

Intrapatent Sequence Variation of the *gag* Gene of Human Immunodeficiency Virus Type 1 Plasma Virions

FAYTH K. YOSHIMURA,^{1*} KURT DIEM,² GERALD H. LEARN, JR.,³ STANLEY RIDDELL,⁴
AND LAWRENCE COREY²

*Department of Biological Structure, University of Washington,¹ Departments of Laboratory Medicine and Medicine,²
and Department of Microbiology,³ University of Washington, Seattle, Washington 98195, and
Fred Hutchinson Cancer Research Center, Seattle, Washington 98104⁴*

Received 20 June 1996/Accepted 28 August 1996

Because certain regions of the *gag* gene, such as p24, are highly conserved among human immunodeficiency virus (HIV) isolates, many therapeutic strategies have been directed at *gag* gene targets. Although intrapatent variation of segments of *gag* have been determined, little is known about the variability of the full-length *gag* gene for HIV isolated from a single individual. To evaluate intrapatent full-length *gag* variability, we derived the nucleotide sequences of at least 10 cDNA *gag* clones of virion RNA isolated from plasma for each of four asymptomatic HIV type 1-infected patients with relatively high CD4⁺ T-cell counts (300 to 450 cells per mm³). Mean values of intrapatent *gag* nucleotide variation obtained by pairwise comparisons ranged from 0.55 to 2.86%. For three subjects, this value was equivalent to that reported for intrapatent full-length *env* variation. The greatest range of intrapatent mean nucleotide variation for individual protein-coding regions was observed for p7. We did not detect any G-to-A hypermutation, as A-to-G and G-to-A transitions occurred at similar frequencies, accounting for 29 and 25%, respectively, of the changes. Mean variation values and phylogenetic analysis suggested that the extent of nucleotide variation correlated with the length of viral infection. Furthermore, no distinct subpopulations of quasispecies were detectable within an individual. The predicted amino acid sequences indicated that there were no regions within a *gag* protein that were comprised of clustered changes.

Important humoral and cellular immune responses to human immunodeficiency virus type 1 (HIV-1) *gag* products have been noted (1, 2, 8–10, 15, 25, 30, 31, 35, 36, 39, 42). Antibodies to p24 are inversely correlated with progression of disease (1, 9, 10, 15, 25, 39, 42), and much of the cytotoxic T-lymphocyte response to HIV-1 is directed at *gag* gene products (2, 8, 35, 36). Thus, several groups have suggested *gag* gene targets for more universal approaches to immunotherapy.

Intrapatent variation has been previously determined for short segments of the HIV-1 *gag* gene (4, 20, 21, 28, 32, 52, 53). Balfe et al. (4) reported that the range of the mean intrapatent variation for portions of p24 was between 0.5 and 4.1%. In contrast, others have observed that the p17 protein-coding region has a much greater intrapatent variation (20, 21, 28, 32, 52, 53), and as a consequence, it has been a useful indicator of HIV evolution. Although such variation for segments of *gag* has been analyzed for HIV isolated from a person, there is no comparable study of the entire *gag* gene.

We therefore examined nearly full-length *gag* sequences from several HIV-1-infected individuals to gain a more complete understanding of intrapatent variation in this gene and to enable us to compare each *gag* protein-coding region for a single genome. Full-length *gag* sequences were more closely related within individuals than between individuals. Significant variation in *gag* gene products was seen in all four patients studied. Our data indicate that within each *gag* protein-coding region, there are no segments of clustered changes analogous to the V1 to V5 regions of envelope (4, 7, 29, 43, 44). Interestingly, it appeared that the degree of sequence variation in

gag correlated with the length of HIV-1 infection. Moreover, phylogenetic analysis suggested that single major clades predominate in individuals.

MATERIALS AND METHODS

Assays of viral load. HIV-1 levels in peripheral blood mononuclear cells (PBMCs) were determined by cocultivating serial dilutions of patient PBMCs (1×10^6 to 3.2×10^3 /ml) with 10^6 donor PBMCs in 2 ml of RPMI 1640 medium (GIBCO/BRL) containing 20% fetal bovine serum, 50 μ g of interleukin-2 per ml, and 0.001% DEAE-dextran for 14 days. At that time, assays for p24 antigen were performed with the Abbott HIV AG-1 enzyme-linked immunosorbent assay kit, using previously published methods. The HIV-1 RNA in plasma was measured by the branched-DNA signal amplification (bDNA) assay as described by Pacht et al. (38).

Virion RNA isolation. One milliliter of plasma from each patient was centrifuged at $20,000 \times g$ at 4°C for 1 h, and the virus pellet was resuspended in 200 μ l of 4 M guanidinium isothiocyanate (Fluka). Ten micrograms of yeast tRNA (Sigma) was added to the sample, which was then extracted by sequentially adding 20 μ l of 2 M sodium acetate (pH 4.0), 200 μ l of phenol (pH 5.0), and 40 μ l of chloroform, with vigorous vortexing between each addition. The RNA was precipitated by the addition of an equal volume of isopropanol and incubation overnight at –20°C.

RT-PCR of *gag* sequences from virion RNA. One-tenth of extracted RNA was resuspended in 10 μ l of diethyl pyrocarbonate-treated water, heated at 80°C for 5 min, and placed on ice. The heat-treated RNA was subsequently added to 15 μ l of reverse transcriptase (RT) mix (1 \times RT mix contains 2.5 mM MgCl₂, 0.6 mM dithiothreitol, 2.0 mM total deoxynucleoside triphosphates [dNTPs] [Perkin-Elmer], 50 pmol of primer, 24 U of RNasin [Promega], 10 mM Tris-HCl [pH 8.3], 50 mM KCl, and 10 U of Moloney murine leukemia virus RT [Gibco/BRL]). Samples were incubated at 37°C for 1 h. The RT enzyme was heat inactivated by incubation at 95°C for 5 min, and the tubes were chilled on ice. For patients 1 and 2, the sequence of the primer was 5'-GGTTTCATCTCTCTGGCAA-3'; for patients 3 and 4, it was 5'-CGAGGGGTCGTTGCCAAAGA-3'.

The cDNA product was PCR amplified by *Taq* polymerase (Boehringer Mannheim) in a reaction mixture containing 0.25 mM MgCl₂, 0.4 mM total dNTPs, 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 50 pmol of second primer, and 5 U of *Taq* polymerase. PCR conditions were 95°C for 1 min, 55°C for 1 min, and 72°C for 2 min for 35 cycles, with a 10-min final extension at 72°C in a Perkin-Elmer thermocycler. For all patients, the sequence of the primer was 5'-GACTAGCG GAGGCTAGAAGG-3'. Amplification of the initial product was performed under identical conditions as before with the internal primers 5'-CGTTCGGA

* Corresponding author. Present address: Wayne State University, Department of Immunology and Microbiology, Gordon H. Scott Hall, 540 E. Canfield Ave., Detroit, MI 48201. Phone: (313) 577-1571. Fax: (313) 577-1155. Electronic mail address: fyoshi@med.wayne.edu.

