UNIVERSITY OF CALIFORNIA, SAN DIEGO

Effects of Anatomic Compartmentalization on HIV-1 Evolution

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Biology

by

Satish Kumar Pillai

Committee in charge:

Professor Douglas D. Richman, Chair
Professor Lin Chao
Professor John Guatelli
Professor Mitchell Kronenberg
Professor Chris Wills
Professor Joseph Wong

2005

The dissertation of Satish Kumar Pillai is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2005

DEDICATION

This work is dedicated to my Mother, Father, and Sister, who have always supported me unfalteringly in absolutely everything I do, and to the countless millions who have suffered at the hands of the AIDS epidemic.

"New knowledge is the most valuable commodity on earth.
The more truth we have to work with, the richer we become."
-- Kurt Vonnegut, Jr. (from *Breakfast of Champions*)

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Douglas Richman. Thanks for going out on a limb and taking a student under your wing who emerged from The Other Side (of Gilman Drive, of course). Doug, you are truly an inspiration to all scientists. Your knowledge of the field (HIV biology), and your capacity and motivation to spread that knowledge are unparalleled. Perhaps one the greatest benefits that I have enjoyed in your lab is the access to collaborators around the world; the Richman SuperGroup is tapped into an extensive network of investigators that have helped me considerably in my own research efforts. It has been an honor to work with someone who has been dealing with HIV in a clinical setting since before it even had a name of its own.

Next I would like to thank the UCSD CFAR (Center for AIDS Research) subset of my thesis committee: Drs. John Guatelli, and Joseph Wong. As far as I'm concerned, I have been lucky enough to have not one but THREE brilliant advisors watching over me during my graduate years here at UCSD. John, I will never be able to thank you enough for adopting me as an unofficial member of your Nef clan and giving me a grad school family to share in all the highs and lows that inevitably come with this kind of adventure (not to mention for giving me a clue about benchtop virology). For now, I'll just thank you "*in silico*". Joe, on top of being a top-notch mentor and (not-self-admitted) evolutionary biologist, you represent the extremely rare fusion of scientific brilliance and humility that I hope will rub off on me somehow. Well, the brilliance will do.

I am grateful to the remaining members of my committee, Drs. Lin Chao, Christopher Wills, and Mitch Kronenberg, for taking the time to offer me their invaluable insights into the worlds of population genetics, evolutionary virology, and immunology, respectively.

Now to extend my heartfelt thanks to my aforementioned graduate foster family, the Guatelli Lab: Scott ("Scootie and the Blowfish") Coleman, John ("Juan Dia") Day, Colleen ("Natch") Noviello, Megan ("The Meganator") Dueck, Nanette ("Sweet. . .") Van Damme, Doug ("Doog") Hitchin, and Rick ("no nickname as yet") Mitchell. An extra special shout-out to Juan Dia, Doog, and The Meganator: Juan, thanks to your wisdom, patience, and kindness, I am now compatible with the Microsoft-saturated 21st century. Doog, thanks to your shutterbug nature, I now have volumes of photographs and videos documenting my musical pursuits over the last half-decade. Meganator, you were instrumental in helping me charge through the finish line with this project (and you enabled me to retain my sanity in the process). Gracias.

Many thanks to the Richman SuperGroup in general. There are too many of you to thank individually, but I'd like to highlight a few names for sure: Major props to my buddy Ben Good ("B'Jamgo") for introducing me to the whole universe of artificial intelligence / machine learning, and for taking me to the geekiest conference I've ever been to in my whole life. Well, at least it was in San Fran. . . Thanks to Nancy Keating and Sherry Rostami for cranking out all those p24 numbers! Muchas gracias, Karole and Theresa ("Super T"), for being the best benchtop mafia this side of

the Mississip' and generating sequence data like the world's coming to an end. Thanks to Darica Smith, Sharon Wilcox, Roma Sysyn, and Rita Mendez for putting up with my incessant botheration and keeping the lab from coming apart at the seams! Thanks to Judy Nordberg and Peggy O'Keefe in the flow cytometry core for all those tiny little dots (there must be like a bazillion of them). Thanks to my co-conspirators Celsa Spina, Davey Smith, Matt Strain, Sergei Kosakovsky Pond, Simon Frost, and Christopher Woelk for sharing their expertise with me countless times during my graduate career.

I'd like to take a second to thank all of my previous mentors for (unwittingly?) setting me on this path: Bill Calder, if you had not been my Ecology professor at the University of Arizona, I would probably be a disgruntled astronomer right now! Your passion for biology and your respect for the environment was the most powerful recruitment effort I have ever witnessed. Thanks to my gurus Sanford Eigenbrode and Joe Spencer for introducing me to the wonderful world of host-parasite interactions and showing me that science and humor make excellent bedfellows! Bette Korber, how can I possibly thank you enough??? Thanks for being a wonderful mentor and friend all these years, and for always having faith in my abilities (musical, scientific, and otherwise).

Finally, I'd like to thank my friends and family for all their support and kindness. A huge round of applause for all of my fellow musicians for keeping me artistically satisfied out here in San Diego (Dan, Brad, Brendan, James, etc.). Thanks to Kevin the Hare Krishna for keeping my belly full of curry. Thanks to Dr. Brad

Buchman for keeping my body on the mend after my Alfa Romeo got crushed like a pancake. Thanks to all my buddies (especially Heather "MacFreeds" Cartwright and her lizard Zoë "MacFreeds" Iguana) for much cavorting and frolicking. And as always, I owe absolutely everything to my Mom, Dad, and sister ("Dummy-Pads") for being fountains of infinite love, encouragement, and compassion.

The text of Chapter Two, in full, is a reprint of the material as it appears in AIDS Research and Human Retroviruses: Pillai, S.K., B. Good, D. Richman, and J. Corbeil, "A New Perspective on V3 Phenotype Prediction", vol. 19, pp. 145-149, February 2003. The text of Chapter Three, in full, is a reprint of the material as it appears in Virology: Pillai, S.K., S. L. Kosakovsky Pond, C.H. Woelk, D.D. Richman, and D.M. Smith, "'Codon Volatility' Does Not Reflect Selective Pressure on the HIV-1 Genome" (in press). The text of Chapter Four, in full, is a reprint of the material as it appears in the Journal of Virology: Pillai, S.K., B. Good, S. Kosakovsky Pond, J.K. Wong, M.C. Strain, D.D. Richman, and D.M. Smith, "Semen-Specific Genetic Characteristics of Human Immunodeficiency Virus Type 1 *env*", vol. 79, pp. 1734-1742, February 2005. The text of Chapter Five, in full, has been submitted for publication: Pillai S. K., S. Kosakovsky Pond, B. Good, M. C. Strain, R. Ellis, S. Letendre, H. Gunthard, I. Grant, T. Marcotte, J. A. McCutchan, D. D. Richman, and J. K. Wong, "Genetic Attributes of Cerebrospinal Fluid-Derived HIV-1. I was the primary researcher and author, and the co-authors listed in these publications supervised and/or contributed to the research which forms the basis for these chapters.

CURRICULUM VITAE

Satish Kumar Pillai
E-mail: supersatish@mail.com
Website: www.supersatish.com

OBJECTIVE

To investigate the evolution of HIV-1 by designing and implementing computational and experimental tools.

EDUCATION

Ph.D., Biology, University of California, San Diego, March 2005
Dissertation: Effects of anatomic compartmentalization on HIV-1 evolution.
Advisor: Dr. Douglas D. Richman, School of Medicine.

B.S., Ecology & Evolutionary Biology, University of Arizona, Tucson, AZ, May 1996

EXPERIENCE

2000-2005 ***Doctoral Student, Richman Lab*** *[University of California, San Diego]*

- Design and implement *in vitro* assays to investigate the effects of HIV-1 gene expression on host cell surface characteristics
- Design and implement computational tools to investigate the relationship between spatiotemporal heterogeneity and HIV-1 sequence variation

1997-1999 ***Graduate Research Asst., HIV Database*** *[Los Alamos National Lab]*

- Curate, analyze, and disseminate genetic sequences of HIV-1, HIV-2, SIV and related animal retroviruses using UNIX and MAC platforms
- Write web-based software applications in Perl, C++ to investigate retroviral evolution
- Reconstruct viral phylogenies using parsimony, distance, and maximum-likelihood methods
- Assimilate and analyze data for the HIV Immunology Compendium and accompanying website

1996-1997 ***Staff Tutor, SALT Center for Learning Disabilities*** *[U. of Arizona]*

- Tutor students in soil and water sciences, hydrology, physical geography, ecology, evolution, math, and jazz history
- Teach workshops on basic writing skills

- Lead group review sessions

1994-1996 ***Research Assistant, E. Bernays Lab*** *[U. of Arizona Entomology]*

- Observe and analyze the larval behavior of Diamondback moth (*Plutella xylostella*) on various plant lipid substrates
- Isolate individual plant surface lipid components using chromatography

1993-1994 ***Research Assistant, K. Feldmann Lab*** *[U. of Arizona Plant Sciences]*

- Generate cDNA library for *Arabidopsis thaliana* using biochemical and recombinant DNA techniques
- Propagate plants and perform tissue culture of *A. thaliana*

PUBLICATIONS

Pillai S, Woelk C, Kosakovsky Pond S, Richman D, Smith D. "Codon volatility" does not reflect selective pressure on the HIV-1 genome. *Virology*. (in press)

Pillai S, Good B, Kosakovsky Pond S, Wong JK, Strain MC, Richman DD, Smith DM. Semen-specific genetic characteristics of HIV-1 *env*. *J Virol*. 2005 Feb;79(3):1734-42.

Strain MC, Letendre S, **Pillai S**, Russell T, Ignacio CC, Gunthard HF, Good B, Smith DM, Wolinsky SM, Furtado M, Marquie-Beck J, Durelle J, Grant I, Richman DD, Marcotte T, McCutchan JA, Ellis RJ, Wong JK. Genetic composition of HIV-1 in CSF and plasma without treatment and during failing combination antiretroviral therapy. *J Virol*. 2005 Feb;79(3):1772-88.

Yu Q, Konig R, **Pillai S**, Chiles K, Kearney M, Palmer S, Richman D, Coffin J, Landau NR. Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. *Nat Struct Mol Biol*. 2004 May;11(5):435-42.

Kunstman KJ, Puffer B, Korber BT, Kuiken C, Smith UR, Kunstman J, Stanton J, Agy M, Shibata R, Yoder AD, **Pillai S**, Doms RW, Marx P, Wolinsky SM. Structure and function of CC-chemokine receptor 5 homologues derived from representative primate species and subspecies of the taxonomic suborders Prosimii and Anthropoidea. *J Virol*. 2003 Nov;77(22):12310-8.

Ali A, **Pillai S**, Ng H, Lubong R, Richman DD, Jamieson BD, Ding Y, McElrath MJ, Guatelli JC, Yang OO. Broadly increased sensitivity to cytotoxic T lymphocytes resulting from Nef epitope escape mutations. *J Immunol*. 2003 Oct 15;171(8):3999-4005.

Pillai S, Good B, Richman D, Corbeil J. A new perspective on V3 phenotype prediction. *AIDS Res Hum Retroviruses*. 2003 Feb;19(2):145-9.

Kalish ML, Korber BT, **Pillai S**, Robbins KE, Leo YS, Saekhou A, Verghese I, Gerrish P, Goh CL, Lupo D, Tan BH, Brown TM, Chan R. The sequential introduction of HIV-1 subtype B and CRF01AE in Singapore by sexual transmission: accelerated V3 region evolution in a subpopulation of Asian CRF01 viruses. *Virology*. 2002 Dec 20;304(2):311-29.

Good B, Peay J, Corbeil J, **Pillai S**. Class prediction based on gene expression: applying neural networks via a genetic algorithm wrapper. pp. 122-130 in *Proceedings of the 2001 Genetic and Evolutionary Computing Conference* (San Francisco, CA.)

Thakallapally R, Rose P, Vasil S, **Pillai S**, Kuiken C. Reagents for HIV/SIV vaccine studies. pp. 506-516 in *Human Retroviruses and AIDS 1999*. Edited by: Kuiken CL, Foley B, Hahn B, Korber B, McCutchan F, Marx PA, Mellors JW, Mullins JI, Sodroski J, and Wolinksy S. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. (1999)

Spencer JL, **Pillai S**, Bernays EA. Synergism in the oviposition behavior of *Plutella xylostella*: sinigrin and wax compounds. J *Insect Behav.* 1999 Jul;12(4):483-500.

Korber B, Foley BT, Kuiken C, **Pillai S**, Sodroski JG. Numbering positions in HIV relative to HXB2CG. pp. III-102-111 in *Human Retroviruses and AIDS 1998*. Edited by: Korber B, Kuiken CL, Foley B, Hahn B, McCutchan F, Mellors JW, and Sodroski J. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM. (1998)

Eigenbrode SD, **Pillai S**. Neonate *Plutella xylostella* Responses to Surface Wax Components of a Resistant Cabbage (*Brassica oleracea*). *J Chem Ecol*. 1998 Oct; 24(10):1611-27

Kuiken C, Kunstman K, Kunstman J, Bhattacharya T, Kommander K, Good B, **Pillai S**, Wolinsky S. Successive dominance of human immunodeficiency virus type 1 sequences in patients receiving combination antiretroviral therapy. (submitted)

Coleman S, Madrid R, Van Damme N, Bouchet J, Servant C, **Pillai S**, Benichou S, Guatelli J. HIV-1 Nef uses different AP complex family members to modulate specific cellular proteins. (submitted)

Pillai S, Kosakovsky Pond S, Good B, Strain MC, Ellis R, Letendre S, Gunthard

H, Grant I, Marcotte T, McCutchan JA, Richman DD, Wong JK. Genetic attributes of cerebrospinal fluid-derived HIV-1. (in prep)

PRESENTED ABSTRACTS

CSF-specific genetic characteristics of HIV-1 *env*.
11th International Workshop on HIV Dynamics & Evolution (5/2004 - Stockholm, Sweden)

Semen-specific genetic characteristics of HIV-1 *env*.
11th Conference on Retroviruses and Opportunistic Infections (2/2004 - San Francisco, CA)

Compartment-specific patterns associated with CSF-derived V3 sequences.
10th Conference on Retroviruses and Opportunistic Infections (2/2003 - Boston, MA)

Determination of HIV-1 coreceptor usage via machine learning.
HIV Pathogenesis: Recent Advances in the Biology and Pathogenesis of Primate Lentiviruses (4/2002 - Keystone, CO)

Effects of compartmentalization on HIV-1 Nef evolution.
9th Conference on Retroviruses and Opportunistic Infections (2/2002 - Seattle, WA)

AWARDS

Recipient of **Neal Nathanson, M.D. Investigator in Training** award, presented by the **International Society for Neurovirology** in Sardinia, Italy (9/2004)

HOBBIES

-Singer/songwriter/pianist/guitarist/percussionist in three San Diego, CA bands:

- **The Satish Collective** (original rock -> see www.supersatish.com)
- **The Will Edwards Band** (folk rock -> see www.willedwards.net)
- **Gato Papacitos** (Latin jazz -> see www.catdaddies.com)

ABSTRACT OF THE DISSERTATION

Effects of Anatomic Compartmentalization on HIV-1 Evolution

by

Satish Kumar Pillai

Doctor of Philosophy in Biology

University of California, San Diego, 2005

Douglas D. Richman, Chair

Human immunodeficiency virus Type 1 (HIV-1) resides in a wide variety of tissues including the brain, blood, lung, spleen, lymph nodes, and genital tract within infected individuals. Each anatomical niche is characterized by region-specific immunological surveillance, host cell characteristics, and antiretroviral drug penetration. Several reports suggest that the trafficking of virus between anatomic compartments is minimal and infrequent. Therefore, it is expected that HIV should evolve independently in each anatomic compartment, adapting to local immunologic, cellular, and pharmacokinetic characteristics.

The goal of this dissertation is to assess the degree of viral compartmentalization between tissues, and furthermore, to identify viral genetic characteristics that are specific to particular cell types and organs. The first chapter is a brief introduction to the basic biology of HIV, focusing on epidemiology and evolution. The second chapter is an exploration of the genetic determinants underlying differential HIV-1 chemokine receptor usage (CCR5 vs. CXCR4), which

largely defines a strain's cellular tropism. In chapter three, the extent of natural selection on CCR5- and CXCR4-using HIV-1 strains is assessed using a variety of analytical techniques, focusing mainly on the deficiencies of the recently described "codon volatility" method. The fourth chapter investigates the biology of HIV-1 transmission by systematically comparing the population genetics of semen- and blood plasma-derived HIV variants, using previously published HIV-1 envelope RNA sequences. In chapter five, HIV neurotropism and neurovirulence are explored by generating and analyzing several hundred envelope sequences representing cerebrospinal fluid- (CSF) and plasma-derived viruses from paired CSF and plasma samples of eighteen chronically infected donors with available neuropsychiatric data. Finally, in chapters six and seven, the *in vitro* phenotypic correlates of HIV neurotropism are examined by comparing the effects of CSF- and plasma-derived HIV-1 Nef proteins on major histocompatibility class I (MHC-I) expression in the host cell. In addition, preliminary data on the *in vitro* infection of fetal brain aggregates are presented, as a model for studying HIV neuropathogenesis in a controlled laboratory setting.

In summary, the data presented within this dissertation support the theory that distinct viral genetic and evolutionary characteristics are associated with compartment-specific HIV-1 populations.

Chapter 1

Introduction

HIV-1 AND THE AIDS EPIDEMIC

Human immunodeficiency virus type I (HIV-1) is the etiologic agent of acquired immune deficiency syndrome (AIDS). To date, AIDS has taken the lives of over 20 million people worldwide, an estimated 40 million people are living with HIV/AIDS currently, and more than 16,000 are infected with HIV each day (29). Despite over two decades of intense research by clinicians and basic scientists, the AIDS epidemic is still spreading in an uncontrolled fashion across the globe (Fig. 1). To date, there is no cure, and no effective vaccine.

The origin of HIV/AIDS is a topic of great interest to researchers and general public alike. HIV-1 belongs to the lentiviral genus of the family retroviridae. The lentiviral clade includes HIV-2, simian, feline and bovine immunodeficiency viruses, visna virus, and equine infectious anemia virus (3). Phylogenetic analysis of the lentiviral lineage has revealed that the virus most closely resembling HIV-1 is the simian immunodeficiency virus (SIV) strain "SIVcpz" that infects the African *Pan troglodytes troglodytes* chimpanzee subspecies (4)(Fig. 2). This has been interpreted as compelling evidence that HIV-1 entered the human population via zoonotic cross-species transmission from an infected chimpanzee (or chimpanzees). HIV-2, on the other hand, is most closely related to the SIV strain "SIVsm" that infects sooty mangabeys and has been less of a public health concern due to its relatively attenuated pathogenicity and apparent confinement to West Africa. It has been estimated based on inferred rates of molecular evolution that the ancestor of the HIV-1 M group (the subgroup principally responsible for the current pandemic) entered the human

population around 1930, with an error window of 10 years in either direction (12). Although more scandalous scenarios have been proposed, the most plausible mechanism of zoonotic transfer involves the consumption of infected chimpanzee meat ("bushmeat") by humans, which is a relatively common practice in equatorial Africa (6).

It is worth mentioning that the vast majority of HIV research has been focused on HIV-1 subtype "B", a single subclade of the M group, due to its prevalence in the Americas and Europe where most of the research is financed and conducted. Subtype B is one out of over 10 circulating HIV-1 M group subtypes, and represents only a fraction of worldwide diversity. Subtype C is currently the driving force behind the epidemic, covering large stretches of Africa and India where HIV prevalence sometimes surpasses 25% (29). Moreover, much of the basic research has involved particular subtype B strains that are adapted to the transformed cell lines commonly used in a laboratory setting. Although to date there have not been any concrete phenotypic or pathogenic differences identified between subtype designations (which are based purely on genetic distance), it is likely that our window into HIV biology suffers from some degree of myopia.

HIV-1 EVOLUTION AND GENETIC DIVERSITY

HIV-1 evolves via two distinct mechanisms: mutation and recombination. RNA viruses in general tend to have relatively high mutation rates, and HIV is by no means an exception with its estimated rate of $5x10^{-5}$ mutations/site/generation, owing to the poor fidelity of reverse transcriptase and a lack of proofreading machinery (14).

Considering that the genome itself is approximately 10 kilobases long (Fig. 3), it is expected that each progeny virus differs from its parent by one nucleotide on average due to mutation alone.

HIV-1, like all retroviruses, is diploid and contains two copies of the viral RNA genome within each particle (Fig. 3). If a host cell is infected with multiple viruses, the possibility exists that heterozygous virions will be formed and recombinants will be generated during reverse transcription (7). Recombination is rampant within HIV populations, and several reports suggest that it may be a much more powerful and relevant force in shaping HIV evolutionary patterns than mutation. Zhuang et al have observed up to 2.8 crossovers per replication cycle within their *in vitro* system (32). This rate is reflected in the ever growing number of circulating intersubtype recombinant viruses (approaching 20 at the time of this publication), which emerge in regions where multiple subtypes overlap geographically and demographically (18).

The true magnitude of these evolutionary processes becomes apparent when discussed in the context of HIV population biology. Stochastic models suggest that 10^{10} viral particles are produced each day within an infected individual, and generation time is in the neighborhood of 1.8 days (19). This rate of production and turnover coupled with the aforementioned rates of recombination and mutation allow the virus to explore vast reaches of sequence space in short periods of time. As a result, *env* genetic diversity within the HIV-1 M group is as high as 35% (23). Viral diversity within a single individual may reach levels of 10% or more during chronic infection

(25). These numbers are in striking contrast to the rates of evolution and diversification observed within influenza virus populations; worldwide influenza hemaglutinnin (HA) diversity at any given time is usually lower than observed HIV-1 *env* diversity within a single city (11) (Fig. 6).

The rapid evolution and diversification of HIV-1 is of tremendous consequence to its clinical management. One of the greatest challenges facing the design of an effective vaccine is genetic diversity (11). Designing an immunogen that will elicit a vigorous antiviral response against such a wide spectrum of circulating variants is a daunting task, especially when numerous studies demonstrate that a single amino acid substitution within an antigenic sequence is often enough to abrogate a cellular or humoral immune response. For similar reasons, the natural anti-HIV immune responses mounted by an infected host are usually insufficient in the long run. Data from studies involving humans and nonhuman primates demonstrate that when a viral antigenic sequence (epitope) becomes the target of neutralizing anibodies or cellular immunity within an infected host, the epitope accumulates mutations that allow the virus to escape from these responses (5). The host, in turn, may generate new antibodies or T cells that recognize the mutant viral sequence. The virus, once again, usually escapes via mutation. This game of cat-and-mouse may persist for several years, although an untreated host typically loses control eventually, perhaps due to the extraordinary rate at which HIV generates genetic variation.

Antiretroviral drug therapy suffers at the hands of viral evolution as well. The administration of a single drug (e.g. AZT) is almost universally insufficient; HIV

rapidly accumulates mutations that render it resistant to the drug, although these mutations often come at a significant fitness cost to the virus (8). However, resistant strains have been catalogued that are comparable to wildtype fitness in the absence of drug due to the acquisition of compensatory mutations that neutralize the deficits associated with the primary resistance mutations (27). As a result clinicians have resorted to flooding infected patients with multiple drug cocktails that target HIV at several stages of its life cycle. Evolving resistance to multiple classes of antiretrovirals simultaneously is much more difficult for the virus, although evolution and transmission of multi-drug resistant strains has been documented in recent years (22). Even in the absence of multi-drug resistance, the drugs themselves take a significant toll on patients in the long run (patients are often placed on six drugs or more, and interactions and additive effects are not uncommon). The toxicity of antiretrovirals manifests itself in several forms, including pain, fatigue, nausea, lipodystrophy, and neurological syndromes (17).

Another intriguing facet of HIV epidemiologic and intrapatient evolution (which we will revisit in the next chapter) arises from the fact that HIV-1 can exploit different chemokine receptors to gain entry into host cells (Fig. 4). Although several potential coreceptors have been identified in laboratory experiments, CCR5 and CXCR4 appear to be most widely used by primary HIV-1 isolates. The coreceptor usage of an HIV strain (CCR5 vs. CXCR4) largely defines its cytopathology, replication kinetics, and tissue tropism in *in vitro* culture. CXCR4-using (X4) isolates tend to replicate rapidly, induce the formation of syncytia (giant multinucleated cells),

and have the capacity to infect transformed T cell lines. CCR5-using (R5) isolates, on the other hand, replicate slowly, do not induce syncytia in T cell lines, and can often infect monocyte-derived macrophages in a laboratory setting (2). Coreceptor usage is also indicative of *in vivo* pathogenicity and transmissibility. CXCR4 strains are associated with more rapid progression to AIDS, in line with their accelerated replication rates *in vitro* (13). It has been observed that HIV-infected individuals experiencing opportunistic infections associated with end-stage disease (Fig. 5) are several times more likely to harbor a syncitium-inducing (X4) strain than asymptomatic patients (21). In addition, people homozygous for a common 32 base-pair deletion in the gene encoding the CCR5 receptor may only be productively infected by CXCR4-using HIV variants (15). Coreceptor usage also modulates viral access to various compartments within the human body, due to tissue-specific cellular characteristics. For example, HIV infection in the central nervous system is highly correlated with the preferential usage of CCR5, while the thymus is mainly colonized by X4 variants (10, 26). In light of the AIDS epidemic as a whole, the majority of HIV-positive individuals are initially infected with R5 HIV strains, suggesting that there may be a selective advantage of such variants with respect to sexual, parenteral, and vertical transmission (30).

As my chosen title indicates, this dissertation focuses on one particular aspect of HIV evolutionary biology: the effects of anatomic compartmentalization on HIV-1 evolution. Let me begin by listing a few definitions of the term “compartment”:

- "One of the parts or spaces into which an area is subdivided." (dictionary.com)
- ". . . an anatomical site in which the virus. . . evolves distinctively from other anatomical sites. . . because of differences between the major cell types sustaining viral replication." (9)
- "In vivo virologic compartments are cell types or tissues between which there is a restriction of virus flow. . ." (16)
- The particular area within a habitat occupied by an organism (ecological definition of "niche"). (dictionary.com)

The meaning of the word compartment as I use it within this dissertation is essentially an amalgam of all four definitions listed above. The presence of HIV (RNA and DNA) has been catalogued in several anatomic compartments within human hosts, including the brain, blood, lung, spleen, lymph nodes, and genital tract. Each anatomical niche is characterized by region-specific immunological surveillance, host cell characteristics, and antiretroviral drug penetration. Several reports suggest that the trafficking of virus between anatomic compartments is minimal and infrequent. Therefore, it is expected that HIV should evolve independently in each anatomic compartment, adapting to local immunologic, cellular, and pharmacokinetic characteristics. The goal of this dissertation is to assess the degree of viral compartmentalization between tissues, and furthermore, to identify viral genetic characteristics that are specific to particular cell types and organs.

As a final point, investigating the effects of anatomic compartmentalization on HIV-1 evolution is not a purely academic pursuit. Viral compartmentalization has

significant consequences in the clinical world. For instance, several sites (e.g. central nervous system) are characterized by suboptimal drug penetration and act as sanctuary sites, compromising antiretroviral therapy (28). Another example involves HIV transmission; the vast majority of HIV-1 infections result from exposure to virus in male genital secretions (20). Therefore, cataloguing the specific genotypic and phenotypic characteristics of virus in the male genital tract (and genital secretions) may be crucial in designing an effective prophylactic vaccine that selectively targets transmitted HIV-1 variants. My sincerest hope is that the information contained within the next couple of hundred pages will shed some light on the relationship between anatomic compartment and viral genetics that will eventually have some beneficial clinical impact.

REFERENCES

1. Coffin, J. M., S. H. Hughes, and H. E. Varmus (eds). 1997. *Retroviruses*. Cold Spring Harbor Lab Press, Plainview, NY, p.34.

2. Fenyo, E. M., H. Schuitemaker, B. Asjo, and J. McKeating. 1997. The History of HIV-1 Biological Phenotypes Past, Present, and Future. pp. III-13-18 in *Human Retroviruses and AIDS 1997*. Edited by: Korber, B., Hahn, B., Foley, B., Mellors, J. W., Leitner, T., Myers, G., McCutchan, F., and Kuiken, C. L. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

3. Fields, B. N., D.M. Knipe, and P.M. Howley (eds). 1996. *Fields Virology*. Lippincott-Raven Publishers, Philadelphia, PA. 3rd Edition, Volume 2.

4. Gao F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn. 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. Nature 397:436-41.

5. Goulder P. J., and D. I. Watkins. 2004. HIV and SIV CTL escape: implications for vaccine design. Nat Rev Immunol 4:630-40.

6. Hahn, B. H., G. M. Shaw, K. M. De Cock, and P. M. Sharp. 2000. AIDS as a zoonosis: scientific and public health implications. Science 287:607-14.

7. Hu, W. S., and H. M. Temin. 1990. Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. Proc Natl Acad Sci USA 87:1556-60.

8. Larder, B. A., G. Darby, and D. D. Richman. 1989. HIV with reduced sensitivity to zidovudine (AZT) isolated during prolonged therapy. Science 243:1731-4.

9. Lowe, S. H., S. U. Sankatsing, S. Repping, F. van der Veen, P. Reiss, J. M. Lange, and J. M. Prins. 2004. Is the male genital tract really a sanctuary site for HIV? Arguments that it is not. AIDS 18:1353-62.

10. Kitchen, S. G., and J. A. Zack. 1997. CXCR4 expression during lymphopoiesis: implications for human immunodeficiency virus type 1 infection of the thymus. J Virol 71:6928-34.

11. Korber, B., B. Gaschen, K. Yusim, R. Thakallapally, C. Kesmir, and V. Detours. 2001. Evolutionary and immunological implications of contemporary HIV-1 variation. Br Med Bull 58:19-42.

12. Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science 288:1789-96.

13. Kreisberg, J. F., D. Kwa, B. Schramm, V. Trautner, R. Connor, H. Schuitemaker, J. I. Mullins, A. B. van't Wout, and M. A. Goldsmith. 2001. Cytopathicity of human immunodeficiency virus type 1 primary isolates depends on coreceptor usage and not patient disease status. J Virol 75:8842-7.

14. Mansky, L. M. and H. M. Temin. 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol 69:5087-5094.

15. Michael, N. L., J. A. Nelson, V. N. KewalRamani, G. Chang, S. J. O'Brien, J. R. Mascola, B. Volsky, M. Louder, G. C. White 2nd, D. R. Littman, R. Swanstrom, and T. R. O'Brien. 1998. Exclusive and persistent use of the entry coreceptor CXCR4 by human immunodeficiency virus type 1 from a subject homozygous for CCR5 delta32. J Virol 72:6040-7.

16. Nickle, D. C., M. A. Jensen, D. Shriner, S. J. Brodie, L. M. Frenkel, J. E. Mittler, and J. I. Mullins. 2003. Evolutionary indicators of human immunodeficiency virus type 1 reservoirs and compartments. J Virol 77:5540-6.

17. Nolan, D., and S. Mallal. 2004. Complications associated with NRTI therapy: update on clinical features and possible pathogenic mechanisms. Antivir Ther 9:849-63.

18. Peeters, M. 2000. Recombinant HIV sequences: Their role in the global epidemic. pp. I-39-54 in *HIV Sequence Compendium 2000*. Edited by: Kuiken, C. L., Foley, B., Hahn, B., Korber, B., McCutchan, F., Marx, P. A., Mellors, J. W., Mullins, J. I., Sodroski, J., and Wolinksy, S. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

19. Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho. 1996. HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582-1586.

20. Piot, P., M. Bartos, P. D. Ghys, N. Walker, and B. Schwartlander. 2001. The global impact of HIV/AIDS. Nature 410:968-973.

21. Richman, D. D., and S. A. Bozzette. 1994. The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression. J Infect Dis 169:968-74.

22. Richman, D. D., S. C. Morton, T. Wrin, N. Hellmann, S. Berry, M. F. Shapiro, and S. A. Bozzette. 2004. The prevalence of antiretroviral drug resistance in the United States. AIDS 18:1393-401.

23. Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. L. Kalish, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber. 1999. HIV-1 Nomenclature Proposal. pp. 492-505 in *Human Retroviruses and AIDS 1999*. Edited by: Kuiken, C. L., Foley, B., Hahn, B., Korber, B., McCutchan, F., Marx, P. A., Mellors, J. W., Mullins, J. I., Sodroski, J., and Wolinksy, S. Published by: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.

24. Sabin, C. A., H. Devereux, A. N. Phillips, A. Hill, G. Janossy, C. A. Lee, and C. Loveday. 2000. Course of viral load throughout HIV-1 infection. J Acquir Immune Defic Syndr 23:172-7.

25. Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins. 1999. Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection. J Virol 73:10489-10502.

26. Shieh, J. T., A. V. Albright, M. Sharron, S. Gartner, J. Strizki, R. W. Doms, and F. Gonzalez-Scarano. 1998. Chemokine receptor utilization by human immunodeficiency virus type 1 isolates that replicate in microglia. J Virol 72:4243-9.

27. Simon, V., N. Padte, D. Murray, J. Vanderhoeven, T. Wrin, N. Parkin, M. Di Mascio, and M. Markowitz. 2003. Infectivity and replication capacity of drug-resistant human immunodeficiency virus type 1 variants isolated during primary infection. J Virol 77:7736-45.

28. Smit, T. K., B. J. Brew, W. Tourtellotte, S. Morgello, B. B. Gelman, and N. K. Saksena. 2004. Independent evolution of human immunodeficiency virus (HIV) drug resistance mutations in diverse areas of the brain in HIV-infected patients, with and without dementia, on antiretroviral treatment. J Virol 78:10133-48.

29. UNAIDS/WHO. 2004. AIDS epidemic update : December 2003. UNAIDS/ World Health Organization, Geneva, Switzerland.

30. Van't Wout, A. B., N. A. Kootstra, G. A. Mulder-Kampinga, N. Albrecht-van Lent, H. J. Scherpbier, J. Veenstra, K. Boer, R. A. Coutinho, F. Miedema, and H. Schuitemaker. 1994. Macrophage-tropic variants initiate human immunodeficiency virus type 1infection after sexual, parenteral, and vertical transmission. J Clin Invest 94:2060-7.

31. Weiss, R. A. 2001. Gulliver's travels in HIVland. Nature 410:963-7.

32. Zhuang, J., A. E. Jetzt, G. Sun, H. Yu, G. Klarmann, Y. Ron, B. D. Preston, and J. P. Dougherty. 2002. Human Immunodeficiency Virus Type 1 Recombination: Rate, Fidelity, and Putative Hot Spots. J Virol 76:11273-11282.

FIGURES AND TABLES

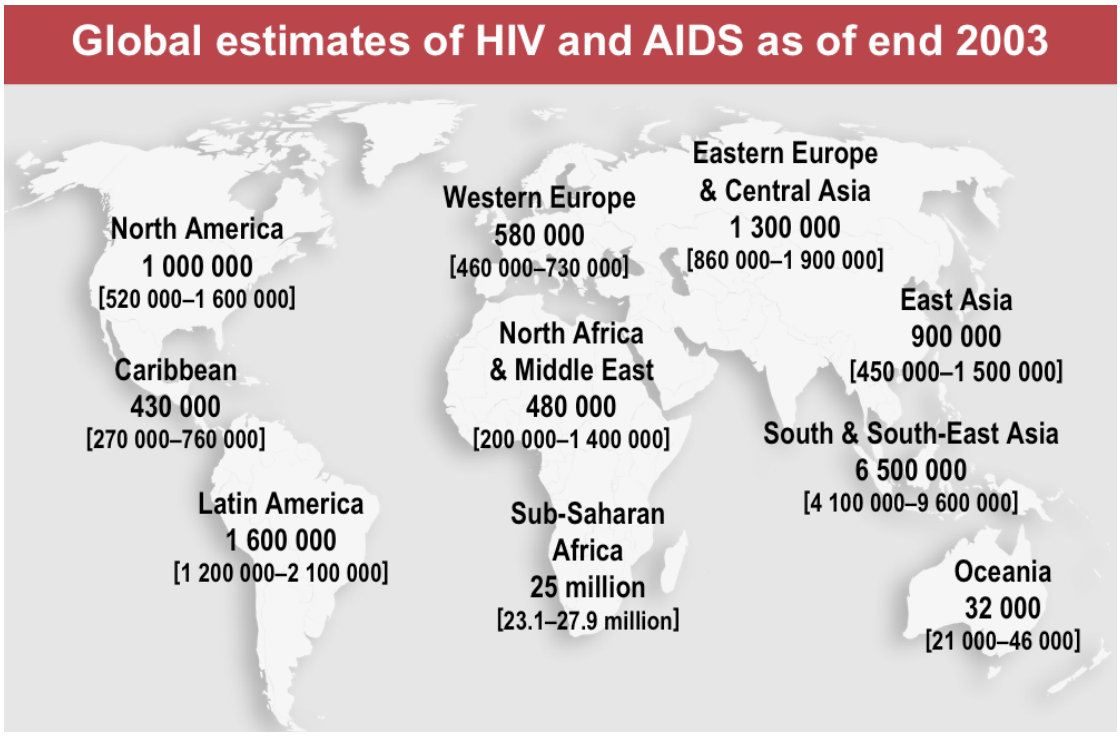


Figure 1: Prevalence of HIV infection on a continent-by –continent basis across the world (World Health Organization, 2003)

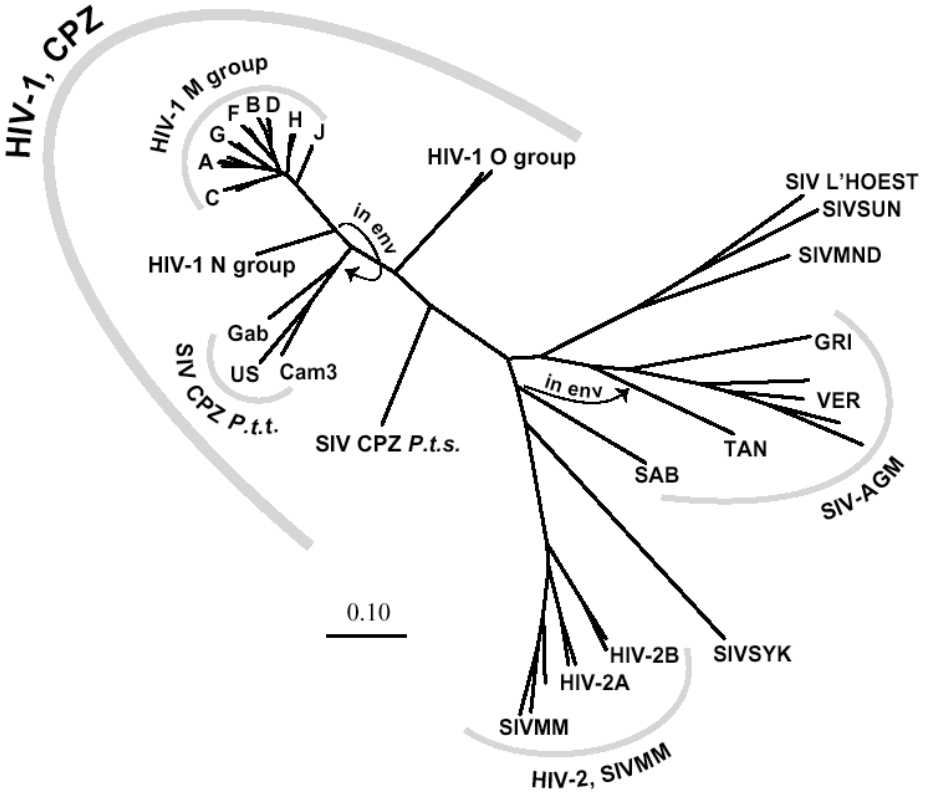


Figure 2: Phylogenetic tree of the primate lentiviral subclade of retroviruses. Scale bar represents 10% genetic distance.

