Numbering Positions in HIV Relative to HXB2CG

Bette T. Korber,[1] **Brian T. Foley,**[1] **Carla L. Kuiken,**[1] **Satish K. Pillai,**[1] **and Joseph G. Sodroski**[2]

[1] *Theoretical Biology and Biophysics, Group T-10, MS K710, Los Alamos National Laboratory, Los Alamos, NM 87545;*
[2] *Howard Hughes Medical Institute, Columbia University, New York, NY 10032*

In this section we present a simple numbering scheme to facilitate the identification of the position number or precise location of interest in HIV DNA or proteins.

Inconsistent and inaccurate numbering of locations in HIV DNA and protein sequences is a serious problem in the HIV literature. Therefore we decided to provide a practical guide to help circumvent these problems in the future, and to attempt to bring a common language into discussions in the field. We present a clearly numbered set of proteins, and the full length genome, for HIV HXB2, GenBank accession number K03455. HIV HXB2 is also known as: HXBc2, for HXB clone 2; HXB2R, in the Los Alamos HIV database, with the R for revised, as it was slightly revised relative to the original HXB2 sequence; and HXB2CG in GenBank, for HXB2 complete genome. Our web site has an interactive program to further facilitate obtaining position numbers relative to HXB2CG (`http://hiv-web.lanl.gov/NUM-HXB2/HXB2.MAIN.html`).

HXB2 was selected as the prototype, because this virus is the most commonly used reference strain for many different kinds of functional studies. Importantly, all of the envelope structural data published to date translates residue numbers into the HXB2 numbering scheme. Now that a core HIV-1 gp120 structure is solved (for review, see Wyatt et al, this compendium), it has become apparent that conservation in core sequences, especially in hydrophobic interior domains, exists to preserve similar folding in gp120 variants. As the envelope protein is riddled with insertions and deletions, it is particularly problematic for numbering. The current system, of sequentially numbering proteins from any strain, lacks a common way to refer to specific locations in a protein. We propose the following system to circumvent this problem:

1) ***Case of insertion in sequence relative to HXB2CG***. Use residue number/alphabet (e.g., 131a, 131b, 131c, etc.) to refer to residues in variable regions that are "extra" compared to what HXB2 has. A similar scheme has been used for immunoglobulin complementarity-determining region (CDR) loops (see Lucas et al., *J Immunol* 1998 **161**:3776–80 (1998) for an example).

Example: If the region under study is `LLLTRDGGSNRSEPEVEIFRP` of ENVB, gp120,

```
452          465    470  HXB2 amino acid position from start of gp160
 |            |      |
 LLLTRDGGNSNNES--EIFRP   HXB2
 LLLTRDGGSNRSEPEVEIFRP   ENVB
```

one could refer to it as corresponding to HXB2 gp160 position numbers 452–470 with a two base insertion (465a = E and 465b = V)

2) ***Case of deletion in sequence relative to HXB2***. Indicate the deleted residues.

Example: If the region under study is LLLTRDGGNN of 92RW020.5,

```
452         463  HXB2 amino acid position from start of gp120
 |           |
 LLLTRDGGNSNN   HXB2
 LLLTRDGG..NN   92RW020.5
```

one refers to it as corresponding to HXB2 gp160 position numbers 452–463 with a two base deletion at positions 460–461. We suggest using the annotation 452-463(del 460-461) to make this explicit.

The sequential numbering relative to either 92RW020.5 or ENVB could also be provided in the above two examples, but the HXB2 numbering should also be provided as a reference.

The benefit of this numbering strategy is that, for example, aspartate 368, which is involved in CD4 binding, or gp160 368 D, means the same thing to everyone working on envelope glycoproteins, regardless of the reference strain they used in their particular studies.

Also, when working with a short functional domain, epitope, or primer, researchers should publish the precise amino acid or nucleotide string that they are working with, as well as the HXB2 numbered positions, to ensure that there is no confusion (for example, write out ENVB `LLLTRDGGSNRSEPEVEIFRP` as well as give the boundary position numbers).

