Numbering Positions in SIV Relative to SIVMM239(revised*)

Charles Calef[1], John Mokili[1], David H. O'Connor[2], David I. Watkins[2], Bette Korber[1]
[1]Theoretical Biology and Biophysics, T10, MS K710, Los Alamos National Laboratory, Los Alamos NM 87545, USA
[2]Wisconsin Regional Primate Research Center, 1220 Capitol Court, Madison, WI USA 53715

* This article has been revised (Oct. 22, 2002). The cleavage sites within the p2, p8, p6 and p1 segments of Gag have been corrected based on Henderson et al., 1988 *J. Virol.* **62**:2587–2595. The SDS-Page mobility values of the Gag proteins have also been modified to agree with those in Ref [2]. We thank Dr. Robert J. Gorelick (Retroviral Mutagenesis Laboratory, AIDS Vaccine Program) for bringing the errors to our attention. The terminal A nucleotide was deleted Aug. 10, 2005.

Introduction

The use of HIVHXB2 as the prototype reference strain for numbering nucleic acid and amino acid sequences has provided a useful strategy for consistent and accurate determination of the locations of nucleic and amino acid sequences of HIV-1 in the literature [1]. Because of the high frequency of insertions and deletions, different HIV sequences have genes and proteins of varying lengths. Specifying the sequence position relative to a unique reference strain, HIVHXB2, allows direct comparisons between studies that use different strains, and easy retrieval of sequences of the gene of protein regions of interest from the databases. Specification of sequence positions is often included in papers where epitopes are defined, where primers are used, or where key functional elements are localized, and in these settings the HXB2 numbering engine is a quick way to determine the precise location of the region of interest.

This exercise is manageable for sequences that are relatively closely related to HIVHXB2, but the more divergent the sequence under study is from HIVHXB2, the harder it is to do the alignment to determine accurately the relative positions vis-a-vis the prototype or reference strain. HXB2 can be used readily for numbering sequences within the M group of HIV-1 viruses, and reasonably efficiently for the more diverse viral sequences from chimpanzee, and the human O and N groups (Figure 1). But the numbering of SIVs isolated from sooty mangabeys illustrates a situation where an alternative approach for numbering the nucleic and amino acid sequences is required. The deduced amino acid sequence of SIVmm239 is similar to that of SIVsmH4 by 91% in Gag, 92% in Pol, 84% in Env, 83% in Vif, 65% in Tat, 73% in Rev and 66% in Nef. Within the same regions, SIVmm239 has a similarity score of 52%, 56%, 31%, 25%, 23% 28% and 29%, respectively, to HXB2 [2]. In addition, most SIVmm, SIV and HIV-2 strains have a vpx ORF instead of vpu, a region of potential problems for numbering SIVs relative to HXB2 (Figure 2). Thus it is more practical to align and number SIVmm and HIV-2 isolates relative to a strain that has the same genomic organization and which is more closely related. Another rationale for adopting a new numbering prototype sequence for SIV is its increasing use in primate vaccine research.

After some deliberation and external consultation, we selected SIVMM239 as the prototype reference sequence for numbering SIV strains at the Los Alamos database. There are reasonable arguments for the use of different strain as the prototype. But the high frequency with which SIVMM239 is used in vaccine studies and the comparatively large number of epitopes that have been defined for SIVMM239 was the determining factor for this choice. However, the original SIVMM239 clone [2] deposited in GenBank (accession number M33262) has 256 nucleotides of flanking non-SIVMM sequence. We have removed the flanking sequence and stored the resulting file as SIVMM239R in our database. The original sequence of SIVMM239 contains a premature stop codon, TAA, at position 9353–9355 within the nef coding sequence. In SIVMM239R we have replaced the TAA stop with the SIVMM consensus codon GAA which codes for glutamate. Finally we have deleted the final “A” at position 10279 because we consider it a PCR artifact, giving the complete genome a length of 10278.

In dealing with deletions and insertions relative to SIVMM239, we have used the same methodology as for the numbering of HIV-1 relative to HIVHXB2 [1].The computer program at Los Alamos that numbers HIV-1 sequences in relation to HXB2, known as the "HXB2 Numbering Engine," has now been extended to number SIV, or closely-related HIV-2 sequences, in relation to SIVMM239R. It can be found at http://hiv-web.lanl.gov/content/hiv-db/LOCATE_SEQ/locate.html

[1] Korber, B. T., Foley, B. F., Kuiken, C. l., Pillai, S. K., and Sodroski, J. G., Numbering Positions in HIV Relative to HXB2CG, in Korber *et al*., eds., *Human Retroviruses and AIDS 1998*, pp. III-102–III-111, Los Alamos National Laboratory, Los Alamos, NM, report LA-UR 99-1704. Available online at http://hiv-web.lanl.gov/NUM-HXB2/NUMBERING.html.

[2] Regier, D. A., and Desrosiers, R. C., The Complete Nucleotide Sequence of a Pathogenic Molecular Clone of Simian Immunodeficiency Virus, *AIDS Research and Human Retroviruses*, **6**(11):1221-1231.

