Published in final edited form as:

Infect Genet Evol. 2010 April ; 10(3): 365–372. doi:10.1016/j.meegid.2009.06.002.

Association between Specific HIV-1 Env Traits and Virologic Control *in vivo*

Beda Joos¹, Philip Rieder¹, Marek Fischer¹, Herbert Kuster¹, Peter Rusert², Alexandra Trkola², Satish K. Pillai³, Joseph K. Wong³, Rainer Weber¹, and Huldrych F. Günthard¹

¹University Hospital Zürich, Department of Medicine, Division of Infectious Diseases and Hospital Epidemiology, Zürich, Switzerland ²Institute of Medical Virology, University of Zürich, Switzerland

³San Francisco Veterans Affairs Medical Center, University of California San Francisco, USA.

Abstract

HIV RNA levels are influenced by genetic characteristics of both the host and the virus. Here we applied machine learning techniques to determine if plasma-derived HIV-1 amino acid sequences can be used to predict spontaneous virologic control. We studied the relationship between HIV-1 *env* genotype and viral load in 20 chronically-infected patients undergoing treatment interruptions (SSITT, Swiss-Spanish Intermittent Treatment Trial) and in 104 primary HIV infected (PHI) patients before antiretroviral therapy (cART) and where applicable also after treatment stop. Extensive longitudinal sampling during the interruptions was performed in nine SSITT patients. Sequences obtained from these nine patients during the first virus rebound were used as a training data set and revealed a strong genetic signature (accuracy 98.6% in cross-validation) associated with control of viremia at levels below 5000 copies/mL of viral RNA maintained for at least two months after the final cART stop. The simple sequence pattern at gp120 positions 268E/358T was confirmed to be predictive of control in the clonal sequences originating from these patients during all subsequent rebounds. Sequences from the remaining 11 SSITT patients with less frequent sampling and from the PHI patients were used for external validation. High sensitivities (71-100%) and negative predictive values (80-100%) but low positive predictive values (12-40%) were achieved in the patient-wise analysis which was based on presence of the genetic pattern in all clones. These results suggest that presence of virus lacking the amino acid pattern 268E/358T is associated with VL >5000 at baseline of PHI and with low probability of spontaneous virologic control after treatment stop. Conversely, however, presence of 268E/358T does not predict control of viremia. These residues in HIV gp120 might affect *in vivo* HIV-1 fitness either at the level of Env function or influence susceptibility to adaptive or innate immune response.

Keywords

HIV-1; envelope gp120; evolution; machine learning

INTRODUCTION

Empirical approaches to development of HIV vaccines have so far, despite huge efforts, not been successful. Among the major obstacles are the enormous genetic variability and the high degree of glycosylation of the outer HIV envelope proteins enabling the virus to escape

Please address correspondence to: Beda Joos University Hospital Zürich, Department of Medicine, Division of Infectious Diseases and Hospital Epidemiology, Rämistrasse 100, CH-8091 Zürich, Switzerland. Phone: ++41 1 255 21 84, Fax: ++41 1 255 32 91, beda.joos@usz.ch..

from immune recognition. Moreover, there is a narrow window of opportunity to clear the initial infection because latency is established within the first weeks, and destruction of CD4⁺ T cells begins early after infection (Johnston and Fauci, 2008). HIV induces a natural immune response involving neutralizing antibodies and cytotoxic T lymphocytes. However, these responses are typically too weak to eliminate the virus and the targeted epitope sequences are continuously mutated to evade the immune system. The envelope glycoproteins gp120 and gp41 comprise the key determinants presenting their surface to the host's humoral immune system. They inevitably undergo continuous evolution driven by positive selection, and are therefore the most variable gene products of HIV. Eventually mutations arise resulting in escape from host immune recognition (Yamamoto and Matano, 2008). These mutations often confer considerable fitness cost to the virus which may subsequently be alleviated by selection of compensatory mutations (Goulder and Watkins, 2004).

To explore inpatient HIV evolution in the region of gp120 we conducted longitudinal clonal sequencing of the envelope C2-V3-C3 domain in HIV-1 infected patients with long-term suppression of plasma viremia who interrupted combination antiretroviral therapy (cART) within a well-controlled structured treatment interruption trial (Swiss Spanish Intermittent Treatment Trial). After successful cART for years, these patients underwent four cycles consisting of two weeks treatment interruption and eight weeks retreatment before cART was completely discontinued for a prolonged time period. During the strictly defined short interruptions HIV RNA usually rebounded within two weeks (Fischer et al., 2003a). Patients who maintained plasma viremia at levels below 5000 copies/mL of viral RNA for at least 2 months after the final treatment stop were classified as "controllers" (Fagard et al., 2003). Analysis of the sequences derived from plasma before the initiation of antiretroviral therapy (cART) revealed that lower viral diversity was associated with spontaneous control of viremia after treatment stop as well as with reduced *in vitro* replication capacity and higher plasma titers of neutralizing antibodies against autologous virus (Joos et al., 2005). Thus, viral properties appeared to influence disease progression being at least in part responsible for spontaneous control of viremia after cessation of cART. We also performed extensive longitudinal clonal studies of plasma HIV RNA to characterize the rebounding virus during the short interruptions and after the prolonged treatment break in a subset of patients (Joos et al., 2008). Phylogenetic analysis exploiting the temporal relation of rebounding virus to pre-treatment sequences was in accord with mono- or oligoclonal reactivation of distinct virus populations from long-lived latently infected cells in contrast to expansion of virus populations replicating at a low-level. Furthermore, viral diversity was restored only very slowly after the final treatment stop. This implies that a sustained evolutionary bottleneck was induced by suppressive cART. In the long term, evolution of the envelope sequences spanning the V3 loop and its flanking regions appeared to be driven by positive selection (Joos et al., 2007). Interestingly the adaptive mutations occurred almost exclusively at solvent exposed amino acid positions on the virion surface which confer a high external accessibility according to the 3D structure of the envelope gp120 (Huang et al., 2005) suggesting that these sites evolved in response to selection pressure induced by neutralizing antibodies or adaptations in the receptor interaction necessitated by shifting cellular tropism. However, in this analysis we detected widely different patterns of positively selected sites within the C2-C3 region of gp120 in individual patients by codon-based likelihood methods (Kosakovsky Pond and Frost, 2005).