HXB2 Amino Acid Sequence Numbering:

Gag Pr55 Gag precursor (Assemblin)

```
MGARASVLSG GELDRWEKIR LRPGGKKKYK LKHIVWASRE LERFAVNPGL LETSEGCRQI LGQLQPSLQT GSEELRSLYN TVATLYCVHQ RIEIKDTKEA  100
LDKIEEEQNK SKKKAQQAAA DTGHSNQVSQ NYPIVQNIQG QMVHQAISPR TLNAWVKVVE EKAFSPEVIP MFSALSEGAT PQDLNTMLNT VGGHQAAMQM  200
LKETINEEAA EWDRVHPVHA GPIAPGQMRE PRGSDIAGTT STLQEQIGWM TNNPPIPVGE IYKRWIILGL NKIVRMYSPT SILDIRQGPK EPFRDYVDRF  300
YKTLRAEQAS QEVKNWMTET LLVQNANPDC KTILKALGPA ATLEEMMTAC QGVGGPGHKA RVLAEAMSQV TNSATIMMQR GNFRNQRKIV KCFNCGKEGH  400
TARNCRAPRK KGCWKCGKEG HQMKDCTERQ ANFLGKIWPS YKGRPGNFLQ SRPEPTAPPE ESFRSGVETT TPPQKQEPID KELYPLTSLR SLFGNDPSSQ  500
```

Gag p17 Matrix

```
MGARASVLSG GELDRWEKIR LRPGGKKKYK LKHIVWASRE LERFAVNPGL LETSEGCRQI LGQLQPSLQT GSEELRSLYN TVATLYCVHQ RIEIKDTKEA  100
LDKIEEEQNK SKKKAQQAAA DTGHSNQVSQ NY                                                                          132
```

Gag p24 Capsid

```
PIVQNIQGQM VHQAISPRTL NAWVKVVEEK AFSPEVIPMF SALSEGATPQ DLNTMLNTVG GHQAAMQMLK ETINEEAAEW DRVHPVHAGP IAPGQMREPR  100
GSDIAGTTST LQEQIGWMTN NPPIPVGEIY KRWIILGLNK IVRMYSPTSI LDIRQGPKEP FRDYVDRFYK TLRAEQASQE VKNWMTETLL VQNANPDCKT  200
ILKALGPAAT LEEMMTACQG VGGPGHKARV L                                                                            231
```

Gag p2

```
AEAMSQVTNS ATIM                                                                                               14
```

Gag p7 Nucleocapsid

```
MQRGNFRNQR KIVKCFNCGK EGHTARNCRA PRKKGCWKCG KEGHQMKDCT ERQAN                                                  55
```

Gag p1

```
FLGKIWPSYK GRPGNF                                                                                             16
```

Gag p6

```
LQSRPEPTAP PEESFRSGVE TTTPPQKQEP IDKELYPLTS LRSLFGNDPS SQ                                                     52
```

Pol polyprotein:

```
FFREDLAFLQ GKAREFSSEQ TRANSPTRRE LQVWGRDNNS PSEAGADRQG TVSFNFPQVT LWQRPLVTIK IGGQLKEALL DTGADDTVLE EMSLPGRWKP  100
KMIGGIGGFI KVRQYDQILI EICGHKAIGT VLVGPTPVNI IGRNLLTQIG CTLNFPISPI ETVPVKLKPG MDGPKVKQWP LTEEKIKALV EICTEMEKEG  200
KISKIGPENP YNTPVFAIKK KDSTKWRKLV DFRELNKRTQ DFWEVQLGIP HPAGLKKKKS VTVLDVGDAY FSVPLDEDFR KYTAFTIPSI NNETPGIRYQ  300
YNVLPQGWKG SPAIFQSSMT KILEPFRKQN PDIVIYQYMD DLYVGSDLEI GQHRTKIEEL RQHLLRWGLT TPDKKHQKEP PFLWMGYELH PDKWTVQPIV  400
LPEKDSWTVN DIQKLVGKLN WASQIYPGIK VRQLCKLLRG TKALTEVIPL TEEAELELAE NREILKEPVH GVYYDPSKDL IAEIQKQGQG QWTYQIYQEP  500
FKNLKTGKYA RMRGAHTNDV KQLTEAVQKI TTESIVIWGK TPKFKLPIQK ETWETWWTEY WQATWIPEWE FVNTPPLVKL WYQLEKEPIV GAETFYVDGA  600
ANRETKLGKA GYVTNRGRQK VVTLTDTTNQ KTELQAIYLA LQDSGLEVNI VTDSQYALGI IQAQPDQSES ELVNQIIEQL IKKEKVYLAW VPAHKGIGGN  700
EQVDKLVSAG IRKVLFLDGI DKAQDEHEKY HSNWRAMASD FNLPPVVAKE IVASCDKCQL KGEAMHGQVD CSPGIWQLDC THLEGKVILV AVHVASGYIE  800
AEVIPAETGQ ETAYFLLKLA GRWPVKTIHT DNGSNFTGAT VRAACWWAGI KQEFGIPYNP QSQGVVESMN KELKKIIGQV RDQAEHLKTA VQMAVFIHNF  900
KRKGGIGGYS AGERIVDIIA TDIQTKELQK QITKIQNFRV YYRDSRNPLW KGPAKLLWKG EGAVVIQDNS DIKVVPRRKA KIIRDYGKQM AGDDCVASRQ 1000
DED                                                                                                          1003
```

