LA-UR-91-3715

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

LA-UR--91-3715

DE92 003782

TITLE HIV-1 INTRAPATIENT SEQUENCE DIVERSITY IN THE IMMUNOGENIC
V3 REGION

AUTHOR(S) BETTE KORBER
STEVEN WOLINSKY
BARTON HAYNES
KEVIN KUNSTMAN
ROBERT LEVY
MANOHAR FURTADO
PATRICIA OTTO
GERALD MYERS

SUBMITTED TO

For publication in a special edition of AIDS RESEARCH AND
HUMAN RETROVIRUSES: PROCEEDINGS FROM THE INTERNATIONAL
CONFERENCE ON ADVANCES IN AIDS VACCINE DEVELOPMENT.
FOURTH ANNUAL MEETING OF THE NATIONAL COOPERATIVE VACCINE
DEVELOPMENT GROUP FOR AIDS.

By acceptance of this article the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce
the published form of this contribution or to allow others to do so, for U.S. Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

FORM NO. 836 RA
SI NO. 2829 5/81

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

MASTER

HIV-1 Inpatient Sequence Diversity in the Immunogenic V3 Region

Bette Korber* Steven Wolinsky† Barton Haynes†
Kevin Kunstman† Robert Levy† Manohar Furtado†
Patricia Otto† Gerald Myers*

November 12, 1991

*Theoretical Division, LANL, Los Alamos, NM 87545

†Department of Medicine, Northwestern University Medical School, Chicago, IL 60611

†Department of Medicine, Duke University Medical Center, Durham, NC 27710

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Author to receive proofs:

Bette Korber

Theoretical Division, T-10, MS K710

LANL

Los Alamos, NM 87545

(505) 665-4453

Abstract: The third hypervariable domain (V3) of the human immunodeficiency virus type-1 (HIV-1) envelope protein (env) can serve as an epitope for potent type-specific neutralizing antibodies (NAbs) – thus short peptides predicated on a the most commonly found variants of the antigenic tip of the V3 loop have been considered as potential candidates for an HIV peptide vaccine. To evaluate the extent of inpatient variation in the immunogenic crest of the V3 loop, sequence sets were analyzed from individuals for whom multiple V3 sequences were available. Several strategies for selecting the best sets of hexapeptides to represent the variable tip of the V3 loop were considered (including those of Holley, Goudsmit, and Karplus, Proc. Natl. Acad. Sci. USA, 88:6800-04 (1991)), and their effectiveness was evaluated by comparing them with the sequence sets from individuals. Most individuals carried at least one, and frequently many, variants that did not match any of the sequences from among the ten most common hexapeptides. Inpatient viral sequence variation was increased by including sequences derived from brain biopsy specimens as well as from blood. Additionally, sequences obtained from brain specimens of different individuals had common elements which were not conserved in the corresponding blood samples, suggesting that certain amino acids in the V3 loop may be requisite for viral propagation in the CNS.

MATERIALS AND METHODS

Sequence Sets. The sequences used for this study are summarized in the Human Retroviruses and AIDS database¹; most have been published, although some were kindly provided to the database prior to publication for use in consensus analyses. If multiple sequences from a single individual were identical across the V3 loop, only one representative was included in the data sets. There were 582 V3 sequences from 233 people in the North American and

European (NA-E) set; the African set includes 108 sequences from 37 people. Forty sequence sets from individuals, or sibling sequence (sib-seq) sets, were taken from the NA-E data set and used to study intrapatient variation in the tip of the V3 loop. The HIV-1 sib-seq sets were selected on the basis of having a minimum of four sequences available per individual. The amino acid sequence alignment and similarity comparisons were done using MASE¹, and various UNIX C-shell scripts were used to perform the analyses.

Derivation of the Most Common Hexapeptides From the Tip of the V3 Loop.

Hexapeptides from the tip of the V3 loop can elicit potent NAb^{2,3}; in particular, the hexapeptide GPGR^{AF} can stimulate NAb that are cross-reactive with divergent HIV-1 strains that share GPGR^{AF} but have different surrounding amino acids². To select the most common hexapeptides found in the defined antigenic region, we simply extracted the six positions that align with either IIIGPGR or GPGR^{AF} of the V3 consensus sequence, sorted them, and determined the frequency of each unique type of hexapeptide found in the NA-E sequence set. The five or ten most common forms were chosen as a hypothetical hexapeptide vaccine (Table 1), and compared with the sib-seq sets to determine how many individuals would have all their V3 sequences matched by the hexapeptides in the cocktail (Table 2). As of most of the NAb epitopes defined within the V3 loop overlap with the four positions that align with GPGR^{2,3,4}, we also looked at the overall variation in those four positions in both the African and the NA-E sets.