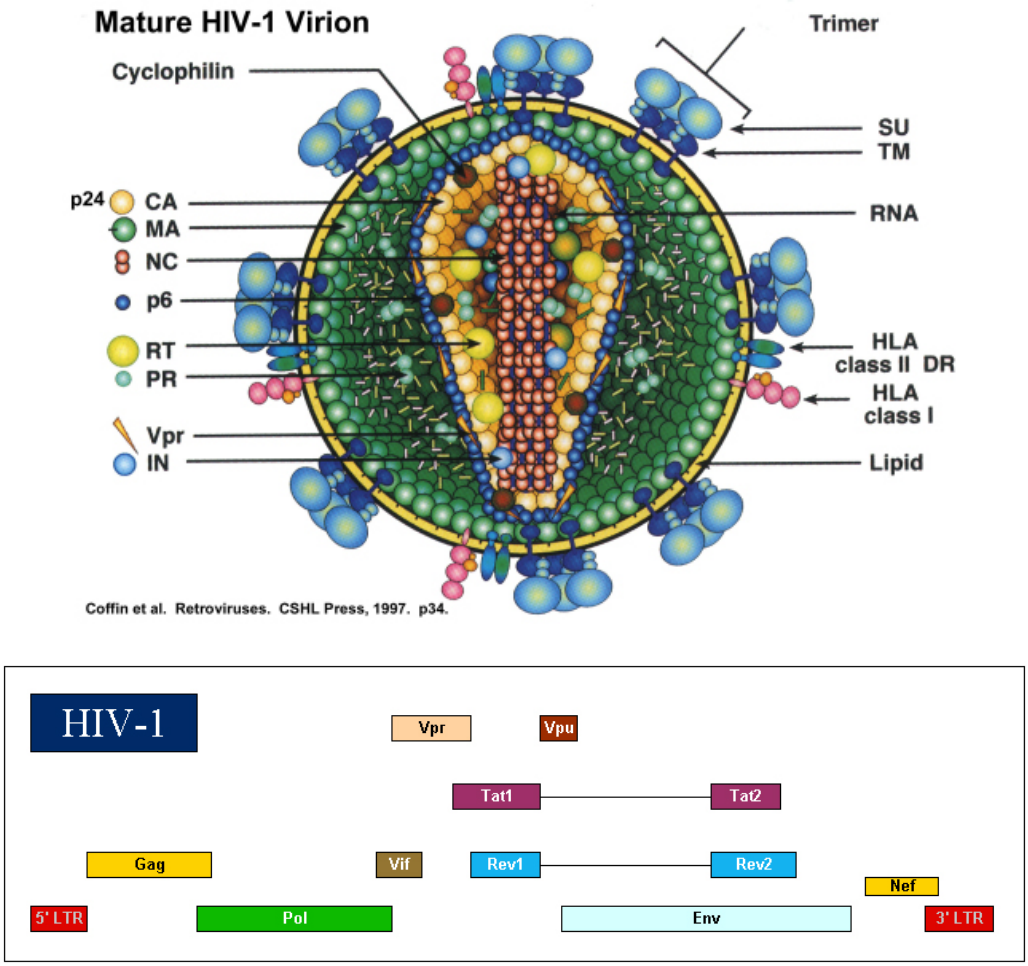


Figure 3: Virion structure and genomic organization of HIV-1. (Top) Schematic of the mature HIV-1 virion (1). The packaging of two copies of the viral RNA genome in each particle allows for viral recombination to occur during each replicative cycle. (Bottom) Map of the ~9.8 kb HIV-1 genome. A total of nine viral genes are encoded (gag, pol, vif, vpr, tat, rev, vpu, env, and nef). Gag, pol, and env are the defining features of the retroviral family, encoding capsid structural proteins, replicative enzymes, and surface glycoproteins, respectively.

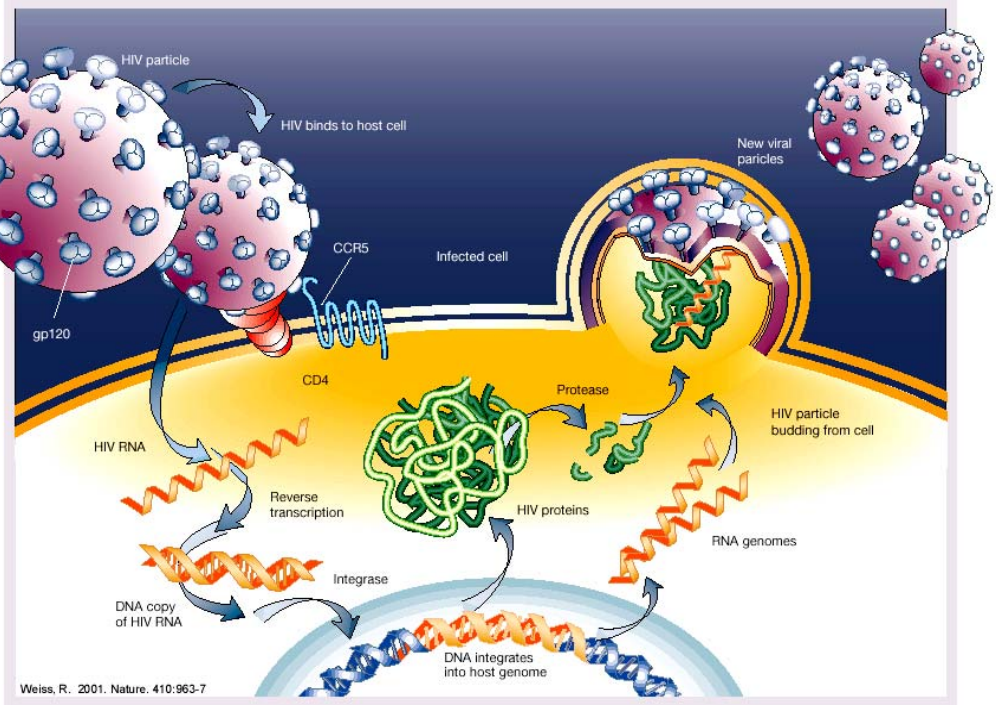


Figure 4: HIV-1 life cycle (31). HIV primarily infects CD4+ T cells, macrophages, and dendritic cells. Initial infection requires the engagement of the viral envelope glycoprotein to the CD4. receptor. Secondary attachment to a chemokine coreceptor (typically CCR5 or CXCR4) permits membrane fusion and entry into the host cell. Following entry, the RNA genome contained in the viral particle is reverse transcribed. The viral DNA genome is then integrated into the host cellular DNA backbone, at which point viral proteins will be transcribed and translated by the host cellular machinery. Newly synthesized viral components are packaged at the plasma membrane, and then bud from the cell to produce a round of infectious viral progeny particles.

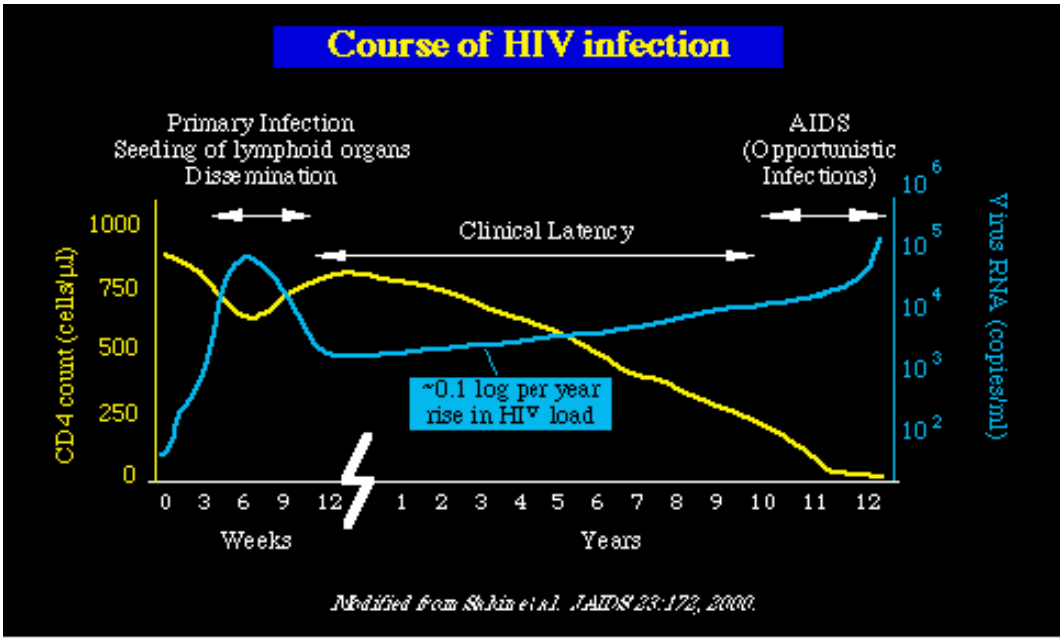


Figure 5: Natural history of HIV-1 infection (24). Blue line represents viral load (RNA copies/ml), and yellow line represents CD4+ cell counts. Newly infected individuals are typically characterized by aggressive primary viremia during which viral titres may exceed 10^6 particles/ml. Once the host immune response begins to combat the infection, viral titres rapidly drop down to a set point which is usually orders of magnitude below peak primary viremia concentration. Chronic infection is characterized by perpetually dropping CD4+ counts and a gradual increase in viral load. End-stage disease (AIDS) is reached when CD4+ counts fall below 200 cells /ml, leaving the host vulnerable to opportunistic infections. The time span between initial infection and death varies dramatically between untreated individuals, most likely due to variation in host genotype.

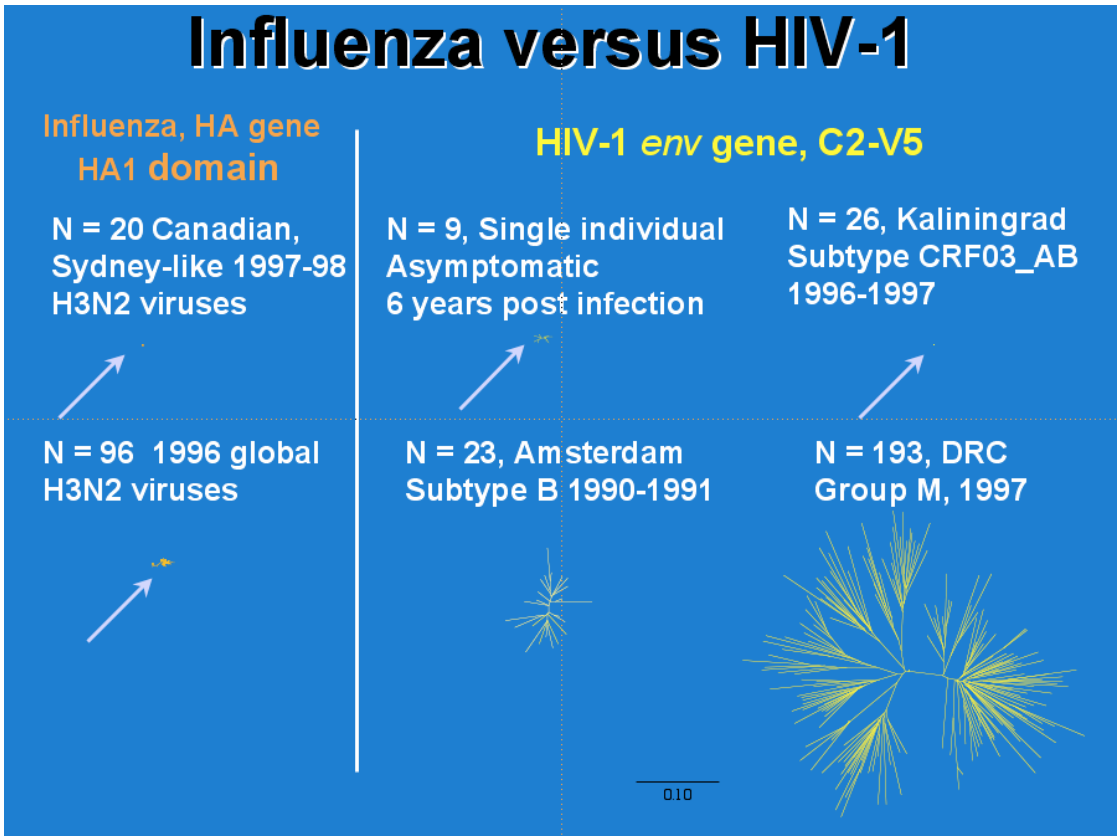


Figure 6: Comparison of genetic diversity within HIV-1 and influenza. Scale bar represents 10% genetic distance (Kindly provided by Dr. Bette Korber, Los Alamos National Lab).

Chapter 2

A New Perspective on V3 Phenotype Prediction

ABSTRACT

The particular coreceptor used by a strain of HIV-1 to enter a host cell is highly indicative of its pathology. HIV-1 coreceptor usage is primarily determined by the amino acid sequence of the V3 loop region of the viral envelope glycoprotein. The canonical approach to sequence-based prediction of coreceptor usage was derived via statistical analysis of a less reliable and significantly smaller data set than what is presently available. We aimed to produce a superior phenotypic classifier by applying modern machine learning (ML) techniques to the current database of V3 loop sequences with known phenotype. The trained classifiers along with the sequence data are available for public use at the supplementary website: http://genomiac2.ucsd.edu:8080/wetcat/v3.html.

INTRODUCTION

The entry of HIV-1 into a host cell is a two-stage process. First, the viral envelope glycoprotein binds to the cell surface molecule CD4, inducing a conformational change in the gp120 ectodomain of the protein. Second, the glycoprotein docks to a seven-transmembrane chemokine coreceptor on the cell surface, triggering the presentation of its gp41 transmembrane segment. This sequence of events results in membrane fusion and penetration of the virus into the cytosol (14). The two principal coreceptors used by HIV-1 are CXCR4 and CCR5, members of the CXC and CC chemokine receptor families, respectively (2).

The particular coreceptor used by a strain of HIV-1 (CXCR4 vs. CCR5) largely defines its replication kinetics and cytopathology *in vitro*. Moreover, coreceptor usage is indicative of the pathogenicity, tissue tropism, and transmissibility of a virus *in vivo*. Unsurprisingly, the determination of this viral phenotype is critical in a wide variety of HIV research contexts.

Several experiments have been conducted on HIV isolates to pinpoint the genetic basis underlying coreceptor preference. The generation and analysis of chimeric (recombinant) viruses have localized the primary determinant of coreceptor usage to the 35 amino acid V3 loop subregion of the HIV envelope glycoprotein (1).

Earlier work involving statistical analysis of V3 loop amino acid sequences and their respective phenotypes suggested that the presence of a positively charged residue at positions 11 and/or 25 of the V3 loop (numbered according to the North American consensus; see Fig. 1) conferred the ability to dock with CXCR4, while CCR5 binding is the default condition (3). To date, this "charge rule" is the most accepted method of sequence-based prediction. However, prediction based on this rule does not always align with experimental determination of coreceptor usage (6). The inaccuracy of the charge rule is most likely due to the comparatively sparse and unreliable data that were available at the time of its creation. Since then, the number of sequences with known phenotype has increased substantially, and the laboratory-based assays used to generate the data have improved. Another possible candidate for a deficiency in this predictive scheme is the consideration of only 2 of the 35 available amino acid positions in the V3 loop.

Modern machine learning (ML) techniques for class prediction can provide advantages over traditional statistics in terms of their abilities to identify and exploit interactions between feature variables. In addition, the rules they generate can often be interpreted with relative ease (7,13). ML has already proven extremely useful in segregating biological sequence data into functional classes (4,9). We used a variety of ML approaches to develop a better classifier of coreceptor usage and to assess the impact of other V3 loop attributes on viral tropism.

MATERIALS AND METHODS

All the V3 loop sequence entries containing documentation of experimentally determined coreceptor usage were downloaded from the Los Alamos National Laboratory HIV Sequence Database, and duplicate sequences were removed. V3 loops less than 34 or more than 36 amino acids in length were deleted from the data set in the interests of producing a relatively gap-free alignment (Table 1). CLUSTAL W (11) was used to generate an automated multiple sequence alignment of the remaining 271 sequences, using the default parameter settings.

Classifiers were trained to make the distinction between viruses capable of using CXCR4 as a coreceptor, versus those that were incapable. Dual-tropic (R5X4) viruses were therefore pooled into the X4 class. Each sample in the initial training set included the amino acid character (or gap) at each of the 40 positions in the V3 loop alignment.

All the experiments in this analysis were conducted using WEKA (the Waikato Environment for Knowledge Analysis), an open source collection of data-processing

and machine-learning algorithms (13). Written in Java, it runs on most platforms and is available for free download at http://www.cs.waikato.ac.nz/ml/weka. Of the many techniques for classification included in the WEKA package, we chose to focus on an implementation of the Quinlan C4.5 decision tree inducer called "j48," an algorithm that builds rules from partial decision trees constructed with C4.5, called "PART," and a sequential minimal optimization-based implementation of support vector machines (SVM) (8,12).

One hundred iterations of stratified 10-fold cross-validation were used to evaluate the different classifiers and training set compositions. For each of 100 trials, the data set was randomly divided into 10 groups of approximately equal size and class distribution. For each "fold," the classifier was trained using all but 1 of the 10 groups and then tested on the unseen group. This procedure was repeated for each of the 10 groups. The cross-validation score for 1 trial was the average performance across each of the 10 training runs. The reported score is the average across the 100 trials. The same divisions of the data set were used for each type of classifier (including the charge rule) to allow for direct comparison.

RESULTS AND DISCUSSION

In the first trial, we compared the abilities of four classifiers, the charge rule, SVM, C4.5, and PART, to accurately predict the coreceptor usage of HIV-1 V3 loop amino acid sequences. The second trial compared the performance of these classifiers on the same data set but with positions 11 and 25 deleted. To reiterate, positions 11 and 25 of the V3 loop constitute the entire basis of prediction for the canonical charge

rule predictor. We eliminated this information from the training data for the second trial in the interests of both informatics and biology; we aimed to assess the capacity of machine learning to unearth novel information content, while concomitantly identifying new areas within the loop that influence HIV coreceptor usage.

The results presented in Table 2 indicate that we can generate a more reliable sequence-based predictor of HIV coreceptor usage by employing a variety of ML techniques. In addition, classifiers trained on sequences lacking positions 11 and 25 produce results competitive to the conventional method and to classifiers constructed using the entire available feature set. Trials conducted with a variety of different sequence attributes resulted in fairly consistent construction of classifiers performing near 90% accuracy in cross-validation trials (data not shown), suggesting that information regarding coreceptor usage is widely distributed throughout the V3 loop.

Throughout all our trials, the SVM was consistently the best phenotypic classifier. Table 3 summarizes its class-specific performance in cross-validation on the full sequence set.

An obvious benefit of the conventional charge rule is its simplicity. Implementing a simpler method may be desirable, especially if significantly more complex schemes provide only marginally better results. In the interest of succinctness, we constructed decision trees using only two V3 loop attributes. The tree in Fig. 2 was constructed with the two sequence attributes identified as having the highest information content with regard to coreceptor usage. Using only positions 7 and 11 (positions 8 and 12 in our alignment), we were able to automatically construct

a theoretical framework that consistently outperformed the charge rule (Fig. 2). Moreover, this rule set had the second highest cross-validation score, 89.97%, and constituted the fourth-best classifier overall by correctly classifying 90.77% of the training set.

It is worth noting that when limiting the search to predictors formed from only two V3 positions, the combination of positions 11 and 25 does not provide the greatest information content. Position 11 is certainly the most predictive, but position 25 does not contribute much additional information (Table 2). Surprisingly, rules generated using only position 11 result in slightly better classification performance than rules that include position 25 (data not shown).

A small but significant subset of the isolates within the data set was consistently misclassified by most of the classifiers, even when they were included in the training set. The final step in our analysis concentrated on unearthing a common element between those isolates that were misclassified when using the entire data set for both training and testing. As mentioned earlier, in addition to R5 and X4 strains there is a third phenotypic class of "dual-tropic" R5X4 variants that can utilize either chemokine receptor to enter a target cell. It has been hypothesized that dual-tropic viruses may represent an intermediate evolutionary stage between full-fledged X4 strains and their CCR5-utilizing ancestors (5). Our classifiers were trained to make the distinction between viruses capable of using CXCR4, dual-tropic viruses included, versus those that were incapable. It is likely that dual-tropic viruses use a different "sequence key" to unlock the X4 door than conventional (monotropic) CXCR4-using

strains, because they have retained the capacity to bind CCR5 as well. Therefore, these isolates may be incorrectly labeled as "X4 incapable" by our classification methods, and hence contribute disproportionately to the various error sets.

We investigated this possibility by tallying the monotropic and dual-tropic sequences in each error set and comparing these numbers against the total in each category (250 monotropic and 21 dual-tropic isolates). The data in Table 4 unequivocally demonstrate that dual-tropic variants are significantly overrepresented in all error sets ($p<0.01$, Fisher exact test), speaking to the biological uniqueness of the R5X4 class, and to the efficacy and sensitivity of the classifiers themselves.

To determine whether there was a consistent sequence pattern associated with this third phenotypic class, each of the classifiers was implemented to classify the data into R5, X4, and R5X4 (dual-tropic) isolates. However, the classifiers could not satisfactorily perform this task, likely because of the inadequate number of available training cases in the dual-tropic category.

The two primary goals of this project were to create a better classifier of coreceptor usage based on V3 sequence and to identify new biologically meaningful positions within this region. Our results indicate a marginal improvement in performance over the established charge rule, and demonstrate conclusively that positions within V3 other than positions 11 and 25 can be substantially informative in determining HIV phenotype. Furthermore, examination of the linkage between these newly implicated positions and the two relied on by the conventional classifier

suggests that they contain novel, independent information pertaining to coreceptor usage.

A limitation to classifier performance stems from the minor subset of V3 loop sequences that violate the sequence-phenotype relationships exhibited by the vast majority of training cases. We determined that a large proportion of these cases represent a third, biologically distinct phenotypic class of dual-tropic isolates that can utilize either chemokine receptor to enter a target cell. It is likely, therefore, that these errors were reflections of a conflict between the two-way classification task and the tripartite structure of the phenotypic data, rather than a shortcoming in the classifiers themselves.

Our revisitation of the conventional prediction scheme suggests that the charge rule may in fact be somewhat obsolete. Predictions based on position 11 alone were on average more accurate than those based on positions 11 and 25. Considering that the statistical derivation of the charge rule was performed several years ago, it is possible that the inclusion of position 25 reflects a sampling bias in the significantly smaller data set available at that time.

Experiments involving mutagenesis of the HIV-1 envelope glycoprotein have introduced the possibility that positions outside the V3 loop may also influence viral tropism (10). We would like to apply the same ML techniques to systematically determine how sequence positions within the HIV-1 envelope but outside the V3 loop subregion modulate coreceptor usage. This will depend on the large-scale generation of full-length envelope sequences with corresponding phenotypic data. In addition, as

in vitro assays become more sophisticated, it should be possible to describe coreceptor usage on a continuous scale, rather than by categorizing the data into discrete, arbitrary classes. This information will allow for a high-resolution map of sequence against phenotype, whereby subtle changes in sequence could be predictive of minor effects on coreceptor preference.

The most significant contributions of this work are the elucidation of predictive V3 positions other than positions 11 and 25, and the demonstration of the power of machine learning for rapid knowledge discovery based on protein sequence. In an age when sequence data are being generated at an astonishing pace, machine learning is an invaluable tool for fluently bridging the gap between genotype and phenotype.

ACKNOWLEDGMENTS

The text of this chapter, in full, is a reprint of the material as it appears in AIDS Research and Human Retroviruses: Pillai, S.K., B. Good, D. Richman, and J. Corbeil, "A New Perspective on V3 Phenotype Prediction", vol. 19, pp. 145-149, February 2003. I was the primary author, and the co-authors listed in this publication supervised and/or contributed to the research which forms the basis for this chapter.

We thank Brian Gaschen at the HIV Sequence Database (Los Alamos, NM) for assistance in consolidating the training data. We also extend our gratitude to Drs. Joe Wong, Simon Frost, and Andrew LeighBrown for critical reviews of this manuscript.

This work was supported by the National Institute of Allergy and Infectious Diseases (A146237 and AI47703; J.C.), the Center for AIDS Research Genomics Core Laboratory (AI36214), the Universitywide AIDS Research Program (IS99-SD213 and

PH97-SD-201), and the San Diego Veterans Medical Research Foundation, as well as by NIH grants AI27670, AI38858, AI43638, and AI29164 (D.D.R.) and the San Diego Veterans Affairs Healthcare System.

REFERENCES

1. Cheng-Mayer C., M. Quiroga, J. W. Tung, D. Dina, and J. A. Levy. 1990. Viral determinants of human immunodeficiency virus type 1 T-cell or macrophage tropism, cytopathogenicity, and CD4 antigen modulation. J Virol 64:4390–4398.

2. Fenyö E. M., H. Schuitemaker, B. Åsjö, J. McKeating, Q. Sattentau, and EC Concerted Action on HIV Variability. 1997. The history of HIV-1 biological phenotypes past, present and future. *In*: *Human Retroviruses and AIDS 1997: A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences* (Korber B, Hahn B, Foley B, Mellors JW, Leitner T, Myers G, McCutchan F, and Kuiken CL, eds. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, pp. III13–III18.

3. Fouchier R. A. M., M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker. 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J Virol 66:3183–3187.

4. Hua, S. and Z. Sun. 2001. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17:721–728.

5. Lu, Z., J. F. Berson, Y. Chen, J. D. Turner, T. Zhang, M. Sharron, M. H. Jenks, Z. Wang, J. Kim, J. Rucker, J. A. Hoxie, S. C. Peiper, and R. W. Doms. 1997. Evolution of HIV-1 coreceptor usage through interactions with distinct CCR5 and CXCR4 domains. Proc Natl Acad Sci USA 94:6426–6431.

6. McDonald, R. A., G. Chang, and N. L. Michael. 2001. Relationship between V3 genotype, biologic phenotype, tropism, and coreceptor use for primary isolates of human immunodeficiency virus type 1. J Hum Virol 4:179–187.

7. Mjolsness, E. and D. DeCoste. 2001. Machine learning for science: State of the art and future prospects. Science 293:2051–2055.

8. Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, Calif.

9. Resch, W., N. Hoffman, and R. Swanstrom. 2001. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. Virology 288:51–62.

10. Rizzuto, C. D., R. Wyatt, N. Hernandez-Ramos, Y. Sun, P. D. Kwong, W. A. Hendrickson, and J. Sodroski. 1998. A conserved HIV gp120 glycoprotein structure involved in chemokine receptor binding. Science 280:1949–1953.

11. Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

12. Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, New York.

13. Witten IH and Frank E. 2000. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, Calif.

14. Wyatt, R., and J. Sodroski. 1998. The HIV-1 envelope glycoproteins: Fusogens, antigens, and immunogens. Science 280:1884-1888.

FIGURES AND TABLES

Table 1: Composition of dataset

R5	X4	R5X4
168 (62%)	103 (38%)	21 (8%)

Table 2: Classifier performance. Percent correct for 100 rounds of 10-fold cross-validation. Values in boldface indicate a statistically significant improvement over the charge rule. The default settings from WEKA were used in all cases.

Classifier	Full Sequence	11&25 removed
Charge	87.45%	0
SVM	**90.86%**	88.79%
C4.5	**89.51%**	84.54%
PART	**89.37%**	85.95%

Table 3: Class-based statistics for the Support Vector Machine on the full sequence set. Precision (predictive power) is equal to the number of true positive predictions for a class divided by the number of predicted positives.

Class	True- Positive Rate	False-Positive Rate	Precision
CXCR4	0.757	0.024	0.951
CCR5	0.976	0.243	0.868

Table 4: Misclassification of mono- and dual-tropic sequences.

Classification method	% of mono-tropics misclassified	% of dual-tropics misclassified
Charge rule	7	33
C4.5	6	43
PART	5	38
PART-no 11&25	6	19
SVM	0	5

ctrpnnntrk**s**ihigpgrafytag**e**iigdirqahc

Figure 1: North American V3 loop sequence with positions 11 and 25 (basis of the "charge rule") underlined.

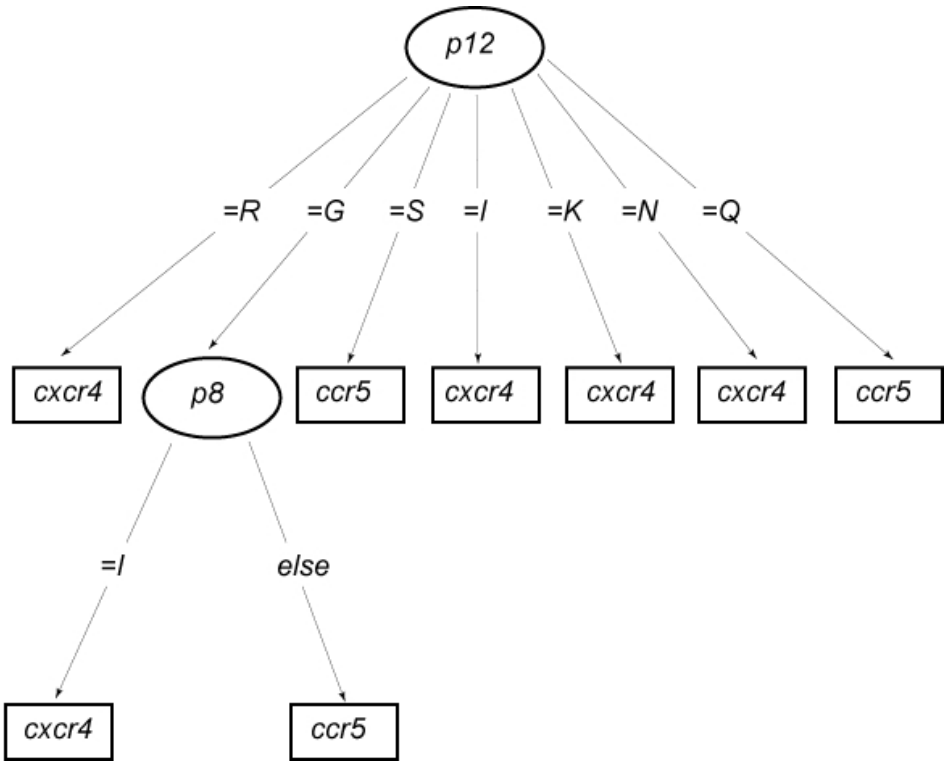


Figure 2: Rule set constructed using C4.5 trained on positions 7 and 11 of the V3 sequence (8 and 12 in the alignment).

Chapter 3

"Codon Volatility" Does Not Reflect Selective Pressure on the HIV-1 Genome

ABSTRACT

Codon volatility is defined as the proportion of a codon's point-mutation neighbors that encode different amino acids. The cumulative volatility of a gene in relation to its associated genome was recently reported to be an indicator of selection pressure. We used this approach to measure selection on all available full-length HIV-1 subtype B genomes in the Los Alamos HIV Sequence Database, and compared these estimates against those obtained via established likelihood- and distance-based comparative methods. Volatility failed to correlate with the results of any of the comparative methods demonstrating that it is not a reliable indicator of selection pressure.

Keywords: HIV, codon volatility, selective pressure, evolution

INTRODUCTION

Natural selection is defined as the process resulting in the evolution of organisms best adapted to their environment (18). Measuring natural selection (positive and negative) across the HIV-1 genome is of tremendous interest to theoreticians and clinicians alike. Evidence of positive selection obtained via nucleotide sequence analysis likely reflects Darwinian adaptation of the virus in response to environmental pressure (37). Cellular and humoral immunity, host-generated antiviral factors such as APOBEC3G, and antiretroviral drug therapy are all

purported to contribute to this pressure and select for adaptive amino acid substitutions in the HIV genome. Evidence of negative selection, on the other hand, reflects the inflexibility of amino acid sequence resulting from functional constraint. The detection and accurate empirical assessment of both processes are critical to the clinical management of HIV infection. Positive selection detection can provide us a window into the evolution of immunologic escape and drug resistance, while negative selection can provide relatively immutable targets for possible therapeutic intervention (28).

Natural selection of protein-coding genes is typically assessed by comparing at least two homologous nucleotide sequences. Positive selection is usually defined as having more nonsynonymous substitutions per nonsynonymous site (dN) than synonymous substitutions per synonymous site (dS), while negative selection is defined as the converse (19). Plotkin et al recently described a novel method to detect positive selection using only a single genome sequence (26). This approach is based entirely on the concept of differential "codon volatility", described as the probability that a point mutation within a codon results in an amino acid change. For example, the triplets AGA and CGA both encode arginine. The codon AGA is assigned a volatility of 0.75, since 6/8 point mutations in AGA result in an amino acid change, while CGA is assigned a lower volatility of 0.5, because only 4/8 mutations are nonsynonymous. The method proposed by Plotkin and colleagues rests on the assumption that regions undergoing extensive amino acid substitution should on average contain an excess of highly volatile codons in comparison to the genome at

large. In essence, a highly volatile codon is regarded as fossil evidence of a recent episode of positive Darwinian selection.

Although the notion of using a single sequence to detect selection may be attractive to many investigators, there are concerns about the volatility approach arising from the very nature of the genetic code itself. Mean codon volatilites for each amino acid vary considerably (from 0.653 for leucine to 1.0 for the non-degenerate methionine and tryptophan). As a result, a gene's observed volatility is greatly influenced by its amino acid composition (6). Controlling for overall amino acid composition does not adequately solve this problem, since codons for only 4 out of the 20 amino acids exhibit any variation in volatility whatsoever. The frequency of these four amino acids (arginine, glycine, leucine, and serine) inevitably has a disproportionate effect on a gene's adjusted relative volatility (*P* value).

An additional caveat stems from the observation that GC content and codon usage may vary considerably within genomes due to factors unrelated to selection at the protein level (32,40). These intragenomic fluctuations are often driven by the effects of nucleotide sequence on DNA or RNA structure (2), as well as the relative abundance of transfer RNA molecules which modulate the rate at which a given codon is translated (20). Nevertheless, Plotkin et al demonstrate that volatility P values and dN/dS estimates obtained using comparative methods align quite well in their test cases of M. tuberculosis and P. falciparum (26). In this study, we apply the codon volatility approach to measure selection on the HIV-1 genome, and systematically

compare this technique against established maximum likelihood- and distance-based comparative methods.

MATERIALS AND METHODS

Acquisition and preparation of sequence data. All available full-length subtype B HIV-1 genomes in the Los Alamos National Laboratory HIV Sequence Database were downloaded and aligned using Multalin (5). Sequences containing frameshifts, premature stop codons, or ambiguities were removed from the data set. Open reading frames for the remaining 92 sequences were determined for gag, protease, reverse transcriptase, integrase, env, and nef. These coding regions were then extracted from each sequence and aligned using Clustal X (36), with default gap parameters and the "DNA 5-0" substitution matrix. Subsequent manual aligning was performed using the Se-Al sequence alignment editor (27). All gene regions associated with overlapping reading frames were deleted from the data sets. Sequence alignments are available for download at: http://supersatish.com/volatility. Genbank accession numbers involved in our analysis are: A04321, AB078005, AB097870, AF003887-AF003888, AF004394, AF042100-AF042101, AF049494-AF049495, AF069140, AF070521, AF146728, AF256206, AF286365, AF538302-AF538304, AF538306-AF538307, AJ006287, AJ271445, AY173951-AY173954, AY173956, AY180905, AY308761-AY308762, AY314044-AY314063, AY331283-AY331284, AY331296, AY332236, AY352275, AY423384, AY560107-AY560108, D10112, D86068-D86069, K02007, K02013, K02083, L02317, L31963, M17451, M19921,

M26727, M38429, M93258, U12055, U23487, U26942, U34603-U34604, U39362, U43096, U43141, U63632, U69584-U69593, U71182, Z11530.

Calculation of codon volatility *P* values. Gene-specific volatility *P* values for each genome were obtained by implementing the command-line version of the volatility server, kindly provided by Dr. J. Plotkin. In brief, the volatility of each codon, v(C), is calculated as follows:

$$v(C) = N/T$$

where N is the number of nonsynonymous codons that differ from codon C at a single nucleotide position and T is the total number of (non-termination) codons that differ from codon C at a single nucleotide position. A gene's volatility, v(G), is the sum of the volatilities of its codons. The volatility *P* value for each gene is calculated by comparing v(G) against the volatility of the remaining genome, adjusting for amino acid composition and length (26).

Estimation of dN/dS using comparative methods. Three separate programs were implemented to obtain gene-specific estimates of dN/dS: the command line version of the Synonymous NonSynonymous Analysis Program (SNAP) (13), Hypothesis Testing Using Phylogenies (HyPhy) (15), and Phylogenetic Analysis by Maximum Likelihood (PAML) (38).

The Synonymous NonSynonymous Analysis Program (SNAP) is a convenient implementation of the method originally developed by Nei and Gojobori (21) that calculates the number of synonymous and nonsynonymous base substitutions for all pairwise comparisons of sequences in an alignment. The number of actual

synonymous and nonsynonymous codon changes between each pair of sequences are counted, as well as the number of potential synonymous and nonsynonymous changes. The reported dN/dS ratio for each comparison is the proportion of observed nonsynonymous substitutions divided by the proportion of observed synonymous substitutions, adjusting for multiple hits using the Jukes-Cantor correction (11).

The REL (Random Effects Likelihood) method (14) fits two independent distributions to synonymous and nonsynonymous substitution rates and infers whether a site is under selection by computing empirical Bayes Factors for the event that dN>dS (or dN<dS) at any fixed sites. Recent results (14,15) suggest that failure to allow silent substitution rates to vary among codon sites may lead to biased estimated of overall dN/dS and misidentification of hypervariable sites as being under selection. A web implementation of the REL method is available at: http://www.datamonkey.org/

CODEML (version 3.13) is available in the PAML package of programs and utilizes a number of different models of codon evolution within a maximum likelihood framework to estimate selection pressures for each codon site in a multiple alignment (39). The average dN/dS for each alignment was estimated by the M8 model (positive selection); M7 (null model) was rejected in all cases using a likelihood ratio test. Since CODEML requires phylogenetic trees as input, the PAUP* package (35) was used to construct maximum likelihood phylogenetic trees under the HKY85+G model using nearest neighbor interchange branch swapping on an initial tree constructed by the neighbor joining method.

Method comparison. A Spearman rank correlation coefficient was computed using JMP Version 5.1 (30) for all pairwise comparisons between selection detection methods.

Prediction of coreceptor usage. A support vector machine-based method was employed to predict the coreceptor usage of viruses based on V3 loop amino acid sequence (24). This method is reported to predict CXCR4 usage with a specificity of 93% (10). The coreceptor classifier is available for public use at: http://genomiac2.ucsd.edu:8080/wetcat/tropism.html

Codon Usage Analysis. The General Codon Usage Analysis (GCUA) package was implemented to look for coreceptor phenotype-specific codon usage patterns (17).

RESULTS AND DISCUSSION

To investigate how gene volatility varies in relation to comparative estimates of selection intensity across the HIV-1 genome, we compiled and analyzed a data set consisting of all available full-length subtype B HIV-1 sequences. Multiple sequence alignments were generated for *gag*, *protease*, *reverse transcriptase* (*RT*), *integrase*, *env*, and *nef*, excluding all coding regions associated with overlapping reading frames. Mean volatility *P* values, ranked from highest to lowest, were as follows (Fig. 1): *nef*, 0.97; *protease*, 0.63; *gag*, 0.62; *RT*, 0.40; *env*, 0.37; *integrase*, 0.12. This hierarchy is in conflict with our basic understanding of HIV-1 biology. For example, the rate of evolution of *env* has been estimated at 1-2% a year based on longitudinal intrapatient data and is the highest of all HIV-1 genes (31). Positive selection intensity is expected to be highest on Env due to its exposed location on the virion surface with its rapid

evolution in response to neutralizing antibodies and its role as the primary determinant of cellular tropism in a diverse target cell environment (1,29).

Selection pressure is typically described as the ratio between nonsynonymous substitutions per nonsynonymous site (dN) and synonymous substitutions per synonymous site (dS) (21). We calculated global, gene-specific estimates of dN/dS using three established comparative methods: the maximum likelihood-based approaches of Nielsen-Yang and Kosakovsky Pond, and the distance-based method of Nei and Gojobori (15,21,39). The mean gene-specifc estimates of dN/dS across all three methods, ranked from highest to lowest, were as follows (Fig. 2): *env*, 1.30; *nef*, 0.97; *gag*, 0.42; *protease*, 0.32; *integrase*, 0.28; *RT*, 0.22. These findings are in accordance with our concept of viral biology; the surface antigen is under the strongest positive selection due to immune pressure, while structural proteins and enzymes are conserved due to functional constraint.

We calculated a Spearman's rank correlation coefficient for pairwise comparisons between all methods. Results obtained using the volatility approach failed to correlate with dN/dS estimates derived via all three comparative methods in our study ($0.50<p<1$). In contrast, the comparative methods were internally consistent, with p-values ranging from 0.016 to 0.058 (Table 1). The tight correlation between dN/dS estimates derived via Nei-Gojobori, Nielsen-Yang and REL speaks to the robustness of the comparative framework, since there is considerable methodological divergence between these techniques (14).

The chemokine receptor preference of an HIV-1 strain is primarily determined by the third variable region (V3) of the envelope glycoprotein (7). The V3 sequences from CCR5-using (R5) strains have been reported to be more resistant to positive selection pressure than CXCR4-using (X4) variants (33). As an additional test of the volatility method, we attempted to replicate this finding by determining the coreceptor phenotype of our data set and measuring positive selection pressure on both classes of V3 sequences. We predicted the chemokine receptor usage of all 92 viral strains in our data set based on V3 amino acid sequence using a previously trained machine learning algorithm (24). Twenty-six sequences were predicted to use the CXCR4 receptor, while the remaining sixty-six were classified as R5 variants. We measured selection intensity in both viral populations (R5 and X4) using the Nei-Gojobori, REL, and codon volatility approaches. Our dN/dS estimates were significantly higher in the X4 subset, in accordance with earlier reports (Table 2). Once again, observed volatility *P* values failed to correlate with dN/dS estimates; mean *P* values were significantly lower for the CCR5-using subset of V3 sequences, suggesting that there was *less* evidence of positive selection in the X4 class (Table 2).

