Cross-reactive neutralizing humoral immunity in HIV-1 disease: dynamics of host-pathogen interactions
van Gils, M.J.

Citation for published version (APA):
Genetic composition of replication competent clonal HIV-1 variants isolated from peripheral blood mononuclear cells (PBMC), HIV-1 proviral DNA from PBMC and HIV-1 RNA in serum in the course of HIV-1 infection
ABSTRACT

The HIV-1 quasispecies in peripheral blood mononuclear cells (PBMC) is considered to be a mix of actively replicating, latent, and archived viruses and may be genetically distinct from HIV-1 variants in plasma that are considered to be recently produced.

Here we analyzed the genetic relationship between gp160 env sequences from replication competent clonal HIV-1 variants that were isolated from PBMC and from contemporaneous HIV-1 RNA in serum and HIV-1 proviral DNA in PBMC of four longitudinally studied therapy naïve HIV-1 infected individuals.

Replication competent clonal HIV-1 variants, HIV-1 RNA from serum, and HIV-1 proviral DNA from PBMC formed a single virus population at most time points analyzed. However, an under-representation in serum of HIV-1 sequences with predicted CXCR4 usage was sometimes observed implying that the analysis of viral sequences from different sources may provide a more complete assessment of the viral quasispecies in peripheral blood in vivo.
INTRODUCTION

During the course of HIV-1 infection virus replication and viral turnover are high \(^1\)\(^2\) which, in combination with the error-prone nature of the HIV-1 reverse transcriptase and its lack of proofreading, contribute to the high genetic variability of HIV-1 and result in a continuous emergence of new HIV-1 variants \(^3\)\(^4\). The random generation of viral mutants facilitates escape from host immune pressure and antiviral drugs, and selection for biological properties such as co-receptor use and replication capacity.

The half life of HIV-1 in plasma is about 1.3 hours indicating that virions present in this compartment have been produced very recently \(^5\). The CD4\(^+\) T cell subset in peripheral blood mononuclear cells (PBMC) is one of the targets for HIV-1, and, once infected, produces new viral progeny, part of which will be replication competent and likely to contribute to the viral quasispecies found in plasma. Viruses produced in other anatomical compartments may however also contribute to the composition of the viral population in plasma, at least to some extent or in certain disease stages. While plasma is considered to harbor recently produced virus variants, PBMC are considered to harbor a combination of recently produced and archived virus variants as they accumulate integrated viral DNA \(^6\)\(^-\)\(^8\). Taken together, this would imply that the genetic composition of the viral populations derived from plasma and PBMC may be different.

Whereas sections of the viral genome can be functionally evaluated using recombinant pseudovirus assays, the study of biological properties of full-length replication competent HIV-1 requires virus isolation in vitro. Given that the direct isolation of full-length replication competent viruses from plasma is not very efficient \(^9\)\(^-\)\(^12\), virus isolation is usually done by coculturing of patient PBMC with stimulated healthy donor PBMC \(^13\)\(^14\). However, during virus isolation in bulk culture, the patient’s HIV-1 variant that is most fit for replication in PBMC in vitro will be rapidly selected and outgrow the less fit virus variants that coexisted in the patient in vivo \(^15\)\(^-\)\(^18\). To overcome this, we designed a protocol in which limiting numbers of HIV-1 infected patient PBMC and stimulated healthy donor PBMC are mixed in multiple parallel cocultures. This procedure allows for the isolation of multiple replication competent clonal HIV-1 variants (CV) from a single PBMC sample \(^19\) and avoids outgrowth and loss of slowly replicating variants \(^14\).

To investigate the genetic relationship between viral variants derived from the cell-free (serum) and cell-associated (PBMC) virus pool and to examine whether CV are representative for the replication competent viral quasispecies in peripheral blood, we compared HIV-1 gp160 \textit{env} sequences from replication competent CV with contemporaneous viral gp160 \textit{env} sequences derived directly from serum (viral RNA) and PBMC (proviral DNA) in a longitudinal study of 4 combined antiretroviral therapy (cART) naïve HIV-1 infected individuals throughout their course of infection.
Chapter 6

MATERIALS AND METHODS

Patients and samples
Longitudinal blood samples from eight participants of the Amsterdam Cohort Studies on HIV-1 infection and AIDS (http://www.amsterdamcohortstudies.org) were used for this study. Patients A, B, C and D were selected on the basis of available CV isolated from PBMC at multiple time points in the course of infection, and the availability of serum and DNA samples from most of these same time points. Additionally, patients A and B were selected because they developed CXCR4-using CV in their course of infection while patients C and D were selected because they did not develop CXCR4-using HIV-1 variants. To substantiate our findings on prevalence of X4/CXCR4-using viruses in cell-free and cell-associated HIV-1 populations, four additional patients, E, F, G and H, were selected because they developed CXCR4-using HIV-1 variants during their course of infection and because viral sequences from serum-RNA and CV were available from multiple time points covering their course of HIV-1 infection. Time from documented or imputed seroconversion (SC) at the moment of sampling and number of sequences analyzed are indicated in Supplementary Table S6.1 for each patient. Patients did not receive cART during the study period. For each patient longitudinal data on CD4 counts and viral load are shown in Supplementary Figure S6.1.

The Amsterdam Cohort Studies are conducted in accordance with the ethical principles set out in the declaration of Helsinki and written informed consent was obtained prior to data collection. The study was approved by the Academic Medical Center institutional medical ethics committee.

Isolation of replication competent clonal HIV-1 variants
Replication competent clonal HIV-1 variants (CV) were obtained in cocultures of longitudinally sampled patient PBMC and 2- to 3- day phytohemagglutinin (PHA) stimulated PBMC from a healthy donor (PHA-PBMC) as described previously. In brief, PBMC from a healthy donor were stimulated for 2-3 days in Iscove’s Modified Dulbecco’s Medium (IMDM, Lonza) supplemented with 10% Fetal Calf Serum (FCS; Hyclone), 1mg/ml PHA (Welcome), 100U/ml Penicillin and 100U/ml Streptomycin (Pen/Strep; Gibco Brl), 5mg/ml Ciprofloxacin (Bayer) in a culture flask at a cell density of 5x10^6/ml. Increasing numbers of patient PBMC (range 2,500-40,000) were added to 1x10^5 PHA-PBMC (48 parallel micro-cocultures per patient PBMC number) in a final volume of 200µl IMDM-IL2 medium (IMDM supplemented with 10% FCS, 100U/ml Penicillin and 100U/ml Streptomycin, 10U/ml rIL-2 (proleukin; Chiron Benelux BV), 5mg/ml Ciprofloxacin and 5mg/ml polybrene (Sigma)) for 28 days in a 96-well flat-bottom microtiter plate. Every week, culture supernatants were tested for virus production in an in-house Gag p24 antigen capture enzyme-linked immunosorbent assay (ELISA). At the same time, half of the remaining resuspended culture volume was transferred to new 96-well plates and fresh PHA-stimulated healthy donor PBMC were added to propagate the culture. If less than
one-third of the 48 parallel micro-cocultures per patient PBMC number is positive for p24 production, then according to Poisson distribution these cultures can be assumed to be infected with progeny virus of one infected patient cell and, therefore, they are highly likely to be clonal. Cultures of clonal viruses were expanded by cocultivation of PBMC from p24 positive micro-cocultures with 5x10⁶ PHA-PBMC at a density of 1x10⁶/ml IMDM-IL2 medium in a culture flask. From all patients, virus isolations were performed at 4 to 6 time points spanning the course of infection.

