Development of a homology model for clade A human immunodeficiency virus type 1 gp120 to localize temporal substitutions arising in recently infected women

Mary Poss,1 David C. Holley,2 Roman Biek,1 Harold Cox3 and John Gerdes3

Division of Biological Sciences1, Department of Pharmaceutical Sciences2 and Department of Chemistry3, University of Montana, Missoula, MT 59812, USA

The virus population transmitted by a human immunodeficiency virus type 1 (HIV-1) infected individual undergoes restriction and subsequent diversification in the new host. However, in contrast to men, who have limited virus diversity at seroconversion, there is measurable diversity in viral envelope gene sequences in women infected with clade A HIV-1. In this study, virus sequence diversity in three unrelated, clade A infected women preceding and shortly after seroconversion was evaluated. It was demonstrated that there is measurable evolution of envelope gene sequences over this time interval. Furthermore, in each of the three individuals, amino acid substitutions arose at five or six positions in sequences derived at or shortly after seroconversion relative to sequences obtained from the seronegative sample. Presented here is a model of clade A gp120 to determine the location of substitutions that appeared as the virus population became established in three clade A HIV-1 infected women.

Transmission of human immunodeficiency virus type 1 (HIV-1) from an infected to a naïve host causes significant changes in the genetic make-up of the virus population. HIV-1-infected individuals may harbour up to 10 % virus diversity in the viral envelope gene (env), depending on the length of infection, but only a subset of these variants initiate the new infection. The inoculum transmitted to a recipient contains diverse genotypes (Learn et al., 2002) but in men the variant pool detected close to the time of seroconversion is essentially homogeneous in env (Wolfs et al., 1992; Zhang et al., 1993; Zhu et al., 1993). In contrast, env diversity in clade A-infected women is 1–5 % at seroconversion (Long et al., 2000; Poss et al., 1995). The factors responsible for different profiles of virus diversity in men and women near seroconversion are unknown.

Diversity in the HIV-1 glycoprotein, gp120, encoded by env can significantly impact viral fitness by altering cell tropism and neutralization sensitivity. Substitutions in clade B gp120 that affect receptor and co-receptor binding and antibody interactions have been determined by mutagenesis studies (Boussard et al., 2002; Kwong et al., 2002; Pantophlet et al., 2003; Saphire et al., 2001; Wyatt et al., 1998; Zwick et al., 2003). However, it is not known whether any key functional sites change as the virus establishes infection in a new host. Furthermore, despite the large number of clade A HIV-1-infected individuals worldwide, there has been no research on the structural biology of gp120 of non-clade B viruses. To gain a better understanding of forces acting on gp120 of a colonizing virus population, we developed a homology model of clade A HIV-1 gp120 to locate substitutions that arose between pre- and post-seroconversion samples in each of three clade A-infected women for whom extensive analysis of HIV-1 diversity (Poss et al., 1995), plasma viral load and post-seroconversion env evolution (Poss et al., 1998), and viral phenotype (Painter et al., 2003) have been described.

To determine if there was measurable evolution near the time of infection, we recovered full-length env sequences from pre-seroconversion plasma viral RNA samples from each individual and from samples obtained at 20 (Q23), 7 (Q45) or 10 and 49 weeks (Q47) from the pre-seroconversion sample (Poss et al., 1995). The analysis employed a Markov Chain Monte Carlo (MCMC) framework and a general time reversible substitution model for the Bayesian estimation of evolutionary rates (Drummond et al., 2002), incorporating a prior of time to most recent common ancestor estimated from a larger dataset of V1–2 and V3 sequences from these subjects (Poss et al., 1998). The estimated evolutionary rate for Q23 sequences was 2.5 % per site per year (95 % confidence intervals 1.61–3.52) based on 16 plasma viral RNA sequences from two samples obtained 20 weeks apart. Rates of 0.28 % (0.07–0.51) were derived for Q45 based on 17 plasma viral sequences obtained 7 weeks apart. Viral RNA sequences were not available for

Received 13 January 2004
Accepted 23 February 2004
Q47 at seroconversion and thus estimated rates of 0.13% (0.05–0.22) were based on eight viral RNA pre-seroconversion sequences, seven proviral DNA sequences obtained 10 weeks later at the time seroconversion was detected, and 10 proviral sequences obtained 7 months after seroconversion. In all cases, confidence intervals excluded zero indicating that measurable rates of evolution occurred in env. There was no correlation between plasma viral loads and evolutionary rates in these individuals.