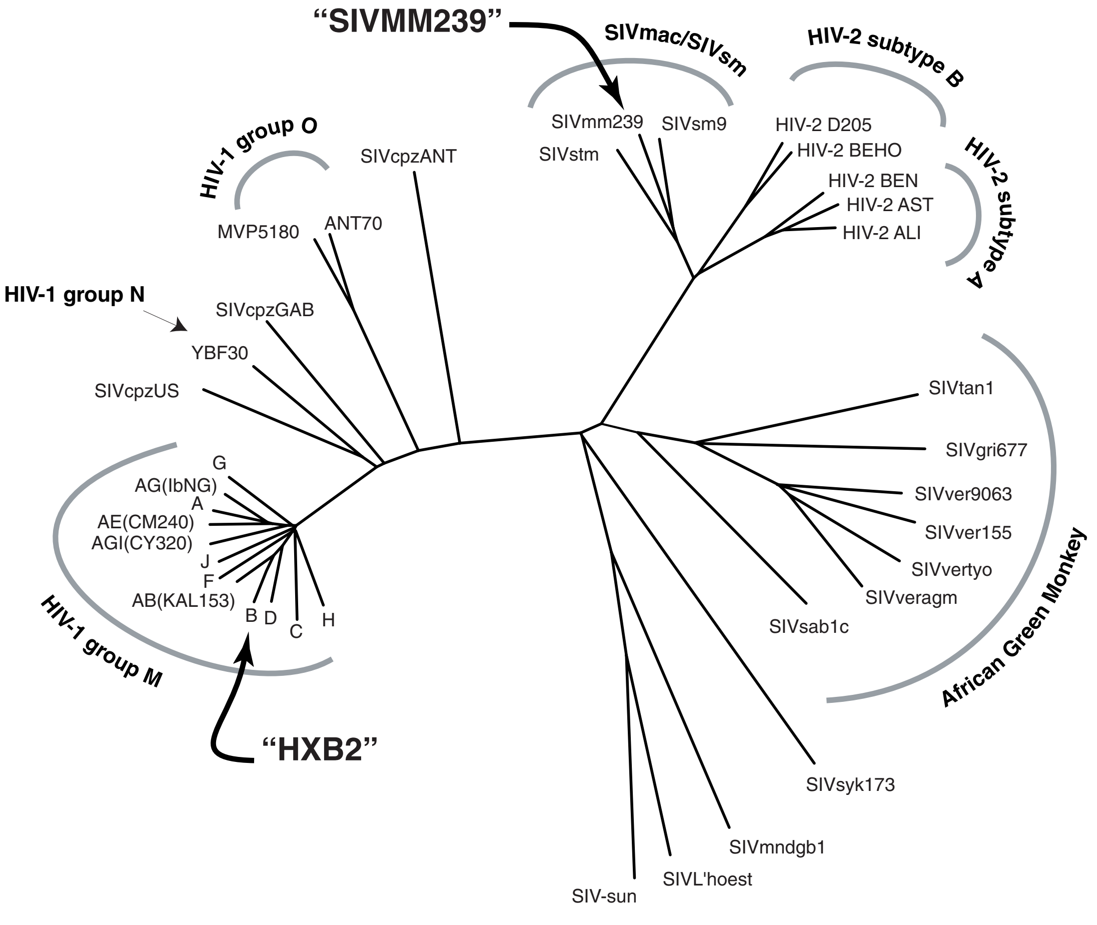

Figure 1. Phylogenetic tree of the primate lentiviruses showing the large distance between the SIVmac group and the HIV-1 M group. Note also the wide divergence of SIVmac from other SIVs.

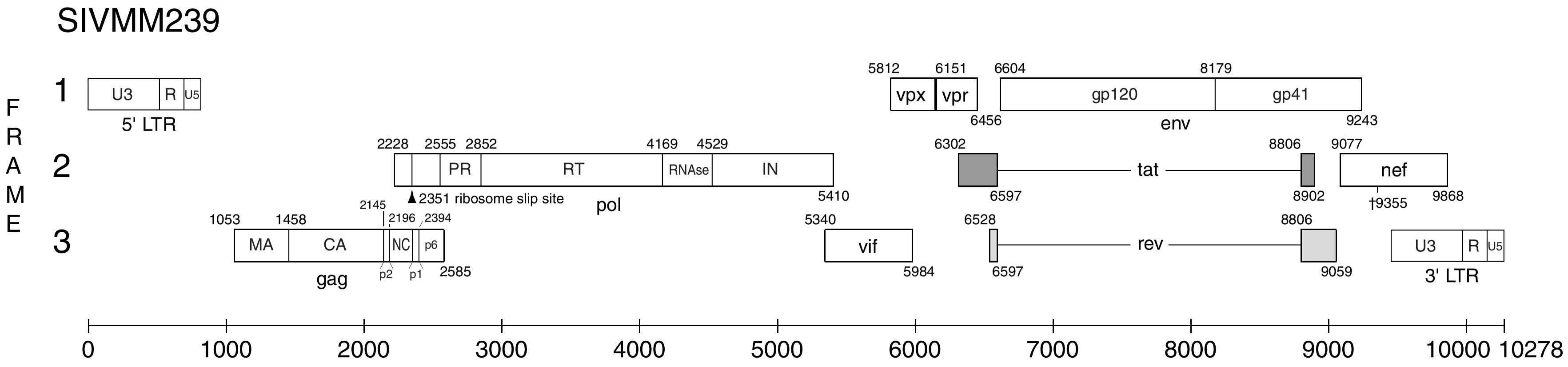

***Figure 2. Landmarks of SIVMAC239 genome**. The gene start, indicated by the small number in the upper left corner of each rectangle normally records the position of the a in the atg start codon for that gene while the number in the lower right records the last position of the stop codon. For pol, the 5' end at position 2228 is the start of the open reading frame. The start of the Pol polyprotein is taken to be the first t in the sequence ttttttag which forms part of the stem loop that potentiates ribosomal slippage on the RNA and a resulting –1 frameshift and the translation of the gag-pol polyprotein. The tat and rev spliced exons are shown as shaded rectangles. †9355 marks a premature stop codon in nef found in the original SIVMM239 strain sequenced and deposited in GenBank. This TAA stop codon has been replaced by a GAA glutamate codon in the reference SIVMM239 sequence annotated on the pages that follow. The putative boundaries of the constituent proteins of the gag, pol, and env polyproteins are tentative having been selected partly by alignment with HIV-1 strain HXB2R. Abbreviations: MA matrix, CA capsid, NC nucleocapsid, PR protease, RT reverse transcriptase, IN integrase.*