HIV-1 RNA isolation from serum, cDNA synthesis and PCR amplification
Serum RNA was isolated with the QIAamp Viral RNA Mini Kit (Qiagen) according to the manufacturer's protocol using 140µl of serum. Isolated RNA was eluted in a final volume of 40µl. Viral RNA (10µl) was amplified in an RT-PCR reaction with a total volume of 50µl using Superscript III One-step RT-PCR with High Fidelity Platinum Taq (Invitrogen) and primers EnvA (fw) (5'-GGCTTAGGCATCTCCTATGGCAGGAAGAA-3') and Env3Rlong (rev) (5'-GGTGTGTGATGGTCTCTATGAGCAGGAAGAAAGTGTTTTTTGTAAC-3'). The RT step was performed at 55˚C for 1 hour, followed by a 2 min denaturation at 95˚C and 40 PCR amplification cycles of 95˚C for 20 sec, annealing for 20 sec, and 68˚C extension for 4 min. Annealing temperatures were 65˚C for 3 cycles, 60˚C for 11 cycles and 55˚C for 26 cycles. A 25 cycle nested PCR reaction was performed on samples that generated insufficient first round product to be visualized by electrophoresis using Advantage 2 Polymerase mix (Clontech) with internal primers Env_2Flong (fw) (5'-GGTTAATTGATAGAAATGCAGAAGAGCGAGACGAGTTGGCAATG-3') and Nef5 (rev) (5'-CCCWTCCAGTCCCCCCTTTTCTTTTAAAAAG-3'), annealing at 55˚C. Gp160 env PCR products were gel purified using the QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's protocol.

DNA isolation from patient PBMC and PCR amplification of proviral DNA
Total DNA was isolated from 3x10⁶ PBMC from HIV-1 infected patients using a modification of the L6 isolation method. Precipitated DNA was dissolved in 100µl of distilled water and 5µl were used for gp160 env gene amplification with the Advantage 2 Polymerase Mix (Clontech) and primers targeting env (EnvA and Env_3Rlong) in a total reaction volume of 50µl. Reactions were as described for the RT-PCR with omission of the 55˚C RT step. A 25 cycle nested PCR reaction was performed with internal primers Env_2Flong and Nef5 on samples that generated insufficient first round product to be visualized by electrophoresis as described in the previous section. The PCR product was gel purified using the QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer’s protocol.

Cloning and sequencing of gp160 env PCR products from serum and PBMC
Purified PCR products were cloned into the pCR®4-TOPO vector from the TOPO
TA Cloning® Kit for Sequencing (Invitrogen) according to the manufacturer’s protocol. Following transformation into One-shot TOP10 chemically competent *E. coli*, positive transformants were selected on LB plates with ampicillin. Eight colonies per sample were picked and shipped to Functional Biosciences, Inc, Madison, WI, USA for sequencing. Plasmid DNA was extracted and sequences were analyzed on an Applied Biosystems 3730xl Genetic Analyzer. Primers used for sequencing are specified in the Supplementary Methods section. Viral load at the time of sampling did not correlate with the diversity of the sequences obtained from serum-RNA or PBMC-DNA, excluding template resampling as a reason for low diversity in some samples.

**DNA isolation, PCR amplification and sequencing of gp160 env from clonal HIV-1 variants**

Total DNA was isolated from 1x10⁶ healthy donor PBMC in vitro infected with clonal HIV-1 isolates, using a modification of the L6 isolation method²² and gp160 env gene amplification was performed with one outer PCR with primers TB3 (fw) (5'-GGCCCTTATTAG-GACACATAGTTAGCC-3') and TBC (rev) (5'-GCTGCTTTGTAAGTCTAGTTGGTCT-TAAAGG-3') and a nested PCR with primers TB2 (fw) (5'-CTCTACAATACCTTG-GCAGTTAGCAGC-3') and TBA (rev) (5'-CCTCTTGTCTTCTTAGGCAGGCACA-3') using the expand high fidelity Taq polymerase kit (Roche) and the following amplification cycles: 2 min 30s 94˚C, 9 cycles of 15s 94˚C, 45s 53˚C, 6 min 68˚C, 30 cycles of 15s 94˚C, 45s 58˚C, 6 min 68˚C, followed by a 10 min extension at 68˚C and subsequent cooling to 4˚C. PCR products were purified using ExoSAP-IT (USB) according to the manufacturer’s protocol. Sequencing conditions were 5’at 94˚C, 30 cycles of 15” at 94˚C, 10” at 50˚C, 2’ at 60˚C and a 10’ extension at 60˚C. Sequencing was performed using BigDye Terminator v1.1 Cycle Sequencing kit (ABI Prism, Applied Biosystems) according to the manufacturer’s protocol. Primers used for sequencing are specified in the Supplementary Methods section. Sequences were analyzed on the 3130 xl Genetic Analyzer (Applied Biosystems).

**Prediction and determination of co-receptor usage**

All study participants were routinely tested at approximately 3-monthly intervals for the presence of replication competent CXCR4-using HIV-1 variants in PBMC using the MT-2 assay²³. In brief, 1x10⁶ patient’s PBMC were cocultured with 1.6x10⁶ MT-2 cells during 3-4 weeks in duplicate and cultures were periodically screened for syncytia formation and for p24 antigen production in the culture supernatant. The midpoint between the last negative and first positive MT-2 test was estimated as the date of first emergence of CXCR4-using viruses in the patient. All isolated clonal HIV-1 variants were tested for their ability to use CXCR4 in the MT-2 assay¹⁴. For patient E, CXCR4-using variants were isolated via clonal virus isolation whilst the MT-2 test performed on the same PBMC sample was still negative at that time point, which is most likely related to the lower number of cells that is used for the MT-2 assay.
(1x10⁶ patient PBMC) as compared to the clonal viral isolation procedure (up to 5x10⁶ patients PBMC). Indeed, the relative contribution of CXCR4 using HIV-1 to the total viral load may initially be low. CXCR4-using phenotype of CV was confirmed with V3 amino acid sequences and the PSSM matrix and using the geno2pheno[coreceptor] method (FPR = 5%) [26]. The same prediction methods were used to predict co-receptor use of viral sequences obtained from viral RNA in serum or from proviral DNA in patient PBMC. Throughout the manuscript we use the term “CXCR4-using” for CV for which actual CXCR4 usage has been demonstrated in the MT2 cell-line, not excluding the possibility that these CV also have the ability to use CCR5 or any other coreceptor. For sequences with a predicted ability to use CXCR4 we use the term “X4” since actual usage of this coreceptor has not been experimentally proven. The term X4 does not exclude the possibility that the envelope has the ability to also use CCR5 or any other coreceptor. The term “CCR5-using” is used for CV which failed to infect the MT2 cell line and the term “R5” is used for sequences with a predicted ability to use CCR5.

**Phylogenetic analysis**

Nucleotide sequences were aligned using ClustalW included in the software package of BioEdit v.7.0.9 [27] (BioEdit v 7.0.5, Tom Hall, Ibis Therapeutics, Carlsbad, CA) and edited manually.