Pol p10 Protease

```
PQVTLWQRPL VTIKIGGQLK EALLDTGADD TVLEEMSLPG RWKPKMIGGI GGFIKVRQYD QILIEICGHK AIGTVLVGPT PVNIIGRNLL TQIGCTLNF    99
```

Pol p66 Reverse Transcriptase (RT/RNAse)

```
PISPIETVPV KLKPGMDGPK VKQWPLTEEK IKALVEICTE MEKEGKISKI GPENPYNTPV FAIKKKDSTK WRKLVDFREL NKRTQDFWEV QLGIPHPAGL  100
KKKKSVTVLD VGDAYFSVPL DEDFRKYTAF TIPSINNETP GIRYQYNVLP QGWKGSPAIF QSSMTKILEP FRKQNPDIVI YQYMDDLYVG SDLEIGQHRT  200
KIEELRQHLL RWGLTTPDKK HQKEPPFLWM GYELHPDKWT VQPIVLPEKD SWTVNDIQKL VGKLNWASQI YPGIKVRQLC KLLRGTKALT EVIPLTEEAE  300
LELAENREIL KEPVHGVYYD PSKDLIAEIQ KQGQGQWTYQ IYQEPFKNLK TGKYARMRGA HTNDVKQLTE AVQKITTESI VIWGKTPKFK LPIQKETWET  400
WWTEYWQATW IPEWEFVNTP PLVKLWYQLE KEPIVGAETF YVDGAANRET KLGKAGYVTN RGRQKVVTLT DTTNQKTELQ AIYLALQDSG LEVNIVTDSQ  500
YALGIIQAQP DQSESELVNQ IIEQLIKKEK VYLAWVPAHK GIGGNEQVDK LVSAGIRKVL                                              560
```

Pol p51 RT

```
PISPIETVPV KLKPGMDGPK VKQWPLTEEK IKALVEICTE MEKEGKISKI GPENPYNTPV FAIKKKDSTK WRKLVDFREL NKRTQDFWEV QLGIPHPAGL  100
KKKKSVTVLD VGDAYFSVPL DEDFRKYTAF TIPSINNETP GIRYQYNVLP QGWKGSPAIF QSSMTKILEP FRKQNPDIVI YQYMDDLYVG SDLEIGQHRT  200
KIEELRQHLL RWGLTTPDKK HQKEPPFLWM GYELHPDKWT VQPIVLPEKD SWTVNDIQKL VGKLNWASQI YPGIKVRQLC KLLRGTKALT EVIPLTEEAE  300
LELAENREIL KEPVHGVYYD PSKDLIAEIQ KQGQGQWTYQ IYQEPFKNLK TGKYARMRGA HTNDVKQLTE AVQKITTESI VIWGKTPKFK LPIQKETWET  400
WWTEYWQATW IPEWEFVNTP PLVKLWYQLE KEPIVGAETF                                                                   440
```

Pol p15 RNAse

```
YVDGAANRET KLGKAGYVTN RGRQKVVTLT DTTNQKTELQ AIYLALQDSG LEVNIVTDSQ YALGIIQAQP DQSESELVNQ IIEQLIKKEK VYLAWVPAHK  100
GIGGNEQVDK LVSAGIRKVL                                                                                        120
```

