Insertion of N-Linked Glycosylation Sites in the Variable Regions of the Human Immunodeficiency Virus Type 1 Surface Glycoprotein through AAT Triplet Reiteration

MARNIX L. BOSCH,1* ARNO C. ANDEWEG,1† RONALD SCHIPPER,2 AND MARCEL KENTER1‡

Laboratory of Immunobiology, National Institute for Public Health and Environmental Protection, 3720 BA Bilthoven, and Laboratory of Immunohaematology and Blood Bank, University Hospital E3-Q, 2333 AA Leiden, The Netherlands

Received 20 May 1994/Accepted 26 July 1994

Variable regions with sequence length variation in the human immunodeficiency virus type 1 envelope exhibit an unusual pattern of codon usage with AAT, ACT, and AGT together composing >70% of all codons used. We postulate that this distribution is caused by insertion of AAT triplets followed by point mutations and selection. Accumulation of the encoded amino acids (asparagine, serine, and threonine) leads to the creation of new N-linked glycosylation sites, which helps the virus to escape from the immune pressure exerted by virus-neutralizing antibodies.

The surface glycoprotein gp120 of the human immunodeficiency virus type 1 (HIV-1) contains a number of regions with a relatively high degree of amino acid sequence variation, designated variable regions 1 through 5 (V1 to V5 [11]). Recently, evidence has been presented that, like variation in, e.g., the V3 region, variation in the V1 and V2 regions contributes to the determination of viral tropism and cytopathic potential (2, 3, 8, 17, 19), indicating that these regions are probably involved in the process of envelope-mediated membrane fusion resulting in virus entry or in syncytium formation. Sequence variation in these regions can then be reflected in these processes, resulting in variations in cell tropism and cytopathicity as described above. Virus-neutralizing antibodies to the V2 region that probably directly interfere with virus entry have also been described previously (7, 10a, 14), and sequence variation could enable the virus to escape from such antibodies. The addition of carbohydrates can alter the recognition of envelope glycoproteins by the immune system (1, 6). Here we describe a process of sequence length variation in the HIV-1 envelope gene that results in the selective insertion of new N-linked glycosylation sites in the variable regions of the envelope glycoproteins as a mechanism to escape immune pressure exerted by neutralizing antibodies.

The alignments of the variable-length sequences found in V1 and V2 are shown in Fig. 1. We have used the HIV-1 envelope sequence alignment presented by Myers et al. (12) as a starting point and readjusted the alignment by eye. The numbers at the top indicate the codons incorporated in the regions of variable sequence length. The underlined blocks of three codons indicate encoded potential N-linked glycosylation sites (see below). We have noticed an apparent skewed usage of codons in these regions. A relatively high number of codons begin with A

To confirm this observation, we have counted the frequencies of the four nucleotides adenosine (A), cytidine (C), guanosine (G), and thymidine (T), at the first, second, and third positions of each codon in the variable-length sequences of the V1 and V2 regions. We find sharp deviations from the distributions as found in the whole of gp120 (we have used the HXB2R gp120 sequence for comparison). Most notably, we find that approximately 84% of the codons in these regions start with A (<1% start with C); only 1 to 4% of codons have a T at position 2 (compared with 26% in gp120); and 67 to 76% of codons end with T, with sharp decreased frequencies of A and G at position 3. We have subsequently analyzed the frequency distributions of all four nucleotides at each codon position in the whole envelope gene alignment by a sliding window method (see legend to Fig. 2 for details). The results obtained for A at position 1 (A1), T at position 2 (T2), and T at position 3 (T3) are plotted in Fig. 2 for a window size of 20 codons shifted by 1 codon, as are the number of sequences counted in each window to map the regions with sequence length variation (bottom panel). It can be seen that sharp increases in frequencies of A1 and T3 with concomitant decreases in T2 occur in the V1 and V2 regions and also in two other regions that display a high degree of sequence length variation, the V4 and V5 regions. No other regions of the envelope gene display similar concomitant nucleotide frequency distributions. The A1 and T3 peaks are not as pronounced for V5 because of the relatively small size of the region of variable sequence length in V5 relative to the window size.

We postulate here that the skewed distribution of nucleotides at the first, second, and third positions of the codons that make up the regions of variable sequence length is the result of the insertion of AAT triplets at these sites followed by point mutations and selection for sequences that provide a selective advantage to the virus by the insertion of N-linked glycosylation sites, resulting in escape from the immune system, as well as selection against disadvantageous (in this case, hydrophobic) sequences. The implications of this mechanism will be discussed below.

The use of a severely limited set of nucleotides at position 1

and end with T, with an especially high frequency of AAT codons.

^{*} Corresponding author. Present address: Regional Primate Research Center and Department of Pathobiology SC-38, University of Washington, Seattle, WA 98195. Phone: (206) 543-7619. Fax: (206) 543-3873.

[†] Present address: Laboratory of Viral Pathogenesis, Biomedical Primate Research Center, Rijswijk, The Netherlands.

[‡] Present address: Institute of Virology, Erasmus University, Rotterdam, The Netherlands.