Some of the most extreme examples of positive selection in nature are found in the variable regions of the HIV-1 envelope (34). Our observation that the codon volatility method fails to appropriately detect positive selection within the V3 region of HIV-1 *env* further refutes the claim that volatility is a reflection of selection pressure. To identify the foundation of the volatility method, we investigated the relationship between sequence variation in our data set and observed volatility *P*

values. We looked for evidence of differential codon usage using the GCUA (General Codon Usage Analysis) package (17). There were no significant differences in codon usage patterns between R5 and X4 V3 loop sequences (data not shown), pointing to another cause for the observed discrepancy in mean volatility *P* values. We calculated correlation coefficients between the relative composition of each of the twenty amino acids and the mean volatility *P* values for the six HIV-1 genes involved in our earlier analysis (Table 3). The frequency of arginine was most strongly correlated with volatility *P* values (correlation coefficient = 0.691). We then compared the amino acid compositions of R5 and X4 V3 loop sequences. Arginine was the amino acid that exhibited the greatest average difference in relative composition between these two classes. The mean arginine content of X4 V3 loops was 18.4%, in contrast to 10.5% for the R5 variants, reflecting the higher net positive charge of the X4 class (7,24). The codon AGA was used to encode arginine preferentially in all V3 loop sequences. Taken together, these data strongly suggest that differential amino acid composition, rather than codon composition, was responsible for the observed discrepancies in volatility *P* values.

Our observations are consistent with recent reports demonstrating that volatility fails to correlate with selection as measured by comparative methods in the cases of *M. tuberculosis*, *M. bovis* and *E. coli* (6,40). In addition, several reports have emerged discrediting the codon volatility method based on theoretical concerns. Computer simulations of sequence evolution demonstrate that directional selection has no effect on codon volatility, and volatility can increase in the absence of positive

selection (6,22,40). The inherent methodological limitation associated with considering only 4 out of the 20 amino acids (arginine, glycine, leucine, and serine) is exacerbated by the observation of Chen and colleagues that serine codon usage exerts a disproportionately large influence on volatility *P* values (3). Hahn *et al* observed that a gene's codon adaptation index (CAI), used to predict its expression level, explains a much larger proportion of variance in volatility than selection (as measured by comparative methods) (9). Similarly, the recent analysis of eukaryotic genomes by Friedman and Hughes revealed that nucleotide content at the second codon position was a much more powerful correlate of elevated codon volatility than selection intensity (8).

There are features particular to the HIV-1 genome that make it an especially unattractive subject for the codon volatility method. Firstly, HIV, like RNA viruses in general, has a relatively high mutation rate of ~3.4*10-5 mutations/site/generation (16). Given that 10^{10} virus particles are produced each day within an infected host, there is a considerable probability that a nonsynonymous substitution in the HIV genome will be masked by a subsequent synonymous substitution at the same site prior to being sampled (23). Therefore, positive selection is likely to be underestimated on average using the volatility approach. Another complicating factor stems from the observation that mutation rate itself varies across the HIV-1 genome (14); evidence of recent positive selection would be expected to erode at different rates at different sites, skewing volatility scores. Moreover, HIV-1 undergoes recombination at a minimum rate of 2.8 crossovers per genome per cycle (41). This

undoubtedly influences per-gene estimates of volatility, although the direction and magnitude of this effect likely depend on the precise location of break points as well as the genetic distance between parental strains. Lastly, the HIV-1 genome is short (<10kb) and contains only nine genes. There is very little statistical power available for intragenomic comparisons.

Determining which genes and which sites within genes are under the greatest or least selection pressures will be important in the rational design of a HIV vaccine (4,12,25). A selection detection method that requires only a single representative genome such as 'codon volatility' would be attractive. However, we have demonstrated that codon volatility is not a reliable indicator of selective pressure on the HIV genome.

ACKNOWLEDGMENTS

The text of this chapter, in full, is a reprint of the material as it appears in Virology: Pillai, S.K., S. L. Kosakovsky Pond, C.H. Woelk, D.D. Richman, and D.M. Smith, "'Codon Volatility' Does Not Reflect Selective Pressure on the HIV-1 Genome" (in press). I was the primary author, and the co-authors listed in this manuscript supervised and/or contributed to the research which forms the basis for this chapter.

These data were presented in part at the 12th Conference on Retroviruses and Opportunistic Infections, Boston, MA, February 2005. We are grateful to Joseph Wong, MD and Simon Frost, Ph.D. for their insightful comments, and Sharon Wilcox and Darica Smith for helping with the preparation of the manuscript. This work was

supported by grants 5K23 AI055276, AI27670, AI043638, the Adult AIDS Clinical Trials Group funded by the National Institute of Allergy and Infectious Diseases, and the AACTG Central Group Grant (U01AI38858), the UCSD Center for AIDS Research (AI 36214), AI29164, AI047745, AI07384, from the National Institutes of Health, and the Research Center for AIDS and HIV Infection of the San Diego Veterans Affairs Healthcare System.

REFERENCES

1. Baribaud, F., Edwards, T. G., Sharron, M., Brelot, A., Heveker, N., Price, K., Mortari, F., Alizon, M., Tsang, M., and Doms, R. W. 2001. Antigenically Distinct Conformations of CXCR4. J Virol 75:8957-8967.

2. Bram, S. 1971. Secondary Structure of DNA Depends on Base Composition. Nat New Biol 232:174-176.

3. Chen, Y., Emerson, J. J., and Martin, T. M. 2005. Evolutionary genomics: codon volatility does not detect selection. Nature 433:E6-7

4. Choisy, M., Woelk, C. H., Guegan, J. F., and Robertson, D. L. 2004. Comparative Study of Adaptive Molecular Evolution in Different Human Immunodeficiency Virus Groups and Subtypes. J Virol 78:1962-1970.

5. Corpet, F. 1988. Multiple Sequence Alignment with Hierarchical-Clustering. Nucleic Acids Res 16:10881-10890.

6. Dagan, T. and Graur, D. 2004. The Comparative Method Rules! Codon Volatility Cannot Detect Positive Darwinian Selection Using a Single Genome Sequence. Mol Biol Evol [Epub].

7. Fouchier, R. A., Groenink, M., Kootstra, N. A., Tersmette, M., Huisman, H. G., Miedema, F., and Schuitemaker, H. 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J Virol 66:3183-3187.

8. Friedman, R., and Hughes, A. L. 2005. Codon volatility as an indicator of positive selection: data from eukaryotic genome comparisons. Mol Biol Evol 22:542-6.

9. Hahn, M. W., Mezey, J. G., Begun, D. J., Gillespie, J. H., Kern, A. D., Langley, C. H., and Moyle, L. C. 2005. Codon bias and selection on single genomes. Nature 433:E5-6

10. Jensen M. A. and van 't Wout, A. 2003. Predicting HIV-1 coreceptor usage with sequence analysis. AIDS Rev 5:104-112.

11. Jukes, T. and and Cantor, C. 1969. Evolution of protein molecules. In "Mammalian Protein Metabolism." (H. Munro, Ed.), pp. 21-132. Academic Press, New York.

12. Kemal, K. S., Foley, B., Burger, H., Anastos, K., Minkoff, H., Kitchen, C., Philpott, S. M., Gao, W., Robison, E., Holman, S., Dehner, C., Beck, S., Meyer, W. A., Landay, A., Kovacs, A., Bremer, J., and Weiser, B. 2003. HIV-1 in genital tract and plasma of women: Compartmentalization of viral sequences, coreceptor usage, and glycosylation. Proc Natl Acad Sci USA 100:12972-12977.

13. Korber, B. 2001. HIV Signature and Sequence Variation Analysis. In (Allen G. Rodrigo and Gerald H. Learn, Eds.), pp. 55-72. Kluwer Academic Publishers, Dordrecht, Netherlands.

14. Kosakovsky Pond S. L. and Frost, S. D. W. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol and Evol (in press).

15. Kosakovsky Pond, S. L., Frost, S. D. W., and Muse, S. V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-679.

16. Mansky, L. M. and Temin, H. M. 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol 69:5087-5094.

17. McInerney, J. O. 1998. GCUA: general codon usage analysis. Bioinformatics 14:372-373.

18. Miller G. A., Beckwith R., Fellbaum, C. D., Gross, D., and Miller, K. Five Papers on WordNet. 1993. Princeton, N.J., Princeton University.

19. Miyata, T. and Yasunaga, T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol 16:23-36.

20. Moriyama, E. N. and Powell, J. R. 1997. Codon usage bias and tRNA abundance in Drosophila. J Mol Evol 45:514-523.

21. Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426.

22. Nielsen, R. and Hubisz, M. J. 2005. Evolutionary genomics: Detecting selection needs comparative data. Nature 433:E7-E8.

23. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. 1996. HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582-1586.

24. Pillai S. K., Good, B., Richman, D., and Corbeil, J. 2003. A new perspective on V3 phenotype prediction. AIDS Res Hum Retroviruses. 19:145-149.

25. Pillai, S. K., Good, B., Pond, S. K., Wong, J. K., Strain, M. C., Richman, D. D., and Smith, D. M. 2005. Semen-Specific Genetic Characteristics of Human Immunodeficiency Virus Type 1 *env*. J Virol 79:1734-1742.

26. Plotkin, J. B., Dushoff, J., and Fraser, H. B. 2004. Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. Nature 428:942-945.

27. Rambaut A. 2002. Se-Al sequence alignment editor v2.0 (Software). Department of Zoology, University of Oxford.

28. Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. 2004. The Causes and Consequences of HIV Evolution. Nat Rev Genet 5:52-61.

29. Richman, D. D., Wrin, T., Little, S. J., and Petropoulos, C. J. 2003. Rapid evolution of the neutralizing antibody response to HIV type 1 infection. Proc Natl Acad Sci USA 100:4144-4149.

30. Sall, J., Lehman, A., and Creighton, L. JMP Start Statistics. 2001. Pacific Grove, CA: Duxbury Press.

31. Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C. R., Learn, G. H., He, X., Huang, X. L., and Mullins, J. I. 1999. Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection. J Virol 73:10489-10502.

32. Sharp, P. M. 2004. Gene "volatility" is Most Unlikely to Reveal Adaptation. Molecular Biology and Evolution [Epub].

33. Shiino, T., Kato, K., Kodaka, N., Miyakuni, T., Takebe, Y., and Sato, H. 2000. A Group of V3 Sequences from Human Immunodeficiency Virus Type 1 Subtype E Non-Syncytium-Inducing, CCR5-Using Variants Are Resistant to Positive Selection Pressure. J Virol 74:1069-1078.

34. Simmonds, P., Zhang, L. Q., McOmish, F., Balfe, P., Ludlam, C. A., and Leigh Brown, A. J. 1991. Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 *env* sequences in plasma viral and lymphocyte-associated proviral populations in vivo: implications for models of HIV pathogenesis. J Virol 65:6266-6276.

35. Swofford D.L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Sunderland, Mass, Sinauer Associates.

36. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876-4882.

37. Yang, W., Bielawski, J. P., and Yang, Z. H. 2003. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J Mol Evol 57:212-221.

38. Yang, Z. and Nielsen, R. 2000. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Mol Biol Evol 17:32-43.

39. Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A. M. K. 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics 155:431-449.

40. Zhang, J. 2004. On the evolution of codon volatility. Genetics 169:495-501.

41. Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B. D., and Dougherty, J. P. 2002. Human Immunodeficiency Virus Type 1 Recombination: Rate, Fidelity, and Putative Hot Spots. J Virol 76:11273-11282.

FIGURES AND TABLES

Table 1: Pairwise comparison of selection detection methods

comparison	Spearman's rank correlation coefficient	p-value	Z-score
Volatility vs. REL	$R_s = 0.0286$	p<= 1	Z = 0.0639
Volatility vs. Yang	$R_s = 0.3143$	p<= 0.5639	Z = 0.7028
Volatility vs. Nei	$R_s = 0.3286$	p<= 0.4972	Z = 0.7347
REL vs. Yang	$R_s = 0.8286$	p<= 0.0583	Z = 1.8527
REL vs. Nei*	$R_s = 0.9000$	p<= 0.0167	Z = 2.0125
Yang vs. Nei*	$R_s = 0.9286$	p<= 0.0167	Z = 2.0763

The results of all four methods were compared in a pairwise fashion using a Spearman's rank test. The rank correlation coefficient, $Rs = 1-(6\Sigma d^2/n^3-n)$, where n = number of ranks and d = difference between ranks. Asterisks indicate significant correlations ($p<0.05$).

Table 2: Inferred positive selection pressure on R5 and X4 V3 loop sequences

Statistic	R5 mean (c.v.)	X4 mean (c.v.)	p-value
Volatility *P* value	0.052 (1.154)	0.288 (0.861)	$p < 0.0001$
dN/dS (Nei-Gojobori)	0.565 (1.619)	1.282 (1.184)	$p < 0.0001$
dN/dS (REL)	0.670 (1.110)	1.467 (0.678)	$p < 0.012$

Positive selection pressure was estimated using the Nei-Gojobori, REL, and codon volatility approaches. Coefficients of variation for each mean estimate are listed in parentheses. Method-specific estimates were compared between R5 and X4 classes using either a two-tailed Mann-Whitney test (volatility and Nei-Gojobori methods) or a likelihood ratio test (REL).

Table 3: Correlations between amino acid frequencies and gene-specific volatility *P* values

Amino acid	Gag (0.62)	Protease (0.63)	RT (0.40)	Integrase (0.12)	Env (0.37)	Nef (0.97)	Correlation coefficient
arg	6.1	3.5	2.9	4.2	4.7	7.7	0.6906
met	3.5	2.3	1.6	1.7	2	2.7	0.6439
glu	7.9	4.7	8.6	6.3	5.5	11.8	0.6189
pro	4.8	4.7	7.9	3.5	3.3	7.2	0.5200
leu	7	12.8	8.1	5.2	8.4	7.7	0.3665
his	2	1.2	1.6	2.8	1.4	3.6	0.3344
gly	7.2	15.1	5.4	8	5.9	7.7	0.2304
ala	10.3	3.5	3.8	8.7	6.1	8.6	0.1218
tyr	1.8	1.2	3.8	2.8	2.4	3.6	0.0168
ser	5.7	0	3.2	4.5	5.9	5	-0.0611
phe	1.5	2.3	3.2	2.8	2.9	3.2	-0.0683
asn	5.5	4.7	2.9	3.1	9.7	3.6	-0.0758
trp	1.8	1.2	4.1	2.4	3	3.2	-0.0783
cys	2.2	2.3	0.5	2.1	2.6	1.4	-0.1479
thr	5.7	7	7	4.2	8.7	3.2	-0.3317
asp	2.4	4.7	4.3	6.6	3	4.1	-0.4556
ile	5.3	12.8	7.2	8.3	7.7	1.8	-0.4807
lys	7.2	5.8	11.3	9	5.2	4.5	-0.6138
val	5.3	5.8	6.6	7.6	6.6	6.3	-0.6590
gln	6.8	4.7	5.9	6.3	4.9	3.2	-0.6760

Amino acid frequencies reported as percentages of overall composition. Mean volatility *P* values for each gene are indicated in parentheses. Pearson correlation coefficients were calculated between residue frequencies and gene volatilities.

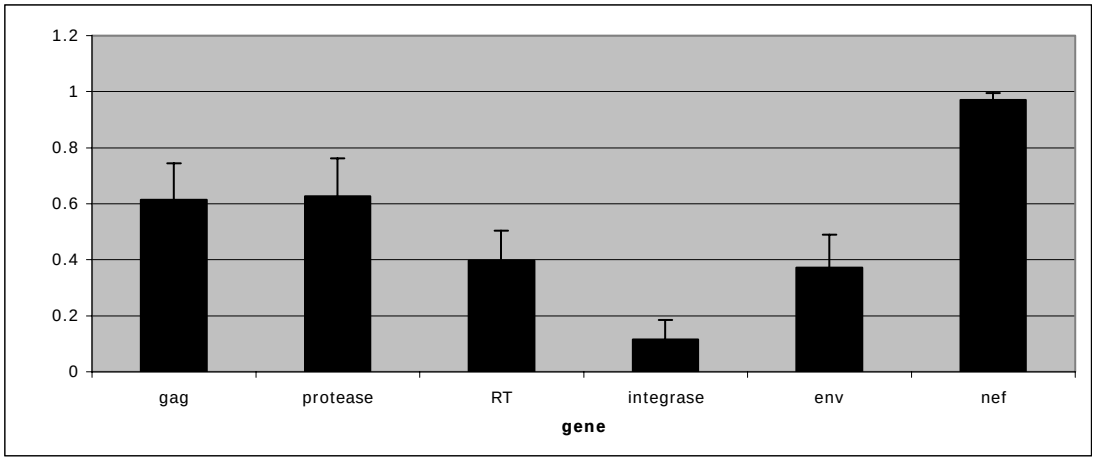


Figure 1: Mean volatility *P* values across the HIV-1 genome. Gene-specific volatilities were calculated for 92 full-length subtype B HIV-1 genomes. Error bars represent s.d..

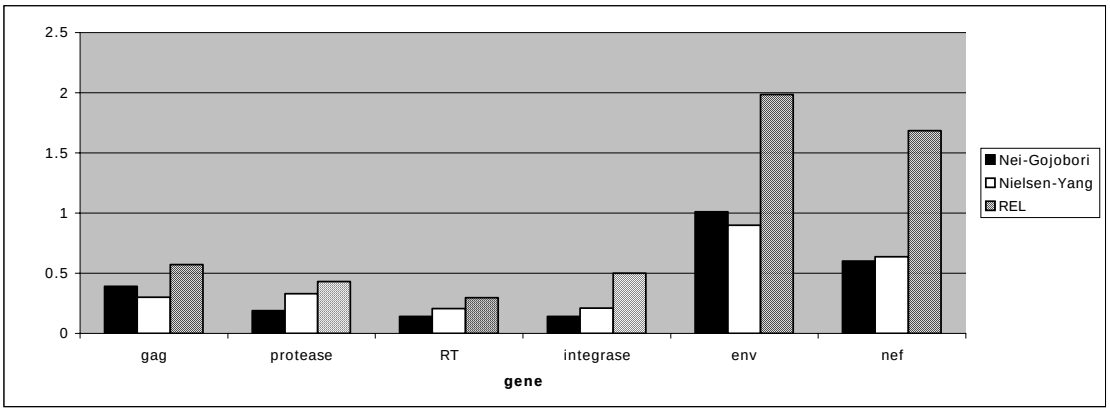


Figure 2: Comparative estimates of selection intensity across the HIV-1 genome. Three comparative methods (Relative Effects Likelihood, Nielsen-Yang, and Nei-Gojobori) were employed to calculate gene-specific dN/dS estimates for 92 full-length subtype B HIV-1 genomes.

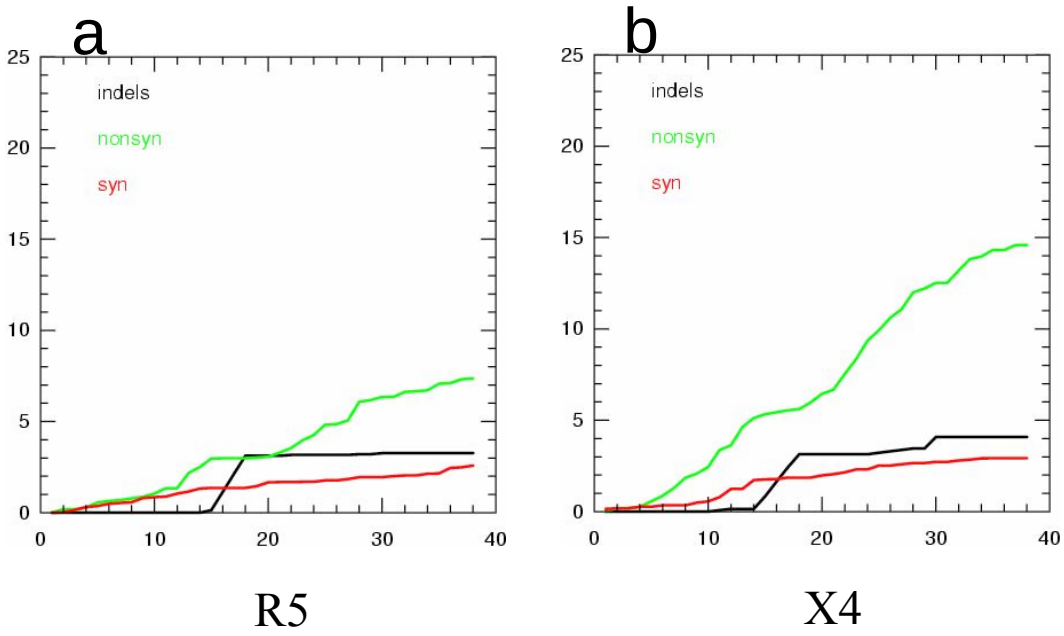


Figure 3: Cumulative behavior of the average synonymous and nonsynonymous substitutions, moving codon by codon across (a) CCR5-using (R5) and (b) CXCR4-using (X4) V3 sequences.

Chapter 4

Semen-Specific Genetic Characteristics of Human Immunodeficiency Virus Type 1 *env*

ABSTRACT

Human immunodeficiency virus type 1 (HIV-1) in the male genital tract may comprise virus produced locally in addition to virus transported from the circulation. Virus produced in the male genital tract may be genetically distinct, due to tissue-specific cellular characteristics and immunological pressures. HIV-1 *env* sequences derived from paired blood and semen samples from the Los Alamos HIV Sequence Database were analyzed to ascertain a male genital tract-specific viral signature. Machine learning algorithms could predict seminal tropism based on *env* sequences with accuracies exceeding 90%, suggesting that a strong genetic signature does exist for virus replicating in the male genital tract. Additionally, semen-derived viral populations exhibited constrained diversity ($P < 0.05$), decreased levels of positive selection ($P < 0.025$), decreased CXCR4 coreceptor utilization, and altered glycosylation patterns. Our analysis suggests that the male genital tract represents a distinct selective environment that contributes to the apparent genetic bottlenecks associated with the sexual transmission of HIV-1.

INTRODUCTION

Most human immunodeficiency virus (HIV) transmission events globally occur via mucosal exposure to male genital secretions carrying the virus (34, 46). Although the risk of sexual HIV transmission correlates with the amount of virus present in the blood of the source partner (36), the correlation between the viral load in the blood and genital compartment is inconsistent (3, 23, 24). The biological

determinants that influence the transmissibility of different viral variants from within the genital tract of the HIV-infected source are still incompletely understood. Since transmitted virus represents the initial virus that the immune system encounters, the understanding of its composition will be critical in our attempts to develop a successful HIV vaccine (1, 7, 54).

HIV in each chronically infected person exists as a diverse population of related genetic variants (5, 12, 20). Anatomic compartmentalization of these variants has been described in blood, lung, central nervous system, and genital tract (10, 16, 17, 20, 21, 32, 41, 50, 53). Male genital tract tissues (e.g., the prostate, seminal vesicles, and epididymis) serve as sites of viral replication and are likely to differ from peripheral tissues in immunological surveillance, target cell characteristics, and efficiencies of drug penetration (10, 17, 43). Virus replicating within the male genital tract could therefore develop distinct, compartment-specific characteristics in response to these local selective pressures (10, 16, 17, 20, 21, 32, 41, 50, 53). Although genetic differences between blood- and semen-derived HIV in an individual have been documented, a seminal signature sequence remains elusive (6, 10). This failure to identify a signature sequence could be attributable to the fact that previous efforts mainly focused on proviral DNA sequences, which often represent archival viral genotypes rather than contemporary, actively replicating variants (4, 44).

We investigated viral genetics and compartmentalization within the male genital tract by applying a battery of computational techniques to paired semen- and blood-derived HIV-1 RNA *env* sequences. Our results suggest that the male genital

tract can represent a legitimate viral compartment, although this compartmentalization is not absolute. Furthermore, when viral migration between blood plasma and the male genital tract is minimal and infrequent, there are several distinct genetic features associated with semen-derived HIV variants. Understanding these tissue-specific properties of HIV type 1 (HIV-1) will likely be crucial for the development of an effective vaccine.

MATERIALS AND METHODS

Sequence data. All of the semen-derived HIV-1 *env* sequences from the Los Alamos National Lab HIV Sequence Database with accompanying subject identification were downloaded. Blood-derived sequences from the same individuals were downloaded; semen sequences without matching blood data were removed from the set. GenBank database accession numbers included in our analysis are AF098718 to AF098734, AF256230 to AF256465, AF373037 to AF373043, AF535219 to AF535859, AY005164 to AY005179, U00821 to U00843, U13381 to U13388, and U96502 to U96608. Duplicates, sequences derived by direct PCR sequencing, proviral DNA sequences, and nonfunctional open reading frames (containing frameshifts, premature stop codons, etc.) were deleted. The final set consisted of 659 *env* C2-V3 RNA sequences (spanning HXB2 coordinates 799 to 1410) from a total of 12 patients (376 plasma and 283 semen samples).

Phylogenetic reconstruction. Initial multiple sequence alignments were generated by using Multalin (8), with default gap parameters and the DNA 5-0 substitution matrix. Subsequent manual aligning was performed by using the Se-Al

sequence alignment editor (37). Phylogenies describing sequences from each individual host were built by using FastDNAml (30), estimating base frequencies from the data and a transition/transversion ratio of 2.0. All diversity and divergence measurements were calculated by using dnadist (14). The absolute rate of molecular evolution (molecular clock) was estimated by running TipDate (38) on maximum likelihood phylogenies with dated tips. A master tree describing the entire data set was built by implementing dnadist and neighbor within the PHYLIP version 3.5c software package (14) by using the F84 model, gamma distributed rates across sites, and a transition/transversion ratio of 2.0. Trees were viewed with TreeView X (31).

Evaluation of compartmentalization. The degree of segregation between compartments was assessed by testing for panmixis by using gene phylogenies (18, 42) as implemented in the MacClade program (Sinauer, Sunderland, Mass.). In brief, the minimum possible number of intercompartment migration events was tallied, based on the maximum likelihood trees for each individual subject's C2-V3 sequences and their characterization according to compartment of origin. This result was compared to the distribution of migration events for 1,000 randomly generated trees. Evidence of restricted gene flow (compartmentalization) was documented when <1% of the random trees required the same or fewer number of migration events as for the sample data (29).

Machine learning classification. A machine learning approach was employed to look for a tissue-specific genetic signature. All classification experiments in this analysis were conducted by using WEKA (Waikato environment for knowledge

analysis), an open source collection of data processing and machine learning algorithms (49). The J48 decision tree inducer, based on the C4.5 algorithm (35) was implemented with the parameter "MinNumObj" set at a value of 7 to limit the complexity of theories and minimize the risk of overfitting. Classifiers were evaluated by using 100 iterations of stratified 10-fold cross-validation, a procedure designed to reflect the performance of classification models on novel data sets. For each of 100 trials, the data set was randomly divided into 10 groups of approximately equal size and class distribution. For each "fold," the classifier was trained by using all but 1 of the 10 groups and then tested on the unseen group. This procedure was repeated for each of the 10 groups. The cross-validation score for one trial was the average performance across each of the 10 training runs. The reported score is the average across the 100 trials (49). In addition, we have reported the true positive rate (TPR) and precision for these classification experiments: TPR = [number of true positives/(number of true positives + number of false negatives)]; precision = [number of true positives/(number of true positives + number of false positives)].

Analysis of Selection. A maximum likelihood method was used to detect and quantify positive and negative selection. All data sets were first evaluated by using a model selection procedure (22) to identify and correct for strong nucleotide substitution biases which are ubiquitous in HIV. The fixed-effects likelihood (FEL) approach (22) was employed to test for selective pressure at a given site. Maximum likelihood estimates of branch lengths and nucleotide substitution rate parameters were derived from the entire alignment. A full codon model, using a modified MG94

(28) rate matrix with site-specific instantaneous synonymous (alphas) and nonsynonymous (betas) rates was then fitted independently to every codon position in the data, under two hypotheses: H_0, neutral evolution (alphas equal betas); H_A, nonneutral evolution (alphas and betas are free to vary independently).

When the hypothesis of neutrality was rejected at site s, it was called positively selected if betas was estimated to be greater than alphas. The FEL method was implemented on a cluster of computers by using the HyPhy package (22).

Coreceptor usage prediction. A support vector machine-based method was employed to predict the coreceptor usage of viruses based on the V3 loop amino acid sequence (33). This method is highly reliable and is reported to predict CXCR4 usage with a specificity of 93% (19). The coreceptor classifier is available for public use at: http://genomiac2.ucsd.edu:8080/wetcat/tropism.html.

Glycosylation. GlycoTracker.pl (S. Pillai, unpublished data) was used to identify N-linked glycosylation sites within each sequence. The Perl script provides a tally of all sequons, along with their respective locations (numbered according to HXB2 gp160). We compared the extent and distribution of N-linked glycosylation across the C2-V3 region in both compartments by identifying NXS and NXT (where X is some other residue) motifs in plasma- and semen-derived sequences (25). All statistical comparisons were performed by using a Wilcoxon Mann-Whitney test (11).

Codon usage analysis. The general codon usage analysis (GCUA) package was implemented to look for compartment-specific codon usage biases (26).

RESULTS

Compartmentalization of semen-derived virus.

To determine if the male genital tract represents a viral compartment, we used systematic phylogenetic comparison of matched blood- and semen-derived HIV-1 RNA *env* sequences from 12 individuals. We hypothesized that if the male genital tract is indeed a viral compartment, semen-derived sequences within each individual should cluster independently, while exhibiting similar levels of diversity and divergence as matching plasma sequences given comparable effective population sizes (29). Maximum likelihood trees describing contemporaneous variants from both tissues revealed that the male genital tract represented a distinct virologic compartment in six individuals (identified as A to F) (Fig. 1a; see Fig. S1 in the supplemental material), based on phylogenetic segregation between blood and semen virus. In five of the individuals, sequences did not cluster with respect to compartment (Fig. 1b; see Fig. S3 in the supplemental material). In one individual, G, there were longitudinal data that showed compartmentalization at the earlier time points but then apparent panmixis at later time points (see Fig. S2 in the supplemental material). In accordance with previous reports, a neighbor-joining tree comprising pooled data from all compartmentalized patients revealed that host, rather than compartment of origin, was the strongest phylogenetic determinant (see Fig. S4 in the supplemental material).

Genetic diversity in plasma- and semen-derived viral populations.

Genetic diversity was characterized by calculating the average pairwise distance within a population, based on distance measurements obtained by using the

F84 matrix. Data across multiple time points were pooled when available. Individuals with phylogenetically distinct virus in blood and semen consistently exhibited lower genetic diversity in semen-derived viral populations ($P < 0.01$ by a paired Wilcoxon test). Conversely, individuals with noncompartmentalized virus failed to demonstrate any significant differences in viral diversity between tissues (Fig. 2).

Analysis of longitudinal sequence data.

Longitudinal sequence data spanning multiple years were available for five individuals (identified as F, G, I, J, and K). We first evaluated tissue-specific longitudinal genetic diversity in these individuals by computing average pairwise genetic distances for each time point where blood and semen sequences were available. The longitudinal data reinforced our aforementioned results; individual F, characterized by compartmentalized virus at all available time points, exhibited constrained viral diversity in semen throughout the 2-year monitored period (Fig. 3a). Individual G, who transitioned from compartmentalized to noncompartmentalized virus, showed considerable variation in tissue-specific diversity; semen diversity bounced between being greater and less than contemporaneous plasma diversity, in accordance with inconsistent trafficking between these tissues. Individuals I, J, and K were consistently characterized by noncompartmentalized virus and exhibited similar levels of viral diversity in blood and semen at nearly all sample points (see Fig. S5 in the supplemental material).

We next looked at longitudinal divergence in these five individuals, by calculating the average genetic distance from sequences at each time point to an

artificial, tissue-specific baseline consensus sequence. On average, the observed level of divergence was comparable across tissues in individuals with both compartmentalized and noncompartmentalized virus, consistent with actively replicating viral populations in both blood and male genital tract (see Fig. S5 in the supplemental material). We also calculated the divergence between blood- and semen-derived virus by computing the average genetic distance between these populations at each time point. Individual F as expected demonstrated continually increasing divergence between tissue-specific populations, most probably due to a combination of genetic drift and compartment-specific viral adaptation. Intercompartment genetic distance exceeded 5% at the last available sample point (Fig. 3b). Individual G showed declining intercompartment divergence at each time point, mirroring the increased contribution of systemic virus to the seminal viral population. Divergence steadily diminished from approximately 8% at the onset to 2% at the final sampling time. Finally, hosts I, J, and K characterized by noncompartmentalized virus maintained low levels of intercompartment divergence throughout the monitored period; distances stayed below 2% at nearly all time points (see Fig. S5 in the supplemental material).

Estimation of molecular clock.

We used dated maximum likelihood phylogenies of sequences from host F, the only individual with compartmentalized virus and with available longitudinal data, to compare the viral molecular clock between plasma and semen. The estimated absolute rates of molecular evolution based on these phylogenies were 0.01004877 and

0.00637917 substitutions/site/year for plasma- and semen-derived sequences, respectively.

Semen-specific *env* genetic signature.

Although phylogenetic evidence suggests that semen- and blood-derived viruses from a given host are more closely related to each other than to virus from corresponding tissues in other individuals, semen-derived viruses may still share genetic characteristics across individuals due to tissue-specific selective pressures that are common across hosts. We employed a machine learning approach (27, 33, 39) to identify a genetic signature associated with seminal tropism. The J48 decision tree inducer (based on the C4.5 algorithm) used in our analysis has been relied on extensively as an alternative to traditional discriminant analysis, due largely to its capacity to detect and exploit interactions between feature variables in training data sets (27). We first applied this algorithm to classify *env* sequences from all individuals based on tissue of origin. The training data for this experiment drew samples from the entire available sequence set, consisting of 376 plasma sequences and 283 from semen. Our results (Table 1) indicate that in this first classification only 65% of sequences were classified correctly, and seminal tropism was predicted with a true positive rate of 0.48.

It is likely that a lack of apparent viral compartmentalization is due to persistent trafficking between blood and semen. To determine if these low scores were due to the presence of viral sequence data classified as semen-derived that actually represented a recent introgression of plasma virus into the male genital tract, we

purged the training set of all data associated with noncompartmentalized hosts. We retained the sequence data from individual G at compartmentalized time points. This pruned set consisted of 143 plasma sequences and 122 from semen. Our results for this second trial (Table 1) demonstrate a strong genetic signature associated with semen-derived sequences; 82% of sequences were classified accurately based on tissue of origin, and seminal tropism was predicted with a precision of 0.842 and a TPR of 0.818 (well over 90% of sequences were classified accurately when the entire training set was used for testing). It is important to point out that the cross-validation procedure used to evaluate this model is quite conservative; the classifier is always tested on a subset of the sequence data that it did not encounter during the training process. The signature underlying seminal tropism comprises a total of four positions within the C2-V3 region (numbered from the start of HXB2 gp160): 270, 291, 387, and 464 (Fig. 4; see Fig. S6 in the supplemental material). The bulk of the signature focuses on either the amino acid character at position 464 or its immediate linkage with a single other Env residue.

Identification of positively selected sites.

We used a maximum likelihood approach to identify sites within *env* that were under positive selection in both compartments, focusing on individuals with compartmentalized virus. We sought to determine if the overall extent of selection and the array of sites under selection varied between compartments, consistent with our finding of a male genital tract-specific genetic signature. Sequence data from hosts A to G (including only data from the initial compartmentalized points associated with

subject G) were first individually evaluated on a per compartment basis by using a model selection procedure to account for any existing mutational biases. Next the FEL approach (22) was employed to test for selective pressure at a given site. All sites in both compartments that appeared to be under positive selection were cataloged and compared. The number of positively selected sites was universally lower in semen-derived viral populations ($P < 0.01$ by a paired Wilcoxon test) (Table 2). Four out of seven individuals failed to exhibit positive selection at any sites within the C2-V3 region in their seminal virus. Additionally, in most cases the sites determined to be under positive selection varied between compartments. Only 3 out of 10 sites identified in seminal populations were also positively selected in corresponding plasma populations (Table 2).

N-linked glycosylation in plasma- and semen-derived viral populations.

To investigate variation in selection pressure from the neutralizing antibody response, we examined glycoslyation patterns across the viral envelope (48). If the antibody response is attenuated in the male genital tract, we might expect fewer glycosylation sites within semen-derived viral sequences. If the response is equivalent, but targeting different epitopes, we might expect a reassortment of sites though the overall number may remain constant. Our results demonstrate that the extent of glycosylation differs significantly in six out of seven patients characterized by compartmentalized virus, but the direction of the discrepancy is inconsistent ($P < 0.05$ for six intrapatient comparisons; Mann-Whitney test). Individuals A, E, and G have

higher average numbers of sequons in semen-derived sequences, while the opposite condition holds true for individuals C, D, and F (Fig. 5).

The distribution of glycosylation sites over time was tracked in the two individuals with compartmentalized virus and with associated longitudinal sequence data. Semen-derived sequences from individual F gradually acquired a single additional sequon at a site (position 411) that was never glycosylated in plasma populations. Plasma sequences demonstrated a continual reassortment of sites with negligible fluctuation in overall number, in accordance with the notion of an evolving "glycan shield" (48). Individual G exhibited a gradual increase in net number of glycosylation sites in both seminal and plasma-derived *env* sequences, with little reassortment in either compartment.

Prediction of coreceptor usage.

We predicted the chemokine receptor preference for all sequences derived from patients with compartmentalized virus to determine if seminal tropism was correlated with altered coreceptor usage. Our results suggest that a trend towards reduced CXCR4 usage in the male genital tract exists, although it is not statistically significant due to the rarity of the CXCR4 phenotype across individuals and compartments; only three out of seven hosts harbored variants predicted to use the CXCR4 receptor (Fig. 6).

Evaluation of codon usage bias.

It has previously been reported that the differential availability of nucleotide precursor pools in target cells may influence HIV-1 codon usage patterns.

Additionally, the cytidine deaminase APOBEC3G, found in lymphocytes, induces G to A mutations that skew codon usage towards A-rich triplets (51). If viral target cells within the male genital tract differ from peripheral tissues in precursor frequencies and APOBEC3G expression levels, an altered codon usage bias may evolve in seminal virus. Our analysis revealed no significant differences in codon usage between blood and semen virus (data not shown).

DISCUSSION

In these investigations we applied a battery of computational techniques to paired semen- and blood-derived HIV-1 *env* sequences, which confirmed previous reports that HIV within the genital tract is different from that within the bloodstream (10, 20). This study extends those observations with findings important to the understanding of how HIV adapts to the male genital tract. First, the male genital tract can function as a viral compartment, but the extent of compartmentalization differs between individuals and within individuals over time. Second, there are discordant selective pressures operating in the male genital tract and blood. Third, semen-derived viruses share a genetic signature across individuals due to tissue-specific selective pressures that are common across hosts.

Viral compartments are characterized by a restriction of gene flow between cells or tissues, usually identified by phylogenetic analysis (29). In this study, viral compartmentalization between blood and the male genital tract was identified in 6 out of 12 individuals, and another individual demonstrated compartmentalization of virus only at the earliest sampling times. Viral migration between blood plasma and the

male genital tract was minimal and infrequent in these individuals, which reinforces the concept that a significant fraction of virus shed in semen is produced locally in the male genital tract. Furthermore, there was a lower genetic diversity and rate of molecular evolution in seminal sequences, probably reflecting a lower effective population size within the male genital tract. This lower effective population size may contribute to the genetic bottleneck associated with HIV-1 transmission. We cannot exclude the possibility, however, that sampling issues contributed to this phenomenon; the efficiency of RNA extraction and reverse transcription-PCR may be lower in semen than plasma, increasing the potential for resampling.

The degree of compartmentalization varied among individuals and also within individuals over time. This may explain the observations of intermittent viral shedding in the semen of HIV-infected men (15, 47) and the increased viral shedding when the urethra is inflamed by concomitant bacterial or viral infection (40). Local inflammation is a likely explanation for increased trafficking of HIV from the circulation to the genital compartment. Future studies examining the relationship between sexually transmitted infections and seminal viral loads may provide valuable insight into viral adaptation and dynamics within the male genital tract. This understanding could be crucial in the development of methods to interrupt HIV transmission such as vaccines, microbicides, and antiretroviral suppression.

Seeding of genital tissues occurs very early in infection before the development of any anti-HIV immune response (13). Once the host mounts an anti-HIV immune response, it most likely varies in strength and nature between

compartments (29). We investigated the degree of selection on the virus within the two compartments and found that there was greater positive selection on virus in the blood than virus in the male genital tract. In six out of the seven individuals with compartmentalized virus, there were highly significant differences in *env* glycosylation but not in a consistent direction. While this reinforces the theory that virus is produced locally in the male genital tract and responds to local humoral immunity, it does not explain the recent reports that HIV transmission through heterosexual exposure involves viruses with fewer envelope glycans (11).

Since cellular tropism may also play a role in viral compartmentalization and adaptation to the male genital tract, we investigated the coreceptor usage of viruses in blood and semen. It is provocative that in all individuals who harbored CXCR4-using viruses, these viruses were underrepresented in the genital tract. Selection favoring R5 variants in the male genital tract may explain the observation that newly infected individuals are disproportionately infected with CCR5-using viruses (54, 55).

Although HIV within the male genital tract is often different from that within the bloodstream (10, 17, 32), the initially infecting virus (founding virus) and the individual's immune responses determine viral genetics more than tissue of origin (29). Therefore, it has been difficult to determine if semen-derived virus shares common genetic characteristics among individuals (10). Using machine learning techniques, we have found that semen-derived HIV-1 has a strong genetic signature among individuals with compartmentalized virus. The signature comprises several positions across C2-V3; however, the residue at position 464 appears to be the most

critical in determining viral tropism to the male genital tract. This particular position, to the best of our knowledge, has not previously been reported within the context of tissue tropism or viral compartmentalization. Nevertheless, this classification trial presents convincing evidence that the male genital tract environment selects for similar, predictable genetic changes in *env* across individuals.