SMM239 Amino Acid Sequence Numbering:

Gag precursor [Assemblin] (p57)

```
MGVRNSVLSG KKADELEKIR LRPNGKKKYM LKHVVWAANE LDRFGLAESL LENKEGCQKI LSVLAPLVPT GSENLKSLYN TVCVIWCIHA EEKVKHTEEA 100
KQIVQRHLVV ETGTTETMPK TSRPTAPSSG RGGNYPVQQI GGNYVHLPLS PRTLNAWVKL IEEKKFGAEV VPGFQALSEG CTPYDINQML NCVGDHQAAM 200
QIIRDIINEE AADWDLQHPQ PAPQQGQLRE PSGSDIAGTT SSVDEQIQWM YRQQNPIPVG NIYRRWIQLG LQKCVRMYNP TNILDVKQGP KEPFQSYVDR 300
FYKSLRAEQT DAAVKNWMTQ TLLIQNANPD CKLVLKGLGV NPTLEEMLTA CQGVGGPGQK ARLMAEALKE ALAPVPIPFA AAQQRGPRKP IKCWNCGKEG 400
HSARQCRAPR RQGCWKCGKM DHVMAKCPDR QAGFLGLGPW GKKPRNFPMA QVHQGLMPTA PPEDPAVDLL KNYMQLGKQQ REKQRESREK PYKEVTEDLL 500
HLNSLFGGDQ                                                                                                 510
```

Gag Matrix (p15)

```
MGVRNSVLSG KKADELEKIR LRPNGKKKYM LKHVVWAANE LDRFGLAESL LENKEGCQKI LSVLAPLVPT GSENLKSLYN TVCVIWCIHA EEKVKHTEEA 100
KQIVQRHLVV ETGTTETMPK TSRPTAPSSG RGGNY                                                                     135
```

Gag Capsid (p27)

```
PVQQIGGNYV HLPLSPRTLN AWVKLIEEKK FGAEVVPGFQ ALSEGCTPYD INQMLNCVGD HQAAMQIIRD IINEEAADWD LQHPQPAPQQ GQLREPSGSD 100
IAGTTSSVDE QIQWMYRQQN PIPVGNIYRR WIQLGLQKCV RMYNPTNILD VKQGPKEPFQ SYVDRFYKSL RAEQTDAAVK NWMTQTLLIQ NANPDCKLVL 200
KGLGVNPTLE EMLTACQGVG GPGQKARLM                                                                            229
```

Gag “Spacer” (p2)

```
AEALKEALAP VPIPFAA                                                                                          17
```

Gag Nucleocapsid [NC] (p8)

```
AQQRGPRKPI KCWNCGKEGH SARQCRAPRR QGCWKCGKMD HVMAKCPDRQ AG                                                   52
```

Gag “Spacer” (p1)

```
FLGLGPWGKK PRNF                                                                                             14
```

Gag (p6)

```
PMAQVHQGLM PTAPPEDPAV DLLKNYMQLG KQQREKQRES REKPYKEVTE DLLHLNSLFG GDQ                                        63
```

Pol polyprotein

```
FFRPWSMGKE APQFPHGSSA SGADANCSPR GPSCGSAKEL HAVGQAAERK AERKQREALQ GGDRGFAAPQ FSLWRRPVVT AHIEGQPVEV LLDTGADDSI 100
VTGIELGPHY TPKIVGGIGG FINTKEYKNV EIEVLGKRIK GTIMTGDTPI NIFGRNLLTA LGMSLNFPIA KVEPVKVALK PGKDGPKLKQ WPLSKEKIVA 200
LREICEKMEK DGQLEEAPPT NPYNTPTFAI KKKDKNKWRM LIDFRELNRV TQDFTEVQLG IPHPAGLAKR KRITVLDIGD AYFSIPLDEE FRQYTAFTLP 300
SVNNAEPGKR YIYKVLPQGW KGSPAIFQYT MRHVLEPFRK ANPDVTLVQY MDDILIASDR TDLEHDRVVL QSKELLNSIG FSTPEEKFQK DPPFQWMGYE 400
```

```
LWPTKWKLQK IELPQRETWT VNDIQKLVGV LNWAAQIYPG IKTKHLCRLI RGKMTLTEEV QWTEMAEAEY EENKIILSQE QEGCYYQEGK PLEATVIKSQ 500
DNQWSYKIHQ EDKILKVGKF AKIKNTHTNG VRLLAHVIQK IGKEAIVIWG QVPKFHLPVE KDVWEQWWTD YWQVTWIPEW DFISTPPLVR LVFNLVKDPI 600
EGEETYYTDG SCNKQSKEGK AGYITDRGKD KVKVLEQTTN QQAELEAFLM ALTDSGPKAN IIVDSQYVMG IITGCPTESE SRLVNQIIEE MIKKSEIYVA 700
WVPAHKGIGG NQEIDHLVSQ GIRQVLFLEK IEPAQEEHDK YHSNVKELVF KFGLPRIVAR QIVDTCDKCH QKGEAIHGQA NSDLGTWQMD CTHLEGKIII 800
VAVHVASGFI EAEVIPQETG RQTALFLLKL AGRWPITHLH TDNGANFASQ EVKMVAWWAG IEHTFGVPYN PQSQGVVEAM NHHLKNQIDR IREQANSVET 900
IVLMAVHCMN FKRRGGIGDM TPAERLINMI TTEQEIQFQQ SKNSKFKNFR VYYREGRDQL WKGPGELLWK GEGAVILKVG TDIKVVPRRK AKIIKDYGGG 1000
KEVDSSSHME DTGEAREVA                                                                                        1019
```

