206. Liao WW, Arthur JW (2011) Predicting peptide binding to Major Histocompatibility Complex molecules. Autoimmun Rev 10: 469-473.
207. Willerth SM, Pedro HA, Pachter L, Humeau LM, Arkin AP, et al. (2010) Development of a low bias method for characterizing viral populations using next generation sequencing technology. PLoS One 5: e13564.
208. Chang MW, Oliveira G, Yuan J, Okulicz JF, Levy S, et al. (2013) Rapid deep sequencing of patient-derived HIV with ion semiconductor technology. J Virol Methods 189: 232-234.
209. Gibson RM, Meyer AM, Winner D, Archer J, Feyertag F, et al. (2014) Sensitive deep-sequencing-based HIV-1 genotyping assay to simultaneously determine susceptibility to protease, reverse transcriptase, integrase, and maturation inhibitors, as well as HIV-1 coreceptor tropism. Antimicrob Agents Chemother 58: 2167-2185.
210. Weinstock GM (2012) Genomic approaches to studying the human microbiota. Nature 489: 250-256.
211. Bertelli C, Greub G (2013) Rapid bacterial genome sequencing: methods and applications in clinical microbiology. Clin Microbiol Infect 19: 803-813.
212. Segata N, Boernigen D, Tickle TL, Morgan XC, Garrett WS, et al. (2013) Computational meta'omics for microbial community studies. Mol Syst Biol 9: 666.
213. Nikolaki S, Tsiamis G (2013) Microbial diversity in the era of omic technologies. Biomed Res Int 2013: 958719.
214. Desai N, Antonopoulos D, Gilbert JA, Glass EM, Meyer F (2012) From genomics to metagenomics. Curr Opin Biotechnol 23: 72-76.
215. Gray G, Buchbinder S, Duerr A (2010) Overview of STEP and Phambili trial results: two phase IIb test-of-concept studies investigating the efficacy of MRK adenovirus type 5 gag/pol/nef subtype B HIV vaccine. Curr Opin HIV AIDS 5: 357-361.
216. D'Souza MP, Frahm N (2010) Adenovirus 5 serotype vector-specific immunity and HIV-1 infection: a tale of T cells and antibodies. AIDS 24: 803-809.