Holley et al.⁴ used two alternate approaches (simulated annealing or a greedy algorithm) to select hexapeptides that best represent the tip of the loop, based on the 245 V3 sequences from LaRosa et al.⁵. They imposed the constraint that the hexapeptides contain the part of

the alignment corresponding to the GPG at the crest of the loop. Using these two methods, they obtained slightly differing sets of hexapeptides (see the legend to Table 2). The set of the ten best hexapeptides from either method covered approximately 88% of the LaRosa sequences⁴, and 82% of the larger NA-E data set described here (which includes the LaRosa set). In the inpatient study presented here, we compared the ten best hexapeptides selected by their two algorithms⁴ with the 40 viral sib-seq sets (Table 2).

Blood-Brain Sequence Sets. Samples were simultaneously taken from the peripheral blood and from brain biopsy specimens of four different patients. DNA was extracted from peripheral blood mononuclear cells or homogenized brain tissue, and HIV-1 DNA was amplified and cloned as previously described⁶. Between 15 and 55 sequences were obtained from each individual. These sib-seq sets formed the basis for the following discussion on blood-brain variation. Statistical analysis of the blood-brain sequence similarities used SAS release 6.02 (SAS Institute, Inc, Cary, NC).

RESULTS

Inpatient Variation in the Tip of the V3 loop. By far the most common tetrapeptide sequence found in the tip of the loop in the NA-E set is GPGR, present in 79.2% of the sequences. The second most common is GP GK, present in 7.6% of the sequences. There are 22 additional possibilities that occur in these four positions in the NA-E alignment, most of which conserve the propensity to form a type II β -turn^{1,3}. Despite the overall conservation of GPGR, most HIV-1 sib-seq sets showed some variation in this motif; 35/40 sib-seq sets had at least one sequence with GPGR at the tip, but only 15/40 had exclusively GPGR. Among African sequences, the variation was considerably greater^{1,7}; GPGR was not the predominant

form – GSGQ, GPGQ, and GLGQ were more common. Twenty different forms of this motif were found in the African set, and a total of 39 forms were found in the combined NA-E and African sets. While the proline in the second position is highly conserved in the NA-E sequences (97% P), this position is variable in the African set (39% P)^{1,7}. The conserved arginine (86% R) in the fourth position of the NA-E sequences is replaced by glutamine (75% Q) as the most common amino acid in this position in the African data^{1,7}. The majority of the African HIV sib-seq sets also show inpatient variation in this motif.

Inpatient Hexapeptide Epitope Variation. There were 63 unique forms of the hexapeptides that align with GPGRF in the NA-E data set, and 40 unique forms in the African data set; combined, a total of 97 unique forms were found. Among the hexapeptides that align with HGPGR, there were 75 unique forms in the NA-E data set, 41 unique forms in Africa, and a total of 108 unique forms in the combined sets. Table 1 shows the ten most common forms of the hexapeptides found in the NA-E data set, and their frequencies.

In general, the sets of ten optimal hexapeptides selected by the algorithms of Holley et al.⁴, and the set selected here from the GPGRF alignment, matched approximately 80% of the sequences from the 582 NA-E sequences. However, when comparing the hexapeptide sets to sib-seq sets, at most half of the sib-seq sets were covered. The hexapeptide set based on the HGPGR alignment was less efficient than the other three sets, and combining peptides from the HGPGR set and the GPGRF set didn't substantially increase the coverage of sequences (data not shown). An important feature of the GPGRF set, as well as of the sets selected in Holley et al.⁴, was that after the first five or so most prominent peptides, the selection process using any of the algorithms was reduced to choosing rare forms that

were present in only 1-2% of the samples, and some of these were likely to have been derived from a single individual. There was little statistical merit in selecting these peptides over others, and their inclusion was likely to be a consequence of the particular data set used to make the hexapeptide selections. An additional trend was observed: the more sequences in a sibling set, the less likely the sibling set was to be covered by the various hexapeptide sets (Fig. 1). This raises the possibility that if enough HIV sibling sequences were obtained from any individual, “unusual” forms of the tip of the V3 loop would be encountered.

