LA-UR--91-3716

DE92 003781

TITLE ANALYSES OF INTER- AND INTRA-PATIENT VARIATION IN THE V3 LOOP OF THE HIV-1 ENVELOPE PROTEIN

AUTHOR(S) BETTE KORBER
STEVEN WOLINSKY
GERALD MYERS

SUBMITTED TO For publication in VACCINES '92 Cold Spring Harbor Press
Cold Spring Harbor, NY

DISCLAIMER

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

FORM NO 836 R4
ST NO 2629 5/91

Analyses of Inter- and Intrapatient Variation in the V3 loop of the HIV-1 Envelope Protein

Bette Korber* Steven Wolinsky† Gerald Myers*

September 17, 1991

*Theoretical Division, LANL, Los Alamos, NM 87545

†Northwestern University Medical School, Chicago, IL 60611

Running Head: HIV-1 V3 Loop Variation

Author to receive proofs:

Bette Korber

Theoretical Division, T-10, MS K710

LANL

Los Alamos, NM 87545

(505) 665-4453 (phone)

(505) 665-3493 (fax)

email: btk@life.lanl.gov

The third hypervariable domain of the HIV-1 gp120 envelope protein (V3) has been the focus of intensive sequencing efforts. To date, nearly one thousand V3 loop sequences have been stored in the HIV sequence database (Myers et al., 1991). Studies have revealed that the V3 loop elicits potent type-specific immune responses, and that it plays a significant role in cell tropism and fusion (Myers et al., 1991; see section III for references to biological features of the V3 loop, defined epitopes within the loop, and specific sequence data sets). The immunogenic tip of the loop can serve as a type-specific neutralizing antibody epitope, as well as a cytotoxic T-cell epitope. A helper T-cell epitope that lies within the amino terminal half of the V3 loop has also been characterized. Despite the richness of the immunologic response to this region, its potential for variation makes it an elusive target for vaccine design. Analyses of sibling sequence sets (sets of viral sequences derived from one person) show that multiple forms of the immunogenic tip of the loop are found within most HIV-1 infected individuals. Viral V3 sequences obtained from epidemiologically unlinked individuals from North America and Europe show extensive variation. However, some amino acid positions distributed throughout the V3 loop are highly conserved, and there is also conservation of the charge class of amino acid able to occupy certain positions relative to the tip of the loop. By contrast, the sequences obtained from many countries throughout the African continent reveal that V3 is a remarkably fluid region with few absolute constraints on the nature of the amino acids that can occupy most positions in the loop. The high degree of heterogeneity in this region is particularly striking in view of its contribution to biologically important viral functions.

V3 Loop Variation in American, European and African Sequences

V3 loop amino acid sequences obtained from N. America and Europe, while quite variable, nonetheless show amino acid conservation in certain positions in an alignment (Figure 1). The GPGR motif, located at the tip of the loop, is found in 78% of the North American and European sequences. It is a component of the defined neutralizing antibody epitopes and cytotoxic T-cell epitope, and is predicted to form a type II β-turn (LaRosa et al.,1990). Among the N. American sequences, conservation of the propensity to form such a turn is extremely high, even when there are variants in the precise sequence (Myers et al., 1991).

The sequences from Africa, on the other hand, show extreme variability. The range of divergence between pairs of amino acid sequences is illustrated in Figure 2. Two histograms (one for the African set, and one for the North American and European set) show the similarities distributions of pairwise comparisons of all sequences within each set. Among the African set, radically divergent V3 sequences that share as few as 20% of their amino acids can be found, with an average similarity of 59% between sequences. In the North American and European set, by contrast, the most divergent isolates share 44% of their amino acids, and the average similarity is 77%. The frequency of the most common amino acid at a given site in the African alignment is generally less than the frequency of the corresponding most common amino acid in the North American and European alignment (Figure 1). GPGQ, not GPGR, is the most common amino acid motif found at the tip of the loop in African sequences, and this sequence is also predicted to form a type II β-turn. However, GLGQ is common in sequences from all over Africa, and it is *not* predicted to form a type II β-turn. If

the North American and European amino acids found at given positions in the alignment are grouped into classes based on chemical similarities (Smith and Smith, 1990) an amino acid class that encompasses most of the range of amino acids at a given position can usually be discerned (Figure 3). For most positions in the V3 loop in the African sequence alignment, no such amino acid class covering pattern can be obtained (the only exceptions are the sites proximal to the perfectly conserved cysteines at either boundary of the loop, and a single site at the tip (Figure 3)).