Here we aimed at identifying common genetic characteristics of rebounding viruses across twenty chronically HIV-1 infected patients having undergone multiple planned treatment interruptions. We employed machine learning techniques to determine if a shared genetic signature could be found in plasma virus rebounding during the first interruption after years of continuous suppressive therapy. Next, we searched for specific signatures associated with

high or low distance from the most recent common ancestor (MRCA) in the phylogeny which would reflect virus populations replicating at different speed. In addition, we sought to determine if plasma-derived HIV amino acid sequences could be used to predict spontaneous control of viremia. We found evidence of a link between HIV Env sequence and viral load by decision tree learning and aimed to validate this genetic signature by analyzing amino acid sequences originating from different patient groups.

MATERIALS AND METHODS

Study Subjects

Blood samples were obtained from chronically-infected patients enrolled into the Swiss-Spanish Intermittent Treatment Trial (SSITT) at the University Hospital Zürich. Two patients were excluded because they did not complete the study. We selected those patients from whom a pre-treatment plasma sample was available provided that PCR amplification was successful. All patients had long-term combination antiretroviral therapy (cART) and undetectable viral loads (<50 copies/mL) for at least six months before study entry and then underwent four consecutive STI cycles (2 weeks off and 8 weeks on treatment), followed by a longer fifth treatment interruption of indefinite duration. Patients were classified into “controlling” and “noncontrolling” groups. Control of viremia was defined as spontaneous maintenance of viral RNA <5000 copies /mL for at least 2 months during the first 4 months after the final cART stop (Trkola et al., 2003). Twenty patients were eligible for the current analysis, nine of them were sampled intensively during the interruptions. The salient patient characteristics (Joos et al., 2007; Joos et al., 2005; Trkola et al., 2004) and results of the clinical trial (Fagard et al., 2003; Fischer et al., 2003a; Fischer et al., 2004a; Fischer et al., 2004b; Fischer et al., 2003b; Metzner et al., 2003; Oxenius et al., 2002a; Oxenius et al., 2002b) have been reported previously.

We also included 104 primary HIV infected (PHI) patients enrolled in the Zurich PHI Study (<http://clinicaltrials.gov>). All had well-documented acute or recent HIV infection presumably within the last 3 months based on the results of plasma HIV-1 RNA, HIV p24 antigen determination, Western blot, HIV-gp120 avidity assay and/or clinical symptoms of acute retroviral syndrome. The studied subjects include 92 men and 12 women, the vast majority infected with HIV subtype B (77%), CRF01_AE (10%), A (3%), and some with C, F1, G, CRF02_AG, and CRF12_BF. Plasma samples were analyzed as part of an ongoing study for all subjects at the time of enrolment before antiretroviral therapy was started and where applicable also later on when cART was stopped after a median of 18 (range 7-30) months in a subset of 28 patients.

Both studies were approved by the Ethics Committee of the University Hospital Zurich and written informed consent was obtained from all patients.

RNA extraction, RT-PCR, cloning and sequencing

Extraction of viral RNA from plasma and subsequent PCR amplification, molecular cloning and bidirectional sequencing of the C2-V3-C3 region of HIV-1 *env* was performed as described (Joos et al., 2005). By extracting HIV RNA from a total of 1 ml plasma our RT-PCR strategy was designed to maximize the probability that the sequenced clones are representative of the actual viral population present in plasma *in vivo*. The method was evaluated by testing the misincorporation rate during the RT step, the fidelity of the polymerase and reproducibility in analyzing identical patient samples (Joos et al., 2007). Representativity of the isolated viral quasispecies was confirmed by comparison with limiting dilution analysis. Similar quasispecies compositions were found in few samples tested in parallel by molecular cloning and by single genome analysis. This indicates that the

used cloning method was not excessively affected by resampling bias (Joos et al., 2008). Approximately 16 individual clones were obtained at each time point. Overall 1753 clonal sequences were obtained from a total of 109 longitudinal samples across 20 SSITT patients (31-190 clones per patient) and 1610 pre-treatment clones were isolated from 104 PHI patients.