Pol Protease (p10)

```
PQFSLWRRPV VTAHIEGQPV EVLLDTGADD SIVTGIELGP HYTPKIVGGI GGFINTKEYK NVEIEVLGKR IKGTIMTGDT PINIFGRNLL TALGMSLNF  99
```

Pol Reverse Transcriptase (RT/RNAse) (p66)

```
PIAKVEPVKV ALKPGKDGPK LKQWPLSKEK IVALREICEK MEKDGQLEEA PPTNPYNTPT FAIKKKDKNK WRMLIDFREL NRVTQDFTEV QLGIPHPAGL 100
AKRKRITVLD IGDAYFSIPL DEEFRQYTAF TLPSVNNAEP GKRYIYKVLP QGWKGSPAIF QYTMRHVLEP FRKANPDVTL VQYMDDILIA SDRTDLEHDR 200
VVLQSKELLN SIGFSTPEEK FQKDPPFQWM GYELWPTKWK LQKIELPQRE TWTVNDIQKL VGVLNWAAQI YPGIKTKHLC RLIRGKMTLT EEVQWTEMAE 300
AEYEENKIIL SQEQEGCYYQ EGKPLEATVI KSQDNQWSYK IHQEDKILKV GKFAKIKNTH TNGVRLLAHV IQKIGKEAIV IWGQVPKFHL PVEKDVWEQW 400
WTDYWQVTWI PEWDFISTPP LVRLVFNLVK DPIEGEETYY TDGSCNKQSK EGKAGYITDR GKDKVKVLEQ TTNQQAELEA FLMALTDSGP KANIIVDSQY 500
VMGIITGCPT ESESRLVNQI IEEMIKKSEI YVAWVPAHKG IGGNQEIDHL VSQGIRQVL                                             559
```

Pol RT (p51)

```
PIAKVEPVKV ALKPGKDGPK LKQWPLSKEK IVALREICEK MEKDGQLEEA PPTNPYNTPT FAIKKKDKNK WRMLIDFREL NRVTQDFTEV QLGIPHPAGL 100
AKRKRITVLD IGDAYFSIPL DEEFRQYTAF TLPSVNNAEP GKRYIYKVLP QGWKGSPAIF QYTMRHVLEP FRKANPDVTL VQYMDDILIA SDRTDLEHDR 200
VVLQSKELLN SIGFSTPEEK FQKDPPFQWM GYELWPTKWK LQKIELPQRE TWTVNDIQKL VGVLNWAAQI YPGIKTKHLC RLIRGKMTLT EEVQWTEMAE 300
AEYEENKIIL SQEQEGCYYQ EGKPLEATVI KSQDNQWSYK IHQEDKILKV GKFAKIKNTH TNGVRLLAHV IQKIGKEAIV IWGQVPKFHL PVEKDVWEQW 400
WTDYWQVTWI PEWDFISTPP LVRLVFNLVK DPIEGEETY                                                                   439
```

Pol RNAse (p15)

```
YTDGSCNKQS KEGKAGYITD RGKDKVKVLE QTTNQQAELE AFLMALTDSG PKANIIVDSQ YVMGIITGCP TESESRLVNQ IIEEMIKKSE IYVAWVPAHK 100
GIGGNQEIDH LVSQGIRQVL                                                                                        120
```

Pol Integrase (p31)

```
FLEKIEPAQE EHDKYHSNVK ELVFKFGLPR IVARQIVDTC DKCHQKGEAI HGQANSDLGT WQMDCTHLEG KIIIVAVHVA SGFIEAEVIP QETGRQTALF 100
LLKLAGRWPI THLHTDNGAN FASQEVKMVA WWAGIEHTFG VPYNPQSQGV VEAMNHHLKN QIDRIREQAN SVETIVLMAV HCMNFKRRGG IGDMTPAERL 200
INMITTEQEI QFQQSKNSKF KNFRVYYREG RDQLWKGPGE LLWKGEGAVI LKVGTDIKVV PRRKAKIIKD YGGGKEVDSS SHMEDTGEAR EVA        293
```

Vif

```
MEEEKRWIAV PTWRIPERLE RWHSLIKYLK YKTKDLQKVC YVPHFKVGWA WWTCSRVIFP LQEGSHLEVQ GYWHLTPEKG WLSTYAVRIT WYSKNFWTDV 100
TPNYADILLH STYFPCFTAG EVRRAIRGEQ LLSCCRFPRA HKYQVPSLQY LALKVVSDVR SQGENPTWKQ WRRDNRRGLR MAKQNSRGDK QRGGKPPTKG 200
ANFPGLAKVL GILA                                                                                           214
```

Vpx

```
MSDPRERIPP GNSGEETIGE AFEWLNRTVE EINREAVNHL PRELIFQVWQ RSWEYWHDEQ GMSPSYVKYR YLCLIQKALF MHCKKGCRCL GEGHGAGGWR 100
PGPPPPPPPG LA                                                                                             112
```

Vpr

```
MEERPPENEG PQREPWDEWV VEVLEELKEE ALKHFDPRLL TALGNHIYNR HGDTLEGAGE LIRILQRALF MHFRGGCIHS RIGQPGGGNP LSAIPPSRSM 100
L                                                                                                         101
```

Tat

```
METPLREQEN SLESSNERSS CISEADASTP ESANLGEEIL SQLYRPLEAC YNTCYCKKCC YHCQFCFLKK GLGICYEQSR KRRRTPKKAK ANTSSASNKP 100
ISNRTRHCQP EKAKKETVEK AVATAPGLGR                                                                          130
```

Rev

```
MSNHEREEEL RKRLRLIHLL HQTNPYPTGP GTANQRRQRK RRWRRRWQQL LALADRIYSF PDPPTDTPLD LAIQQLQNLA IESIPDPPTN TPEALCDPTE 100
DSRSPQD                                                                                                   107
```