Protein sequences from pre-seroconversion samples were compared to those obtained at or near seroconversion to determine if amino acid replacement occurred in the evolving virus population. Although intra-sample sequence diversity was high (Fig. 1), in all three subjects amino acids at several positions in sequences derived at or shortly after seroconversion were replaced when compared with the sequences detected at the time of infection (Table 1). Six substitutions were fixed in the seroconversion virus population of Q23. Four of these sites have been shown by mutagenesis studies of clade B gp120 to affect the binding of CD4 and one to affect binding of antibodies to the CD4-binding site (CD4BS) (Table 1). Using a null distribution of 66 positions in gp120 affecting CD4 binding (Kwong et al., 1998; Pantophlet et al., 2003; Rizzuto et al., 1998), there was a significantly higher number of substitutions at sites affecting binding to the CD4-binding site than would be expected if substitutions were randomly distributed (Fisher’s exact test, $P=0.009$). Substitution T430A is of particular interest because alanine mutagenesis of this site in clade B gp120 strongly increased binding of antibody IgG1b12 (b12) (Pantophlet et al., 2003; Saphire et al., 2001) but decreased binding of CD4 (Rizzuto et al., 1998), suggesting that this substitution could affect the structure of the receptor-binding domain. Three out of five temporal substitutions from Q45 viruses involved sites that affected CD4 and b12 or CCR5 binding to clade B gp120, which is significantly more than would be expected by chance ($P=0.037$). Temporal changes in Q47 sequences did not involve sites that have been investigated for impact on receptor, co-receptor or antibody binding, although sites 278 and 364 are adjacent to residues that contact CD4 (Kwong et al., 1998). None of the temporal substitutions in either Q23 or Q45 sequences affected potential N-linked glycosylation sites or involved V1, V2 or V3, and none occurred in the gp41 portion of the glycoprotein (data not shown). In contrast, temporal substitutions in Q47 sequences did involve one site in V1 and gp41 (data not shown) and did result in two new potential N-linked glycosylation sites (Table 1).

Of the temporal substitutions identified in these subjects, 50% or more fell outside variable regions. It is particularly noteworthy that none of the early temporal changes affected V3 because fixation of amino acids in V3 did occur over a 2 year period post-seroconversion in sequences obtained from these individuals (Poss et al., 1998). Substitutions that arise in V3 often correlate with a phenotypic change in the virus population (Scarlati et al., 1997; Speck et al., 1997). The fact that the V3 region in the viruses from these individuals does not evolve near the time of infection.

---

**Fig. 1.** Distribution of gp120 amino acid substitutions. A consensus sequence was determined for the pre-seroconversion sequences obtained from each of three clade A-infected women. The percentage of sequences that contained an amino acid that differed from the pre-seroconversion consensus at each position is shown for pre-seroconversion sequences (grey lines, ‘I’) and sequences obtained at or after seroconversion (dashed lines, ‘Sc’). The secondary structure elements are relative to the HXBc2 crystal and include only the regions that appear in the crystal structure. The locations of the deleted V1–2 and V3 loops are indicated in the bottom panel by arrowheads. The numbers of sequences in each dataset are; Q23–8 and 8, Q45–9 and 8 and Q47–8 and 17 for pre-seroconversion and post-seroconversion samples, respectively.
and seroconversion suggests that there may be selection against change in V3 as the virus population becomes established.

Despite the high prevalence of non-clade B HIV-1 infection globally, most structural, evolutionary and therapeutic research on HIV-1 is based on clade B viruses. Thus, it is not clear whether results obtained from clade B gp120 mutagenesis studies are applicable to clade A gp120. A crystal structure for clade B gp120 is available (Kwong et al., 1998, 2000a) and provides the opportunity to determine the structural relatedness of gp120 from clades A and B. Using the clade B structure, we first identified substitutions specific to clade A gp120 from an alignment of consensus sequences of clade B and clade A (available at http://hiv-web.lanl.gov/), which are shown in red in Fig. 2(A, B). Although clade A-specific substitutions are distributed throughout the sequence, there is a preponderance of differences in the ‘silent face’ of the gp120 outer domain in β sheet 12 and 13 and α helix 2, which flank the V3 loop (Fig. 2A). This α helix is known to be variable both within and between clades, which can lead to presentation of distinct antigenic surfaces to the immune system (Kwong et al., 2000a). In the 3-D structure, substitutions characteristic of clade A occur at periodic intervals in the α helix and in a continuous stretch of the apposing β12 sheet to form a clade A-specific planar surface. In addition, there are several clade A changes within the CD4-binding pocket (Fig. 2B). Clade A HIV-1 gp120s, including all representatives discussed herein, contain a substitution at P369, which is flanked by conserved CD4 contact residues D368 and E370. Substitutions of P369 decrease the binding of