Pol p31 Integrase

```
FLDGIDKAQD EHEKYHSNWR AMASDFNLPP VVAKEIVASC DKCQLKGEAM HGQVDCSPGI WQLDCTHLEG KVILVAVHVA SGYIEAEVIP AETGQETAYF  100
LLKLAGRWPV KTIHTDNGSN FTGATVRAAC WWAGIKQEFG IPYNPQSQGV VESMNKELKK IIGQVRDQAE HLKTAVQMAV FIHNFKRKGG IGGYSAGERI  200
VDIIATDIQT KELQKQITKI QNFRVYYRDS RNPLWKGPAK LLWKGEGAVV IQDNSDIKVV PRRKAKIIRD YGKQMAGDDC VASRQDED              288
```

Vif

```
MENRWQVMIV WQVDRMRIRT WKSLVKHHMY VSGKARGWFY RHHYESPHPR ISSEVHIPLG DARLVITTYW GLHTGERDWH LGQGVSIEWR KKRYSTQVDP  100
ELADQLIHLY YFDCFSDSAI RKALLGHIVS PRCEYQAGHN KVGSLQYLAL AALITPKKIK PPLPSVTKLT EDRWNKPQKT KGHRGSHTMN GH           192
```

Vpr HXB2 frameshift \/

```
MEQAPEDQGP QREPHNEWTL ELLEELKNEA VRHFPRIWLH GLGQHIYETY GDTWAGVEAI IRILQQLLFI HFRIGCRHSR IGVTRQRRAR NGASRS       96
```

Tat (premature HXB2 stop codon indicated by $) | Primary splice site

```
MEPVDPRLEP WKHPGSQPKT ACTNCYCKKC CFHCQVCFIT KALGISYGRK KRRQRRRAHQ NSQTHQASLS KQPTSQPRGD PTGPKE$KKK VERETETDPF  100
D                                                                                                               101
```

Rev | Primary splice site

```
MAGRSGDSDE ELIRTVRLIK LLYQSNPPPN PEGTRQARRN RRRRWRERQR QIHSISERIL GTYLGRSAEP VPLQLPPLER LTLDCNEDCG TSGTQGVGSP  100
QILVESPTVL ESGTKE                                                                                               116
```

Vpu (defective start codon)

```
TQPIPIVAIV ALVVAIIIAI VVWSIVIIEY RKILRQRKID RLIDRLIERA EDSGNESEGE ISALVEMGVE MGHHAPWDVD DL                       82
```

Envelope (Env) gp160

Env signal peptide |

```
MRVKEKYQHL WRWGWRWGTM LLGMLMICSA TEKLWVTVYY GVPVWKEATT TLFCASDAKA YDTEVHNVWA THACVPTDPN PQEVVLVNVT ENFNMWKNDM  100
VEQMHEDIIS LWDQSLKPCV KLTPLCVSLK CTDLKNDTNT NSSSGRMIME KGEIKNCSFN ISTSIRGKVQ KEYAFFYKLD IIPIDNDTTS YKLTSCNTSV  200
ITQACPKVSF EPIPIHYCAP AGFAILKCNN KTFNGTGPCT NVSTVQCTHG IRPVVSTQLL LNGSLAEEEV VIRSVNFTDN AKTIIVQLNT SVEINCTRPN  300
NNTRKRIRIQ RGPGRAFVTI GKIGNMRQAH CNISRAKWNN TLKQIASKLR EQFGNNKTII FKQSSGGDPE IVTHSFNCGG EFFYCNSTQL FNSTWFNSTW  400
STEGSNNTEG SDTITLPCRI KQIINMWQKV GKAMYAPPIS GQIRCSSNIT GLLLTRDGGN SNNESEIFRP GGGDMRDNWR SELYKYKVVK IEPLGVAPTK  500
```

