Semen-Specific Genetic Characteristics of Human Immunodeficiency Virus Type 1 env†
Satish K. Pillai, 1* Benjamin Good, 2 Sergei Kosakovky Pond, 1 Joseph K. Wong, 1, 2 Matt C. Strain, 1 Douglas D. Richman, 1, 2 and Davey M. Smith 1

University of California, San Diego, La Jolla, 1 and Veterans Administration, San Diego Healthcare System, San Diego, 2 California

Received 2 July 2004/Accepted 10 September 2004

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 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.

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 AF089718 to AF098734, AF256230 to AF256465, AF373037 to AF373043, AF35219 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 dna-dist (14). The absolute rate of molecular evolution (molecular clock) was estimated by running TopiDate (38) on maximum likelihood phylogenies with dated tips. A master tree describing the entire data set was built by implementing dndist and neighbor within the PHYLP 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 (alpha_x) and nonsynonymous (beta_x) rates was then fitted independently to every codon position in the data, under two hypotheses: H_0, neutral evolution (alpha_x, beta_x = 0); H_A, nonneutral evolution (alpha_x, beta_x are free to vary independently).

When the hypothesis of neutrality was rejected at site s, it was called positively selected if beta_x was estimated to be greater than alpha_x. 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://genomiacs2.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 sequences based on tissue of origin with 82% accuracy. p, plasma classification; s, semen. The values in parentheses are the number of instances/

<table>
<thead>
<tr>
<th>Parameter</th>
<th>All sequences</th>
<th>Compartmentalized sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly classified instances</td>
<td>434 (65.9%)</td>
<td>217 (81.9%)</td>
</tr>
<tr>
<td>Incorrectly classified instances</td>
<td>225 (34.1%)</td>
<td>48 (18.1%)</td>
</tr>
<tr>
<td>Total number of instances</td>
<td>659</td>
<td>265</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.281</td>
<td>0.636</td>
</tr>
<tr>
<td>True positive rate, blood</td>
<td>79.5%</td>
<td>82.0%</td>
</tr>
<tr>
<td>True positive rate, semen</td>
<td>47.7%</td>
<td>81.8%</td>
</tr>
</tbody>
</table>

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 (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.

To investigate variation in selection pressure from the neutralizing antibody response, we examined glycosylation 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

**FIG. 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.
seven patients characterized by compartmentalized virus, but the direction of the discrepancy is inconsistent \( (P < 0.05 \text{ for six intrapatient comparisons; Mann-Whitney test}) \). 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

![FIG. 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 \( (P < 0.05 \text{ by a Mann-Whitney test}) \). Vertical bars represent standard error.](http://jvi.asm.org/)

<table>
<thead>
<tr>
<th>Sample source</th>
<th>Sites by individual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood</td>
<td>397, 467 461 354, 438 402 335, 336, 337, 340, 343, 354, 446</td>
</tr>
<tr>
<td>Semen</td>
<td>364 364</td>
</tr>
</tbody>
</table>

\( ^a \) Fewer sites were under selective pressure in seminal populations based on the FEL method \( (P < 0.01 \text{ by a paired Wilcoxon test}) \).
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...
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, AI24258, 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 Health Care System.

ACKNOWLEDGMENTS

REFERENCES


(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 microbiicides, 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.