The male genital tract has been characterized as a reservoir (43, 52), a compartment (10), and a drug sanctuary (45). All have significant implications for preventing the transmission of HIV by using various theoretical methods such as microbicides, vaccines, or antiretroviral therapy (2, 9, 10). Our investigations uniquely detail the viral compartmentalization dynamics and differing selection pressures between the blood and male genital tract and document a specific genetic signature of virus compartmentalized in the male genital tract. Taken together, these data offer important insights into the adaptation of HIV to the male genital tract, which may be valuable in the rational design of an effective vaccine.

ACKNOWLEDGMENTS

The text of this chapter, in full, is a reprint of the material as it appears in the Journal of Virology: Pillai, S.K., B. Good, S. Kosakovsky Pond, J.K. Wong, M.C. Strain, D.D. Richman, and D.M. Smith, "Semen-Specific Genetic Characteristics of Human Immunodeficiency Virus Type 1 *env*", vol. 79, pp. 1734-1742, February 2005. I was the primary author, and the co-authors listed in this publication supervised and/or contributed to the research which forms the basis for this chapter.

We are grateful to Susan Little and Simon Frost for their insightful comments. We also thank Brian Gaschen for assistance with assimilating the sequence data, John Day for his technical expertise, and Darica Smith and Sharon Wilcox for helping with the preparation of the manuscript.

This work was supported by grants 5K23AI055276, AI27670, AI38858, AI43638, AI43752, AI36214 (UCSD Center for AIDS Research), AI29164, and AI047745 from the National Institutes of Health. Additional support was provided by the Research Center for AIDS and HIV Infection of the San Diego Veterans Affairs Healthcare System.

REFERENCES

1. Altfeld, M., E. S. Rosenberg, R. Shankarappa, J. S. Mukherjee, F. M. Hecht, R. L. Eldridge, M. M. Addo, S. H. Poon, M. N. Phillips, G. K. Robbins, P. E. Sax, S. Boswell, J. O. Kahn, C. Brander, P. J. Goulder, J. A. Levy, J. I. Mullins, and B. D. Walker. 2001. Cellular immune responses and viral diversity in individuals treated during acute and early HIV-1 infection. J Exp. Med 193:169-180.

2. Auvert, B., S. Males, A. Puren, A. Taljaard, M. Carael, and B. Williams. 2004. Can highly active antiretroviral therapy reduce the spread of HIV?: A study in a township of South. J Acquir Immune Defic Syndr. 36:613-621.

3. Chakraborty, H., P. K. Sen, R. W. Helms, P. L. Vernazza, S. A. Fiscus, J. J. Eron, B. K. Patterson, R. W. Coombs, J. N. Krieger, and M. S. Cohen. 2001. Viral burden in genital secretions determines male-to-female sexual transmission of HIV-1: a probabilistic empiric model. AIDS 15:621-627.

4. Chun, T.-W., L. Carruth, D. Finzi, X. Shen, J. A. DiGiuseppe, H. Taylor, M. Hermankova, K. Chadwick, J. Margolick, T. C. Quinn, Y.-H. Kuo, R. Brookmeyer, M. A. Zeiger, P. Barditch-Crovo, and R. F. Siliciano. 1997. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature 387:183-188.

5. Coffin, J. M. 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis and therapy. Science 267:483-489.

6. Coombs, R. W., P. S. Reichelderfer, and A. L. Landay. 2003. Recent observations on HIV type-1 infection in the genital tract of men and women. AIDS 17:455-480.

7. Coombs, R. W., C. E. Speck, J. P. Hughes, W. Lee, R. Sampoleo, S. O. Ross, J. Dragavon, G. Peterson, T. M. Hooton, A. C. Collier, L. Corey, L. Koutsky, and J. N. Krieger. 1998. Association between culturable human immunodeficiency virus type 1 (HIV- 1) in semen and HIV-1 RNA levels in semen and blood: evidence for compartmentalization of HIV-1 between semen and blood. J Infect Dis 177:320-330.

8. Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16:10881-10890.

9. Davis, C. W. and R. W. Doms. 2004. HIV transmission: closing all the doors. J Exp Med 199:1037-1040.

10. Delwart, E. L., J. I. Mullins, P. Gupta, G. H. Learn, Jr., M. Holodniy, D. Katzenstein, B. D. Walker, and M. K. Singh. 1998. Human immunodeficiency virus type 1 populations in blood and semen. J Virol 72:617-623.

11. Derdeyn, C. A., J. M. Decker, F. Bibollet-Ruche, J. L. Mokili, M. Muldoon, S. A. Denham, M. L. Heil, F. Kasolo, R. Musonda, B. H. Hahn, G. M. Shaw, B. T. Korber, S. Allen, and E. Hunter. 2004. Envelope-Constrained Neutralization-Sensitive HIV-1 After Heterosexual Transmission. Science 303:2019-2022.

12. Drew, W. L., R. C. Miner, D. F. Busch, S. E. Follansbee, J. Gullett, S. G. Mehalko, S. M. Gordon, W. F. Owen, Jr., T. R. Matthews, W. C. Buhles, and B. DeArmond. 1991. Prevalence of resistance in patients receiving ganciclovir for serious cytomegalovirus infection. J Infect Dis 163:716-719.

13. Dyer, J. R., B. L. Gilliam, J. J. Eron, Jr., M. S. Cohen, S. A. Fiscus, and P. L. Vernazza. 1997. Shedding of HIV-1 in semen during primary infection. AIDS 11:543-545.

14. Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c.

15. Fiscus, S. A., P. L. Vernazza, B. Gilliam, J. Dyer, J. J. Eron, and M. S. Cohen. 1998. Factors associated with changes in HIV shedding in semen. AIDS Res. Hum. Retroviruses 14, Suppl 1:S27-S31.

16. Gunthard, H. F., D. V. Havlir, S. Fiscus, Z. Q. Zhang, J. Eron, J. Mellors, R. Gulick, S. D. Frost, A. J. Brown, W. Schleif, F. Valentine, L. Jonas, A. Meibohm, C. C. Ignacio, R. Isaacs, R. Gamagami, E. Emini, A. Haase, D. D. Richman, and J. K. Wong. 2001. Residual human immunodeficiency virus (HIV) type 1 RNA and DNA in lymph nodes and HIV RNA in genital secretions and in cerebrospinal fluid after suppression of viremia for 2 years. J Infect Dis 183:1318-1327.

17. Gupta, P., C. Leroux, B. K. Patterson, L. Kingsley, C. Rinaldo, M. Ding, Y. Chen, K. Kulka, W. Buchanan, B. McKeon, and R. Montelaro. 2000. Human immunodeficiency virus type 1 shedding pattern in semen correlates with the compartmentalization of viral quasispecies between blood and semen. J Infect Dis 182:79-87.

18. Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-589.

19. Jensen, M. A. and A. B. van't Wout. 2003. Predicting HIV-1 coreceptor usage with sequence analysis. AIDS Rev. 5:104-112.

20. Kemal, K. S., B. Foley, H. Burger, K. Anastos, H. Minkoff, C. Kitchen, S. M. Philpott, W. Gao, E. Robison, S. Holman, C. Dehner, S. Beck, W. A. Meyer, III, A. Landay, A. Kovacs, J. Bremer, and B. Weiser. 2003. HIV-1 in genital tract and plasma of women: compartmentalization of viral sequences, coreceptor usage, and glycosylation. Proc Natl Acad Sci U. S. A 100:12972-12977.

21. Kiessling, A. K., G. Zheng, and R. C. Eyre. 1992. Semen producing organs are an isolated reservoir of HIV which may play a significant role in the development of drug resistant strains. J Hum Virol 2:193.

22. Kosakovsky Pond, S. and S. D. W. Frost. 2004. The fast and the punctilious: an integrative approach to detecting amino-acid sites undergoing adaptive or purifying evolution. Submitted.

23. Krieger, J. N., R. W. Coombs, A. C. Collier, D. D. Ho, S. O. Ross, J. E. Zeh, and L. Corey. 1995. Intermittent shedding of human immunodeficiency virus in semen: implications for sexual transmission. J Urol. 154:1035-1040.

24. Krieger, J. N., A. Nirapathpongporn, M. Chaiyaporn, G. Peterson, I. Nikolaeva, R. Akridge, S. O. Ross, and R. W. Coombs. 1998. Vasectomy and human immunodeficiency virus type 1 in semen. J Urol. 159:820-825.

25. Marshall, R. D. 1974. The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins. Biochem. Soc. Symp.17-26.

26. McInerney, J. O. 1998. GCUA: general codon usage analysis. Bioinformatics 14:372-373.

27. Mjolsness, E. and D. DeCoste. 2001. Machine learning for science: state of the art and future prospects. Science 293:2051-2055.

28. Muse, S. V. and B. S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11:715-724.

29. Nickle, D. C., D. Shriner, J. E. Mittler, L. M. Frenkel, and J. I. Mullins. 2003. Importance and detection of virus reservoirs and compartments of HIV infection. Current Opinion in Microbiology 6:410-416.

30. Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek. 2004. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput.Appl.Biosci. 10:41-48

31. Page, R. D. M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comput.Appl.Biosci. 12:357-358

32. Paranjpe, S., J. Craigo, B. Patterson, M. Ding, P. Barroso, L. Harrison, R. Montelaro, and P. Gupta. 2002. Subcompartmentalization of HIV-1 quasispecies between seminal cells and seminal plasma indicates their origin in distinct genital tissues. AIDS Res. Hum. Retroviruses 18:1271-1280.

33. Pillai, S., B. Good, D. Richman, and J. Corbeil. 2003. A new perspective on V3 phenotype prediction. AIDS Res. Hum. Retroviruses 19:145-149.

34. Piot, P., M. Bartos, P. D. Ghys, N. Walker, and B. Schwartlander. 2001. The global impact of HIV/AIDS. Nature 410:968-973.

35. Quinlan, J. R. 1993. C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco.

36. Quinn, T. C., M. J. Wawer, N. Sewankambo, D. Serwadda, C. Li, F. Wabwire-Mangen, M. O. Meehan, T. Lutalo, and R. H. Gray. 2000. Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group. N Engl J Med 342:921-929.

37. Rambaut A. 2002. Se-Al sequence alignment editor v2.0 (Software). Oxford: Department of Zoology, University of Oxford.

38. Rambaut, A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics. 16:395-399.

39. Resch, W., N. Hoffman, and R. Swanstrom. 2001. Improved success phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. Virology 288:51-62.

40. Sadiq, S. T., S. Taylor, S. Kaye, J. Bennett, R. Johnstone, P. Byrne, A. J. Copas, S. M. Drake, D. Pillay, and I. Weller. 2002. The effects of antiretroviral therapy on HIV-1 RNA loads in seminal plasma in HIV-positive patients with and without urethritis. AIDS 16:219-225.

41. Singh, A., G. Besson, A. Mobasher, and R. G. Collman. 1999. Patterns of chemokine receptor fusion cofactor utilization by human immunodeficiency virus type 1 variants from the lungs and blood. J Virol 73:6680-6690.

42. Slatkin, M. and W. P. Maddison. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123:603-613.

43. Smith, D. M., J. D. Kingery, J. K. Wong, C. C. Ignacio, D. D. Richman, and S. J. Little. 2004. The prostate as a reservoir for HIV-1. AIDS 18:6-8.

44. Strain, M. C., H. F. Günthard, D. V. Havlir, C. C. Ignacio, D. M. Smith, A. J. Leigh Brown, T. R. Macaranas, R. Y. Lam, O. A. Daly, M. Fischer, M. Opravil, H. Levine, L. Bacheler, C. A. Spina, D. D. Richman, and J. K. Wong. 2003. Heterogeneous clearance rates of long-lived lymphocytes infected with HIV: intrinsic stability predicts lifelong persistence. Proc Natl Acad Sci U. S. A 100:4819-4824.

45. Taylor, S., R. P. van Heeswijk, R. M. Hoetelmans, J. Workman, S. M. Drake, D. J. White, and D. Pillay. 2000. Concentrations of nevirapine, lamivudine and stavudine in semen of HIV- 1-infected men. AIDS 14:1979-1984.

46. UNAIDS/WHO. 2004. AIDS epidemic update : December 2003. UNAIDS/ World Health Organization, Geneva, Switzerland.

47. Vernazza, P. L., B. L. Gilliam, J. Dyer, S. A. Fiscus, J. J. Eron, A. C. Frank, and M. S. Cohen. 1997. Quantification of HIV in semen: correlation with antiviral treatment and immune status. AIDS 11:987-993.

48. Wei, X., J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A.

Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw. 2003. Antibody neutralization and escape by HIV-1. Nature 422:307-312.

49. Witten, I. H. and E. Frank. 2000. Data mining practical machine learning tools and techniques with java implementations. Morgan Kaufmann, San Francisco.

50. Wong, J. K., C. C. Ignacio, F. Torriani, D. Havlir, N. J. S. Fitch, and D. D. Richman. 1997. In vivo compartmentalization of HIV: evidence from the examination of pol sequences from autopsy tissues. J Virol 70:2059-2071.

51. Yu, Q., R. Konig, S. Pillai, K. Chiles, M. Kearney, S. Palmer, D. Richman, J. M. Coffin, and N. R. Landau. 2004. Single-strand specificity of APOBEC3G accounts for minus-strand deamination of the HIV genome. Nat Struct Mol Biol. 11:435-442.

52. Zhang, H., G. Dornadula, M. Beumont, L. Livornese, B. Van Uitert, K. Henning, and R. J. Pomerantz. 1998. Human immunodeficiency virus type 1 in the semen of men receiving highly active antiretroviral therapy. The New England Journal of Medicine 339:1803-1809.

53. Zhang, L., L. Rowe, T. He, C. Chung, J. Yu, W. Yu, A. Talal, M. Markowitz, and D. D. Ho. 2002. Compartmentalization of surface envelope glycoprotein of human immunodeficiency virus type 1 during acute and chronic infection. J Virol 76:9465-9473.

54. Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. Leigh-Brown, and P. Simmonds. 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. J Virology 67:3345-3356.

55. Zhu, T., H. Mo, N. Wang, D. S. Nam, Y. Cao, R. A. Koup, and D. D. Ho. 1993. Genotypic and phenotypic characterization of HIV-1 in patients with primary infection. Science 261:1179-1184.

FIGURES AND TABLES

Table 1: Classification of *env* C2-V3 sequences based on tissue of origin **(**cross-validation statistics**)**

	All Sequences	Compartmentalized Sequences
Correctly classified instances	434 (65.9%)	217 (81.9%)
Incorrectly classified instances	225 (34.1%)	48 (18.1%)
Total number of instances	659	265
Kappa statistic	.281	.636
True positive rate - blood	79.5%	82.0%
True positive rate - semen	47.7%	81.8%

Table 2: Sites under positive selection in compartmentalized individuals. Fewer sites were under selective pressure in seminal populations based on the Approximate Likelihood Ratio at a Site (ARS) method ($p<0.01$, paired Wilcoxon).

Individual	A	B	C	D	E	F	G
Blood	397, 467	461	-	354, 438	402	335, 336, 337, 340, 343, 354, 446	283, 335, 336, 346, 354, 263, 364, 405, 466, 467
Semen	364	-	-	-	-	354, 446	354, 401, 402, 455, 460, 463, 471

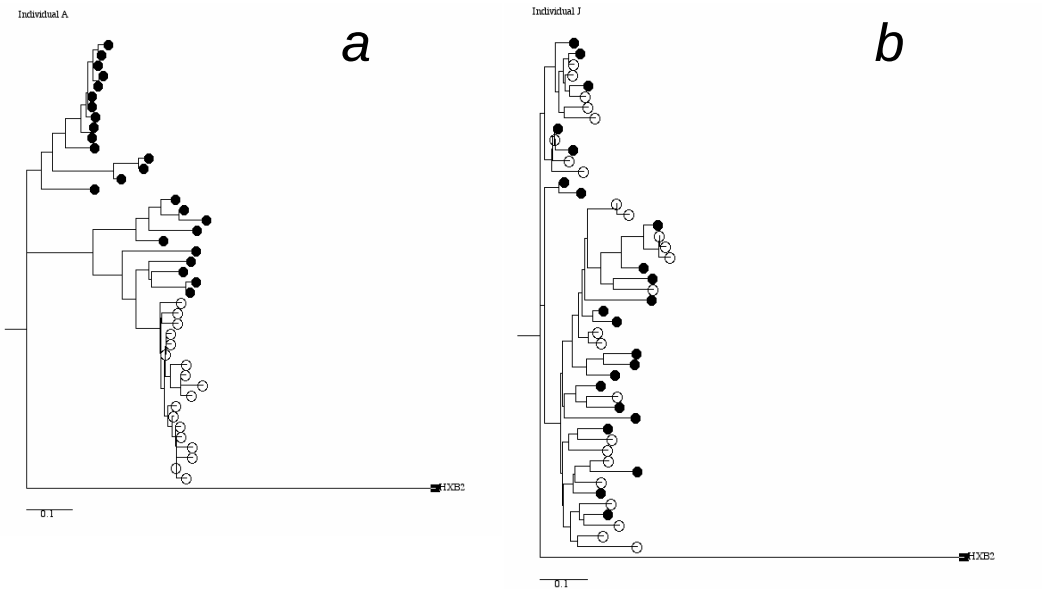


Figure 1: Examples of compartmentalized and noncompartmentalized viral populations. Maximum likelihood phylogenies of C2-V3 *env* sequences. (a) Individual A, compartmentalized virus. (b) Individual J, noncompartmentalized virus. Open circles represent semen sequences, and closed circles indicate plasma-derived sequences. Black squares represent the HXB2 outgroup. Scale bar equals 10% genetic distance.

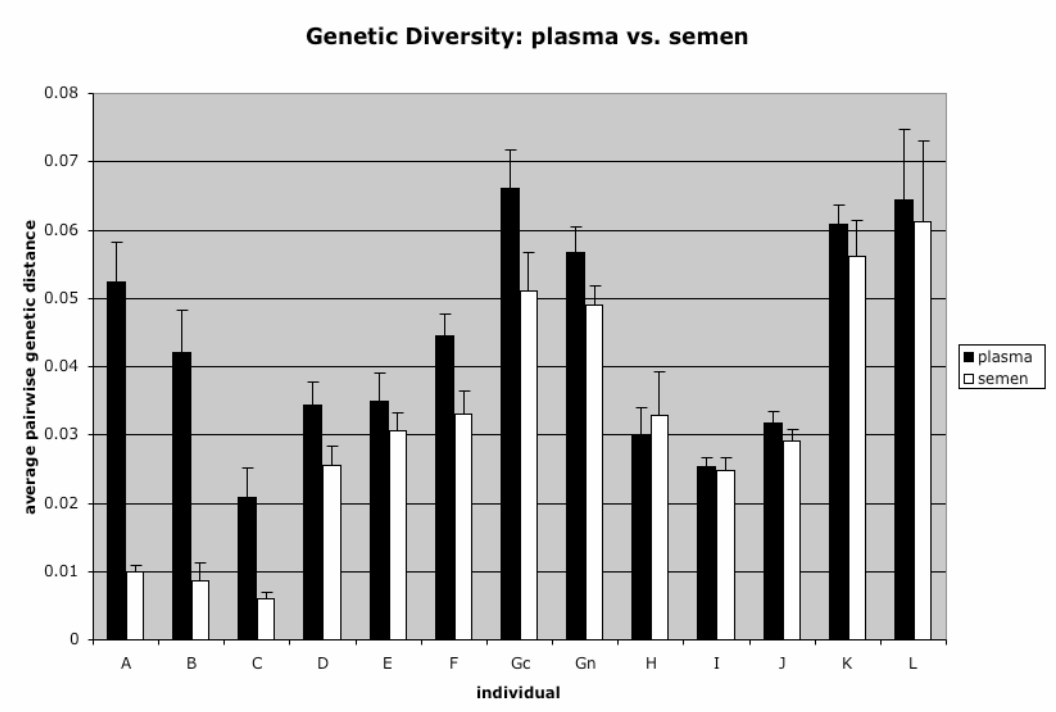


Figure 2: Genetic diversity in semen-derived and blood-derived viral populations. Genetic diversity was significantly lower in semen-derived viral populations within individuals characterized by compartmentalized virus (individuals A to Gc; $P < 0.01$ by a paired Wilcoxon test). No significant difference in viral diversity between blood and semen viral populations was observed in individuals with noncompartmentalized virus (individuals Gn to L). Vertical bars represent standard error.

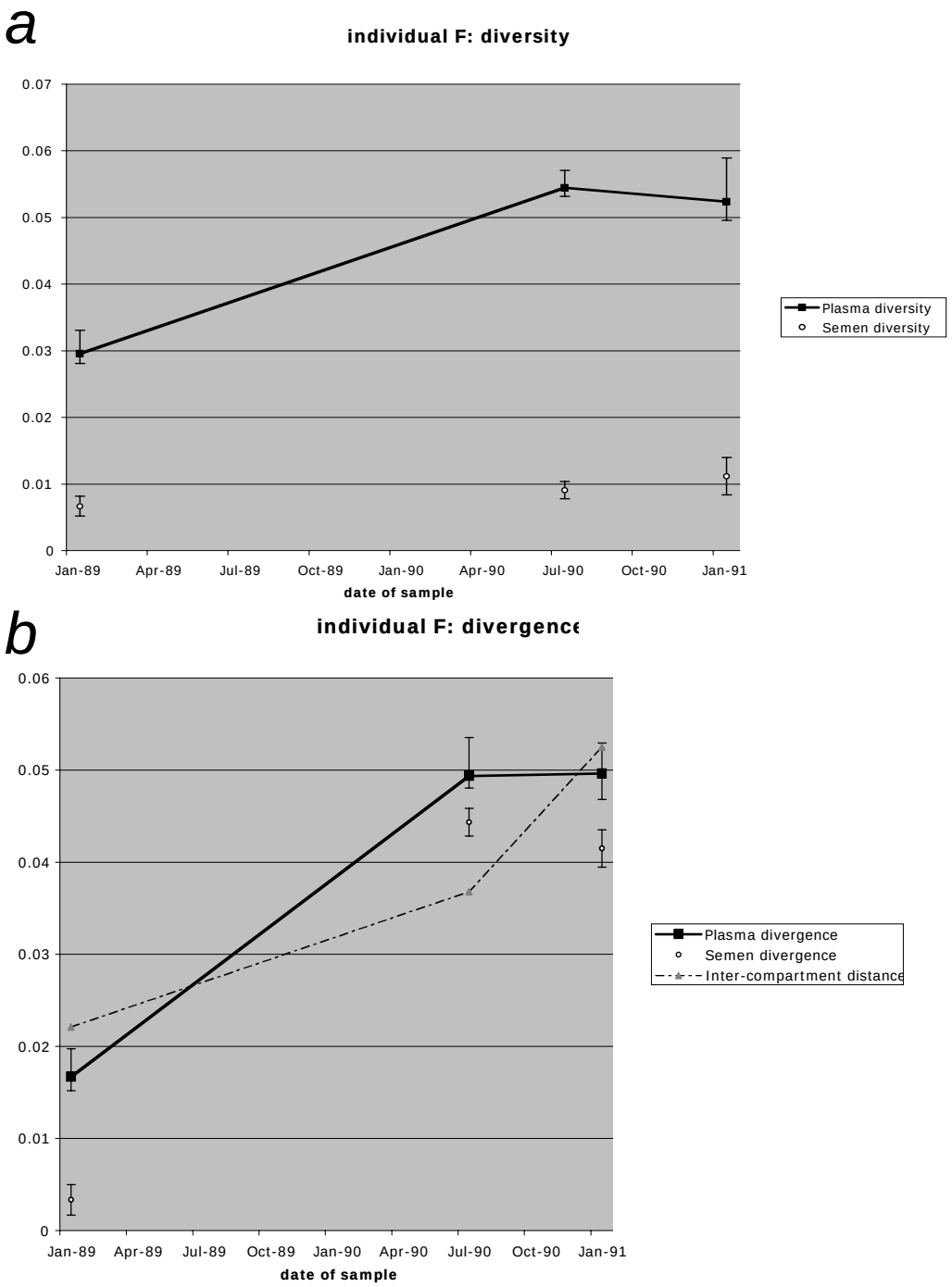


Figure 3: Longitudinal viral diversity and divergence in an individual with compartmentalized virus, individual F. (a) Genetic diversity measured over a 2-year period. (b) Divergence and intercompartment genetic distance measured over a 2-year period. Vertical bars represent standard error

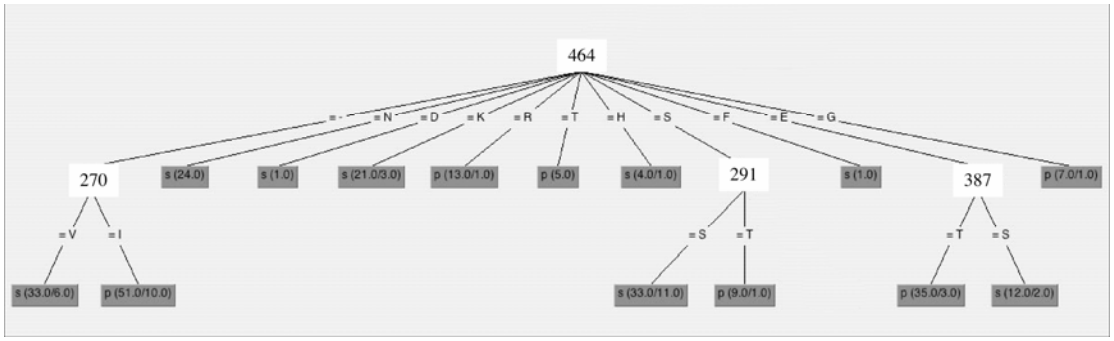


Figure 4. Genetic signature associated with seminal sequences from compartmentalized individuals. Decision tree classifying C2-V3 *env* sequences based on tissue of origin with 82% accuracy. p, plasma classification; s, semen. The values in parentheses are the number of instances/number of incorrect classifications. Residue numbers are based on HXB2 gp160 positions.

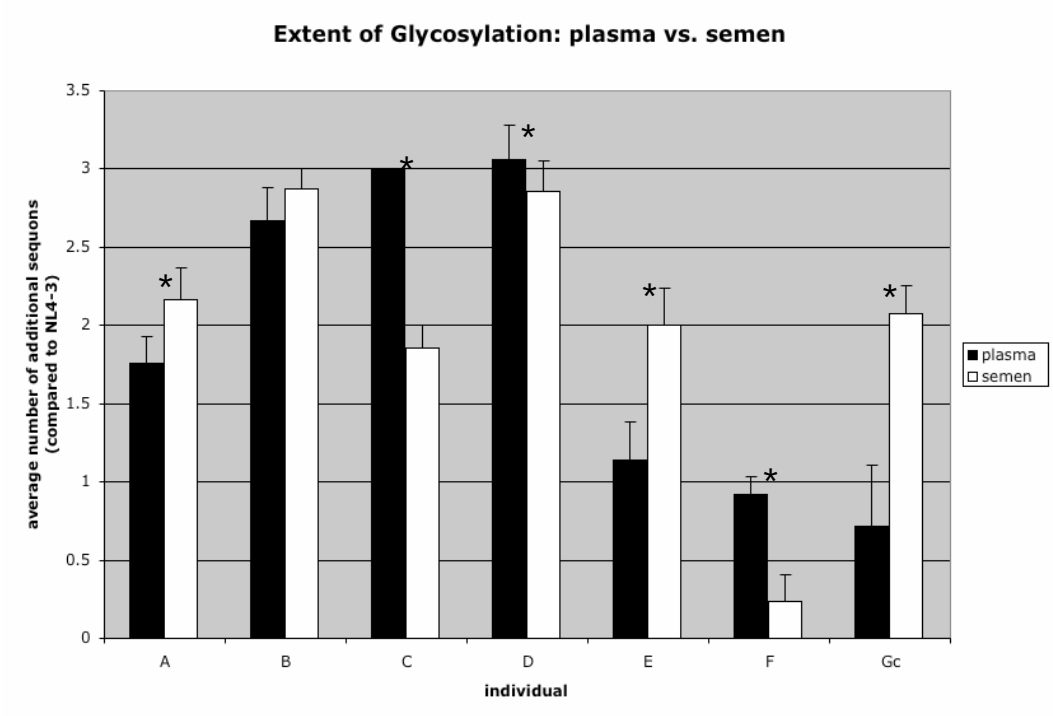


Figure 5. Extent of viral glycosylation in plasma and semen of individuals with compartmentalized virus. N-linked glycosylation sites were predicted based on an NXS or NXT sequence motif. Asterisks indicate significant comparisons ($P < 0.05$ by a Mann-Whitney test). Vertical bars represent standard error.

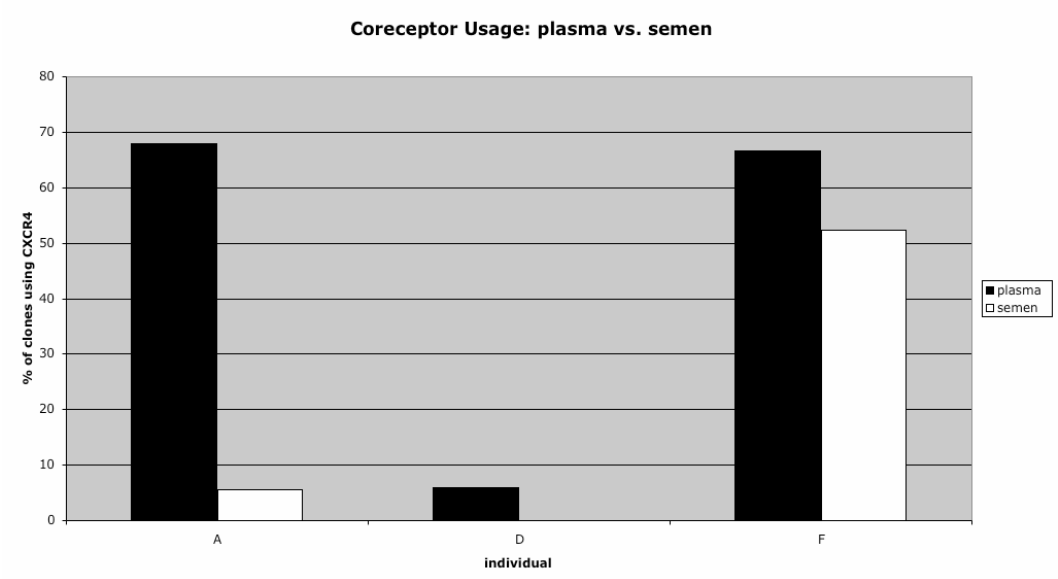
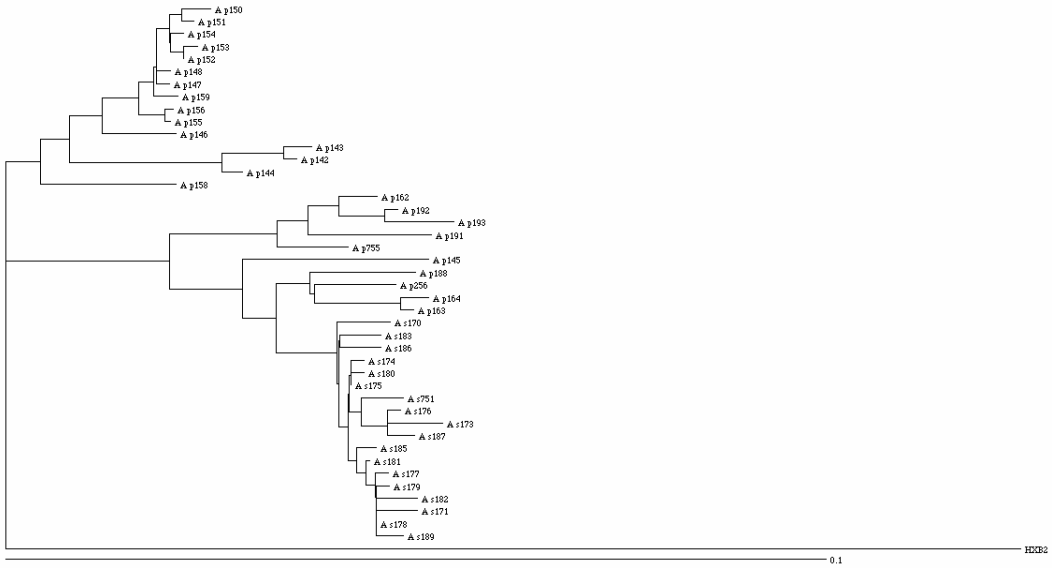


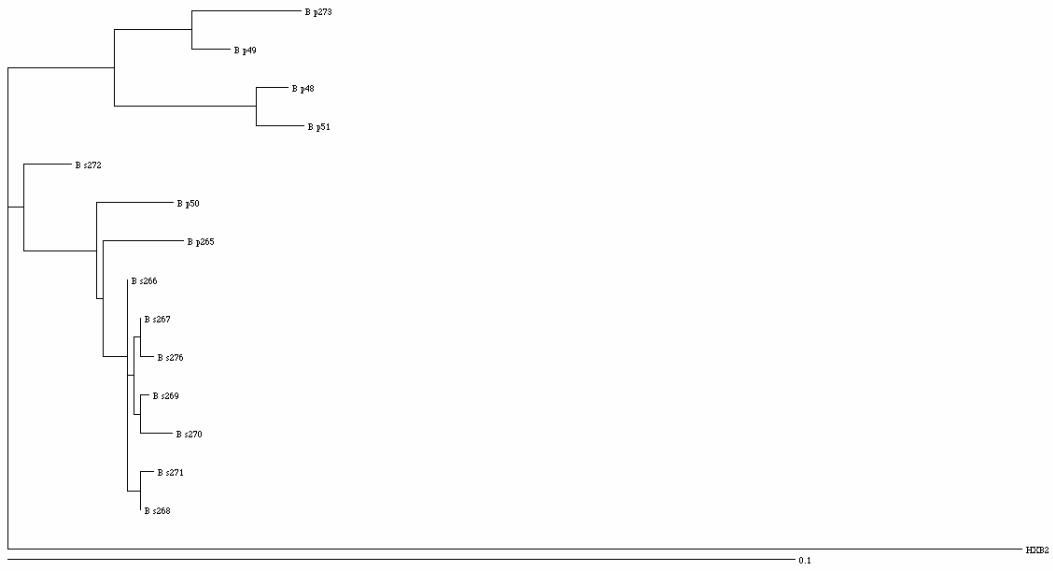
Figure 6. Coreceptor phenotype in plasma and semen of individuals with compartmentalized virus showing evidence of CXCR4 usage in either tissue. Phenotypes were predicted based on V3 genotype by using a machine learning approach

individual A



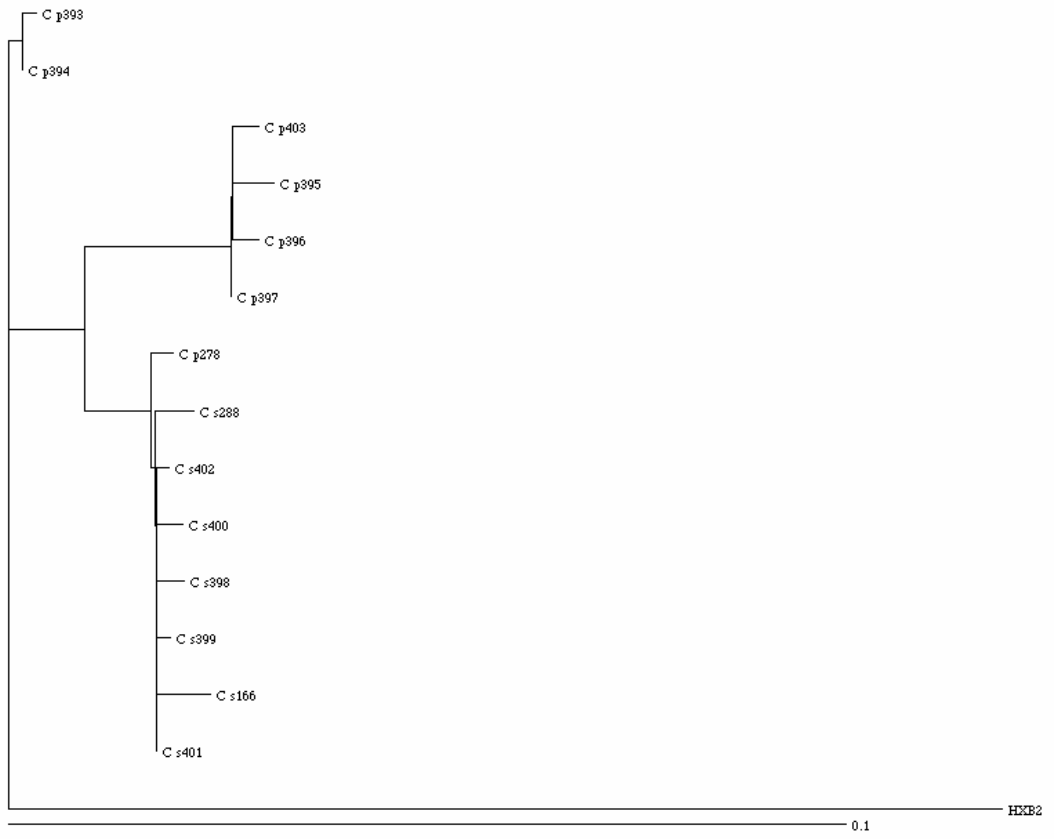
Supplementary Figure 1. Maximum likelihood phylogenetic trees of each compartmentalized individual's sequence data (individulas A-F). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

individual B



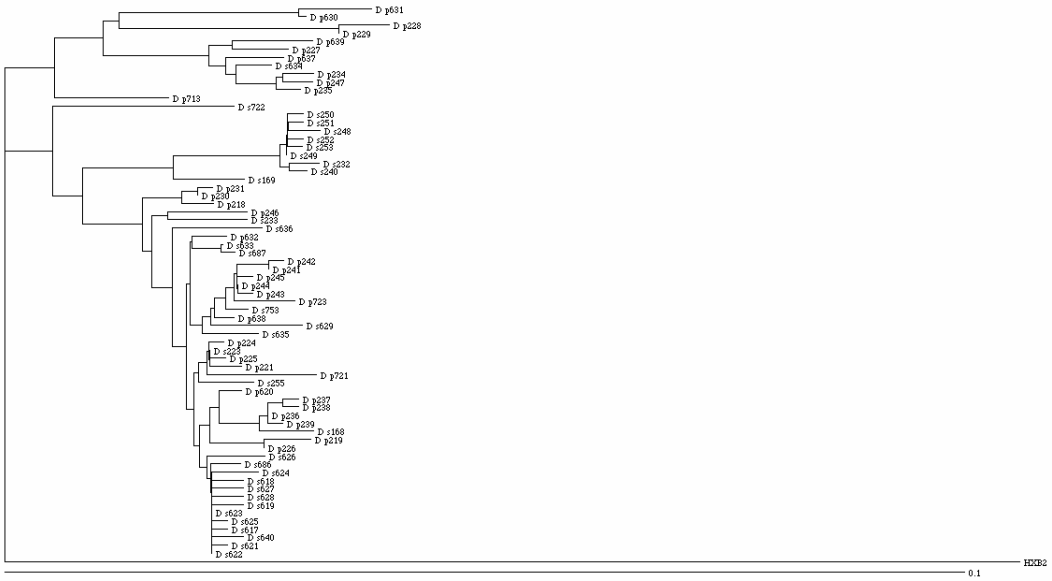
Supplementary Figure 1 (cont'd). Maximum likelihood phylogenetic trees of each compartmentalized individual's sequence data (individulas A-F). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

individual C



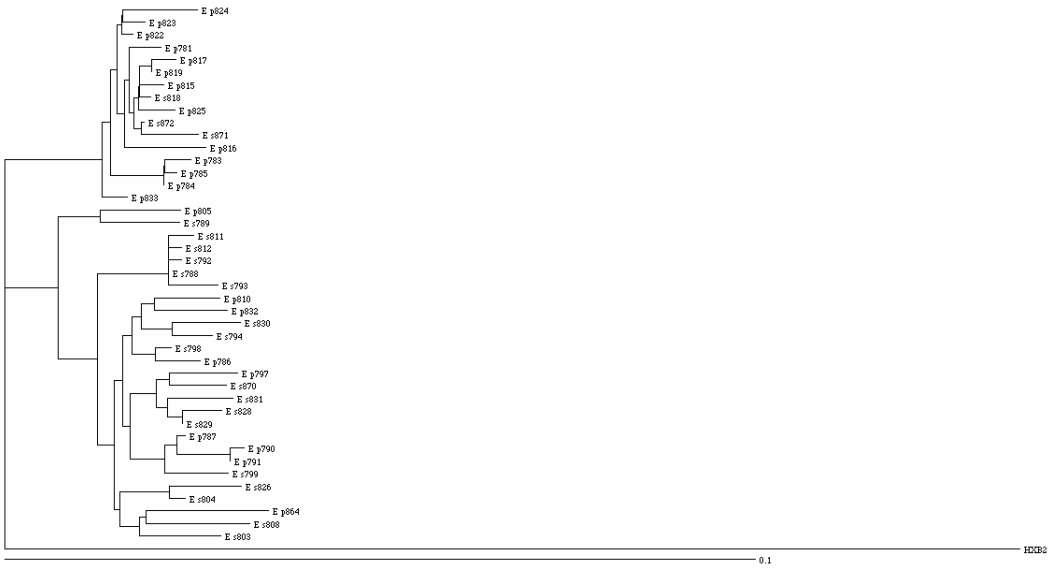
Supplementary Figure 1 (cont'd). Maximum likelihood phylogenetic trees of each compartmentalized individual's sequence data (individulas A-F). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