Vol. 68, 1994 NOTES 7567

λ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	
CONSENSUS	gATttga?g ???	???	???	???	???	???	???	???	???	???	???	222	AAt	act	act	??t	???	???	AAt.	a?t	a?t	agt	202	222	222	222	222	222224424424
HIVJRCSF	GGTG												T	CCT	ACT				T	ACC	ACT	ACT	ACT	ACT				CACCCAATC
HIVJRFL	GGTG												T	GCT	ACT				T	ACC	ACT	AAT	GAT	AGC				GAGGGAACG
HIVALA1 HIVBRVA	GTATTTG												T	GCT	ACT				T	ACC	ACT	AGT	AGT	AAC				CCCACAACA
HIVJH3	GTTC	• • •	227	CAT	ACT	AGC.	ccc	• • •	• • •	• • •	• • •	• • •	T	GCT	ACT	• • •	• • •	• • •	T	GCC	ACT	AGT	AAT	AGC	• • •	• • •	• • •	GGGAAAATG
HIVSC	GTGGGGG ATTGAGG	•••	AAT	GAT	ACT	AGC	ACC	•••	• • • •		•••	• • •	T	CCT	ACT	• • •	• • •	• • •	T	ACC	ACT	AGT	AGT		• • •	• • •	• • •	TGGGGGGAAAAG
HIVBAL1	GTTGAGG		AAT	GCT	ACT	AAT	GGG						Т	GAC	ACT				T	ACC	ACT	ACT	ACT	ACC				ACCCCC A A TC
HIVYU2	GTTAAGG												T	GCT	ACT				T	ACC	ACT	ACT	ACT	ACC				TOGGANACO
HIVMN	GTTGAGG		AAT	ACT	ACT	AAT	ACC	AAT					T	AGT	ACT	CCT	AAT			AAT	ACT	AAT	ACC					CACCCAACA
HIVHXB2R	GTTGAAG												T	CAT	ACT	AAT	ACC		T	ACT	ACT	ACC						CCCACAAMCAMA
HIVLAI HIVNL43																												
HIVMFA	GTTGAAG GTTGAAG	• • •	• • •	•••	• • •	• • • •	•••	• • •	• • •	• • •	• • •	• • •	T	GAT	ACT	AAT	ACC	•••	T	AGT	AGT	AGC	• • •	• • •	•••	• • •	• • •	GGGAGAATGATA
HIVCAM1	AGTA	• • •	• • • •	•••	• • •			• • •	•••	• • •	• • •	• • • •		DAT	ACT	AAI	ACC	• • •	T	AGT	AGT	CAT	• • •	• • •	TCC			GGGAGAATGATA
HIVNY5CG	GTTGACG												T	രസ	ACT	TAT	CCC		T	CCT	ACT	ACT						
HIVADA	GTIGAGG												T	GTT	ACT	AAT	ATC		T	AAT	ACT	ACT						CACCCA
HIVJFL	AIIGAAG												T	CTT	ACT	ΔΔΤ	ATC		T	AAT	ACT							TCCCCCAAAC
HIVSIMI84	AIGGAAG AGI	GCT											T	AGT	ACC	ACT	AAT	ACC	T	ACT	ACT							ተተረ
HIVD31	GCTGAAG												T	GCT	ACT	AAT	ACC	AAT	T	AGT	AGT	TGG	ACG	ATG				ACACCACA A
HIVSF162 HIVOYI	ATTGAAG	• • •	337	300	A CT			3.00	 mmc		• • •	• • •	T	GCT	ACT	• • •	• • •	• • •	T	ACC	AAG	AGT	AGT	AAT	• • •	• • •	• • •	TGGAAAGAG
HIVSF33	GGTT GTATTTG GGG	• • •	WYI	ACC	ACI	AGI	AG I	AGI	116	AUG	• • •	• • •	T	CCT	ACT	 33T	300	• • •	T	ACC	ACA	AGT	AGT	AGT	• • •	• • •	• • •	TGGGAAACG
HIVCDC4	GTTGAAT ACT	AAT	• • •	•••	• • •	•••	• • •	• • •	• • • •	•••	• • •	• • • •	T	ACT	ACT	AAT	ACC	• • •		AAT	ACT	CAA	CTA	TCA		3.073		GGGAGAACG
HIVSF2	GTTGGGG												6	CCT	ልሮተ	λλΤ	ACC		T	ACT	ACT	እአጥ						TOCAL ACAR
HIVSF2B13																												
HIVHAN	GIMIMAI												7	GCC:	ACT					ACT	ACT							TOCOCONNOC
HIVRF																												
HIVWMJ2 HIVTB132	GAAG	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	C	ATT	ACT	GAT	TGG	AAG	T	ACA	ACA	ATA	• • •					
HIVGUN	ATTGAGG	• • • •	• • •	•••	• • •	• • • •	• • • •	• • •	• • •	• • •	• • •	• • •	T	GAT	ACT			• • •	• • •	• • •	:::	• • •	• • •	• • •	• • •	• • •	• • •	• • • • • • • • • • • • • • • • • • • •
HIVACH9	ATTGGTG GAATTTG	•••	•••	•••	•••	• • •	• • •	• • •	• • •	• • •	• • • •	•••	T	ACT	ACT	TAT	AGT	ccc	T	AGT	AGT	CAT	• • •	• • •	• • •	• • •	• • •	TGGGAAAAG
HIVACP1	GGATTTG AGG										• • •	• • •	T	GAT	ACI	W	110	GGG		AAI	AC1	GAI	• • •	• • •	• • •	• • •	• • •	TCCCAACAC
HIVJM	GGCCTCG AGG		AAT	GTT	ACT	aag	AGG	AAT	GAT	ACT	AAT	GAG	T	GCT	ACT	GAT	AGT	AAT	T	GGG	ACT	GCT	ACT	AGT	GGG	AAT		AAAACTATC
HIVMA1CON	GGCCACT ATT	AGA	AAT	GCC	ACT	ATT	AGA			ACT	AAT	ACC	T	ACT	ACT				T	ACT	ACT	ACT	ACC					CCCCAACTC
HIVMA208	GGCCACT ATT	AGA								ACT	AAT	ACC	T	ACT	ACT	AAT	ACT		T	ACT	ACC							CCCCAACTC
HIVRJS HIVWM	ATGGAAG GTGTGTA ACT	TCA	• • • •	• • • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	• • •	T	GAT	ACT	GAT	ACT	•••	T	ACC	ACT	AAT	AGT	AGT	AGT	• • •	• • •	GGGGAAAAA
111 4441	GIGIGIA ACI	ICA	•••	•••	• • • •	• • • •	•••	• • • •	• • • •	• • • •	• • •	• • • •	T	TGC	ACT	AAT	TTG	AGG	T	GTT	ACC	AAT	ACC	AAA	AGT	AGT	AGC	AAGGAAATA
В	1	2	3	4			7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22						
CONSENSUS	cCAatAgat ?a?	???	???	???	???	???	222	???	???	???	???	???	???	???	222	a At	act	age	222	222	222	277	TaTA	lggTT	' g			
CONSENSUS HIVJRCSF	cCAatAgat ?a? CAT-GAT AAT	??? AAG	???	???	???	???	???	???	???	???	???	???	???	???	???	aAt A-T	act AAT	agc ACC	???	???	???	a?? AAA	-A	GG	À			
CONSENSUS HIVJRCSF HIVJRFL	CCAatAgat ?a? CAT-GAT AAT CAT-GAT AAT	??? AAG	???	???	???	???	???	???	???	???	???		???	???	???	A-T	AAT	ACC	???	???	???	a?? AAA	-A	GG	À			
CONSENSUS HIVJRCSF HIVJRFL HIVALA1	CCAatAgat ?a? CAT-GAT AAT CAT-GAT AAT CAT-GAT AAT	??? AAG	???	???	???	???	???	???	???	???	???	???	???	???	???	aAt $A-T$ $A-T$ $A-T$	AAT AAT AAT	ACC ACC ACC	???	???	???	a?? AAA AGC AGC	-A -A	GG GG TA	·A ·G ·G			
CONSENSUS HIVJRCSF HIVJRFL	CCAatAgat ?a? CAT-GAT AAT CAT-GAT AAT CAT-GAT AAT CAT-GAT AAT CAT-AAT	??? AAG GAT	??? AAA	??? ACT	???	??? 	???	???	???	???	???	???	???	???	???	A-T A-T A-T A-T	AAT AAT AAT ACT	ACC ACC ACC	???	???	???	A?? AAA AGC AGC AGA	-A -A -A	GG GG TA GG	A G G			
