UNIVERSITY OF CALIFORNIA, SAN DIEGO

Detecting and Characterizing HIV-1 Intraclade Dual Infection

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

in

Bioinformatics

by

Mary Elizabeth Pacold

Committee in charge:

Professor Douglas Richman, Chair
Professor Glenn Tesler, Co-Chair
Professor Davey Smith
Professor Robert Schooley
Professor Shankar Subramaniam

2010

The dissertation of Mary Elizabeth Pacold is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2010

DEDICATION

To my parents, with gratitude.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to sincerely thank my advisors, Profs. Doug Richman and Davey Smith, for their mentorship and the many opportunities they have provided me. The guidance of my other committee members, Profs. Chip Schooley, Shankar Subramaniam, and Glenn Tesler, has also been invaluable. I am indebted to Prof. Sergei Kosakovsky Pond for lending his expertise at many points along the way and to Prof. Susan Little for the use of her study cohort.

I am extremely grateful for the help of my friends and colleagues in the Translational Virology Core: Karole Ignacio, Parris Jordan, Stephen Espitia, George Hightower, Josue Perez-Santiago, Sanjay Mehta, Millie Vargas, Pok Cheng, Sara Gianella Weibel, Gabe Wagner, and Winston Tilghman. Thanks also to my past and present labmates David Butler, Lalo Cachay, Sherry Rostami, Nancy Keating, Deya Collier, Steffney Rought, Pinyi Du, David Looney, Paula Soto, Topher Woelk, Rick Mitchell, Laura Martinez, and Dr. Zhang from the fourth floor, as well as past and present members of the Antiviral Research Center Simon Frost, Art Poon, Selene Zarate, Wayne Delport, Anya Umlauf, and Gina Osorio.

I would like to thank my parents, my siblings Mike, Martha, Christine, Joe, and Luke, my aunt Giok, and my boyfriend Levon for their inestimable support. I also thank my classmates in the bioinformatics program, especially Merril Gersten, Jason Chan, Harish Nagarajan, and Vipul Bharghava, and my friends, especially Andrea Kmicikewycz, Marisol Chang, Patrick Verkaik, Gjergji Zyba, Didem Unat, and Panos Voulgaris.

Finally, I have to thank my elderly neighbor Mrs. Micks for asserting that "A Ph.D. is the kind of doctor who doesn't do anything!"

Chapter 2 is, in part, a reprint of the paper "Butler DM, Pacold ME, Jordan PS, Richman DD, Smith DM. The efficiency of single genome amplification and sequencing is improved by quantitation and use of a bioinformatics tool. *J Virol Methods*, 2009 Dec;162(1-2):280-3." The dissertation author was the second author and investigator of this paper.

Chapter 3 is, in part, a reprint of the paper "Pacold M, Smith D, Little S, Cheng PM, Jordan P, Ignacio C, Richman D, Pond SK. Comparison of Methods to Detect HIV Dual Infection. *AIDS Res Hum Retroviruses*, 2010 Oct 18. [Epub ahead of print]." The dissertation author was the primary investigator and author of this paper.

Chapter 4 is, in part, a reprint of the manuscript in preparation "Pacold ME, Pond SK, Wagner GA, Delport W, Bourque DL, Richman DD, Little SJ, Smith DM. Clinical, Virologic, and Immunologic Correlates of HIV-1 Dual Infection." The dissertation author was the primary investigator and author of this paper.

VITA AND PUBLICATIONS

2004	B.S. in Computer Science, University of Illinois at Urbana-Champaign
2010	Ph.D. in Bioinformatics, University of California, San Diego

Pacold ME, Pond SK, Wagner GA, Delport W, Bourque DL, Richman DD, Little SJ, Smith DM. Clinical, Virologic, and Immunologic Correlates of HIV-1 Dual Infection. Manuscript in preparation.

Pacold M, Smith D, Little S, Cheng PM, Jordan P, Ignacio C, Richman D, Pond SK. Comparison of Methods to Detect HIV Dual Infection. *AIDS Res Hum Retroviruses*, 2010 Oct 18. [Epub ahead of print].

Manosuthi W, Butler DM, Perez-Santiago J, Poon AF, Pillai SK, Mehta SR, Pacold ME, Richman DD, Pond SK, Smith DM. Protease polymorphisms in HIV-1 subtype CRF01_AE represent selection by antiretroviral therapy and host immune pressure. *AIDS*, 2010 Jan 28;24(3):411-6.

Butler DM, Pacold ME, Jordan PS, Richman DD, Smith DM. The efficiency of single genome amplification and sequencing is improved by quantitation and use of a bioinformatics tool. *J Virol Methods*, 2009 Dec;162(1-2):280-3.

Smith DM, May SJ, Tweeten S, Drumright L, Pacold ME, Kosakovsky Pond SL, Pesano RL, Lie YS, Richman DD, Frost SD, Woelk CH, Little SJ. A public health model for the molecular surveillance of HIV transmission in San Diego, California. *AIDS*, 2009 Jan 14;23(2):225-32.

Cachay ER, Frost SD, Poon AF, Looney D, Rostami SM, Pacold ME, Richman DD, Little SJ, Smith DM. Herpes simplex virus type 2 acquisition during recent HIV infection does not influence plasma HIV levels. *J Acquir Immune Defic Syndr*, 2008 Apr 15;47(5):592-6.

Tseng TT, McMahon AM, Johnson VT, Mangubat EZ, Zahm RJ, Pacold ME, Jakobsson E. Sodium channel auxiliary subunits. *J Mol Microbiol Biotechnol*, 2007;12(3-4):249-62. Review.

Tseng TT, McMahon AM, Zahm RJ, Pacold ME, Jakobsson E. Calcium channel auxiliary subunits. *J Mol Microbiol Biotechnol*, 2006;11(6):326-44. Review.

ABSTRACT OF THE DISSERTATION

Detecting and Characterizing HIV-1 Intraclade Dual Infection

by

Mary Elizabeth Pacold

Doctor of Philosophy in Bioinformatics

University of California, San Diego, 2010

Professor Douglas Richman, Chair
Professor Glenn Tesler, Co-Chair

HIV-1 dual infection occurs when the same individual is infected by two different strains of HIV-1. Although the high prevalence of circulating and unique recombinant forms of HIV-1, which can only be generated in the setting of dual infection, indicates that dual infection is not rare, relatively few cases of it have been documented. Hypothesizing that existing methods to detect dual infection were insufficiently sensitive, we developed and validated a novel technique utilizing 454 ultra-deep sequencing to detect low minority viral populations. We then applied this method to determine the prevalence of dual infection in a well-characterized study cohort and investigated the identified dual infection cases for clinical, virologic, and immunologic correlates of dual infection. At approximately 40% of the cost and 20% of the laboratory and analysis time of previous methods, ultra-deep sequencing identified three times more dual infection cases than pre-

viously documented in this study cohort. Dually infected participants had significantly faster viral load increases than monoinfected controls and displayed multiple patterns of viral recombination and CTL escape.

Chapter 1

Introduction

1.1 HIV Dual Infection

In a small but measurable minority of HIV-infected individuals, concurrent (coinfection) or subsequent (superinfection) infections with different HIV-1 strains establish productively replicating viral populations [67]. These instances of dual infection (DI) are characterized by molecular evidence of two or more viral subpopulations that are too divergent to be explained by typical within-host HIV-1 evolution from a single founder strain. Coinfection (CI) refers to infection of the second strain before seroconversion occurs in response to the first strain, while superinfection (SI) refers to infection of the second strain after seroconversion (Fig. 1.1). HIV-1 superinfection is of particular interest because, while not precisely reflecting primary infection following vaccination, it offers a unique opportunity to investigate the (inadequate) immunologic protection conferred by the first infection.

Another significant attribute of HIV-1 DI is its contribution to global HIV-1 genetic diversity. Recombination is a key evolutionary strategy HIV-1 employs to expand the diversity of progeny viruses [13], and this diversity can be greatly augmented in the presence of DI. In each replication cycle, reverse transcriptase switches RNA templates an average of 2.8 times [83], thus producing recombinant variants within the host. If they are sufficiently fit, these mosaic viruses may then persist within the host and, if transmitted, within the population. The large number of circulating recombinant forms (48 in the established nomenclature[1]) provides strong circumstantial evidence that DI

[1] http://www.hiv.lanl.gov/, accessed November 2010

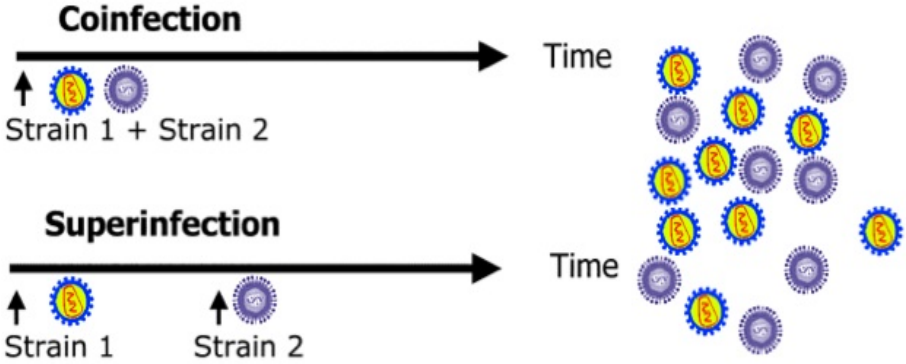

Figure 1.1: Types of dual infection.

with different clades is not rare and that the mosaic variants it produces can frequently persist at the population level. In fact, one study found that a high-risk group of bar workers in Tanzania had a higher prevalence of both dual infection and recombinant strains than a normal-risk control population [31]. According to mathematical models, a limited prevalence of DI can still result a high prevalence of circulating recombinant forms if a small, high-risk population is linked with a large, low-risk one [27].

Dual infection may be termed interclade, when the two infecting strains come from different clades (subtypes), or intraclade, when the strains come from the same clade. Clearly, interclade DI is easier to detect than intraclade DI, because of the large genetic differences (up to 30% in the envelope (*env*) gene [55]) between viral clades. Intraclade DI is likely more frequent than interclade DI, because of the usual predominance of a single viral subtype in a population or geographic area.

HIV-1 SI was first demonstrated in a chimpanzee model in 1987 [20]. The first cases of HIV-1 DI in humans were documented in 1994 and 1995 [3, 16, 62, 78, 82]. Questions remained about the individual and global effects of HIV-1 SI, and incidence studies were undertaken.

1.2 Incidence of Superinfection

The majority of DI screening methods described in the literature involve sequence analyses of one or a few HIV-1 coding regions to determine if phylogenetically distinct viral populations are present. Coding regions have been sampled using population-based sequencing of HIV RNA [69] or DNA [25] populations, or clonal and single genome

sequencing of HIV RNA [63] or DNA [5] populations. Table 1.1 summarizes studies of HIV-1 superinfection incidence in humans published to date.

As described above, the greater genetic similarity between infecting strains in intraclade DI renders it more difficult to detect than interclade DI [67]. Other challenges to identifying DIs arise when one strain composes a small minority of the total circulating viral population [67, 31], or when the two infecting strains recombine, making it impossible to detect DI on the basis of a single genomic fragment that has been homogenized by recombination [31, 52]. This notion is supported by studies from Piantadosi et al. [52], who detected additional cases of DI using a second coding region of HIV.

Table 1.1: Published studies of incidence of HIV-1 superinfection in humans. IDU: injection drug use, WSM: women who have sex with men, MSM: men who have sex with men.

Citation	**Incidence reported**	**Average length of follow-up**	**Inter or intra-clade**	**Risk factor**	**Regions examined**	**Methods**
Ramos et al. (2002) [59]; Hu et al. (2005) [34]	2.2%/year (2/130)	12 months	Inter	IDU	*pol* and *env*	Restriction fragment length polymorphism of *pol*, cloning of *env*
Gonzales et al. (2003) [22]	0% (0/3155)	12.2 months	NA	Not stated	*pol*, *gag*, *tat*	Multistep filtering for likely cases: genetic distance between *pol* sequences, GPS, and *gag* and/or *tat* sequencing
Fang et al. (2004) [18]	Not stated (1/7)	10 years	Intra	WSM	V1-V3 *env*	Heteroduplex tracking assay

Continued on next page

Table 1.1 – continued from previous page

Citation	Incidence reported	Average length of follow-up	Inter or intra-clade	Risk factor	Regions examined	Methods
Gottlieb et al. (2004) [23]; Gottlieb et al. (2007) [24]	Not stated (1/64)	3 years	Not stated	MSM	C2-V5 *env*	Heteroduplex mobility analysis; cloning
Manigart et al. (2004) [43]	Not stated (2/147)	Not stated	Inter	WSM	*env*	Heteroduplex mobility assay of *env*
Smith et al. (2004) [69]	5% per year (3/78 ART-naive)	6-12 months	Intra	MSM	*env* C2-V3	*Pol* longitudinal separation; C2-V3 cloning
Tsui et al. (2004) [75]	0% (0/37)	Not stated (duration of study 13 years)	Intra	IDU	p17 *gag*, V3-V5 *env*, and/or *tat*	Cloning of p17, V3-V5, and/or *tat*
Yerly et al. (2004) [80]	5% per year (3/58)	Not stated	Inter	IDU	RT-PRO, C2-V3	Bulk sequencing of RT, protease, and C2-V3; clade-specific PCR in blood plasma and PBMC
Chohan et al. (2005) [14]; Piantadosi et al. (2008) [52]	4% per year	6 years	Intra	WSM	*gag*, *pol*, *env*	Strain-specific PCR
Piantadosi et al. (2007) [50]	3.7% per year (7/36)	6 years	Intra	WSM	*gag*, *pol*, *env*	Strain-specific PCR

Continued on next page

Table 1.1 – continued from previous page

Citation	Incidence reported	Average length of follow-up	Inter or intra-clade	Risk factor	Regions examined	Methods
Campbell et al. (2009) [9]	Not stated (1/16)	3 years	Intra	MSM	*env*	Cloning
Templeton et al. (2009) [72]	Not stated (17/58)	3 years	Both	IDU, WSM	*pol* and *env*	Genetic screens

Because clonal and single genome sequencing of viral populations from a single host are expensive, labor intensive, and subject to possible sampling bias, new lower-cost and higher-throughput methods are needed to screen large cohorts for DI.

1.3 Superinfection's Effect on Pathogenesis

Evidence exists that DI causes faster disease progression [23], as well as several case reports of SI identified because of a spike in viral load. A central question to this topic is whether DI itself or with a dual-tropic virus causes faster disease progression, or whether individuals predisposed to rapid disease progression fail to resist the challenge of a second infection. However, comprehensive studies of the effect of SI on disease progression have been lacking, largely due to lack of suitable cohorts and sufficient numbers of identified SI cases.

1.4 Goals and Overview of the Remainder of This Dissertation

The goals of this thesis project were: 1) develop and validate a sensitive, high-throughput method to detect HIV-1 intraclade DI; 2) apply this method to a large, well-characterized study cohort to determine its prevalence of CI and incidence of SI; and 3) determine the clinical, virologic, and immunologic correlates of DI by comparing the identified DI cases to MI controls. Each of these goals is described in Chapters 2-4 of this dissertation, and Chapter 5 presents conclusions.

Chapter 2

The Use of Published Methods to Detect HIV-1 Intraclade Dual Infection

2.1 Introduction

As described in Chapter 1, a variety of techniques have been used to identify HIV dual infection (DI) since it was first recognized in humans. We applied several of these methods to a well-characterized, longitudinally followed study cohort to determine the prevalence of dual infection. A high proportion of ambiguous base calls or mixtures, e.g. "R" (A or G) and "Y" (C or T) in a population-based sequence, can be used as a marker for DI [15]. However, because non-synonymous mixtures are often a hallmark of selection by the immune response or HAART in a mono-infected HIV host [58], we evaluated a version of the method focusing on synonymous (silent) mixtures. To that end, we have developed a simple descriptive measure–synonymous mixture index or "SM-Index"–and demonstrated how it can be applied to discriminate between dually and singly infected participants. We then performed single genome sequencing to confirm the presence of multiple strains.