Phylogenetic inference

Sequences were edited and aligned using Lasergene software version 5.08 (DNASTAR Inc., Madison, WI) and manually adjusted in the correct reading frame to the following GenBank sequences: HIV-1 subtype B isolates HXB2 (accession number K03455), ADA (M60472), YU2 (M93258), BAL (M68893), JRFL (U63632), NL4-3 (U26942), subtype A isolate U455 (M62320), subtype C isolate ETH2220 (U46016), subtype D isolate ELI (K03454), subtype F isolate BR020 (AF005494), and the HIV group M envelope fragment ZR59a-2 (AF030527) isolated from a historical plasma sample collected in 1959 (Zhu et al., 1998). Neighbor-joining phylogenetic trees were inferred by MEGA version 4 (Tamura et al., 2007) using genetic distance estimates obtained from Tamura-Nei 6-parameter model. Bootstrap analysis was performed using 1000 replicates. Maximum likelihood phylogenetic analyses were performed by DNAML on a parallel computing cluster using randomized input order, global rearrangements, and multiple jumble options (PHYLIP Phylogeny Inference Package version 3.6 distributed by J. Felsenstein, Department of Genetics, University of Washington, Seattle). The Slatkin-Maddison test (Slatkin and Maddison, 1989), a cladistic measure of inter-population gene flow, was used to evaluate the degree of viral population shift between the different timepoints in the longitudinal analysis of clonal sequences from the 20 SSITT patients. Screening for intrapatient recombination was performed by GARD (Genetic Algorithm for Recombination Detection), using the single breakpoint detection method (Kosakovsky Pond et al., 2006). These analyses showed no evidence of recombination in the sequences from any individual in this study. No putative breakpoints could be identified, and a single phylogeny rather than multiple segment-specific phylogenies best described the data in all cases (AIC scores were not improved with multiple trees).

GenBank accession numbers

The reported sequences were deposited in GenBank under accession numbers EF424791 to EF426096, DQ002058 to DQ002345, AY656534 to AY656549, AY656440 to AY656471, AY656249 to AY656312, AY375663 to AY375678, AY375616 to AY375630, and AY375568 to AY375583.

Coreceptor usage prediction

Predictions of coreceptor use were made by the WetCat algorithm (Pillai et al., 2003) and by use of position-specific scoring matrices trained on the syncytium-inducing phenotype (PSSM SI/NSI) and on the X4 or R5 coreceptor phenotype (PSSM X4/R5) (Jensen et al., 2003). The former method applies a support vector machine-based learning technique to classify the V3 loop amino acid sequences of the envelope glycoprotein according to sequences from viruses with known coreceptor phenotype. The latter derives a score by using precalculated matrices for the 35 amino acids forming the V3 loop. Both methods yielded comparable results. CXCR4-using virus was sporadically detected in the longitudinal sequences but in the combined analysis no clear switch in coreceptor usage was found.

Glycosylation sites

Potential N-linked glycosylation sites were assessed with the N-GlycoSite tool provided by Los Alamos HIV Sequence database (Zhang et al., 2004).

Positive selection detection

Identification of positively selected amino acid positions was performed by Single Likelihood Ancestor Counting, Fixed Effects Likelihood, and Random Effects Likelihood methods using the HyPhy software package (Kosakovsky Pond and Frost, 2005). Duplicate sequences were removed and the optimal time-reversible substitution model was determined for each patient before the analysis.

Machine learning

Open source machine learning software Weka 3.5 (Waikato environment for knowledge analysis) was used to develop decision trees by the J48 algorithm (Frank et al., 2004). The minimum number of objects per leaf node was set at high values in all classification experiments to confine the complexity of the models and reduce overfitting bias (since total numbers of sequences varied the best fitting models with minNumObj values of 10 to 20 were selected). Prediction performance was evaluated by stratified 10-fold cross-validation method. The sensitivity (recall) is defined as $TP/(TP+FN)$, specificity (precision) by $TP/(TP+FP)$, where TP is the number of the true positives, FN the number of underpredictions and FP the number of overpredictions. The balanced F-score is the weighted harmonic mean of recall and precision given by the formula $F=2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$. Kappa coefficients (observed coincidence-expected coincidence)/([1-expected coincidence]) and the area under the ROC curve (receiver operating characteristic) were also used to measure the performance of the classification algorithms and are also reported.

Statistical analyses

For standard statistical analyses GraphPad Prism version 5 (GraphPad Software, San Diego, CA) was used. Nonparametric tests with Bonferroni adjustment for multiple comparisons were applied for group comparisons (Mann-Whitney) and correlations (Spearman). Two sided Fisher's exact test was used to determine associations between categories. Positive predictive value (PPV) is defined as $TP/(TP+FP)$, negative predictive value (NPV) as $TN/(FN+TN)$ (TP = true positives; FP = false positives, TN = true negatives, FN = false negatives).