CONSENSUS HIVJRCSF HIVJRFL HIVALA1 HIVBRVA HIVJH3 HIVSC	CCAatAgat ?a? CAT-GAT AAT CAT-GAT AAT CAT-GAT AAT CAT-AAT CAT-GAT	??? AAG GAT	??? AAA	??? ACT	??? AAT	??? AGT	??? ACC	??? AAG	???	??? AAT	???	??? AAA	??? AAT	??? GAT	???	A-T A-T A-T A-T A-T A-T	AAT AAT AAT ACT AGT	ACC ACC ACC ACC	???	???	???	a?? AAA AGC AGC AGA AGA	-A -A -A -A	GG GG TA GG	G G A			
CONSENSUS HIVJRCSF HIVJRFL HIVALA1 HIVBRVA HIVJH3 HIVSC HIVBAL1	CCAATAGAT ?A? CAT-GAT AAT CAT-GAT AAT CAT-GAT AAT CAT-GAT AAT CAT-GAT CAT-GAT CAT-GAT	??? AAG GAT 	??? AAA	??? ACT	???	???	??? ACC	???	???	???	???	???	??? AAT	???	???	A-T A-T A-T A-T A-T A-T A-T	AAT AAT ACT AGT ACG	ACC ACC ACC ACC ACC ACC ACC	???	???	???	a?? AAA AGC AGC AGA AGA	-A -A -A -A	GG GG GG GA	A G G A			
CONSENSUS HIVJRCSF HIVJRFL HIVALA1 HIVBRVA HIVJH3 HIVSC HIVBAL1 HIVYU2	CCAATAGAT ?A? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT	??? AAG GAT	???	??? ACT	???	???	??? ACC	???	???	???	??? ATA	??? AAA	??? AAT	???	???	AAT A-T A-T A-T A-T A-T A-T A-T	AAT AAT ACT AGT ACG AAT	ACC ACC ACC ACC ACC ACC AGC	???	???	??? AAT	a?? AAA AGC AGC AGA AGA	-A -A -A -A	GG GG GG GA GG	A G G A G			
CONSENSUS HIVJRCSF HIVJRFL HIVALA1 HIVBRVA HIVJH3 HIVSC HIVBAL1 HIVPAL1 HIVYU2 HIVMN	CCAatAgat ?a? CAT-GAT AAT CAT-GAT AAT CAT-GAT AAT CAT-GAT AAT CAT-GAT CAT-GAT CAT-GAT CAT-GAT TAT-GAT CAT-GAT TAT-GAT	??? AAG GAT 	???	???	???	???	???? ACC	??? AAG	???	??? AAT 	??? ATA	???	???	???	???	AAT A-T A-T A-T A-T A-T A-T A-T	AAT AAT ACT ACG ACG AAT GCT GAT	ACC ACC ACC ACC ACC AGC AGT AGC	???	???	???	a?? AAA AGC AGC AGA AGA 	-A -A -A -A -A -A	GG GG GG GG GG	A G G A G G			
CONSENSUS HIVJRCSF HIVJRFL HIVALA1 HIVBRVA HIVJH3 HIVSC HIVBAL1 HIVVB2L1 HIVVY2 HIVMN HIVHXB2R	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT T-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT	???? AAG GAT	???	???	???	???	??? ACC	???	???	??? AAT 	??? ATA 	??? AAA 	??? AAT 	???	??? 	AAT A-T A-T A-T A-T A-T A-T A-T	AAT AAT ACT ACG AAT GCT GAT GAT	ACC ACC ACC ACC ACC AGC AGT AGC AGT ACT	???	???	???	a?? AAA AGC AGC AGA AGA AGA AGC	-A -A -A -A -A -A -A	GG GG GG GA GG GG	A G G G A G G G G G G G G G G G G G G G			
CONSENSUS HIVJRCSF HIVJRFL HIVALA1 HIVBRVA HIVJH3 HIVSC HIVBAL1 HIVPAL1 HIVYU2 HIVMN	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT T-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT	???? AAG GAT	???	???	???	???	??? ACC	???	???	??? AAT 	??? ATA 	??? AAA 	??? AAT 	???	??? 	AAT A-T A-T A-T A-T A-T A-T A-T	AAT AAT ACT ACG AAT GCT GAT GAT	ACC ACC ACC ACC ACC AGC AGT AGC AGT ACT	???	???	???	a?? AAA AGC AGC AGA AGA AGA AGC	-A -A -A -A -A -A -A	GG GG GG GA GG GG	A G G G A G G G G G G G G G G G G G G G			
CONSENSUS HIVJRCSF HIVJRFL HIVALA1 HIVBRVA HIVJH3 HIVSC HIVBAL1 HIVYU2 HIVMN HIVHXB2R HIVLAI	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AC C-AT-GAT	??? AAG GAT	???	???	????	???	2??	???	???	???	??? ATA	???	??? AAT	???	???	AAT A-T A-T A-T A-T A-T A-T A-T	ACT AAT ACT ACG ACG AAT GCT GAT GAT GAT ACC ACC	ACC ACC ACC ACC ACC AGC AGT AGC ACT ACT ACC ACC		???	???	a?? AAA AGC AGC AGA AGA AGC AGC AGC	-A -A -A -A -A -A -A -A -A	GG GG GG GA GG GG CG CG	A G G G A G G G G G G G G G G G G G G G			
CONSENSUS HIVURCSF HIVURFL HIVALAI HIVBRVA HIVYH3 HIVSC HIVBAII HIVYU2 HIVMN HIVHXB2R HIVMN HIVHXB2R HIVLAI HIVNIAI HIVNIAI HIVMFA HIVCAMI	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT T-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT	??? AAG GAT 	???	???	????	???	???? ACC	???	???	??? AAT	??? ATA	???	???	???	???	AAT A-T A-T A-T A-T A-T A-T A-T	ACT AAT ACT ACT ACG AAT GCT GAT GAT GAT ACC ACC	ACC ACC ACC ACC ACC AGC AGC AGT ACT ACC ACC ACC ACC ACC ACC ACC		???	???	a?? AAA AGC AGC AGA AGA AGC AGC AGC	-AAAAAAAA-	GG GG GG GG GG GG GG GG-	A G G G A G G G G G G G G G G G G G G G			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVSC HIVBALI HIVYU2 HIVYN HIVHXB2R HIVMN HIVHXB2R HIVLAI HIVNL43 HIVMFA HIVCAM1 HIVNYSCG	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AG C-AT AG C-AT AG AT AG C-AT AG AT AG	??? AAG GAT GCT GAT	???	2??? ACT	2???	???	???? ACC	???	???	??? AAT	??? ATA	???	???	???	???	AAT A-T A-T A-T A-T A-T A-T A-T	ACT AAT ACT ACG ACG AAT GCT GAT GAT GAT ACC ACC ACC	ACC ACC ACC ACC AGC AGC AGT ACT ACT ACC AGC AGT ACC ACC AGC	???	???	???	a?? AAA AGC AGC AGA AGA AGC AGC AGC	-AAAAAAAA	GG GG GA GG GG GG GG CG C	A G G G A G G G G G G G G G G G G G G G			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVSC HIVBAL1 HIVPUL HIVMN HIVHRBER HIVHAI HIVNILAI HIVNILAI HIVNILAI HIVNILAI HIVNILAI HIVNILAI HIVNISC HIVCAMI HIVNYSCG HIVADA	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT T-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AGC-AT-GAT C-AT-GAT AGC C-AT-GAT AGC C-AT-GAT AGC C-AT-GAT AGT	???? AAG GAT GCT GAT GAT	2??? AAA	2??? ACT	2??? AAT	???	??? ACC	???	???	???	??? ATA	???	???	???	???	AAT A-T A-T A-T A-T A-T A-T A-T	ACT AAT AAT ACT AGT ACG AAT GCT GAT GAT ACC ACC ACC ACA ACA ACA	ACC ACC ACC ACC AGC AGC AGT ACT ACT ACC AGC AGT ACC AGC ACC AGC	???	???	???	a?? AAA AGC AGC AGA AGA AGA AGC AGC AGC	-AAAAAAAA	GG GG GA GG GG GG GG GG	A G G G A G G G G G G G G G G G G G G G			