individual D



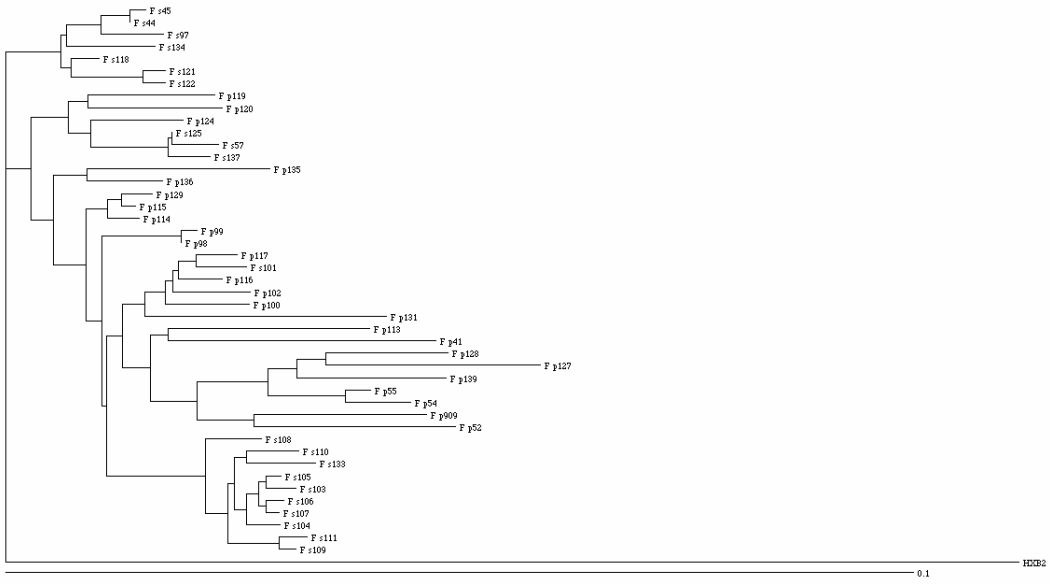
Supplementary Figure 1 (cont'd). Maximum likelihood phylogenetic trees of each compartmentalized individual's sequence data (individulas A-F). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

individual E



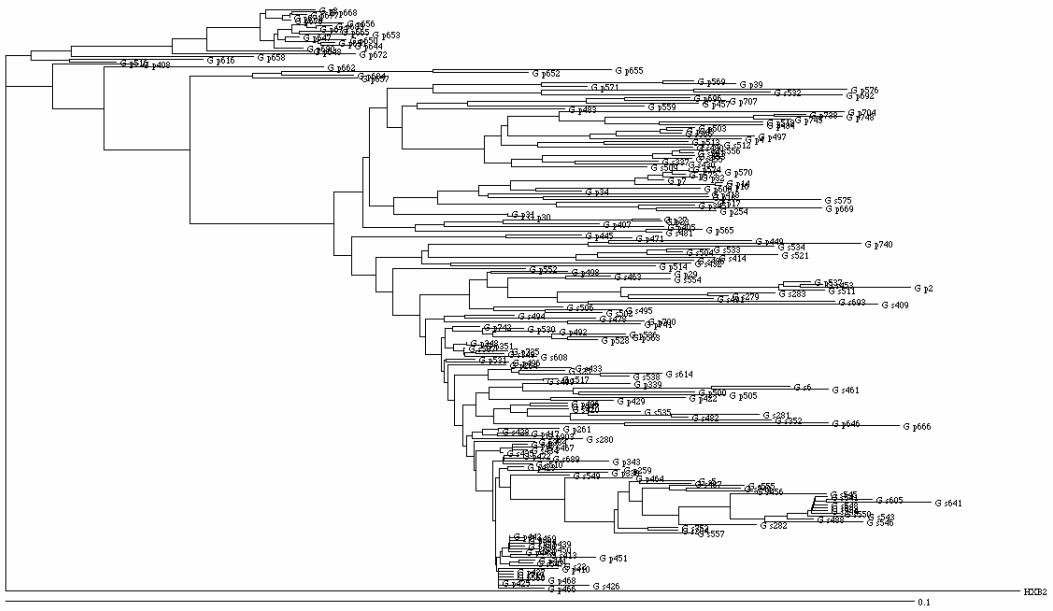
Supplementary Figure 1 (cont'd). Maximum likelihood phylogenetic trees of each compartmentalized individual's sequence data (individulas A-F). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

individual F



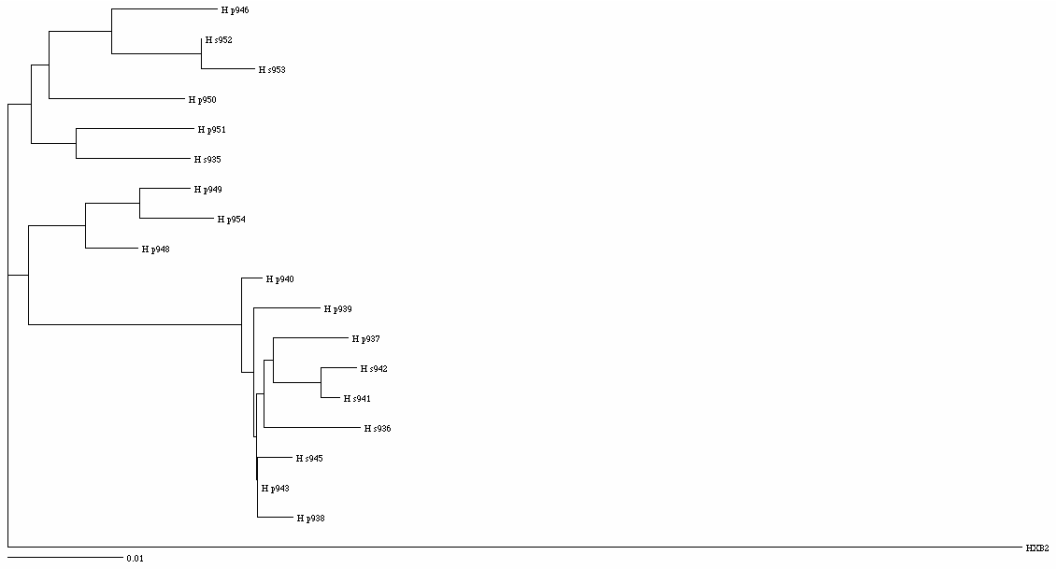
Supplementary Figure 1 (cont'd). Maximum likelihood phylogenetic trees of each compartmentalized individual's sequence data (individulas A-F). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

individual G



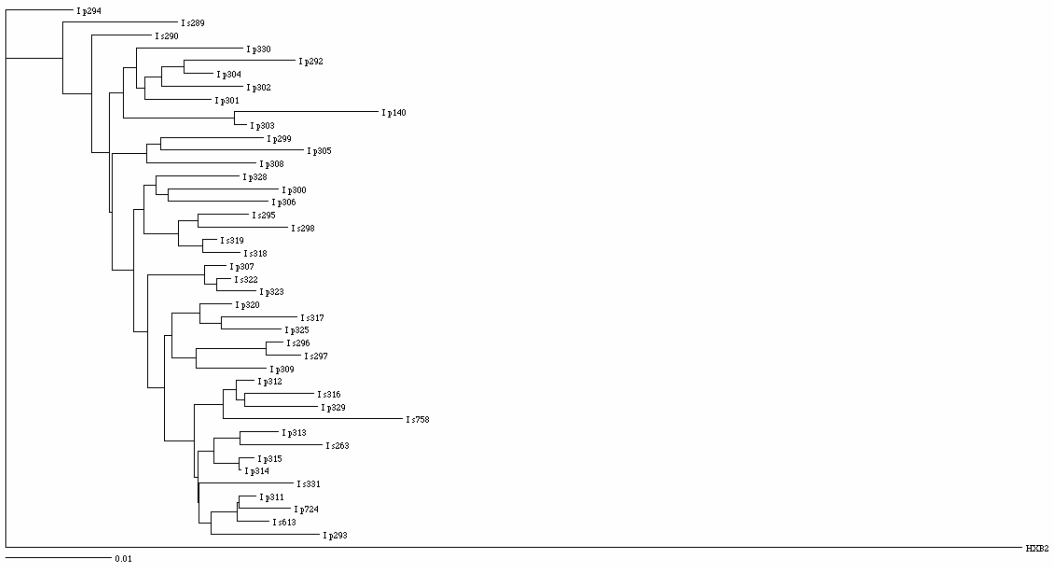
Supplementary Figure 2. Maximum likelihood phylogenetic tree of all sequence data associated with individual G, who demonstrates inconsistent viral compartmentalization between blood and semen. HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

individual H



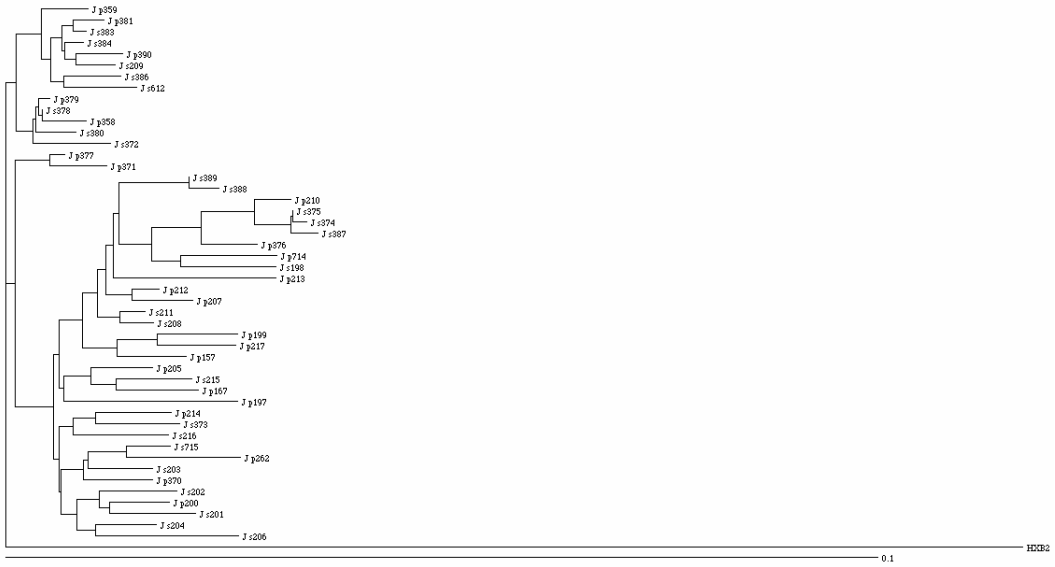
Supplementary Figure 3. Maximum likelihood phylogenetic trees of each non-compartmentalized individual's sequence data (individulas H-L). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

individual I



Supplementary Figure 3 (cont'd). Maximum likelihood phylogenetic trees of each non-compartmentalized individual's sequence data (individulas H-L). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

individual J

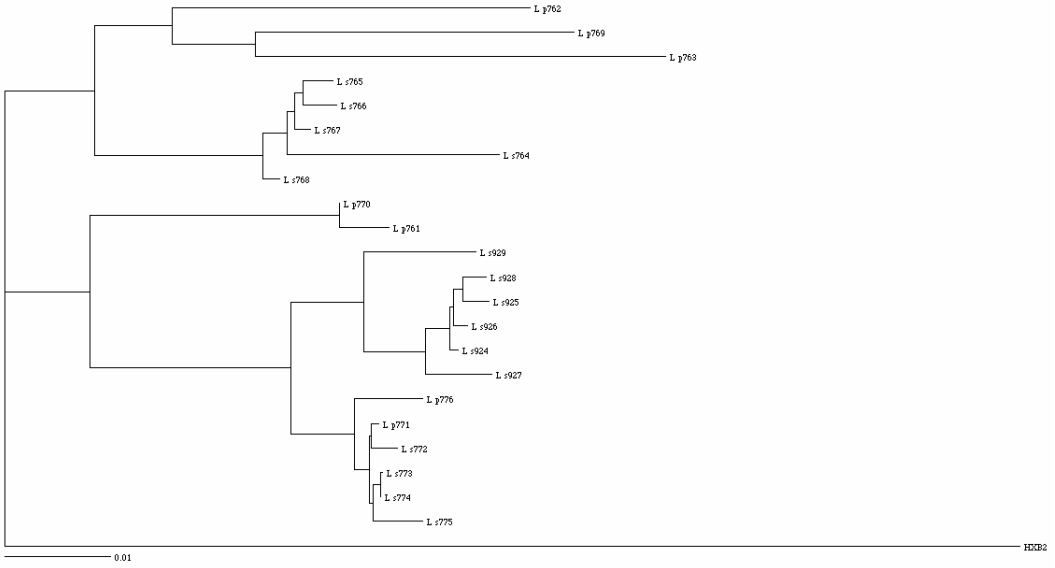


Supplementary Figure 3 (cont'd). Maximum likelihood phylogenetic trees of each non-compartmentalized individual's sequence data (individulas H-L). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

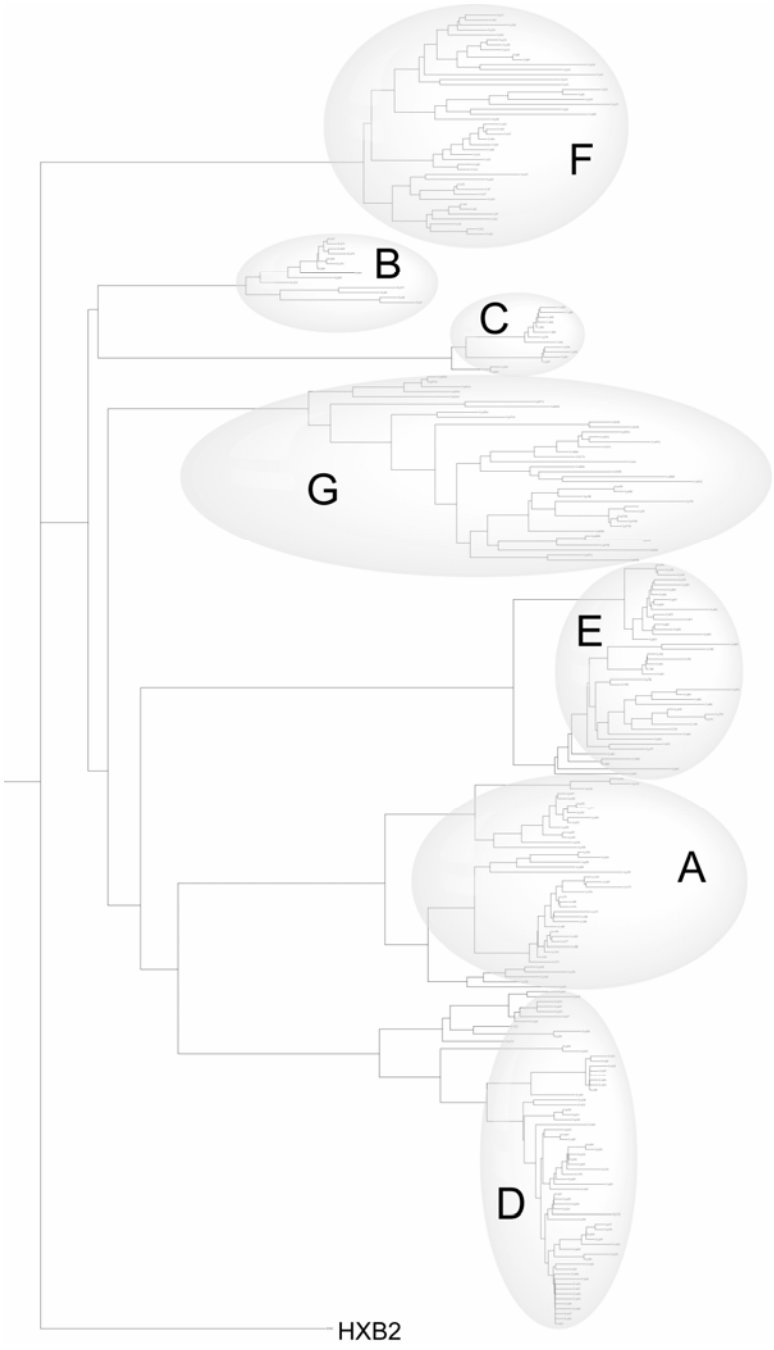
individual K

Supplementary Figure 3 (cont'd). Maximum likelihood phylogenetic trees of each non-compartmentalized individual's sequence data (individulas H-L). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.

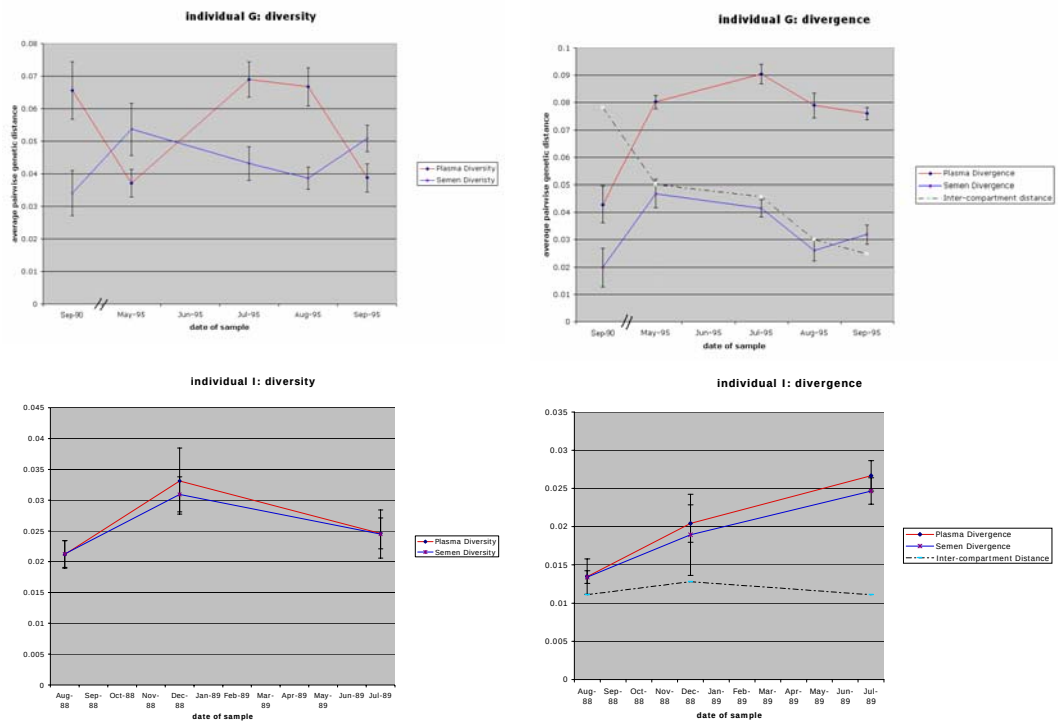
individual L



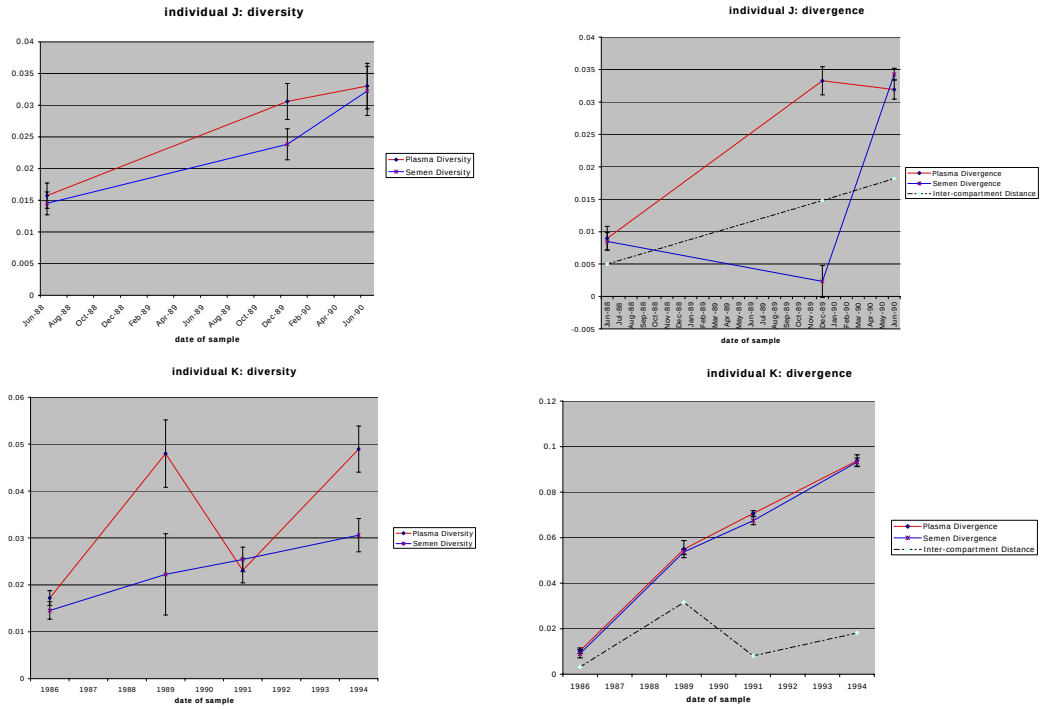
Supplementary Figure 3 (cont'd). Maximum likelihood phylogenetic trees of each non-compartmentalized individual's sequence data (individulas H-L). HIV-1 HXB2 included as outgroup sequence. Scale bar represents 10% genetic distance.



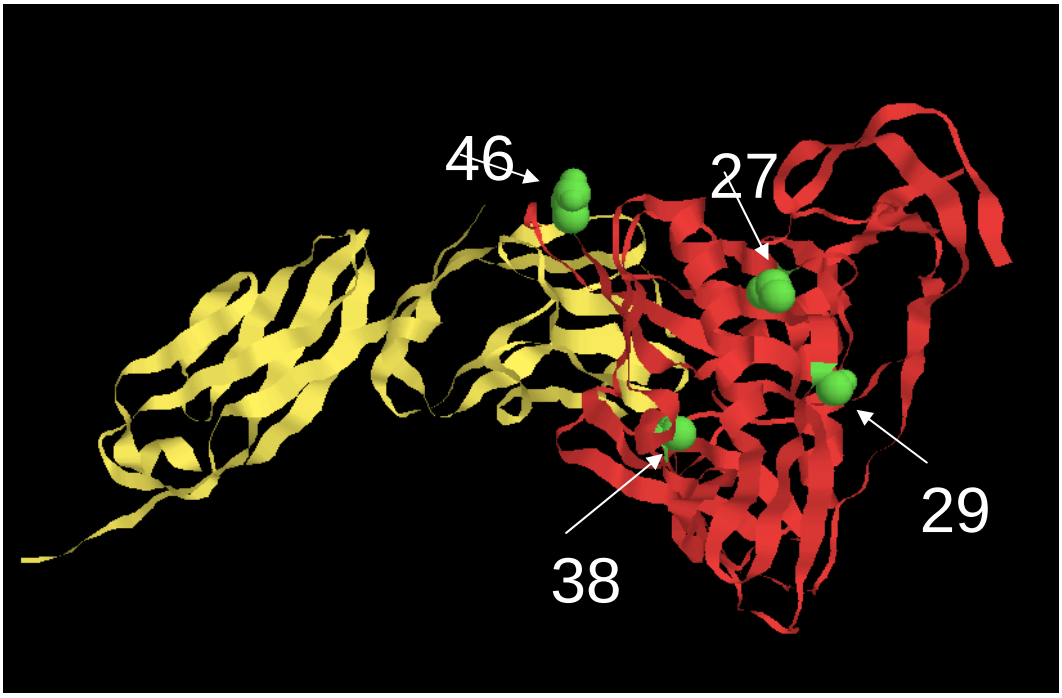
Supplementary Figure 4. Neighbor-joining tree of all individuals demonstrating viral compartmentalization between blood and semen (individuals A-G, only compartmentalized subset of sequence data from individual G is included here). Bubbles represent sequences from a single individual. HIV-1 HXB2 included as outgroup sequence.



Supplementary Figure 5. Longitudinal diversity and divergence in four individuals (G, I, J, and K).



Supplementary Figure 5 (cont'd). Longitudinal diversity and divergence in four individuals (G, I, J, and K).



Supplementary Figure 6. HIV-1 HXB2 gp120 structure with semen-specific signature residues highlighted. Red = gp120, yellow = CD4, green = signature residue.

Chapter 5

Genetic Attributes of Cerebrospinal Fluid-Derived HIV-1

ABSTRACT

Background: HIV-1 often invades the central nervous system (CNS) during primary infection, eventually resulting in neurological disorders in up to 50% of untreated patients. The CNS is a distinct viral reservoir, differing from peripheral tissues in immunological surveillance, target cell characteristics, and antiretroviral penetration. Neurotropic HIV-1 likely develops distinct genotypic characteristics in response to this unique selective environment.

Objective: To catalog the genetic features of neurotropic HIV-1, and to evaluate the contribution of viral genetics to neurovirulence.

Methods: 456 clonal HIV-1 RNA sequences of the C2-V3 *env* subregion were generated from CSF and plasma of 18 chronically infected patients. Neuropsychological performance of all subjects was evaluated and summarized as a global deficit score. A battery of phylogenetic, statistical, and machine learning tools was applied to these data to identify genetic features associated with HIV-1 neurotropism and neurovirulence.

Results: 11 of 18 patients exhibited significant compartmentalization between blood and CSF-derived virus ($p<0.01$, Slatkin-Maddison test). A CSF-specific genetic signature was identified, comprising positions 9, 13, and 32 of the V3 loop. CSF-derived sequences exhibited constrained diversity, while containing fewer glycosylated and positively selected sites. In addition, the presence of serine at position 5 of the V3 loop was highly correlated with neurocognitive deficit ($p<0.0025$, Fisher's Exact test).

Conclusions: There are several genetic features that distinguish CSF- and plasma-derived HIV-1 populations. CSF-specific characteristics of HIV-1 *env* likely reflect altered cellular entry requirements and decreased immune pressure in the CNS. Furthermore, neurological impairment may be influenced by mutations within the V3 loop.

Keywords: HIV-1, evolution, compartmentalization, central nervous system

INTRODUCTION

HIV-1 and SIV cross the blood-brain barrier during primary infection, eventually resulting in neurological complications in up to 50% of untreated individuals (5,14,16,17,36,41,62). HIV-associated dementia, encephalopathy, and sensory neuropathies contribute significantly to morbidity and mortality (27). In addition, the central nervous system (CNS) serves as a sanctuary site for long-term viral persistence due to the suboptimal penetration of several antiretroviral agents (28).

The CNS is a distinct viral reservoir, differing from peripheral tissues in immunological surveillance, cytokine milieu, target cell characteristics, and antiretroviral penetration. Evidence from humans and chimpanzees suggests that selective pressure from anti-HIV neutralizing antibodies and cytotoxic T cells may be diminished in the brain and cerebrospinal fluid (13,26,34,45,51,56). Cytokines such as IL-16, TNF-alpha, and RANTES that modulate HIV replication may differ in relative concentrations between the CNS and blood plasma (11). The predominant targets of HIV-1 infection in the CNS are brain-derived macrophages and microglial

cells, rather than the CD4+ lymphocytes that serve as targets in the periphery (10). Several antiretroviral drugs do not efficiently cross the blood-brain barrier, resulting in only partial suppression of viral replication in the central nervous system (28,52).

The uniqueness of the CNS environment is often reflected in compartment-specific HIV-1 genotypic and phenotypic characteristics. Contemporaneous CNS- and blood-derived viruses are frequently compartmentalized based on phylogenetic analysis of intrapatient sequences (12,22,32,39,53,60), and brain-derived *env* sequences may share signature mutations across individuals (22,40). Differences in CD4 dependence, coreceptor usage phenotype, and LTR sequence have been observed between brain-and blood-derived HIV isolates, reflecting differences in chemokine receptor expression and transcriptional environment between microglia and peripheral host cells (1,2,18,44,48,61). The presence of discordant drug resistance mutations in CNS and plasma viral populations likely results from tissue-specific variation in drug efficacies due to the poor penetration of certain antiretroviral agents (53,60). Moreover, recent evidence suggests that the evolution of resistance may differ between brain subcompartments (50).

HIV-1 RNA is detectable in the cerebrospinal fluid (CSF) of most infected individuals throughout disease. Due to the obvious sampling difficulties associated with brain tissue, CSF virus has often been investigated as a proxy for brain-derived HIV-1 (studies of brain virus are almost exclusively limited to post-mortem samples). This indirect sampling strategy is supported by phylogenetic evidence that CSF and brain-derived viral populations are more closely related to each other than with

populations derived from bone marrow, kidney, liver, lung, lymph nodes, and spleen (46). We sought to determine the genetic basis of HIV-1 neurotropism by systematically comparing CSF- and plasma-derived *env* sequences from eighteen chronically infected donors. In addition we investigated the genetic basis of neurovirulence by comparing CSF-derived sequences from several individuals with known global deficit scores (GDS) based on a comprehensive neuropsychological evaluation.

MATERIALS AND METHODS

Subjects. Twenty-one individuals enrolled in longitudinal clinical studies at the HIV Neurobehavioral Research Center (HNRC) between 1998 and 2002 were studied. All subjects had stable or no antiviral therapy for at least 2 months prior to study, had plasma and CSF HIV RNA of >500 copies/ml and had no evidence of systemic or CNS opportunistic infections or malignancy based on clinical, laboratory and neuro-imaging studies. Data were available on past and present therapy, current HIV RNA and CD4 counts, nadir CD4 counts and CSF cell counts. All studies were conducted in compliance with local IRB guidelines and with subjects' written informed consent.

Specimen processing. Paired blood from peripheral venipuncture in ACD tubes and CSF from lumbar punctures were collected (typically collected within one hour of each other) and processed within 2 hours of collection. Plasma and cell free CSF were aliquoted, frozen and stored at -70^{o} C until processing. All subsequent

plasma and CSF processing was performed separately to minimize the within-subject cross contamination of samples.

Nucleotide sequencing. Sequencing methods were previously described in full (53). In brief, reverse transcription and PCR amplification of C2-V3 *env* for each sample was performed in triplicate or quadruplicate using the Finnzyme one step RT-PCR kit (MJ Research, Waltham, MA) and primers V3Fout and V3Bout as previously described (1) in a 25µl reaction volume. 2.5µl of first step RT-PCR product was used in the second, nested PCR reaction with primers V3Fin and V3Bin (1). All assays were conducted in conditions to minimize the potential for PCR contamination utilizing aerosol resistant pipet tips, dedicated PCR reagents and laminar flow hoods. All assays included negative controls. Replicate PCR products were proportionately pooled and cloned using the TOPO-TA cloning system (Invitrogen, Carlsbad, CA). Purified plasmids were sequenced in both directions with –20M13 primer (5'-gtaaaacgacggccag-3') and Topo Forward primer (5'-tggatatctgcagaattcg-3') using Prism Dye terminator kits (ABI, Foster City, CA) on an ABI 3100 Genetic Analyzer. Sequences were compiled, aligned, and edited using Sequencher 4.0 (Genecodes, Ann Arbor, MI) and Clustal (version 1.81).

Neuropsychological assessment. Subjects completed a detailed neuropsychological assessment measuring their functioning in 8 cognitive ability domains: verbal functioning, abstraction, complex perceptual-motor skills, attention, learning, memory, motor skills and sensory functioning. Test results were summarized as "deficit scores" which reflect the number and severity of impaired

performances throughout the test battery, and give relatively less weight to test performances within or above the average range. The demographically corrected T-score for each test measure is converted to a zero to five point deficit rating, as follows: T>39 = 0 (no impairment), 35-39T = 1 point (mild impairment); 30-34T = 2 points (mild to moderate impairment); 25-29T = 3 points (moderate impairment); 20-24T = 4 points (moderate to severe impairment); T<20 = 5 points (severe impairment). A Global Deficit Score (GDS) is computed by adding the deficit ratings of the component test measures and dividing by the total number of measures. Deficit scores are sensitive to the presence and pattern of NP impairments in HIV+ individuals (16). Statistical classification of NP impaired/NP normal was made through the use of a GDS cut-off score that demonstrates high accuracy in predicting clinician ratings of NP status. A GDS of 0.5 or greater is considered to be in the impaired range. This represents at least mild impairment on half of the tests of the NP battery (3).

Phylogenetic reconstruction. Initial multiple sequence alignments were generated using ClustalX (54), with default gap parameters and the "IUB" DNA weight matrix. Subsequent manual aligning was performed using the Se-Al sequence alignment editor (43). Phylogenies describing sequences from each individual host were built using FastDNAml (33), estimating base frequencies from the data, Ts/Tv ratio of 2.0. Diversity measurements were calculated using dnadist and protdist (7). A master tree describing the entire data set was built by implementing dnadist and neighbor within the Phylip 3.5c package (7) using the F84 model, Gamma distributed rates across sites, and Ts/Tv ratio of 2.0. Trees were viewed using TreeView X (35).

Evaluation of compartmentalization. The degree of segregation between compartments was assessed by testing for panmixis using gene phylogenies (19,49) as implemented in MacClade (Sinauer, Sunderland, MA). In brief, the minimum possible number of inter-compartment migration events was tallied, based on the maximum likelihood trees for each individual subject's C2-V3 sequences and their characterization according to compartment of origin. This result was compared to the distribution of migration events in 1000 trees in which the taxa have been randomly shuffled across tips, retaining the original topology and associated polytomies (31). Evidence of restricted gene flow (compartmentalization) was documented when <1% of the randomized trees required the same or a fewer number of migration events as for the sample data (49).

Calculation of Shannon entropy. Residue specific entropy was computed from the frequency f(Ai) of amino acid A at position i according to $-\sum A\ f(Ai)\ \ln[f(Ai)]$.

Machine learning classification. A machine learning approach was employed to look for a tissue-specific genetic signature. All classification experiments in this analysis were conducted using WEKA (Waikato Environment for Knowledge Analysis), an open- source collection of data-processing and machine learning algorithms (58). The J48 decision tree inducer, based on the C4.5 algorithm (42) was implemented with the parameter "MinNumObj" set at a value of 11 to limit the complexity of theories and minimize the risk of over-fitting. Classifiers were evaluated using one hundred iterations of stratified ten-fold cross-validation, a

procedure designed to reflect the performance of classification models on novel data sets. For each of 100 trials, the data set was randomly divided into 10 groups of approximately equal size and class distribution. For each "fold," the classifier was trained using all but 1 of the 10 groups and then tested on the unseen group. This procedure was repeated for each of the 10 groups. The cross-validation score for 1 trial was the average performance across each of the 10 training runs. The reported score is the average across the 100 trials (58).

Analysis of selection. We used maximum likelihood methods that fit independent synonymous (alpha) and non-synonymous (beta) rate parameters to each site in the codon alignment. For codon site s, we tested the hypothesis of differential selection between two populations (alpha_1, beta_1, alpha_2 and beta_2 estimated by maximum likelihood), versus the null hypothesis of identical selection (beta_1 = R alpha_1, beta_2 = R alpha_2, with alpha_1, alpha_2 and R estimated by maximum likelihood) using the likelihood ratio test, and the chi-squared (one degree of freedom) distribution to assess significance (25). Other phylogenetic parameters, such as branch lengths, base frequencies and nucleotide substitution biases were estimated from each individual alignment and held constant during subsequent site comparisons.

Coreceptor usage prediction. A previously trained machine learning algorithm (support vector machine) was employed to predict the coreceptor usage of viruses based on V3 loop amino acid sequence (37). This method is reported to predict CXCR4 usage with a specificity of 93% (20). The coreceptor classifier is available for public use at: http://genomiac2.ucsd.edu:8080/wetcat/tropism.html

RESULTS

Compartmentalization of CSF-derived virus.

To determine the genetic characteristics of cerebrospinal fluid-derived HIV-1, we generated and analyzed 456 clonal CSF and plasma *env* sequences from eighteen chronically infected individuals. If the CSF represents a distinct viral compartment, contemporaneous CSF- and plasma-derived sequences are expected to cluster independently (31). We previously demonstrated that independent clustering of tissue-specific populations was observed in eleven of these eighteen individuals (53), determined by applying the parsimony-based cladistic method of Slatkin and Maddison to maximum likelihood phylogenetic reconstructions (49). However, inter-population gene flow may be underestimated by this approach, due to the potential loss of polytomies in the randomly generated trees used to evaluate statistical significance (30). We circumvented this issue by modifying the Slatkin-Maddison test; we generated 1000 random trees in which the taxa have been randomly shuffled across tips, retaining the original topology and associated polytomies. Evidence of restricted gene flow (compartmentalization) was documented when <1% of the randomized trees required the same or a fewer number of migration events as for the sample data. The results of this more stringent test of panmixis did not conflict with our original findings. Eleven of eighteen individuals (individuals A, F, H, J, M-S) in this cohort exhibited independent clustering of tissue-specific populations ($p<0.01$), while two of the remaining seven individuals (B, L) harbored variants that exhibited partial compartmentalization between tissues ($0.01<p<0.05$) (Fig. 1).

Amino acid diversity in plasma- and CSF-derived viral populations.

We next calculated the amino acid diversity of C2-V3 sequences from CSF and plasma, focusing on average pairwise distances derived using the Dayhoff PAM substitution matrix. There was no significant difference in overall protein diversity between CSF- and plasma-derived sequences pooled across all eighteen individuals (data not shown). When focusing exclusively on the V3 loop subregion, however, there was a significant reduction of diversity in CSF-derived sequences ($p<0.05$, paired Wilcoxon) (Fig. 2). Tissue-specific patterns of diversity did not differ between compartmentalized and non-compartmentalized individuals.

Correlation between infection date and HIV-1 genetic diversity.

Several reports suggest that HIV-1 intrapatient diversity is correlated with duration of infection and disease stage (47,59). We investigated this relationship by comparing CSF- and plasma-derived viral genetic diversity at the nucleotide level against date of infection (as reported by the infected individual). A moderately positive correlation existed between plasma viral diversity and infection date ($R^2=0.27$) (Fig. 3). No correlation, however, was apparent between CSF viral diversity and infection date.

CSF-specific Env genetic signature.

CSF-derived viruses may share genetic characteristics across individuals due to tissue-specific selective pressures that are common across hosts. We employed a previously described machine learning approach to look for evidence of a genetic

signature shared by CSF-derived sequences pooled across individuals (38). The j48 decision tree inducer (based on the Quinlan C4.5 algorithm) was implemented to classify Env sequences from all individuals based on tissue of origin. The training data for this experiment drew samples from the entire available sequence set, consisting of 231 plasma sequences and 225 from CSF. Our results (data not shown) indicated that there was no evidence of a signature, due to the high error rate associated with the classification trial; sequence tropism was misclassified in nearly 50% of test cases. However, when we limited our analysis to the 94 plasma and 130 CSF sequences associated with compartmentalized individuals (A, F, H, J, M-S), classification accuracy (conservatively estimated using a cross-validation procedure) increased dramatically to 87%. The genetic signature underlying the classification model consisted of positions 5, 9, 13, and 19 of the V3 loop (HXB2 gp160 positions 300, 304, 308, and 314, respectively) (Fig. 4). The presence of proline or histidine at V3 loop position 13 (gp160 position 308) was significantly correlated with compartmentalization in the CNS ($p<0.044$, Fisher's Exact test). Compartmental differences in relative amino acid composition at these signature sites are evident in Env alignments of each individual's sequences (Supplementary Fig. 1) and in (pooled) consensus sequence logos (Fig. 5).

Comparison of site-specific entropy.

To determine if any specific Env residues exhibited discordant levels of conservation between tissues, we assessed the variability at each site in our amino acid alignment by calculating site-specific Shannon entropy scores. Entropy estimates

account for both the total number and relative frequencies of different residues at a site. Our analysis of the C2-V3 region revealed that several sequence positions had highly divergent Shannon entropy scores in CSF and plasma (Fig. 6). The five highest net differences in entropy were observed at positions 304, 308, 340, 341, and 360 (numbered according to the HXB2 gp160 sequence). Sites 304 and 308 are located in the V3 loop region.

Identification of discordantly selected sites.

We used a maximum likelihood approach to identify codons within *env* that were under discordant selection pressure in CSF and plasma. Selection pressure is described as the ratio between nonsynonymous substitutions per nonsynonymous site (dN) and synonymous substitutions per synonymous site (dS) (29). A total of seven sites in the C2-V3 region exhibited discordant levels of selection pressure (dN/dS) in CSF and plasma, based on a differential p-value cutoff of 0.1 (Table 1). Five out of these seven sites were under strong negative selection in CSF while corresponding sites were being positively selected or evolving neutrally in blood plasma (Table 1). These data are in alignment with our compartment-specific maximum likelihood estimates of global dN/dS across the C2-V3 region. CSF-derived HIV-1 sequences exhibited considerably lower dN/dS values than plasma-derived sequences (data not shown).

N-linked glycosylation in plasma- and CSF-derived viral populations.