RESULTS

To study the association between HIV-1 *env* genotype and viral load in chronically infected patients we took advantage of the Swiss-Spanish Intermittent Treatment Trial (SSITT) and followed 20 patients longitudinally. The first samples were obtained during the chronic phase of infection before any treatment was started. Subsequently patients received combination antiretroviral therapy for a median of 2.7 (range 1.5 - 3.4) years and then underwent 4 treatment interruption cycles (2 weeks interruption and 8 weeks retreatment) followed by cART stop. Clonal HIV-1 *env* sequences encompassing the C2-V3-C3 region of gp120 were isolated from pre-treatment plasma and from the last available sample following treatment stop. 9 of the 20 patients also had intense sampling during the interruptions. Overall, 1753 clonal sequences were derived from 109 samples collected over 6.5 (range 3.3 - 8.6) years (Joos et al., 2008; Joos et al., 2007). A Neighbor-joining phylogenetic tree including all sequences failed to show any evidence of cross contamination between study subjects or contamination with laboratory strains (Figure 1). Phylogenetic trees were also constructed with the sequences of each patient alone by using much more computer intensive maximum likelihood method. This allowed us to infer the individual most recent

common ancestral sequences (MRCA) and to explore the structure of the phylogenetic trees in regard to sampling time.

The relationship between sampling time and viral population diversity was determined in the 9 patients analyzed before treatment as well as during the four structured treatment interruptions and sequentially thereafter (Figure 2). Relatively homogenous virus populations emerged during the short 2 week interruptions (time 0 - 0.8 years) but at different time points often phylogenetically distinct virus populations appeared. After treatment cessation, diversity increased gradually with a substantial delay. On average the level of pre-treatment diversity was not achieved until 2.5 years after beginning of the study, when appreciable divergence from the virus populations observed before treatment was also apparent (Figure 3). Evolution of the genetic distances to the MRCA in the 9 frequently sampled patients is shown in Figure 4. While the average branch lengths to the MRCA remained relatively stable over the observed time period, transient drops were observed during STI in all individuals. In comparison with the genetic distance to the MRCA at pretreatment, marked reductions by >1% occurred in 7/9 patients (range 1.1 to 2.7%) suggesting rebound of archived viral variants which were already present generations before the initiation of cART or even at the time of infection. Increasing branch lengths to the MRCA indicating ongoing evolution were observed in the majority of patients after the final treatment stop.

Genetic characteristics of rebounding virus

The distinct phylogenies and low diversities observed in plasma virus rebounding during the first treatment interruption after years of continuous suppressive therapy suggested potential reservoirs as the source of viral rebound. We posited that such virus might share some compartment-specific genetic signatures and features that could be obscured by interpatient sequence differences as has been shown for CSF virus and virus in genital secretions (Pillai et al., 2005; Pillai et al., 2006). To explore this further, we employed machine learning techniques to search for a specific signature associated with first rebound among the clonal sequences originating from the 9 patients with frequent longitudinal sampling.

Initially we compared the C2-V3-C3 sequences (126 amino acids) obtained during the first treatment interruption with those collected at pre-treatment, at later structured interruptions and during the last prolonged treatment interruption. In addition we also compared the first rebound sequences with those from all 20 patients at all time points. Testing showed no evidence of a genetic signature shared by the amino acid sequences derived from the first virus rebounds in plasma. Sensitivities of the predictions ranged from 0.357 to 0.736 and specificities from 0.71 to 0.893 resulting in balanced F-scores of 0.51 to 0.747. Kappa coefficients (chance-corrected proportional agreement) ranged from 0.488 to 0.641. Next, we applied decision tree learning to classify the sequences collected during the first STI on the basis of their distance to the MRCA. However, no clear genetic signature associated with high or low branch length to the MRCA could be detected.

Strong evidence for a genetic signature was found when the sequences derived from rebounding virus in the 9 patients with frequent sampling were classified by control of viremia (defined by maintenance of viral loads <5000 RNA copies/mL for at least two months) after treatment stop (Kappa coefficients >0.95, Table 1). Somewhat different classification models were inferred by using the sequences isolated during the 1st rebound alone (Figure 5A) and those from the 1st to 5th rebounds (Figure 5B) as training data sets. Interestingly, the common feature consisting of just glutamic acid at amino acid position 268 in combination with threonine at position 358 (268E/358T) was present in 647 out of 715 sequences obtained during all time points from controlling patients as opposed to only 63 of 687 sequences in non-controllers (Table 2, Fisher's exact $p < 0.0001$, Kappa coefficient of

agreement 0.813). The accuracy of this association was consequently even slightly improved when the entire classification model (as illustrated in Figure 5B) was linked to control of viremia in these patients (Kappa 0.886; 670/715 versus 35/687 sequences). A more complex, but also highly accurate decision tree was achieved when all envelope sequences from all individuals were tested. The signatures associated with control or no control of viremia primarily comprised the surface exposed amino acid residues at positions 268, 358, 269, 352 and 278 where significant positive selection has partially been detected before (relative surface accessibility $48.0 \pm 18.8\%$, data shown elsewhere (Joos et al., 2007)). Of note, this genetic signature found in patients spontaneously controlling plasma viremia was already present in the pre-treatment sequences and thus was not attributable to treatment interruptions.