Env

```
MGCLGNQLLI AILLLSVYGI YCTLYVTVFY GVPAWRNATI PLFCATKNRD TWGTTQCLPD NGDYSEVALN VTESFDAWNN TVTEQAIEDV WQLFETSIKP 100
CVKLSPLCIT MRCNKSETDR WGLTKSITTT ASTTSTTASA KVDMVNETSS CIAQDNCTGL EQEQMISCKF NMTGLKRDKK KEYNETWYSA DLVCEQGNNT 200
GNESRCYMNH CNTSVIQESC DKHYWDAIRF RYCAPPGYAL LRCNDTNYSG FMPKCSKVVV SSCTRMMETQ TSTWFGFNGT RAENRTYIYW HGRDNRTIIS 300
LNKYYNLTMK CRRPGNKTVL PVTIMSGLVF HSQPINDRPK QAWCWFGGKW KDAIKEVKQT IVKHPRYTGT NNTDKINLTA PGGGDPEVTF MWTNCRGEFL 400
YCKMNWFLNW VEDRNTANQK PKEQHKRNYV PCHIRQIINT WHKVGKNVYL PPREGDLTCN STVTSLIANI DWIDGNQTNI TMSAEVAELY RLELGDYKLV 500
                gp120 end \/ gp41 start
EITPIGLAPT DVKRYTTGGT SRNKRGVFVL GFLGFLATAG SAMGAASLTL TAQSRTLLAG IVQQQQLLD VVKRQQELLR LTVWGTKNLQ TRVTAIEKYL 600
KDQAQLNAWG CAFRQVCHTT VPWPNASLTP KWNNETWQEW ERKVDFLEEN ITALLEEAQI QQEKNMYELQ KLNSWDVFGN WFDLASWIKY IQYGVYIVVG 700
VILLRIVIYI VQMLAKLRQG YRPVFSSPPS YFQQTHIQQD PALPTREGKE RDGGEGGGNS SWPWQIEYIH FLIRQLIRLL TWLFSNCRTL LSRVYQILQP 800
ILQRLSATLQ RIREVLRTEL TYLQYGWSYF HEAVQAVWRS ATETLAGAWG DLWETLRRGG RWILAIPRRI RQGLELTLL                    879
```

Premature stop in original SIVMM239 sequence, changed to consensus glutamate, E.

Nef

```
                                                                                         |
MGGAISMRRS RPSGDLRQRL LRARGETYGR LLGEVEDGYS QSPGGLDKGL SSLSCEGQKY NQGQYMNTPW RNPAEEREKL AYRKQNMDDI DEEDDDLVGV 100
SVRPKVPLRT MSYKLAIDMS HFIKEKGGLE GIYYSARRHR ILDIYLEKEE GIIPDWQDYT SGPGIRYPKT FGWLWKLVPV NVSDEAQEDE EHYLMHPAQT 200
SQWDDPWGEV LAWKFDPTLA YTYEAYVRYP EEFGSKSGLS EEEVRRRLTA RGLLNMADKK ETR                                     263
```