ATTCATGGGTGCGAGAGCGTCAGTA-3' and 5'-CGTTCGGTTCGACTTA GAAGCTCTCTTCTGGTGGG-3' (patients 1 and 2) or 5'-CGTTCGGAATT CATGGGTGCGAGAGCGTCAGTA-3' and 5'-CGTTCGGTTCGACCTAATA GAGCTTCCTTTAGTTGCC-3' (patients 3 and 4).

gag cloning into plasmid vectors. The products of three to four nested PCRs were pooled and treated with 10 U of sequencing-grade Klenow enzyme (Boehringer Mannheim) in 50 mM Tris-HCl (pH 7.5)–10 mM MgCl₂–0.8 mM total dNTPs. DNA was precipitated, treated again with Klenow enzyme, and electrophoresed through a 1% SeaPlaque (FMC) gel in TAE buffer (40 mM Tris-acetate [pH 8.0], 1 mM EDTA). The band of interest was excised, isolated, digested with *EcoRI* and *SalI*, and inserted into the same sites of pMal (New England Biolabs) for patients 1 and 2 or pUC19 for patients 3 and 4.

DNA sequence determination. Template DNA was isolated by suspending bacterial pellets in 200 μ l of 50 mM glucose–25 mM Tris-HCl (pH 8.0)–10 mM EDTA to which 300 μ l of 0.2 N NaOH–1% sodium dodecyl sulfate was added. Samples were mixed by inversion and placed on ice for 5 min. Then 300 μ l of 3 M potassium acetate (pH 5.0) was added, and tubes were mixed by inversion and incubated on ice for 5 min. Tubes were spun in a microcentrifuge at maximum speed for 2 min. The supernatant was removed, and 4 μ l of RNase A (10 mg/ml; Sigma) was added. Samples were incubated at 37°C for 20 min. Three chloroform extractions were performed, and DNA was precipitated by adding an equal volume of isopropanol, vortexing for 1 min, and spinning in a microcentrifuge at maximum speed for 10 min at RT. Pellets were washed in 70% ethanol and resuspended in 32 μ l of water. Eight microliters of 4 M NaCl and 40 μ l of 13% polyethylene glycol 8000 were added before the tubes were placed on ice for at least 1 h. Tubes were spun at 4°C in a microcentrifuge, and pellets were washed in 70% ethanol and resuspended in 11 μ l of water. The nucleotide sequence of both DNA strands was determined by using the *Taq* dyeDeoxy terminator cycle sequencing kit (Applied Biosystems). Labeled samples were put through CentriSep columns (Princeton Separations) to remove unincorporated terminators before drying in a Speedvac. The samples were resuspended in formamide and analyzed on a sequencing gel by a model 373A DNA Sequencer (Applied Biosystems).

Error rate for nucleotide sequence determination. To determine the error rate of *Taq* polymerase for our PCR amplifications involving the use of nested primers, we performed amplifications of the HIV-1 LAI *gag* gene (14) by using similar procedures. We determined the *gag* sequence of a clone that was isolated from each of three independent nested PCRs performed at different times. We detected one base pair change in 4,160 sequenced bases. Based on these data and the published error rate (1 in 3,470 nucleotides) of the Moloney murine leukemia virus RT (16) (Bethesda Research Laboratories) that was used to reverse transcribe the HIV-1 RNA, we calculated our error rate to be 1 in 3,800 nucleotides, or less than 1 base for the entire *gag* gene.

Sequence analysis and phylogenetic programs. Clustal V (18) and the Genetics Computer Group program PILEUP were used to align the deduced *gag* amino acid sequences. PRETTY was used to generate a consensus sequence for each patient. Alignments were verified by visual inspection. The sequences were not gap stripped for additional analyses since the alignments were unambiguous, but insertion/deletion regions were omitted on a pairwise basis for all distance calculations. Intra- and interpatient distances were calculated by using the MEGA package (24). Distances were corrected for multiple substitutions by using the method of Jukes and Cantor (22). Because analyses of inferred protein sequences and synonymous and nonsynonymous differences required complete reading frames, the proper reading frame for some sequences was restored by removing single nucleotide insertions or inserting an ambiguous nucleotide. Stop codons were omitted from the protein and nonsynonymous/synonymous distance (d_n/d_s) analyses. Phylogenetic analysis of the sequences was done with the PHYLIP package (13). Specifically, phylogenies were estimated by the neighbor-joining method (41) as implemented in NEIGHBOR with two-parameter Kimura distances (23) (transition/transversion ratio = 2.0) generated by the DNADIST program. Bootstrap confidence values were generated from 1,000 bootstrap resampled sequences from SEQBOOT. Distances were calculated as described above, using DNADIST; phylograms for each replicate were estimated with NEIGHBOR; bootstrap confidence consensus values (12) shown represent trees seen in greater than 70% of the resulting trees as determined with CONSENSE.

Nucleotide sequence accession numbers. The nucleotide sequences that we obtained for the *gag* clones isolated from each patient have been submitted to Genbank with accession numbers U29242, U29246 to U29265, and U29403 to U29422.

RESULTS

We isolated molecular clones of the HIV-1 *gag* gene directly from plasma virions because this population should represent actively replicating virus (19, 48). Plasma was obtained from four seropositive patients who were being evaluated for a protocol involving the infusion of HIV-1-specific CD8⁺ T-cell clones. All four patients were asymptomatic, had no history of opportunistic infection, and were receiving antiretroviral treatment with zidovudine only. All four were infected with HIV-1

TABLE 1. Virological and immunological markers of HIV-1 infection

Patient	Yr of known HIV seropositivity	CD4 ⁺ lymphocytes/mm ³	Titer of HIV in PBMCs ^a	HIV RNA copies/ml in plasma ^b
1	2	406	ND ^c	<10,000
2	5	450	5,000	45,000
3	9	325	6	15,000
4	12	364	40	14,000

^a Expressed as infectious units per million cells. Serial dilutions of patient PBMCs were cocultivated with 10⁶ donor PBMCs in 2 ml of RPMI 1640 medium containing 20% fetal bovine serum, 50 μ g of interleukin-2 per ml, and 0.001% DEAE-dextran for 14 days, at which time assays for p24 were performed with the Abbott HIV AG-1 enzyme-linked immunosorbent assay kit.

^b Viral RNA was detected by a bDNA assay (37). Level of detection by the bDNA assay is 10⁴ RNA copies per ml.

^c ND, not detectable.

for various lengths of time (Table 1). The CD4⁺ lymphocyte counts for these patients ranged between 325 and 450 cells per mm³. Cocultivation assays of patient PBMCs with normal PBMCs yielded virus titers with a range from nondetectable for patient 1 to 5,000 infectious units per million cells for patient 2. We measured the level of HIV-1 RNA in plasma with a bDNA assay (37) and obtained values that ranged from less than detectable by this assay (10 × 10³) to 45 × 10³ RNA copies per ml.

Nucleotide sequences of *gag* clones. By using reverse transcription and PCR amplification of *gag* sequences with nested primers as detailed in Materials and Methods, we obtained molecular clones of *gag* for each of the four patients. The *gag* sequences that we obtained corresponded to full-length *gag* except for the last 111 bp in p6. For clarification, we shall nevertheless refer to these sequences as full-length *gag* in this report to distinguish them from the individual protein-coding regions. Because the *gag* genes were cloned into a plasmid vector, we determined the nucleotide sequence of both DNA strands of each clone for each individual. Each of the clones was similar to prototype subtype B viruses (see Fig. 3). We also used a multiple alignment program to generate a consensus sequence for the *gag* clones for each patient (data not shown). Sequence similarities of the consensus *gag* sequence for each individual to prototypes HIV-1 LAI, MN, and SF2 (33) were 95, 93, and 94%, respectively.