Association of genetic signature with control of viremia

The depicted exploratory analysis of sequences isolated from the 9 frequently sampled SSITT patients revealed that presence of the simple pattern 268E/358T in HIV envelope protein gp120 could be strongly associated with virologic control. Next we tried to validate this finding rigorously in different populations. For this purpose we evaluated clonal sequences from the remaining 11 SSITT patients with less frequent sampling and from a group of well-characterized primary HIV infected patients. These analyses were done by categorizing individual patients upon presence of 268E/358T in all clones isolated before treatment was started. We chose this restrictive condition under the assumption that a minority of fitter viruses might eventually overgrow the "beneficial" variants which presumably have reduced replication capacity.

While the predictive performance of the genetic pattern was naturally in the clonal analyses of the training data sets (Table 3, A - C), a different picture was found in the patient-wise analyses (Table 3, D - G). The very strong positive predictive values of the signature 268E/358T associated with viral control in the 9 frequently sampled SSITT patients were not confirmed in the 11 SSITT patients with less frequent sampling (Table 3, D). However, high negative predictive values were retained in this group as well as in the combined analysis of all chronically infected patients of the SSITT study (Table 3, E). Basically the same result, very high sensitivity and negative predictive value, was achieved by analyzing the sequences isolated from the first plasma samples of 104 primary HIV infected patients before treatment (Table 3, F). Participating in an ongoing study, most of the PHI patients currently remain under combined antiretroviral therapy. In 28 of them treatment was meanwhile stopped and viremia was contained at levels below 5000 copies/mL in some patients. Acceptable sensitivity and negative predictive value but low positive predictive value was also obtained in this group (Table 3, G). Interestingly, raising the cut-off for virologic failure from 5000 to 10000 RNA copies/mL did practically not influence the result of the analysis. 5000 was chosen here in order to be consistent with the definitions of the SSITT study.

The signature 268E/358T is present in 50% of the Env sequences which are collected in the Los Alamos National Laboratory HIV Sequence Database and represent the full spectrum of worldwide HIV-1 and SIVcpz sequences (year 2007 web alignment, 1766 sequences, <http://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html>). A fraction of 73% was found among the currently available sequences of subtype B viruses, in particular 71% and 79% of those with predicted CCR5 and CXCR4 coreceptor usage, respectively.

DISCUSSION

It is well accepted that the replicative capacity of HIV *in vivo* is influenced by both genetic characteristics of the host and those of the virus. Interactions between virus and the host's immune system are complex and better understanding of the underlying mechanisms is

important for planning optimized treatment strategies and for vaccine development. In a comprehensive longitudinal study we examined the virus characteristics of HIV envelope in a well-defined group of chronically HIV infected patients undergoing structured treatment interruptions. Patients were enrolled in the Swiss HIV-cohort study for years and were studied prior to cART, during continuous cART and subsequently during a structured treatment interruption trial (Swiss Spanish Intermittent Therapy Trial) (Fagard et al., 2003; Fischer et al., 2003a; Fischer et al., 2004a; Joos et al., 2007; Joos et al., 2005; Oxenius et al., 2002a; Oxenius et al., 2002b; Trkola et al., 2003). Here, we specifically attempted to find a common pattern in the phylogeny of the rebounding virus, to relate HIV-1 *env* genotype with HIV RNA levels and to test whether spontaneous virologic control after cessation of cART can be predicted from plasma-derived amino acid sequences.

Remarkable features could be deduced from the phylogenetic analyses (Joos et al., 2008). During the short treatment interruptions distinct mono-, or oligoclonal virus rebounded but no significant evolution with regard to the MRCA was detected between the investigated time points. This indicates random reactivation of archived virus populations from latently infected cells without any significant residual replication in patients receiving potent cART. Unexpectedly, after the final treatment stop it took 1 to 2 years until the virus diversity reached the extent from before treatment. This long lasting effect shows that treatment introduced a strong bottleneck in viral evolution. Total divergence rate from the first to the last time points, averaged over the drug free periods only, was about 1% per year which is in agreement with previous reports (Shankarappa et al., 1999).

By applying machine learning techniques we further explored whether specific sequence patterns of HIV-1 *env* are associated with spontaneous virologic control. A decision tree inducer revealed in the sequences derived from the first virus rebound in plasma a strong genetic signature associated with maintenance of viremia at low levels after treatment stop. The signature was very simple and comprised two externally accessible amino acid sites located in the N-terminal and C-terminal flanking regions adjacent to the V3 loop (Figure 6). The pattern at gp120 positions 268E/358T was also found in the clonal sequences isolated during all subsequent rebounds. However, since all sequences originated from the same patients, this could also be due to overfitting. Indeed, attempts to verify the predictive capacity of the genetic pattern by external validation in a different group of SSITT patients and in primary HIV infected patients failed because specificity was lost. However, high sensitivity and negative predictive value were retained even in the large group of 104 PHI patients. This implies that presence of virus lacking the amino acid pattern 268E/358T was clearly associated with high viral loads at baseline and after treatment interruptions in both primary and chronically HIV infected patients. Conversely, however, presence of 268E/358T was not able to predict spontaneous control of viremia after treatment stop. It is conceivable that mutations at these two positions in HIV envelope gp120 affect *in vivo* fitness of the virus either at the level of Env function or by influencing susceptibility to adaptive or innate immune control.