Neighbor-Joining (NJ) tree [28] for the gp160 envelope sequences from serum, proviral DNA from patient PBMC and isolated clonal HIV-1 variants from the 4 patients under study and a reference sequence panel from different HIV-1 subtypes obtained from Los Alamos Database was constructed under the Hasegawa-Kishino-Yano (HKY85) model of evolution [29] in PAUP* 4.0 beta 8 software package [30] (http://paup.csit.fsu.edu/). Phylogenetic confidence was assessed by bootstrap with 1000 replicates (data not shown).

The best-fit nucleotide substitution model for every patient sequence data set was selected by hierarchical likelihood ratio test (hLTR) in Model Test 3.7 [31] and implemented in the construction of maximum likelihood (ML) trees for the gp160 env region per patient. The heuristic search for the best tree was performed using a NJ tree as starting tree and the TBR branch-swapping algorithm. NJ trees were constructed under the HKY85 model with a transition/transversion ratio and the shape of the γ-distribution estimated using maximum likelihood. ML trees were rooted using the root that maximized the correlation of root-to-tip divergence as a function of sampling time.

**Analysis of genetic distance**

To estimate diversity and divergence, pairwise nucleotide distances were calculated with the Kimura-2 parameter model of evolution in the software package MEGA 4 [32]. Mean pairwise distances were compared using the Mann-Whitney test for independent samples (SPSS 16.0 software package).
Statistical tests for compartmentalization

Six different compartmentalization tests were used to determine compartmentalization between sequences from viral RNA in serum, from proviral DNA in PBMC, and from clonal HIV-1 variants. Four of these tests are based on the tree topology (Slatkin-Maddison (SM)\textsuperscript{33}, Simmonds Association Index (AI)\textsuperscript{34} and Correlation Coefficients (r and r\textsubscript{b})\textsuperscript{35}) and two of them are based on pairwise genetic distances between viral sequences (Wright’s measure of population subdivision (F\textsubscript{ST})\textsuperscript{36-38} and Nearest-neighbor statistic (Snn)\textsuperscript{39}). Compartmentalization between sequences derived from the three sources was analyzed with four tests (SM, AI, r and r\textsubscript{b}). In addition, a pairwise comparison of sequences from two sources was performed with all six tests. Analyses were implemented in ML trees or alignments that included only the gp160 env sequences from time points at which the sources to be compared were available. Analyses were implemented in HyPhy as described\textsuperscript{40}. Methods and parameters implemented in Hyphy for each method are described in Supplementary Material and Methods.

Results

Phylogenetic analysis of gp160 env sequences from viral RNA in serum, PBMC proviral DNA and clonal HIV-1 variants isolated from PBMC

Neighbor-joining phylogenetic tree analysis of gp160 env nucleotide sequences from all four patients and a reference sequence panel from different HIV-1 subtypes showed monophyletic clustering per patient of sequences derived from viral RNA in serum (serum-RNA), proviral DNA in PBMC (PBMC-DNA) and replication competent PBMC-derived clonal HIV-1 variants (CV) with high bootstrap support (values between 80 and 100), indicating absence of cross-contamination between patient samples (data not shown).

To better examine the intra-host genetic relationship between the HIV-1 quasispecies in serum-RNA, PBMC-DNA, and CV, maximum likelihood (ML) trees were constructed with gp160 env sequences from each patient separately. Tree topologies revealed temporal structure in all four patients: sequences derived from serum-RNA, PBMC-DNA or CV that had been sampled at the same time tended to cluster together, and sequences from samples collected later in infection showed greater divergence from the root of the tree. The degree of intermingling of sequences from the three different sources varied between patients and per time point. For patient A (Figure 6.1), we observed separate clustering of PBMC-DNA sequences at the first and third time point, respectively, but intermingling of the PBMC-DNA sequences, mainly with CV sequences, at later time points. Sequences from serum-RNA and CV from the first four time points intermingled, but later serum-RNA sequences clustered separately from both contemporary PBMC-DNA and CV sequences, especially at the last time point where they formed a well supported monophyletic cluster that considerably diverged from the other sequences. In patient B (Figure 6.2), PBMC-DNA
sequences collected at the first and second time point intermixed with CV sequences. At the third and fourth time point, viral sequences from all three sources were interspersed. The tree also showed that the majority of serum-RNA sequences from the first two time points and two PBMC-DNA sequences from the third time point formed separate phylogenetic lineages. Only two serum-RNA sequences from the first and second time point, respectively, clustered with sequences from later serum-RNA samples, and with all the CV and the majority of PBMC-DNA sequences. All serum samples were processed separately from the PBMC samples and the CV, excluding contamination between samples from the three different sources and implying that the two clusters of viral \textit{env} sequences from early serum-RNA samples and the two PBMC-DNA sequences from the third time point of patient B had indeed evolved independently from the other sequences of this patient. In patient C and D (Figures 6.3 and 6.4), intermingling of sequences from the three different sources could be observed at all time points. However, similar to the observation for patient B, five sequences from serum-RNA of the first three time points of patient C clustered separately from the remainder of the sequences of that patient.

**Patient A**

![Diagram](image)

**Figure 6.1**: Maximum-likelihood tree of gp160 \textit{env} sequences from viral RNA in serum, PBMC proviral DNA and clonal HIV-1 variants.

ML tree was rooted using the root that maximized the correlation of root-to-tip divergence as a function of sampling time. Bootstrap support with value $>70\%$ is shown. * indicate archived sequences. The scale bar (horizontal line) indicates branch length corresponding to 0.01 substitutions per site.
Figure 6.2: Maximum-likelihood tree of gp160 env sequences from viral RNA in serum, PBMC proviral DNA and clonal HIV-1 variants
ML tree was rooted using the root that maximized the correlation of root-to-tip divergence as a function of sampling time. Bootstrap support with value >70% is shown. * indicate archived sequences. The scale bar (horizontal line) indicates branch length corresponding to 0.01 substitutions per site.
Patient C

Figure 6.3: Maximum-likelihood tree of gp160 env sequences from viral RNA in serum, PBMC proviral DNA and clonal HIV-1 variants

ML tree was rooted using the root that maximized the correlation of root-to-tip divergence as a function of sampling time. Bootstrap support with value >70% is shown. * indicate archived sequences. The scale bar (horizontal line) indicates branch length corresponding to 0.01 substitutions per site.

1: 34 months after SC
2: 48 months after SC
3: 72 months after SC
4: 87 months after SC
Figure 6.4: Maximum-likelihood tree of gp160 env sequences from viral RNA in serum, PBMC proviral DNA and clonal HIV-1 variants

ML tree was rooted using the root that maximized the correlation of root-to-tip divergence as a function of sampling time. Bootstrap support with value >70% is shown. * indicate archived sequences. The scale bar (horizontal line) indicates branch length corresponding to 0.01 substitutions per site.
CXCR4-using variants, which were detected in patients A and B, formed a monophyletic cluster irrespective of the time of sampling. Interestingly, sequences with an X4 signature were only obtained from PBMC-DNA and CV.