> gp41 start

```
AKRRVVQREK RAVGIGALFL GFLGAAGSTM GAASMTLTVQ ARQLLSGIVQ QQNNLLRAIE AQQHLLQLTV WGIKQLQARI LAVERYLKDQ QLLGIWGCSG  600
KLICTTAVPW NASWSNKSLE QIWNHTTWME WDREINNYTS LIHSLIEESQ NQQEKNEQEL LELDKWASLW NWFNITNWLW YIKLFIMIVG GLVGLRIVFA  700
VLSIVNRVRQ GYSPLSFQTH LPTPRGPDRP EGIEEEGGER DRDRSIRLVN GSLALIWDDL RSLCLFSYHR LRDLLLIVTR IVELLGRRGW EALKYWWNLL  800
QYWSQELKNS AVSLLNATAI AVAEGTDRVI EVVQGACRAI RHIPRRIRQG LERILL                                                   856
```

Nef (premature HXB2 stop codon indicated by $)

```
MGGKWSKSSV IGWPTVRERM RRAEPAADRV GAASRDLEKH GAITSSNTAA TNAACAWLEA QEEEEVGFPV TPQVPLRPMT YKAAVDLSHF LKEKGGLEGL  100
IHSQRRQDIL DLWIYHTQGY FPD$QNYTPG PGVRYPLTFG WCYKLVPVEP DKIEEANKGE NTSLLHPVSL HGMDDPEREV LEWRFDSRLA FHHVARELHP  200
EYFKNC                                                                                                          206
```