CONSENSUS HIVURCSF HIVURFL HIVALAI HIVBRVA HIVVBALI HIVYBALI HIVYBL HIVMN HIVHXB2R HIVMN HIVHXB2R HIVMIAI HIVNLAI HIVNLAI HIVNLAI HIVNSCG HIVADA HIVNSCG HIVADA	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG	???? AAG GAT GCT GAT GAT GAT	???? AAA	2??	????	???? AGT	???	???	???	???	???	???	???	???	???	AAT A-T A-T A-T A-T A-T A-T A-T	ACT AAT AAT ACT ACG AAT GCT GAT GAT ACC ACC ACC ACA ACA	AGC AGC AGC AGC AGC AGC AGT AGC AGT AGC AGC AGC AGC AGC	???	????	???	a?? AAA AGC AGC AGA AGA AGA AGC AGC AGC	-AAAAAAAA	GG GG GA GG GG GG GG GG	A G G G A G G G G G G G G G G G G G G G			
CONSENSUS HIVURCSF HIVURST HIVALAI HIVBRVA HIVSC HIVBALI HIVYU2 HIVYN HIVHXB2R HIVAN HIVHXB2R HIVLAI HIVNL43 HIVMFA HIVNSCG HIVADA HIVNYSCG HIVADA HIVYSLHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVSELHIVINTELHIVSELHIVSELHIVSELHIVAHIVSELHIVSELHIVANTINTELHIVTELHIVSELHIVTELHIVTELHIVTELHIVTELHIVTELHIVS	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AG C-AT-GAT AG C-AT-GAT AAT C-AT-GAT GAT C-AT-GAT AAT C-AT-C-AT C-AT-C-AT C-AT C-AT C-A	???? AAG GAT GAT GCT GAT GAT GAT	2??? AAAA	2??	2???	???	??? ACC	2??	???	???	2??	2???	???	????	???	AAT A-T A-T A-T A-T A-T A-T A-T	ACT AAT AAT ACT ACG AAT GCT GAT GAT ACC ACC ACC ACA ACA ACA ACT ACA	agc ACC ACC ACC ACC AGC AGT AGC AGT ACT ACC AGC AGC AGC AGC AGC AGC AGC AGC AGC	???	???	????	a?? AAA AGC AGC AGA AGA AGC AGC AGC AGC	-AAAAAAAA	GG GG GG GG GG GG GG GG-	ଜୁନ୍ଦ୍ର ବ୍ୟବର ବ୍ୟବର ବ୍ୟବର ଜୁନ୍ଦ୍ର ଜୁନ୍			
CONSENSUS HIVURCSF HIVURFL HIVALAI HIVBRVA HIVVBALI HIVYBALI HIVYBL HIVMN HIVHXB2R HIVMN HIVHXB2R HIVMIAI HIVNLAI HIVNLAI HIVNLAI HIVNSCG HIVADA HIVNSCG HIVADA	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT T-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAT	AAG GAT GAT GAT GAT GAT GAT GAT GAT GAT	???? AAA	2??	???? AAT	2??	2??	???	???	2??	2??	???	???	???	???	AAT A-T A-T A-T A-T A-T A-T A-T	act AAT AAT ACT ACG ACG AAT GCT GAT GAT ACC ACC ACA ACA ACA ACA ACT ACT	agc ACC ACC ACC ACC AGC AGC AGT AGC AGC AGC AGC AGC AGC AGC AGC AGC AGC	???	???	???? AAT ACC ACC	a?? AAA AGC AGC AGA AGA AGA AGC AGC	-AAAAAAAAAAAAAAAAA	GG GG GG GG GG GG GG GG-	A G G G A G G G G G G G G G G G G G G G			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVSC HIVBALI HIVYU2 HIVWN HIVHRB2R HIVLAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNSCG HIVADA HIVUSTL HIVSTL HIVSTL HIVSTL HIVSTL HIVSTL HIVSTL HIVSTL HIVSTL HIVSTL	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT AAT C-AT-GAT AAG	PROPERTY OF THE PROPERTY OF TH	????	???? ACT	????	????	???	2??	???	2??	2??	???? AAAA	???	???	???? 	aAt A-T A-T A-T A-T A-T A-T A-T A-T	act AAT AAT AAT ACT AGG ACG ACG ACA ACC ACC ACC ACC ACC ACC	agc ACC ACC ACC AGC AGC AGC AGC AGC AGC AGC	???	???	???? AAT ACC ACC	a?? AAA AGC AGC AGA AGA AGA AGA AGC AGC	-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	GG GG GG GG GG GG GG CG CG CG CG CA CA	A			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVSC HIVBALI HIVYU2 HIVWN HIVHXB2R HIVLAI HIVNLAI HIVNLAI HIVNLAI HIVNYSCG HIVADA HIVJFL HIVSIMIB4 HIVSF162 HIVOYI HIVSF33	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAT	PROPERTY OF THE PROPERTY OF TH	AAA	2??? ACT	2???	???	2??	2??	???	2??	2??	2???	2??	???	277	aAt A-T A-T A-T A-T A-T A-T A-T A-T	act AAT AAT ACT AGG AAG GCT GAT ACC AACA ACCA AC	agc ACC ACC ACC AGC AGC AGC AGC AGC AGC AGC	???	????	???? AAT ACC ACC	a?? AAA AGC AGC AGA AGA AGA AGA AGA AGA AGC AGC	-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	GG GG TA GG GG GG GG GG	A			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVYBAI HIVYBAI HIVYBAI HIVYBA HIVWBA HIVWBBA HIVWBBA HIVBSB162 HIVWSF133 HIVWBCA	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAT C-AT-GAT GAT	PROPERTY OF THE PROPERTY OF TH	AAA	???? ACT	2???	AGT	2??	222	???	2??	2??	????	2??	???	2??	aAt A-T A-T A-T A-T A-T A-T A-T A-T	act AAT AAT ACT AGG ACG GAT GAT ACC ACC ACA ACA ACT ACT	agc ACC ACC ACC ACC AGC AGC AGC AGC ACC AGC AG	???	???	???? AAT ACC ACC ACC	a?? AAA AGC AGC AGA AGA AGA AGA AGC AGC AGC	-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	GG GG GG GG GG GG GG CG CG CA CA	A O O O O A O O O O O O O O O O O O O O			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVJH3 HIVSC HIVWAN HIVHXEZR HIVLAI HIVNL43 HIVNL43 HIVVLAI HIVNL43 HIVMADA HIVUFL HIVSTIG2 HIVSTIG2 HIVSTIG2 HIVSTIG2 HIVSTIG2 HIVSTG3 HIVSTG3 HIVSTG4 HIVSTG4 HIVSTG4 HIVSTG4 HIVSTG64 HIVSTG664 HIVSTG664 HIVSTG666	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAT C-AT-GAT	PROPERTY OF THE PROPERTY OF TH	2???	ACT	???? AAT AAT AAT AAT AAT	AGT	PROPERTY ACC	AAG	???	AAT	2??	????	2??	???	2??	AAT A-T A-T A-T A-T A-T A-T A-T	act AAT AAT AAT ACT ACG ACG AAT ACG AAT ACC ACC ACA ACA ACT ACC ACC ACC ACC	agc ACC ACC ACC AGC AGC AGT AGC AGT AGC AGC AGC AGC AGC AGC AGC AGC AGC AGC	???	