Few sequences per patient were identified as archived (marked with * in the phylogenetic tree) as they clustered with sequences from earlier time points and had diverged less from the root of the tree than other sequences from the same time point. We found a total of 2 archived sequences in serum-RNA (in patients C and D), 4 in PBMC-DNA (1 in patient A and 3 in patient C) and 6 in CV (2 in patient B and 4 in patient D).

**Gene flow between sequences from viral RNA in serum, PBMC proviral DNA and clonal HIV-1 variants**

Compartmentalization tests are generally used to detect a restriction of gene flow between viral subpopulations, in which case each virus subpopulation, generally coming from different anatomical compartments, can evolve independently and become genetically distinct. Here, we used these tests to determine whether there was evidence for frequent gene trafficking between virus variants derived from serum-RNA, PBMC-DNA and CV as if they belonged to a single virus population, or whether gene trafficking was restricted implying differences in genetic composition and independent evolution between virus variants of the different sources.

ML trees and alignments based on envelope sequences of each patient were analyzed with six different compartmentalization tests: SM, r, rb and AI that are all based on tree topology, and F_St and Snn that are based on pairwise genetic distances between viral sequences. The 4 tree-based tests, which allow for simultaneous comparison of more than two sequence populations, were used for the analysis of gene trafficking between serum-RNA, PBMC-DNA and CV. In addition, a pairwise comparison of two sources (serum-RNA vs CV, CV vs PBMC-DNA and serum-RNA vs PBMC-DNA) was performed with all six tests (Table 6.1). Comparisons included only the time points at which sequences from all the sources compared were available. Given the different sensitivities and frequently discordant predictions of the different methods, for each comparison we took a majority consensus approach meaning that the signal for compartmentalization was only considered positive when more than half of the tests gave significant P-values.

For patients A, C and D, borderline significant compartmentalization was detected in the simultaneous analysis of all three sources of viral sequences (2 of 4 tests with a significant P-value) suggesting that the viruses from serum-RNA, PBMC-DNA and CV in these patients are likely to form a single virus population. This was supported by the comparisons of each pair of sequence sources separately in which ≤3 of 6 tests gave significant P-value. For patients A and C, we obtained a borderline significant result (3 of 6 tests positive for compartmentalization) for the compartmentalization analysis between PBMC-DNA and either serum-RNA or CV. This coincided with a certain degree of segregation, at some time...
points, between PBMC-DNA and either serum-RNA or CV sequences in the phylogenetic tree of those two patients. For patient B, significant $P$-values were obtained in all 4 compartmentalization tests in which sequences from serum-RNA, PBMC-DNA, and CV were analyzed together. Analysis of gene flow for each pair of sequence sources separately showed that compartmentalization occurred between serum-RNA and either PBMC-DNA or CV ($\geq 5$ of 6 tests with a significant $P$-value). This was supported by the tree topology in which only two serum sequences from the first and second time points clustered with the sequences from later serum-RNA samples and all the CV and the majority of PBMC-DNA sequences, while all other serum-RNA sequences from the first and second time point and two PBMC-DNA sequences from the third time point formed a separate cluster.

Table 6.1: Results of compartmentalization tests per patient and time point

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td>$r_1$</td>
<td>$r_2$</td>
<td>$r_3$</td>
<td>$r_4$</td>
</tr>
<tr>
<td>ALL</td>
<td>0.001</td>
<td>0.019</td>
<td>ns</td>
<td>ns</td>
<td>0.013</td>
<td>ns</td>
</tr>
<tr>
<td>SE vs CV</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>0.003</td>
<td>0.009</td>
</tr>
<tr>
<td>CV vs PB</td>
<td>0.003</td>
<td>0.009</td>
<td>ns</td>
<td>ns</td>
<td>0.003</td>
<td>ns</td>
</tr>
<tr>
<td>SE vs PB</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.003</td>
<td>0.003</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td>$r_1$</td>
<td>$r_2$</td>
<td>$r_3$</td>
<td>$r_4$</td>
</tr>
<tr>
<td>ALL</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>SE vs CV</td>
<td>&lt;0.0001</td>
<td>0.016</td>
<td>0.009</td>
<td>ns</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>CV vs PB</td>
<td>&lt;0.0001</td>
<td>0.001</td>
<td>ns</td>
<td>ns</td>
<td>&lt;0.0001</td>
<td>ns</td>
</tr>
<tr>
<td>SE vs PB</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>$r_1$</td>
<td>$r_2$</td>
<td>$r_3$</td>
<td>$r_4$</td>
</tr>
<tr>
<td>ALL</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>SE vs CV</td>
<td>0.030</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>0.002</td>
<td>ns</td>
</tr>
<tr>
<td>CV vs PB</td>
<td>&lt;0.0001</td>
<td>0.008</td>
<td>ns</td>
<td>ns</td>
<td>&lt;0.0001</td>
<td>ns</td>
</tr>
<tr>
<td>SE vs PB</td>
<td>&lt;0.0001</td>
<td>0.002</td>
<td>ns</td>
<td>ns</td>
<td>&lt;0.0001</td>
<td>ns</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td>$r_1$</td>
<td>$r_2$</td>
<td>$r_3$</td>
<td>$r_4$</td>
</tr>
<tr>
<td>ALL</td>
<td>0.001</td>
<td>0.025</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>SE vs CV</td>
<td>0.041</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>0.015</td>
<td>ns</td>
</tr>
<tr>
<td>CV vs PB</td>
<td>0.008</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>&lt;0.0001</td>
<td>ns</td>
</tr>
<tr>
<td>SE vs PB</td>
<td>0.008</td>
<td>ns</td>
<td>ns</td>
<td>ns</td>
<td>0.003</td>
<td>ns</td>
</tr>
</tbody>
</table>

Compartmentalization analysis was performed with 4 tests (SM, AI, $r$ and $r_s$) when sequences of all three sources (Serum-RNA (SE); clonal HIV-1 variants (CV); PBMC-DNA (PB)) were compared (ALL). Subsequently, all 6 tests were implemented on the comparison of each pair of sources separately (SE vs CV, CV vs PB and SE vs PB) including only the time points at which sequences from the two sources compared were available. The signal for compartmentalization was only considered positive when more than half of the tests gave a significant $P$-value (filled cells). $P$-values $< 0.05$ were considered significant. ns: not significant.

SM: Slatkin-Madison; AI: Simmonds Association Index; $r$: distance between to sequences as the number of tree branches separating them in the phylogenetic tree; $r_s$: cumulative genetic distance between the sequences; Snn: Nearest-neighbor statistic; F_Ст: Wright’s measure of population subdivision.
None of the statistical tests reported here were adjusted for multiple comparisons. While it is customary to adjust $P$-values for multiple comparisons when making positive claims, our use of uncorrected $P$-values is conservative with respect to the negative conclusions in this study (i.e., that compartmentalization was minimal).

Diversity of HIV-1 gp160 env sequences from RNA in serum, PBMC proviral DNA and clonal HIV-1 variants

We next compared the mean pairwise genetic distance per time point (diversity) within the viral gp160 env sequences derived from serum-RNA, PBMC-DNA, and CV (Figure 6.5, and summary in Table 6.2 and Supplementary Table S6.2).