The mean percent nucleotide variation for full-length *gag* and the p17, p24, p7, and p6 protein-coding regions was determined by performing pairwise comparisons of clones for each patient (Table 2). Within an individual, nucleotide differences for full-length *gag* varied from 0.55% for patient 1 to 2.86% for patient 3. The ranges of the mean percent variation for the protein-coding regions were 0.80 to 2.55% for p17, 0.4 to 2.5% for p24, 0.13 to 3.88% for p7, and 0.74 to 2.43% for p6. Overall, the mean *gag* variation between individuals was 5.9% (Table 3). Thus, differences in *gag* sequences from a single individual varied less than between individuals. As with full-length *gag*, the interpatient variation for each of the protein coding regions was also greater than the inpatient divergence (Table 3).

To examine the rates at which nonsynonymous and synonymous nucleotide changes have occurred within an individual, we calculated d_n/d_s ratios for full-length *gag* as well as for the individual *gag* genes as summarized in Table 4. We noted considerable variation in the inpatient d_n/d_s values that we obtained. However, between patients, the most strongly constrained protein-coding region was p24, where the synonymous

TABLE 2. Nucleotide sequence variation of *gag* genes

Patient	Mean % variation ^a (range)				
	<i>gag</i>	p17	p24	p7	p6
1	0.55 (0.14–1.01)	0.80 (0.00–1.54)	0.40 (0.15–0.87)	0.13 (0.00–0.61)	1.92 (0.00–6.89)
2	1.16 (0.58–1.97)	1.47 (0.51–2.33)	0.78 (0.00–1.61)	2.20 (0.62–4.45)	0.74 (0.00–2.26)
3	2.86 (1.89–3.84)	2.55 (0.77–4.46)	2.50 (1.32–4.17)	3.88 (0.62–6.44)	2.43 (0.00–6.98)
4	2.05 (0.94–3.69)	1.81 (0.00–5.27)	1.69 (0.58–3.26)	2.60 (0.00–6.32)	1.67 (0.00–4.58)

^a Intrapatient distances were calculated by using the MEGA package of Kumar et al. (24). Distances were corrected for multiple substitutions by the method of Jukes and Cantor (22). The mean value was calculated by dividing the sum of the percent variation for each pairwise comparison of *gag* clones by the number of pairwise comparisons analyzed for each patient. *gag* refers to all *gag* sequences, and p17, p24, p7, and p6 refer to the protein-coding regions of *gag*.

changes were 10 times more frequent than those at nonsynonymous sites (Table 3). The p17-coding region was the least constrained when interpatient ratios were compared, with approximately twice as many synonymous changes as nonsynonymous changes. Interestingly, the intrapatient d_n/d_s values and length of infection showed a relationship suggesting that d_n/d_s decreases with length of infection. This is most clearly seen for the complete *gag* value for patients 1, 2, and 3. Patient 4, however, departs from this trend for unknown reasons. Interestingly, the d_n/d_s values for patient 3 are clearly the lowest for all four *gag* reading frames; this patient showed the greatest nucleotide sequence variation, as may be seen both in Table 2 and the phylogenetic tree (Fig. 3).

The nucleotide changes we detected in *gag* were predominately A-to-G and G-to-A transitions, which accounted for 29 and 25%, respectively, of the changes. The next most common changes were represented by T-to-C (17%) and C-to-T (13%) transitions. Transitions occurred roughly five times more frequently than transversions, similar to what has been reported for *env* (17).

Amino acid sequence variation. The amino acid differences for the *gag* clones in comparison with the consensus sequence for each patient are shown in Fig. 1. We did not detect any amino acid residue in any of the *gag* proteins that corresponded to a preferred nonsilent mutation site. As a consequence, within each protein-coding region there were no detectable regions where amino acid changes were clustered as has been seen for the hypervariable regions of *env* (4, 7, 29, 43, 44).

Eight of the forty-one *gag* clones had mutations that would lead to the production of an aberrant *gag* polyprotein. Two clones had a single nucleotide change that would lead to pre-

mature termination. One of these point mutations was in p17 (clone 1-6), and the other was in p24 (clone 4-20). The mutation in clone 1-6 involved a Lys-to-ochre change, and in clone 4-20 it involved a Trp-to-opal change. Three clones had frameshift mutations that were produced by the insertion of one nucleotide (clones 2-49, 3-25, and 4-9). These insertions occurred at two different sites in p17 (clones 2-49 and 3-25) and at one site in p24 (clone 4-9). Three additional frameshift mutations were the result of a single nucleotide deletion at one site in p17 (clone 3-13) and two nonidentical sites in p7 (clones 1-30 and 1-5). The rest of the clones had open reading frames for *gag*.

As with the nucleotide sequences, we performed a pairwise comparison of the amino acid sequences of the *gag* proteins for each patient's clones (Table 5) and observed a wide variation for each patient. Calculations of the interpatient variation (Table 3) indicated that p24 had the lowest level (3.14%) and p17 and p7 the highest levels (11.66 and 11.46%, respectively) of amino acid variation. An amino acid identity plot of mean similarity versus the amino acid position in *gag* provided a more detailed map of regions of homology within each protein-coding region (Fig. 2) compared with the mean variation values. The most extreme sequence variations in *gag* were detectable in the C-terminal region of p17, the p2 region, and the N-terminal half of p7.

Comparison of the amino acid sequences of the individual patient clones with each other revealed that only the p24 region of *gag* had relatively large regions (10 or more residues) of contiguous amino acids that were invariant. There were five of these regions in p24 (indicated by numbered boxed regions in the consensus sequence for patient 1 in Fig. 1B). These five regions are also identical for the 25 subtype B isolates in the Los Alamos Human Retrovirus and AIDS database (33). Region 4 is contained within the major homology region (MHR; indicated by the box with dashed lines) that has been shown to be important for HIV-1 assembly and infectivity (11, 27, 40, 46, 49). Interestingly, the other four regions in p24 are even more highly conserved among subtype B clones than the MHR.

Relatedness of quasispecies. We compared the patients' clones with *gag* sequences of HIV-1 isolates of subtypes A

TABLE 3. Interpatient sequence variation

Sequence	Mean % interpatient sequence variation (range)		Mean interpatient d_n/d_s ^a (range)
	Nucleotide ^b	Amino acid ^c	
<i>gag</i>	5.89 (5.10–7.09)	7.11 (5.27–9.01)	0.24 (0.16–0.33)
p17	7.45 (4.46–10.35)	11.66 (7.63–16.79)	0.51 (0.08–1.07)
p24	4.02 (2.66–5.88)	3.14 (0.87–6.52)	0.10 (0.02–0.30)
p7	9.49 (3.09–15.74)	11.46 (3.64–21.82)	0.16 (0.06–0.49)
p6	3.78 (0.00–9.46)	6.75 (0.00–13.33)	0.43 (0.00–0.57)

^a Calculated for each pairwise comparison of each clone between patients by the method of Nei and Gojobori (34) as implemented in the MEGA computer program (24).

^b Calculated for all pairwise comparisons of each sequence with the *gag* clones for the other three patients as described for Table 2.