SMM239 Nucleic Acid Sequence Numbering:

```
/  5′ LTR U3 region start
tggaagggat ttattacagt gcaagaagac atagaatctt agacatatac ttagaaaagg aagaaggcat cataccagat tggcaggatt acacctcagg 100
accaggaatt agatacccaa agacatttgg ctggctatgg aaattagtcc ctgtaaatgt atcagatgag gcacaggagg atgaggagca ttatttaatg 200
catccagctc aaacttccca gtgggatgac ccttggggag aggttctagc atggaagttt gatccaactc tggcctacac ttatgaggca tatgttagat 300
acccagaaga gtttggaagc aagtcaggcc tgtcagagga agaggttaga agaaggctaa ccgcaagagg ccttcttaac atggctgaca agaaggaaac 400
tcgctgaaac agcagggact ttccacaagg ggatgttacg ggaggtact ggggaggagc cggtcgggaa cgcccacttt cttgatgtat aaatatcact 500

      5′ LTR U3 region end \ / 5′ LTR R repeat region start
                             / putative mRNA start
gcatttcgct ctgtattcag tcgctctgcg gagaggctgg cagattgagc cctgggaggt tctctccagc actagcaggt agagcctggg tgttccctgc 600
                                                                                                    5′ LTR U5
                                                                  5′ LTR R repeat region end \/ region start
tagactctca ccagcacttg gccggtgctg ggcagagtga ctccacgctt gcttgcttaa agccctcttc aataaagctg ccattttaga agtaagctag 700
tgtgtgttcc catctctcct agccgccgcc tggtcaactc ggtactcaat aataagaaga ccctggtctg ttaggaccct ttctgctttg ggaaaccgaa 800
gcaggaaaat ccctagcaga ttggcgcctg aacagggact tgaaggagag tgagagactc ctgagtacgg ctgagtgaag gcagtaaggg cggcaggaac 900
caaccacgac ggagtgctcc tataaaggcg cgggtcggta ccagacggcg tgaggagcgg gagaggaaga ggcctccggt tgcaggtaag tgcaacacaa 1000

                                                 / Gag p17 start
aaaagaaata gctgtctttt atccaggaag ggtaataag atagagtggg agatgggcgt gagaaactcc gtcttgtcag ggaagaaagc agatgaatta 1100
gaaaaaatta ggctacgacc caacggaaag aaaaagtaca tgttgaagca tgtagtatgg gcagcaaatg aattagatag atttggatta gcagaaagcc 1200
tgttggagaa caaagaagga tgtcaaaaaa tactttcggt cttagctcca ttagtgccaa caggctcaga aaatttaaaa agcctttata atactgtctg 1300
cgtcatctgg tgcattcacg cagaagagaa agtgaaacac actgaggaag caaaacagat agtgcagaga cacctagtgg tggaaacagg aacaacagaa 1400

                                            Gag p17 end \/Gag p24 start
actatgccaa aaacaagtag accaacagca ccatctagcg gcagaggagg aaattaccca gtacaacaaa taggtggtaa ctatgtccac ctgccattaa 1500
gcccgagaac attaaatgcc tgggtaaaat tgatagagga aaagaaattt ggagcagaag tagtgccagg atttcaggca ctgtcagaag gttgcacccc 1600
ctatgacatt aatcagatgt taaattgtgt gggagaccat caagcggcta tgcagattat cagagatatt ataaacgagg aggctgcaga ttgggacttg 1700
cagcacccac aaccagctcc acaacaagga caacttaggg agccgtcagg atcagatatt gcaggaacaa ctagttcagt agatgaacaa atccagtgga 1800
tgtacagaca acagaacccc ataccagtag gcaacattta caggagatgg atccaactgg ggttgcaaaa atgtgtcaga atgtataacc caacaaacat 1900
tctagatgta aaacaagggc caaaagagcc atttcagagc tatgtagaca ggttctacaa aagtttaaga gcagaacaga cagatgcagc agtaaagaat 2000
tggatgactc aaacactgct gattcaaaat gctaacccag attgcaagct agtgctgaag ggctgggtg tgaatcccac cctagaagaa atgctgacgg 2100

                                  Gag p24 end \/ Gag p2 start                        Gag p2 end \/ Gag NC (p7) start
cttgtcaagg agtagggggg ccgggacaga aggctagatt aatggcagaa gccctgaaag aggccctcgc accagtgcca atccctttg cagcagccca 2200
acagagggga ccaagaaagc caattaagtg ttggaattgt gggaaagagg gacactctgc aaggcaatgc agagccccaa gaagacaggg atgctggaaa 2300
```

```
                                         Gag NC (p7) end \/ Gag p1 start
                                            ribosome -1 slip site Gag to Gag-Pol
                                                       / Pol start                          Gag p1 end \/ Gag p6 start
tgtggaaaaa tggaccatgt tatggccaaa tgcccagaca gacaggcggg tttttaggc cttggtccat ggggaaagaa gccccgcaat ttccccatgg 2400
ctcaagtgca tcaggggctg atgccaactg ctcccccaga ggacccagct gtggatctgc taaagaacta catgcagttg ggcaagcagc agagagaaaa 2500

                                                           /Pol protease start     Gag p6 end \
gcagagagaa agcagagaga agccttacaa ggaggtgaca gaggatttgc tgcacctcaa ttctctcttt ggaggagacc agtagtcact gctcatattg 2600
aaggacagcc tgtagaagta ttactggata caggggctga tgattctatt gtaacaggaa tagagttagg tccacattat accccaaaaa tagtaggagg 2700
aataggaggt tttattaata ctaaagaata caaaaatgta gaaatagaag ttttaggcaa aaggattaaa gggacaatca tgacagggga caccccgatt 2800

                                           Pol protease end \/ Pol p66 & p51 RT  start
aacattttg gtagaaattt gctaacagct ctggggatgt ctctaaattt tcccatagct aaagtagagc ctgtaaaagt cgccttaaag ccaggaaagg 2900
atggaccaaa attgaagcag tggccattat caaaagaaaa gatagttgca ttaagagaaa tctgtgaaaa gatggaaaag gatggtcagt tggaggaagc 3000
tcccccgacc aatccataca acaccccac atttgctata aagaaaaagg ataagaacaa atggagaatg ctgatagatt ttagggaact aaatagggtc 3100
actcaggact ttacggaagt ccaattagga ataccacacc ctgcaggact agcaaaaagg aaaagaatta cagtactgga tataggtgat gcatatttct 3200
ccatacctct agatgaagaa tttaggcagt acactgcctt tactttacca tcagtaaata atgcagagcc aggaaaacga tacatttata aggttctgcc 3300