???	????	a?? AAA AGC AGC AGA AGA AGA AGA AGC AGC AGC	-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	GG GG GG GG GG GG GG GG-	A O O O A O O O O O O O O O O O O O O O			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVSC HIVBALI HIVYU2 HIVWAN HIVHAE HIVHAE HIVHAE HIVHAE HIVHAE HIVHAE HIVCAMI HIVMYSCG HIVADA HIVUFL HIVSIMIB4 HIVSFI HIVSF162 HIVOFC4 HIVSF2 HIVSF2 HIVSF2 HIVSF2 HIVSF2	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAT C-AT-GAT GAT C-AT-GAT	AAG GAT GAT GAT GAT GAT GAT GAT GAT GAT	2???	???? ACT	???? AAT AAT AAT	ACT GCT GCT	ACC ACC ACT ACC ACT ACC ACT ACT ACT ACT	AAG	???	AAT	2??	2???	2??? AAT	???	2??	aAt A-T A-T A-T A-T A-T A-T A-T A-T	act AAT AAT ACT ACG ACG AAT ACG AAT ACC ACC ACC ACC ACC ACC ACC ACC ACC	agc ACC ACC ACC AGC AGC AGT AGC AGT AGC ACC AGC AGC AGC AGC AGC AGC AGC AGC	???	????	????	a?? AAA AGC AGC AGC AGA AGA AGA AGA AGA AGC AGC AGC AGC AGC AGC AGC AGC AGC	-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	GG GG GG GG GG GG GG GG-	A O O O O A O O O O O O O O O O O O O O			
CONSENSUS HIVURCSF HIVURFL HIVALAI HIVBRVA HIVALAI HIVYBALI HIVSF162 HIVYBF133 HIVYBC4 HIVSF2B11 HIVSF2B11 HIVSF2B13 HIVYBAN	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT T-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAT C-AT-GAT GAT C-AT-GAT GAT C-AT-GAT GAT C-AT-GAT GAT C-AT-GAT GAT C-AT-GAT AAT	AAG GAT GAT GAT GAT GAT GAT GAT GAT GAT	2???	2??? ACT	2???	ACT ACT ACT ACAT ACAT ACAT ACAT ACAT AC	ACC AGT AGT AGT AGT AGT AGT AGT AGT AGT	AAG ACT AGT	???	AAT AAT AAC AACC AAGG	???	???? AAA	2??	???	277	AAT A-T A-T A-T A-T A-T A-T A-T A-T A-T	act AAT AAT AAT ACG ACG AAT ACG ACA ACC ACC ACC ACC ACC ACC ACC ACC	agc ACC ACC ACC AGC AGC AGC AGT AGC ACT ACT ACT ACT ACT ACC ACC ACC AGC AGC AGC AGC AGC AGC AGC	???	????	????	a?? AAA AGC AGC AGA AGA AGA AGA AGA AGC AGC	-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	GG GG GG GG GG GG GG GG-	A			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVYBA HIVYBC HIVWAN HIVHXBCR HIVLAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNSTA HIVADA HIVUFL HIVSTHABA HIVODAI HIVSF162 HIVSCOAI HIVSF2 HIVSF3	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAT C-AT-GAG AAG	???? AAG GAT	2???	2??? ACT	AAT	AGT ACT ACT ACT ACT ACT AAAG AAAAAAAAAA	ACC AGT AGT AGC	AAG ACT ACT ACT CCT	???	AAT ACC ACC ACC ACC ACC	ATA	???? AAA	2?? AAT	???	277	AAT A-T A-T A-T A-T A-T A-T A-T A-T A-T	act AAT AAT ACT AGG AAT ACG AAT GGAT GGAT ACC ACC ACA ACT ACC ACC ACA ACT ACT A	agc ACC ACC ACC AGC AGC AGC AGC AGC ACC AGC AG	???	????	???? AAT AACC ACC ACC GGGT	a?? AAA AGC AGA AGA AGA AGA AGA AGA AGC AGC	-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA	GG GG GG GG GG GG GG GG-	A O O O O O A O O O O O O O O O O O O O			
CONSENSUS HIVURCSF HIVURFL HIVALAI HIVBRVA HIVALAI HIVYBALI HIVSF162 HIVYBF133 HIVYBC4 HIVSF2B11 HIVSF2B11 HIVSF2B13 HIVYBAN	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAG C-AT-GAT AAT C-AT-GAT AT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AT C-AT-GAT AAT C-AT-GAT A	???? AAG GAT GCT GAT GAT GAT GAT GAT GAT GAT GAT GAT GA	2???	????	AAT AAT AAT AAT AAT AAT AAT	AGT ACT GCT GCT AAG	ACC AGT ACT AGC AGT AGC AGT AGC	AAGT ACT ACT	???	AAC ACC ACG	2??	???? AAA	2??? AAT	???	277	AAT A-T A-T A-T A-T A-T A-T A-T	act AAT AAT ACT AGG AAT GGAT GGAT GGAT ACC ACC ACC ACC ACC ACC ACC ACC ACC A	agc ACC ACC ACC AGC AGC AGC AGT ACC AGT AGC AGC AGC AGC AGC AGC AGC AGC AGC AGC	???	????	???? AAAT ACC ACC ACC	a?? AAA AGC AGA AGA AGA AGA AGA AGA AGA AGC AGC	-A	GG GG GG GG GG GG GG GG-	A			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVSC HIVBALI HIVYU2 HIVWN HIVHXB2R HIVLAI HIVNLA3 HIVGAMI HIVNYSCG HIVADA HIVJFL HIVSIMI84 HIVGYI HIVSF162 HIVOYI HIVSF283 HIVCDC4 HIVSF2813 HIVHAN HIVRF	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAG C-AT-AAT AAT C-AT-GAT	???? AAG GAT GCT GAT GAT GAT GAT GAT GAT GAT GAT GAT GA	7???	????	AAT	AGT AGT AGT ACT ACT ACT ACT ACT	ACC AGC AGT AGC AGT AGC AGT AGC AGC AGT AGC	AAG ACT ACT ACT CCT	???	AAT ACC ACC ACC	2??	2???	2??? AAT	???	272	AAT A-T A-T A-T A-T A-T A-T A-T	act AAT AAT ACT AGG AGG GAT ACC ACC ACC ACC ACC ACC ACC ACC ACC A	age ACC ACC ACC AGC AGC AGT AGT ACC AGT ACC AGC AGC AGC AGC AGC AGC AGC AGC AGC	????	????	????	a?? AAA AGC AGA AGA AGA AGA AGA AGA AGA AGC AGC	-A	GG GG GG GG GG GG CG CG	\$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVYBAI HIVYSCG HIVADA HIVJFL HIVSIMIBA HIVJFI HIVSIMIBA HIVSF162 HIVOYI HIVSF2 HIVGYI HIVSF2 HIVFSAI HIVHAN HIVRF HIVWMJ HIVAN HIVRF HIVWMJ LIVTBIJ LIVTB	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAG C-AT-GAT AAT C-AT-GAT	???? AAG GAT GCT GAT GAT GAT GAT GAT GAT GAT GAT GAT GA	7???	???? ACT	AAT	AGT AGT AGT AGT ACT GCT GCT AAGT AAGT	ACC AGT ACT ACT ACT ACT ACT ACT ACT ACT ACT AC	AAGC ACT AGT AGT AGT AGT AGT AGT AGT AGT AGT AG	???	