^c Calculated for all pairwise comparisons of the amino acid sequence of each sequence with those for the *gag* clones for the other three patients. Values were obtained similarly to the calculations for nucleotide variation as presented in Table 2.

TABLE 4. Mean d_n/d_s values

Patient	Mean intrapatient d_n/d_s ^a				
	<i>gag</i>	p17	p24	p7	p6
1	0.56	0.25	0.23	0.00	0.34
2	0.15	0.36	0.15	0.36	0.00
3	0.12	0.14	0.09	0.08	0.03
4	0.26	0.40	0.21	0.12	0.18

^a Calculated as described in Table 3, footnote a.

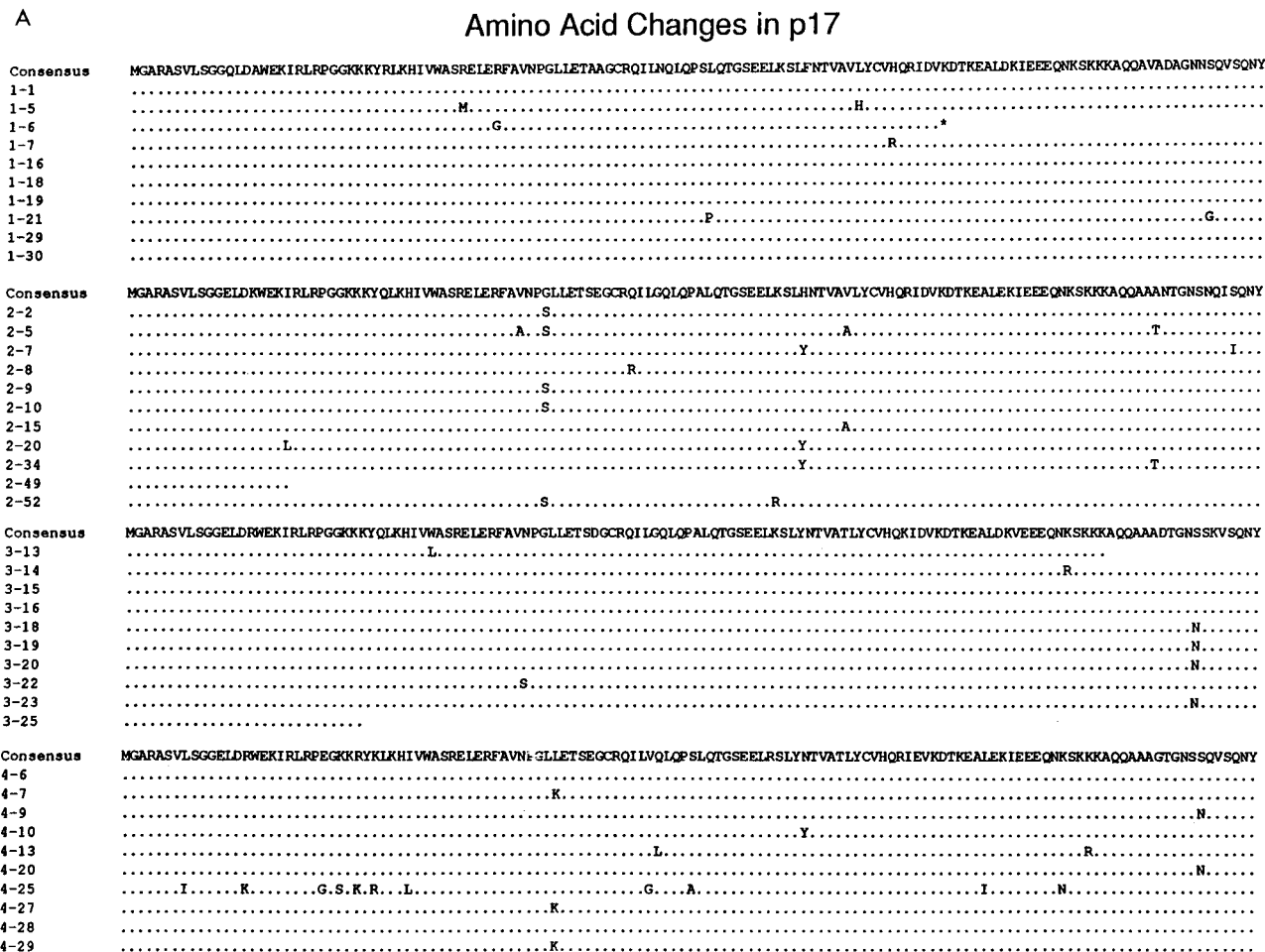


FIG. 1. Deduced amino acid sequence alignment of the *gag* clones for each patient. The consensus amino acid sequence for each patient is shown above that patient's clones. Dots indicate identical residues, and changes are shown by letters. Δ indicates a missing residue. Boxes with solid lines identify five regions of 10 or more amino acids that are conserved in all four patients' clones and subtype B HIV-1 isolates in the Los Alamos database (33). The box with dashed lines designates the MHR (11, 27, 40, 46, 49). Termination codons are represented by asterisks, and frameshift mutations caused by insertions or deletions are indicated by truncated sequences.

(K124), B (D31, NY5, NL43, CAM1, HXB2, and JH31), D (ELI), and F (VI325) (Fig. 3) (33). This phylogenetic analysis revealed that all of the patients' clones were most closely related to the subtype B clade. Our data suggested that all four patients' *gag* sequences formed monophyletic groups that were more closely related to each other than to the other patients' clones or to the representative subtype B isolates that were included in this analysis (Fig. 3). Moreover, the branch lengths of the patient phylograms appeared to be closely related to the length of infection. The depth of the tree for patient 1 (infected for 2 years), for example, is less than that of the other patients, while patients 3 and 4 (infected 9 and 12 years, respectively) showed the greatest tree depth. Although most of these patients' *gag* sequences did not show pronounced within-individual differentiation into subpopulations, clone 4-25, while still most closely related to the other patient 4 sequences, was clearly divergent from other patient 4 clones. The divergence of clone 4-25 from the rest of the patient 4 clade was supported at a bootstrap level of 100%.

DISCUSSION

Our detailed study of the nucleotide and amino acid sequences of the HIV-1 *gag* gene revealed a surprisingly high

degree of inpatient variation for the entire gene. The large degree of inpatient *gag* variation was also indicated by the overlapping ranges of nucleotide variation of the inpatient and interpatient values for each of the protein-coding regions (compare Tables 2 and 3). Although there were overlaps in the distributions, the sequences showed expected phylogenetic relationships (data not shown), thereby ruling out any contamination of samples. While we studied only four patients, our data suggested that duration of HIV infection is a factor in *gag* gene diversity. Patient 1, who had been infected only 2 years, has the smallest variation, and patients 3 and 4, who had been infected 9 and 12 years, respectively, have the greatest (Tables 1 and 2). This is further illustrated by our phylogenetic analysis where the branch lengths of the patient phylograms correlated with the length of infection (Fig. 3).