Table 1. Summary of temporal substitutions

| Subject | Position* | Amino acid infection† (%) | Amino acid post-SC† (%) | Secondary structure§ | Predicted effect||
|---------|-----------|--------------------------|------------------------|---------------------|------------------|
| Q23     | T/K232    | E (100)                  | G (100)                | £ A                 | NR§              |
| Q23     | K/Q337    | S (100)                  | R (100)                | ±2                  | B6†              |
| Q23     | V430      | T (100)                  | A (100)                | β21                 | CD4 & b12†§§     |
| Q23     | N/K460    | N (100)                  | D (100)                | V5                  | CD4 & b12††      |
| Q23     | S/D461    | D (100)                  | N (100)                | V5                  | CD4 & b12††      |
| Q23     | N463      | N (100)                  | F (100)                | V5                  | CD4 & b12††      |
| Q45     | N92       | E (56)                   | D (85)                 | β1                  | NR               |
| Q45     | W395      | F (56)                   | W (85)                 | β18-V4              | CD4 & b12††      |
| Q45     | T415      | D (100)                  | T (62)                 | β19                 | NR               |
| Q45     | I424      | V (100)                  | I (54)                 | β19-β20             | CCR5 & CD4§§     |
| Q45     | S/D461    | K (100)                  | E (77)                 | V5                  | CD4 & b12††      |
| Q47     | N241      | H (75)                   | N (88)                 | β6-β7               | NR               |
| Q47     | T278      | L (88)                   | S (100)#               | £ D                 | NR               |
| Q47     | S364      | H (75)                   | P (100)                | β14-β15             | NR               |
| Q47     | ?         | S (88)                   | N (65)                 | V4-β19              | NR               |
| Q47     | S/G411    | D (88)                   | N (100)                | V4-β19              | NR               |
| Q47     | T415      | I (88)                   | T (78)#                | β19                 | NR               |

*Amino acid and position number is relative to HXBc2 (Kwong et al., 1998). Where two amino acids are shown, the second is from primary isolate YU2 (Kwong et al., 2000a).
†Amino acid in single letter nomenclature that was present in sequences obtained from viral RNA in plasma samples obtained before seroconversion (Poss et al., 1995). The percentage shown in parentheses is the proportion of sequences that contain the amino acid shown.
‡Amino acid in single letter nomenclature that was present in sequences obtained from viral RNA (Q23) or RNA and proviral DNA (Q45 and Q47) at or shortly after seroconversion (Poss et al., 1995). The percentage shown in parentheses is the proportion of sequences that contain the amino acid shown.
§Location of the substitution is referenced to the secondary structure determined from the HXBc2 crystal (Kwong et al., 1998).
||Sites determined by mutagenesis to affect CD4BS antibody, receptor or co-receptor binding.
§§Mutagenesis at this site is not reported.
#Substitution creates a new potential N-linked glycosylation site.
††Pantophlet et al. (2003).
‡‡Rizzuto et al. (1998).
§§Saphire et al. (2001).
NR, not reported.
CD4 binding site antibody, b12 (Saphire et al., 2001), and are a feature of clade B neutralization escape mutants (Mo et al., 1997). Clade A-specific substitutions are also present in the bridging sheet, which is involved in co-receptor binding (Kwong et al., 1998). It is noteworthy that there are few clade differences in the inner domain of gp120, a region of contact with gp41 (Kwong et al., 1998).

We developed a homology model of clade A gp120 based on the clade B structure, 1G9N (Kwong et al., 2000a) to determine the spatial distribution of temporal substitutions from three subjects. The prototype clade A model employed the gp120 sequence of Q23Sc4 (AY069928) because previous phylogenetic analysis demonstrated that Q23 viruses were basal in the clade A tree (Poss et al., 1997), suggesting that they were suitable representatives of clade A gp120. The clade A homology model was built at the Molecular Computational Core Facility at the University of Montana, utilizing an Octane SGI (Silicon Graphics) workstation operated with software SYBYL 6.8 employing Biopolymer,