Standard (or "population") genetic sequencing utilizes gene amplification by PCR of the analyte population to create one sequence representing all the amplicons generated. When population sequencing reports a mixture of bases at a given position, this indicates diversity or the existence of multiple variants at that position within the

genetic population. Minority variants usually comprise greater than 20% of the viral population to be detected by bulk/population sequencing [29]. A high proportion of ambiguous base calls or mixtures, e.g. "R" (A or G) and "Y" (C or T) in a population-based sequence, may indicate DI [15]. Because non-synonymous mixtures are often a hallmark of selection by the immune response or HAART in a mono-infected HIV host [58], we evaluated a version of the method focusing on synonymous (silent) mixtures. To that end, we developed a simple descriptive measure synonymous mixture index or SM-Index and applied it to determine if it could discriminate between dually and singly infected participants. We also used the same population-based sequences to determine whether the *pol* population for each host was stable over time. An overall population shift could indicate replacement by a second strain that either appeared later than the first (superinfection) or was present from baseline as a small minority (coinfection).

Population-based sequencing is a rapid and effective tool for DI screening, but it is insufficiently sensitive for confirmation. To increase the detection of minority genetic variants within a sample population, we used single genome sequencing (SGS), a terminal dilution technique that has been used frequently in HIV research [49, 64, 65, 66]. This technique attempts to isolate a single molecule of viral DNA or copy DNA (cDNA) generated by reverse transcription (RT) of viral RNA through serial dilution testing. Specifically, DNA or cDNA is diluted serially over a range of concentrations, and the concentration at which $\leq$30% of reactions contain amplifiable cDNA may be expected, assuming a Poisson distribution, to yield product generated from a single template in approximately 80% of those samples. By using a single molecule of DNA or cDNA as the template for amplification and sequencing, the risk of nucleotide misincorporations or template switching introduced during PCR amplification is reduced [19, 44, 49, 64, 65, 79], and with repeated sampling of the viral population, SGS can detect minority populations of less than 20% of the total population. We applied SGS to isolate individual genomes from the circulating HIV population in each sample and then assessed their relatedness to each other by phylogenetic analysis. Because SGS is a time-consuming and laborious process, we proposed two methods to improve its efficiency: 1) measurement of cDNA concentration to expedite the identification of the terminal dilution and 2) use of a bioinformatics application for interpreting the experimental results of a dilution test.

Table 2.1: Cohort Demographics. Age, HIV viral load, and CD4 values are shown as median (interquartile range).

Age at enrollment	**31 (26-38)**
Sex	**N**
Male	114
Female	3
Race	**N**
White	95
Black or African American	8
Asian or Pacific Islander	3
American Indian	9
Unknown	1
Ethnicity	**N**
Hispanic or Latino/Latina	21
Not Hispanic or Latino/Latina	47
Unknown	48
HIV Risk Factors	**N**
MSM	106
Heterosex	5
MSM and Injection Drug Use	4
Unknown	1
HIV viral load at enrollment (copies/mL)	**59,150 (6,225-219,000)**
CD4 at enrollment (cells/μL)	**540 (460-704)**
Length of ART-naive followup (months)	**19 (8-36)**

2.2 Methods

2.2.1 Study Cohort

All participants in the San Diego Acute Infection and Early Disease Research Program [69] who had deferred antiretroviral therapy (ART) for at least the first 6 months after infection and had at least two blood samples available were included for DI screening (N=117). Demographics of participants are shown in Table 2.1.

2.2.2 Population-based Sequencing of *pol* from Blood Plasma

HIV RNA extraction and population-based *pol* (HXB2 coordinates 2253-3554) sequencing (Viroseq version 2.0, Celera Diagnostics, Foster City, CA, USA) were per-

formed for at least two time points for each of the study participants as previously described [69].

The SM-Index: A Screening Method for HIV Dual Infection

The SM-Index descriptive measure was calculated as the number of synonymous base pair mixtures in a *pol* sequence divided by the number of synonymous sites in it. The sequences were ranked for likelihood of DI according to the SM-Index, i.e. higher SM-Index indicated greater synonymous population heterogeneity, and hence a greater probability of DI.

Identification of Distinct *pol* Populations Over Time

All population *pol* sequences from all study participants were aligned and manually edited, and a single phylogeny was generated. We identified a population shift within the same individual over time when sequences from the same participant at different dates were less closely related to each other than to at least one other epidemiologically unrelated isolate.

2.2.3 Single Genome Sequencing of C2-V3 and RT from Blood Plasma

HIV RNA was extracted from the blood plasma samples (QIAamp Viral RNA Mini Kit, Qiagen, Hilden, Germany) and cDNA produced (RETROscript Kit, Applied Biosystems/Ambion, Austin, TX, USA). The expected concentration of cDNA (copies/μL) after reverse transcription is given by the formula:

$$[cDNA] = \frac{BPR}{EF}$$

where

B = blood plasma viral load of the specimen (copies/μL)
P = plasma volume of sample used for reverse transcription (μL)
R = volume of RNA elution used in reverse transcription (μL)
E = volume into which extracted RNA has been eluted (μL)
F = final volume of reverse transcription reagents plus R (μL)

The standard protocol for SGS begins with an estimation of the amount of DNase-free water necessary to add to the sample of cDNA to dilute its concentration to approximately 10 copies per μL and then dilutes this concentration threefold three times.

This results in a range of dilutions for testing with hypothetical concentrations of 10, 3.3, 1.1, and 0.4 cp/μL. Each dilution is used in 16 wells for PCR, and if a dilution yields product in 4 wells, then that dilution is used for a full 96-well plate. If none of the dilutions yield 4 positive reactions, then more or less DNase-free water is added to alter the concentrations of cDNA, and the experiment is repeated until the right dilution is found based on 4 positive reactions. Second-round products are electrophoresed on 1% agarose gels to assess the fraction of positive reactions for a given dilution of cDNA.

Nested PCR reactions are performed using 10 μL of diluted cDNA template added to 40 μL of reaction mixture for the first round. The reaction mixture for amplifying the C2-V3 portion of *env* (HXB2 coordinates 6928-7344) consists of 5.0 μL of 10× PCR Buffer containing magnesium chloride and 1.0 μL of 10 nM dNTP Mix (GeneAmp, Applied Biosystems, Foster City, CA, USA), 0.25 μL of Taq DNA Polymerase (Roche Diagnostics, Indianapolis, IN, USA), 31.75 μL of molecular grade water, and 1 μL each of two 20 μM primers:

V3-F_{out} (5'-CAAAGGTATCCTTTGAGCCAAT-3')
V3-B_{out} (5'- ATTACAGTAGAAAAATTCCCCT-3')

The 50 μL samples are heated to 95 °C for 2 minutes and then subjected to 35 cycles of 30 seconds at 95 °C followed by 30 seconds at 50 °C followed by 60 seconds at 72 °C. After this, the samples are heated to 72 °C for 10 minutes and then held at 4 °C until used.

The second round PCR utilizes 5 μL of the first round product as template added to 45 μL of reaction mixture, for a total volume of 50 μL. This reaction mixture consists of the same reagents, but the volume of molecular grade water is increased to 36.75 μL. For this round, the primers used are:

V3-F_{in} (5'-GAACAGGACCAGGATCCAATGTCAGCACAGTACAAT-3')
V3-B_{in}(5'-GCGTTAAAGCTTCTGGGTCCCCTCCTGAG-3')

The cycling parameters are the same as for the first round.

The protocol for SGS of the RT portion of *pol* (HXB2 coordinates 2708-3242) was identical to the nested C2-V3 PCR protocol, including the thermal cycler settings,

with the following primer substitutions:

First round
CI-POL1 (5'-GGAAGAAATCTGTTGACTCAGATTGG-3')
3RT (5'-ACCCATCCAAAGGAATGGAGGTTCTTTC-3')

Second round
5RT (5'-AAATCCATACAATACTCCAGTATTTGC-3')
3RT (5'-ACCCATCCAAAGGAATGGAGGTTCTTTC-3')

15-30 single genome sequences per coding region were generated for each of the selected blood plasma samples.

Phylogenetic Analysis: Separation by Background Sequences

All SGS reads were checked for inter-sample and lab strain contamination by performing Megablast [81] homology searches against public HIV databases and each other. The SGS reads for each coding region per sample were added to a set of HIV clade B background sequences, aligned, and manually edited. Dual infection was interpreted by phylogenetic analysis when sequences from the same sample were no more closely related to one another than to an epidemiologically unlinked sequence.

Proposed Improvements

Addition of cDNA Quantification Step Real-time quantitative PCR was performed to quantify HIV-1 cDNA copies in a TaqMan-based approach as described previously [30]. The forward and reverse primers (HXB2 coordinates 4809-4829 and 4957-4974) and probe sequence (HXB2 coordinates 4896-4922) are as follows:

Forward: 5'-TACAGTGCAGGGGAAAGAATA-3'
Reverse: 5'-CTGCCCCTTCACCTTTCC-3'
Probe: 5'-TTTCGGGTTTATTACAGGGACAGCAG-3'

They were made (Integrated DNA Technologies Inc., Coralville, IA, USA) with specificity to the p31 integrase domain of *pol* [60]. TaqMan standards were derived

from a linearized, full-length HIV-1 clone, pNL-EX (courtesy of Dr. Yoshiharu Miura, Tohoku University Graduate School of Medicine, Sendai, Japan) in dilutions ranging from 1×10^6 copies/reaction to 20 copies/reaction. Each reaction consisted of 5 μL of HIV-1 standard template or sample cDNA and 12.5 μL of 2× Universal PCR Master Mix (Applied Biosystems, Foster City, CA, USA). Primers and probe were present in final concentrations of 200 and 900 nM, respectively. All amplifications, including negative controls, were performed in duplicate with the ABI 7900HT Sequence Detection System (Applied Biosystems) using cycling parameters of 50 °C for 2 minutes, then 95 °C for 10 minutes, followed by 45 cycles of 95 °C for 10 seconds, then 60 °C for 1 minute.

SGS Dilution Calculator An application was developed that uses the Poisson distribution to calculate the real concentration of the terminal dilution (D) based on the number of positive PCR reactions (P) and the total number of PCR reactions run (T).

According to the Poisson distribution,

$$P(X = x) = \frac{\lambda^x \cdot e^{-\lambda}}{x!}$$

where

P(X = x) = probability of getting x template copies in one PCR reaction
λ = average number of cDNA template copies in each PCR reaction
X = number of cDNA template copies in one PCR reaction (random variable [0,1,2 ...])

Thus

$$\frac{P}{T} = P(X \geq 0) = 1 - P(X = 0) = 1 - \frac{\lambda^0 \cdot e^{-\lambda}}{0!} = 1 - e^{-\lambda}$$

Therefore

$$\lambda = -ln\left(1 - \frac{P}{T}\right)$$

Also,

$$D = \frac{\lambda}{10}$$

i.e., the actual concentration of the terminal dilution, as the PCR protocol calls for 10 μL template per reaction.

This application, which is called "SGS Calculator", also uses the real terminal dilution concentration (D) combined with two other inputs (the putative concentration of the terminal dilution and the putative concentration of the cDNA) to determine the real concentration of the cDNA. This is accomplished by solving for the unknown quantity after setting the following proportion:

$$\frac{\text{Actual cDNA concentration [unknown]}}{\text{Actual terminal dilution concentration [D, known]}} = \frac{\text{putative cDNA concentration [known]}}{\text{putative terminal dilution concentration [known]}}$$

The user must input the number of positive PCR reactions (P) and the total number of PCR reactions (T) from a trial run as well as the putative concentration (C) of the cDNA sample, as determined by quantitative real-time PCR, and the putative dilution (D) of the sample that was used in the trial run. The outputs of the SGS Calculator are the actual dilution of cDNA in the user's trial run and the actual concentration of the sample. This application is designed for large PCR sample sizes, i.e., the total number of PCR reactions run (T) should be at least 95. SGS Calculator is increasingly inaccurate for lower values of T. The application is implemented in Javascript and currently available at http://sgscalc.ucsd.edu.

2.3 Results

2.3.1 Population-based Sequencing of *pol* from Blood Plasma

We obtained a total of 405 population-based *pol* sequences for the 117 study participants (median sequences per participant: 2, range 2-21). Table 2.2 shows sequencing results.

Table 2.2: Samples completed with each sequencing method.

Sequencing type	**Regions sequenced**	**Number of reads**	**Length of reads**	**Number of samples completed**	**Number of participants sampled**
Population based, "bulk"	*pol*	1	1300 bp	405	117
Single genome sequencing	C2-V3 and RT	25 per region	300-400 bp	72	37

The SM-Index: A Screening Method for HIV Dual Infection

The majority of samples had SM-Index values near zero (Figure 2.1: median 0.0026, range 0-0.0766).

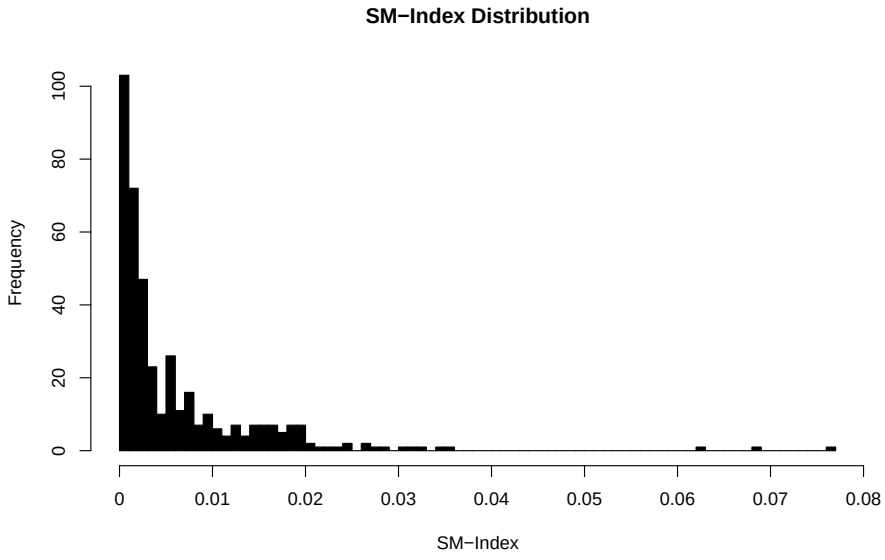

Figure 2.1: SM-Index distribution for 405 population-based *pol* sequences.

Identification of Distinct *pol* Populations Over Time

Separate *pol* populations over time were observed in 4 participants (I447, K613, K908, S155). All other *pol* sequences clustered by participant.

2.3.2 Single Genome Sequencing of C2-V3 and RT from Blood Plasma

SGS of two coding regions was obtained for 72 samples from 37 participants (Table 2.2). When one time point from a particular participant was identified as DI, the baseline time point underwent SGS to differentiate between coinfection and superinfection. When no evidence of DI was observed at baseline, intermediate time points underwent SGS to determine the timing of superinfection. DI was identified in 10 of the 117 participants, for an overall prevalence of 8.5%. Figure 2.2 shows phylogenetic evidence of DI at 2 time points for participant L537 in the RT coding region.

Longitudinal results for each of the 10 DI participants are shown in Figure 2.3. Coinfection was identified in 3 participants (L537, Q294, and U189). Superinfection was identified in 7 participants (D224, K613, K908, P265, P853, S155, and U796). Superinfection occurred in the first year of initial infection for 5 participants and in the second year of initial infection for 2 participants. P853 had transient superinfection in C2-V3 at two years of followup, but the superinfecting strain was not observed at 31 months. At 34 months, superinfection was observed in RT but not C2-V3.

SGS confirmed superinfection for 3 of the 4 participants identified as superinfected by the *pol* population turnover method (K613, K908, and S155, but not I447).