Intrapatient viral sequence variation in the V3 loop

Many of the V3 sequences in the sets described above are generated through PCR amplification of viral DNA from peripheral mononuclear cells, and subsequent sequencing of multiple viral variants from a single individual. Such viral sibling sets tend to be more closely related to each other than to other non-epidemiologically linked sequences from the general population (Balfe et al., 1990), and often contain multiple sequences that are virtually identical throughout the V3 loop (Myers et al., 1991). Despite this tendency towards conservation relative to the population as a whole, examination of large sibling sequence sets (from four to fifty V3 sequences per person) reveals that most infected individuals carry viruses which vary in the immunogenic tip of the loop and its neighboring positions. These positions comprise the principal neutralizing antibody sites and a cytotoxic T-cell epitope.

Intrapatient variation in viral V3 sequences is extended by sequencing samples derived from brain biopsy as well as from peripheral blood. Pairwise similarity comparisons (discounting gaps inserted to maintain alignment) of multiple nucleotide sequences de-ived

from simultaneously taken blood and brain samples from three patients illustrate this point (Wolinsky et al., 1991a). From 972 pairwise comparisons of viral sibling sequences from the blood of three patients, the median similarity between sibling sequences was found to be 95%, and the interquartile range of similarities was 93-97%. On the other hand, a comparison of intrapatient sequences derived from the blood with sequences derived from the brain (1153 comparisons) yielded a median similarity of 92% with an interquartile range of 90-94%. Therefore, at least in the three patients sampled, inclusion of brain sequences as well as blood sequences in sibling sets significantly increased the overall intrapatient heterogeneity of V3 sequences.

Multiple V3 sequences from three mother-infant transmission pairs have been examined to investigate perinatal HIV-1 infection (Wolinsky et al., 1991b). Three interesting features of the sequences obtained from the infants soon after birth suggest that selection plays an important role in determining which viral forms from the mother are transmitted to her baby. First, the V3 loop sequences derived from a given infant are far more homogeneous than those derived from its mother. Second, in two of the three mother-infant pairs, an unusual form of the virus from the mother is the predominant form in the infant. And third, a highly conserved glycosylation site, present in most infected individuals and some of the sequences from each of the mothers, was absent in all of the sequences from the infants. Taken together, these results suggest that only a small subset of sequences are successfully transmitted from mother to infant. Alternative selective processes may determine which of the viruses are transmitted: the mother's immune response may inhibit transfer of some of

the viral forms, cell tropism may dictate which forms are transferred across the placenta, or viral replication rates may cause certain forms to prevail in the infant.

Summary

In general, the high degree of variability seen in the V3 loop makes it a difficult target for vaccine design. As evidenced by the extraordinary degree of heterogeneity in this region, it is apparent that the viral envelope can successfully adopt many unique and highly divergent V3 sequence variants. Sufficient variation in the V3 region exists within individuals such that a type-specific vaccine designed to trigger an immune response to the most common forms of V3 may select unusual forms from the pool of variants which coexist in an individual. However, because of the biological significance of this region in terms of viral tropism and fusion, there are probably patterns of variation and structural elements within the V3 loop that are conserved despite overall sequence variation. A new direction we are exploring looks for positions which co-vary, and thus show evolutionary interdependence. We have begun such analyses considering the very large database of V3 sequences, as well as the smaller set of complete envelope sequences, which allows us to study V3 variation in the context of the intact env protein.

Acknowledgements

We thank the many contributors to the AIDS and Human Retroviruses Database who made their sequences available for inclusion in consensus sequence analyses prior to publication. In particular, Dr. Ellen Murphy made a large set of African sequences available and Dr. Chin-Yi Ou contributed a large North American set. Also many thanks to April Gifford

for help with analyses and graphics and to James Theiler for editorial comments.