AAC ACC ACC	ATA	2???	2??? AAT	????	272	8At A-T A-T A-T A-T A-T A-T A-T A-T	act AAT AAT ACT ACG GAT GGAT ACC ACC ACC ACC ACC ACC ACC ACC ACC A	age ACC ACC ACC ACC ACC ACC ACC ACC ACC AC	????	????	????	a?? AAA AGC AGA AGA AGA AGA AGA AGA AGA AGC AGC	-A	GG GG GG GG GG GG GG GG-	ର ବର ବର ବର ବର ବଳ ବର			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVJH3 HIVSC HIVMN HIVHXEZR HIVLAI HIVNL43 HIVVLAI HIVNL43 HIVVLAI HIVNL43 HIVGAMI HIVNL5GG HIVADA HIVJFL HIVSF162 HIVSF162 HIVSF28 HIVSF28 HIVSF28 HIVSF28 HIVSF28 HIVFF HIVMM12 HIVMM12 HIVMM12 HIVGUN HIVACH9 HIVACH9 HIVACH9	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-	???? AAG GAT GCT GAT GAT GAT GAT GAT GAT GAT GAT GAT GA	7???	2???	AAT AAT AAT AAT AAT AAT AAT AAT AAT	AGT ACT ACT ACT ACT ACT ACT ACT	ACC	AAGT CCT	???	AAT AAT ACC ACC ACC ACC ACC ACC ACC AC	2??	2???	2??? AAT	???? GAT	222	## A	act AAT AAT ACT AGT GAT GAT ACC ACA ACA ACA ACT ACC ACC ACC ACC A	age ACC ACC ACC AGC AGC AGC AGC AGC AGC AGC	????	???	???? AAT ACC ACC GGT	a?? AAA AGC AGA AGA AGA AGA AGA AGC AGC AGC	-A	GG GG GG GG GG GG GG GG-	ଜଗର ର ର ର ର ର ର ର ର କଳ ର ର ର ର ର ର ର ର ର ର			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVSC HIVBALI HIVYU2 HIVWA HIVHAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNSCG HIVADA HIVUSTCA HIVSSIMBA HIVASSIMBA HI	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAT	???? AAG GAT GCT GAT GAT GAT GAT GAT GAT GAT GAT AAT AA	7???	7??? ACT	AAT	AGT AGT AGT ACT ACT ACT ACT ACT	ACC AGT AGC AGT AGG AGT AGC AGT AGG AGT AGC AGT AGG AGG	AAG ACT AGT CCT AAAT AAAT AAAT AAAT AAAT	GAT GAT ACT AAT AAAAAAAAAAAAAAAAAAAAAAAAAAA	AAT AAC ACC ACC ACC ACC ACC ACC	ATA GAT AAT	2??? AAA	2222 AAT	???? GAT	222	## A	act AAT AAT ACT AGG AAT GGAT ACC ACC ACC ACC ACC ACC ACC ACC ACC A	age ACC ACC ACC AGC AGC AGC AGC AGC AGC AGC	????	????	????	a?? AAA AAGC AGC AGA AGA AGA AGC AGC AGC AG	-A	GG GG GG GG GG GG CG CG	ଜୟରରରରରରରରରେ ବର୍ଷ ବର୍ଷ ବର୍ଷ ବର୍ଷ ବର୍ଷ ବର୍ଷ ବର୍ଷ ବର୍ଷ			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVYBAI HIVYBL HIVYBL HIVYBL HIVYBL HIVYBL HIVYBL HIVYBL HIVYBC HIVABAI HIVYBC HIVABAI HIVYBC HIVABAI HIVYBC HIVYBC HIVYBC HIVYBC HIVYBC HIVSF18 HIVSF18 HIVSF2 HIVGE HIVSF2 HIVGE HIVSF2 HIVFSF2 HIVGEN HIVACPI HIVACPI HIVACPI	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT T-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAT C-AT-GAT GAT C-AT-GAT GAT C-AT-GAT C-AT-GAT GAT	???? AAG GAT GCT GAT GAT GAT GAT GAT GAT GAT GAT GAT GA	7??? AAA	7???	AAT	AGT ACT ACT ACT ACT ACT ACT ACT ACT ACT AC	ACC	AAG ACT	???	AAT ACC ACC ACG ACC AAGG AAT AAAT AAAT AAAT AAAT AAAT AAAT AAAT	2??	2???	2??? AAT ACT	????	????	## A T A T A T A T A T A T A T A T A T A	act AAT AAT ACT AGG AAT GCT ACC ACA ACA ACC ACA ACC ACA ACT ACC ACC	age ACC ACC ACC ACC ACC ACC ACC ACC ACC AC	????	????	????	a?? AAA AGC AGC AGA AGA AGA AGC AGC AGC AGC	-A	GG GG GG GG GG GG GG GG-	ଜଗ୍ୟ ଦର୍ଶର ବ୍ରବ୍ୟ			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVJH3 HIVSC HIVWAN HIVHSER HIVLAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNLAI HIVNSC HIVADA HIVUFL HIVSTIGE HIVSTIGE HIVSTIGE HIVSTE HIVSTE HIVSF2 HIVSF2 HIVSF2 HIVSF2 HIVSF2 HIVSF2 HIVSF2 HIVMM2 HIVMM2 HIVMM2 HIVMM2 HIVMM1 HIVMACP1 HIVJM HIVMALCON HIVMALCON	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAT C-AT-GAT	???? AAG GAT GCT GAT GAT GAT GAT GAT AAT GGT GAT AAT GGT GAT AAT GGT GAT AAT	7???	????	AAT AAT AAT AAT AAT AAT AAT AAT	AGT ACT ACT ACT ACT ACT ACT ACT ACT ACT AC	ACC AGT	AAG AGT ACT AGT AGT AGA AGA AGA AGA AGA AGA AGA AG	???? GAT ACT AAT AAAG AAT	AAT ACC ACC ACG ACA AAT AAT AAT	ATA GAT AAT AAT AAT	2???	AAT AAT AAT AAT ACT	???? GAT AGC AGT	????	## A	act AAT AAT ACT AGG ACC ACC ACC ACC ACC ACC ACC ACC ACC	age ACC ACC ACC AGC AGC AGC AGC AGC AGC AGC	???	????	????	a?? AAAA AGC AGA AGA AGA AGA AGA AGA AGC AGC	-A	GG	ର ଦର ୪ ୪ ବର			
CONSENSUS HIVURCSF HIVURCSF HIVURFL HIVALAI HIVBRVA HIVYBAI HIVYBL HIVYBL HIVYBL HIVYBL HIVYBL HIVYBL HIVYBL HIVYBC HIVABAI HIVYBC HIVABAI HIVYBC HIVABAI HIVYBC HIVYBC HIVYBC HIVYBC HIVYBC HIVSF18 HIVSF18 HIVSF2 HIVGE HIVSF2 HIVGE HIVSF2 HIVFSF2 HIVGEN HIVACPI HIVACPI HIVACPI	CCAatAgat ?a? C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT AAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT T-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT C-AT-GAT AAT C-AT-GAT GAT C-AT-GAT GAT C-AT-GAT C-AT-GAT GAT	???? AAG GAT GGT GAT GAT GAT GAT GAT AAT AAT AAT GAT G	7??? AAA AAA AAA AAA	????	AAT	AGT ACT GCT GCT AAGT AAT AGT AAT AGT AAT GAT	ACC AGT	222 AAG ACT AGC AACT ACT ACT ACT ACT ACT ACT ACT ACT A	???	AAT AAT AAT AAT	ATA GAT AAT AAT AAT	AAAA AAT AAAT	2??? AAT	????	????	## A	act AAT AAT ACT AGG AAT GGT GAT GAT ACC ACC ACC ACC ACC ACC ACC ACC ACC A	age ACC ACC AGC AGC AGC AGC AGC AGC AGC AGC	????	????	????	a?? AAA AGC AGC AGC AGC AGC AGC AGC AGC AGC	-A	- GG	ବ୍ୟରର ୧୯ର ବ୍ୟର ବ୍ୟର ବ୍ୟର ବ୍ୟର ବ୍ୟର ବ୍ୟର ବ୍ୟର ବ୍ୟ			