Proposed Improvements

Using the standard SGS protocol, the mean number of PCRs required to obtain an average of 30 (range: 26-34, SD: 3) SGS products per sample was 245 (range: 218-266, SD: 20) after an average of 8 trial dilutions. With the use of qRT-PCR and the bioinformatics tool, 135 PCRs (range: 135-135, SD: 0) produced 30 (range: 27-30, SD: 1) SGS products per sample using exactly 2 dilutions. The turnaround time for generating SGS product for sequencing was reduced from 8 days using the standard approach to 2 days with the new method (Fig. 2.4).

2.4 Discussion

In these experiments, we generated two types of sequences, population-based *pol* and SGS of C2-V3 and RT, in order to screen for and confirm DI. The SM-Index identified samples likely to harbor DI, although the SM-Index alone is not sufficiently powerful or accurate to confirm the presence of two strains. Population *pol* sequencing

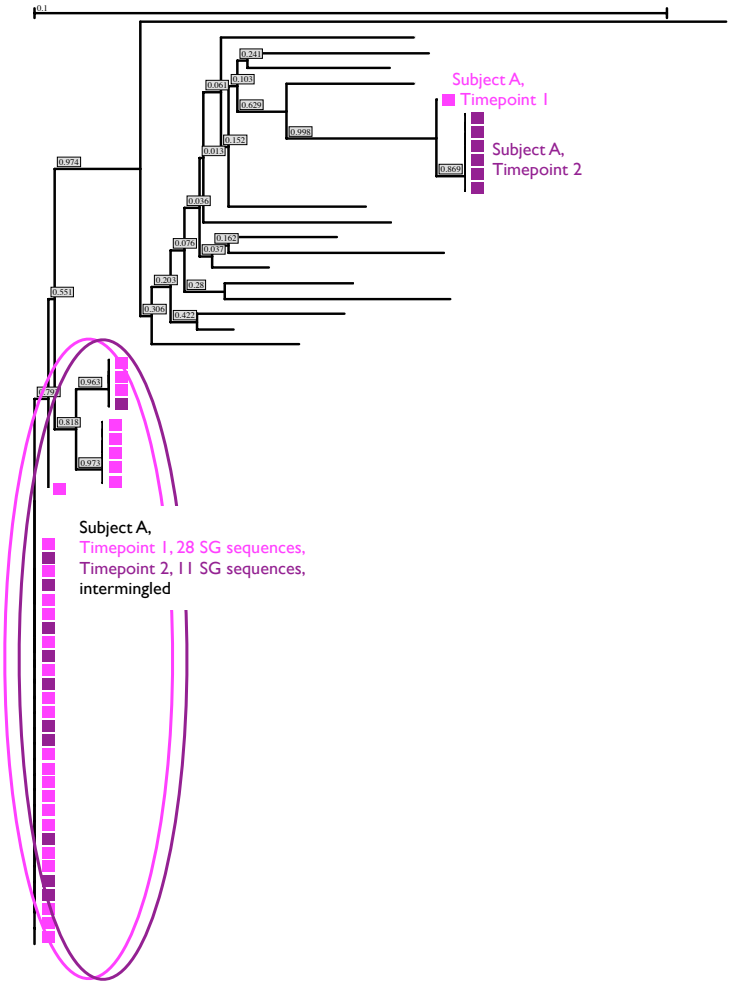

Figure 2.2: Evidence of the same 2 populations within participant L537 is visible at time point 1 (baseline) and time point 2 (1 year after initial infection).

is comparatively cheap and frequently performed for routine drug resistance testing, so SM-Index scoring based on pol genotypes remains a useful initial DI screening method. However, it alone cannot confirm DI, in part because it examines only one coding region. In our study set, the SM-Index appears to be most accurate at predicting DI for values on the extremes of its distribution.

SGS of C2-V3 and RT identified DI in 10 of the 117 participants. The combination of SM-Index with SGS identified a superinfection incidence significantly higher

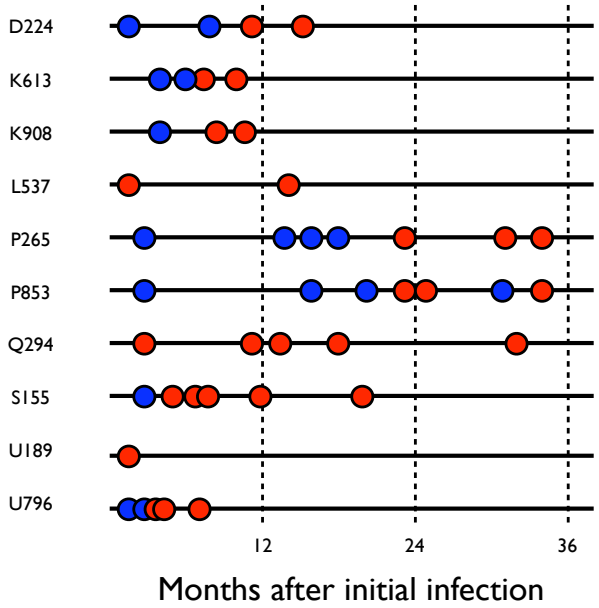

Figure 2.3: SGSed samples for the 10 DI participants. Blue circles represent no evidence of DI in either coding region sequenced, and red circles represent DI in at least one of the two coding regions.

than the 5% per year previously estimated for this cohort [69]. However, the 8.5% DI prevalence identified is a conservative estimate, since only 37 of the 117 participants underwent SGS at least once.

To improve efficiency of SGS, we proposed a new method for determining the appropriate dilution of cDNA for the terminal dilution. Employing qRT-PCR to quantitate the nominal copies of cDNA after RT can lessen the observed discrepancies between the theoretically calculated and empirically determined optimal cDNA concentration to use for end-point dilution testing. The reasons for these discrepancies include: 1) a particular specimen may contain a concentration of HIV-1 RNA that is outside the dynamic range for which the viral load assay has optimal accuracy; 2) the number of freeze/thaw cycles a specimen undergoes will affect the integrity of viral RNA available for participation in reverse transcription; 3) extraction of viral RNA from blood plasma may not capture all of the RNA measured by the viral load assay; 4) reverse transcription of RNA to cDNA is less than 100% efficient regardless of the procedure used, the number of freeze/thaw cycles, or the accuracy of the viral load assay. Furthermore, depending upon the type of research being performed, the amount of clinical material available for study can be a limiting factor with the standard method of SGS. If viral populations from compartments

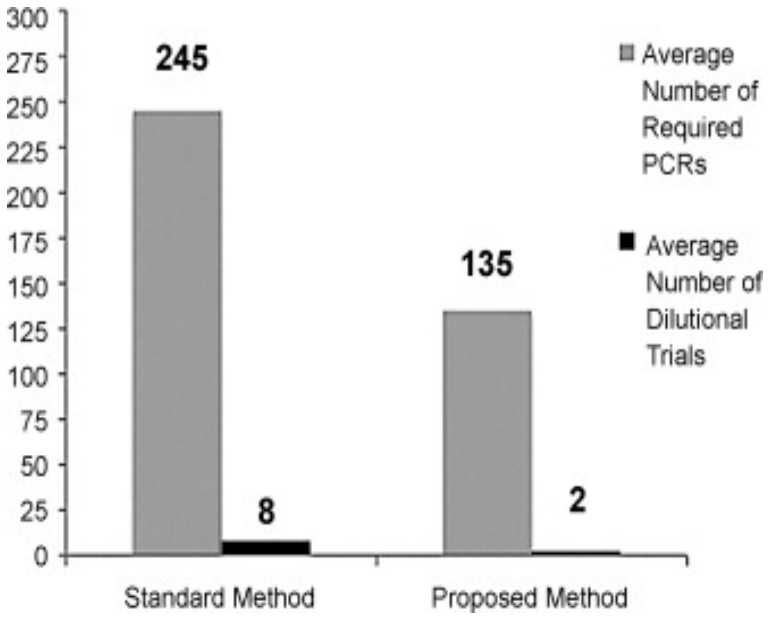

Figure 2.4: Comparison of standard and proposed methods. Using the standard method of single genome sequencing (SGS), the mean number of PCRs to produce an average of 30 (26-34, SD = 3) SGS per sample was 245 (218-266, SD = 20) after an average of 8 trial dilutions. With the proposed method, 135 PCRs (135-135, SD = 0) produced 30 (27-30, SD = 1) SGS using exactly 2 dilutions.

other than the blood (e.g., cerebrospinal fluid) are being characterized, then the quantity of sample limits the number of trials that may be performed searching for the optimal dilution to use for SGS. Using the methods proposed here, SGS was accomplished using on average only 2 rather than 8 trial dilutions. Although some of the original cDNA was used for the qRT-PCR test itself, less is used for this purpose than for a typical trial dilution in standard SGS. In conclusion, the method proposed here will increase the efficiency of the SGS procedure. This can reduce cost by decreasing the amount of reagents and labor involved, and it also may allow for application of this research tool to a broader range of investigations, as the amount of clinical material used for determining the optimal dilution is less than required previously.

Although our improvements to the SGS method allowed for substantially faster sequencing, higher-throughput methods were still needed to examine the HIV population structure of all study participants. While incurring significant cost, labor, and time overheads, we still were only able to perform SGS on samples from less than one third of the study cohort. In the next chapter, we describe a sequencing method more suitable

for determining the prevalence of DI in large cohorts.

Chapter 2 is, in part, a reprint of the paper "Butler DM, Pacold ME, Jordan PS, Richman DD, Smith DM. The efficiency of single genome amplification and sequencing is improved by quantitation and use of a bioinformatics tool. *J Virol Methods*, 2009 Dec;162(1-2):280-3." The dissertation author was the second author and investigator of this paper.

Chapter 3

Validation and Use of Ultra-deep Sequencing to Detect HIV-1 Dual Infection

3.1 Introduction

The advent of next-generation or ultra-deep sequencing (UDS) technologies has made it feasible to generate a high-resolution snapshot of viral diversity in a biological sample rapidly and relatively inexpensively by direct sequencing. This approach appears particularly promising for studying rapidly mutating RNA viruses such as HIV-1 [17, 33]. A number of recent studies have successfully used UDS to detect HIV minority variants with drug-resistant mutations [33, 76, 39, 74], different chemokine co-receptor usage [2], various integration sites [77], and distinctive novel variants [7]. Given the ability of UDS to identify HIV minority variants as low as 1% in experiments with reconstructed samples [33, 74], we hypothesized that UDS would be similarly adept at screening for DI. To test this hypothesis, we analyzed three HIV-1 genomic coding regions with UDS sequencing and wrote a custom bioinformatics pipeline to filter, align, and analyze sequence reads for evidence of DI. The performance of SM-Index and UDS were then compared for DI screening, using single genome sequencing (SGS) as the gold-standard reference method.

Having validated UDS as a method to detect DI, we applied it to a set of samples from our study cohort in order to determine the prevalence of DI. We integrated the UDS results with the SGS results described in Chapter 2.

3.2 Methods

3.2.1 Validation: Study Participants

All participants in the San Diego Acute Infection and Early Disease Research Program 2 who had deferred antiretroviral therapy (ART) for at least the first 6 months after infection and had at least two blood samples available were included for DI screening. We screened male participants who were infected with HIV-1 subtype B and reported sex with men as an HIV risk behavior. To evaluate whether UDS could distinguish DI from the natural history of viral evolution in a host, we evaluated 5 samples (D1, D2, E, F, and H) collected from four individuals with an estimated infection duration of 30 months or more. To evaluate the utility of screening methods for individuals who had received ART, we also chose samples from two participants (F and H) who were initially ART-naive for at least one year and then underwent ART but had detectable, amplifiable HIV RNA. In one of the ART-experienced participants (F), the virus had a large number of drug resistance associated mutations. The last criterion for inclusion in UDS screening and comparison study was to select samples that represented low, medium, and high SM-Index scores (see below).

3.2.2 Screening Methods

Population-based *pol* sequences for the SM-Index were generated as described in Section 2.2.2. For UDS, HIV RNA was extracted from the blood plasma samples (QIAamp Viral RNA Mini Kit, Qiagen, Hilden, Germany) and cDNA produced (RETROscript Kit, Applied Biosystems/Ambion, Austin, TX, USA). Three coding regions (Fig. 3.1)–*gag* p24 (HXB2 coordinates 1366-1619), *pol* RT (HXB2 coordinates 2708-3242), and *env* C2-V3 (HXB2 coordinates 6928-7344)–were amplified by PCR with region-specific primers as described in Section 2.2.3 for RT and C2-V3.

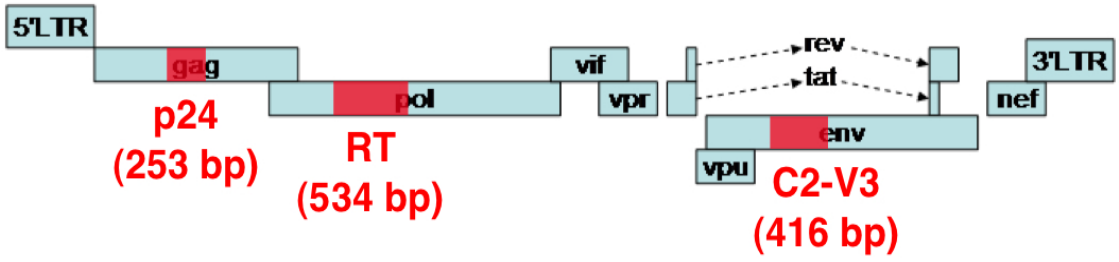

Figure 3.1: HIV-1 genomic regions sequenced.

The *gag* p24 PCR methodology was as follows: Nested polymerase chain reactions were performed using 2.5 μL of diluted cDNA template added to 47.5 μL of reaction mixture for the first round. The reaction mixture consisted of 5.0 μL of 10X PCR Buffer containing magnesium chloride and 1.0 μL of 10 mM dNTP Mix (GeneAmp, Applied Biosystems, Foster City, CA, USA), 0.25 μL of Taq DNA Polymerase (Roche Diagnostics, Indianapolis, IN, USA), 39.25 μL of molecular grade water, and 1 μL of each of two 20 μM primers:

CI-p24gag1312F$_{out}$ (5'-TATCAGAAGGAGCCACCC-3')
CI-p24gag1846B$_{out}$ (5'-CTCCCTGACATGCTGTCATCA-3')

The 50 μL samples were heated to 94 °C for 2 minutes and then subjected to 35 cycles of 30 seconds at 94 °C followed by 30 seconds at 58 °C followed by 60 seconds at 72 °C. After this, the samples were heated to 72 °C for 10 minutes and then held at 4 °C until used. The second round PCR utilized 2.5 μL of the first round product as template added to 47.5 μL of reaction mixture for a total volume of 50 μL. This reaction mixture consisted of the same reagents in the same volumes. For this round, the primers used were:

CI-p24gag1366F$_{in}$ (5'-GGACATCAAGCAGCCATGCAAATG-3')
CI-p24gag1619B$_{in2}$ (5'-TACATTCTTACTATTTTATT-3')

The 50 μL samples were heated to 94 °C for 2 minutes and then subjected to 35 cycles of 30 seconds at 94 °C followed by 30 seconds at 42 °C followed by 60 seconds at 72 °C. After this, the samples were heated to 72 °C for 10 minutes and then held at 4 °C until used.

Rubber gaskets were used to physically separate 16 concurrently sequenced samples on a single 454 GS FLX Titanium picoliter plate (454 Life Sciences, a Roche company, Branford, CT, USA). A custom bioinformatics pipeline was designed, as described below, to select high-quality UDS reads, generate consensus sequences, align reads to the consensus, and perform phylogenetic analysis of specific coding regions, used to identify DI.