A comparison of the mean amino acid variation of the protein-coding regions of *gag* for each patient studied did not yield any significant differences among these regions, and we attribute this result to the small sample number of this study. However, when we compared the sequences of the *gag* clones between these patients, we observed that the greatest variations were in the p17 and p7 regions, while p24 had the lowest sequence diversity. Our results for p17 and p24 are in agree-

B

Amino Acid Changes in p24

Consensus PIVQLGQMVHQAI SPRTLNAWVKVVEEKAFSP EVIPMFSA LSEGATPQDLN TMLNTVGGHQAA MQLKETINEEAAEWDRLHPVHAGPVAPGQ MREPRGSDIAGTSTLQEQIGWMTNPP IPVGEIYKRWIILG

1-1
 1-5
 1-7
 1-16 P
 1-18 S
 1-19
 1-21
 1-29
 1-30

Consensus PIVQLGQMVHQAI SPRTLNAWVKVVEEKAFSP EVIPMFSA LSEGATPQDLN TMLNTVGGHQAA MQLKETINEEAAEWDRLHPVHAGPIAPGQ MREPRGSDIAGTSTLQEQIGWMTNPP IPVGEIYKRWIILG

2-2
 2-5 A
 2-7 R
 2-8
 2-9
 2-10
 2-15
 2-20 K
 2-34
 2-52

Consensus PIVQLGQMVHQAI SPRTLNAWVKVVEEKAFSP EVIPMFSA LSEGATPQDLN TMLNTVGGHQAA MQLKETINEEAAEWDRLHPVHAGPVAPGQ MREPRGSDIAGTSTLQEQIGWMTNPP IPVGEIYKRWIILG

3-14 L H
 3-15 L Y T
 3-16 L H
 3-18
 3-19
 3-20 R S
 3-22 R S
 3-23 Y S

Consensus PIVQLGQMVHQAI SPRTLNAWVKVVEEKAFSP EVIPMFSA LSEGATPQDLN TMLNTVGGHQAA MQLKETINEEAAEWDRLHPVHAGPVAPGQ MREPRGSDIAGTSTLQEQIGWMTNPP IPVGEIYKRWIILG

4-6
 4-7 G
 4-9
 4-10 R S
 4-13 V
 4-20 *
 4-25 L P
 4-27
 4-28 K A
 4-29 Q I

Consensus LNKIVRMYSPTSILDI RQGPKEPFRDYVDRFYKTLRAEQASQEVKNMMTETLLVQNANP DCKTI LKALGPAATLEEMMTACQGVGGPGHKARVL

1-1
 1-5
 1-7
 1-16
 1-18
 1-19
 1-21 A
 1-29
 1-30

Consensus LNKIVRMYSPTSILDI RQGPKEPFRDYVDRFYKTLRAEQASQEVKNMMTETLLVQNANP DCKTI LKALGPAATLEEMMTACQGVGGPGHKARVL

2-2
 2-5
 2-7
 2-8
 2-9 A
 2-10
 2-15 M V
 2-20 L A
 2-34
 2-52

Consensus LNKIVRMYSPTSILDI RQGPKEPFRDYVDRFYKTLRAEQASQEVKNMMTETLLVQNANP DCKTI LKALGPAATLEEMMTACQGVGGPGHKARVL

3-14
 3-15
 3-16
 3-18 G
 3-19
 3-20
 3-22
 3-23

Consensus LNKIVRMYSPTSILDI RQGPKEPFRDYVDRFYKTLRAEQASQEVKNMMTETLLVQNANP DCKTI LKALGPAATLEEMMTACQGVGGPGHKARVL

4-6
 4-7
 4-9
 4-10 V
 4-13 I S
 4-25 A
 4-27 T S
 4-28 S
 4-29 A

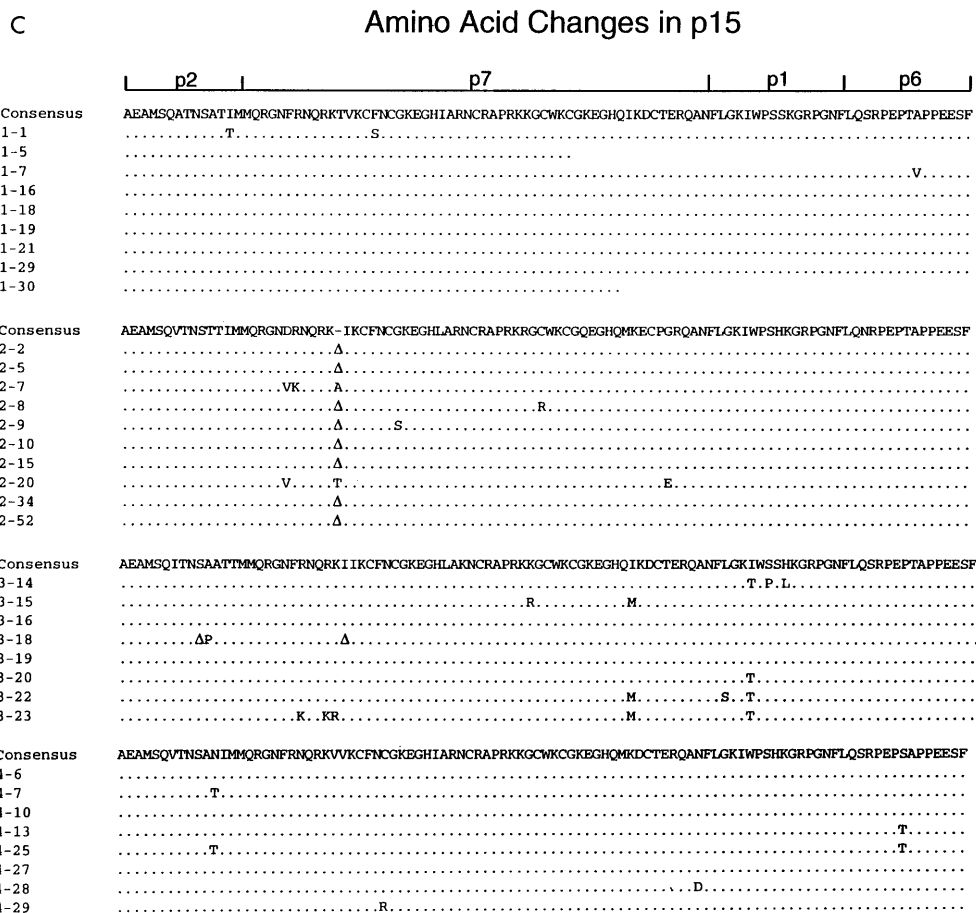


FIG. 1—Continued.

ment to what has been observed before by others (3–5, 20, 26, 28, 32, 33, 52, 53). The interpatient comparison of the mean amino acid variation of the full-length *gag* gene yielded a value of 7.1%, which is similar to a previously reported value for interpatient full-length *gag* variation (26).

The amino acid changes that were detectable within a protein-coding region of *gag* were scattered throughout the protein. Nowhere in *gag* did our analysis identify any residue that corresponded to a site of preferred amino acid substitution. As such, there were no detectable hypervariable regions comparable to the V1 to V5 regions of *env* (4, 7, 29, 43, 44, 50). The implication for these findings on the persistence or effectiveness of *gag*-specific humoral or cellular responses remains to be determined.