HXB2 Nucleotide Sequence Numbering:

```
> 5′ LTR U3 region start
tggaagggct aattcactcc caacgaagac aagatatcct tgatctgtgg atctaccaca cacaaggcta cttccctgat tagcagaact acacaccagg  100
gccagggatc agatatccac tgacctttgg atggtgctac aagctagtac cagttgagcc agagaagtta gaagaagcca acaaaggaga gaacaccagc  200
ttgttacacc ctgtgagcct gcatggaatg gatgacccgg agagagaagt gttagagtgg aggtttgaca gccgcctagc atttcatcac atggcccgag  300
agctgcatcc ggagtacttc aagaactgct gacatcgagc ttgctacaag ggactttccg ctggggactt tccagggagg cgtggcctgg gcgggactgg  400

                                    5′ LTR U3 region end \/ 5′ LTR R repeat start
ggagtggcga gccctcagat cctgcatata agcagctgct ttttgcctgt actgggtctc tctggttaga ccagatctga gcctgggagc tctctggcta  500

                                          5′ LTR R      5′ LTR U5
                                          repeat end \/ region start
actagggaac ccactgctta agcctcaata aagcttgcct tgagtgcttc aagtagtgtg tgcccgtctg ttgtgtgact ctggtaacta gagatccctc  600

               5′ LTR U5 region end <
agacccttttt agtcagtgtg gaaaatctct agcagtggcg cccgaacagg gacctgaaag cgaaagggaa accagaggag ctctctcgac gcaggactcg  700

                                                                                                  > Gag p17 start
gcttgctgaa gcgcgcacgg caagaggcga ggggcggcga ctggtgagta cgccaaaaat tttgactagc ggaggctaga aggagagaga tgggtgcgag  800
agcgtcagta ttaagcgggg gagaattaga tcgatgggaa aaaattcggt taaggccagg gggaaagaaa aaatataaat taaaacatat agtatgggca  900
agcagggagc tagaacgatt cgcagttaat cctggcctgt tagaaacatc agaaggctgt agacaaatac tgggacagct acaaccatcc cttcagacag 1000
gatcagaaga acttagatca ttatataata cagtagcaac cctctattgt gtgcatcaaa ggatagagat aaaagacacc aaggaagctt tagacaagat 1100

                                                                                    Gag p17 end \/ Gag p24 start
agaggaagag caaaacaaaa gtaagaaaaa agcacagcaa gcagcagctg acacaggaca cagcaatcag gtcagccaaa attaccctat agtgcagaac 1200
atccaggggc aaatggtaca tcaggccata tcacctagaa ctttaaatgc atgggtaaaa gtagtagaag agaaggcttt cagcccagaa gtgataccca 1300
tgttttcagc attatcagaa ggagccaccc cacaagattt aaacaccatg ctaaacacag tgggggggca tcaagcagcc atgcaaatgt taaaagagac 1400
catcaatgag gaagctgcag aatgggatag agtgcatcca gtgcatgcag ggcctattgc accaggccag atgagagaac caaggggaag tgacatagca 1500
ggaactacta gtaccttca ggaacaaata ggatggatga caaataatcc acctatccca gtaggagaaa tttataaaag atggataatc ctgggattaa 1600
ataaaatagt aagaatgtat agccctacca gcattctgga cataagacaa ggaccaaagg aacccttttag agactatgta gaccggttct ataaaactct 1700
aagagccgag caagcttcac aggaggtaaa aaattggatg acagaaacct tgttggtcca aaatgcgaac ccagattgta agactatttt aaaagcattg 1800

                                                                  Gag p24 Capsid end \/ Gag p2 start
ggaccagcgg ctacactaga agaaatgatg acagcatgtc agggagtagg aggacccggc cataaggcaa gagttttggc tgaagcaatg agccaagtaa 1900
```

```
         Gag p2 end \ / Gag p7 Nucleocapsid start
caaattcagc taccataatg atgcagagag gcaattttag gaaccaaaga aagattgtta agtgtttcaa ttgtggcaaa gaagggcaca cagccagaaa 2000

                                                                                 ribosome -1 slip Gag to Gag-Pol
                                                                                               ______
                                                                       Gag p7 nucleocapsid end \/Gag p1 start
                                                                                     Pol start >
ttgcagggcc cctaggaaaa agggctgttg gaaatgtgga aaggaaggac accaaatgaa agattgtact gagagacagg ctaatttttt agggaagatc 2100

                        Gag p1 end \/ Gag p6 start
tggccttcct acaagggaag gccagggaat tttcttcaga gcagaccaga gccaacagcc ccaccagaag agagcttcag gtctggggta gagacaacaa 2200

                                                            > Pol protease start        Gag p6 end <