Bioinformatics Platform to Analyze Ultra-deep Sequencing Results

Initial read files filtering UDS generates both the set of called bases (reads) in FASTA file format and a quality score for each base. The quality scores are industry standard PHRED values that provide a confidence level that a given base call is correct. For this study, we used a PHRED cutoff value of 20, i.e. 1 expected error in 100 bases. We designed a filtering program (following the procedure described in Kosakovsky Pond et al. [56]) that examines each read and its accompanying base-by-base PHRED score to select fragments with runs of good quality bases. Filtering employs the following algorithm: i) Each retained fragment must have a continuous run of PHRED scores of 20 or greater for 50 or more bases; ii) The only exception to the above rule is made for homopolymers, a known source of error for the Roche 454 pyrosequencing platform. In this case, if a base with a poor score follows the same base with a good score, the run is extended; iii) If the original read contains multiple discontinuous high-quality fragments, then each output is delivered as a separate (shorter) read.

Read alignment and filtering An iterative HIV-1 gene-specific alignment and filtering procedure was implemented as a collection of scripts for the HyPhy software package [53] (available from the HyPhy subversion system code repository) to construct a high quality region consensus sequence and map individual reads. The procedure works in 3 steps:

1. A starting reference sequence was used to protein align each of the 6 possible translations of each read (using the 5% divergence HIV scoring matrix from Nickle et al. [46]) and select the frame with the highest alignment score for each read. The best score per position for each read was compared against the expected value for a random sequence with an HIV-like residue composition, and only the reads exceeding the threshold by a factor of 5 or greater by high protein-alignment scoring (HPAS) were included in building the codon sample reference sequence (SRS).

2. To recover sequences with frameshifts (e.g. due to a homopolymer length error), we computed the median of the distribution of nucleotide alignment scores (per position) of each read from step 1 to the SRS. This median defines a lower threshold for filtering sequences initially excluded in step 1 (M). The reads excluded in step 1 were nucleotide-aligned to the SRS and included in the analysis if their nucleotide scores/position exceeded M. Note that steps 1 and 2 automatically separate mixed genomic regions, because only the reads that align well with the reference gene of

interest are retained.

3. A final consensus sequence was generated from HPAS reads and reads retained in step 2. This consensus was used as a coordinate system to tabulate the position of each residue in high-scoring sequence reads.

The result of each filtering run was an SQL database with a variety of metrics, consensus sequences, and high-scoring sequence reads aligned to and mapped onto the consensus. Each participant read set was run through the pipeline using HXB2 *gag*, *pol*, and *env* sequences as initial (Step 1) references, resulting in region-specific alignments.

Individual sample analyses Databases of curated and mapped reads for each genomic region from individual patient samples were examined for molecular evidence of DI. We analyzed sliding sequence windows of length L $\geq$ 125 bp (L determined based on the median read length in the database) with stride 25, which were covered by at least 400 reads. Individual reads were required to completely cover the window to be included in the analysis. We did not perform contig assembly, partly because sufficient signal was obtained directly from shorter reads, and partly because HIV-1 is known to have very high rates of recombination, complicating the assembly. We condensed reads identical within a single window to unique variants and the corresponding copy numbers. Only the variants with at least 5 copies or 0.5% of the reads (whichever was greater) were used for further analyses, in order to further reduce the influence of sequencing errors. Maximum likelihood pairwise nucleotide distances (Tamura-Nei 93) were computed for the variants, and 95% confidence intervals of each distance estimate were obtained via non-parametric bootstrap. If the distance estimate between a pair of variants exceeded a preset threshold D (see below), and the lower bound corresponding confidence interval was greater than D, then the sample was classified as putatively dually infected. When more than 3 variants were present, putative dually infected windows were further examined using standard phylogenetic analyses (bootstrap) to confirm the presence of two or more genetically divergent populations. Genetic distance cutoffs for potential dual infection were chosen to exceed typical within-sample divergence produced by chronic monoinfection–1.7% in *gag* and 3.1% in *env* [51]. Divergence thresholds were set at 2% for RT and p24 and 5% for C2-V3. A sample was classified as dually infected if the divergence of at least one of its coding regions exceeded the threshold and if the phylogenetic structure of at least one sliding window in that region indicated dual infection, i.e. two viral subpopulations

separated by a branch with high bootstrap support.

3.2.3 Confirmation Method: Single Genome Sequencing (SGS)

Using the same viral cDNA produced for UDS, we generated SGS of *env* C2-V3 (HXB2 coordinates 6928-7344) and *pol* RT (HXB2 coordinates 2708-3242), as previously described in Section 2.2.3. The RT and C2-V3 regions amplified were identical to the RT and C2-V3 regions amplified for UDS. 15-30 single genome sequences per coding region were generated for each of the selected blood plasma samples. Sequences were subjected to the same phylogenetic analyses and genetic distance cutoffs for DI as UDS reads. All UDS and SGS reads were checked for inter-sample and lab strain contamination by performing MEGABLAST [81] homology searches against public HIV databases and each other.

3.2.4 Cost and Time Analyses

We calculated the cost of reagents, disposable materials, kits, sequencing runs, and labor for obtaining SM-Index, UDS, and SGS. Time per sample was calculated as the labor time plus instrument time required to perform each experimental step of the methods.

3.2.5 Application of Ultra-deep Sequencing to Determine the Prevalence of DI

Our algorithm for determining the prevalence of dual infection and incidence of superinfection was as follows:

1. Obtain the SM-Index for at least two time points per participant.

2. Perform UDS or SGS on the sample with the highest SM-Index per participant. If multiple samples have the highest SM-Index value, choose the sample latest in followup.

3. If the sample with the highest SM-Index is a baseline sample, DI is not observed in it, and the participant has > 1 year of ART-naive followup , perform UDS or SGS on the last available time point sample.

4. If at least 2 coding regions are successfully sequenced with UDS or SGS for the highest SM-Index sample and, if appropriate, the last time point sample, and none demonstrate DI, classify the participant as singly infected and sequence no further samples.

5. If a sample demonstrates DI and the minority population is $\geq$10% of the total, perform either SGS or UDS on a sample from the same date or within 3 months of it to confirm DI. If a sample demonstrates DI and the minority population is $<$10% of the total, perform UDS for confirmation.

6. If DI is not confirmed with additional sequencing at the same or a neighboring date, classify the participant as singly infected and sequence no further samples.

7. If DI is confirmed and the sample is either within 3 months of the estimated date of infection or is the first available sample from the estimated date of infection, classify the participant as coinfected and sequence no further samples.

8. If DI is confirmed and the sample is not a baseline sample as defined in step 7, perform either UDS or SGS on a baseline sample according to the size of the minority population, as described in step 5.

9. If DI is observed in the baseline sample, classify the participant as coinfected and sequence no further samples.

10. If DI is not observed in the baseline sample, classify the participant as superinfected and sequence further samples using a binary search approach to determine the timing of superinfection within a 1-year window.

The prevalence of DI was calculated as the total number of DI cases divided by the cohort size. The prevalence of coinfection was calculated as the number of coinfections divided by the cohort size. The incidence of superinfection was calculated as the number of new superinfection cases per year for each year of followup after initial infection.

3.3 Results

3.3.1 Validation of Ultra-deep Sequencing as a Method to Detect DI: Screening and Confirmation Methods

SM-Index To select specimens for analysis, the SM-Index was calculated for all participants in the cohort (n=116) who had population based *pol* sequences available from multiple time points (n=405 sequences), as described in Section 2.2.2. We then chose ten samples from nine participants with a range of SM-Index values for further comparison with UDS screening methods, as described below.

UDS and SGS Ten blood plasma samples were selected based on their SM-Index values: low (0-0.0013, Samples A-C), medium (0.0168-0.0270, Samples D1-H), and high (0.0766, Sample I). To assess the utility of proposed methods in clinical cohorts, samples were also selected to span a range of viral loads (3.05-6.36 log10 HIV RNA copies/ml). Demographics of the participants and clinical data associated with the nine samples are shown in Table 3.1, with two of the samples, D1 and D2, obtained from the same participant at different time points.

Table 3.1: Participant characteristics and clinical data. CD4 count and viral load refer to the dates shown.

Participant	**Date**	**Age**	**Race/ Ethnicity**	**Estimated duration of infection (months)**	**CD4 count (cells/ μL)**	**Viral load (log copies/ mL)**	**Anti-retroviral naive?**	**SM-Index**
A	7/19/00	21	White	3.1	535	5.05	Yes	0.0000
B	11/30/01	30	White	7.3	527	4.54	Yes	0.0000
C	12/21/05	26	White	1.5	746	6.36	Yes	0.0013
D1	10/18/05	24	Hispanic	31	366	4.26	Yes	0.0168
E	4/13/06	49	Hispanic	39.9	796	4.26	Yes	0.0174
F	9/15/06	40	White	49.6	298	4.36	No	0.0179
D2	1/10/06	24	Hispanic	33.8	468	5.00	Yes	0.0204
G	8/17/04	35	White	2.8	321	4.58	Yes	0.0263
H	7/24/03	40	Black	70.6	733	3.05	No	0.0270
I	9/2/05	19	White	1.5	744	5.42	Yes	0.0766

In order to evaluate the efficacy of the SM-Index for participants who had undergone at least some ART, we chose one sample (H) from a participant who was ART-naive

for the first 15 months, and then was placed on a Nelfinavir, Zidovudine and Lamivudine regimen for 4.6 years. As would be expected for an individual receiving ART and having ongoing viral replication (i.e. detectable viral load), we identified a mutation associated with resistance to the ART he was taking–protease inhibitor major resistance mutation M46LM.

Another participant (F) was chosen to evaluate if pre-existing HIV drug resistance and continued antiretroviral pressure with resistance influenced molecular methods of detecting DI. Specifically, participant F had three-drug class resistance mutations at baseline, identifying transmitted drug resistance (protease inhibitor major resistance mutations: I54V, I84V, L90M and minor resistance mutations: L10I, A71V; nucleoside reverse transcriptase inhibitor resistance mutations: M41L, D67N, T215Y; and non-nucleoside reverse transcriptase inhibitor resistance mutations: K101P, K103N). He then underwent a variety of dual (Tenofovir and Emtricitabine) and quadruple (Didanosine, Ritonavir, Atazanavir, Tenofovir) therapy regimens that never completely suppressed his viral load, and at the time of study evaluation his population-based *pol* sequence contained all of his baseline drug resistance mutations, with the addition of two nucleoside reverse transcriptase inhibitor resistance mutations: K70E and M184V.

UDS was performed in duplicate for the seven samples with enough cDNA to run parallel reactions (all samples except E, F, and G). UDS produced an average of 4,650 high-quality UDS reads per sample region, while SGS averaged 25 reads. One UDS sample region (RT of sample C) had too few high-quality reads to infer DI status. Both SGS and UDS identified samples A, B, C, E, F, and G as singly infected and samples D1, D2, H, and I as dually infected. DI results specific to the coding regions of each sample are shown in Table 3.2.

For nearly all the samples, the high read coverage of UDS identified greater maximum divergence than SGS (Table 3.2). Duplicate UDS runs performed on the same sample cDNA for the same coding regions agreed in DI status for all 20 cases. Combined phylogenies of UDS and SGS for Sample I are shown in Figure 3.2. The one sample (H) in which the divergence found by SGS in both C2-V3 and RT exceeded that of UDS was the sample with the lowest viral load tested, 1,113 HIV RNA copies/ml, in which the calculated input copy number that was interrogated by UDS was only 52.3. UDS of the *gag* p24 region identified DI only for sample I, which had the highest SM-Index of the cohort and was also the only sample whose UDS and SGS of the C2-V3 and RT coding

Table 3.2: Dual infection per-region analysis. N/A: Not applicable, since *gag* SGS was not performed. Samples E, F, and G lacked sufficient cDNA to run UDS duplicates. Divergence values are shown in parentheses as the bootstrapped bottom 5% quantile of the divergence distribution.

Sample	**Estimated duration of infection (months)**	**Genetic region**	**SGS: dual infection?**	**UDS first duplicate: dual infection?**	**UDS second duplicate: dual infection?**
A	3.1	*env* C2-V3	N (1.0%)	N (2.6%)	N (2.1%)
		pol RT	N (1.0%)	N (1.7%)	N (1.6%)
		gag p24	N/A	N (0.8%)	N (0.8%)
B	7.3	*env* C2-V3	N (0%)	N (1.7%)	N (0.8%)
		pol RT	N (0%)	N (1.6%)	N (0.8%)
		gag p24	N/A	N (0.8%)	N (1.6%)
C	1.5	*env* C2-V3	N (0.3%)	N (0%)	N (0.8%)
		pol RT	N (0.8%)	N/A (poor quality reads)	N/A (poor quality reads)
		gag p24	N/A	N (0%)	N (0.8%)
D1	31	*env* C2-V3	Y (12.5%)	Y (18%)	Y (18.4%)
		pol RT	Y (2.4%)	Y (4.1%)	Y (4.9%)
		gag p24	N/A	N (3.2%)	N (2.0%)
E	39.9	*env* C2-V3	N (3.4%)	N (5.9%)	N/A
		pol RT	N (2.0%)	N (4.1%)	N/A
		gag p24	N/A	N (1.6%)	N/A
F	49.6	*env* C2-V3	N (3.9%)	N (7.0%)	N/A
		pol RT	N (1.8%)	N (3.2%)	N/A
		gag p24	N/A	N (3.2%)	N/A
D2	33.8	*env* C2-V3	Y (11%)	Y (20.2%)	Y (20.4%)
		pol RT	Y (5.2%)	Y (3.3%)	Y (5.5%)
		gag p24	N/A	N (3.2%)	N (3.2%)
G	2.8	*env* C2-V3	N (0%)	N (0.9%)	N (0.8%)
		pol RT	N (0.4%)	N (0.8%)	N/A
		gag p24	N/A	N (0.8%)	N/A
H	70.6	*env* C2-V3	N (1.2%)	N (0%)	N (0%)
		pol RT	Y (4.4%)	Y (2.4%)	Y (4.0%)
		gag p24	N/A	N (0.8%)	N (0%)
I	1.5	*env* C2-V3	Y (16.4%)	Y (27%)	Y (27%)
		pol RT	Y (5.7%)	Y (5.8%)	Y (8.2%)
		gag p24	N/A	Y (8.2%)	Y (7.4%)

regions both identified DI (Figure 3.2).

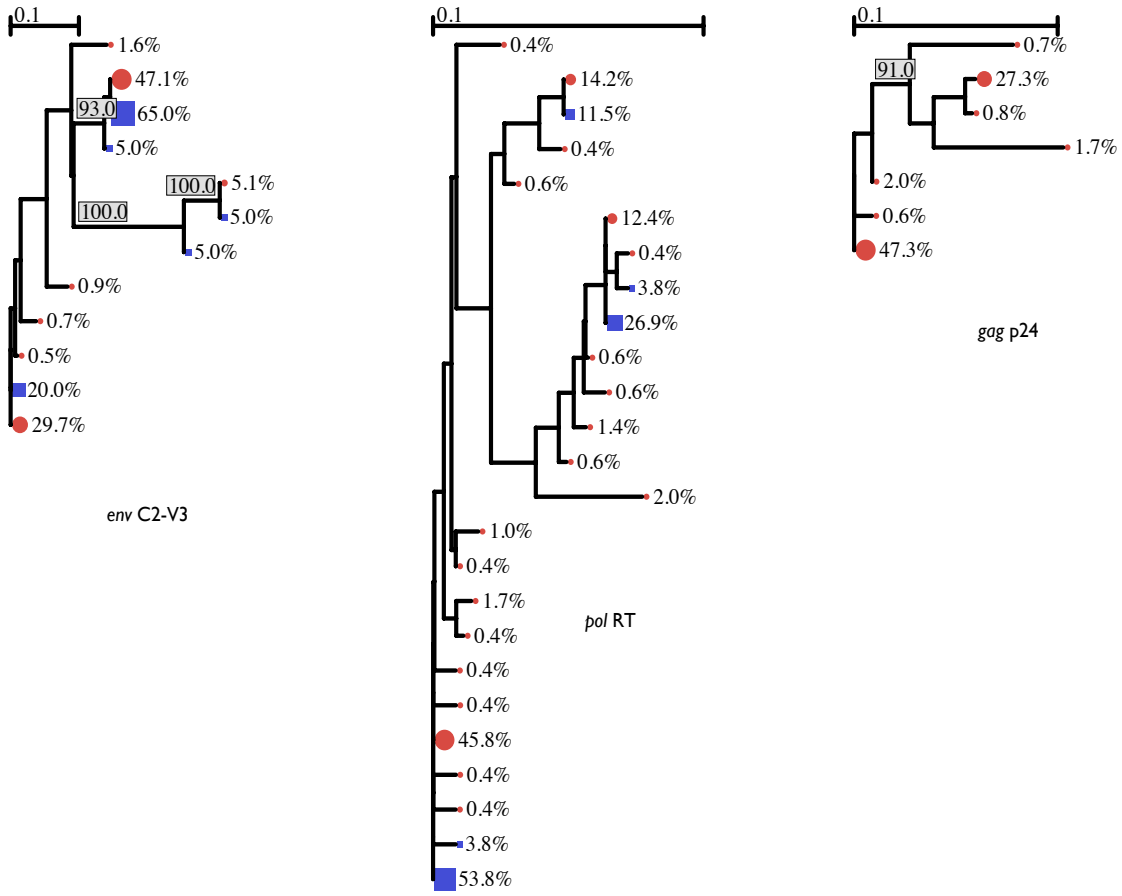

Figure 3.2: U189 phylogenies for 3 coding regions. SGS are represented as blue squares and UDS as red circles. Variant abundances per node and branches with $\geq 90\%$ bootstrap support are labeled.