Sequence diversity within serum-RNA, CV, and PBMC-DNA, which was compared at time points where sequences from all three sources were available, was similar at 1 of 3 time points in patients A and D and at 2 of 4 time points in patients B and C (Table 6.2). In patient A and C, sequence diversity was significantly lower in PBMC-DNA at one of the time points (first and fourth time point, respectively). Sequence diversity in serum-RNA was significantly lower at the last time point of patient A and significantly higher at the second time point of patient C. For patient B and D, sequence diversity in serum-RNA was significantly lower at one of the time points (third and first time point, respectively) and PBMC-DNA had the highest, and CV the lowest sequence diversity at the other time point (first and third time point for patients B and D, respectively; Table 6.2).

Diversity between serum-RNA and CV could be compared at 5, 4, 4, and 3 time points for patients A, B, C, and D, respectively (Table 6.2). At 9 of these 16 time points, corresponding to 33-75% of the time points analyzed per patient (Supplementary Table S6.2), sequence diversity between serum-RNA and CV was similar.

Diversity between CV and PBMC-DNA could be compared at 4 time points for patients A, B and C and at 5 time points for patients D (Table 6.2). At 11 of those 17 time points, corresponding to 50-75% of the time points analyzed per patient, PBMC-DNA and CV had similar sequence diversity (Supplementary Table S6.2).

Diversity in serum-RNA and PBMC-DNA could be compared at 3 time points for patients A and D, and at 4 time points for patients B and C (Table 6.2). Similar sequence diversity in serum-RNA and PBMC-DNA was found at 6 of the in total 14 time points at which the comparison was performed, corresponding to 33-50% of the time points that were analyzed per patient (Supplementary Table S6.2).

Differences in sequence diversity between serum-RNA, CV and PBMC-DNA were detected in the pairwise comparisons of the different sources under study at 25-67% time points analyzed per patient. Those differences could not be attributed to a higher or lower sequence heterogeneity of one of the sources in particular, neither to any patient-specific patterns. No correlation was found between the viral load at a certain time point and the diversity of sequences from serum-RNA, PBMC-DNA and CV (data not shown).
Divergence of HIV-1 gp160 env sequences from RNA in serum, PBMC proviral DNA and clonal HIV-1 variants

Subsequently, we compared per time point the mean pairwise genetic distance between gp160 env sequences from serum-RNA and PBMC-DNA, serum-RNA and CV, and PBMC-DNA and CV (divergence) with the diversity of the gp160 env sequences from each source separately (Figure 6.5, and summary in Table 6.2 and Supplementary Table S6.2). For patients A, C and D, the divergence between sequences from serum-RNA and CV was similar to the diversity within sequences from either serum-RNA or CV in 33-60% of the time points analyzed. For patient B, divergence between serum-RNA and CV was always

Table 6.2: Comparison of mean gp160 env pairwise genetic distances within (diversity) and between (divergence) sequences from serum-RNA, PBMC-DNA and clonal HIV-1 variants per time point and patient

<table>
<thead>
<tr>
<th>Patient</th>
<th>Time point</th>
<th>Diversity</th>
<th>serum-RNA / CV</th>
<th>CV / PBMC-DNA</th>
<th>serum-RNA / PBMC-DNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>SE = CV &gt; PB</td>
<td>SEvsCV = SE = CV</td>
<td>CVvsPB = CV &gt; PB</td>
<td>SEvsPB = SE &gt; PB</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SE = CV</td>
<td>SEvsCV = SE = CV</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>CV &gt; PB</td>
<td>SEvsCV = CV *</td>
<td>CVvsPB = CV &gt; PB</td>
<td>SEvsPB &gt; PB *</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SE &lt; CV</td>
<td>SEvsCV = CV &gt; SE</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>SE = CV = PB</td>
<td>SEvsCV = SE = CV</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB = SE = PB</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>SE &lt; CV = PB</td>
<td>SEvsCV &gt; CV &gt; SE</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB &gt; PB &gt; SE</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>PB &gt; SE &gt; CV</td>
<td>SEvsCV &gt; SE &gt; CV</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB = SE &gt; PB</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SE = CV = PB</td>
<td>SEvsCV &gt; SE = CV</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB = SE &gt; PB</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>CV = PB &gt; SE</td>
<td>SEvsCV = CV &gt; SE</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB &gt; SE &gt; PB</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SE = CV = PB</td>
<td>SEvsCV &gt; SE / SEvsCV = CV</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB = SE = PB</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>SE = CV = PB</td>
<td>SEvsCV = SE = CV</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB &gt; SE / SEvsPB = PB</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SE &gt; CV = PB</td>
<td>SEvsCV = SE &gt; CV</td>
<td>CVvsPB = BC = PB</td>
<td>SEvsPB = SE &gt; PB</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>SE = CV = PB</td>
<td>SEvsCV = SE = CV</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB = SE = PB</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>SE = CV &gt; PB</td>
<td>SEvsCV &gt; SE / SEvsCV = CV</td>
<td>CVvsPB = BC = PB</td>
<td>SEvsPB = SE &gt; PB</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>CV = PB &gt; SE</td>
<td>CV &gt; SEvsCV &gt; SE</td>
<td>CVvsPB = CV / CVvsPB = PB</td>
<td>SEvsPB = PB &gt; SE</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>SE = CV = PB</td>
<td>SEvsCV = SE = CV</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB = SE = PB</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>PB &gt; SE &gt; CV</td>
<td>SEvsCV = SE &gt; CV</td>
<td>CVvsPB = CV &gt; PB</td>
<td>SEvsPB = SE / SEvsPB = PB</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>CV &gt; PB</td>
<td>n.a.</td>
<td>CVvsPB = CV &gt; PB</td>
<td>n.a.</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>CV = PB *</td>
<td>SEvsCV = CV *</td>
<td>CVvsPB = CV = PB</td>
<td>SEvsPB &gt; PB *</td>
</tr>
</tbody>
</table>

SE: diversity in serum-RNA; CV: diversity in clonal HIV-1 variants; PB: diversity in PBMC-DNA; SEvsCV: divergence between serum-RNA and CV; CVvsPB: divergence between CV and PBMC-DNA; SEvsPB: divergence between serum-RNA and PBMC-DNA. = indicates no significant difference between mean pairwise genetic distances; < or > indicate significant differences between mean pairwise genetic distances (P-value < 0.05); bold indicates no significant differences between any of the comparisons. * Time point with only one serum sequence available.
higher than diversity in either serum-RNA and/or CV. Divergence between CV and PBMC-DNA sequences was similar to the diversity within the sequences from either CV or PBMC-DNA in 40-100% of the time points analyzed per patient. Divergence between sequences from serum-RNA and PBMC-DNA was higher than the diversity within serum-RNA and/or PBMC-DNA in 33-75% of the time points analyzed per patient. The increased divergence between serum-RNA and both CV and PBMC-DNA for patients A and B coincided with the presence of CXCR4-using CV and sequences with an X4 genotype in PBMC-DNA but not in serum-RNA. Indeed, sequences with an X4 signature clustered separately from R5 sequences in the ML tree (Figure 6.1 and 6.2).