ctccccctca gaagcaggag ccgatagaca aggaactgta tcctttaact tccctcaggt cactctttgg caacgacccc tcgtcacaat aaagataggg 2300
gggcaactaa aggaagctct attagataca ggagcagatg atacagtatt agaagaaatg agtttgccag gaagatggaa accaaaaatg atagggggaa 2400
ttggaggttt tatcaaagta agacagtatg atcagatact catagaaatc tgtggacata aagctatagg tacagtatta gtaggaccta cacctgtcaa 2500

                                          Pol protease end \/ Pol p66 and p51 RT start
cataattgga agaaatctgt tgactcagat tggttgcact ttaaattttc ccattagccc tattgagact gtaccagtaa aattaaagcc aggaatggat 2600
ggcccaaaag ttaaacaatg gccattgaca gaagaaaaaa taaaagcatt agtagaaatt tgtacagaga tggaaaagga agggaaaatt tcaaaaattg 2700
ggcctgaaaa tccatacaat actccagtat ttgccataaa gaaaaaagac agtactaaat ggagaaaatt agtagatttc agagaactta ataagagaac 2800
tcaagacttc tgggaagttc aattaggaat accacatccc gcagggttaa aaaagaaaaa atcagtaaca gtactggatg tgggtgatgc atatttttca 2900
gttcccttag atgaagactt caggaagtat actgcattta ccatacctag tataaacaat gagacaccag ggattagata tcagtacaat gtgcttccac 3000
agggatggaa aggatcacca gcaatattcc aaagtagcat gacaaaaatc ttagagcctt ttagaaaaca aaatccagac atagttatct atcaatacat 3100
ggatgatttg tatgtaggat ctgacttaga aatagggcag catagaacaa aaatagagga gctgagacaa catctgttga ggtggggact taccacacca 3200
gacaaaaaac atcagaaaga acctccattc ctttggatgg gttatgaact ccatcctgat aaatggacag tacagcctat agtgctgcca gaaaaagaca 3300
gctggactgt caatgacata cagaagttag tggggaaatt gaattgggca agtcagattt acccagggat taaagtaagg caattatgta aactccttag 3400
aggaaccaaa gcactaacag aagtaatacc actaacagaa gaagcagagc tagaactggc agaaaacaga gagattctaa aagaaccagt acatggagtg 3500
tattatgacc catcaaaaga cttaatagca gaaatacaga agcaggggca aggccaatgg acatatcaaa tttatcaaga gccatttaaa aatctgaaaa 3600
caggaaaata tgcaagaatg aggggtgccc acactaatga tgtaaaacaa ttaacagagg cagtgcaaaa aataaccaca gaaagcatag taatatgggg 3700
aaagactcct aaatttaaac tgcccataca aaaggaaaca tgggaaacat ggtggacaga gtattggcaa gccacctgga ttcctgagtg ggagtttgtt 3800

                                                   Pol p51 end p66 RT continue \/ Pol p15 RNAse H start
aatacccctc ccttagtgaa attatggtac cagttagaga aagaacccat agtaggagca gaaaccttct atgtagatgg ggcagctaac agggagacta 3900
aattaggaaa agcaggatat gttactaata gaggaagaca aaaagttgtc accctaactg acacaacaaa tcagaagact gagttacaag caatttatct 4000
agctttgcag gattcgggat tagaagtaaa catagtaaca gactcacaat atgcattagg aatcattcaa gcacaaccag atcaaagtga atcagagtta 4100
gtcaatcaaa taatagagca gttaataaaa aaggaaaagg tctatctggc atgggtacca gcacacaaag gaattggagg aaatgaacaa gtagataaat 4200
```

```
      Pol RNAse H, p66 RT end \/ Pol p31 Integrase start
tagtcagtgc tggaatcagg aaagtactat ttttagatgg aatagataag gcccaagatg aacatgagaa atatcacagt aattggagag caatggctag 4300
tgattttaac ctgccacctg tagtagcaaa agaaatagta gccagctgtg ataaatgtca gctaaaagga gaagccatgc atggacaagt agactgtagt 4400
ccaggaatat ggcaactaga ttgtacacat ttagaaggaa aagttatcct ggtagcagtt catgtagcca gtggatatat agaagcagaa gttattccag 4500
cagaaacagg gcaggaaaca gcatattttc ttttaaaatt agcaggaaga tggccagtaa aaacaataca tactgacaat ggcagcaatt tcaccggtgc 4600
tacggttagg gccgcctgtt ggtgggcggg aatcaagcag gaatttggaa ttccctacaa tccccaaagt caaggagtag tagaatctat gaataaagaa 4700
ttaaagaaaa ttataggaca ggtaagagat caggctgaac atcttaagac agcagtacaa atggcagtat tcatccacaa ttttaaaaga aaaggggga 4800
ttgggggta cagtgcaggg gaaagaatag tagacataat agcaacagac atacaaacta aagaattaca aaaacaaatt acaaaaattc aaaattttcg 4900
ggtttattac agggacagca gaaatccact ttggaaagga ccagcaaagc tcctctggaa aggtgaaggg gcagtagtaa tacaagataa tagtgacata 5000

                                          > Vif start                        Pol, p31 Integrase end <
aaagtagtgc caagaagaaa agcaaagatc attagggatt atggaaaaca gatggcaggt gatgattgtg tggcaagtag acaggatgag gattagaaca 5100
tggaaaagtt tagtaaaaca ccatatgtat gtttcaggga aagctagggg atggttttat agacatcact atgaaagccc tcatccaaga ataagttcag 5200
aagtacacat cccactaggg gatgctagat tggtaataac aacatattgg ggtctgcata caggagaaag agactggcat ttgggtcagg gagtctccat 5300