3.3.2 Cost and Time Analyses

We estimated cost and time per sample for SM-Index, SGS, and UDS based on a batch of 16 samples (corresponding to a single UDS run). The cost per sample for population-based *pol* sequence was \$278.18, for SGS of two coding regions \$2,646.39, and for UDS of three coding regions \$1,075.10. Costs of each sequencing type are summarized in Table 3.3. It took 3 hours to produce one sample's population-based *pol* sequence, 42 hours for one samples SGS, and 9.5 hours for one samples UDS. Cost and time estimates for parallel steps like RNA extraction are highly throughput-dependent. UDS can be customized to produce fewer reads per sample at a lower cost. As previously noted [11], many factors (such as price reductions related to quantity) influence cost estimates and

may cause large price differences for experiments using the same technologies.

Table 3.3: Cost of sequencing per sample for 3 sequencing methods. All costs were calculated in US dollars.

	Population-based *pol*	**SGS C2-V3, RT**	**UDS C2-V3, RT, p24**
Kits and miscellaneous reagents	$187.05	$550.10	$263.82
Disposable materials	$32.63	$313.54	$62.57
Sequencing run	$21.75	$1,305.00	$593.83
Labor	$36.75	$477.75	$154.88
Total	$278.18	$2,646.39	$1,075.10

3.3.3 Application of Ultra-deep Sequencing to Determine the Prevalence of DI

UDS was performed on a total of 94 samples from 92 participants. DI results from both UDS and SGS for completed participants are summarized in Table 3.4.

The 10 confirmed DI cases resulted in an overall DI prevalence of 8.5%. The prevalence of coinfection was 2.6%. 5 of the 7 confirmed superinfections took place in the first year of initial infection, for an incidence of 4.3% in the first year, and the other 2 superinfections took place in the second year of initial infection, for an incidence of 1.7% in the second year. Of the suspected and confirmed DI cases identified by UDS, 75% harbored minority populations comprising $\leq$10% of the total, and 29% harbored minority populations comprising $\leq$5% of the total.

3.4 Discussion

Systematic identification of HIV DI in large cohorts has previously relied on a variety of screening methods, including population-based sequencing analysis from different time points [69], counting sequencing ambiguities [15], heteroduplex mobility assays [26], and molecular analysis of a single coding region [69]. Single genome sequencing is

Table 3.4: Dual infection results.

	Number of participants
Monoinfections	35
Coinfections	3
Superinfections	7
Suspected coinfections (confirmation pending)	11
Suspected DI (confirmation and timing pending)	26
Suspected DI (confirmation pending)	5
Unknown status (baseline monoinfection)	7
Unknown status (no samples completed)	25

the current standard to identify distinct strains in a viral population; however, SGS is too slow, expensive, and labor-intensive to be used as a screening method for the presence of DI in hundreds or thousands of biological samples. In this validation study, two alternative methods to detect DI were assessed. The SM-Index identified samples likely to harbor DI, although the SM-Index alone is not sufficiently powerful or accurate to confirm the presence of two strains. Population *pol* sequencing is comparatively cheap and frequently performed for routine drug resistance testing, so SM-Index scoring based on *pol* genotypes remains a useful initial DI screening method. However, it alone cannot confirm DI, in part because it examines only one coding region. In our study set, the three samples in the low SM-Index group were singly infected, and the one sample in the high SM-Index group was dually infected. However, the SM-Indices of the six samples in the medium SM-Index group were not ordered by DI status, suggesting that the SM-Index may be most useful for values on the extremes of its distribution.

Previous HIV DI studies have usually discerned DI via phylogenetic analysis when sequences from the same sample are no more closely related to one another than to epidemiologically unlinked (background) sequences. This approach allows inference of clade support for subpopulations, which provides additional information about the plausibility of the variants having come from a single infection event. However, it has the disadvantage of dependence on the diversity of the unlinked background sequences to

show clade separation within the study sample. In the current study, we use a bootstrap estimate of the simple metric of population diversity (the length of the longest path in the sample tree), which is easy to automate and interpret, and hence more appropriate for a high-throughput screen.

UDS is a massively parallel analog of SGS, and it has not yet been evaluated as a potential approach to the detection of DI. Because UDS can efficiently generate many more sequences than SGS, in this study it matched or exceeded the performance of SGS. For most samples, UDS also identified additional minority variants not present in the SGS results, which may be useful for inferring the evolutionary and population history of HIV populations. This degree of resolution was obtained because UDS produced over 500 reads for each of the sequences obtained by SGS in this study. Further, a single UDS run of 16 samples with three coding regions sequenced can also be performed in approximately a fifth of the time required to generate SGS for the same number of samples and only two coding regions. In our analyses, the SM-Index was 9.5 times cheaper than SGS, and UDS was 2.5 times cheaper than SGS per sample investigated (Table 3.3).

Shortcomings of the validation study include limited sample size, a large number of reads lost to gasketing, and a large number of low-quality reads that had to be excluded from the analysis. Furthermore, there was one sample whose SGS divergence exceeded its UDS divergence, despite the higher number of reads obtained by UDS. Sample H's anomalous results indicate that any DI screening technique must interrogate a sufficient number of input molecules to discern minority species in a representative manner. Samples with low viral loads may, therefore, require multiple replicates to compensate for initial template amplification bias, but a reliable viral load cut-off was not determined by this study. Samples C and I also had poorer coverage than the other samples, with about 50% fewer raw reads when compared to the others. This is somewhat unexpected, as all gasket-delineated regions should have the same read density, but perhaps demonstrates imperfections of the current UDS platform. One sample (C) also had a region with insufficient quality to infer DI. Nevertheless, this UDS run produced over 500 times more high-quality reads than the SGS procedures.

The higher sequencing volume and less time required for UDS might have other benefits in clarifying unresolved issues concerning HIV DI. For example, if UDS can identify superinfections sooner after the second transmission, when the new viral vari-

ants population is still low, then it may facilitate a more accurate determination of the incidence of superinfection. Taken together, these results demonstrate great promise in the use of UDS to confirm samples for DI, optionally preceded by a SM-Index screen. Especially because the per-base costs of existing and new UDS platforms are expected to continue decreasing and their accuracy and read lengths to continue increasing, we anticipate that UDS will eventually supplant SGS as the method of choice for dual infection screening.

Using the methods of SM-Index screening combined with SGS and UDS confirmation, we determined the overall DI prevalence to be 8.5%, which is over twice the previously identified prevalence for this study cohort [Smith Jama 04]. Nevertheless, 8.5% is likely a conservative estimate of DI, because 42 participants (an additional 36% of the study cohort) have suspected but unconfirmed DI. Our methods may also miss cases of transient superinfection, in which the second strain is completely replaced by the initial strain by the end of followup. The high percentage of DI cases identified by UDS to have small ($\leq$5%) minority populations highlights the need for sensitive DI screening and confirmation methods.

Chapter 3 is, in part, a reprint of the paper "Pacold M, Smith D, Little S, Cheng PM, Jordan P, Ignacio C, Richman D, Pond SK. Comparison of Methods to Detect HIV Dual Infection. *AIDS Res Hum Retroviruses*, 2010 Oct 18. [Epub ahead of print]." The dissertation author was the primary investigator and author of this paper.

Chapter 4

Clinical, Virologic, and Immunologic Correlates of HIV-1 Dual Infection

4.1 Introduction

Although it has been 16 years since the first cases of HIV-1 DI have been identified, many questions remain about its clinical, virologic, and immunologic correlates. On an individual level, identified DI has been associated with accelerated disease progression, including CD4 decline and time to AIDS diagnosis [23]. A few case reports of SI have identified a jump in viral load after acquisition of the second HIV-1 strain [1, 35, 69]. In the presence of DI, the interplay between the two strains can lead to various outcomes, including complete replacement of one viral strain by the other, transient presence of the second strain, low-level persistence of one strain, and production of recombinant populations. What influences these virologic dynamics and clinical consequences remains unclear but likely includes both host immune responses and viral characteristics, like replication capacity and the presence of immune epitopes [68]. We have performed a case-control study among well-characterized HIV-infected individuals followed since primary infection to investigate the virologic and clinical consequences of HIV-1 intraclade B dual infection and the impact of HLA haplotype on these findings.

4.2 Methods

4.2.1 Study Population

This study was approved by our local ethics committee included all participants of the San Diego Primary HIV Infection Program between January 1998 and January 2007 who had deferred antiretroviral therapy (ART) for at least the first 6 months after infection and had at least two blood samples available for DI screening. Male participants who were infected with HIV-1 subtype B and reported sex with men as an HIV risk behavior were screened. Blood plasma samples were aliquoted and stored at -80°C without previous thaws. All participants received baseline drug resistance testing (Geneseq, Monogram Biosciences, South San Francisco, CA USA). Estimated duration of infection (EDI) was calculated at baseline for each participant [68]. Each participant was HLA-A, B, and C haplotyped to two digits using collected blood samples. For those participants found to have HLA B35, we haplotyped to four digits. For the case control objective, we included participants in the CI and SI groups if they met CI or SI criteria as described below, and mono-infected (MI) controls in the cohort were matched to these CI and SI cases based on: 1) follow-up >6 months, 2) men who reported sex with other men as their HIV risk factor, 3) ultra-deep sequencing (UDS) or single genome sequencing (SGS) at a time point <1 year from the last date of follow-up in two or more HIV-1 coding regions that included *env*, 4) within-sample divergence cut-offs for *pol* and *gag* <2.5% and *env* <5%, and 5) phylogenetic structure of all regions indicating MI (lack of two or more genetically divergent populations supported by ≥95% bootstrap).

4.2.2 Sequencing Methods

Population-based *pol* sequences (HXB2 coordinates 2253-3554), SGS of *env* C2-V3 and *pol* RT, and UDS of *gag* p24 (HXB2 coordinates 1366-1619), *pol* RT (HXB2 coordinates 2708-3242), and *env* C2-V3 (HXB2 coordinates 6928-7344) were generated as previously described [48]. UDS reads were generated in batches of 16 samples physically separated with rubber gaskets on a 454 GS FLX Titanium picoliter plate (454 Life Sciences, a Roche company, Branford, CT). Read alignment and filtering were performed using a collection of scripts for the HyPhy software package [53] that selected high-quality UDS reads, generated consensus sequences, aligned reads to the consensus, and performed phylogenetic analysis of specific coding regions.

4.2.3 Dual Infection Screening and Confirmation

Dual infection was identified in a sample when the divergence of at least one coding region exceeded the DI thresholds of 5% for *env* and 2.5% for *pol* and *gag*, and when the phylogenetic structure of variants from that region supported DI (i.e. included two branches separated with bootstrap support ≥95%). Our algorithm for classifying a participant as MI, CI, or SI was:

1. Evaluate population-based pol sequences of baseline and last time point samples. If they are >5% divergent, then evaluate baseline, intervening, and last time point samples with UDS and/or SGS to investigate viral dynamics of DI, including CI and SI with and without viral replacement.

2. Perform UDS or SGS on the last available time point sample.

3. If at least 2 coding regions are successfully sequenced with UDS or SGS for the last time point sample, and none demonstrate DI, classify the participant as MI.

4. If a sample demonstrates DI and the minority population is <10% of the total, perform either SGS or UDS on a sample from the same date or within 3 months of it to confirm DI. If a sample demonstrates DI and the minority population is <10% of the total, perform UDS for confirmation.

5. If DI is not confirmed with additional sequencing at the same or a neighboring date, classify the participant as neither DI nor MI, i.e. ambiguous, and exclude from the case-control study.

6. If DI is confirmed and the sample is either <3 months from the EDI or is the first available sample at baseline, conservatively classify the participant as CI.

7. If DI is confirmed and the sample is not a baseline sample, perform either UDS or SGS on a baseline sample according to the size of the minority population, as described in step 4.

8. If DI is not observed in the baseline sample, classify the participant as SI and sequence further samples using a binary search approach to determine the timing of SI within a 1-year window.

9. Check DI sequences for contamination with BLAST and, in cases of complete sequence replacement, HLA typing at ≥2 time points to verify the sample identity.

10. Participants with baseline and last time point viral populations that were more than 5% divergent in V3, 2.5% in RT, or 2.5% in p24 but no evidence of mixed populations were evaluated for complete viral replacement by performing UDS and SGS

on intervening time point samples as available.

4.2.4 Clinical Consequences Analysis

All individual log viral load and square root CD4 progressions were plotted based on EDI. Similar to previous reports [4, 61] HIV-1 viral load dynamics of each group (MI, CI, and SI) were assessed with a linear mixed-effects model, using the nlme package in R. The infection group was included as an indicator variable, and both intercepts and slopes were estimated.

4.2.5 HLA Frequency and Linkage Disequilibrium

To assess the HLA allele frequencies in the MI, CI, and SI groups, we used the online HLA Analysis Tools[1]. The HLA Comparison tool provided HLA frequency graphs of two different populations (MI vs. SI, and MI vs. CI) and for each HLA comparison, it assigned a p-value as well as a q-value to account for false-positive results due to multiple testing. To assess for HLA Linkage Disequilibrium within each group, we used the HLA Linkage Disequilibrium tool, which uses the Fishers exact test to find statistically significant HLA pairs. Of note, linkages with a p-value of less than 0.05 were shown for MI and SI samples, but for CI samples, a p-value of less than 0.5 was used given the low power in the set due to small sample size.

4.2.6 CTL and Sequence Analysis

Epitope mapping: To study the CTL epitopes inside the *pol* and *env* sequences of DI participants, we used the consensus HIV HXB2 CTL epitope maps for *pol*/RT/env found at the Los Alamos National Laboratory website[2]. For superinfected individuals, we selected baseline and SI sample time-point SGS; for coinfected individuals, we selected a single sample time-point with the two different predominant virus variants in the SGS sampling. We compared the coding regions to the CTL epitope maps and identified all the epitopes, according to each participant's two-digit HLA haplotype.