TABLE 5. Amino acid variation of *gag* genes

Patient	Mean intrapatient % variation ^a				
	<i>gag</i>	p17	p24	p7	p6
1	0.79	1.07	0.52	0.37	4.00
2	1.12	2.08	0.61	2.03	0.19
3	2.36	2.41	1.73	3.12	1.33
4	2.97	4.40	2.14	3.24	4.85

^a Calculated by using the programs used to calculate the nucleotide mean percent variation as described for Table 2. The percent variation was determined by dividing the number of amino acid differences by the number of amino acids in each *gag* region and multiplying by 100.

Our observation that p24 had the smallest amino acid sequence variation was supported by our calculation of the ratio of nonsynonymous to synonymous nucleotide changes in this region of *gag*. This ratio indicated that p24 was the most highly constrained *gag* protein. This is similar to what has been observed before for short segments of *gag* (5). The interpatient nonsynonymous-to-synonymous ratio for p17 indicated that this protein-coding region was the least constrained, supporting previous observations (3, 20, 21, 26, 28, 32, 52, 53). This may be due mainly to the extreme amino acid sequence variation that exists in the C-terminal region of p17 as indicated by the similarity plot (Fig. 3). Our data support the previously proposed notion that the structure of p24 must be more strictly conserved for viral replication compared with other *gag* proteins, such as matrix.

We observed no evidence of hypermutation in *gag* of the kind that has been detected for *env* sequences (17, 45). Instead of predominately G-to-A transitions, we observed both G-to-A and A-to-G transversions at approximately equal frequencies. Similar results have been reported for changes in the *env* gene of simian immunodeficiency virus isolates (6, 37). As pointed out by Overbaugh et al. (36a), G-to-A hypermutations can often result from *Taq* polymerase errors. We observed that transitions were significantly more frequent than transversions, which is similar to what has been seen for HIV-1 *env* (17, 45).

Our phylogenetic analysis shows that the HIV-1 isolates from each of the four patients were closely related to each other. These data support the idea that a single viral clone of

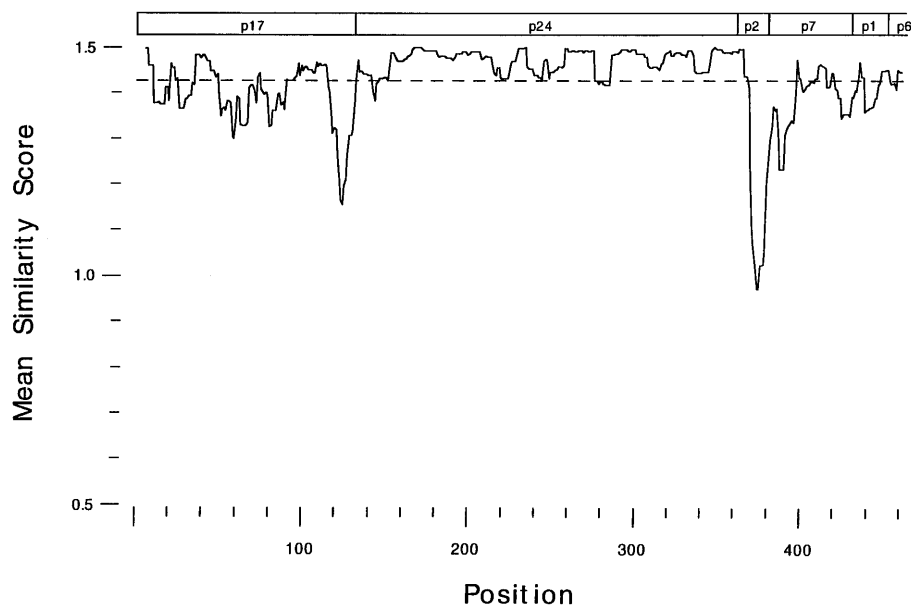


FIG. 2. Amino acid identity plot for *gag*. The solid line represents the running average sequence identity for 30 adjacent amino acid residues centered at the residue position on the abscissa. Amino acid sequence identity was obtained from a multiple sequence alignment of the inferred amino acid sequences for the *gag* clones that we sequenced along with representative sequences from the Los Alamos database (33) (see Results). The dashed line is the overall mean identity for these sequences over their entire lengths.

HIV-1 predominates in an individual. Whether this has occurred as a result of a bottleneck effect after an initial infection by multiple genotypes as discussed by others (32, 51–53) or because only a single genotype can successfully initiate infection is not known, and we are not able to distinguish between these possibilities from our data. Bootstrap analysis indicated that only one patient had a clone (4-25) that diverged significantly from the major clade that was detectable for that patient. Because this outlying clone is still most closely related to the rest of the *gag* clones from that patient, it is unlikely that it arose by superinfection from a different source. Although a sampling bias as a consequence of analyzing only 10 clones per patient may be another possible explanation for our data, we think that this is unlikely as we did not detect a single clone that was unrelated to the major clade for all four patients. In our analysis of a total of 41 clones, there should be a good chance of detecting at least one clone if it were more closely related to other subtype B viruses than to a patient's clade. On the other hand, we cannot eliminate the possibility that plasma viral RNA may reflect the expression of only a limited subset of integrated proviruses since this has been observed by others (44).

The five invariant regions in our clones that we detected (Fig. 1) are identical for all subtype B sequences in the Los Alamos database (33). These regions are also conserved in other subtypes as well, with regions 3 and 4 having the greatest similarity (33). Region four corresponds to part of the MHR, which has been shown to be essential for HIV-1 replication (11, 27, 40, 46, 49). Amino acid residues in two other regions (3 and 5) have also been shown to affect HIV-1 replication (40, 47). A role in HIV-1 replication of regions 1 and 2, however, has not yet been assessed. Given the importance of amino acids in the other three conserved regions, we predict that regions 1 and 2 will also play significant roles in viral replication.

Our study has implications for both molecular epidemiologic studies involving *gag* gene products as well as studies using

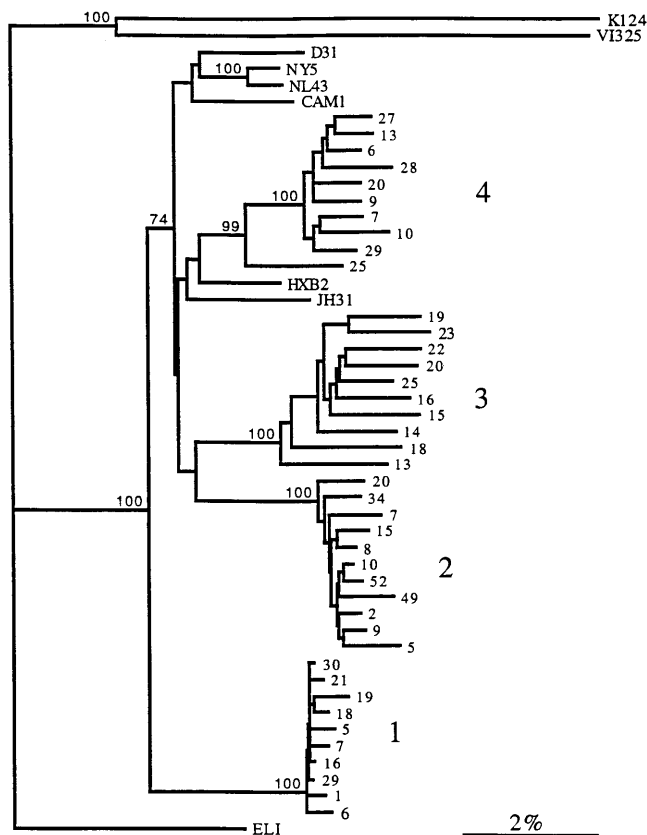


FIG. 3. Phylogenetic analysis of *gag* clones. Phylogenetic analysis was performed with the PHYLIP package of Felsenstein (13). Phylogenies were estimated with the neighbor-joining method (41) on a matrix of pairwise nucleotide sequence distances (23). Support values for the trees are based on 1,000 bootstrap replicates (12). Included in the analysis were HIV-1 subtype A (K124), B (D31, NY5, NL43, CAM1, HXB2, and JH31), D (ELI), and F (VI325) sequences (33). Patient numbers are shown to the right of the clone numbers. The scale bar indicates 2% sequence divergence.