tcagggatgg aaggggtcac cagccatctt ccaatacact atgagacatg tgctagaacc cttcaggaag gcaaatccag atgtgacctt agtccagtat 3400
atggatgaca tcttaatagc tagtgacagg acagacctgg aacatgacag ggtagtttta cagtcaaagg aactcttgaa tagcataggg ttttctaccc 3500
cagaagagaa attccaaaaa gatcccccat ttcaatggat ggggtacgaa ttgtggccaa caaaatggaa gttgcaaaag atagagttgc cacaaagaga 3600
gacctggaca gtgaatgata tacagaagtt agtaggagta ttaaattggg cagctcaaat ttatccaggt ataaaaacca aacatctctg taggttaatt 3700
agaggaaaaa tgactctaac agaggaagtt cagtggactg agatggcaga agcagaatat gaggaaaata aaataattct cagtcaggaa caagaaggat 3800
gttattacca agaaggcaag ccattagaag ccacggtaat aaagagtcag gacaatcagt ggtcttataa aattcaccaa gaagacaaaa tactgaaagt 3900
aggaaaattt gcaaagataa agaatacaca taccaatgga gtgagactat tagcacatgt aatacagaaa ataggaaagg aagcaatagt gatctgggga 4000
caggtcccaa aattccactt accagttgag aaggatgtat gggaacagtg gtggacagac tattggcagg taacctggat accggaatgg gattttatct 4100

                                                    Pol p51 end p66 RT continues \/ Pol p15 RNAse start
caacaccacc gctagtaaga ttagtcttca atctagtgaa ggaccctata gagggagaag aaacctatta tacagatgga tcatgtaata aacagtcaaa 4200
agaagggaaa gcaggatata tcacagatag gggcaaagac aaagtaaaag tgttagaaca gactactaat caacaagcag aattggaagc atttctcatg 4300
gcattgacag actcagggcc aaaggcaaat attatagtag attcacaata tgttatggga ataataacag gatgccctac agaatcagag agcaggctag 4400
ttaatcaaat aatagaagaa atgattaaaa agtcagaaat ttatgtagca tgggtaccag cacacaaagg tataggagga aaccaagaaa tagaccacct 4500

         Pol p15 RNAse, p66 RT end \/ Pol p31 integrase start
agttagtcaa gggattagac aagttctctt cttggaaaag atagagccag cacaagaaga acatgataaa taccatagta atgtaaaaga attggtattc 4600
aaatttggat tacccagaat agtggccaga cagatagtag acacctgtga taaatgtcat cagaaaggag aggctataca tgggcaggca aattcagatc 4700
tagggacttg gcaaatggat tgtacccatc tagagggaaa aataatcata gttgcagtac atgtagctag tggattcata gaagcagagg taattccaca 4800
agagacagga agacagacag cactatttct gttaaaattg gcaggcagat ggcctattac acatctacac acagataatg gtgctaactt tgcttcgcaa 4900
gaagtaaaga tggttgcatg gtgggcaggg atagagcaca cctttggggt accatacaat ccacagagtc agggagtagt ggaagcaatg aatcaccacc 5000
```

```
    tgaaaaatca aatagataga atcagggaac aagcaaattc agtagaaacc atagtattaa tggcagttca ttgcatgaat tttaaaagaa ggggaggaat 5100
    aggggatatg actccagcag aaagattaat taacatgatc actacagaac aagagataca atttcaacaa tcaaaaaact caaaatttaa aaattttcgg 5200
    gtctattaca gagaaggcag agatcaactg tggaagggac ccggtgagct attgtggaaa ggggaaggag cagtcatctt aaaggtaggg acagacatta 5300

                                              / Vif start
    aggtagtacc cagaagaaag gctaaaatta tcaaagatta tggaggagga aaagaggtgg atagcagttc ccacatggag gataccggag aggctagaga 5400

Pol, Gag-Pol, and
p31 integrase end \
    ggtggcatag cctcataaaa tatctgaaat ataaaactaa agatctacaa aaggtttgct atgtgcccca ttttaaggtc ggatgggcat ggtggacctg 5500
    cagcagagta atcttcccac tacaggaagg aagccattta gaagtacaag ggtattggca tttgacacca gaaaaagggt ggctcagtac ttatgcagtg 5600
    aggataacct ggtactcaaa gaacttttgg acagatgtaa caccaaacta tgcagacatt ttactgcata gcacttattt cccttgcttt acagcgggag 5700
    aagtgagaag ggccatcagg ggagaacaac tgctgtcttg ctgcaggttc ccgagagctc ataagtacca ggtaccaagc ctacagtact tagcactgaa 5800

                / Vpx start
    agtagtaagc gatgtcagat cccagggaga gaatcccacc tggaaacagt ggagaagaga caataggaga ggccttcgaa tggctaaaca gaacagtaga 5900

                                                                                          Vif end \
    ggagataaac agagaggcgg taaaccacct accaagggag ctaattttcc aggtttggca aaggtcttgg gaatactggc atgatgaaca agggatgtca 6000
    ccaagctatg taaaatacag atacttgtgt ttaatacaaa aggctttatt tatgcattgc aagaaaggct gtagatgtct aggggaagga catggggcag 6100

                                                  Vpx end \ / Vpr start
    ggggatggag accaggacct cctcctcctc cccctccagg actagcataa atggaagaaa gacctccaga aaatgaagga ccacaaaggg aaccatggga 6200
    tgaatgggta gtggaggttc tggaagaact gaaagaagaa gctttaaaac attttgatcc tcgcttgcta actgcacttg gtaatcatat ctataataga 6300

     / Tat exon 1 start
    catggagaca cccttgaggg agcaggagaa ctcattagaa tcctccaacg agcgctcttc atgcatttca gaggcggatg catccactcc agaatcggcc 6400

                                                          Vpr end \
    aacctggggg aggaaatcct ctctcagcta taccgccctc tagaagcatg ctataacaca tgctattgta aaaagtgttg ctaccattgc cagttttgtt 6500

                                 / Rev exon 1start                                             Tat, Rev exon 1 end \/ Tat, Rev intron
    ttcttaaaaa aggcttgggg atatgttatg agcaatcacg aaagagaaga agaactccga aaaaggctaa ggctaataca tcttctgcat caaacaagta 6600

       / Env gp120, gp160 start, signal peptide
    agtatgggat gtcttgggaa tcagctgctt atcgccatct tgcttttaag tgtctatggg atctattgta ctctatatgt cacagtcttt tatggtgtac 6700
    cagcttggag gaatgcgaca attcccctct tttgtgcaac caagaatagg gatacttggg gaacaactca gtgcctacca gataatggtg attattcaga 6800
    agtggccctt aatgttacag aaagctttga tgcctggaat aatacagtca cagaacaggc aatagaggat gtatggcaac tctttgagac ctcaataaag 6900