Binding affinity: To predict binding affinities for each identified epitope, we entered each one into online prediction tools MHC-1 Binding Prediction[3] and NetMHC[4]

[1] http://www.hiv.lanl.gov/content/immunology/hla/
[2] http://www.hiv.lanl.gov/content/immunology/, accessed August 2010
[3] http://tools.immuneepitope.org/analyze/html/mhc_binding.html
[4] http://www.cbs.dtu.dk/services/NetMHC/

[41]. The binding affinity values were in units of half-maximal inhibitory concentration, IC_{50}nM, so a lower value indicates a higher affinity (peptides with IC_{50} values <50 are considered high affinity, <500 intermediate affinity, and <5000 low affinity). For epitopes that were 8, 10, or 11 amino acids, i.e. not 9, we used the online tools that allowed for affinity prediction of different combinations of length for a particular epitope, and we chose the peptide combination with the lowest (i.e. strongest) IC_{50}nM value. To investigate if binding affinities changed in relation to amino acid differences flanking putative epitopes, we used the online tool NetChop[5].

4.2.7 Recombination Analysis

Recombination breakpoint analysis of the complete RT and V3 sequence alignments for each sample point was performed using the online GARD (Genetic Algorithm Recombination Analysis) and single breakpoint tools[6] [54].

4.2.8 Selection Analysis

Positive and negative selection analysis of the complete RT and V3 sequence alignment for each sample viral population was performed using the SLAC (Single Likelihood Ancestor Counting) model of selection [57] to evaluate if observed amino acid changes in the DI viral populations demonstrated selection pressure. This provided an estimation of the ratio of non-synonymous substitutions per non-synonymous site (dN) to synonymous substitutions per synonymous site (dS). Although average dN/dS values for genes are uninformative for potential positively selected codons, they do provide a means of assessing adaptive evolutionary pressures at the gene level. A p-value of 0.05 was used for inference of positive and negative selection for individual codons.

4.3 Results

4.3.1 Participants

After screening 116 participants, 19 were identified with MI, 7 with SI, and 4 with CI. All were men who reported sex with other men as their initial and on-going HIV risk factor, and the average age at study enrollment was 31 years. The majority

[5]http://tools.immuneepitope.org/stools/netchop/netchop.do
[6]http://www.datamonkey.org/

Table 4.1: Entry clinical data of MI, CI, and SI groups.

	Monoinfected (N=19)	**Coinfected (N=4)**	**Superinfected (N=7)**
HIV viral load at enrollment: mean $\log_{10}$ copies/mL (range)	4.74 (1.70-7.14)	4.94 (2.89-5.42)	3.86 (2.25-6.36) [p>0.05]
CD4 at enrollment: mean cells/μL (range)	582 (210-1119)	622 (546-744)	623 (321-866)
HLA B35: N (%)	5 (26%)	0	5 (71%) [p=0.07]

were white (83%); 17% reported Hispanic ethnicity. There was no difference between the CI, SI, and MI groups based on baseline CD4 count (582 for MI, 622 for CI, 623 for SI), but there was a trend for the SI group to have a lower log10 viral load (3.86 for SI, 4.74 for MI, 4.94 for CI) [p>0.05] (Table 4.1).

4.3.2 Clinical Consequences

Longitudinal viral loads of participants were plotted over three years of follow-up for the MI, CI, and SI groups (Fig. 4.1). The SI group had a lower average viral load than CI and MI at baseline, but reached and overtook the other groups at approximately one year of followup. Compared to the MI group, the SI group had a significantly faster viral load increase over time (p=0.0004). The CI group also had a faster viral load increase over time than the MI group, but the difference did not attain statistical significance (p=0.06).

Longitudinal CD4 measurements were plotted over three years of follow-up to assess progression to the CD4 cell count falling below 400 cells/μL (Fig. 4.2). Too many participants elected to initiate antiretroviral therapy to assess the status of lower cell counts. Compared to MI, neither the CI nor the SI group had a significantly faster CD4 decline (p>0.05).

4.3.3 Virologic Consequences

Similar to previous reports, we increased the ability to detect DI by investigating more than one coding region [32, 52]. Together, UDS and SGS identified CI in 4

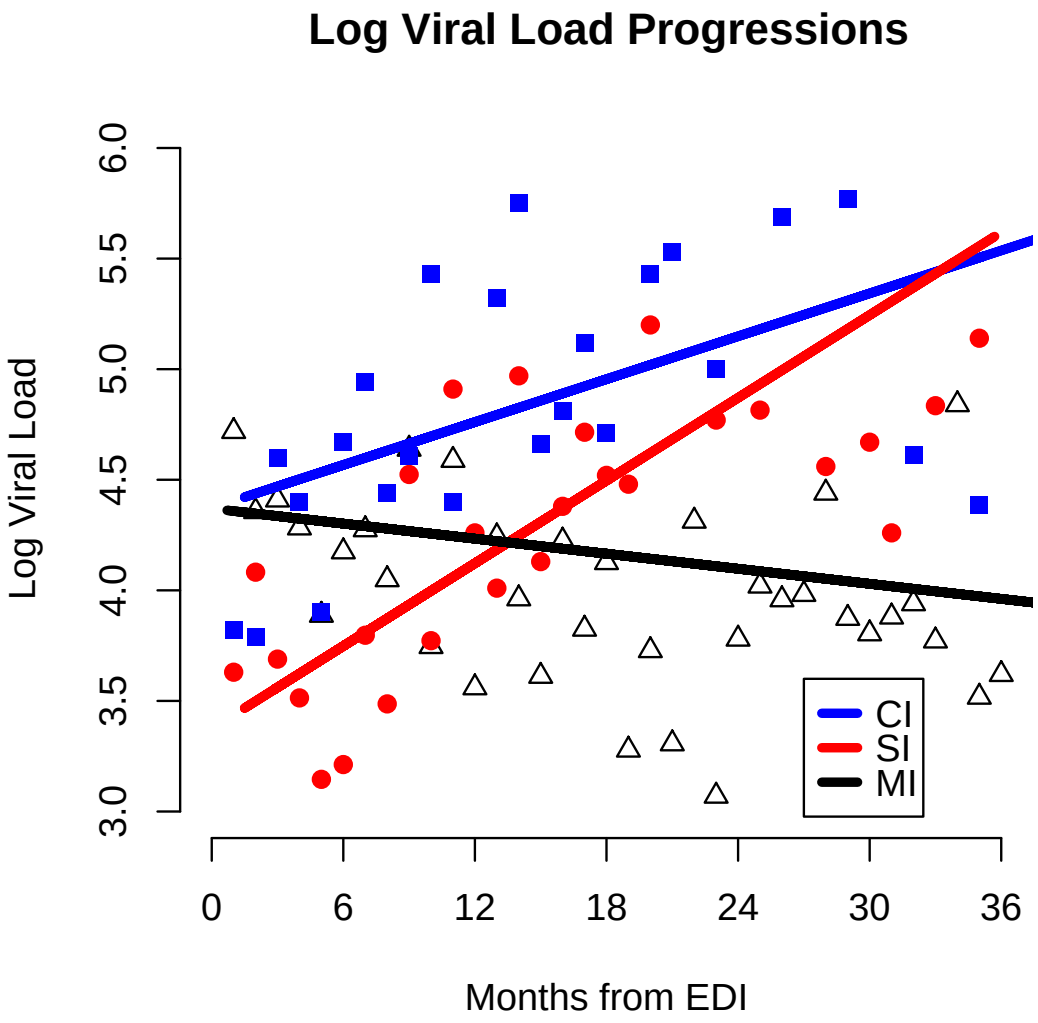

Figure 4.1: Viral load progressions for the 3 groups. EDI: estimated date of infection.

participants and SI in 7 participants. DI was visible in the *env* C2-V3 coding region at $\geq$1 time point for all 11 DI participants, in *pol* RT at $\geq$1 time point for 8/11 DI participants, and in *gag* p24 at $\geq$1 time point for 2/8 DI participants for which p24 was sequenced. From the 11 DI participants, we sequenced the C2-V3 region at a total of 49 time points, RT at a total of 44 time points, and p24 at a total of 11 time points. DI was discerned in 18/49 C2-V3 time points plus 9/49 complete replacements, in 18/44 RT time points plus 1/44 complete replacements, and in 2/11 p24 time points with no complete replacements. Of the 7 SI participants, 5 acquired the second strain during

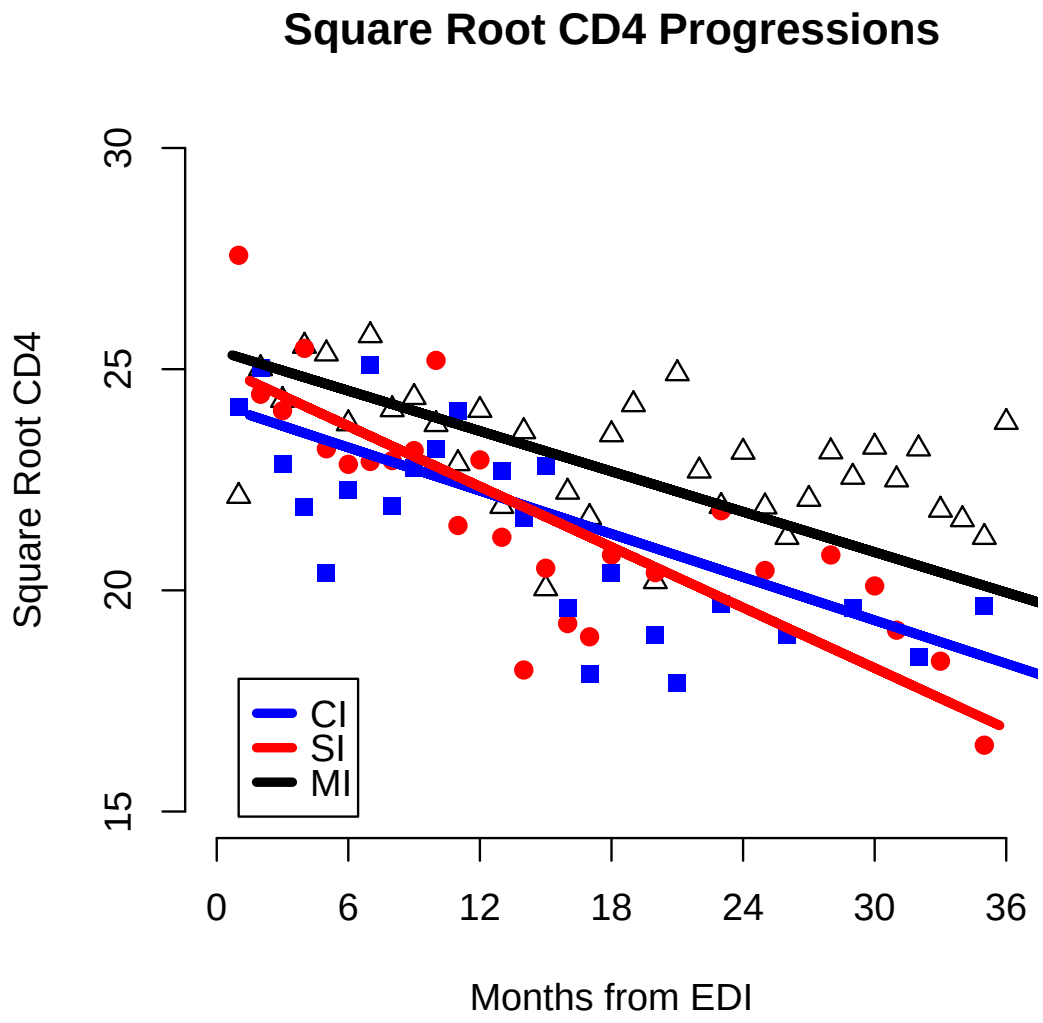

Figure 4.2: CD4 progressions for the 3 groups. EDI: estimated date of infection.

the first year of primary infection, and 2 acquired the second strain during the second year. UDS matched SGS in all cases in which both sequencing methods were applied to the same sample. Unlike many previous reports, most (71%) of the cases of DI that were identified in this study demonstrated transient DI. Two individuals demonstrated complete replacement of their viral populations in RT and C2-V3; these two cases have previously been reported [69].

Given the propensity of HIV-1 to recombine in the setting of DI [36], we examined recombination within sampled viral populations among plasma samples with DI.

We screened the C2-V3 and RT coding regions from all samples sequenced after DI [54]. Recombination was not detected for any sample in the C2-V3 coding region (n=27 samples), and breakpoints were detected in the RT coding region of three of the five SI cases that were identified in the RT coding region. Fig. 4.3 shows the initial, superinfecting, and recombinant populations of SI participant D224 in the RT coding region. Interestingly, identified breakpoints occurred within eight amino acids of each other among the three cases. These breakpoints did not appreciably change putative CTL epitopes for the individual.

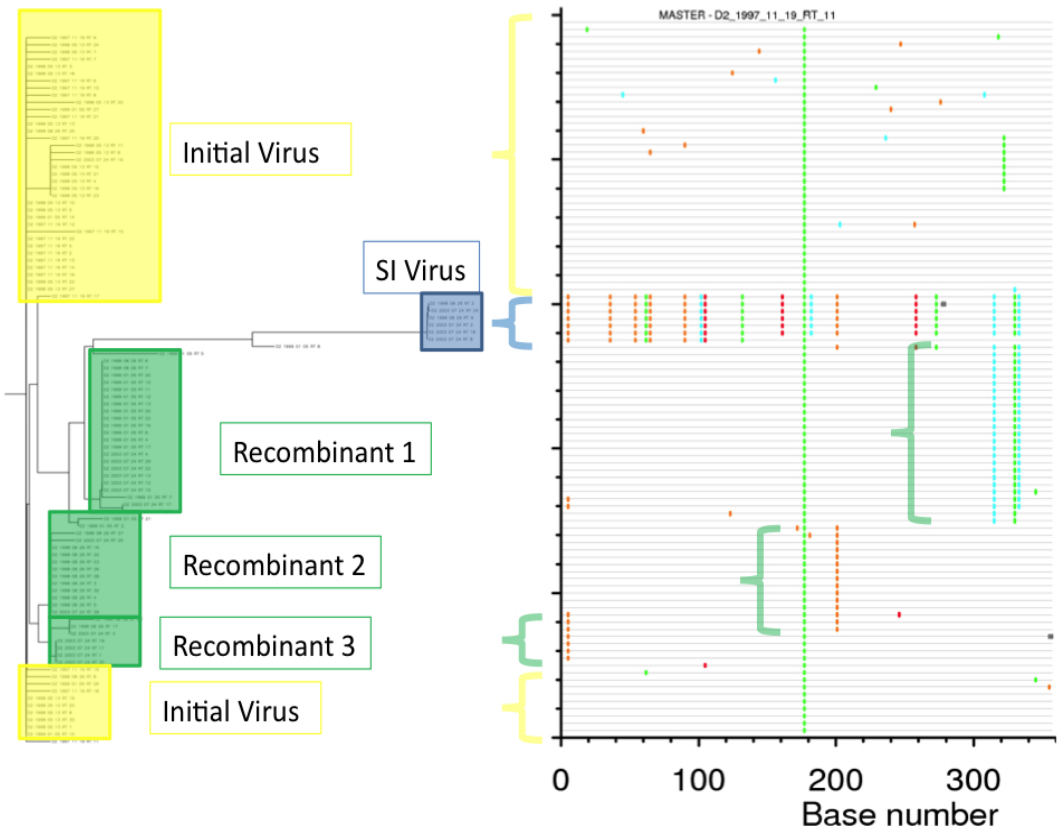

Figure 4.3: Phylogeny of D224 RT longitudinal time point sequences and visualization of recombination breakpoints.

4.3.4 CTL protection

If the CTL immune response influences protection from SI, then the superinfecting strain may display evidence of escape in putative CTL epitopes. For the five SI cases that demonstrated incomplete or transient changes in their viral population after

SI, amino acid differences between the initial and second viruses were no more likely to be inside putative CTL epitopes than outside CTL epitopes (Table 4.2). However, two SI cases (K613 and K908) demonstrated complete replacement of the *pol* and V3 coding regions following SI, and for both of these cases, the amino acid residues conferring CTL escape were observed based on estimations of the differences in binding affinities between initial and SI viruses. Evaluation of the amino acid changes 3-5 residues upstream or downstream that might influence epitope cleavage detected no consistent pattern for any subject in the SI group (data not shown).