Figure 6.5: Pairwise genetic distances per time point within (diversity) and between (divergence) gp160 env sequences from viral RNA in serum (SE), PBMC proviral DNA (PB) and clonal HIV-1 variants (CV)
Mean pairwise distances were calculated per time point and compared using the Mann-Whitney test for independent samples. P-values < 0.05 were considered significant. Patients A (A), patient B (B), patient C (C), patient D (D).
Figure 6.6: Percentage of predicted R5/CCR5-using variants and predicted X4/CXCR4-using variants in serum RNA (SE), PBMC proviral DNA (PB) and clonal HIV-1 variants (CV). All CV were tested for the ability to use CXCR4 in the MT-2 assay and results were confirmed with coreceptor use prediction based on the V3 amino acid sequences and the PSSM matrix and using the geno2pheno[coreceptor] method (FPR = 5%). The same prediction methods were used to predict coreceptor use of viral sequences obtained from viral RNA in serum or from proviral DNA in patient PBMC. Numbers on bars indicate percentages of predicted R5/CCR5-using variants and predicted X4/CXCR4-using variants per time point.
Divergence between serum-RNA and both CV and PBMC-DNA at the first two time points of patient B could be attributed to the separate clustering of most of the early serum-RNA sequences in the ML tree (Figure 6.2).

**Co-receptor use of clonal HIV-1 variants and predicted co-receptor use of sequences derived from viral RNA in serum and PBMC proviral DNA**

CXCR4 usage of CV was determined by their ability to replicate in the MT-2 cell line \(^{14}\) and confirmed with prediction tools that are based on V3 amino acid sequences \(^{24-26}\). The same prediction tools were used to predict co-receptor usage for viral sequences obtained from serum-RNA and PBMC-DNA.

Co-receptor use of CV and viral sequences derived from serum-RNA and PBMC-DNA was compared for patients who harbored both CCR5- and CXCR4-using variants (patients A and B, and four additional patients (E, F, G and H) from whom gp120 (C2-C4) envelope unpublished sequences from serum-RNA and CV were available (see Supplementary Table S6.1 and Supplementary Material and Methods)). Figure 6.6 shows percentages and absolute numbers of R5/CCR5-using and X4/CXCR4-using variants in serum-RNA and CV for all six patients and additionally in PBMC-DNA for patients A and B. Absence or underrepresentation of X4 variants in serum was observed in patients from whom only low numbers of clonal CXCR4-using HIV-1 variants could be isolated (patients A, B, E and F). Patients A, B and E completely lacked X4 env sequences in serum (Figure 6.6A, 6.6B and 6.6C). For patient F, X4 sequences were found in serum at the time point close to the estimated emergence of CXCR4-using variants but were no longer detected in later stages of infection (Figure 6.6D). Interestingly, in patients A and B, X4 sequences were detected in PBMC-DNA at the same time points at which CXCR4-using CV could be isolated. Patients from whom higher numbers of CXCR4-using clonal HIV-1 variants could be isolated (patients G and H; Figure 6.6E and 6.6F) had similar percentages of X4/CXCR4-using variants in serum and CV.

**Discussion**

The cell-free and cell-associated HIV-1 pool in peripheral blood may potentially differ in genetic composition as the virus variants from plasma are generally considered to represent the recently produced virus population in vivo, while PBMC are considered to harbor a mix of recently produced viruses and archived viruses. To address this, we compared evolution and genetic variability of cell-associated virus (proviral DNA from PBMC (PBMC-DNA) and replication competent clonal HIV-1 variants (CV) isolated from PBMC) and cell-free virus in serum (serum-RNA).

Phylogenetic analysis of gp160 env nucleotide sequences revealed that sequences derived from serum-RNA, PBMC-DNA, and CV from the same time point generally clustered together and that divergence to the most recent common ancestor was in general greater.
for sequences collected at later time points of infection, independently of the source from which they were obtained. Intermingling of sequences from the three sources was observed in all patients, although segregation of sequences from serum-RNA or PBMC-DNA was found at some time points. Firstly, in patients B and C, some of the serum-RNA sequences evolved independently with no descendents at late stages of the infection, suggestive for negative selection of these viruses or only transient production of these viruses from another body compartment. Secondly, in patients A and B who showed emergence of CXCR4-using variants, the X4/CXCR4-using variants, which clustered separately from the R5/CCR5-using variants in the phylogenetic tree, were detected in patient PBMC-DNA and CV but not in serum-RNA. The segregation between sequences from serum and PBMC-DNA and/or CV after the emergence of CXCR4-using variants in these patients could be attributed to the absence of X4 variants in serum. Finally, a low degree of intermingling of viral sequences from PBMC-DNA with sequences of the other two sources was observed at time points where PBMC-DNA sequences were very homogeneous. This could either reflect low sequence diversity in PBMC-DNA at specific time points or limitations in the detection of minor variants.

In 3 out of 4 patients (A, C and D), we found evidence for frequent gene trafficking between serum-RNA, PBMC-DNA and CV, albeit that compartmentalization between PBMC-DNA and either serum-RNA or CV was borderline significant for patients A and C, which coincided with a certain degree of segregation between some of the sequences from PBMC-DNA and the other sources in the phylogenetic trees. Overall however, our data support that in each of these patients, sequences from the three sources belonged to the same viral population. In patient B, we detected a restriction of gene trafficking between serum-RNA and both PBMC-DNA and CV. This restriction is probably the result of negative selection of most of the early viruses in serum, in combination with the differential evolution of viral coreceptor use in serum and PBMC. This implies that the cell-free (serum) and cell-associated (PBMC) viral quasispecies can occasionally be different and therefore not always reflecting the same fraction of viruses present in peripheral blood at a certain time point.

Previous studies have shown the effect of PBMC cultures for bulk virus isolation on the composition of virus populations. While some argued that a minor HIV-1 variant present in vivo or a variant not even detected in the patient could dominate the co-culture in vitro, others found that the major variant present in uncultured PBMC was the one persisting in the culture. Indeed, bulk virus isolation in vitro seems to select for one or few HIV-1 variants that have optimal fitness for replication in PBMC in vitro. Although the clonal virus isolation procedure allows for the isolation of multiple variants, there is still the concern that only a fraction of the virus variants originally present in vivo may be selected and that novel mutations may occur during culturing of the virus. We here show that CV exhibit a sequence variability similar to contemporaneous viral populations obtained from viral RNA in serum and PBMC proviral DNA. When diversity between the three different
sources differed, this was not necessarily due to a higher or lower heterogeneity of sequences from CV. Moreover, significant divergence was found between CV and the viral sequences derived from serum-RNA and PBMC-DNA, but comparable differences in divergence were also observed between serum-RNA and PBMC-DNA. Therefore we conclude that CV are not a specific selection from the virus pool in blood at the moment of sampling and that the sequence variation that may have been introduced during culture in CV is limited. We are aware that with the population sequencing methodology not all viral variants present in the sample may have been detected and that we therefore may have underestimated the actual sequence variability in PBMC-DNA and serum-RNA. However, our approach at least demonstrates a similar composition of the CV and the dominant virus populations in serum-RNA and PBMC-DNA.