FIG. 1. Sequence alignments of variable-length sequences in V1 (A) and V2 (B). Sequences and alignment were adopted from the work of Myers et al. (12) and adjusted by hand. Numbers in the top row indicate codons of the variable-length sequence region. Sequences outside of the variable-length sequences discussed here are in italics. Dashes indicate nucleotides identical to the consensus sequence (top row), and dots indicate deletions relative to the consensus sequence. Uppercase letters in the consensus sequence indicate bases conserved throughout all aligned sequences; lowercase letters indicate bases found in at least 50% (but less than 100%) of the aligned sequences in reference 12. Blocks of three codons encoding a potential N-linked glycosylation site (asparagine-X-serine/threonine) are underlined.

(as, in this case, mostly A) is reflected in the encoded amino acid sequence. The preferential use of codons starting with an A (>80% in the V1 and V2 variable-length sequences) results in predominant insertion of isoleucine, methionine, aspara-

gine, serine, threonine, lysine, and arginine. The virtual absence of T's at position 2 results in the exclusion of isoleucine (Ile) and methionine (Met), the only two hydrophobic amino acids whose codon starts with an A. Apparently, the insertion

7568 NOTES J. Virol.

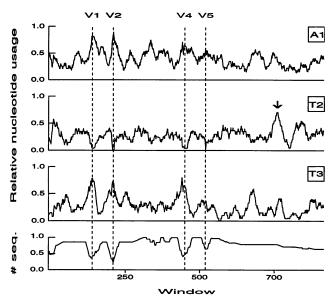


FIG. 2. Frequency distributions of A1 (top panel), T2 (second panel from top), and T3 (third panel from top) in the HIV-1 envelope gene, as well as the relative number of sequences counted in each window (bottom panel). Valleys in the bottom panel plot indicate regions of variable sequence length. Frequencies of each nucleotide at each position were counted in a window of 20 codons, shifted by 1 codon, with the aligned sequences taken from reference 12. Only A1, T2, and T3 are shown. Variable regions V1, V2, V4, and V5 are indicated (except V3). The 95% confidence interval upper and lower limits, calculated as mean $\pm 1.96 \times$ standard deviation following square root transformation of the data to achieve normality, are as follows: A1, 0.71 and 0.15; T2, 0.56 and 0.08; and T3, 0.63 and 0.05, respectively. The A1 and T3 peaks in V1, V2, and V4 exceed these limits, as do the T2 valleys in V1, V2, V4, and V5. A number of A1 peaks are found besides the V1, V2, V4, and V5 regions. The lack of concomitant T2 valleys and T3 peaks indicates that these have been generated through mechanisms different from the one postulated in this paper. A T2 peak is found at positions corresponding to windows 754 to 769 (arrow). All codons with a T at position 2 code for hydrophobic amino acids, and this peak corresponds to the location of the hydrophobic membrane anchor sequences in gp41. Both the alignments and the computer program used to calculate this plot are available from the authors upon request.

of hydrophobic amino acids in this region of env is disadvantageous to the virus. The strong bias towards T, a remnant of the postulated originally inserted AAT, and to a lesser extent towards C as a result of mutation and selection, at position 3 now selectively inserts threonine, asparagine, and serine at these positions. Both threonine and serine codons can be generated from the AAT and AAC asparagine codons through one mutation. Four of the six codons that can potentially code for serine start with a T rather than with an A, namely, T-C-x. These codons are not used in the insertion sequences described here, although they normally compose approximately 60% of all serine codons in mammalian genes (18) and are used throughout the Env open reading frame (not shown). Random insertions and mutations followed by selection would have resulted in a predominant use of T-C-x codons for serine. The exclusive use of AGT and AGC codons to code for serine is therefore a strong point in favor of the postulated mechanism. We have counted the frequency of asparagine, threonine, and serine in V1 and V2 and compared them with those in the gp120 amino acid sequence of HIV_{HXB2R}. In the V1 and V2 variable-length sequences, the combination of Asn, Thr, and Ser makes up 75% of all amino acids, compared with only 22% of all amino acids in gp120.

Insertion of AAT triplets followed by mutation and selection results in the insertion of predominantly asparagine (N), threonine (T), and serine (S) as described above. Random combination of these three amino acids will result in the frequent occurrence of N-x-T and N-x-S sequence blocks, where "x" stands for any amino acid. Both N-x-T and N-x-S are potential N-linked glycosylation sites, and indeed the variablelength sequences are characterized by the occurrence of multiple potential glycosylation sites. Approximately 55% of all amino acids in the variable-length sequences are actually involved in the formation of N-linked glycosylation sites (Fig. 1), demonstrating the efficiency of the postulated mechanism for the generation of such sites. If glycosylation of these regions confers selective advantage to the virus, then it might be expected that virus genotypes that contain such a novel glycosylation site would persist and appear in future virus generations. Examples of such glycosylation sites that have become stable features of the HIV-1 genome can be found in both the V1 and V2 variable-length sequences (V1 codons 13 to 15 and 19 to 21 and V2 codons 16 to 18 [Fig. 1]) and the V4 and V5 regions (not shown). The occurrence of sequences with variable length 5' and 3' of these conserved sites indicates that the mechanism of AAT triplet insertion is ongoing; we postulate that the accumulation of more glycosylation sites provides further selective advantages to the virus through better masking of the epitopes in this region from the immune system. We predict, therefore, that the number of glycosylation sites in these regions of the Env glycoproteins of viruses isolated from infected individuals is directly related to the immune pressure exerted on these regions, which may vary from person to person, but in all probability will increase over time, especially during the asymptomatic period of infection. Others have shown a statistically significant sequence length increase and addition of a novel N-linked glycosylation site at the V2 insertion sequence that correlates with increased viral virulence (8). More virulent virus variants generally arise over time in infected individuals and, as discussed above, so will the length of the insertion sequences because of selective advantages of the viruses that have acquired more glycosylation sites. The observed correlation between viral virulence and sequence length variation in V2, therefore, probably reflects the parallelism in time of these two processes. Accumulation of sequences rich in potential glycosylation sites (both N-linked and O-linked) has been reported for the V1 regions of simian immunodeficiency virus (SIV_{MNE}) isolates obtained from monkeys that progressed towards AIDS (13). This region in SIV shows evidence of ACA triplet reiteration (based on the SIVenv alignment in the work of Myers et al. [12] and on the work of Overbaugh and Rudensey [13], not shown), which results in the accumulation of threonine residues, thereby creating potential targets for O-linked glycosylation (13). It appears that the highly related lentiviruses SIV and HIV have adopted very similar but not identical strategies (AAT versus ACA reiteration) to achieve the same goal: glycosylation of sites on the envelope glycoproteins that are important for the virus life cycle, thereby shielding these sites from recognition by antibodies.