Table 4.2: Comparison of amino acid differences inside vs. outside epitopes between initial and superinfecting viruses. *p-value<0.05.

Subject	**C2-V3**	**RT**	***pol***
K613	In=Out	In>Out*	In>Out*
K908	Out>In	In=Out	In>Out*
D224	Out>In	Out>In	NA
P265	In=Out	NA	NA
P853	In=Out	In=Out	NA
S155	In=Out	In=Out	NA
U796	In=Out	In=Out	NA

Previous studies have analyzed the association of HLA supertypes with rates of HIV disease progression. Although the associations vary according to factors like population ethnicity, HLA B27, B57, and B58 have been associated with slower disease progression [38, 71, 73], while haplotypes B7 and B35 have been associated with faster progression [12, 38, 73] and differential targeting of CTL epitopes during primary HIV infection [70]. A comparison of HLA frequencies in our study cohort revealed that the SI group had trends (p-value range 0.07-0.17) for higher frequencies of A29, C16, B35 and DRB1-07 and less DRB1-11 and C05 than the MI group (Table 4.3). In comparison, the CI group had higher frequencies of A29, C02, C16 and DRB1-07 (p-value range 0.02-0.09) than the MI group (Table 4.4). The SI group had a trend for a higher frequency of HLA B35 and DRB1-07. These are linked haplotypes [28], are associated with faster HIV disease progression [21], and target epitopes less frequently during primary infection [70].

Although not reaching significance at $p<0.05$ when comparing the SI and MI groups, both groups (26 and 71%) had a significantly greater frequency of HLA B35 than would be expected for a similar population in the United States, independent of racial or ethnic group (European Caucasian 8%, African American 7%, Asian Pacific Islander 8% and Hispanic 15%) [42, 45].

Table 4.3: Comparison of HLA Frequencies: SI vs. MI. *B35 and DR07 were found to be in linkage disequilibrium.

HLA	**SI (%)**	**MI (%)**	**p-value**	**q-value**
A29	14	0	0.07	0.87
C16	14	0	0.07	0.87
DR11	0	21	0.09	0.66
B35*	36	13	0.11	0.55
DR07*	36	16	0.14	0.54
C05	0	16	0.17	0.63

Table 4.4: Comparison of HLA Frequencies: CI vs. MI.

HLA	**SI (%)**	**MI (%)**	**p-value**	**q-value**
A29	33	0	0.02	0.12
C02	33	0	0.02	0.12
C16	33	0	0.02	0.12
DR07	50	15	0.09	0.35

The two cases of SI with complete replacement of the viral populations (subjects K613 and K908) were the only two SI cases who lacked HLA B35. Since the type of HLA B35 (PX vs. PY) has been associated with peptide binding specificity and HIV disease progression [21], we compared those with HLA B35 (four-digit HLA haplotyping) in the SI group to those with HLA B35 in the MI group and found no difference in the frequency of the types of HLA B35 between the groups, although these numbers are very small (3 out of 5 of MI and 2 out of 5 of the SI subjects had PX B35).

4.4 Discussion

The clinical, virologic, and immunologic correlates of DI have been poorly characterized, largely due to insufficient numbers of subjects screened and characterized for MI, CI and SI. Understanding these correlates for intraclade DI is important because, although more difficult to identify because of the genetic similarity between viral variants, intraclade DI is likely more common than interclade DI given the propensity of HIV-1 clades to be distributed unevenly throughout the world. (For example, over 90% of HIV-infected people in the United States are infected with clade B [8], and thus if exposed to SI, will most likely be exposed to a second clade B virus.)

The clinical consequences of DI are most likely influenced by the immune capability and reactivity of the individual, and this study found that DI (both SI and CI) was associated with faster viral load increases than the viral load changes observed in MI controls. Interestingly, in this study the SI group had lower baseline viral loads than both CI and MI groups, but the significance of this finding remains unclear. Since virus-specific CTL immune responses that develop during primary HIV infection are responsible for the earliest control of viral replication [6, 10, 11, 37, 47, 70] and viral set-point [40], we investigated if there was evidence of CTL escape in the sequences that were different during DI or after SI but found no evidence for CTL pressure and viral escape in any of the participants demonstrating transient DI. However, escape was identified in two participants who had replaced the two evaluated coding regions (RT and C2-V3) completely. Interestingly, these two participants with complete viral replacement were the only ones with SI who did not have HLA B35. Overall, this weak evidence suggests that CTL responses that develop during HIV infection may protect from some SI exposures, but if the SI virus has existing residues in epitopes that allow escape from the immune responses to the initial virus, then the SI virus replaces the initial virus, at least in the coding regions containing these residues allowing escape. This study also suggests that among individuals with HLA haplotypes that develop later during the course of HIV infection or are associated with less immune responses, then the SI virus may replace the initial virus, at least in the coding regions containing these residues allowing escape. This study also suggests that individuals with HLA haplotypes that develop later during the course of HIV infection or are associated with less CTL control of HIV infection, like B35, may be more susceptible to SI, but again, these pilot observations require evaluation in larger cohort studies.

There are several limitations to the current study. Although unavoidable in a case-control study design, there could be a bias in selecting the MI controls. Potentially, these controls may not adequately represent the natural history of HIV-1 MI, and the inclusion criteria associated with the "confirmation" of MI using sequencing methods could cause a systematic selection bias in the selection of the controls, since it is impossible to rule out that DI never took place. This study is also limited in that it only represents the men who have sex with men risk group in San Diego, California. The confirmation of DI may also be biased towards the detection of those DI individuals who have distinct viral populations that comprise a certain level of co-circulation or where the viral population has been completely replaced over time, and we may have missed DI if co-circulation existed at a time point that was not sampled, a sample not interrogated, or was not confirmed by an additional confirmation method. Our methods for confirmation of DI are relatively conservative and aimed to limit the number of false positives for DI due to a laboratory mistake, i.e. sample mix-up or contamination. Since this study used an observational cohort to select cases and controls, and was not a controlled trial, there is potential confounding by the variability in the initiation of ART based on decisions of patient and health care provider. In addition, only selected regions of the virus were sequenced, which restricts the information regarding both CTL escape and recombination.

Chapter 4 is, in part, a reprint of the manuscript in preparation "Pacold ME, Pond SK, Wagner GA, Delport W, Bourque DL, Richman DD, Little SJ, Smith DM. Clinical, Virologic, and Immunologic Correlates of HIV-1 Dual Infection." The dissertation author was the primary investigator and author of this paper.

Chapter 5

Conclusions

The methods developed in this proposal were used to identify significant numbers of occurrences of intraclade HIV dual infection. At the completion of these studies, we: 1) developed and validated novel methods for the screening and confirmation of intraclade HIV dual infection, 2) identified and characterized the largest number of instances of intraclade dual infection to date, 3) determined the incidence of HIV superinfection and prevalence of coinfection in the largest well-characterized cohort of recently infected individuals, and 4) determined the clinical consequences of dual infection in this California cohort. The methods developed for detection of dual infection in these studies are more sensitive, higher-throughput, and more cost-effective than those of previous studies. They can be readily applied to additional study cohorts.

Based on the results described in Chapter 3, we conclude that ultra-deep sequencing is a suitable DI screening method for large cohorts and that it may become the method of choice for similar studies. Application of this method to a high-exposure study cohort of San Diego participants revealed a DI prevalence including our newly identified intraclade SI cases that likely exceeds 10%, well above the prevalence of 4% previously estimated for this cohort. The studies described in Chapter 4 show that HLA haplotype may influence susceptibility to SI and changes in the viral population after SI, though these findings will be better validated when additional cases of SI are identified.

On a clinical level, the negative clinical consequences of DI demonstrated in Chapter 4 indicate that serosorting (the practice of choosing partners to engage in unprotected behaviors according to similar HIV serostatus) should be discouraged, as it opens HIV-infected people to the possibility of SI and accompanying faster disease pro-

gression.

Bibliography

[1] M Altfeld, T Allen, X Yu, M Johnston, D Agrawal, B Korber, D Montefiori, D O'Connor, B Davis, P Lee, E Maier, J Harlow, P Goulder, C Brander, E Rosenberg, and B Walker. Hiv-1 superinfection despite broad cd8+ t-cell responses containing replication of the primary virus. *Nature*, 420(6914):434–9, Nov 2002.

[2] J Archer, M Braverman, B Taillon, B Desany, I James, P Harrigan, M Lewis, and D Robertson. Detection of low-frequency pretherapy chemokine (cxc motif) receptor 4 (cxcr4)-using hiv-1 with ultra-deep pyrosequencing. *AIDS*, 23(10):1209–18, Jun 2009.

[3] A Artenstein, T VanCott, J Mascola, J Carr, P Hegerich, J Gaywee, E Sanders-Buell, M Robb, D Dayhoff, S Thitivichianlert, and et al. Dual infection with human immunodeficiency virus type 1 of distinct envelope subtypes in humans. *J Infect Dis*, 171(4):805–10, Apr 1995.

[4] Arman A Bashirova, Gabriela Bleiber, Ying Qi, Holli Hutcheson, Traci Yamashita, Randall C Johnson, Jie Cheng, Galit Alter, James J Goedert, Susan Buchbinder, Keith Hoots, David Vlahov, Margaret May, Frank Maldarelli, Lisa Jacobson, Stephen J O'brien, Amalio Telenti, and Mary Carrington. Consistent effects of tsg101 genetic variability on multiple outcomes of exposure to human immunodeficiency virus type 1. *Journal of Virology*, 80(14):6757–63, Jul 2006.

[5] C Blish, O Dogan, N Derby, M Nguyen, B Chohan, B Richardson, and J Overbaugh. Hiv-1 superinfection occurs despite relatively robust neutralizing antibody responses. *J Virol*, Oct 2008. Journal article.

[6] P Borrow, H Lewicki, B H Hahn, G M Shaw, and M B Oldstone. Virus-specific cd8+ cytotoxic t-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. *J Virol*, 68(9):6103–10, Sep 1994.

[7] A Bruselles, G Rozera, B Bartolini, M Prosperi, F Del Nonno, P Narciso, M Capobianchi, and I Abbate. Use of massive parallel pyrosequencing for near full-length characterization of a unique hiv type 1 bf recombinant associated with a fatal primary infection. *AIDS Res Hum Retroviruses*, 25(9):937–42, Sep 2009.

[8] Isolde F Butler, Ivona Pandrea, Preston A Marx, and Cristian Apetrei. Hiv genetic diversity: biological and public health consequences. *Curr HIV Res*, 5(1):23–45, Jan 2007.

[9] M Campbell, G Gottlieb, S Hawes, D Nickle, K Wong, W Deng, T Lampinen, N Kiviat, and J Mullins. Hiv-1 superinfection in the antiretroviral therapy era: are seroconcordant sexual partners at risk? *PLoS One*, 4(5):e5690, 2009.

[10] Jianhong Cao, John McNevin, Sarah Holte, Lisa Fink, Lawrence Corey, and M Juliana McElrath. Comprehensive analysis of human immunodeficiency virus type 1 (hiv-1)-specific gamma interferon-secreting cd8+ t cells in primary hiv-1 infection. *Journal of Virology*, 77(12):6867–78, Jun 2003.

[11] Jianhong Cao, John McNevin, Uma Malhotra, and M Juliana McElrath. Evolution of cd8+ t cell immunity and viral escape following acute hiv-1 infection. *J Immunol*, 171(7):3837–46, Oct 2003.

[12] M Carrington, G W Nelson, M P Martin, T Kissner, D Vlahov, J J Goedert, R Kaslow, S Buchbinder, K Hoots, and S J O'Brien. Hla and hiv-1: heterozygote advantage and b*35-cw*04 disadvantage. *Science*, 283(5408):1748–52, Mar 1999.

[13] Charlotte Charpentier, Tamara Nora, Olivier Tenaillon, François Clavel, and Allan J Hance. Extensive recombination among human immunodeficiency virus type 1 quasispecies makes an important contribution to viral diversity in individual patients. *J Virol*, 80(5):2472–82, Mar 2006.

[14] B Chohan, L Lavreys, S Rainwater, and J Overbaugh. Evidence for frequent reinfection with human immunodeficiency virus type 1 of a different subtype. *J Virol*, 79(16):10701–8, Aug 2005.

[15] M Cornelissen, S Jurriaans, K Kozaczynska, J Prins, R Hamidjaja, F Zorgdrager, M Bakker, N Back, and A van der Kuyl. Routine hiv-1 genotyping as a tool to identify dual infections. *AIDS*, 21(7):807–11, Apr 2007. Journal Article England.

[16] R Diaz, E Sabino, A Mayer, J Mosley, and M Busch. Dual human immunodeficiency virus type 1 infection and recombination in a dually exposed transfusion recipient. the transfusion safety study group. *J Virol*, 69(6):3273–81, Jun 1995.

[17] N Eriksson, L Pachter, Y Mitsuya, S Rhee, C Wang, B Gharizadeh, M Ronaghi, R Shafer, and N Beerenwinkel. Viral population estimation using pyrosequencing. *PLoS Comput Biol*, 4(4):e1000074, Apr 2008.

[18] G Fang, B Weiser, C Kuiken, S Philpott, S Rowland-Jones, F Plummer, J Kimani, B Shi, R Kaul, J Bwayo, O Anzala, and H Burger. Recombination following superinfection by hiv-1. *AIDS*, 18(2):153–9, Jan 2004.

[19] G Fang, G Zhu, H Burger, J S Keithly, and B Weiser. Minimizing dna recombination during long rt-pcr. *J Virol Methods*, 76(1-2):139–48, Dec 1998.

[20] P Fultz, A Srinivasan, C Greene, D Butler, R Swenson, and H McClure. Superinfection of a chimpanzee with a second strain of human immunodeficiency virus. *J Virol*, 61(12):4026–9, Dec 1987.

[21] X Gao, G W Nelson, P Karacki, M P Martin, J Phair, R Kaslow, J J Goedert, S Buchbinder, K Hoots, D Vlahov, S J O'Brien, and M Carrington. Effect of a single amino acid change in mhc class i molecules on the rate of progression to aids. *N Engl J Med*, 344(22):1668–75, May 2001.

[22] M Gonzales, E Delwart, S Rhee, R Tsui, A Zolopa, J Taylor, and R Shafer. Lack of detectable human immunodeficiency virus type 1 superinfection during 1072 person-years of observation. *J Infect Dis*, 188(3):397–405, Aug 2003.

[23] G Gottlieb, D Nickle, M Jensen, K Wong, J Grobler, F Li, S Liu, C Rademeyer, G Learn, S Karim, C Williamson, L Corey, J Margolick, and J Mullins. Dual hiv-1 infection associated with rapid disease progression. *Lancet*, 363(9409):619–22, Feb 2004.

[24] G Gottlieb, D Nickle, M Jensen, K Wong, R Kaslow, J Shepherd, J Margolick, and J Mullins. Hiv type 1 superinfection with a dual-tropic virus and rapid progression to aids: a case report. *Clin Infect Dis*, 45(4):501–9, Aug 2007.

[25] Robert Grant, J McConnell, J Marcus, G Spotts, T Liegler, R Brennan, and F Hecht. High frequency of apparent hiv-1 superinfection in a seroconverter cohort. *Conference on Retroviruses and Opportunistic Infections*, 12, 2005.

[26] J Grobler, C Gray, C Rademeyer, C Seoighe, G Ramjee, S Karim, L Morris, and C Williamson. Incidence of hiv-1 dual infection and its association with increased viral load set point in a cohort of hiv-1 subtype c-infected female sex workers. *J Infect Dis*, 190(7):1355–9, Oct 2004.

[27] K Gross, T Porco, and R Grant. Hiv-1 superinfection and viral diversity. *AIDS*, 18(11):1513–20, Jul 2004.