We next examined N-linked glycosylation patterns across the C2-V3 *env* subregion, to investigate variation in selection pressure from the neutralizing antibody

response (57). If the antibody response is attenuated in the central nervous system, we might expect fewer glycosylation sites within CSF-derived viral sequences. If the response is equivalent, but targeting different epitopes, we might expect a reassortment of sites though the overall number may remain constant. Our analysis revealed that extent of glycosylation, reported as average number of glycosylation sites (sequons) per sequence tended to be lower in CSF-derived viruses, although this trend was not quite significant ($p<0.062$, paired Wilcoxon) (Fig. 7). Virus from seven out of eleven compartmentalized individuals exhibited significantly different levels of glycosylation between compartments ($p<0.05$, Mann-Whitney), and five out of these seven had lower numbers of sequons in CSF-derived sequences (Fig. 7).

Prediction of coreceptor usage.

We predicted the chemokine receptor preferences for all sequences in our data set to determine if neurotropism was correlated with skewed coreceptor usage. We had previously reported that predicted coreceptor usage did not differentiate between CSF- and plasma-derived populations (53) when using the "11/25" rule to predict phenotype (8). However, predicting coreceptor usage based on the presence of a basic residue at V3 loop position 11 and/or 25 is not entirely reliable, and tends to underestimate CXCR4 usage (20). We revisited this issue by predicting phenotypes using a previously trained machine learning algorithm (37) that is significantly more sensitive to the CXCR4 class (20). Our reanalysis reinforced the original finding that coreceptor phenotype does not differ between CSF- and plasma-derived viral populations. The vast majority of sequences from both compartments were predicted

to use CCR5 as a coreceptor, although a subset of plasma clones in two individuals (I and M) were predicted to use CXCR4 (data not shown).

Correlation between Env sequence and neurovirulence.

All of the subjects involved in this study underwent a comprehensive neuropsychological assessment measuring their functioning in 8 cognitive ability domains: verbal functioning, abstraction, complex perceptual-motor skills, attention, learning, memory, motor skills and sensory functioning. Test results were summarized as "deficit scores" which reflect the number and severity of impaired performances throughout the test battery. Global deficit scores (GDS) in this cohort ranged from 0.31 (normal) to 3.5 (moderate to severe impairment) (Table 2). We looked for correlations between HIV-1 Env sequence and cognitive deficit scores using machine learning-based regression analysis. The residue at position 5 of the V3 loop (HXB2 gp160 position 300) was most strongly correlated with GDS. The presence of serine at position 5 was significantly correlated with topmost quartile global deficit scores ($p<.0025$, Fisher's Exact Test) (Table 2).

DISCUSSION

The results of our investigation reveal that the genetics of cerebrospinal fluid- and blood plasma-derived strains of HIV-1 differ on several levels, confirming previous reports that HIV-1 within the central nervous system can differ from virus in peripheral tissues. This study extends those observations by cataloguing the CSF-specific population genetic features of the HIV-1 quasispecies. First, the CNS (as

represented by cerebrospinal fluid) can function as a viral compartment in most, but not all infected individuals. Second, sequence diversity of the V3 loop Env subregion measured at the amino acid level is reduced in CSF-derived viral populations. Third, there is a CSF-specific HIV-1 genetic signature associated with sequences from compartmentalized individuals comprising 4 sites within the V3 loop region. Fourth, several sites in the viral envelope are under different levels of selective constraint in CSF and plasma, and exhibit discordant levels of entropy in these tissues. Lastly, CSF-derived viruses tend to be less glycosylated than blood-derived viruses.

Viral compartments are characterized by a restriction of gene flow between cells or tissues, usually identified by phylogenetic analysis (30,38). In this study, viral compartmentalization between blood and the central nervous system was identified in 13 out of 18 individuals. Viral migration across the blood-brain barrier was minimal and infrequent in these individuals, which reinforces the concept that a significant fraction of virus sampled in CSF is produced locally in the CNS (46). Furthermore, CSF-derived V3 loop sequences were under stronger negative selection and exhibited reduced levels of amino acid diversity. These genetic characteristics likely reflect constraints associated with tissue-specific cellular entry determinants and reduced immune selection pressure in the CNS (1,34,48). A low viral effective population size in the CNS may contribute to the reduced diversity as well (24), and is reinforced by our observed lack of correlation between duration of infection and CSF-derived HIV-1 diversity. The positive correlation between genetic diversity of plasma virus and infection date, however, is concordant with observations that sequence diversity often

diminishes during late stages of disease, due to a lack of diversifying selection attributed to immune collapse and target cell homogeneity (47,59).

The identification of a CSF-specific HIV-1 genetic signature across compartmentalized individuals is strong evidence that the viral quasispecies adapts to the local fitness landscape within the CNS, and moreover, that commonalities in this selective environment exist across individuals. Position 308 (V3 loop position 13) was the most informative sequence position. The contribution of position 308 to HIV-1 neurotropism is highlighted by its discordant entropy scores in CSF and plasma and by the fact that it has been featured in the reports of multiple investigators over the last decade (22,40). In addition, the presence of certain residues at position 308 has been associated with macrophage tropism, which is most likely correlated with microglial tropism due to the extensive similarities between these cell types (4). Understanding the relevance of this sequence position to neurotropism and neuropathogenesis may be achieved by infecting *in vitro* fetal brain aggregates (55) or microglial cell cultures with genetically defined HIV-1 strains to determine the fitness consequences associated with p308 polymorphisms.

Our exploration of the relationship between cognitive deficit in the host and viral genetics suggests that V3 loop sequence may be a genetic determinant of neurovirulence (23). The contribution of V3 loop position 5 to neurovirulence may result from accelerated pathogenesis due to improved replicative capacity within the CNS (6). However, HIV-1 neurotropism and neurovirulence may be disctinct and separable phenomena (39). Although certain Env mutations may not enhance the

replication kinetics of HIV-1 within the CNS, they may increase gp160-mediated neurotoxicity due to alterations in interactions between virion surface glycoproteins and host cell surface molecules (21).

The central nervous system has been characterized as a reservoir, a compartment, and a drug sanctuary (9,15). Our investigations characterize the differing selection pressures between the CSF and blood and document a specific genetic signature of virus compartmentalized in the central nervous system. Taken together, these data offer important insights into the adaptation of HIV to the CNS environment, which may prove valuable in managing HIV-1 infection and preventing the development of neurological disorders.

ACKNOWLEDGMENTS

The text of this chapter, in full, has been submitted for publication: Pillai S. K., S. Kosakovsky Pond, B. Good, M. C. Strain, R. Ellis, S. Letendre, H. Gunthard, I. Grant, T. Marcotte, J. A. McCutchan, D. D. Richman, and J. K. Wong, "Genetic Attributes of Cerebrospinal Fluid-Derived HIV-1. I was the primary author, and the co-authors listed in this manuscript supervised and/or contributed to the research which forms the basis for this chapter.

This work was supported by grants 5K23AI055276, AI27670, AI38858, AI43638, AI43752, AI36214 (UCSD Center for AIDS Research), AI29164, and AI047745 from the National Institutes of Health. Additional support was provided by the Research Center for AIDS and HIV Infection of the San Diego Veterans Affairs Healthcare System.

REFERENCES

1. Argyris E. G., E. Acheampong, G. Nunnari, M. Mukhtar, K. J. Williams, and R. J. Pomerantz. 2003. Human immunodeficiency virus type 1 enters primary human brain microvascular endothelial cells by a mechanism involving cell surface proteoglycans independent of lipid rafts. J Virol 77:12140-51.

2. Burdo, T. H., M. Nonnemacher, B. P. Irish, C. H. Choi, F. C. Krebs, S. Gartner, and B. Wigdahl. 2004. High-affinity interaction between HIV-1 Vpr and specific sequences that span the C/EBP and adjacent NF-kappaB sites within the HIV-1 LTR correlate with HIV-1-associated dementia. DNA Cell Biol 23:261-9.

3. Carey, C. L., S. P. Woods, R. Gonzales, C. E., T. D. Marcotte, I. Grant, R. K. Heaton, and H. Group. 2004. Predictive validity of global deficit scores in detecting neuropsychological impairment in HIV infection. Journal of Clinical and Experimental Neuropsychology:In Press.

4. Chesebro, B., K. Wehrly, J. Nishio, and S. Perryman. 1996. Mapping of independent V3 envelope determinants of human immunodeficiency virus type 1 macrophage tropism and syncytium formation in lymphocytes. J Virol 70:9055-9.

5. Clements, J. E., T. Babas, J. L. Mankowski, K. Suryanarayana, M. Piatak, Jr., P. M. Tarwater, J. D. Lifson, and M. C. Zink. 2002. The central nervous system as a reservoir for simian immunodeficiency virus (SIV): steady-state levels of SIV DNA in brain from acute through asymptomatic infection. J Infect Dis 186:905-13.

6. Ellis, R. J., K. Hsia, S. A. Spector, J. A. Nelson, R. K. Heaton, M. R. Wallace, I. Abramson, J. H. Atkinson, I. Grant, and J. A. McCutchan. 1997. Cerebrospinal fluid human immunodeficiency virus type 1 RNA levels are elevated in neurocognitively impaired individuals with acquired immunodeficiency syndrome. HIV Neurobehavioral Research Center Group. Ann Neurol 42:679-88.

7. Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c.

8. Fouchier, R. A. M., M. Groenink, N. A. Kootstra, M. Tersmette, H. G. Huisman, F. Miedema, and H. Schuitemaker. 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. Journal of Virology 66:3183-3187.

9. Foudraine, N. A., R. M. Hoetelmans, J. M. Lange, F. De Wolf, B. H. van Benthem, J. J. Maas, I. P. Keet, and P. Portegies. 1998. Cerebrospinal-fluid HIV-

1 RNA and drug concentrations after treatment with lamivudine plus zidovudine or stavudine. Lancet 351:1547-1551.

10. Gartner, S., P. Markovits, D. M. Markovitz, M. H. Kaplan, R. C. Gallo, and M. Popovic. 1986. The role of mononuclear phagocytes in HTLV-III/LAV infection. Science 233:215-9.

11. Gonzalez-Scarano, F., and J. Martin-Garcia. 2005. The neuropathogenesis of AIDS. Nat Rev Immunol. Jan;5(1):69-81.

12. Gorry, P. R., G. Bristol, J. A. Zack, K. Ritola, R. Swanstrom, C. J. Birch, J. E. Bell, N. Bannert, K. Crawford, H. Wang, D. Schols, E. De Clercq, K. Kunstman, S. M. Wolinsky, and D. Gabuzda. 2001. Macrophage tropism of human immunodeficiency virus type 1 isolates from brain and lymphoid tissues predicts neurotropism independent of coreceptor specificity. J Virol 75:10073-89.

13. Goudsmit, J., L. G. Epstein, D. A. Paul, H. J. van der Helm, G. J. Dawson, D. M. Asher, R. Yanagihara, A. V. Wolff, C. J. Gibbs Jr, and D. C. Gajdusek. 1987. Intra-blood-brain barrier synthesis of human immunodeficiency virus antigen and antibody in humans and chimpanzees. Proc Natl Acad Sci USA 84:3876-80.

14. Grant, I., J. H. Atkinson, J. R. Hesselink, C. J. Kennedy, D. D. Richman, S. A. Spector, and J. A. McCutchan. 1987. Evidence for early central nervous system involvement in the acquired immunodeficiency syndrome (AIDS) and other human immunodeficiency virus (HIV) infections. Annals of Internal Medicine 107:828-836.

15. Gunthard, H. F., D. V. Havlir, S. Fiscus, Z. Q. Zhang, J. Eron, J. Mellors, R. Gulick, S. D. F. Frost, A. J. Leigh-Brown, W. Schlief, F. Valentine, L. Jonas, A. Meibohm, C. Ignacio, R. Isaacs, R. Gamagami, E. Emini, A. T. Haase, D. D. Richman, and J. K. Wong. 2001. Residual HIV RNA and DNA in lymph node and HIV RNA in genital secretions and in CSF after two years of suppression of viremia in the Merck 035 cohort. Journal of Infectious Diseases 183:1318-1327.

16. Heaton, R. K., I. Grant, N. Butters, D. A. White, D. Kirson, J. H. Atkinson, J. A. McCutchan, M. J. Taylor, M. D. Kelly, R. J. Ellis, and et al. 1995. The HNRC 500--neuropsychology of HIV infection at different disease stages. HIV Neurobehavioral Research Center. J Int Neuropsychol Soc 1:231-51.

17. Ho, D. D., T. R. Rota, R. T. Schooley, J. C. Kaplan, J. D. Allan, J. E. Groopman, L. Resnick, D. Felsenstein, C. A. Andrews, and M. S. Hirsch. 1985. Isolation of HTLV-III from cerebrospinal fluid and neural tissues of patients with neurologic syndromes related to the acquired immunodeficiency syndrome. N Engl J Med 313:1493-7.

18. Hogan, T. H., D. L. Stauff, F. C. Krebs, S. Gartner, S. J. Quiterio, and B. Wigdahl. 2003. Structural and functional evolution of human immunodeficiency virus type 1 long terminal repeat CCAAT/enhancer binding protein sites and their use as molecular markers for central nervous system disease progression. J Neurovirol 9:55-68.

19. Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583-589

20. Jensen, M. A. and A. B. van't Wout. 2003. Predicting HIV-1 coreceptor usage with sequence analysis. AIDS Rev. 5:104-112.

21. Kanmogne, G. D., R. C. Kennedy, and P. Grammas. 2002. HIV-1 gp120 proteins and gp160 peptides are toxic to brain endothelial cells and neurons: possible pathway for HIV entry into the brain and HIV-associated dementia. J Neuropathol Exp Neurol 61:992-1000.

22. Korber, B. T., K. J. Kunstman, B. K. Patterson, M. Furtado, M. M. McEvilly, R. Levy, and S. M. Wolinsky. 1994. Genetic differences between blood- and brain-derived viral sequences from human immunodeficiency virus type 1-infected patients: evidence of conserved elements in the V3 region of the envelope protein of brain-derived sequences. J Virol 68:7467-81.

23. Kuiken, C. L., J. Goudsmit, G. F. Weiller, J. S. Armstrong, S. Hartman, P. Portegies, J. Dekker, and M. Cornelissen. 1995. Differences in human immunodeficiency virus type 1 V3 sequences from patients with and without AIDS dementia complex. J Gen Virol 76:175-80.

24. Leigh Brown, A. J. 1997. Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci USA 94:1862–1865.

25. Leigh Brown, A. J., S. L. Kosakovsky Pond, Z. Grossman, D. D. Richman, and S. D. W. Frost. Adaptation to different individuals and different populations by HIV-1. (submitted).

26. Ljunggren, K., F. Chiodi, P. A. Broliden, J. Albert, G. Norkrans, L. Hagberg, M. Jondal, and E. M. Fenyo. 1989. HIV-1-specific antibodies in cerebrospinal fluid mediate cellular cytotoxicity and neutralization. AIDS Res Hum Retroviruses 5:629-38.

27. McArthur, J. C., N. Haughey, S. Gartner, K. Conant, C. Pardo, A. Nath, and N. Sacktor. 2003. Human immunodeficiency virus-associated dementia: an evolving disease. J Neurovirol 9:205-21.

28. Misra, A., S. Ganesh, A. Shahiwala, and S. P. Shah. 2003. Drug delivery to the central nervous system: a review. J Pharm Pharm 6:252-73.

29. Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426.

30. Nickle, D. C., M. A. Jensen, D. Shriner, S. J. Brodie, L. M. Frenkel, J. E. Mittler, and J. I. Mullins. 2003. Evolutionary indicators of human immunodeficiency virus type 1 reservoirs and compartments. J Virol 77:5540-6.

31. Nickle, D. C., D. Shriner, J. E. Mittler, L. M. Frenkel, and J. I. Mullins. 2003. Importance and detection of virus reservoirs and compartments of HIV infection. Curr Opin Microbiol 6:410-6.

32. Ohagen, A., A. Devitt, K. J. Kunstman, P. R. Gorry, P. P. Rose, B. Korber, J. Taylor, R. Levy, R. L. Murphy, S. M. Wolinsky, and D. Gabuzda. 2003. Genetic and functional analysis of full-length human immunodeficiency virus type 1 env genes derived from brain and blood of patients with AIDS. J Virol 77:12336-45.

33. Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek. 2004. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput.Appl.Biosci. 10:41-48

34. Pachter J. S., H. E. de Vries, and Z. Fabry. 2003. The blood-brain barrier and its role in immune privilege in the central nervous system. J Neuropathol Exp Neurol 62:593-604.

35. Page, R. D. M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comput.Appl.Biosci. 12:357-358

36. Pierson, T., J. McArthur, and R. F. Siliciano. 2000. Reservoirs for HIV-1: mechanisms for viral persistence in the presence of antiviral immune responses and antiretroviral therapy. Annu Rev Immunol 18:665-708.

37. Pillai, S., B. Good, D. Richman, and J. Corbeil. 2003. A new perspective on V3 phenotype prediction. AIDS Res. Hum. Retroviruses 19:145-149.

38. Pillai, S.K., B. Good, S. Kosakovsky Pond, J.K. Wong, M.C. Strain, D.D. Richman, and D.M. Smith. 2005. Semen-Specific Genetic Characteristics of Human Immunodeficiency Virus Type 1 *env*. J Virol 79:1734-1742

39. Power, C., J. C. McArthur, R. T. Johnson, D. E. Griffin, J. D. Glass, R. Dewey, and B. Chesebro. 1995. Distinct HIV-1 env sequences are associated with neurotropism and neurovirulence. Curr Top Microbiol Immunol 202:89-104.

40. Power, C., J. C. McArthur, R. T. Johnson, D. E. Griffin, J. D. Glass, S. Perryman, and B. Chesebro. 1994. Demented and nondemented patients with AIDS differ in brain-derived human immunodeficiency virus type 1 envelope sequences. J Virol 68:4643-49.

41. Price, R. W. 1996. Neurological complications of HIV infection. Lancet 348:445-52.

42. Quinlan, J. R. 1993. C4.5: Programs for machine learning. Morgan Kaufmann, San Francisco.

43. Rambaut A. 2002. Se-Al sequence alignment editor v2.0 (Software). Department of Zoology, University of Oxford.

44. Ross, H. L., S. Gartner, J. C. McArthur, J. R. Corboy, J. J. McAllister, S. Millhouse, and B. Wigdahl. 2001. HIV-1 LTR C/EBP binding site sequence configurations preferentially encountered in brain lead to enhanced C/EBP factor binding and increased LTR-specific activity. J Neurovirol 7:235-49.

45. Ruta, S. M., R. Matusa, and C. C. Cernescu. 1998. Cerebrospinal fluid western Blot profiles in the evolution of HIV-1 pediatric encephalopathy. Rom J Virol 49:61-71.

46. Sanjuan, R., F. M. Codoner, A. Moya, and S. F. Elena. 2004. Natural selection and the organ-specific differentiation of HIV-1 V3 hypervariable region. Evolution Int J Org Evolution 58:1185-94.

47. Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins. 1999. Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection. J Virol 73:10489-10502.

48. Shieh, J. T., A. V. Albright, M. Sharron, S. Gartner, J. Strizki, R. W. Doms, and F. Gonzalez-Scarano. 1998. Chemokine receptor utilization by human immunodeficiency virus type 1 isolates that replicate in microglia. J Virol 72:4243-9.

49. Slatkin, M. and W. P. Maddison. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123:603-613.

50. Smit, T. K., B. J. Brew, W. Tourtellotte, S. Morgello, B. B. Gelman, and N. K. Saksena. 2004. Independent evolution of human immunodeficiency virus (HIV) drug resistance mutations in diverse areas of the brain in HIV-infected patients, with and without dementia, on antiretroviral treatment. J Virol 78:10133-48.

51. Song, B., M. Cayabyab, N. Phan, L. Wang, M. K. Axthelm, N. L. Letvin, and J. G. Sodroski. 2004. Neutralization sensitivity of a simian-human immunodeficiency virus (SHIV-HXBc2P 3.2N) isolated from an infected rhesus macaque with neurological disease. Virology 322:168-81.

52. Staprans, S., N. Marlowe, D. Glidden, T. Novakovic-Agopian, R. M. Grant, M. Heyes, F. Aweeka, S. Deeks, and R. W. Price. 1999. Time course of cerebrospinal fluid responses to antiretroviral therapy: evidence for variable compartmentalization of infection. AIDS 13:1051-61.

53. Strain M. C., S. Letendre, S. Pillai, T. Russell, C. C. Ignacio, H. F. Gunthard , B. Good, D. M. Smith, S. M. Wolinsky, M. Furtado, J. Marquie-Beck, J. Durelle, I. Grant, D. D. Richman, T. Marcotte, J. A. McCutchan, R. J. Ellis, and J. K. Wong. 2005. Genetic composition of HIV-1 in CSF and plasma without treatment and during failing combination antiretroviral therapy. J Virol 79:1772-88.

54. Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

55. Trillo-Pazos, G., A. Kandanearatchi, J. Eyeson, D. King, A. Vyakarnam, and I. P. Everall. 2004. Infection of stationary human brain aggregates with HIV-1 SF162 and IIIB results in transient neuronal damage and neurotoxicity. Neuropathol Appl Neurobiol 30:136-47.

56. von Gegerfelt, A., F. Chiodi, B. Keys, G. Norkrans, L. Hagberg, E. M. Fenyo, K. Broliden. 1992. Lack of autologous neutralizing antibodies in the cerebrospinal fluid of HIV-1 infected individuals. AIDS Res Hum Retroviruses 8:1133-8.

57. Wei, X., J. M. Decker, S. Wang, H. Hui, J. C. Kappes, X. Wu, J. F. Salazar-Gonzalez, M. G. Salazar, J. M. Kilby, M. S. Saag, N. L. Komarova, M. A. Nowak, B. H. Hahn, P. D. Kwong, and G. M. Shaw. 2003. Antibody neutralization and escape by HIV-1. Nature 422:307-312.

58. Witten, I. H. and E. Frank. 2000. Data mining practical machine learning tools and techniques with java implementations. Morgan Kaufmann, San Francisco.

59. Wolinsky, S. M., B. T. Korber, A. U. Neumann, M. Daniels, K. J. Kunstman, A. J. Whetsell, M. R. Furtado, Y. Cao, D. D. Ho, and J. T. Safrit. 1996. Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. Science 272:537-42.

60. Wong, J. K., C. C. Ignacio, F. Torriani, D. Havlir, N. J. S. Fitch, and D. D. Richman. 1997. *In vivo* compartmentalization of HIV: evidence from the examination of *pol* sequences from autopsy tissues. J Virol 70:2059-2071.

61. Zhou, N., X. Zhang, X. Fan, E. Argyris, J. Fang, E. Acheampong, G. C. DuBois, and R. J. Pomerantz. 2003. The N-terminal domain of APJ, a CNS-based coreceptor for HIV-1, is essential for its receptor function and coreceptor activity. Virology 317:84-94.

62. Zink, M. C., K. Suryanarayana, J. L. Mankowski, A. Shen, M. Piatak, Jr., J. P. Spelman, D. L. Carter, R. J. Adams, J. D. Lifson, and J. E. Clements. 1999. High viral load in the cerebrospinal fluid and brain correlates with severity of simian immunodeficiency virus encephalitis. J Virol 73:10480-8.

FIGURES AND TABLES

Table 1: Sites within *env* C2-V3 under differential selective pressure in plasma and CSF

HXB2 env position	CSF (dN/dS)	Plasma (dN/dS)	Differential p-value	Transition type
249	0.0000/ 1.5313	0.1499/ 0.0000	0.0923	Negative->Neutral
250	0.0000/ 0.8244	0.1313/ 0.0000	0.0947	Negative->Neutral
251	0.2539/ 0.0000	0.0000/ 1.3152	0.0499	Neutral->Negative
255	0.0000/ 2.0299	0.4633/ 1.6224	0.0795	Negative->Neutral
286	0.0000/ 0.8520	1.1051/ 0.0000	0.0131	Negative->Positive
330	0.0000/ 3.2736	0.3629/ 0.9094	0.0694	Negative->Neutral
346	1.5486/ 0.0000	0.9555/ 0.4798	0.092	Positive->Neutral

Table 2: Global deficit scores (GDS) and consensus residues at position 5 of the V3 loop in CSF- and plasma-derived sequences from all individuals with neuropsychologic data. Individuals listed from highest to lowest cognitive deficit.

individual	p5-CSF	p5-plasma	GDS
R	S	S	3.5
S	H	N	2.88
B	S	S	2.13
E	S	S	1.41
O	S	N	1.4
A	N	N	1.28
P	N	N	1.18
Q	N	N	0.94
F	N	N	0.88
J	N	N	0.75
D	N	N	0.69
L	N	N	0.63
M	N	N	0.56
H	G	G	0.53
K	N	N	0.5
C	N	N	0.44
N	G	G	0.31

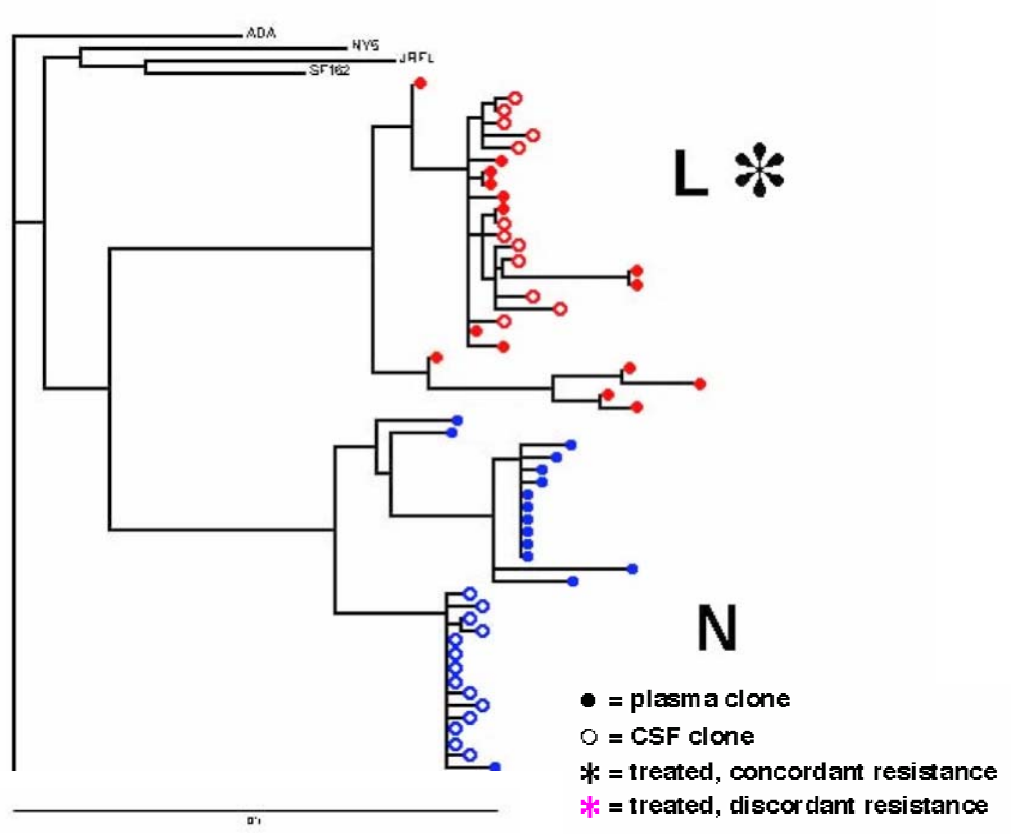


Figure 1: Examples of compartmentalized and noncompartmentalized viral populations. Maximum likelihood phylogenies of C2-V3 *env* sequences. Red circles = individual L (compartmentalized virus) and blue circles = individual N (noncompartmentalized virus). Open circles represent CSF sequences, and closed circles indicate plasma-derived sequences. Strains ADA, NY5, JRFL, and SF162 included as outgroups. Scale bar equals 10% genetic distance.

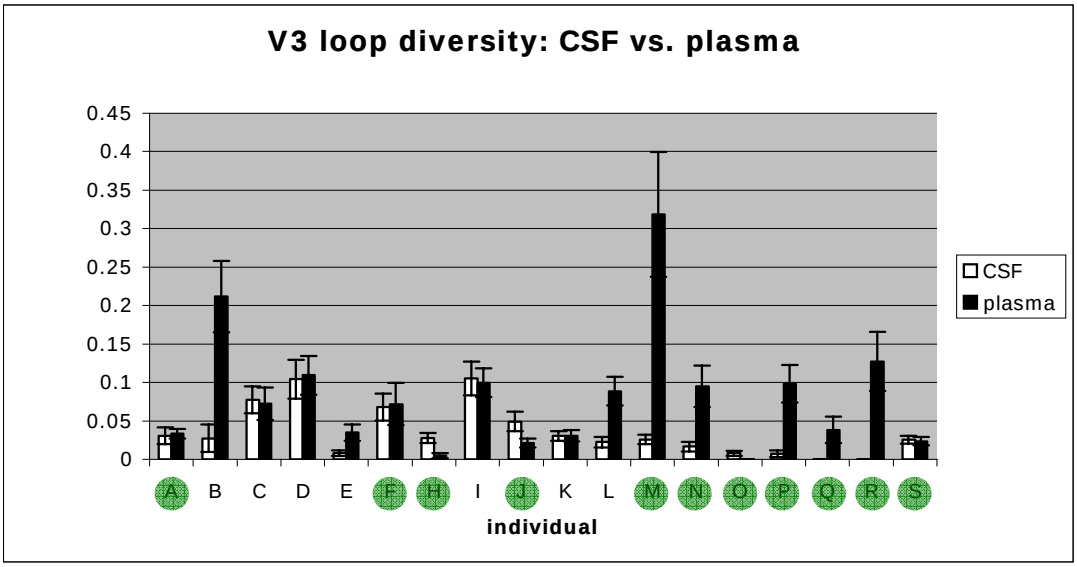


Figure 2: V3 amino acid diversity in CSF- and blood-derived viral populations. Diversity was significantly lower in CSF-derived viral populations across individuals ($P < 0.01$, Wilcoxon test). Green circles indicate compartmentalized individuals. Vertical bars represent standard error.

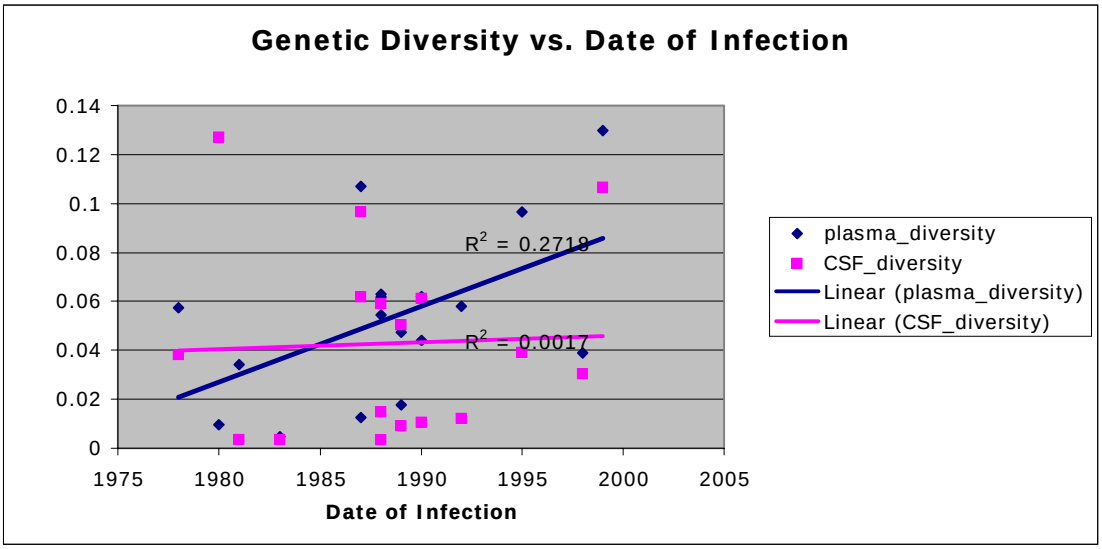


Figure 3. Genetic diversity vs. date of infection (as reported by infected individual). There is a positive correlation between infection date and plasma diversity (R^2=0.27), in accordance with observations that individuals with advanced disease tend to harbor relatively homogeneous viral populations in blood plasma. There is no correlation between infection date and CSF viral diversity.

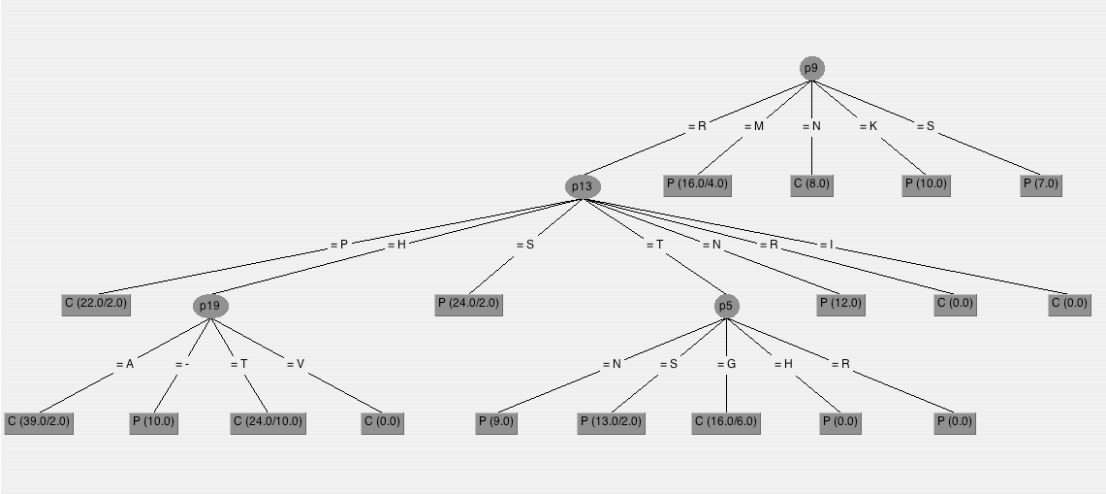
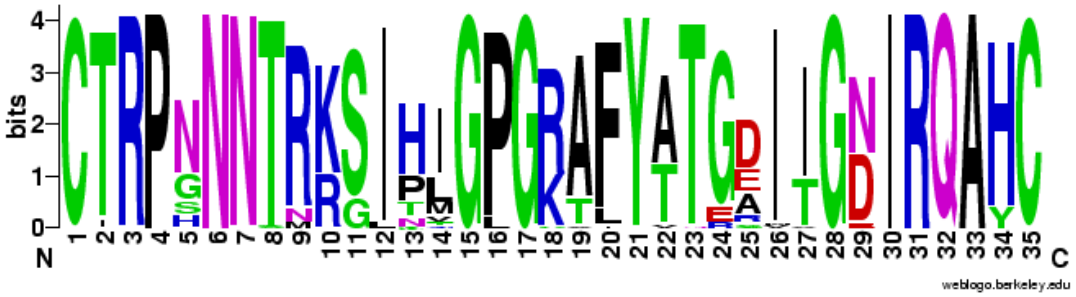


Figure 4. Genetic signature associated with CSF-derived sequences from compartmentalized individuals. Decision tree classifying V3 sequences based on tissue of origin with 87% accuracy. p, plasma classification; c, CSF. The values in parentheses are the number of instances/number of incorrect classifications. Residues are numbered starting from the cysteine at the beginning of the V3 loop.

a)



b)

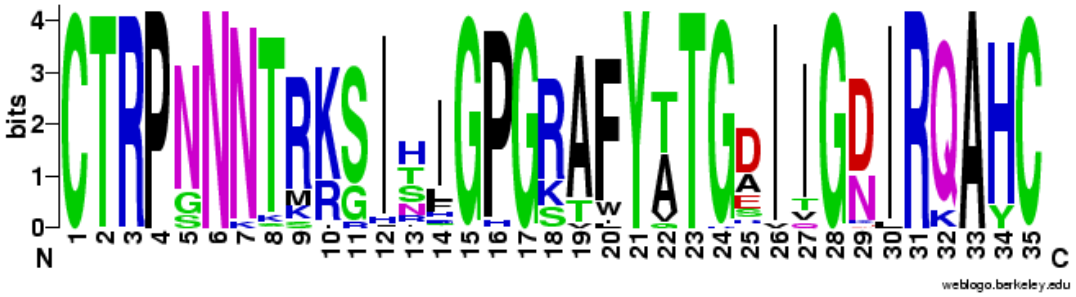


Figure 5. Consensus V3 loop sequences of a) CSF and b) plasma. The overall height of each position is proportional to its conservation. Relative height of each amino acid reflects its prevalence at that site. The prevalence of proline and histidine at position 13 are significantly higher in CSF sequences.

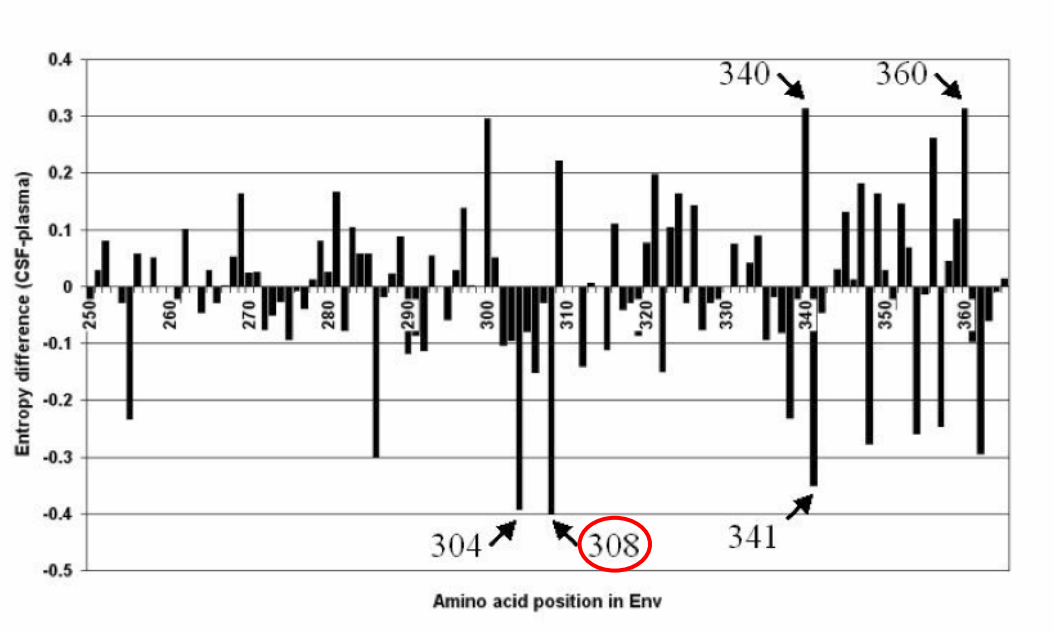


Figure 6: Difference in Shannon entropy between CSF and plasma at sites across the Env C2-V3 region. The five sites with the highest net differences are labeled with position numbers (numbered according to HXB2 gp160). Position 308 (V3 loop position 13), circled in red, exhibited the greatest net difference in entropy out of all C2-V3 amino acid sites.

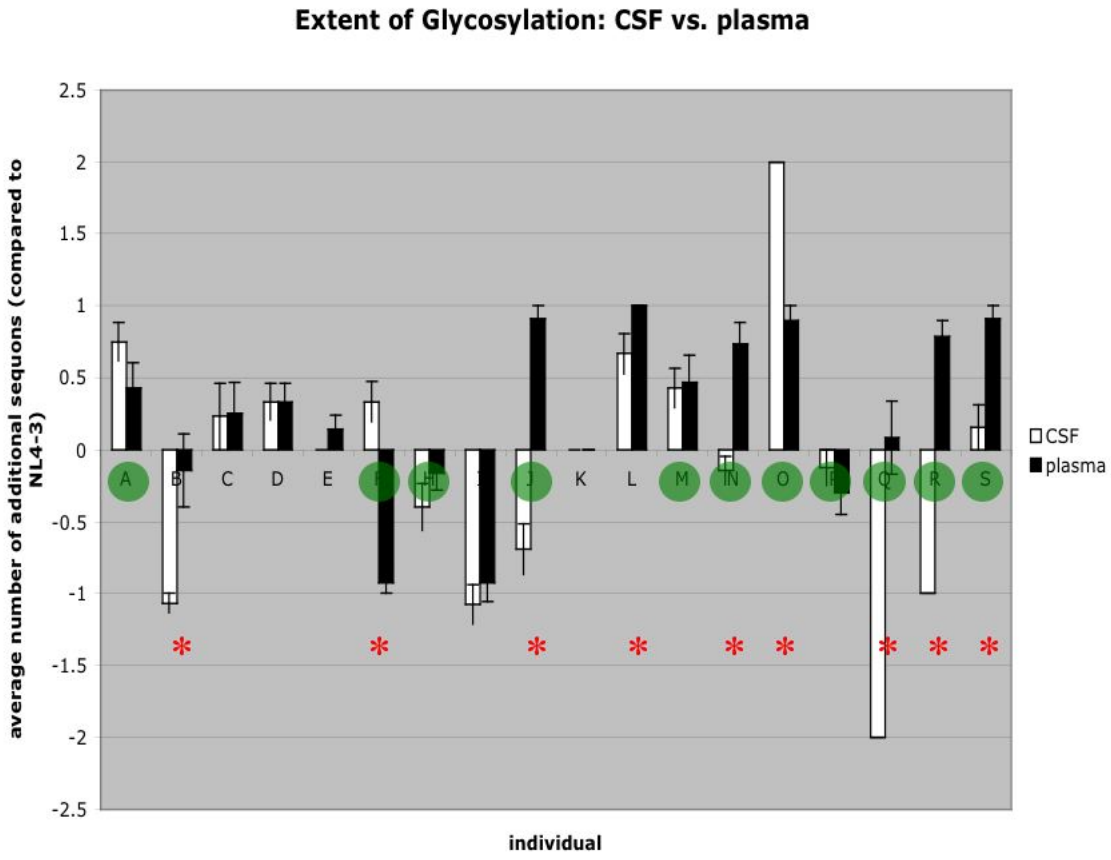


Figure 7. Extent of glycosylation (number of N-linked glycosylation sites) in CSF- and plasma-derived C2-V3 sequences. Green circles indicate compartmentalized individuals. Red asterisks indicate significan differences between compartments ($p<0.05$, Mann-Whitney). Vertical bars represent standard error. 7 out of 11 compartmentalized individuals have significantly different numbers of glycosylation sites in CSF- and plasma-derived sequences, and 5 out of those 7 have greater numbers in plasma populations.