**Fig. 2.** Topological distribution of clade A-specific substitutions and temporal substitutions arising after infection with clade A HIV-1. (A, B) The structure of clade B gp120 (1G9N) is shown as a ribbon diagram with sites differing between consensus clade B and clade A gp120 depicted in red, the truncated V1–2 and V3 loops indicated in green and the V4 and V5 loops shown in cyan. The arrow in (A) indicates the planar face containing clade A-specific substitutions. Perspective in (A) is with the virus membrane upper right and cell membrane lower left (CD4 is not shown), and in (B) the molecule is rotated approximately 110° around the vertical axis to display the CD4-binding pocket and CD4, shown as a violet tube structure. (C, D) Homology model of clade A HIV-1 based on the sequence of Q47S6 is shown in the same orientation as the clade B structure in (A, B). Temporal substitutions that arose in Q23, Q45 and Q47 sequences are displayed in magenta, cyan and purple, respectively. The temporal substitutions that were shared between viruses are indicated with striping. The V3 loop has been added to the model and adjustments have been made to accommodate the longer clade A V4 loop.
Composer and ProTable modules (Tripos). The Q23Sc4 sequence was threaded against the 1G9N scaffold using the Composer module, which considers residues that only align with the scaffold, thereby providing a preliminary homology model. The original 1G9N structure is devoid of the first three variable loops. For the homology model, the V1–V2 loop regions were not considered. However, the clade A V3 loop structure was fashioned by a protocol related to Kwong’s method (Kwong et al., 2000b) utilizing published NMR conformations of the gp120 clade B V3 loop region (Vranken et al., 2001). Whereas there are clade-specific regions of gp120 (shown in Fig. 2A, B), the core clade A gp120 model structure was found to possess few abnormal φ, ψ and ω values (using ProTable; Lovell et al., 2000) relative to those of the X-ray-derived core clade B gp120 structure, suggesting a noteworthy degree of core structural similarity between the protein representations. The clade A V4 loop is longer than that of the 1G9N crystal. Therefore, the additional amino acids were accommodated by clipping at the V4 N-terminal side, inserting the extra clade A amino acids, followed by adjusting select bond rotations at the loop C-terminal end to allow loop attachment to the core structure. To obtain a refined Q23Sc4 homology model (Fig. 2C, D), side chain clashes between the V3 and V4 loops were minimized, first by changing select bond torsions to provide φ, ψ and ω values consistent with established standards (Lovell et al., 2000), followed by local V3 and V4 energetic minimizations (Tripos force field) to their respective nearest energy wells.

The sequence with the longest V4 region, Q47S6 (AY288084), was threaded against the Q23Sc4 homology structure to display the temporal substitution. Virus populations from each of the three individuals evolved a temporal substitution in the inner domain near the truncated N terminus of gp120 (Fig. 2C), a region that is proximal to the viral membrane and that also accommodated clade A-specific changes (Fig. 2A). The remainder of the inner domain did not change in the period following infection. Both Q45 and Q47 viruses had amino acid substitutions in V4 and the V4 stem, a region on the outer domain of the protein that does not contribute to virus neutralization. In both the Q23 and Q45 virus populations, temporal substitutions involved the bridging sheet and the V5 region (Fig. 2D), which forms an upper surface to the CD4-binding pocket and has been implicated by mutagenesis to affect gp120–receptor interaction (Table 1). One of the Q47 substitutions (H364P) also lies on the lateral surface of the binding pocket and is adjacent to key CD4 contact residues (Kwong et al., 1998).

Our data demonstrate that in three independent infections there is measurable evolution in HIV-1 env preceding and following seroconversion. We provide the first model of clade A gp120 and demonstrate that it has significant structural similarity with the clade B glycoprotein and that substitutions arising as the virus population becomes established in three hosts are not randomly distributed in the protein. Furthermore, in two of the three subjects there were significantly more changes at sites shown by mutagenesis of clade B gp120 to affect receptor or coreceptor binding than would be expected by chance. Further studies of the substitutions that arise soon after infection, will be valuable to understand selective forces acting on the infecting virus population and to inform efforts aimed at preventing establishment of new infections.

Acknowledgements

The authors thank Drs Joan Kreiss, Harold Martin, Jr and their collaborators at the Ganjoni Municipal Clinic and Coast Provincial General Hospital, Mombasa, Kenya for sample collection and Dr Julie Overbaugh for providing plasma samples. This research was supported in part by NIH AI44609. Partial support for D. C. H. and H. C. came from Montana NSF-EPSCoR, EPS-0091995 and for J. G. from NIH-NCRR, NIH P20 RR15583. Funding for the Molecular Computational Core Facility, Center for Structural and Functional Neuroscience, was derived from NIH P20 RR15583-01 and the NSF EPS-0091995.

References