A potential limitation of this study is that only a fragment of gp120 and not full length *env* sequences were analyzed. It is not clear whether the two identified positions alone contribute to differential replication capacity since the occurrence of compensatory mutations due to changes elsewhere in the envelope can not be excluded. Additionally, the influence of host genetic polymorphisms and HLA types on the virologic outcomes could also not be assessed. Overall, the relatively small number of patients in this study may limit the power of our analysis and prohibit generalization.

The key finding of our analyses was that patients harbouring virus variants differing from gp120 268E/358T may have a lower probability of spontaneous virologic control after

treatment stop. Intrinsic viral characteristics were, at least in part, responsible for reduced *in vivo* replication capacity. Whether incorporation of baseline genotypic analysis of HIV envelope will help in allocating different treatment options in the future e.g. in patients who are not willing to continue lifelong therapy needs to be assessed in further trials.

Acknowledgments

The study was supported by the Swiss National Science Foundation (SNF) (grant no. 3345-65168 and 3100AO-103748/1 to H.F.G. and grant no 320000-116035 to H.F.G.), by the Swiss HIV Cohort Study (SNF grant no. 33CSC0-10878), SHCS project no. 290 and a grant from the EMDO Foundation (both to H.F.G.), and a research grant from the Kanton of Zürich. The authors are grateful for the training received at the 14th International Bioinformatics Workshop on Virus Evolution and Molecular Epidemiology, September 2008, Cape Town. We thank the participating patients for their commitment, Christine Schneider, Christina Grube and Roland Hafner for excellent patient care, Esther Beerli, Friederike Burgener, Christine Leemann, Tuyet Trinh Lu, Barbara Niederöst and Bärbel Sauer for laboratory support and Ingrid Nievergelt and Christine Vögtli for administrative assistance.

REFERENCES

- Fagard C, Oxenius A, Gunthard H, Garcia F, Le Braz M, Mestre G, Battegay M, Furrer H, Vernazza P, Bernasconi E, et al. A prospective trial of structured treatment interruptions in human immunodeficiency virus infection. *Arch Intern Med*. 2003; 163:1220–1226. [PubMed: 12767960]
- Fischer M, Hafner R, Schneider C, Trkola A, Joos B, Joller H, Hirschel B, Weber R, Gunthard HF. HIV RNA in plasma rebounds within days during structured treatment interruptions. *Aids*. 2003a; 17:195–199. [PubMed: 12545079]
- Fischer M, Joos B, Hirschel B, Bleiber G, Weber R, Gunthard HF. Cellular viral rebound after cessation of potent antiretroviral therapy predicted by levels of multiply spliced HIV-1 RNA encoding nef. *J Infect Dis*. 2004a; 190:1979–1988. [PubMed: 15529263]
- Fischer M, Joos B, Wong JK, Ott P, Opravil M, Hirschel B, Weber R, Gunthard HF. Attenuated and nonproductive viral transcription in the lymphatic tissue of HIV-1-infected patients receiving potent antiretroviral therapy. *J Infect Dis*. 2004b; 189:273–285. [PubMed: 14722893]
- Fischer M, Trkola A, Joos B, Hafner R, Joller H, Muesing MA, Kaufman DR, Berli E, Hirschel B, Weber R, Gunthard HF. Shifts in cell-associated HIV-1 RNA but not in episomal HIV-1 DNA correlate with new cycles of HIV-1 infection in vivo. *Antivir Ther*. 2003b; 8:97–104. [PubMed: 12741621]
- Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. *Bioinformatics*. 2004; 20:2479–2481. [PubMed: 15073010]
- Goulder PJ, Watkins DI. HIV and SIV CTL escape: implications for vaccine design. *Nat Rev Immunol*. 2004; 4:630–640. [PubMed: 15286729]
- Huang CC, Tang M, Zhang MY, Majeed S, Montabana E, Stanfield RL, Dimitrov DS, Korber B, Sodroski J, Wilson IA, et al. Structure of a V3-containing HIV-1 gp120 core. *Science*. 2005; 310:1025–1028. [PubMed: 16284180]
- Jensen MA, Li FS, van 't Wout AB, Nickle DC, Shriner D, He HX, McLaughlin S, Shankarappa R, Margolick JB, Mullins JI. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J Virol*. 2003; 77:13376–13388. [PubMed: 14645592]
- Johnston MI, Fauci AS. An HIV Vaccine -- Challenges and Prospects. *N Engl J Med*. 2008; 359:888–890. [PubMed: 18753644]
- Joos B, Fischer M, Kuster H, Pillai SK, Wong JK, Boni J, Hirschel B, Weber R, Trkola A, Gunthard HF. HIV rebounds from latently infected cells, rather than from continuing low-level replication. *Proc Natl Acad Sci U S A*. 2008; 105:16725–16730. [PubMed: 18936487]
- Joos B, Fischer M, Schweizer A, Kuster H, Boni J, Wong JK, Weber R, Trkola A, Gunthard HF. Positive in vivo selection of the HIV-1 envelope protein gp120 occurs at surface-exposed regions. *J Infect Dis*. 2007; 196:313–320. [PubMed: 17570120]
- Joos B, Trkola A, Fischer M, Kuster H, Rusert P, Leemann C, Boni J, Oxenius A, Price DA, Phillips RE, et al. Low human immunodeficiency virus envelope diversity correlates with low in vitro