    ccttgtgtaa aattatcccc attatgcatt actatgagat gcaataaaag tgagacagat agatggggat tgacaaaatc aataacaaca acagcatcaa 7000
    caacatcaac gacagcatca gcaaaagtag acatggtcaa tgagactagt tcttgtatag cccaggataa ttgcacaggc ttggaacaag agcaaatgat 7100
    aagctgtaaa ttcaacatga cagggttaaa aagagacaag aaaaaagagt acaatgaaac ttggtactct gcagatttgg tatgtgaaca agggaataac 7200
    actggtaatg aaagtagatg ttacatgaac cactgtaaca cttctgttat ccaagagtct tgtgacaaac attattggga tgctattaga tttaggtatt 7300
    gtgcacctcc aggttatgct ttgcttagat gtaatgacac aaattattca ggctttatgc ctaaatgttc taaggtggtg gtctcttcat gcacaaggat 7400
    gatggagaca cagacttcta cttggtttgg ctttaatgga actagagcag aaaatagaac ttatatttac tggcatggta gggataatag gactataatt 7500
    agtttaaata agtattataa tctaacaatg aaatgtagaa gaccaggaaa taagacagtt ttaccagtca ccattatgtc tggattggtt ttccactcac 7600
    aaccaatcaa tgataggcca aagcaggcat ggtgttggtt tggaggaaaa tggaaggatg caataaaaga ggtgaagcag accattgtca aacatcccag 7700
    gtatactgga actaacaata ctgataaaat caatttgacg gctcctggag gaggagatcc ggaagttacc ttcatgtgga caaattgcag aggagagttc 7800
    ctctactgta aaatgaattg gtttctaaat tgggtagaag ataggaatac agctaaccag aagccaaagg aacagcataa aaggaattac gtgccatgtc 7900
    atattagaca aataatcaac acttggcata aagtaggcaa aaatgtttat ttgcctccaa gagagggaga cctcacgtgt aactccacag tgaccagtct 8000
    catagcaaac atagattgga ttgatggaaa ccaaactaat atcaccatga gtgcagaggt ggcagaactg tatcgattgg aattgggaga ttataaatta 8100

                                                                             Env gp120 end \/ Env gp41 start
    gtagagatca ctccaattgg cttggccccc acagatgtga agaggtacac tactggtggc acctcaagaa ataaaagagg ggtctttgtg ctagggttct 8200
    tgggttttct cgcaacggca ggttctgcaa tgggcgcggc gtcgttgacg ctgaccgctc agtcccgaac tttattggct gggatagtgc agcaacagca 8300
    acagctgttg gacgtggtca agagacaaca agaattgttg cgactgaccg tctggggaac aaagaacctc cagactaggg tcactgccat cgagaagtac 8400
    ttaaaggacc aggcgcagct gaatgcttgg ggatgtgcgt ttagacaagt ctgccacact actgtaccat ggccaaatgc aagtctaaca ccaaagtgga 8500
    acaatgagac ttggcaagag tgggagcgaa aggttgactt cttggaagaa aatataacag ccctcctaga ggaggcacaa attcaacaag agaagaacat 8600
    gtatgaatta caaaagttga atagctggga tgtgtttggc aattggtttg accttgcttc ttggataaag tatatacaat atggagttta tatagttgta 8700
    ggagtaatac tgttaagaat agtgatctat atagtacaaa tgctagctaa gttaaggcag ggtataggc cagtgttctc ttccccaccc tcttatttcc 8800

Tat, Rev
intron end \/ Tat, Rev exon 2 start
    agcagaccca tatccaacag gacccggcac tgccaaccag agaaggcaaa gaaagagacg gtggagaagg cggtggcaac agctcctggc cttggcagat 8900

  Tat exon 2
  end \
    agaatatatt catttcctga tccgccaact gatacgcctc ttgacttggc tattcagcaa ctgcagaacc ttgctatcga gagtatacca gatcctccaa 9000

                                                        Rev exon 2 end \                               / Nef start
    ccaatactcc agaggctctc tgcgacccta cagaggattc gagaagtcct caggactgaa ctgacctacc tacaatatgg gtggagctat ttccatgagg 9100
    cggtccaggc cgtctggaga tctgcgacag agactcttgc gggcgcgtgg ggagacttat gggagactct taggagaggt ggaagatgga tactcgcaat 9200

                                   Env gp41, gp160 end \
    ccccaggagg attagacaag ggcttgagct cactctcttg tgagggacag aaatacaatc agggacagta tatgaatact ccatggagaa acccagctga 9300
```

```
                                                 Premature in-frame stop taa in
                                                 original SIVMM239 sequence
agagagagaa aaattagcat acagaaaaca aaatatggat gatatagatg aggaagatga tgacttggta ggggtatcag tgaggccaaa agttccccta 9400

                                                                   /  3′ LTR U3 region start
agaacaatga gttacaaatt ggcaatagac atgtctcatt ttataaaaga aaagggggga ctggaaggga tttattacag tgcaagaaga catagaatct 9500
tagacatata cttagaaaag gaagaaggca tcataccaga ttggcaggat tacacctcag gaccaggaat tagataccca aagacatttg gctggctatg 9600
gaaattagtc cctgtaaatg tatcagatga ggcacaggag gatgaggagc attatttaat gcatccagct caaacttccc agtgggatga cccttgggga 9700
gaggttctag catggaagtt tgatccaact ctggcctaca cttatgaggc atatgttaga tacccagaag agtttggaag caagtcaggc ctgtcagagg 9800
                                                                   Nef end \
aagaggttag aagaaggcta accgcaagag gccttcttaa catggctgac aagaaggaaa ctcgctgaaa cagcagggac tttccacaag ggatgttac 9900

                                                                    3′ LTR U3 region end \/  3′ LTR R repeat start
ggggaggtac tggggaggag ccggtcggga acgcccactt tcttgatgta taaatatcac tgcatttcgc tctgtattca gtcgctctgc ggagaggctg 10000
gcagattgag ccctgggagg ttctctccag cactagcagg tagagcctgg gtgttccctg ctagactctc accagcactt ggccggtgct gggcagagtg 10100

                                              3′ LTR repeat end  \/ 3′ LTR U5 region start
actccacgct tgcttgctta aagccctctt caataaagct gccattttag aagtaagcta gtgtgtgttc ccatctctcc tagccgccgc ctggtcaact 10200

                                                                    3′ LTR U5 region end  \
cggtactcaa taataagaag accctggtct gttaggaccc tttctgcttt gggaaaccga agcaggaaaa tccctagc                        10278
```