Individual A

```
Consensus     CTRPNNNTRR SISIGPGRAF YATGAIIGNI RQAHC
A-01C         ---------- --P------- --I-----D- -----
A-04C         ---------- --H------- ---------- -----
A-05C         ---------- --H------- ---------- -----
A-07C         ---------- ---------- ---------- -----
A-08C         ---------- --H------- ---------- -----
A-09C         ---------- --H------- ---------- -----
A-10C         ---------- --H------- ---------- -----
A-11C         ---------- --H------- ---------- -----
A-12C         ---------- --H------- ---------- -----
A-13C         ---------- --H------- ---------- -----
A-14C         ----S----- ---------- --------D- -----
A-15C         ---------- --H------- ---------- -----
A-01P         ---------- ---------- ------T--- -----
A-02P         ---------- ---------- ------T-D- -----
A-03P         ---------- ---------- ------T--- -----
A-04P         ---------- ---------- ------T-D- -----
A-05P         ---------- ---------- ------T-D- -----
A-06P         ---------- --P------- ------T--- -----
A-08P         ---------- ---------- ---------- -----
A-09P         ---------- ---------- ---------- -----
A-10P         ---------- ---------- ---------- -----
A-11P         ---------- ---------- ---------- -----
A-12P         ---------- ---------- ---------- -----
A-13P         ---------- ---------- ---.G---D- -----
A-14P         ---------- ---------- ---.----D- -----
A-15P         ---------- ---------- ------T--- -----    35
```

Supplementary Figure 1. V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual F

```
Consensus     CTRPNNNTMK SIHLGPGRAF YTTGSIIGDI RQAYC
F-01C         ---------- ---------- ---------- -----
F-02C         ---------- ---------- ---------- -----
F-03C         --------R- ---------- ---------- -----
F-04C         ---------- ---------- ---------- -----
F-06C         ---------- ---------- ---------- -----
F-07C         --------R- -----L---- -A--D----- -----
F-08C         --------R- -----L---- -A--D----- -----
F-09C         --------R- -----L---- -A--D----- -----
F-11C         --------R- -----L---- -A--D----- -----
F-13C         --------R- -----L---- -A--D----- -----
F-14C         --------R- -----L---- -A--D----- -----
F-15C         --------R- -----L---- -A--D----- -----
F-02P         ---------- --RF---S-- ---------- -K---
F-04P         ---------- --RF---S-- ---------- -K---
F-05P         ---------- --RF---S-- ---------- -K---
F-06P         ---------- --RF---S-- --------N- -K---
F-07P         ---------- --RF---S-- ----I---N- -K---
F-08P         ---------- --RF---S-- ---------- -K---
F-09P         ---------- --RF---S-- ---------- -K---
F-10P         ---------- ---------- ---------- -----
F-11P         ---------- --RF---S-- ---------- -K---
F-12P         ---------- --RF---S-- ---------- -K---
F-13P         --------R- -----L---- -A--D----- -----
F-14P         ---------- --RF---S-- ---------- -K---
F-15P         ---------- --RF---S-- ---------- -K---    35
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual H

```
Consensus    CTRPGNNTRK SITIGPGRAF YATGDIIGDI RQAHC
H-01C        ---------- ---------- --------N- -----
H-02C        ---------- -------S-- --------N- -----
H-03C        ---------- ---------- --------N- -----
H-06C        ---------- ---------- ----E?--N- -----
H-08C        ---------- ---------- ---------- -----
H-09C        ---------- ---------- --------?- -----
H-12C        ---------- ---------- ---------- -----
H-13C        ---------- ---------- ---------- -----
H-14C        ---------- ---------- --------N- -----
H-15C        ---------- -------S-- --------N- -----
H-01P        ---------- ---------- ---------- -----
H-02P        ---------- ---------- -----?---- -----
H-04P        ---------- ---------- ---------- -----
H-05P        ---------- ---------- ---------- -----
H-06P        ---------- -V-------- ---------- -----
H-07P        ---------- ---------- ---------- -----
H-08P        ---------- ---------- ---------- -----
H-10P        ---------- ---------- ----.----- -----
H-11P        ---------- ---------- ---------- -----
H-13P        ---------- ---------- ---------- -----
H-14P        ---------- ---------- ---------- -----
H-15P        ---------- ---------- ---------- -----    35
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual J

```
Consensus    CTRPNNNTRK GIPMGPGKFY ATGEIIGNIR QAHC
J-01C        ---------- S--------F -QEQ------ ----
J-02C        ---------- S--------- ---A------ ----
J-03C        ---------- S--------- ---A------ ----
J-04C        ---------- S--------- ---A------ ----
J-05C        ---------- ---------- ---------- ----
J-06C        ---------- S--I------ ---------- ----
J-07C        ---------- S--------- ---A------ ----
J-09C        ---------- S--------- ---A------ ----
J-11C        ---------- ---------- ---------- ----
J-12C        ---------- S--------- ---A------ ----
J-13C        ---------- S--------- ---------- ----
J-14C        ---------- ---------- ---------- ----
J-15C        ---------- S--------- ---A------ ----
J-01P        ---------- --HIE----- ---------- ----
J-02P        ---------- --H------- ---------- ----
J-03P        ---------- --HI------ ---------- ----
J-04P        ---------- --HI---E-- ---------- ----
J-05P        -------I-- --HI------ ---------- ----
J-06P        ---------- --HI------ ---------- ----
J-08P        ---------- --HI------ ---------- ----
J-09P        ---------- --HI------ ---------- ----
J-10P        ---------- --HI------ ---------- ----
J-11P        ---------- --HI------ ---------- ----
J-12P        ---------- --HI------ ---------- ----34
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual M

```
Consensus     CTRPNNNTRR SIHIGPGKAF YTGDIIGNIR QAHC
M-01C         ---------- ---------- -------D-- ----
M-02C         --G------- -------R-- ---------- ----
M-03C         ---------- -------R-- ---------- ----
M-04C         ---------- ---------- ---------- ----
M-05C         ---------- ---------- ---------- ----
M-07C         ---------- -------R-- ---------- ----
M-08C         -A-------- ---------- ---------- ----
M-09C         ---------- ---------- ---------- ----
M-10C         ---------- ---------- ---------- ----
M-11C         ---------- ---------- ---------- ----
M-12C         ---------- ---------- ---------- ----
M-13C         ---------- -------R-- ---------- ----
M-14C         ---------- ---------- ---------- ----
M-15C         ---------- ---------- ---------- ----
M-01P         ---------- --T------- -------D-- ----
M-02P         ---------- --T------- -------D-- ----
M-03P         ------KKKI RHIH-H-RT- -----Q-KL- ----
M-04P         ---------- --T------- ---------- ----
M-05P         ---------- --T------- -------D-- ----
M-06P         ---------- --T------- -------D-- ----
M-07P         ---------- --T------- -------D-- ----
M-08P         ------KKKI RHIH-H-RT- -----Q-KL- ----
M-09P         ---------- -------R-- ---------- ----
M-10P         ------KKKI RHIH-H-RT- -----Q-KL- ----
M-11P         ---------- --T------- -------D-- ----
M-12P         ---------- --T------- -------D-- ----
M-13P         ------KKKI RHIH-H-RT- -----Q-KL- ----
M-14P         ---------- --T------- -------D-- ----
M-15P         ------KKKI RHIH-H-RT- -----Q-EL- ----34
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual N

```
Consensus     CTRPGNNTRK GIHLGPGRTF YATGEITGDI RQAHC
N-C01         ---------- ---------- ---------- -----
N-C02         ---------- ---------- ---------- -----
N-C03         ---------- ---------- ---------- ----R
N-C05         ---------- ---------- ---------- -----
N-C06         ---------- ---------- ---------- -----
N-C07         ---------- ---------- ---------- -----
N-C08         ---------- ---------- ---------- -----
N-C09         ---------- ---------- ---------- -----
N-C10         ---------- ---------- ---------- -----
N-C11         ---------- ---------- -----M---- -----
N-C12         ---------- ---------- ---------- -----
N-C13         ---------- ---------- ---------- -----
N-C14         ---------E ---------- -----M---- -----
N-C15         ---------- ---------- ---------- -----
N-P01         --------S- S-------A- ----R-I-N- -----
N-P02         ---------- S-N-----A- ----N-I--- ---Y-
N-P03         ---------- S-N-----A- ----N-I--- -R-Y-
N-P04         --------S- S-------A- ---------- -----
N-P05         --------S- S-------A- ----R-I-N- -----
N-P06         --G------- ---------- ---------- -----
N-P07         --------S- S-------A- ----R-I-N- -----
N-P08         --------S- S-------A- ----R-I-N- -----
N-P09         --------S- S-------A- ----R-I-A- -----
N-P10         --------S- S-------A- ----R-I-N- -----
N-P11         --------S- S-------A- ----R-I-N- -----
N-P12         --------S- S-------A- ----R-I-N- -----
N-P13         --------S- S-------A- ----R-I-N- -----
N-P14         --------S- S-------A- ----R-I-N- -----
N-P15         --------S- S-------A- ----R-I-N- -----    35
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual O

```
Consensus     CTRPSNNTNR SINIGPGRAF YATERITGDI RQAHC
O-C01         ---------- ---------- ---------- -----
O-C02         ---------- ---------- ---------- -----
O-C03         ---------- ---------- ---------- -----
O-C04         ---------- ---------- ---------- -----
O-C05         ---------- ---------- ---------- -----
O-C06         ---------- ---------- ---------- -----
O-C07         ---------- ---------- ---------- -----
O-C09         ---------- ---------- ---------- -----
O-C10         ---------- ---------- ---------- -----
O-C11         ------D--- ---------- ---------- -----
O-C12         ---------- ---------- ---------- -----
O-C13         R--------- ---------- ---------- -----
O-C14         ---------- ---------- ---------- -----
O-C15         ---------- ---------- ---------- -----
O-P02         ----N---K- ---------W -G-G.-I--- -----
O-P03         ----N---K- ---------W -G-G.-I--- -----
O-P04         ----N---K- ---------W -G-G.-I--- -----
O-P05         ----N---K- ---------W -G-G.-I--- -----
O-P06         ----N---K- ---------W -G-G.-I--- -----
O-P07         ----N---K- ---------W -G-G.-I--- -----
O-P08         ----N---K- ---------W -G-G.-I--- -----
O-P09         ----N---K- ---------W -G-G.-I--- -----
O-P10         ----N---K- ---------W -G-G.-I--- -----
O-P12         ----N---K- ---------W -G-G.-I--- -----    35
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual P

```
Consensus     CTRPNNNTRK GLHTGPGRTL YVTRAIIGDI RQAHC
P-C03         ---------- ---------- ---------- -----
P-C04         ---------- ---------- ---------- -----
P-C05         ---------- ---------- ---------- -----
P-C07         ---------- ---------- ---------- -----
P-C08         ---------- ---------- ---------- -----
P-C10         ---------- ---------- ---------- -----
P-C11         ---------R ---------- ---------- -----
P-C15         ---------- ---------- ---------- -----
P-P03         ---------- -I-I---S-W ---G------ -----
P-P04         ---------- ---------- ---------- -----
P-P06         ---------- SI-------- ---GD----- -----
P-P07         ---------- -I-I---S-W ---G------ -----
P-P08         ---------- -I-I---S-W ---G------ -----
P-P09         ---------- SI-------- ---GD----- -----
P-P10         ---------- -I-I---S-W ---G------ -----
P-P12         ---------- -I-I---S-W ---G------ -----
P-P13         ---------- -I-I---S-W ---G------ -----
P-P15         ---------- ---------- ---------- -----    35
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual Q

```
Consensus     CIRPNNNTRK SIPVGPGKAL YTTGEIIGEI RQAHC
Q-C01         ---------- ---------- ---------- -----
Q-C02         ---------- ---------- ---------- -----
Q-C03         ---------- ---------- ---------- -----
Q-C04         ---------- ---------- ---------- -----
Q-C05         ---------- ---------- ---------- -----
Q-C06         ---------- ---------- ---------- -----
Q-C07         ---------- ---------- ---------- -----
Q-C08         ---------- ---------- ---------- -----
Q-C09         ---------- ---------- ---------- -----
Q-C10         ---------- ---------- ---------- -----
Q-C11         ---------- ---------- ---------- -----
Q-C12         ---------- ---------- ---------- -----
Q-C13         ---------- ---------- ---------- -----
Q-C14         ---------- ---------- ---------- -----
Q-C15         ---------- ---------- ---------- -----
Q-P01         -T-------- --SI---R-F ----D---D- -----
Q-P02         -T-------- --SI---R-F ----D---D- -----
Q-P03         -T-------- --SI---R-F ----D---D- -----
Q-P05         -T-------- --SI---R-F ----D---D- -----
Q-P06         -T-------- --SI---R-F ----D---D- -----
Q-P08         -T-------- --SI---R-F ----D---D- -----
Q-P09         -T-------- --SI---R-F ----D---D- -----
Q-P10         -T-------- --SI---R-F --------D- -----
Q-P12         -T-------- --SI---R-F ----D---D- -----
Q-P13         -T-------- --SI---R-F --------D- -----
Q-P14         -T-------- --SI---R-F --------D- -----
Q-P15         ---------- ---------- ---------- -----    35
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual R

```
Consensus     CTRPSNNTRK GITIGPGRAF YATGDIIGNI RQAHC
R-C01         -------S-- S--R----VI -T---VV-D- -----
R-C03         -------S-- S--R----VI -T---VV-D- -----
R-C04         -------S-- S--R----VI -T---VV-D- -----
R-C09         -------S-- S--R----VI -T---VV-D- -----
R-C13         -------S-- S--R----VI -T---VV-D- -----
R-P01         ---------- ---------- ---------- -----
R-P02         ---------- ---------- ---------- -----
R-P03         ---------- ---------- ---------- -----
R-P04         ---------- ---------- ---------- -----
R-P05         ---------- S--R----VI -T---VV-D- -----
R-P06         ---------- ---------- ---------- -----
R-P07         ---------- ---------- ---------- -----
R-P09         -------S-- S--R----VI -T---VV-D- -----
R-P10         -------S-- S--R----VI -T---VV-D- -----
R-P11         ---------- ---------- ---------- -----
R-P12         ---------- ---------- ---------- -----
R-P13         ---------- ---------- ---------- -----
R-P14         -------S-- S--R----VI -T---VV-D- -----
R-P15         ---------- ---------- ---------- -----    35
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Individual S