- replication capacity and predicts spontaneous control of plasma viremia after treatment interruptions. *J Virol.* 2005; 79:9026–9037. [PubMed: 15994796]
- Kosakovsky Pond SL, Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol.* 2005; 22:1208–1222. [PubMed: 15703242]
- Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. GARD: a genetic algorithm for recombination detection. *Bioinformatics.* 2006; 22:3096–3098. [PubMed: 17110367]
- Metzner KJ, Bonhoeffer S, Fischer M, Karanickolas R, Allers K, Joos B, Weber R, Hirschel B, Kostrikis LG, Gunthard HF. Emergence of minor populations of human immunodeficiency virus type 1 carrying the M184V and L90M mutations in subjects undergoing structured treatment interruptions. *J Infect Dis.* 2003; 188:1433–1443. [PubMed: 14624368]
- Oxenius A, McLean AR, Fischer M, Price DA, Dawson SJ, Hafner R, Schneider C, Joller H, Hirschel B, Phillips RE, et al. Human immunodeficiency virus-specific CD8(+) T-cell responses do not predict viral growth and clearance rates during structured intermittent antiretroviral therapy. *J Virol.* 2002a; 76:10169–10176. [PubMed: 12239291]
- Oxenius A, Price DA, Gunthard HF, Dawson SJ, Fagard C, Perrin L, Fischer M, Weber R, Plana M, Garcia F, et al. Stimulation of HIV-specific cellular immunity by structured treatment interruption fails to enhance viral control in chronic HIV infection. *Proc Natl Acad Sci U S A.* 2002b; 99:13747–13752. [PubMed: 12370434]
- Pillai S, Good B, Richman D, Corbeil J. A new perspective on V3 phenotype prediction. *AIDS Res Hum Retroviruses.* 2003; 19:145–149. [PubMed: 12643277]
- Pillai SK, Good B, Pond SK, Wong JK, Strain MC, Richman DD, Smith DM. Semen-specific genetic characteristics of human immunodeficiency virus type 1 env. *J Virol.* 2005; 79:1734–1742. [PubMed: 15650198]
- Pillai SK, Pond SL, Liu Y, Good BM, Strain MC, Ellis RJ, Letendre S, Smith DM, Gunthard HF, Grant I, et al. Genetic attributes of cerebrospinal fluid-derived HIV-1 env. *Brain.* 2006; 129:1872–1883. [PubMed: 16735456]
- Shankarappa R, Margolick JB, Gange SJ, Rodrigo AG, Upchurch D, Farzadegan H, Gupta P, Rinaldo CR, Learn GH, He X, et al. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol.* 1999; 73:10489–10502. [PubMed: 10559367]
- Slatkin M, Maddison WP. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics.* 1989; 123:603–613. [PubMed: 2599370]
- Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* 2007; 24:1596–1599. [PubMed: 17488738]
- Trkola A, Kuster H, Leemann C, Oxenius A, Fagard C, Furrer H, Battegay M, Vernazza P, Bernasconi E, Weber R, et al. Humoral immunity to HIV-1: kinetics of antibody responses in chronic infection reflects capacity of immune system to improve viral set point. *Blood.* 2004; 104:1784–1792. [PubMed: 15187026]
- Trkola A, Kuster H, Leemann C, Ruprecht C, Joos B, Telenti A, Hirschel B, Weber R, Bonhoeffer S, Gunthard HF. Human immunodeficiency virus type 1 fitness is a determining factor in viral rebound and set point in chronic infection. *J Virol.* 2003; 77:13146–13155. [PubMed: 14645571]
- Yamamoto H, Matano T. Anti-HIV adaptive immunity: determinants for viral persistence. *Rev Med Virol.* 2008; 18:293–303. [PubMed: 18416450]
- Zhang M, Gaschen B, Blay W, Foley B, Haigwood N, Kuiken C, Korber B. Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. *Glycobiology.* 2004; 14:1229–1246. [PubMed: 15175256]
- Zhu T, Korber BT, Nahmias AJ, Hooper E, Sharp PM, Ho DD. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature.* 1998; 391:594–597. [PubMed: 9468138]

Figure 1. Phylogenetic Tree

Inferred neighbor-joining phylogenetic tree of clonal HIV-1 *env* C2-V3-C3 sequences derived from 20 patients. Elongated triangles represent the compressed subtrees containing 31-190 clones isolated from plasma of each patient before antiretroviral therapy and after treatment interruptions. The length of the triangle corresponds to the respective intrapatient diversity, thickness is proportional to the number of taxa. The bar denotes 2% nucleotide divergence and diversity (determined using the MEGA software), bootstrap values corresponding to 1000 replications are indicated below the branches.

Figure 2. Diversity

Longitudinal changes in viral diversity of clonal C2-V3-C3 sequences from 9 patients before treatment and following structured treatment interruptions starting at time 0. The dark line indicates the median time course, patients have the same colour coding in Figures 2-4. Distances were estimated by Tamura-Nei 6-parameter model.

Figure 3. Divergence

Longitudinal changes of average genetic distances (net divergence) from pre-treatment C2-V3-C3 sequences in 9 patients following structured treatment interruptions starting at time 0. The dark line indicates the median time course, patients have the same colour coding in Figures 2-4. Distances were estimated by Tamura-Nei model.