immunotherapy directed against the HIV *gag* gene. The significant variation in many regions of the HIV-1 *gag* gene indicates that the use of T-cell clones to defined regions of *gag* may not eliminate many of the circulating virus strains. We would predict that escape variants may be a concern with such a therapeutic approach. Consideration should be given to using immunotherapies that are directed at the conserved regions of p24.

ACKNOWLEDGMENTS

We thank Eric Peterson and Todd Berard for excellent technical assistance and James Hughes for discussions regarding the manuscript.

This research was supported by National Institutes of Health grants AI36613 (SPIRAT), AI05065 (AVEU), and AI27757 (CFAR) awarded to L.C.

REFERENCES

- Ahearne, P. M., T. J. Matthews, H. K. Lyerly, G. C. White, D. P. Bolognesi, and K. J. Weinhold. 1988. Cellular immune response to viral peptides in patients exposed to HIV. *AIDS Res. Hum. Retroviruses* **4**:259–267.
- Ahearne, P. M., R. A. Morgan, M. W. Sebastian, D. P. Bolognesi, and K. J. Weinhold. 1995. Multiple CTL specificities against autologous HIV-1 infected BLCLs. *Cell. Immunol.* **161**:34–41.
- Albert, J., J. Wahlberg, T. Leitner, D. Escanilla, and M. Uhlen. 1994. Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 *pol* and *gag* genes. *J. Virol.* **68**:5918–5924.
- Balfe, P., P. Simmonds, C. A. Ludlam, J. O. Bishop, and A. J. L. Brown. 1990. Concurrent evolution of human immunodeficiency virus type 1 in patients infected from the same source: rate of sequence change and low frequency of inactivating mutations. *J. Virol.* **64**:6221–6233.
- Brown, A. L., and P. Monaghan. 1988. Evolution of the structural proteins of human immunodeficiency virus: selective constraints on nucleotide substitution. *AIDS Res. Hum. Retroviruses* **4**:399–406.
- Burns, D. P., and R. C. Desrosiers. 1991. Selection of genetic variants of simian immunodeficiency virus in persistently infected rhesus monkeys. *J. Virol.* **65**:1843–1854.
- Burns, D. P., and R. C. Desrosiers. 1994. Envelope sequence variation, neutralizing antibodies, and primate lentivirus persistence. *Curr. Top. Microbiol. Immunol.* **188**:185–219.
- Carmichael, A., X. Jin, P. Sissons, and L. Borysiewicz. 1993. Quantitative analysis of the human immunodeficiency virus type 1 (HIV-1)-specific cytotoxic T lymphocyte (CTL) response at different stages of HIV-1 infection: differential CTL response to HIV-1 and Epstein-Barr virus in late disease. *J. Exp. Med.* **177**:249–256.
- Chargelegue, D., B. T. Colvin, and C. M. O'Toole. 1993. A 7-year analysis of antiGag (p17 and p24) antibodies in HIV-1 seropositive patients with haemophilia: immunoglobulin G titer and avidity are early predictors of clinical course. *AIDS* **7**(Suppl. 2):S87–S90.
- de Wolf, F., J. M. A. Lange, J. T. M. Houweling, R. A. Coutinho, P. T. Schellekens, J. Van Der Noordaa, and J. Goudsmit. 1988. Numbers of CD4 cells and the levels of core antigens and of antibodies to the human immunodeficiency virus as predictors of AIDS among seropositive homosexual men. *J. Infect. Dis.* **158**:615–622.
- Dorfman, T., A. Bukovsky, Å. Öhagen, S. Höglund, and H. G. Göttlinger. 1994. Functional domains of the capsid protein of human immunodeficiency virus type 1. *J. Virol.* **68**:8180–8186.
- Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783–791.
- Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle.
- Fisher, A. G., E. Collalti, L. Ratner, R. C. Gallo, and F. Wong-Staal. 1985. A molecular clone of HTLV-III with biological activity. *Nature (London)* **316**:262–265.
- Forster, M. S., L. M. Osborne, R. Chiengson-Popov, C. Kenny, R. Burnell, D. J. Jeffries, A. J. Pinching, J. R. W. Harris, and J. N. Weber. 1987. Decline of anti-p24 antibody precedes antigenemia as correlate to prognosis in HIV1 infection. *AIDS* **1**:235–240.
- Gerard, G. F. 1986. The error rate of cloned M-MLV reverse transcriptase during DNA synthesis. *BRL Focus* **8**(3):3.
- Goodenow, M., T. Huet, W. Saurin, S. Kwok, J. Sninsky, and S. Wain-Hobson. 1989. HIV-1 isolates are rapidly evolving quasispecies: evidence for viral mixtures and preferred nucleotide substitutions. *J. Acquired Immune Defic. Syndr.* **2**:344–352.
- Higgins, D. G., A. J. Bleasby, and R. Fuchs. 1992. CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biol. Sci.* **8**:189–191.
- Ho, D. D., A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz. 1995. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature (London)* **373**:123–126.
- Holmes, E. C., L. Q. Zhang, P. Robertson, A. Cleland, E. Harvey, P. Simmonds, and A. J. L. Brown. 1995. The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. *J. Infect. Dis.* **171**:45–53.
- Holmes, E. C., L. Q. Zhang, P. Simmonds, A. S. Rogers, and A. J. L. Brown. 1993. Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. *J. Infect. Dis.* **167**:1411–1414.
- Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules, p. 21–132. *In* H. N. Munro (ed.), *Mammalian protein metabolism*. Academic Press, New York.
- Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111–120.
- Kumar, S., K. Tamura, and M. Nei. 1993. MEGA: molecular evolutionary genetic analysis, version 1.02. The Pennsylvania State University, University Park, Pa.
- Lange, J. M. A., D. A. Paul, H. G. Huisman, F. de Wolf H. Van Den Berg, R. A. Coutinho, S. A. Danner, J. Van Der Noordaa, and J. Goudsmit. 1986. Persistent HIV antigenemia and decline of HIV core antibodies associated with transition to AIDS. *Br. Med. J.* **293**:1459–1462.
- Louwagie, J., F. E. McCutchan, M. Peeters, T. P. Brennan, E. Sanders-Buell, G. A. Eddy, G. van der Groen, K. Fransen, G.-M. Gershy-Damet, R. Deleys, and D. S. Burke. 1993. Phylogenetic analysis of *gag* genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS* **7**:769–780.
- Mammano, F., Å. Öhagen, S. Höglund, and H. G. Göttlinger. 1994. Role of the major homology region of human immunodeficiency virus type 1 in virion morphogenesis. *J. Virol.* **68**:4927–4936.
- Markham, R. B., X. Yu, H. Farzadegan, S. C. Ray, and D. Vlahov. 1995. Human immunodeficiency virus type 1 *env* and p17*gag* sequence variation in polymerase chain reaction-positive, seronegative injection drug users. *J. Infect. Dis.* **171**:797–804.
- Martins, L. P., N. Chenciner, and S. Wain-Hobson. 1992. Complex intrapatient sequence variation in the V1 and V2 hypervariable regions of the HIV-1 gp120 envelope sequence. *Virology* **191**:837–845.
- McRae, B., J. A. M. Lange, M. S. Ascher, F. de Wolf, H. W. Sheppard, J. Goudsmit, and J.-P. Allain. 1991. Immune response to HIV p24 core protein during the early phases of human immunodeficiency virus infection. *AIDS Res. Hum. Retroviruses* **7**:637–643.
- Mehta, S. U., K. R. Rupprecht, J. C. Hunt, D. E. Kramer, B. J. McRae, R. G. Allen, G. J. Dawson, and S. G. Devane. 1990. Prevalence of antibodies to the core protein p17, and a serological marker during HIV-1 infection. *AIDS Res. Hum. Retroviruses* **6**:440–454.
- Mulder-Kampinga, G. A., A. Simonon, C. L. Kuiken, J. Dekker, H. J. Scherpbier, P. van de Perre, K. Boer, and J. Goudsmit. 1995. Similarity in *env* and *gag* genes between genomic RNAs of human immunodeficiency virus type 1 (HIV-1) from mother and infant is unrelated to time of HIV-1 RNA positivity in the child. *J. Virol.* **69**:2285–2296.
- Myers, G., B. Rabson, T. F. Smith, J. A. Bertzofsky, and G. N. Pavlakis. 1994. Human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos, N.Mex.
- Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**:418–426.
- Nicholson, J. K., T. J. Spira, C. H. Aloisio, B. M. Jones, M. S. Kennedy, R. C. Holman, and J. S. McDougal. 1989. Serial determinations of HIV-1 titers in HIV-infected homosexual men: association of rising titers with CD4 T cell depletion and progression to AIDS. *AIDS Res. Hum. Retroviruses* **5**:205–215.
- Nixon, D. F., A. R. Townsend, J. G. Elvin, C. R. Rizza, J. Gallwey, and A. J. McMichael. 1988. HIV-1 gag-specific cytotoxic T lymphocytes defined with recombinant vaccinia virus and synthetic peptides. *Nature (London)* **336**:484–487.
- 36a. Overbaugh, J., et al. Unpublished data.
- Overbaugh, J., L. M. Rudensey, M. D. Papanhausen, R. E. Benveniste, and W. R. Morton. 1991. Variation in simian immunodeficiency virus *env* is confined to V1 and V4 during progression to simian AIDS. *J. Virol.* **65**:7025–7031.
- Pachl, C., J. A. Todd, D. G. Kerns, P. J. Sheridan, S. J. Fong, M. Stempien, B. Hoo, D. Besemer, T. Yeghiazarian, B. Irvine, J. Kolberg, R. Kokka, P. Neuwald, and M. S. Urdea. Rapid and precise quantification of HIV-1 RNA in plasma using a branched DNA (bDNA) signal amplification assay. *J. Acquired Immune Defic. Syndr. Hum. Retrovirol.* **8**:446–454.
- Pedersen, C., C. M. Nielsen, B. F. Vestergaard, J. Gersoft, K. Krogsgaard, and J. O. Nielsen. 1987. Temporal relation of antigenemia and loss of antibodies to core antigens to development of clinical disease in HIV infection. *Br. Med. J.* **295**:567–569.
- Reicin, A. S., S. Paik, R. D. Berkowitz, J. Luban, I. Lowy, and S. P. Goff. 1995. Linker insertion mutations in the human immunodeficiency virus type 1 *gag* gene: effects on virion particle assembly, release, and infectivity. *J. Virol.* **69**:642–649.

41. Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406–425.
42. Sei, Y., P. H. Tsang, F. P. Roboz, P. S. Sarin, J. I. Wallace, and J. G. Bekesi. 1988. Neutralizing antibodies as a prognostic indicator in the progression of acquired immune deficiency syndrome (AIDS)-related disorders: a double-blind study. *J. Clin. Immunol.* **8**:464–472.
43. Simmonds, P., P. Balfe, C. A. Ludlam, J. D. Bishop, and A. J. Brown. 1990. Analysis of sequence diversity in hypervariable regions of the external glycoprotein of human immunodeficiency virus type 1. *J. Virol.* **64**:5840–5850.
44. Simmonds, P., L. Q. Zhang, F. McOmish, P. Balfe, C. A. Ludlam, and A. J. Brown. 1991. Discontinuous sequence change of human immunodeficiency virus (HIV) type 1 env sequences in plasma viral and lymphocyte-associated proviral populations in vivo: implications for models of HIV pathogenesis. *J. Virol.* **65**:5266–6276.
45. Vartanian, J. P., A. Meyerhans, B. Åsjö, and S. Wain-Hobson. 1991. Selection, recombination, and G→A hypermutation of human immunodeficiency virus type 1 genomes. *J. Virol.* **65**:1779–1787.
46. von Pöblotzki, A., R. Wagner, M. Niedrig, G. Wanner, H. Wolf, and S. Modrow. 1993. Identification of a region in the Pr55^{gag}-polyprotein essential for HIV-1 particle formation. *Virology* **193**:981–985.
47. Wang, C. T., and E. Barklis. 1993. Assembly, processing, and infectivity of human immunodeficiency virus type 1 *gag* mutants. *J. Virol.* **67**:4264–4273.
48. Wei, X., S. K. Ghosh, M. E. Taylor, V. A. Johnson, E. A. Emini, P. Deutsch, J. D. Lifson, J. D. Bonhoeffer, M. A. Nowak, B. H. Hahn, M. S. Saag, and G. M. Shaw. 1995. Viral dynamics in human immunodeficiency virus type 1 infection. *Nature (London)* **373**:117–122.
49. Wills, J. W., and R. C. Craven. 1991. Form, function, and use of retroviral Gag proteins. *AIDS* **5**:639–654.
50. Wolfs, T. F. W., J. de Jong, H. van der Berg, J. M. G. H. Tunagel, W. J. A. Krone, and J. Goudsmit. 1990. Evolution of sequences encoding the principal neutralization epitope of HIV-1 is host-dependent, rapid and continuous. *Proc. Natl. Acad. Sci. USA* **87**:9928–9942.
51. Wolinsky, S. M., C. M. Wike, B. T. M. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. L. Munoz. 1992. Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. *Science* **255**:1134–1137.
52. Zhang, L. Q., P. MacKenzie, A. Cleland, E. C. Holmes, A. J. L. Brown, and P. Simmonds. 1993. Selection for specific sequences in the external envelope protein of human immunodeficiency virus type 1 upon primary infection. *J. Virol.* **67**:3345–3356.
53. Zhu, T. H. Mo, D. S. Nam, Y. Cao, R. A. Koup, and D. D. Ho. 1993. Genotypic and phenotypic characterization of HIV-1 in patients with primary infection. *Science* **261**:1179–1181.