Blood-Brain Heterogeneity. In the four sib-seq sets that contain HIV-1 sequences derived from brain as well as the blood, we found that when blood sequences were compared to brain sequences there was somewhat greater overall divergence than when only blood sequences were considered. Pairwise similarity comparisons (discounting gaps) of the nucleotide sequences encoding the V3 loop and the region just upstream of the loop (153 bases) were made between sequences in each sibling set. When just the sib-seqs from the blood sample were compared (the number of pairwise similarity scores for the four patients, or N, equals 1950), the mean similarity was 94.8% between sequences. When inpatient blood samples were compared to brain samples (N = 741) the average similarity dropped to 92.5%. Because the distributions of values are not Gaussian, we used a Wilcoxon two sample t test and found that these two sets of similarity scores were significantly different ($p = 0.0001$).

Our preliminary analysis revealed an additional interesting feature worth noting. There were particular positions in the V3 loop which were conserved between the brain samples of the four individuals, but were variable in the blood samples. These sites include the two N-linked glycosylation sites closest to the amino terminal cysteine of the V3 loop, and a set

of several other amino acids scattered throughout the loop (the amino acids that were most strikingly conserved in the brain are shown in capital letters, relative the V3 consensus¹: **SihigpgrAfytTGeli**).

DISCUSSION

There are many limitations to this kind of sequence analysis of intrapatient HIV-1 variation of antigenic regions. Sequence error (introduced either during the amplification or the sequencing reactions) and amplification of biologically inactive clones may result in overestimates of the overall variability in an individual. Conversely, selective amplification imposed by the PCR primer pairs may result in certain viral forms being undetected or under-represented. Also, opposing influences beyond primary sequence diversity can effect the ability of V3 loop sequences to stimulate protective immunity: antigenic cross-reactivity of variants with different amino acid sequences would help, while changes outside of the loop that alter the structure of the tip and hence its antigenic specificity would detract. Finally, the sequence sets used here came from a diverse group of people, and no selection was made based on health status or any other parameter which might be useful for eventual study. However, despite these limitations, the results presented here provide a basis for evaluating what might be expected in terms of the range of variability present in an individual.

The African samples reveal the remarkable potential for variation in the V3 loop. Since the diversity between HIV-1 sequences obtained from different people seems to be increasing with time (for example, two isolates obtained in 1990 from two individuals in the US are likely to be substantially more divergent from each other than two isolates obtained in 1985), and since vaccine trials may actually select for unusual variants, the *potential* for diversity

may be as important to consider as the range of diversity in a given population at the time of sampling.

When HIV-1 sequences obtained from brain as well as from blood were included in sib-seq analysis, greater overall inpatient heterogeneity of sequences was found. This increased variation may be important to consider in studies of vaccination of HIV infected individuals. An additional, though preliminary, observation resulting from the blood-brain comparison was that a pattern of conserved amino acids was found when comparing brain derived sequences from different individuals, relative to blood derived sequences. A conserved brain signature pattern raises the intriguing possibility that differences in an individual's immune response to the subset of viruses that carry the signature (and thus are able to infect the CNS) may dictate that individual's propensity for AIDS related dementia.

Any realistic vaccine design which depends critically upon the immune response to the V3 loop must address the question of variation. Many groups are looking at this issue through studies of antigenic cross reactivity, molecular modeling techniques to estimate the structural consequences of amino acid replacements, and serotyping. The potential for antigenic variation in V3 also serves to reemphasize the importance of studying conserved epitopes, whether CTL epitopes, contiguous NAb epitopes from other regions, or conformational NAb epitopes, which give promise of broad reactivity.

ACKNOWLEDGMENTS: We thank James Theiler and April Gifford for thoughtful reading of the manuscript, and David J. Wright for assistance with statistical analysis.

FIG. 1. Distribution of viral sibling sequence sets. Each box represents an individual. The numbers along the abscissa represent the number of V3 sequences available in the sibling set, while the ordinate shows how many individuals had that number of sequences. The shaded boxes indicate that at least one of sequences in the set was not matched by the ten best hexapeptides from A.) the GPGRAPH alignment, or B.) simulated annealing⁴.