Figure 4. Distance from MRCA

Longitudinal changes in average branch lengths to the most recent common ancestor (MRCA) of clonal C2-V3-C3 sequences from 9 patients before treatment and following structured treatment interruptions starting at time 0. The dark line indicates the median time course, patients have the same colour coding in Figures 2-4. Distances were estimated by Tamura-Nei model.

Figure 5. Decision Tree Learning

Genetic characteristics of rebounding virus in 9 patients with frequent longitudinal sampling. Signature associated with control of viremia after treatment stop (yes or no) in C2-V3-C3 sequences from 1st rebound (A, n=140) and those from 1st to 5th rebound (B, n=614). J48 decision tree learning algorithm and stratified 10-fold cross-validation by Weka. Residues are numbered according to HXB2 gp160 positions. Minus sign (-) denotes an amino acid deletion in the sequence. Numbers in parentheses indicate the total number of instances classified by the node and, if a second value is present, how many instances are incorrectly classified. Controlling patients are preferentially characterized by glutamic acid (E) at position 268 in combination with threonine (T) at 358.

Figure 6. Structure of the HIV envelope protein gp120

Ribbon representation of gp120 JR-FL (Protein Data Bank accession code 2B4C) (Huang et al., 2005). The structure was derived from truncated gp120 crystallized in the CD4-bound conformation. The two amino acid residues 268 and 358 are located at solvent-exposed exterior positions in the regions flanking the V3 loop. The V3 loop itself comprises residues 297 – 330). The figure was generated by use of the program PyMOL (available at <http://www.pymol.org>).

Table 1

Env C2-V3-C3 sequence classification based on control of viremia following treatment stop (cross-validation statistics).

Class variable	control of viremia		
Sequences	first rebound	1st-5th rebound	all sequences
Total number of instances	140	614	1753
Correctly classified instances	138 (98.6 %)	607 (98.9 %)	1743 (99.4 %)
Kappa coefficient	0.971	0.977	0.989
Sensitivity	1.000	0.993	0.996
Specificity	0.975	0.983	0.991
F- score	0.987	0.988	0.994
ROC Area	0.997	0.988	0.997

Table 2

Env C2-V3-C3 sequence classification according to control of viremia following treatment stop. Validation of patterns inferred from the training data sets (rebounding virus) in longitudinal samples collected from nine patients before treatment (pre), during 5 structured interruptions (STI), and 1-2 years as well as 2.5 years following treatment stop. Numbers indicate frequencies of matching clonal sequences.

Patient	Control of viremia	Time	268E & 358T		Signature associated with control	
			present	absent	present	absent
106	yes	pre	32	0	32	0
		STI	78	1	79	0
		1-2y	48	0	48	0
		2.5y	31	0	31	0
107	yes	pre	6	10	14	2
		STI	31	1	32	0
		1-2y	12	4	13	3
		2.5y	11	21	11	21
118	yes	pre	16	0	16	0
		STI	80	0	80	0
		1-2y	41	7	41	7
		2.5y	31	1	31	1
123	yes	pre	12	4	10	6
		STI	63	0	63	0
		1-2y	46	2	46	2
		2.5y	24	8	29	3
124	yes	pre	16	0	16	0
		STI	30	0	30	0
		1-2y	32	0	32	0
		2.5y	7	9	16	0
109	no	pre	10	22	10	22
		STI	1	94	1	94
		1-2y	21	2	21	2
		2.5y	0	16	0	16
113	no	pre	0	16	0	16
		STI	0	63	1	62
		1-2y	3	45	0	48
		2.5y	11	21	0	32
120	no	pre	0	16	0	16
		STI	0	94	0	94
		1-2y	0	47	0	47
		2.5y	0	32	0	32
128	no	pre	3	13	0	16
		STI	7	71	2	76
		1-2y	2	45	0	47

Patient	Control of viremia	Time	268E & 358T		Signature associated with control	
			present	absent	present	absent
		2.5y	5	27	0	32
Total with control of viremia:			647	68	670	45
Total without control of viremia:			63	624	35	652

Table 3

Association of genetic signature with control of viremia

Data set	pt (n)	CTR (n) ^a	clones (n)	type of analysis ^b	sensitivity	specificity	PPV ^c	NPV ^c
A SSITT first rebound	9	5	140	cl	0.987	0.937	0.950	0.983
B SSITT all rebounds	9	5	614	cl	0.993	0.976	0.972	0.994
C SSITT all timepoints	9	5	1402	cl	0.905	0.908	0.911	0.902
D SSITT external validation	11	2	176	pt	1.000	0.444	0.286	1.000
E SSITT all patients	20	7	352	pt	0.714	0.615	0.500	0.800
F PHI baseline	104	8	1610	pt	1.000	0.375	0.118	1.000
G PHI plateau	28	8	429	pt	0.750	0.550	0.400	0.846

^a) control of viremia <5000 copies/mL for at least 2 months after cART stop, resp. viral load <5000 (set F)^b) analysis based on individual clones (cl) or patients (pt)^c) PPV = positive predictive value, NPV = negative predictive value.