Differences in the genetic composition of virus populations from serum-RNA and PBMC-DNA have mostly been reported for patients receiving HAART. Drug-resistance mutations can be detected in plasma before their occurrence in PBMC. More importantly, the recovery of replication-competent wild-type HIV from PBMC despite prolonged suppression of plasma viremia, suggested a reservoir of archived viruses in PBMC. In our present study, only a few sequences from PBMC-DNA and CV, and interestingly in two patients also from serum-RNA, were identified as potentially archived viruses. This suggests that in the absence of therapy, PBMC proviral DNA reflects mostly the actively replicating virus population and that CV can be considered an accurate reflection of the replication competent viruses at a given moment, not excluding that a low percentage of archived latent viruses may be recovered with the clonal virus isolation procedure.

Our observation that CXCR4-using variants may be more common in PBMC than in serum-RNA is in agreement with a previous cross-sectional study from Verhofstede et al. We confirm and extend those observations by showing in longitudinal samples that X4 variants were not always detected in serum when only low numbers of CXCR4-using CV were obtained. We have previously reported that early CXCR4-using HIV-1 variants have a higher sensitivity to antibody neutralization than their co-existing CCR5-using variants and several studies have suggested that HIV-specific immune responses may promote the preferential survival of CCR5-using strains. This may suggest that a cell-free state of CXCR4-using viruses is incompatible with neutralizing humoral immunity, which could explain their absence or underrepresentation in serum and their restriction to cell-to-cell spread in PBMC. As the contribution of CXCR4-using HIV-1 to the total virus population may increase over time, the equal distribution of X4 sequences in serum-RNA and CV in individuals with higher numbers of CXCR4-using variants may reflect the selection of neutralization resistant CXCR4-using variants that can persist in plasma.

The major advantage of working with replication competent CV is that biological properties of the virus can be studied in the context of the original genetic background and the complete viral genome, which obviously is not the case with cloned viral gene fragments.
from plasma in the background of a molecular HIV-1 clone. Our present study shows that clonal HIV-1 variants isolated from PBMC may equally represent the viral quasispecies in blood as sequences obtained from serum and PBMC proviral DNA. However, certain selective forces may drive differential evolution of the cell-free and cell-associated virus pool, in which case, sequences from both sources would be ideally required to obtain a more complete picture of the viral quasispecies in peripheral blood *in vivo*.

**ACKNOWLEDGEMENTS**

The Amsterdam Cohort Studies on HIV infection and AIDS, a collaboration between the Amsterdam Health Service, the Academic Medical Center of the University of Amsterdam, Sanquin Blood Supply Foundation, the University Medical Center Utrecht, and the Jan van Goyen Clinic are part of the Netherlands HIV Monitoring Foundation and financially supported by the Center for Infectious Disease Control of the Netherlands National Institute for Public Health and the Environment. We thank Philippe Lemey for discussion and help in the interpretation of the phylogenetic analysis. We thank Laura Heath for assistance with running the compartmentalization tests. We thank Jim Mullins for stimulating discussion. We thank Georgios Pollakis and William A. Paxton for helpful discussion and critical reading of the manuscript. We are indebted to all participants for their continuous participation in the study.

This research was financially supported by the Landsteiner Foundation for Blood Transfusion Research (LSBR grant 0317) and The European Community’s Seventh Framework Programme NGIN (FP7/2007-2013) under grant agreement nº 201433.

**REFERENCES**


**Supplementary material**

*Primers used for sequencing of gp160 env from clonal HIV-1 variants*

- Seq2 (rev) (5’-TCTCTCATATCTCCTCCACAGGT-3’),
- Seq3 (fw) (5’-TATGGGATCAAAGCCTAAAGGCGATG-3’),
- Seq4 (rev) (5’-CTTGTATTGTTGTTGGGTCTTGTAC-3’),
- Seq5 (fw) (5’-GTCAACTCAACTGCTGTAAATGCG-3’),
- Seq6 (rev) (5’-ATCTAATTTGTCCACCTGATGCGGAGG-3’),
- env9 (rev) (5’-ACAGGGGTTGTAATGACTGA-3’),
- env1aTOPO (fw) (5’-CAACGGTTAGCGCATCTTCTATGGCAGGAAGAA-3’),
- PSC (fw) (5’-CCCTCAGGAGGGGACCAG-3’),
- PSH (rev) (5’-CCATAGTGCTTCCTGCTGCT-3’),
- Han15 (fw) (5’-ACCAAGGCAAGGAGAGGATGG-3’),
- env5 (rev) (5’-GGTGAATATCCCTGCCTAACTCTA-3’),
- envM (fw) (5’-CCACAAACTTGCCACTTCTCA-3’),
- Han16 (rev) (5’-TTCATTTTCTCCCTTACAGTGAGG-3’).

*Primers used for sequencing of gp160 env PCR products from serum and PBMC*

- M13F (fw) (5’-GTAAAACGACGGCAG-3’), M13R (rev) (5’-CAGGAAACAGCTATGAC-3’),
- Senv5 (fw) (5’-GGTAGCTGTGTGAAAG-3’), F112 (fw) (5’-CAGTACAATGYACATGGRAT-3’)
- EnvSeqF (fw) (5’-TTCAGACCTGGAGGAGGARATATGA-3’).

*HIV-1 RNA isolation from serum, cDNA synthesis, cloning and sequencing of gp120 (C2-C4) env PCR products for Patients E, F, G and H*

Viral RNA was isolated from 140µl serum using the QIAgen Viral RNA Mini Kit. Isolated RNA was eluted in a final volume of 50µl. Viral RNA (10µl) was reverse transcribed into cDNA with Superscript II RnaseH Reverse Transcriptase (Invitrogen).

From the synthesized cDNA of patients E, F, G and H the gp120 env region (C2 to C4: corresponding to HxB2 envelope nucleotide positions 811 to 1290) was amplified in a outer PCR (10 independent PCRs per sample) with primers seq2 (5’-TCCCTCATATCTCCTCCACAGGT-3’) and seq3 (5’-TATGGGATCAAAGCCTAAAGGCGATG-3’) and a nested PCR with primers seq5 (5’-GTCAACTCAACTGCTGTAAATGCG-3’) and seq6 (5’-ATCTAATTTGTCCACCTGATGCGGAGG-3’) (outer and nested PCR temperature program: 97°C, 5’, 35 cycles of 97°C 45”, 50°C 45”, 62°C 90”, 62°C 6’, 4°C indefinite).

Multiple bulk PCR products resulting from serum RNA were cloned in the pGEM-Teasy Vector system (Promega) and transformed into DH5α competent cells (invitrogen). A maximum of 2 clones obtained from one independent PCR were picked, gp120 env C2-C4 region was amplified with pGEM-Teasy Vector specific primers and PCR products were purified using EXOSAP-IT (USB) and sequenced using the nested PCR primers and the ABI prism Big Dye Terminator v1.1 Cyclesequencing Kit (Applied Biosystems) according
HIV-1 evolution in PBMC and plasma

to the manufacturer’s protocol. Sequencing conditions were 5’ at 94˚C, 30 cycles of 15” at
94˚C, 10” at 50˚C, 2” at 60˚C and a 10’ extension at 60˚C. Sequences were analyzed on the
Applied Biosystems/Hitachi 3130 xl Genetic Analyzer.

DNA isolation, PCR amplification and sequencing of gp120 (C2-C4) env from clonal HIV-1
variants isolated from Patients E, F, G and H

Total DNA was isolated from 1x10⁶ healthy donor PBMC in vitro infected with clonal HIV-1
isolates, using a modified L6 isolation method. For each clonal HIV-1 variant, the gp120
ev region (C2 to C4: corresponding to HxB2 envelope nucleotide positions 811 to 1290)
was amplified and sequenced as described for the serum samples of those patients with the
omission of the cloning steps.