REFERENCES

1. Myers G, Korber B, Berzofsky J, Smith R, Pavlakis GN, eds. Human Retroviruses and AIDS, A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences 1991; published by Los Alamos National Laboratory, T-10, Los Alamos, NM.
2. Javaherian K, Langlois AJ, LaRosa GJ, Profy AT, Bolognesi DP, Herlihy WC, Putney SD, Matthews TJ: Broadly Neutralizing Antibodies Elicited by the Hypervariable Neutralizing Determinant of HIV-1. *Science* 1990; **250**: 1590-92.
3. Goudsmit J, Debouck C, Meloen RH, Smit L, Bakker M, Asher DM, Wolff AV, Gibbs CJ, and Gajdusek DC: Human immunodeficiency virus type 1 neutralizing epitope with conserved architecture elicits early type-specific antibodies in experimentally infected chimpanzees. *Proc. Natl. Acad. Sci. USA* 1988; **85**:4478-82.
4. Holley LH, Goudsmit J, and Karplus M: Prediction of optimal peptide mixtures to induce broadly neutralizing antibodies to human immunodeficiency virus type 1. *Proc. Natl. Acad. Sci. USA* 1991; **88**, 6800-6804.
5. LaRosa GJ, Davide JP, Weinhold K, Waterbury JA, Profy AT, Lewis JA, Langlois AJ, Dreesman GR, Boswell RN, Shadduck P, Holley LH, Karplus M, Bolognesi DP, Matthews TJ, Emmini EA, and Putney SD: Conserved Sequence and Structural Elements in the HIV-1 Principal neutralizing Determinant. *Science* 1990; **249**: 932-935.
6. Wolinsky SM, Wike CM, Korber B, Hutto C, Parks W, Rosenblum L, Kunstman K, Furtado M, and Munoz J: Selective transmission of Human Immunodeficiency Virus Type 1 (HIV-1) Variants from Mothers to Infants. Submitted to *Science*.
7. Korber BK, Wolinsky SM, and Myers G: Analysis of Inter- and Intrapatient Variation

in the V3 loop of the HIV-1 Envelope Protein. in press, *Vaccines 92*, Cold Spring Harbor press, New York.

Site 1			Site 2		
<u>Hexapeptide</u>	<u>Number</u>	<u>Percentage</u>	<u>Hexapeptide</u>	<u>Number</u>	<u>Percentage</u>
GPGRA ^F	330	56.7	HIGRGR	176	30.2
GPGKAF	36	6.2	PIGPGR	50	8.6
GPGRA ^V	21	3.6	NIGPGR	48	8.2
GPGRV ^F	21	3.6	PMGPGK	28	4.8
GPGRA ^I	20	3.4	SIGPGR	23	4.0
GPGR ^T V	11	1.9	TLGPGR	23	4.0
APGRA ^F	10	1.7	HMGPGR	20	3.4
APGSAW	8	1.4	TIGPGR	19	3.3
GPGRV ^L	7	1.2	HIAPGR	17	2.9
GPGRV ^V	7	<u>1.2</u>	RIGPGR	15	<u>2.6</u>
Total:		80.9%	Total:		72.4%

Table 1. The most common hexapeptides found at the crest of the loop in the North American and European data set. The first column for each site shows the hexapeptide sequence, the second shows the number of times the sequence was found in among the 582 sequences that comprised the set, and the last column shows the percentage of sequences that included each hexapeptide.

Peptide set		% of sequences covered	% of individuals covered
Aligned GPGRAF	10 best	80.9	42.5 (17/40)
	5 best	73.5	37.5 (15/40)
Aligned HIGPGR	10 best	72.0	27.5 (11/40)
	5 best	55.8	25.0 (10/40)
Greedy algorithm Holley et al. (4)	10 best	81.8	50.0 (20/40)
	5 best	75.0	45.0 (18/40)
Simulated annealing Holley et al. (4)	10 best	81.6	47.5 (19/40)
	5 best	76.0	45.0 (18/40)

Table 2. Coverage of the 5 and 10 best hexapeptides in individuals and in the complete set. The peptides shown in Table 1 were compared in this table to the North American and European data set and to the forty individuals for which we had four or more sequences. In addition, the ten best peptides based on Holley et al. were also used for comparison. These were: greedy algorithm, IGPGRA, GPGRAF, IHIGPG, GPGRVY, GPGKAF, IFIGPG, GPGRRF, GAGRAI, IHIAPG, PIGLGQ; and simulated annealing, IGPGRA, GPGRAF, IHIGPG, TMGPGR, GPGKAF, IFIGPG, GPGRRF, LGPGRV, TTGPGR, GAGRAI. "Covered" means that the percentage of the sequences shown had perfect homology with a member of the hexapeptide set, or that the percentage of individuals had all of their sequences perfectly matched by the hexapeptides in the set.

A.

B.