```
Consensus     CTRPHNNTRK SINIGPGRAF YTTGDITGNI RQAHC
S-01C         ---------- --P------- ---------- -----
S-02C         ---------- --H------- ---------- -----
S-03C         ---------- --H----K-- ---------- -----
S-04C         ---------- --H------- ---------- -----
S-05C         ---------- --H------- ---------- -----
S-06C         ---------- --H----K-- ---------- -----
S-07C         ---------- --H----K-- ---------- -----
S-09C         ----R----- --H------- ---------- -----
S-10C         ---------- --H------- ---------- -----
S-11C         ---------- --H------- ---------- -----
S-12C         ---------- --H------- ---------- -----
S-14C         ---------- --P------- ---------- -----
S-15C         ---------- --H----K-- ---------- -----
S-02P         ----N----- ---------- ------V-D- -----
S-03P         ----N----- ---------- ------I-D- -----
S-04P         ----N----- ---------- ------I-D- -----
S-05P         ----N----- ---------- ------V-D- -----
S-06P         ----N----- ---------- ------V-D- -----
S-08P         ----N----- ---------- ------I-D- ----Y
S-09P         ----N----- ---------- ----N-I-D- -----
S-11P         ----N----- ---------- ------I-D- -----
S-12P         ----N----- ---------- ------I-D- -----
S-13P         ----N----- ---------- ------V-D- -----
S-14P         ----N----- ---------- ------I-D- -----    35
```

Supplementary Figure 1 (cont'd). V3 loop amino acid alignments of CSF- and plasma-derived sequences from individuals with compartmentalized virus (A, F, H, J, M-S). The first sequence in each alignment is a consensus of all available clones, dashes represent identity, and dots symbolize gaps (deletions). Sequence names contain clone numbers and tissue of origin ("C" = CSF, "P" = plasma). Compartment-specific consensus residues differed most frequently at position 13.

Chapter 6

Genotypic and Phenotypic Differences between CSF- and Plasma-Derived HIV-1 Nef Proteins

ABSTRACT

One of the primary functions of the HIV-1 Nef protein is downregulation of MHC class I expression at the host cell surface, as a means of avoiding destruction by cytotoxic T lymphocytes (CTL). The central nervous system (CNS) is an immune privileged site, and may harbor fewer CTL than peripheral tissues. HIV replicating within the CNS may therefore eventually lose the ability to downregulate MHC-I, due to a lack of selective benefit associated with the phenotype. We compared the sequences and phenotypes of Nef proteins obtained from the cerebrospinal fluid (CSF) and plasma of three chronically infected donors. Nef sequences were compartmentalized in all three individuals based on phylogenetic evidence and population-level genotyping. Nef function differed between anatomic compartments in one out of three donors; the extent of MHC downregulation conferred by CSF-derived Nef variants was considerably reduced at both available sample time points. Our results suggest that anatomic compartmentalization leads to tissue-specific evolution of the *nef* gene, and functional differences between CNS- and plasma-derived viral gene products may be observed in an *in vitro* subgenomic context.

INTRODUCTION

The 27 kDa HIV-1 Nef protein (Fig. 1) modulates the expression and trafficking of numerous host proteins within infected cells, including CD4 and major histocompatibility (MHC) class I and II molecules (4). MHC class I molecules recruit viral antigens to the host cell surface and display them to CD8+ cytotoxic T

lymphocytes (CTL), which subsequently destroy the infected cell via granzyme-mediated dissolution of the plasma membrane or Fas-mediated apoptosis (Fig. 2). HIV-1 downregulates the expression of MHC class I at the cell surface, thereby reducing the probability that the host cell will be destroyed by CD8+ cytolytic activity (2,5).

The presence of HIV-1 has been detected in several anatomic sites within infected individuals, including brain, blood, lung, lymph nodes, spleen, and genital tract. There is likely to be variation in the nature and extent of immunological surveillance between these tissues. The central nervous system (CNS), for instance, is an immune privileged site due to the selective permeability of the blood-brain barrier, and is believed to contain fewer circulating lymphocytes than peripheral tissues (10). Therefore, the selective benefit of MHC class I downregulation to HIV-1 replicative fitness may be minimal or nonexistent within the CNS. A loss of selective benefit is likely to result in loss of phenotype after several replicative cycles, due to the erosionary influence of the sloppy viral polymerase (reverse transcriptase) on the viral genome. In addition, selective tradeoffs may exist between Nef functions, whereby another phenotype (e.g. CD4 downregulation) may be enhanced at the cost of MHC-I downregulation (3). We investigated this possibility by comparing the genotypes and phenotypes of *nef* alleles associated with cerebrospinal fluid- and blood plasma-derived viruses from three chronically infected donors. Phenotypes were compared using a high-throughput assay involving linear PCR-assembled subgenomic

expression constructs, recently used by Ali, Pillai, Ng *et al* to evaluate the selective benefits of Nef-induced MHC-I donwregulation in an *in vitro* setting (2).

MATERIALS AND METHODS

Subjects, specimen processing, and nucleotide sequencing. Patient cohort selection criteria, specimen processing, sequencing methods, and viral load quantitation have previously been described in full (14). CSF and blood samples from a total of three individuals, hereby referred to as "A", "B", and "C" were involved in this study. Clonal sequencing was performed on virus obtained from individuals A and B, and population (bulk) sequencing was performed on virus from individual C.

Phylogenetic reconstruction. Initial multiple sequence alignments were generated by using Multalin (6), with default gap parameters and the DNA 5-0 substitution matrix. Subsequent manual aligning was performed by using the Se-Al sequence alignment editor (12). A phylogenetic tree describing sequences from individuals A and B was constructed by implementing dnadist and neighbor within the PHYLIP version 3.5c software package (7), using the F84 model, gamma distributed rates across sites, and a transition/transversion ratio of 2.0. Trees were viewed with TreeView X (11).

Evaluation of MHC class I downregulation. Expression vectors expressing Nef variants were constructed using the TAP Express Fragment System (Gene Therapy Systems, San Diego, CA) according to the manufacturer's protocol. Two-step recombinant PCR was used to link the *nef* alleles from the limiting dilution sequencing reactions to CMV promoter and terminator sequences (Fig. 3). Primer

sequences (TAP custom oligos) for the first step were as follows: TapNef-L (5') CTgCAggCACCgTCgTCgACTTAACAACCTABAAgAATAAgACAg and TapNef-R (3') CATCAATgTATCTTATCATgTCTgACCAgCggAAAgTCCCTTgTA. These vectors (2 □ug each) were then colipofected (GenePORTER, Gene Therapy Systems) with the green fluorescence protein (GFP)-expressing vector phGFP-S65T (Clontech, Palo Alto, CA) into HEK-293 cells. MHC-I expression was determined by flow cytometric analysis of GFP-expressing cells 48 h after lipofection using a pan MHC class I (A, B, and C) antibody (Pharmingen, San Diego, CA). All experiments were performed in duplicate. Downregulation of MHC-I by wildtype NL4-3 was calculated by comparison to a *nef* negative control containing two premature stop codons. Downregulation by CSF- and blood-derived Nef variants was normalized against NL4-3 phenotype and reported as percentage of wildtype activity (Fig. 4).

Accession numbers. Genbank accession numbers for the 18 sequences involved in this analysis will be reported upon submission and acceptance of this manuscript.

RESULTS AND DISCUSSION

A total of sixteen clonal sequences were generated representing CSF- and plasma-derived *nef* sequences from individuals A and B (4 clones per sample tissue). Phylogenetic analysis of these data revealed that *nef* sequences were compartmentalized; tissue-specific populations formed distinct, independent clusters (Fig. 5). We then generated population sequences representing CSF and plasma virus from individual C, using two paired CSF and plasma samples isolated in June, 1995

and June, 2000. Individual C was chosen based on the availability of longitudinal samples and clinical characteristics that reflected a selective environment in which Nef phenotype was likely to evolve (3). CSF viral load was relatively high (approaching 50,000 copies/ml) at the first time point, suggesting that virus was actively replicating within the CNS. More importantly, CD4+ T cell numbers dropped from 753 cells/ml at the initial sample time to 68 at the second (Table 1). AIDS is defined by a CD4+ measurement of <200 cells/ml (8). The near lack of immune surveillance surrounding the second sample time would likely select for Nef functions that directly accelerate replication kinetics (e.g. CD4 downregulation, enhancement of infectivity) at the cost of MHC-I downregulation (3). This sample set would allow us to investigate the effects of both anatomic compartmentalization and temporal variation in host immune function on viral evolution. Nef sequences from individual C did in fact demonstrate considerable variation between tissues and between sample times (Fig. 6). A few of the mutations distinguishing CSF and plasma viruses were immediately adjacent to sequence domains governing MHC class I downregulation, suggestive of phenotypic differences. The twin arginines at positions 21 and 22 of the plasma Nef consensus sequence were substituted with lysines in the CSF consensus. The methionine at position 20 is critical for MHC class I downregulation, based on mutagenesis experiments performed by Akari *et al* (1). In addition, the glutamine at position 82 of the plasma sequence was replaced with lysine in the CSF-derived sequence. Position 82 is a few residues downstream of the polyproline region ("PxxP" domain), which is involved in class I downregulation as well (13). It is worth mentioning that in both of

these cases, it is the CSF-derived sequence rather than the plasma variant that contains the same residues as the subtype B consensus sequence.

We benchmarked our flow cytometry-based MHC-I expression assay by characterizing canonical Nef mutants with previously established phenotypes prior to evaluating variants from individuals A-C (13). Constructs expressing three mutants, "LL/AA", "PxxP", and "Delta 62-68" were transfected into 293 (human embryonic kidney) cells, and surface expression of class I was determined 48 hours post transfection via flow cytometry. Results in Table 2 indicate that our assay produced results that were in alignment with earlier reports, and that our subgenomic Nef expression system was a reliable proxy for Nef phenotype observed within the context of the whole viral genome.

We next measured the extent of class I downregulation conferred by Nef variants from CSF and plasma of individuals A-C. All of the Nef variants from individuals A and B failed to show any significant variation in function, ranging from 95 to 102% of NL4-3 phenotype. CSF variants from individual C, however, exhibited 10% attenuation in phenotype with respect to contemporaneous plasma samples. Although this dynamic range is comparable to the observed phenotypic differences between rapid progressor and non-progressor alleles catalogued by Carl *et al* (3), it is difficult to assess the true significance of this discrepancy. The extent to which natural selection can act on each Nef function independently is unknown. Longitudinal variation in either tissue was negligible, dropping from 84% to 81% of wildtype and 95% to 91% of wildtype in CSF and plasma, respectively (Table 3).

Our results demonstrate that compartmentalization of HIV-1 within the central nervous system may select for mutations within the Nef sequence that optimize its function for viral replication in the local fitness landscape. Although sequences were compartmentalized in the cases of individuals A and B, no significant differences were observed with respect to class I downregulation phenotype. The differences in sequence may affect one or more of Nef's countless other phenotypes, or perhaps may simply be attributed to genetic drift resulting from replicative isolation. A third possibility exists as well; although it is convenient for laboratory manipulation for a multitude of reasons, the HEK-293 cell is not always a reliable facsimile of a legitimate primary human host cell. Le Gall *et al* have reported that there are features associated with the MHC class I recycling machinery that differ substantially between lymphocytes and non-lymphocytes (9). Our assay, therefore, may not be sensitive to phenotypic changes that would be observable in a cell type that more closely resembled a natural target cell. We have devised two additional versions of this assay that measure phenotype within the Jurkat T cell line (Supplementary Fig. 1) and within freshly isolated, non-activated CD4+ peripheral blood mononuclear cells (Supplementary Fig. 2). It remains to be seen if future application of these more physiologically relevant systems will yield results that conflict with what we have reported here.

The data associated with individual C are quite provocative. The attenuation of class I phenotype in CSF-derived virus fits in nicely with the paradigm that immune surveillance is reduced in the CNS, and moreover, that HIV-1 adapts specifically to

microenvironments and anatomic compartments within the human body. The observed longitudinal trend (attenuation of phenotype in both compartments over time) is in alignment with the recorded drop in CD4+ cell counts over the sample period; the collapse of immune function within the host would neutralize the selective consequence of MHC class I downregulation (3,5). However, the longitudinal attenuation (4-5%) is minimal, and barely exceeds the noise in our assay system. This relative preservation of class I downregulation activity by Nef variants in both tissue compartments across the five-year sample period is in conflict with the longitudinal modulation in phenotype during the progression to AIDS observed by Carl and colleagues (3). Our data suggest that selective pressure from CD8+ cytolytic activity may persist despite low CD4+ counts, and moreover, Nef functions may not be entirely distinct and separable at the sequence level.

Virus from several additional individuals must be sampled to determine if there are any statistically significant patterns associated with HIV-1 Nef evolution within the central nervous system. For now, we have provided an initial glimpse at this phenomenon, demonstrating that CNS compartmentalization may result in tissue-specific Nef genotypic and phenotypic characteristics.

ACKNOWLEDGMENTS

I am grateful to Drs. John Guatelli, Joseph Wong, Scott Letendre, and Ron Ellis for their mentorship and provision of clinical samples. This work was supported by grants AI27670, AI043638, the Adult AIDS Clinical Trials Group funded by the National Institute of Allergy and Infectious Diseases, and the AACTG Central Group

Grant (U01AI38858), the UCSD Center for AIDS Research (AI 36214), AI29164, AI047745, from the National Institutes of Health, and the Research Center for AIDS and HIV Infection of the San Diego Veterans Affairs Healthcare System.

REFERENCES

1. Akari H., S. Arold, T. Fukumori, T. Okazaki, K. Strebel, and A. Adachi. 2000. Nef-induced major histocompatibility complex class I down-regulation is functionally dissociated from its virion incorporation, enhancement of viral infectivity, and CD4 down-regulation. J Virol 74:2907-12.

2. Ali, A., S. K. Pillai, H. Ng, R. Lubong, D. D. Richman, B. D. Jamieson, Y. Ding, M. J. McElrath, J .C. Guatelli, and O. O. Yang. 2003. Broadly increased sensitivity to cytotoxic T lymphocytes resulting from Nef epitope escape mutations. J Immunol 171:3999-4005.

3. Carl, S., T. C. Greenough, M. Krumbiegel, M. Greenberg, J. Skowronski, J. L. Sullivan, and F. Kirchhoff. 2001. Modulation of different human immunodeficiency virus type 1 Nef functions during progression to AIDS. J Virol 75:3657-65.

4. Coleman, S. H., J. R. Day, and J. C. Guatelli. 2001. The HIV-1 Nef protein as a target for antiretroviral therapy. Expert Opin Ther Targets 5:1-22.

5. Collins K. L., B. K. Chen, S. A. Kalams, B. D. Walker, and D. Baltimore. 1998. HIV-1 Nef protein protects infected primary cells against killing by cytotoxic T lymphocytes. Nature 391:397-401.

6. Corpet, F. 1988. Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16:10881-10890.

7. Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c.

8. From the Centers for Disease Control and Prevention. 1993. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. JAMA 269:729-30.

9. Le Gall S., F. Buseyne, A. Trocha, B. D. Walker, J. M. Heard, and O. Schwartz. 2000. Distinct trafficking pathways mediate Nef-induced and clathrin-dependent major histocompatibility complex class I down-regulation. J Virol 74:9256-66.

10. Pachter J. S., H. E. de Vries, and Z. Fabry. 2003. The blood-brain barrier and its role in immune privilege in the central nervous system. J Neuropathol Exp Neurol 62:593-604.

11. Page, R. D. M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comput Appl Biosci 12:357-358

12. Rambaut A. 2002. Se-Al sequence alignment editor v2.0 (Software). Oxford: Department of Zoology, University of Oxford.

13. Riggs N. L., H. M. Craig, M .W. Pandori, and J. C. Guatelli. 1999. The dileucine-based sorting motif in HIV-1 Nef is not required for down-regulation of class I MHC. Virology 258:203-7.

14. Strain M. C., S. Letendre, S. Pillai, T. Russell, C. C. Ignacio, H. F. Gunthard , B. Good, D. M. Smith, S. M. Wolinsky, M. Furtado, J. Marquie-Beck, J. Durelle, I. Grant, D. D. Richman, T. Marcotte, J. A. McCutchan, R. J. Ellis, and J. K. Wong. 2005. Genetic composition of HIV-1 in CSF and plasma without treatment and during failing combination antiretroviral therapy. J Virol 79:1772-88.

FIGURES AND TABLES

Table 1: Individual C: CD4+ T cell counts and compartment-specific viral load estimates at sample isolation dates.

sample date	CD4+ count	plasma VL (copies/ml)	CSF VL (copies/ml)	disease stage
Jun-95	753	20716	46080	asymptomatic
Jun-00	68	237720	1711	AIDS

Table 2: Benchmarking of MHC class I downregulation assay using canonical Nef mutants with previously determined phenotypes (11).

Nef variant	previously reported phenotype	TAP-Nef results (% of WT downregulation)
LL/AA	similar to wildtype	90%
PxxP	partially impaired	81%
delta 62-68	significantly impaired	45%

Table 3: Individual C: extent of MHC class I downregulation by CSF- and plasma-derived Nef variants at two time points (% of wildtype activity)

sample date	plasma phenotype	CSF phenotype
Jun-95	95%	84%
Jun-00	91%	81%

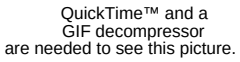


Figure 1: Structure of HIV-1 Nef protein (starting at W57)

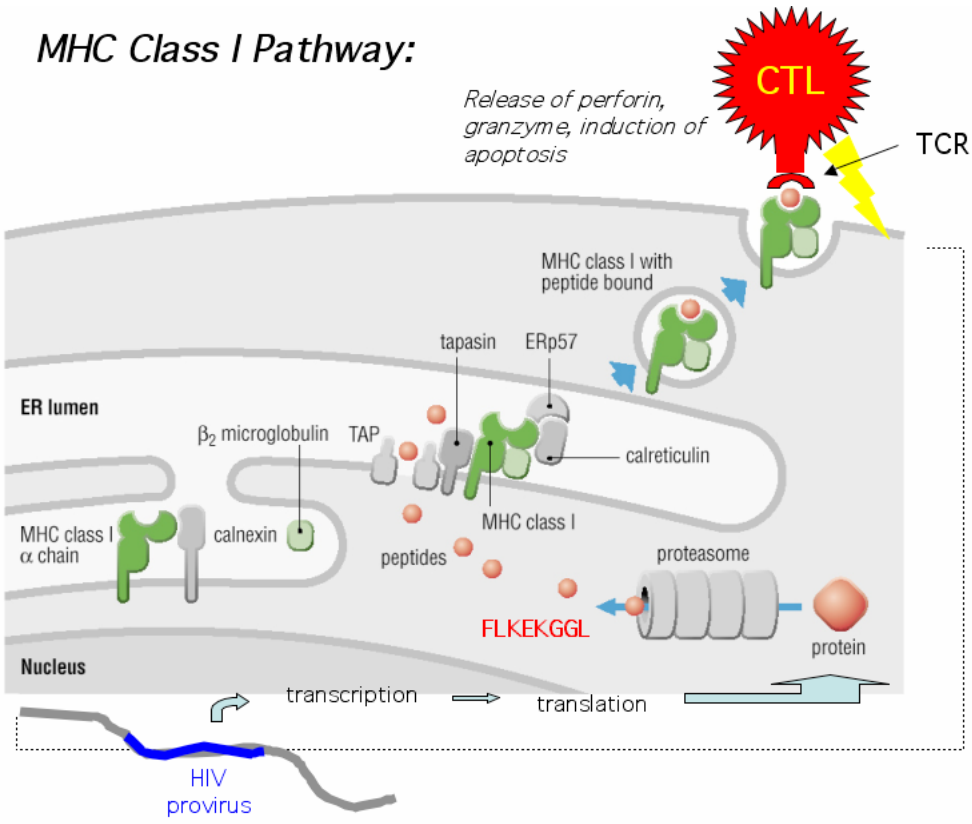


Figure 2: Schematic of the MHC class I antigen presentation pathway and induction of cytolytic activity by CD8+ T cells (CTL) resulting from the processing and display of HIV peptides (adapted from Janeway *et al*, 2001).

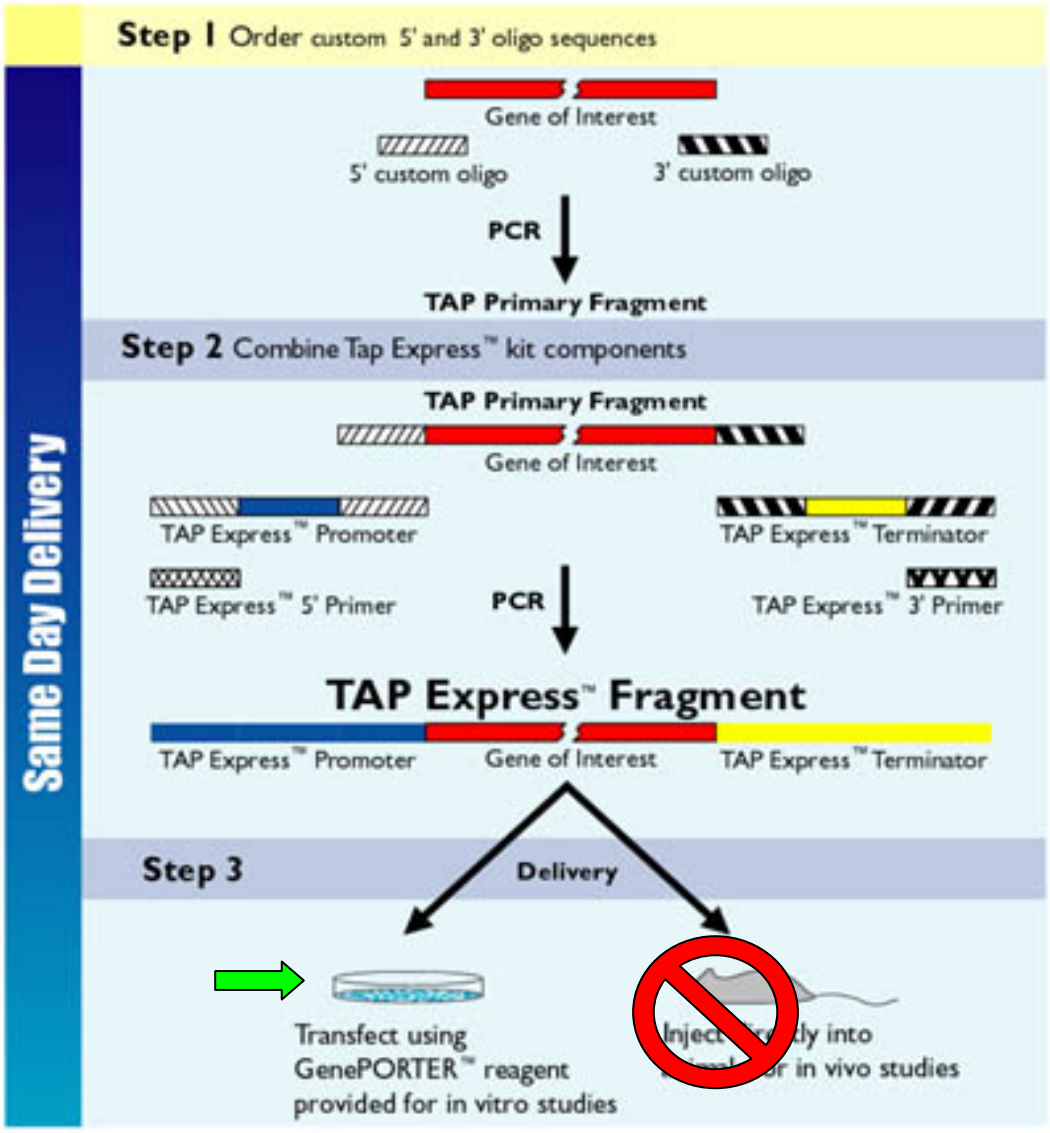


Figure 3: Breakdown of TAP (Transcriptionally Active PCR fragment) expression system (courtesy of Gene Therapy Systems, inc.).

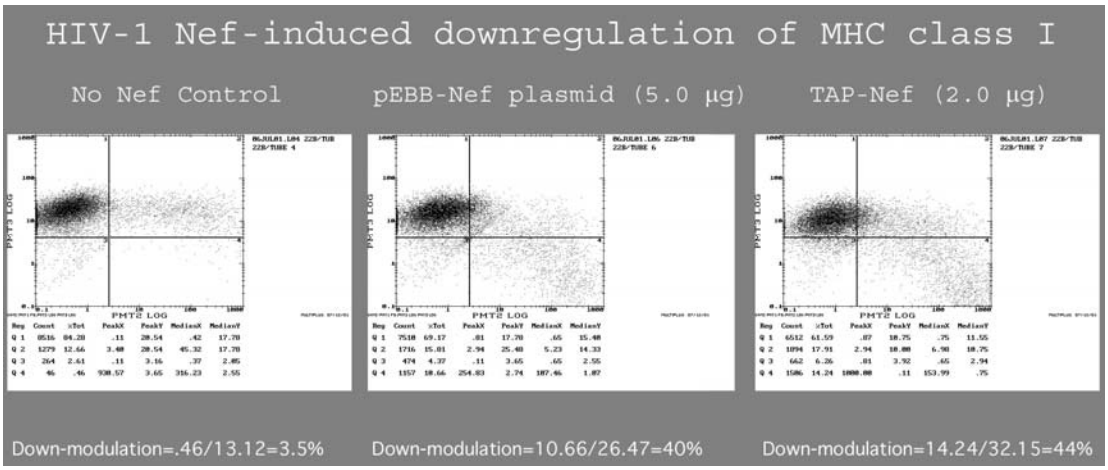


Figure 4: Flow cytometric analysis of MHC class I expression in 293 cells transfected with a GFP co-transfection marker and (a) mock DNA control, (b) a conventional CMV-driven plasmid (pEBB) expressing Nef, and (c) TAP-Nef. X-axis represents GFP expression, Y-axis represents MHC class I expression.

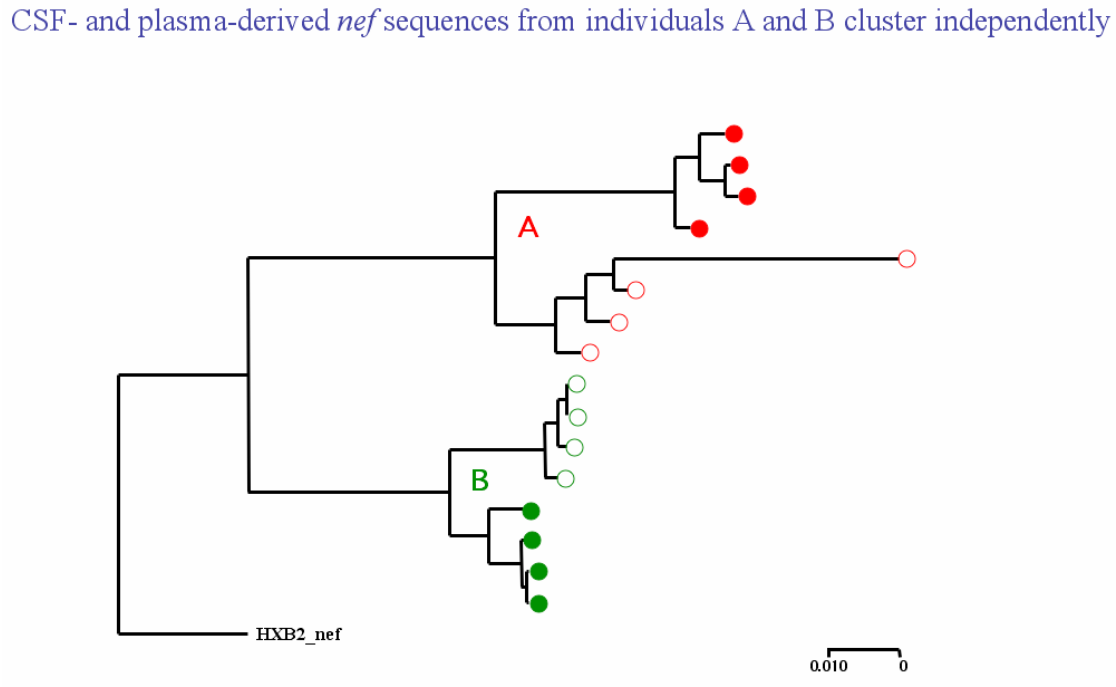


Figure 5: Maximum likelihood phylogeny of *nef* sequences from individuals A and B. Red circles = individual A and green circles = individual B. Open circles represent CSF sequences, and closed circles indicate plasma-derived sequences. HXB2 strain included as outgroup. Scale bar equals 10% genetic distance.

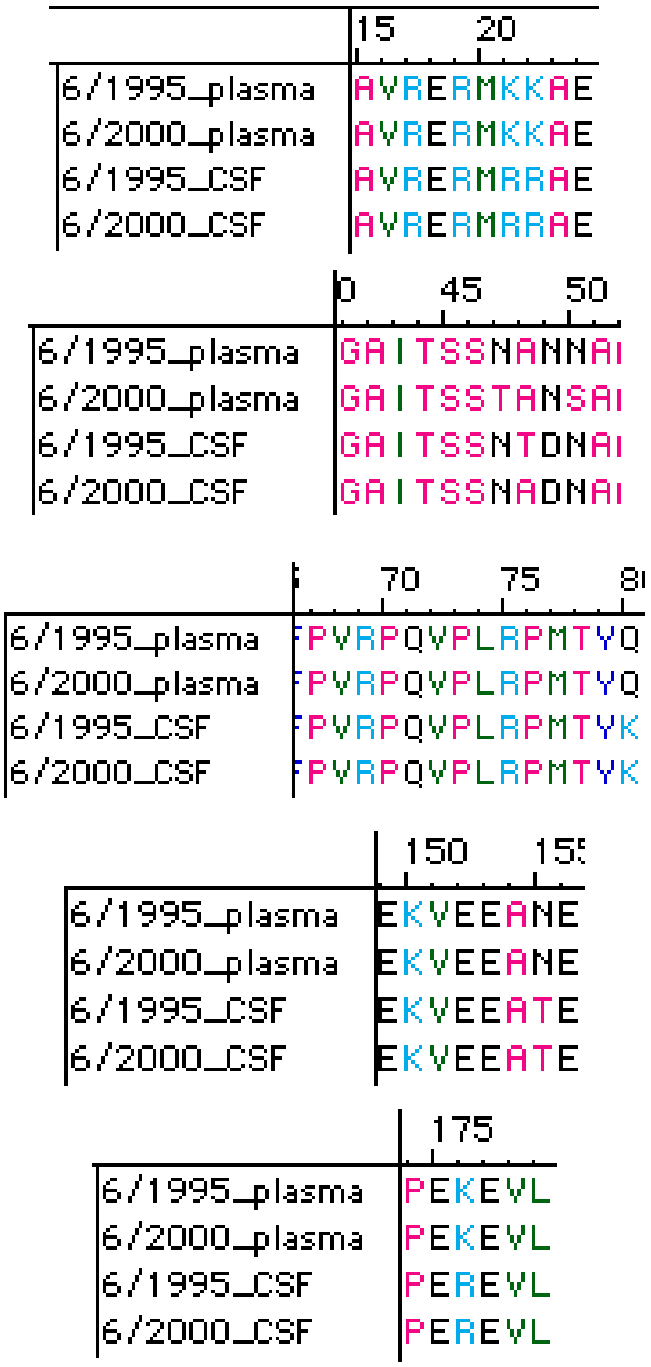
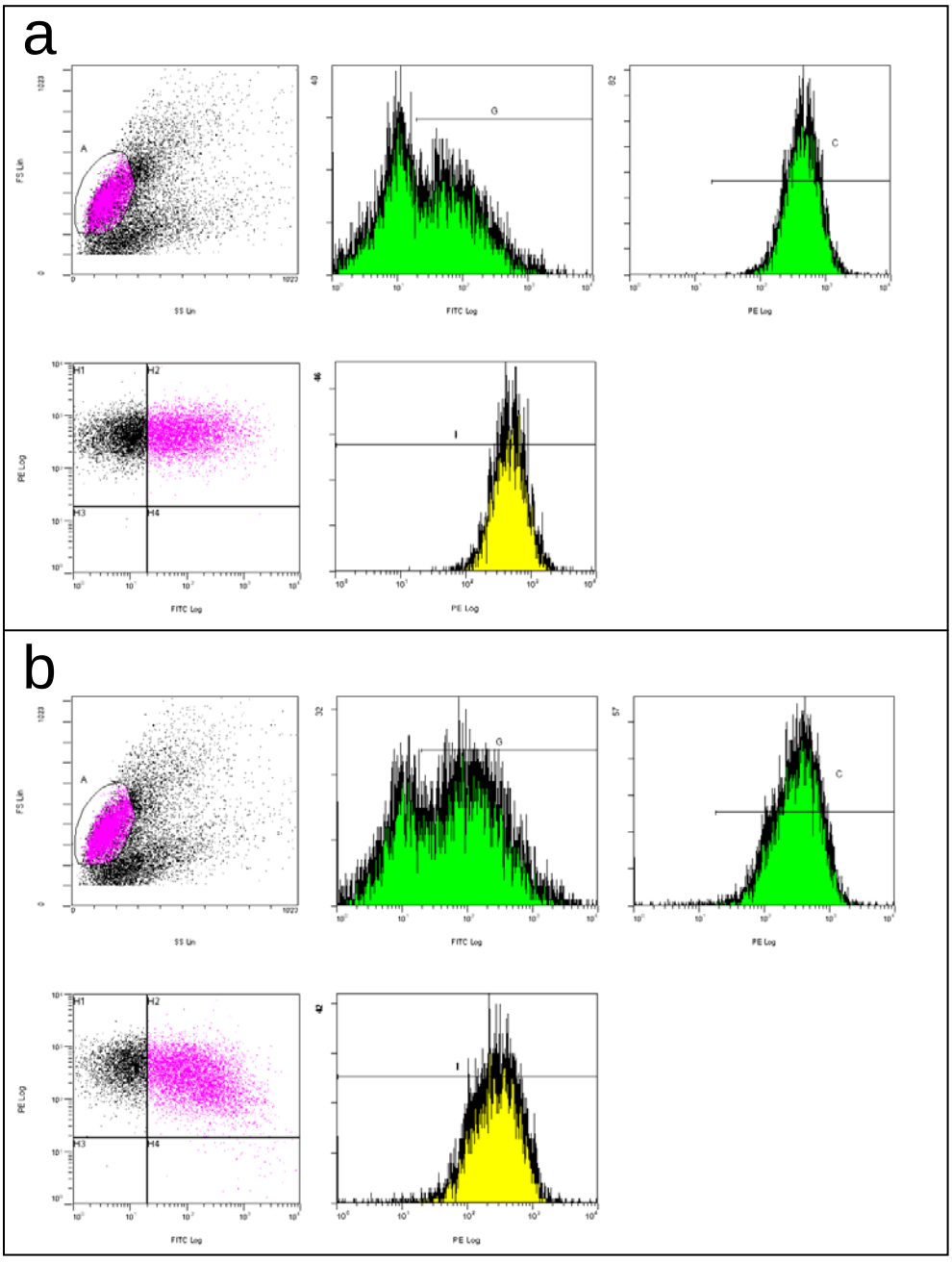
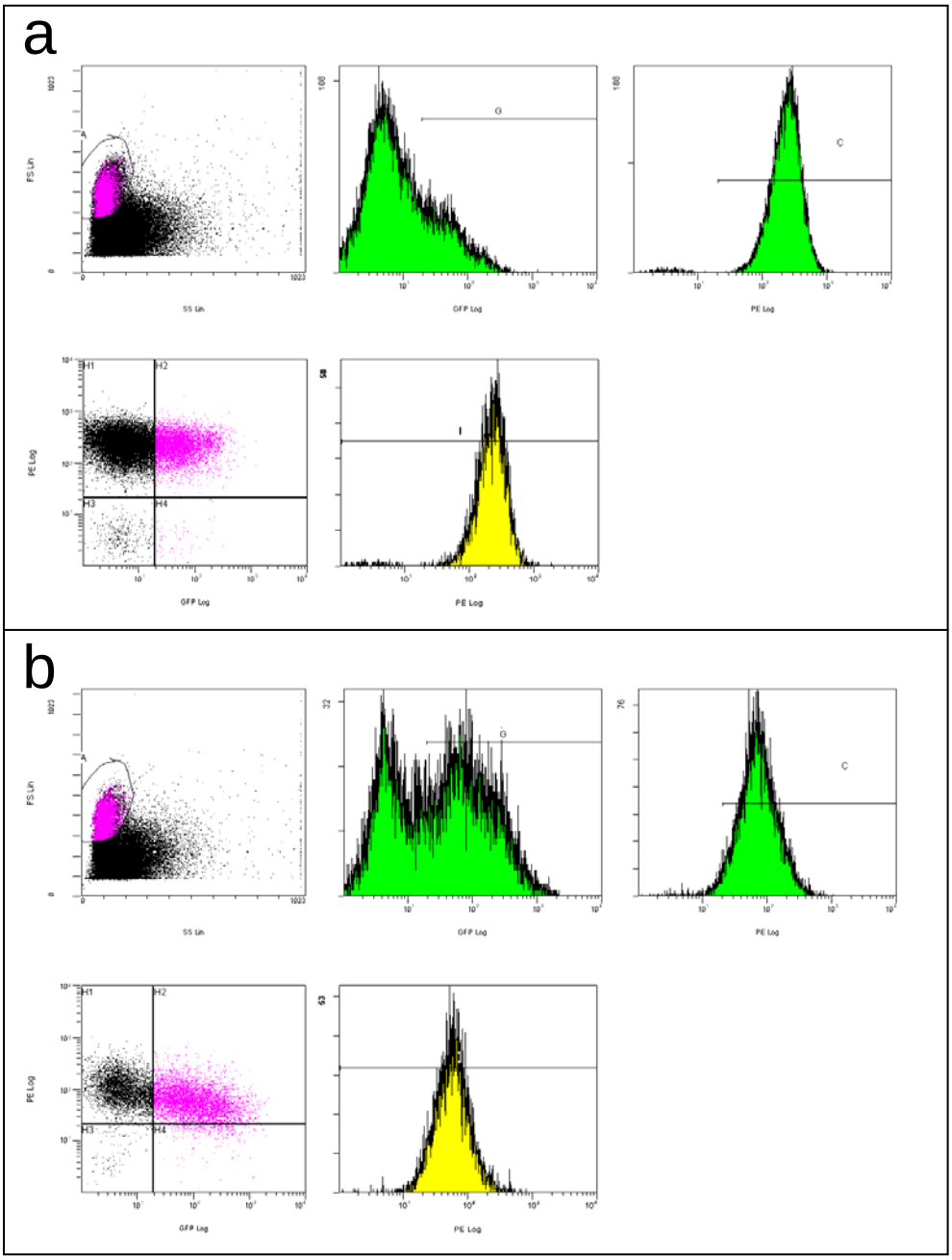


Figure 6: Nef subregions showing discordant mutations between CSF- and plasma-derived population sequences from individual C. Date of tissue collection indicated at left and Nef position number indicated on top of each chart.



Supplementary Figure 1: HIV-1 NL4-3 Nef-induced downregulation of MHC class I in PMA-activated Jurkat E6.1 T cells. (a) 10 ug mock DNA vector control, (b) 10 ug pCI-NL Nef-expressing (CMV-driven) plasmid. Cells were transfected using the Amaxa "nucleofection" method at the determined optimal density of $3*10^5$ cells/ml.



Supplementary Figure 2: HIV-1 NL4-3 Nef-induced downregulation of MHC class I in primary non-activated CD4+ T cells. (a) 5 ug mock DNA vector control, (b) 5 ug pCI-NL Nef-expressing (CMV-driven) plasmid. In brief, PBMC's were extracted from whole blood using the Ficoll separation technique, then CD4+ cells were isolated via negative selection using the "RosetteSep" antibody cocktail. Cells were transfected immediately after isolation using the Amaxa "nucleofection" method.

Chapter 7

Conclusions and Future Directions

SUMMARY

In this dissertation I have used a combination of computational and experimental tools to systematically compare sequences of HIV-1 derived from several anatomic sites within the human body. My data support the theory that distinct viral genetic and evolutionary characteristics are associated with anatomic compartment-specific HIV-1 populations.

My analysis of HIV-1 V3 sequences in Chapter 2 revealed that information regarding chemokine receptor preference (CCR5 vs. CXCR4) was spread throughout the V3 sequence, conflicting with the pre-existing dogma that only 2 sequence positions governed this phenotype. This insight resulted in the development a novel, machine learning-based coreceptor usage prediction algorithm that is available for public use at: http://genomiac2.ucsd.edu:8080/wetcat/tropism.html. Since the publication of Chapter 2, two different articles have been published by other groups comparing coreceptor prediction algorithms (2,4). Both articles reported that our prediction scheme had the highest positive predictive value (PPV) out of all available methods. In addition, the work described in Chapter 3 revealed that V3 sequences from HIV-1 strains predicted to use the CXCR4 receptor (based on our algorithm) showed more evidence of positive selection than CCR5-using variants, in line with the concept that CXCR4-using viruses tend to be more immunogenic (1).

Chapters 4, 5, and 6 systematically compared HIV-1 sequences derived from semen, blood plasma, and cerebrospinal fluid. Tissue-specific populations typically clustered independently in phylogenetic reconstructions, and differed on several levels

including extent of genetic diversity, glycosylation patterns, intensity of positive selection pressure, and coreceptor usage phenotype. Machine learning analysis of *env* sequences revealed that there may be specific signature mutations associated with viruses from different anatomic sites. Additionally, comparison of CSF- and plasma-derived HIV-1 *nef* sequences demonstrated that phenotypic discrepancies between tissue-specific viral proteins may be observed in a reductionist *in vitro* setting. Taken together, these results strongly suggest that immunological surveillance and target cell characteristics differ between the central nervous system, male genital tract, and peripheral tissues, and these differences select for divergent locally adapted HIV-1 populations.

FUTURE DIRECTIONS

The work described in Chapter 5 of this dissertation suggested that certain mutations may be overrepresented in central nervous system (CNS)-derived HIV-1 strains. The selective advantage of these mutations within the CNS environment has not been determined as yet, due to a lack of appropriate experimental systems. Trillo-Pazos *et al* have recently developed a three-dimensional *in vitro* model of the human brain, by culturing human fetal brain tissue on Noble agar-coated plates in Dulbecco's modified Eagle's serum and 5% human serum for 4 weeks (5). This *in vitro* model has been employed to evaluate the neuropathogenic consequences of HIV-1 infection, and to determine how effectively nucleoside reverse transcriptase inhibitors suppress HIV-1 replication in human brain tissue (3,5). These studies demonstrate that the fetal brain aggregate, or "neurosphere" is in fact permissive to HIV-1 infection, based on

measurements of p24 (capsid protein) in culture supernatants and detection of integrated viral DNA in homogenized tissue. I propose using the neurosphere model to investigate the effects of HIV-1 sequence variation on replicative fitness within the CNS environment. My plan consists of five main stages:

1) Inoculate fetal brain aggregates with CCR5- and CXCR4-tropic laboratory-adapted HIV-1 strains at several concentrations to determine the optimal infection model (highest peak p24 production).
2) Mutagenize the fittest lab strain at positions in *env* that appear to be correlated with HIV-1 neurotropism.
3) Compare the replication kinetics of wildtype and mutagenized viruses in the brain aggregate model. The superior performance of mutagenized strains would be a clear demonstration of the selective benefit associated with the CNS-specific *env* mutations.
4) Use existing viral entry assays to determine if the improved replicative fitness of mutagenized strains can be attributed to enhanced viral entry.
5) Histochemically analyze infected brain aggregates to characterize variation in neuropathogenesis resulting from HIV-1 sequence variation.

We have generated preliminary data regarding stage 1 of this plan. Fetal brain aggregates were infected with the CXCR4-tropic strain NL4-3 and "NL-BAL", a mutant version of NL4-3 containing the CCR5-using *env* gene sequence derived from the macrophage-tropic HIV-1 "BAL" strain (Fig. 1). There was no convincing evidence of viral replication at any of the tested inoculum concentrations and

adsorption times. Although p24 values did reach detectable levels within 3 days post-infection in most treatments, the levels were not high or sustained enough to exclude the possibility that released input virus was being measured instead of *de novo* produced virus (Fig. 2).

Microglial cells are believed to be the principal target of HIV-1 in the CNS. An average fetal brain aggregate consists of approximately 500,000 cells, 2-3% of which are microglia. It is hard to imagine that 10,000 sparsely distributed target cells would provide sufficient fuel to propagate HIV-1 infection. The administration of GM-CSF (granulocyte macrophage colony stimulating factor) to forming aggregates reportedly increases microglial cell composition to 10-15% (5), increasing the likelihood of a spreading infection. For our second trial, we inoculated GM-CSF-treated aggregates with NL4-3 and NL-BAL. GM-CSF had no apparent effect on p24 production. Once again, we were not able to reject the possibility that input virus rather than produced virus was being detected (Fig. 3).

To confirm our suspicions that input virus was being detected in p24 assays, brain aggregates were infected with two different strains of HIV-1 containing GFP reporter sequences, "NL-EGFP" and "NL-CSF-GFP". NL-EGFP is a mutant version of NL4-3 containing an internal ribosomal entry site (IRES) and GFP sequence adjacent to the *nef* reading frame. NL-CSF-GFP is a variant of NL-EGFP containing the CCR5-tropic *env* gene from the CSF-derived primary isolate JR-CSF. The advantage associated with these reporter viruses is that infection can be monitored via flow cytometry. The number of cells expressing GFP is a direct indication of the

number of infected cells. In addition, the presence of residual input virus in the brain aggregate is of no consequence; only viruses that productively infect host cells will generate GFP. Our results did in fact confirm our suspicions; neither strain (at any of the tested inoculum concentrations) showed any detectable expression of GFP in the fetal brain aggregate model (Fig. 4).

An *in vitro* model of the brain would certainly be invaluable for studying the effects of HIV-1 sequence variation on neurotropism and neurovirulence. However, we have not been able to productively infect fetal brain aggregates as yet. Further manipulation of this model system will be necessary to make it a viable option for my proposed studies.

CONCLUSIONS

As I mentioned in Chapter 1 of this dissertation, the influence of anatomic compartmentalization on HIV-1 evolution has significant consequences in the clinical world. Several tissues (e.g. the brain) are relatively inaccessible to drugs and act as viral sanctuary sites, compromising antiretroviral therapy. The presence of HIV-1 within the central nervous system results in debilitating neurological disorders in up to 50% of untreated patients. Certain cell types (e.g. macrophages) are more long-lived than others, and enable HIV to establish latent infection that results in indefinite viral persistence. Virus residing in the genital tracts of infected individuals ultimately drives the AIDS epidemic, since HIV-1 is spread predominantly via genital secretions. Understanding how HIV-1 sequence variation correlates with tissue tropism should help considerably in the clinical management of HIV disease and in the prevention of

its transmission. The data presented within this dissertation reinforce the notion that distinct viral genetic and evolutionary characteristics are associated with anatomic compartment-specific HIV-1 populations. The ultimate goal of these studies is to translate information on intrapatient viral genetic variation into effective pharmacological agents and prophylactic intervention schemes. Many, many nucleotides, computer bytes, and pipet tips lie on the long road ahead.

ACKNOWLEDGMENTS

I am grateful to Drs. Gursharan Chana and Ian Everall for their mentorship and provision of fetal brain aggregates.

REFERENCES

1. Harouse J. M., C. Buckner, A. Gettie, R. Fuller, R. Bohm, J. Blanchard, and C. Cheng-Mayer. 2003. CD8+ T cell-mediated CXC chemokine receptor 4 simian/human immunodeficiency virus suppression in dually infected rhesus macaques. Proc Natl Acad Sci U S A. 100:10977-82.

2. Jensen M. A., and A. B. van 't Wout. 2003. Predicting HIV-1 coreceptor usage with sequence analysis. AIDS Rev 5:104-12.

3. Kandanearatchi, A., A. Vyakarnam, S. Landau, and I. P. Everall. 2004. Suppression of human immunodeficiency virus replication in human brain tissue by nucleoside reverse transcriptase inhibitors. J Neurovirol 10:136-9.

4. Sing, T. 2004. Learning localized rule mixtures by maximizing the area under the ROC curve, with an application to the prediction of HIV-1 coreceptor usage, Master's thesis, Max Planck Institute for Informatics.

5. Trillo-Pazos, G., A. Kandanearatchi, J. Eyeson, D. King, A. Vyakarnam, and I. P. Everall. 2004. Infection of stationary human brain aggregates with HIV-1 SF162 and IIIB results in transient neuronal damage and neurotoxicity. Neuropathol Appl Neurobiol 30:136-47.

FIGURES AND TABLES

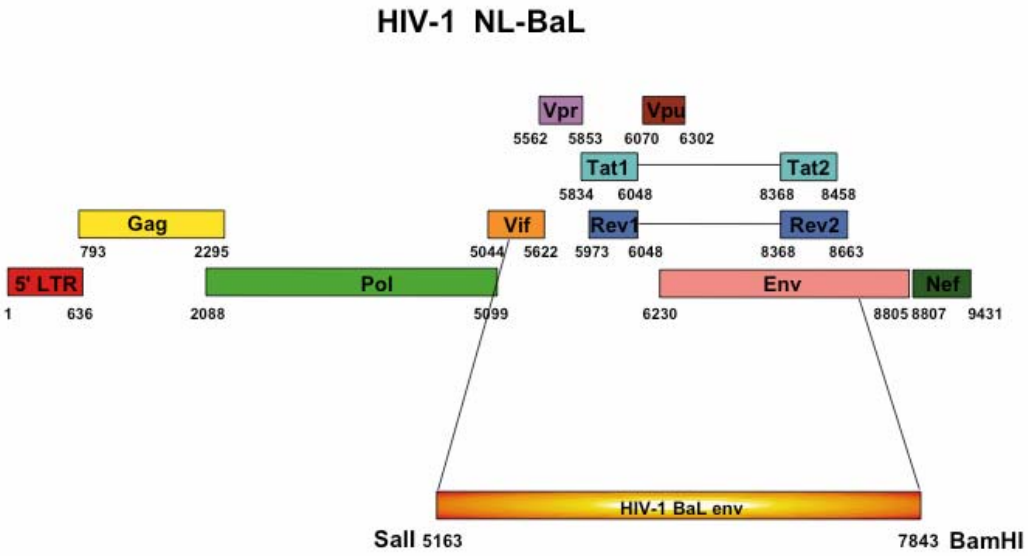


Figure 1: Genome map of the HIV-1 "NL-BaL" strain, which contains the R5, macrophage-tropic BaL *env* gene in place of the original X4 NL4-3 sequence.

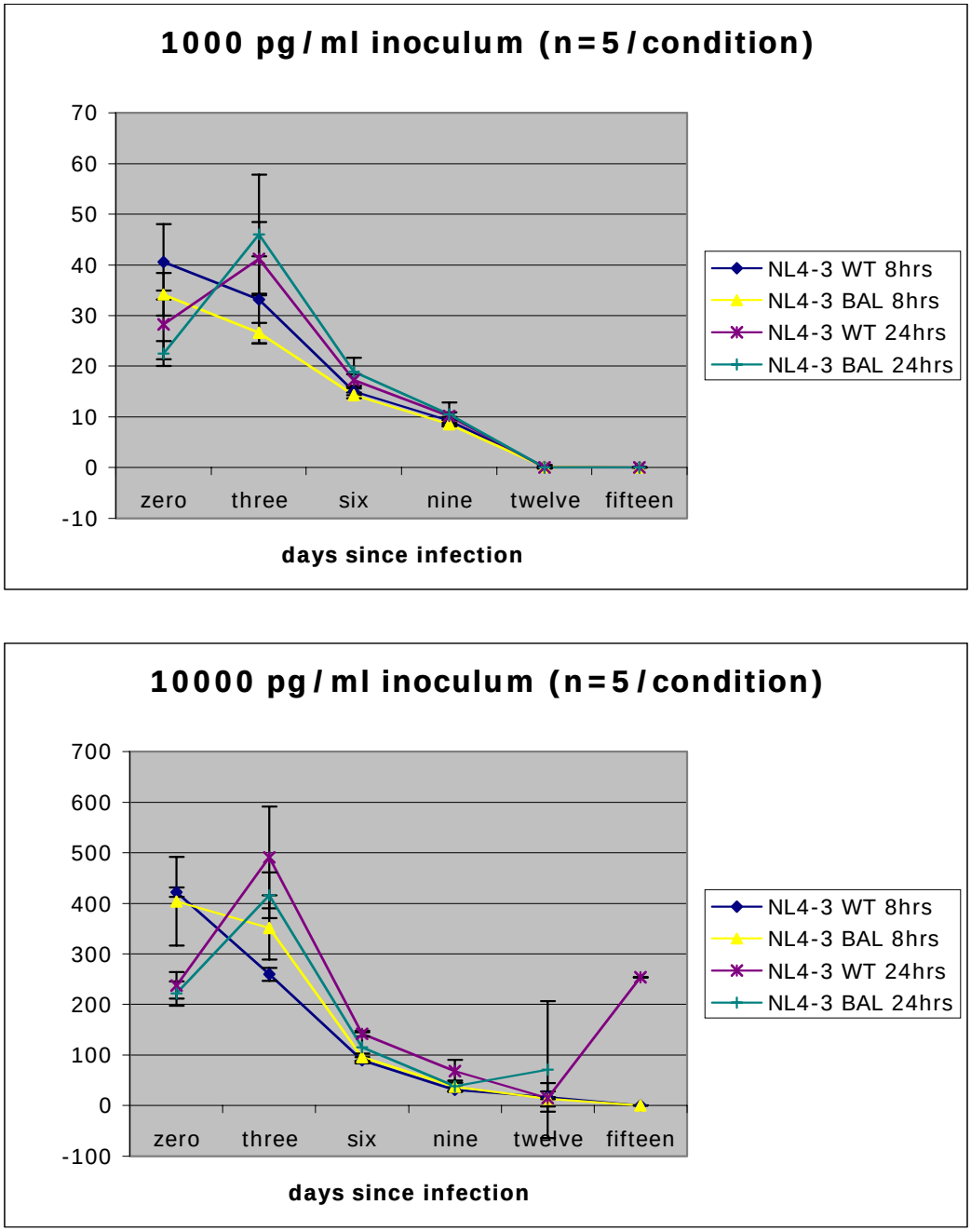
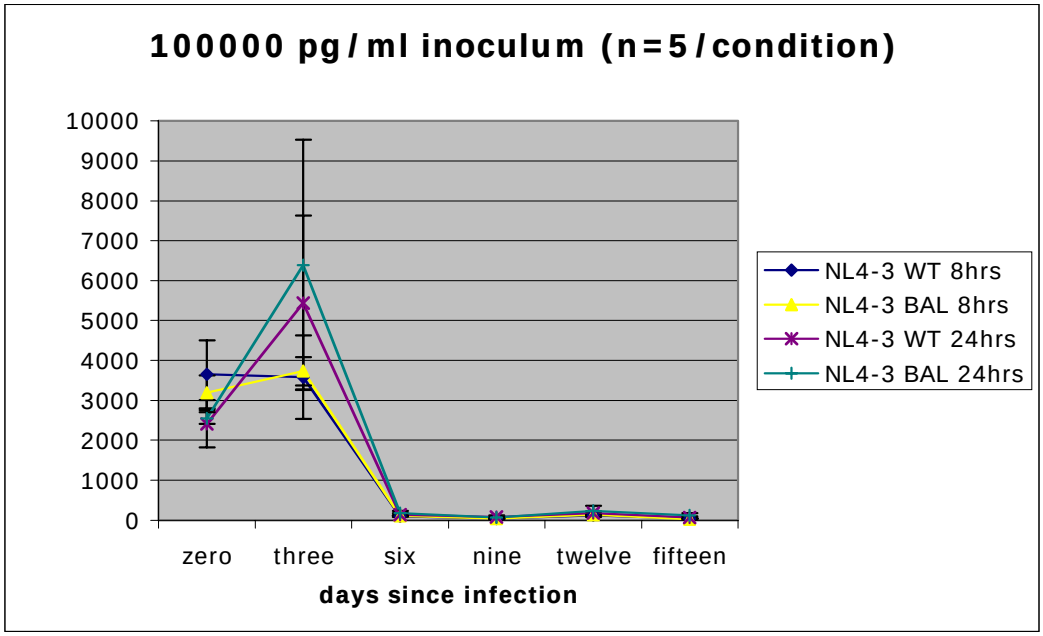


Figure 2: p24 values (pg/ml) in brain aggregate culture supernatants measured at three-day intervals. Input virus washed off at either 8 or 24 hours post infection.



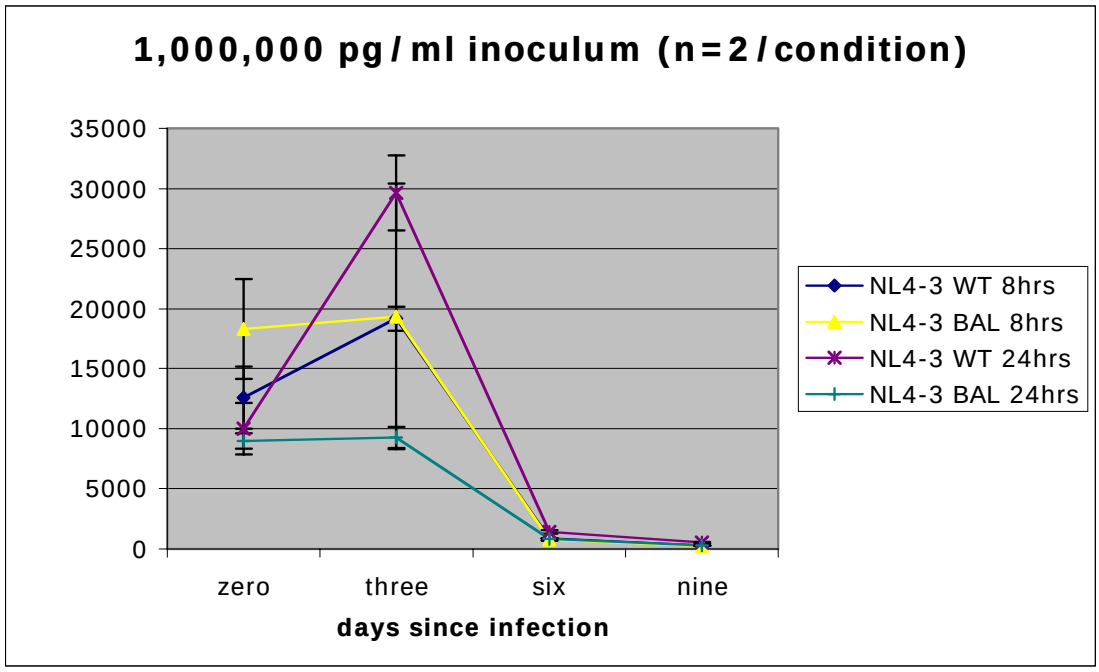


Figure 2 (cont'd): p24 values (pg/ml) in brain aggregate culture supernatants measured at three-day intervals. Input virus washed off at either 8 or 24 hours post infection.

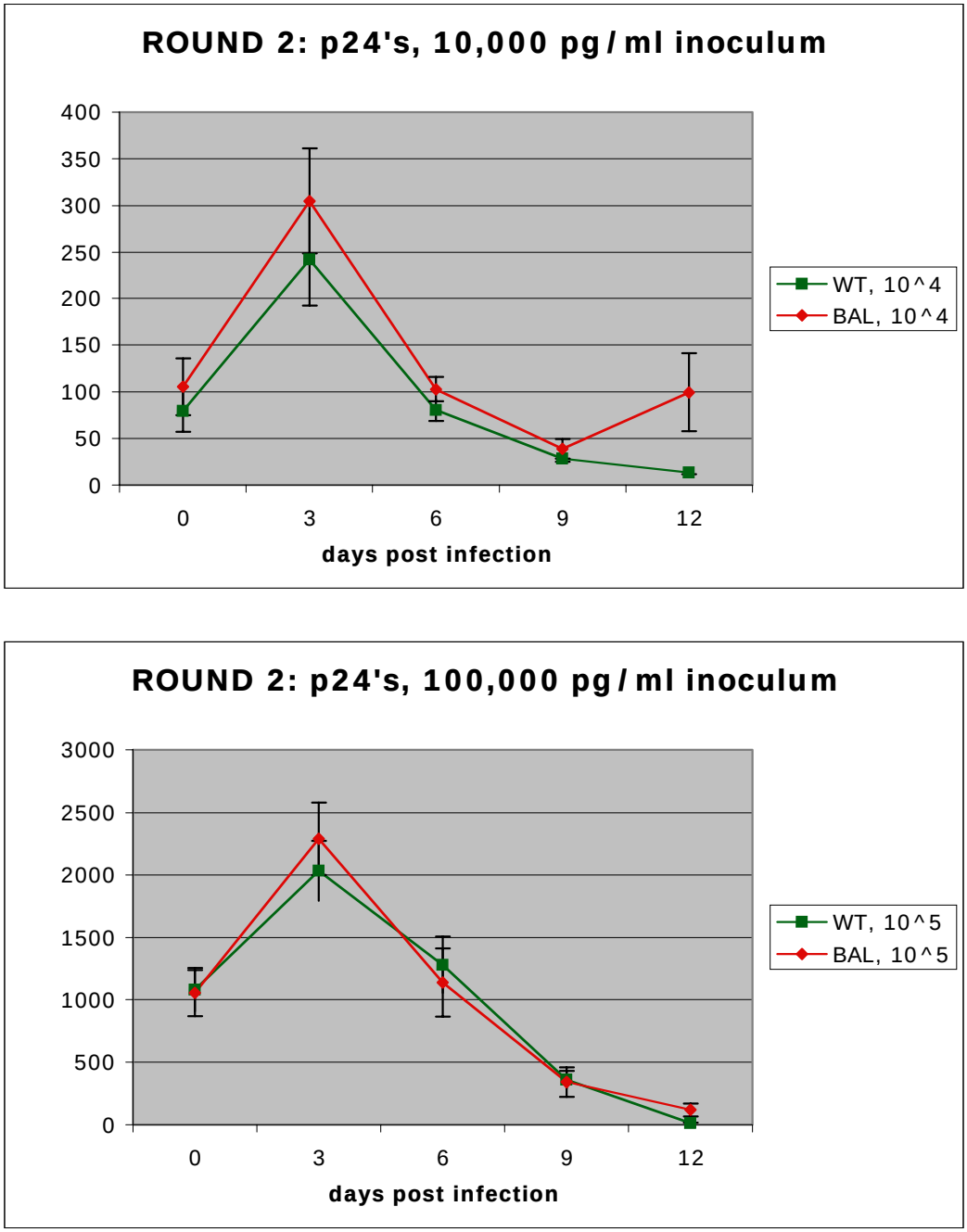


Figure 3: p24 values (pg/ml) in brain aggregate culture supernatants measured at three-day intervals. Input virus washed off 8 hours post infection.

-

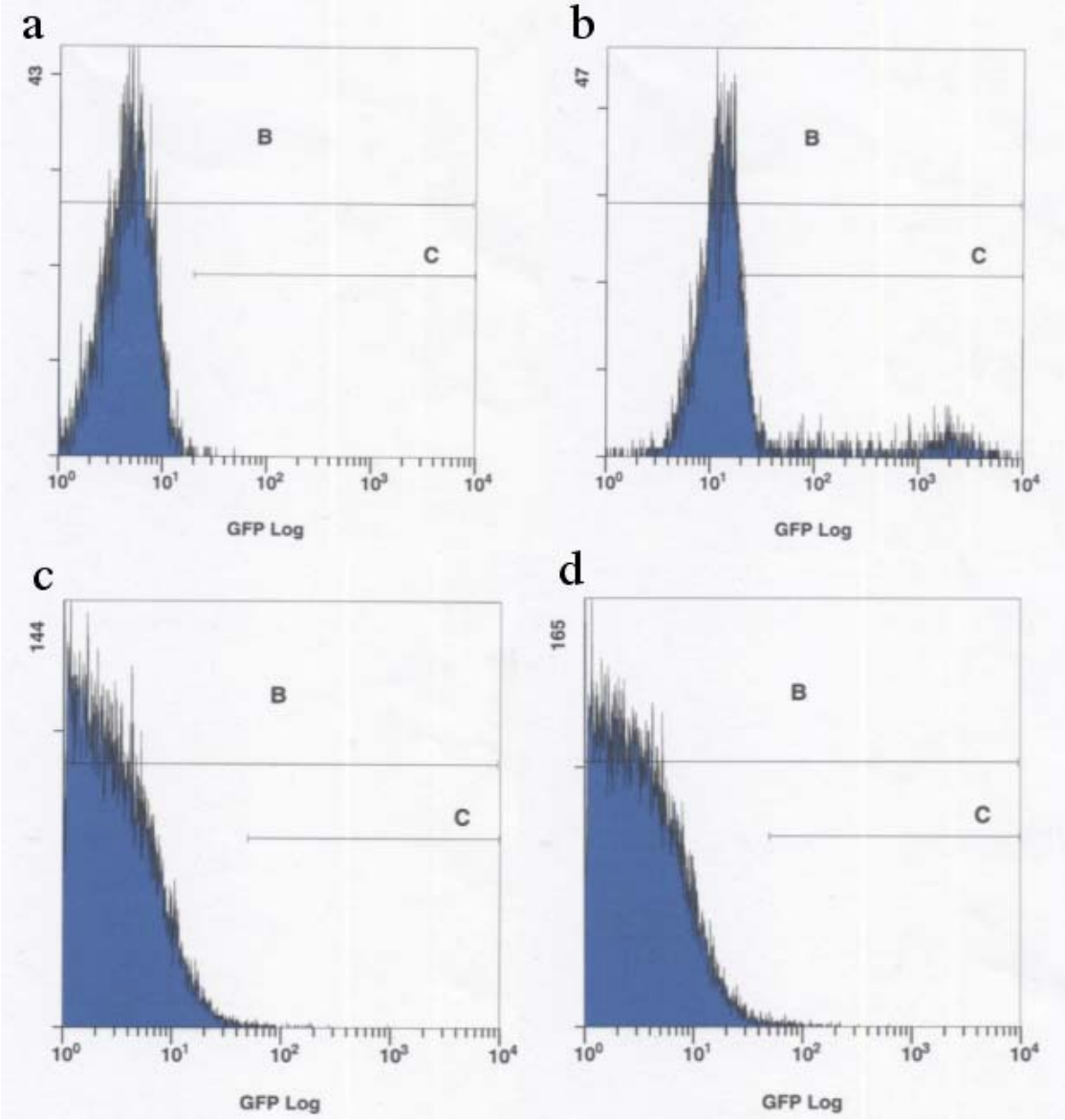


Figure 4: Example of flow cytometric analysis of GFP expression by a CCR5-tropic HIV strain containing a GFP reporter (NL-CSF-EGFP). (a) uninfected "P4R5" cells (HeLa cell line stably transfected with CD4 and CCR5), (b) P4R5 cells 3 days after infection with 100,000 pg/ml NL-CSF-GFP, (c) uninfected homogenized brain aggregate, (d) homogenized brain aggregate 3 days after infection with 100,000 pg/ml NL-CSF-GFP. P4R5's show marked increase in GFP expression, while brain aggregates show no detectable difference.