The sequences in the V1 and V2 region play a direct role in virus entry, as discussed above. No such functional properties have been described for either the V4 or the V5 region, and yet it appears from our analysis that they may be subject to the same selective pressures as V1 and V2. The high degree of variation in these regions hampers functional studies, e.g.,

Vol. 68, 1994 NOTES 7569

because of the difficulties of raising cross-reactive antibodies. It is conceivable that functional roles for V4 and V5 will be described in the future. Alternatively, glycosylation at V4 and V5 could help mask other functional regions that are brought into the proximity of V4 and/or V5 in the folded molecular complex.

The mechanism by which the described triplet reiterations arise is unclear. Recently, a number of human genetic diseases characterized by trinucleotide repeats have been described (reviewed in reference 15), and strand slippage during DNA replication (5) has been proposed as a mechanism that could introduce such repeats (16). Tandemly repeated sequences are also observed in human hypervariable minisatellite sequences (10), and an alternative mechanism involving double-stranded breaks and gap repair has been proposed (9). Although it is tempting to speculate that the viral reverse transcriptase is primarily responsible for the observed phenomenon, in vitro studies of the HIV-1 reverse transcriptase have not revealed direct evidence for triplet insertion, although strand slippage during replication of DNA or RNA is observed (4). The examples quoted above are compatible with the idea that cellular polymerases may be involved.

We are greatly indebted to Kees Siebelink for helpful suggestions and critical reading of the manuscript and to Albert Osterhaus for continued support.

REFERENCES

- Alexander, S., and J. H. Elder. 1984. Carbohydrate dramatically influences immune reactivity of antisera to viral glycoprotein antigens. Science 226:1328-1330.
- Andeweg, A. C., P. Leeflang, A. D. M. E. Osterhaus, and M. L. Bosch. 1993. Both the V2 and V3 regions of the human immuno-deficiency virus type 1 surface glycoprotein functionally interact with other envelope regions in syncytium formation. J. Virol. 67: 3232-3239.
- Boyd, M. T., G. R. Simpson, A. J. Cann, M. A. Johnson, and R. A. Weiss. 1993. A single amino acid substitution in the V1 loop of human immunodeficiency virus type 1 gp120 alters cellular tropism. J. Virol. 67:3649–3652.
- Boyer, J. C., K. Bebenek, and T. A. Kunkel. 1992. Unequal human immunodeficiency virus type 1 reverse transcriptase error rates with RNA and DNA template. Proc. Natl. Acad. Sci. USA 89: 6919-6923.
- Chamberlin, M., and P. Berg. 1962. Deoxyribonucleic acid-directed synthesis of ribonucleic acid by an enzyme from Escherichia Coli. Proc. Natl. Acad. Sci. USA 48:81–88.
- Davis, D., D. M. Stephens, C. Willers, and P. Lachmann. 1990. Glycosylation governs the binding of antipeptide antibodies to regions of hypervariable amino acid sequence within recombinant gp120 of human immunodeficiency virus type 1. J. Gen. Virol. 71:2889–2898.
- 7. Fung, M., C. Sun, W. Gordon, R. S. Liou, T. W. Chang, W. Sun, E.

- **Daar, and D. Ho.** 1992. Identification and characterization of a neutralization site within the second variable region of human immunodeficiency virus type 1 gp120. J. Virol. **66:**848–856.
- Groenink, M., R. A. M. Fouchier, S. Broersen, C. H. Baker, M. Koot, A. B. van 't Woudt, H. G. Huisman, F. Miedema, M. Tersmette, and H. Schuitemaker. 1993. Relation of phenotype evolution of HIV-1 to envelope V2 configuration. Science 260: 1513-1516.
- Jeffreys, A. J., K. Tamaki, A. MacLeod, D. G. Monckton, D. L. Neil, and J. A. L. Armour. 1994. Complex gene conversion events in germline mutation at human minisatellites. Nat. Genet. 6:136– 145.
- Jeffreys, A. J., V. Wilson, and S. L. Thein. 1985. Hypervariable 'minisatellite' regions in human DNA. Nature (London) 314:67– 74
- 10a.McKeating, J. A., C. Shotten, J. Cordell, S. Graham, P. Balfe, N. Sullivan, M. Charles, M. Page, A. Bolmstedt, S. Olofsson, S. C. Kayman, Z. Wu, A Pinter, C. Dean, J. Sodroski, and R. A. Weiss. 1993. Characterization of neutralizing monoclonal antibodies to linear and conformation-dependent epitopes within the first and second variable domains of human immunodeficiency virus type 1 gp120. J. Virol. 67:4932–4944.
- Modrow, S., B. H. Hahn, G. S. Shaw, R. C. Gallo, F. Wong-Staal, and H. Wolf. 1987. Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. J. Virol. 61:570–578.
- Myers, G., B. Korber, S. Wain-Hobson, R. F. Smith, and G. N. Pavlakis (ed.). 1993. Human retroviruses and AIDS. Los Alamos Laboratory, Los Alamos, N. Mex.
- Overbaugh, J., and L. M. Rudensey. 1992. Alterations in potential sites for glycosylation predominate during evolution of the simian immunodeficiency virus envelope gene in macaques. J. Virol. 66:5937-5948.
- Pinter, A., S. Kayman, Z. Wu, W. Honnen, H. Chen, J. McKeating, C. Shotten, S. Warrier, and S. Tilley. 1993. Epitope mapping and functional characterization of neutralizing monoclonal antibodies directed against the V2 domain of HIV-1 gp120, abstr. Q443. J. Cell. Biochem. 1993(Suppl. 17E):69.
- Richards, R. I., and G. R. Sutherland. 1992. Dynamic mutations: a new class of mutations causing human disease. Cell 70:709-712.
- Richards, R. I., and G. R. Sutherland. 1994. Simple repeat DNA is not replicated simply. Nat. Genet. 6:114-116.
- Sullivan, N., M. Thali, C. Furman, D. Ho, and J. Sodroski. 1993.
 Effect of amino acid changes in the V1/V2 region of the human immunodeficiency virus type 1 gp120 glycoprotein on subunit association, syncytium formation, and recognition by a neutralizing antibody. J. Virol. 67:3674-3679.
- Wada, K., S. Aota, R. Tsuchiya, F. Ishibashi, T. Gojobori, and T. Ikemura. 1990. Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res. 18:2367–2411.
- Westervelt, P., D. B. Trowbridge, L. G. Epstein, B. M. Blumberg, Y. Li, B. H. Hahn, G. M. Shaw, R. W. Price, and L. Ratner. 1992. Macrophage tropism determinants of human immunodeficiency virus type 1 in vivo. J. Virol. 66:2577-2582.