[28] C Grundschober, A Sanchez-Mazas, L Excoffier, A Langaney, M Jeannet, and J M Tiercy. Hla-dpb1 dna polymorphism in the swiss population: linkage disequilibrium with other hla loci and population genetic affinities. *Eur J Immunogenet*, 21(3):143–57, Jun 1994.

[29] H F Günthard, J K Wong, C C Ignacio, D V Havlir, and D D Richman. Comparative performance of high-density oligonucleotide sequencing and dideoxynucleotide sequencing of hiv type 1 pol from clinical samples. *AIDS Res Hum Retroviruses*, 14(10):869–76, Jul 1998.

[30] C A Heid, J Stevens, K J Livak, and P M Williams. Real time quantitative pcr. *Genome Res*, 6(10):986–94, Oct 1996.

[31] K Herbinger, M Gerhardt, S Piyasirisilp, D Mloka, M Arroyo, O Hoffmann, L Maboko, D Birx, D Mmbando, F McCutchan, and M Hoelscher. Frequency of hiv type 1 dual infection and hiv diversity: analysis of low- and high-risk populations in mbeya region, tanzania. *AIDS Res Hum Retroviruses*, 22(7):599–606, Jul 2006.

[32] M Hoelscher, W Dowling, E Sanders-Buell, J Carr, M Harris, A Thomschke, M Robb, D Birx, and F McCutchan. Detection of hiv-1 subtypes, recombinants,

and dual infections in east africa by a multi-region hybridization assay. *AIDS*, 16(15):2055–64, Oct 2002.

[33] C Hoffmann, N Minkah, J Leipzig, G Wang, M Arens, P Tebas, and F Bushman. Dna bar coding and pyrosequencing to identify rare hiv drug resistance mutations. *Nucleic Acids Res*, 35(13):e91, 2007.

[34] D Hu, S Subbarao, S Vanichseni, P Mock, A Ramos, L Nguyen, T Chaowanachan, F Griensven, K Choopanya, T Mastro, and J Tappero. Frequency of hiv-1 dual subtype infections, including intersubtype superinfections, among injection drug users in bangkok, thailand. *AIDS*, 19(3):303–8, Feb 2005.

[35] S Jost, M Bernard, L Kaiser, S Yerly, B Hirschel, A Samri, B Autran, L Goh, and L Perrin. A patient with hiv-1 superinfection. *N Engl J Med*, 347(10):731–6, Sep 2002.

[36] G Kijak and F McCutchan. Hiv diversity, molecular epidemiology, and the role of recombination. *Curr Infect Dis Rep*, 7(6):480–8, Nov 2005.

[37] R A Koup, J T Safrit, Y Cao, C A Andrews, G McLeod, W Borkowsky, C Farthing, and D D Ho. Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. *J Virol*, 68(7):4650–5, Jul 1994.

[38] Aleksandr Lazaryan, Wei Song, Elena Lobashevsky, Jianming Tang, Sadeep Shrestha, Kui Zhang, Lytt I Gardner, Janet M McNicholl, Craig M Wilson, Robert S Klein, Anne Rompalo, Kenneth Mayer, Jack Sobel, Richard A Kaslow, HIV Epidemiology Research Study, Reaching for Excellence in Adolescent Care, and Health Study. Human leukocyte antigen class i supertypes and hiv-1 control in african americans. *Journal of Virology*, 84(5):2610–7, Mar 2010.

[39] T Le, J Chiarella, B Simen, B Hanczaruk, M Egholm, M Landry, K Dieckhaus, M Rosen, and M Kozal. Low-abundance hiv drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS One*, 4(6):e6079, 2009.

[40] Mathias Lichterfeld, Xu G Yu, Stanley K Mui, Katie L Williams, Alicja Trocha, Mark A Brockman, Rachel L Allgaier, Michael T Waring, Tomohiko Koibuchi, Mary N Johnston, Daniel Cohen, Todd M Allen, Eric S Rosenberg, Bruce D Walker, and Marcus Altfeld. Selective depletion of high-avidity human immunodeficiency virus type 1 (hiv-1)-specific cd8+ t cells after early hiv-1 infection. *Journal of Virology*, 81(8):4199–214, Apr 2007.

[41] Claus Lundegaard, Kasper Lamberth, Mikkel Harndahl, Søren Buus, Ole Lund, and Morten Nielsen. 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(Web Server issue):W509–12, Jul 2008.

[42] Martin Maiers, Loren Gragert, and William Klitz. High-resolution hla alleles and haplotypes in the united states population. *Hum Immunol*, 68(9):779–88, Sep 2007.

[43] O Manigart, V Courgnaud, O Sanou, D Valea, N Nagot, N Meda, E Delaporte, M Peeters, and P Van de Perre. Hiv-1 superinfections in a cohort of commercial sex workers in burkina faso as assessed by an autologous heteroduplex mobility procedure. *AIDS*, 18(12):1645–51, Aug 2004.

[44] A Meyerhans, J P Vartanian, and S Wain-Hobson. Dna recombination during pcr. *Nucleic Acids Res*, 18(7):1687–91, Apr 1990.

[45] D Middleton, L Menchaca, H Rood, and R Komerofsky. New allele frequency database: http://www.allelefrequencies.net. *Tissue Antigens*, 61(5):403–7, May 2003.

[46] D Nickle, L Heath, M Jensen, P Gilbert, J Mullins, and S Kosakovsky Pond. Hiv-specific probabilistic models of protein evolution. *PLoS One*, 2(6):e503, 2007.

[47] Annette Oxenius, David A Price, Alexandra Trkola, Charles Edwards, Emma Gostick, Hua-Tang Zhang, Philippa J Easterbrook, Tin Tun, Andrew Johnson, Anele Waters, Edward C Holmes, and Rodney E Phillips. Loss of viral control in early hiv-1 infection is temporally associated with sequential escape from cd8+ t cell responses and decrease in hiv-1-specific cd4+ and cd8+ t cell frequencies. *J Infect Dis*, 190(4):713–21, Aug 2004.

[48] Mary Pacold, Davey Smith, Susan Little, Pok Man Cheng, Parris Jordan, Caroline Ignacio, Douglas Richman, and Sergei Kosakovsky Pond. Comparison of methods to detect hiv dual infection. *AIDS research and human retroviruses*, Oct 2010.

[49] S Palmer, M Kearney, F Maldarelli, E Halvas, C Bixby, H Bazmi, D Rock, J Falloon, R Davey, R Dewar, J Metcalf, S Hammer, J Mellors, and J Coffin. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. *J Clin Microbiol*, 43(1):406–13, Jan 2005.

[50] A Piantadosi, B Chohan, V Chohan, R McClelland, and J Overbaugh. Chronic hiv-1 infection frequently fails to protect against superinfection. *PLoS Pathog*, 3(11):e177, Nov 2007.

[51] A Piantadosi, B Chohan, D Panteleeff, J Baeten, K Mandaliya, J Ndinya-Achola, and J Overbaugh. Hiv-1 evolution in gag and env is highly correlated but exhibits different relationships with viral load and the immune response. *AIDS*, 23(5):579–87, Mar 2009.

[52] A Piantadosi, M Ngayo, B Chohan, and J Overbaugh. Examination of a second region of the hiv type 1 genome reveals additional cases of superinfection. *AIDS Res Hum Retroviruses*, 24(9):1221, Sep 2008.

[53] S Pond, S Frost, and S Muse. Hyphy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676–9, Mar 2005.

[54] S Kosakovsky Pond, D Posada, M Gravenor, C Woelk, and S Frost. Gard: a genetic algorithm for recombination detection. *Bioinformatics*, 22(24):3096–8, Dec 2006.

[55] S Kosakovsky Pond and D Smith. Are all subtypes created equal? the effectiveness of antiretroviral therapy against non-subtype b hiv-1. *Clin Infect Dis*, 48(9):1306–9, May 2009.

[56] S Kosakovsky Pond, S Wadhawan, F Chiaromonte, G Ananda, W Chung, J Taylor, and A Nekrutenko. Windshield splatter analysis with the galaxy metagenomic pipeline. *Genome Res*, 19(11):2144–53, Nov 2009.

[57] Sergei Pond and Simon Frost. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol*, 22(5):1208–1222, 2005.

[58] Art Poon, Sergei Pond, Phil Bennett, Douglas Richman, Andrew Brown, and Simon Frost. Adaptation to human populations is revealed by within-host polymorphisms in hiv-1 and hepatitis c virus. *PLoS Pathog*, 3(3):e45, 2007.

[59] A Ramos, D Hu, L Nguyen, K Phan, S Vanichseni, N Promadej, K Choopanya, M Callahan, N Young, J McNicholl, T Mastro, T Folks, and S Subbarao. Intersubtype human immunodeficiency virus type 1 superinfection following seroconversion to primary infection in two injection drug users. *J Virol*, 76(15):7444–52, Aug 2002.

[60] Christine M Rousseau, Ruth W Nduati, Barbra A Richardson, Grace C John-Stewart, Dorothy A Mbori-Ngacha, Joan K Kreiss, and Julie Overbaugh. Association of levels of hiv-1-infected breast milk cells and risk of mother-to-child transmission. *J Infect Dis*, 190(10):1880–8, Nov 2004.

[61] Manish Sagar, Erin Kirkegaard, E Long, Connie Celum, Susan Buchbinder, Eric Daar, and Julie Overbaugh. Human immunodeficiency virus type 1 (hiv-1) diversity at time of infection is not restricted to certain risk groups or specific hiv-1 subtypes. *J Virol*, 78(13):7279–7283, 2004.

[62] M Sala, G Zambruno, J Vartanian, A Marconi, U Bertazzoni, and S Wain-Hobson. Spatial discontinuities in human immunodeficiency virus type 1 quasispecies derived from epidermal langerhans cells of a patient with aids and evidence for double infection. *J Virol*, 68(8):5280–3, Aug 1994.

[63] J Salazar-Gonzalez, E Bailes, K Pham, M Salazar, M Guffey, B Keele, C Derdeyn, P Farmer, E Hunter, S Allen, O Manigart, J Mulenga, J Anderson, R Swanstrom, B Haynes, G Athreya, B Korber, P Sharp, G Shaw, and B Hahn. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. *J Virol*, 82(8):3952–70, Apr 2008.

[64] Daniel Shriner, Allen G Rodrigo, David C Nickle, and James I Mullins. Pervasive genomic recombination of hiv-1 in vivo. *Genetics*, 167(4):1573–83, Aug 2004.

[65] P Simmonds, P Balfe, C A Ludlam, J O Bishop, and A J Brown. Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *J Virol*, 64(12):5840–50, Dec 1990.

[66] P Simmonds, P Balfe, J F Peutherer, C A Ludlam, J O Bishop, and A J Brown. Human immunodeficiency virus-infected individuals contain provirus in small numbers

of peripheral mononuclear cells and at low copy numbers. *J Virol*, 64(2):864–72, Feb 1990.

[67] Davey Smith, Douglas Richman, and Susan Little. Hiv superinfection. *J Infect Dis*, 192(3):438–444, 2005.

[68] Davey Smith, Matthew Strain, Simon Frost, Satish Pillai, Joseph Wong, Terri Wrin, Yang Liu, Christos Petropolous, Eric Daar, Susan Little, and Douglas Richman. Lack of neutralizing antibody response to hiv-1 predisposes to superinfection. *Virology*, 355(1):1–5, 2006.

[69] Davey Smith, Joseph Wong, George Hightower, Caroline Ignacio, Kersten Koelsch, Eric Daar, Douglas Richman, and Susan Little. Incidence of hiv superinfection following primary infection. *JAMA*, 292(10):1177–1178, 2004.

[70] Hendrik Streeck, Jonathan S Jolin, Ying Qi, Bader Yassine-Diab, Randall C Johnson, Douglas S Kwon, Marylyn M Addo, Chanson Brumme, Jean-Pierre Routy, Susan Little, Heiko K Jessen, Anthony D Kelleher, Frederick M Hecht, Rafick-Pierre Sekaly, Eric S Rosenberg, Bruce D Walker, Mary Carrington, and Marcus Altfeld. 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. *Journal of Virology*, 83(15):7641–8, Aug 2009.

[71] Jianming Tang, Rakhi Malhotra, Wei Song, Ilene Brill, Liangyuan Hu, Paul K Farmer, Joseph Mulenga, Susan Allen, Eric Hunter, and Richard A Kaslow. Human leukocyte antigens and hiv type 1 viral load in early and chronic infection: predominance of evolving relationships. *PLoS ONE*, 5(3):e9629, Jan 2010.

[72] Alan R Templeton, Melissa G Kramer, Joseph Jarvis, Jeanne Kowalski, Stephen Gange, Michael F Schneider, Qiujia Shao, Guang Wen Zhang, Mei-Fen Yeh, Hua-Ling Tsai, Hong Zhang, and Richard B Markham. Multiple-infection and recombination in hiv-1 within a longitudinal cohort of women. *Retrovirology*, 6:54, Jan 2009.

[73] Elizabeth Trachtenberg, Bette Korber, Cristina Sollars, Thomas B Kepler, Peter T Hraber, Elizabeth Hayes, Robert Funkhouser, Michael Fugate, James Theiler, Yen S Hsu, Kevin Kunstman, Samuel Wu, John Phair, Henry Erlich, and Steven Wolinsky. Advantage of rare hla supertype in hiv disease progression. *Nature Medicine*, 9(7):928–35, Jul 2003.

[74] A Tsibris, B Korber, R Arnaout, C Russ, C Lo, T Leitner, B Gaschen, J Theiler, R Paredes, Z Su, M Hughes, R Gulick, W Greaves, E Coakley, C Flexner, C Nusbaum, and D Kuritzkes. Quantitative deep sequencing reveals dynamic hiv-1 escape and large population shifts during ccr5 antagonist therapy in vivo. *PLoS One*, 4(5):e5683, 2009.

[75] R Tsui, B Herring, J Barbour, R Grant, P Bacchetti, A Kral, B Edlin, and E Delwart. Human immunodeficiency virus type 1 superinfection was not detected following 215 years of injection drug user exposure. *J Virol*, 78(1):94–103, Jan 2004.

[76] C Wang, Y Mitsuya, B Gharizadeh, M Ronaghi, and R Shafer. Characterization of mutation spectra with ultra-deep pyrosequencing: application to hiv-1 drug resistance. *Genome Res*, 17(8):1195–201, Aug 2007.

[77] G Wang, A Ciuffi, J Leipzig, C Berry, and F Bushman. Hiv integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res*, 17(8):1186–94, Aug 2007.

[78] K Q Xin, X H Ma, K A Crandall, H Bukawa, Y Ishigatsubo, S Kawamoto, and K Okuda. Dual infection with hiv-1 thai subtype b and e. *Lancet*, 346(8986):1372–3, Nov 1995.

[79] Y L Yang, G Wang, K Dorman, and A H Kaplan. Long polymerase chain reaction amplification of heterogeneous hiv type 1 templates produces recombination at a relatively high frequency. *AIDS Res Hum Retroviruses*, 12(4):303–6, Mar 1996.

[80] S Yerly, S Jost, M Monnat, A Telenti, M Cavassini, J Chave, L Kaiser, P Burgisser, and L Perrin. Hiv-1 co/super-infection in intravenous drug users. *AIDS*, 18(10):1413–21, Jul 2004.

[81] Z Zhang, S Schwartz, L Wagner, and W Miller. A greedy algorithm for aligning dna sequences. *J Comput Biol*, 7(1-2):203–14, Feb 2000.

[82] T Zhu, N Wang, A Carr, S Wolinsky, and D Ho. Evidence for coinfection by multiple strains of human immunodeficiency virus type 1 subtype b in an acute seroconvertor. *J Virol*, 69(2):1324–7, Feb 1995.

[83] J Zhuang, A Jetzt, G Sun, H Yu, G Klarmann, Y Ron, B Preston, and J Dougherty. Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J Virol*, 76(22):11273–82, Nov 2002.