Compartmentalization tests

Five methods (Slatkin-Maddison, Simmonds Association Index, Correlation Coefficients,
Wright’s measure of population subdivision (FST), Nearest-neighbor statistic) were used to
determine compartmentalization between sequences from serum, PBMC proviral DNA and
clonal HIV-1 variants isolated from PBMCs. Methods and parameters implemented in Hyphy for each method are described below:

1) Slatkin-Maddison (SM) determines the minimum number of migration events between
the separated populations based on the tree topology. Statistical support is based on the
number of migration events that would be expected in a randomly structured population,
derived by permuting sequences between compartments. 1000 permutations were used in
our analysis.

2) Simmonds Association Index (AI) assesses the degree of population structure by
weighting the contribution of each internal node based on its depth in the tree. 10 relabelings
per sample and 1000 tree bootstrap samples were used in our analysis.

3) Correlation coefficients (r and r̄) correlate distances between two sequences in a
phylogenetic tree with the information about whether or not they were isolated from
the same compartment. The distance between sequences can be either the number of
tree branches separating the sequences (r̄) or the cumulative genetic distance between
sequences (r). To assess whether the computed coefficient was statistically significant, we
estimated the distribution of these coefficients by permuting 1000 times sequences between
compartments and 10 relabelings per samples were used.

4) Wright’s measure of population subdivision (FST) compares the mean pairwise genetic
distance between two sequences sampled from different compartments to the mean distance
between sequences samples from the same compartment. Statistical significance is derived
via a population-structure randomization test. This score is calculated using two estimates
of FST and an estimate of KST. Distance matrices were calculated under the HKY85
substitution model.
5) Nearest Neighbor Statistic (Snn) measures how often the nearest neighbors of each sequence were isolated from the same or different compartments. Distance matrices were calculated under the HKY85 substitution model.

REFERENCES


Supplementary Figure S6.1

- Patient C
- Patient D
- Patient E
- Patient F
- Patient G
- Patient H
- Patient A
- Patient B
- Patient C

CD4+ T cells/ul serum
RNA load (copies/ml)
Zidovudine (AZT)
Stavudine (d4T) Lamivudine (3TC)
Time point of sampling
AIDS diagnosis
Zalcitabine (ddC)
Sanquinavir / Invirase (HGC)
Didanosine (ddI)
Indinavir
Estimated date of emergence of CXCR4-using viruses
### Supplementary Table S6.1: Patients, time points and number of sequences analyzed

<table>
<thead>
<tr>
<th>Patient cohort ID</th>
<th>Time point</th>
<th>Sampling time after SC or imputed SC date (months)</th>
<th>Time to X4 emergence (months)</th>
<th>Number of sequences analyzed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R5</td>
<td>X4</td>
<td>Total</td>
</tr>
<tr>
<td>A 19642</td>
<td>1</td>
<td>-107</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-87</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-62</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>-28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B 19554</td>
<td>1</td>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C 19298</td>
<td>1</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>48</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>72</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D 18969</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>46</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>67</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E 19829</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>51</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>66</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F 18869</td>
<td>1</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G 19296</td>
<td>1</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>61</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H 19308</td>
<td>1</td>
<td>26</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>29.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>37</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>41</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>53</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>59</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* time to emergence of CXCR4-using viruses (estimated date between last negative and first positive MT-2 test).

b sequences obtained from viral RNA from serum.

c sequences obtained from clonal HIV-1 variants isolated from patient PBMC.

d sequences obtained from patient PBMC proviral DNA.

R5: number of sequences with predicted and/or experimentally tested CCR5 use.

X4: number of sequences with predicted and/or experimentally tested CXCR4 use.
### Supplementary table S6.2: Comparison of mean gp160 env pairwise genetic distances within (diversity) and between (divergence) sequences from viral RNA in serum, PBMC proviral DNA and clonal HIV-1 variants per patient

<table>
<thead>
<tr>
<th></th>
<th>Patient A</th>
<th>Patient B</th>
<th>Patient C</th>
<th>Patient D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>freq.</td>
<td>%</td>
<td>freq.</td>
<td>%</td>
</tr>
<tr>
<td>Diversity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE = CV</td>
<td>3/5</td>
<td>60</td>
<td>2/4</td>
<td>50</td>
</tr>
<tr>
<td>SE &gt; CV</td>
<td>0/5</td>
<td>0</td>
<td>1/4</td>
<td>25</td>
</tr>
<tr>
<td>SE &lt; CV</td>
<td>2/5</td>
<td>40</td>
<td>1/4</td>
<td>25</td>
</tr>
<tr>
<td>CV = PB</td>
<td>2/4</td>
<td>50</td>
<td>3/4</td>
<td>75</td>
</tr>
<tr>
<td>CV &gt; PB</td>
<td>2/4</td>
<td>50</td>
<td>0/4</td>
<td>0</td>
</tr>
<tr>
<td>CV &lt; PB</td>
<td>0/4</td>
<td>0</td>
<td>1/4</td>
<td>25</td>
</tr>
<tr>
<td>SE = PB</td>
<td>1/3</td>
<td>33</td>
<td>2/4</td>
<td>50</td>
</tr>
<tr>
<td>SE &gt; PB</td>
<td>1/3</td>
<td>33</td>
<td>0/4</td>
<td>0</td>
</tr>
<tr>
<td>SE &lt; PB</td>
<td>1/3</td>
<td>33</td>
<td>2/4</td>
<td>50</td>
</tr>
<tr>
<td>Divergence</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEvsCV = SE = CV</td>
<td>3/5</td>
<td>60</td>
<td>0/4</td>
<td>0</td>
</tr>
<tr>
<td>SEvsCV &gt; SE and/or CV</td>
<td>2/5</td>
<td>40</td>
<td>4/4</td>
<td>100</td>
</tr>
<tr>
<td>CVvsPB = CV = PB</td>
<td>2/4</td>
<td>50</td>
<td>4/4</td>
<td>100</td>
</tr>
<tr>
<td>CVvsPB &gt; CV and/or PB</td>
<td>1/4</td>
<td>25</td>
<td>0/4</td>
<td>0</td>
</tr>
<tr>
<td>SEvsPB = SE = PB</td>
<td>1/3</td>
<td>25</td>
<td>1/4</td>
<td>25</td>
</tr>
<tr>
<td>SEvsPB &gt; SE and/or PB</td>
<td>2/3</td>
<td>75</td>
<td>3/4</td>
<td>75</td>
</tr>
</tbody>
</table>

SE: diversity in serum-RNA; CV: diversity in clonal HIV-1 variants; PB: diversity in PBMC-DNA; SEvsCV: divergence between serum-RNA and CV; CVvsPB: divergence between CV and PBMC-DNA; SEvsPB: divergence between serum-RNA and PBMC-DNA; = indicates no significant difference between mean pairwise genetic distances; < or > indicate significant differences between mean pairwise genetic distances (P-value < 0.05).

*tp3 was not included because only 1 serum-RNA sequence was available.

*tp5 was not included because only 1 serum-RNA sequence was available.