Figure Legends

1. V3 loop amino acid consensus sequences. The most common amino acid found in each position in the V3 loop is shown; the frequency with which it occurs is given as a percentage directly below the amino acid. (Note the generally lower frequencies in the African set.) Sites where only small fraction of sequences have inserted amino acids, and most sequences have only gaps which were included to maintain the alignment, are not shown. The average length of the V3 loop is 35 amino acids; the range of the African set is 30 to 40 amino acids. Positions where the most common amino acid in the African set differs from the North American and European set are framed by a darkened box. The data sets used for these consensus sequences are comprised of only unique V3 sequences from any given individual. There are 551 sequences included in the North American and European set, and 108 sequences in the African data set (Myers et al., 1991)

2. Histograms illustrating the range of divergence in V3 loop amino acid sequences. Pairwise comparisons were made to determine the extent of sequence similarity (discounting gaps inserted to maintain the alignment) between all V3 sequences in each of the data sets described in Figure 1. The similarity score for each pair of sequences was rounded off to the nearest percentage, and the number of comparisons with that score was tallied. The histograms plot the percent similarity versus the fraction of the total number of comparisons which had that similarity. 9180 comparisons were made between African sequences to generate the histogram on the left; 134,421 comparisons were made for the North American and European set shown on the right. The highest similarity scores can be

generally be attributed to comparisons of unique but similar V3 sequences from intrapatient sequence sets.

3. Amino acid class covering patterns for V3 loop sequences. A. Amino acid classification scheme used to group amino acids by chemical properties (Smith and Smith, 1990, and Myers et al., 1991). The character "g" is written when a gap occurs in a given position. B. Amino acid class covering patterns for 95% of the amino acids at any given site in the alignments described in Figure 1. 95% was used rather than 100% to allow for minor discrepancies due to sequencing error, sequencing of biologically inactive clones, or sub-optimal alignments. Directly beneath the class covering pattern is a column of amino acids which can occupy each position, from most common to least common, such that 95% of the amino acids are accounted for in each site.

References

Balfe, P., P. Simmonds, C.A. Ludlam, J.O. Bishop, and A.J. Leigh 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.

LaRosa, G.J., J.P.Davide, K. Weinhold, J.A. Waterbury, A.T. Profy, J.A. Lewis, A.J. Langlois, G.R. Dreesman, R.N. Boswell, P. Shadduck, L.H. Holley, M. Karplus, D.P. Bolognesi, T.J. Matthews, E.A. Emini, and S.D. Putney. 1990. Conserved Sequence and Structural Elements in the HIV-1 Principal neutralizing Determinant. *Science* **249**, 932-935.

Myers, G., B. Korber, J. Berzofsky, R. Smith, G.N. Pavlakis, eds. 1991. Human Retroviruses and AIDS, A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences, published by Los Alamos National Laboratory, T-10, Los Alamos, NM.

Smith, R.F. and T.F. Smith. 1990. Automatic Generation of Primary Sequence Patterns from Sets of Related Protein Sequences. *Proc. Natl. Acad. Sci. USA* **87**, 118-122.

Wolinsky, S.M., C.M. Wike, B. Korber, C. Hutto, W. Parks, L. Rosenblum, K. Kunstman, M. Furtado, and J. Munoz. 1991a. Selective transmission of Human Immunodeficiency Virus Type 1 (HIV-1) Variants form Mothers to Infants. Submitted to *Science*.

Wolinsky, S.M., K. Kuntsman, M. Furtado, R. Levy, and B. Korber. 1991b. HIV-1 Sequence Variation Between Simultaneously Obtained Blood and Brain Isolations. Manuscript in preparation.

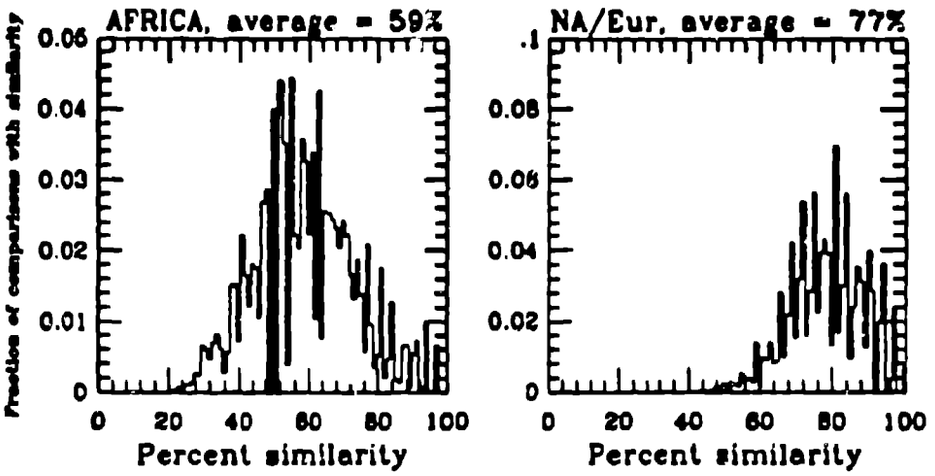
V3 consensus sequence, North America and Europe:

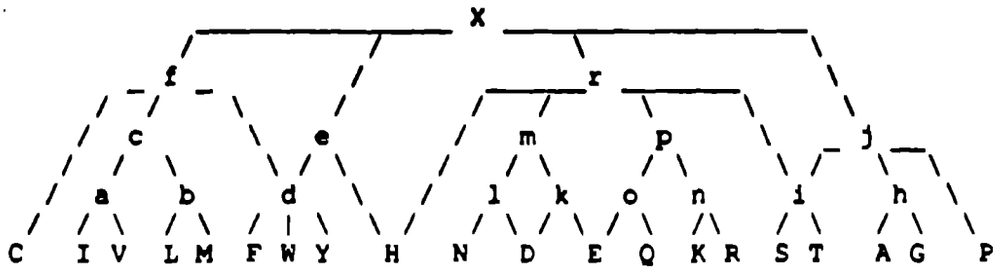
C T R P N N N T R K S I H I G P G R A F Y T T G E I I G D I R Q A H C

100 87 99 98 85 93 94 94 88 75 64 94 45 76 95 97 93 86 86 76 82 54 88 75 32 95 81 98 84 94 98 81 100 90 99

Africa:

100 62 94 100 62 52 51 60 80 39 35 44 37 72 90 39 92 75 69 42 81 46 82 44 27 56 44 77 60 85 70 69 94 67 10(


AFRICA, average = 59%
NA/Eur, average = 77%
Fraction of comparisons with similarity
0.06
0.04
0.03
0.02
0.01
0
.1
0.08
0.06
0.04
0.02
0
0 20 40 60 80 100
Percent similarity
0 20 40 60 80 100
Percent similarity


X
f
r
c
e
m
p
j
a
b
d
l
k
o
n
i
h
C
I V
L M
F W Y
H
N
D
E
Q
K R
S T
A G
P

AACC pattern for ≥ 95% of sequences, North America and Europe:

C	X	R	P	r	r	r	j	r	p	X	a	X	c	h	P	G	p	X	f	X	j	X	g	g	I	g	G	l	c	R	p	A	e	C
T				N	N	N	T	R	K	S	I	H	I	G			R	A	F	Y	T	T	G	E		I		D	I		Q		H	
I				S	T	T	A	K	R	G	V	P	M	A			K	V	I	H	A	A	--	D		V		N	M		K		Y	
								S	Q	R		N	L				Q	T	W	V		I	E	Q		T					R			
												T							V	F		R	K	R		--								
												R							L				R	K										
												S											D	A										
												Y												N										
																								H										
																								--										

AACC pattern for ≥ 95% of sequences, Africa:

C	X	R	P	X	X	g	X	X	g	g	g	X	X	X	X	X	p	g	X	X	X	X	g	g	X	g	X	X	g	X	p	j	e	C
T				Y	N	N	T	R	Q	R	T	R	I	G	P	G	Q	A	F	Y	T	T	G	K	I	I	G	D	I	R	Q	A	H	
S				N	K	K	I	I	K	S	I	P	R	K	S	K	R	T	L	H	A	Y	-	D	T	T	R	Y	T	G	K	P	Y	
I				S	R	--	K	K	R	N	V	H	M		L		K	V	Y	Q	R	I	K	R	K	V	K	N	M	K	R			
				G	Y	S	M	S	T	K	--	S	F		Q			--	W	V	K	S	N	N	G	R	N	V	--						
					T	Y	V		I	G	A		L						S		V		R	G	P	Q		R						
					S	A			--	--													R	A	L	--		I						
					D					T														--	R	G		K						
					H																			L		E								
																								S		F								
																								Y		S								
																								V										
																										t								