agaatggagg aaaaagagat atagcacaca agtagaccct gaactagcag accaactaat tcatctgtat tactttgact gtttttcaga ctctgctata 5400
agaaaggcct tattaggaca catagttagc cctaggtgtg aatatcaagc aggacataac aaggtaggat ctctacaata cttggcacta gcagcattaa 5500

                                                          > Vpr start
taacaccaaa aaagataaag ccacctttgc ctagtgttac gaaactgaca gaggatagat ggaacaagcc ccagaagacc aagggccaca gagggagcca 5600

        Vif end <
cacaatgaat ggacactaga gcttttagag gagcttaaga atgaagctgt tagacatttt cctaggattt ggctccatgg cttagggcaa catatctatg 5700
aaacttatgg ggatacttgg gcaggagtgg aagccataat aagaattctg caacaactgc tgtttatcca ttttcagaat tgggtgtcga catagcagaa 5800

                      Tat start >        Vpr end <
taggcgttac tcgacagagg agagcaagaa atggagccag tagatcctag actagagccc tggaagcatc caggaagtca gcctaaaact gcttgtacca 5900

                                                            Rev start >
attgctattg taaaaagtgt tgctttcatt gccaagtttg tttcataaca aaagccttag gcatctccta tggcaggaag aagcggagac agcgacgaag 6000

                              Tat, Rev exon end \/Tat, Rev intron  > Vpu start (defective ACG start codon)
agctcatcag aacagtcaga ctcatcaagc ttctctatca aagcagtaag tagtacatgt aacgcaacct ataccaatag tagcaatagt agcattagta 6100
gtagcaataa taatagcaat agttgtgtgg tccatagtaa tcatagaata taggaaaata ttaagacaaa gaaaaataga caggttaatt gatagactaa 6200

                           > Env gp160 start, signal peptide
tagaaagagc agaagacagt ggcaatgaga gtgaaggaga aatatcagca cttgtggaga tgggggtgga gatggggcac catgctcctt gggatgttga 6300
```

```
Vpu end, signal peptide end
      <
tgatctgtag tgctacagaa aaattgtggg tcacagtcta ttatggggta cctgtgtgga aggaagcaac caccactcta ttttgtgcat cagatgctaa 6400
agcatatgat acagaggtac ataatgtttg ggccacacat gcctgtgtac ccacagaccc caacccacaa gaagtagtat tggtaaatgt gacagaaaat 6500
tttaacatgt ggaaaaatga catggtagaa cagatgcatg aggatataat cagtttatgg gatcaaagcc taaagccatg tgtaaaatta accccactct 6600
gtgttagttt aaagtgcact gatttgaaga atgatactaa taccaatagt agtagcggga gaatgataat ggagaaagga gagataaaaa actgctcttt 6700
caatatcagc acaagcataa gaggtaaggt gcagaaagaa tatgcatttt tttataaact tgatataata ccaatagata atgatactac cagctataag 6800
ttgacaagtt gtaacacctc agtcattaca caggcctgtc caaaggtatc ctttgagcca attcccatac attattgtgc cccggctggt tttgcgattc 6900
taaaatgtaa taataagacg ttcaatggaa caggaccatg tacaaatgtc agcacagtac aatgtacaca tggaattagg ccagtagtat caactcaact 7000
gctgttaaat ggcagtctag cagaagaaga ggtagtaatt agatctgtca atttcacgga caatgctaaa accataatag tacagctgaa cacatctgta 7100
gaaattaatt gtacaagacc caacaacaat acaagaaaaa gaatccgtat ccagagagga ccagggagag catttgttac aataggaaaa ataggaaata 7200
tgagacaagc acattgtaac attagtagag caaaatggaa taacacttta aaacagatag ctagcaaatt aagagaacaa tttggaaata ataaaacaat 7300
aatctttaag caatcctcag gaggggaccc agaaattgta acgcacagtt ttaattgtgg aggggaattt ttctactgta attcaacaca actgtttaat 7400
agtacttggt ttaatagtac ttggagtact gaagggtcaa ataacactga aggaagtgac acaatcaccc tcccatgcag aataaaacaa attataaaca 7500
tgtggcagaa agtaggaaaa gcaatgtatg cccctcccat cagtggacaa attagatgtt catcaaatat tacagggctg ctattaacaa gagatggtgg 7600
taatagcaac aatgagtccg agatcttcag acctggagga ggagatatga gggacaattg gagaagtgaa ttatataaat ataaagtagt aaaaattgaa 7700

                                          Env gp120 end \/ Env gp41 start
ccattaggag tagcacccac caaggcaaag agaagagtgg tgcagagaga aaaaagagca gtgggaatag gagctttgtt ccttgggttc ttgggagcag 7800
caggaagcac tatgggcgca gcctcaatga cgctgacggt acaggccaga caattattgt ctggtatagt gcagcagcag aacaatttgc tgagggctat 7900
tgaggcgcaa cagcatctgt tgcaactcac agtctggggc atcaagcagc tccaggcaag aatcctggct gtggaaagat acctaaagga tcaacagctc 8000
ctggggattt ggggttgctc tggaaaactc atttgcacca ctgctgtgcc ttggaatgct agttggagta ataaatctct ggaacagatt tggaatcaca 8100
cgacctggat ggagtgggac agagaaatta acaattacac aagcttaata cactccttaa ttgaagaatc gcaaaaccag caagaaaaga atgaacaaga 8200
attattggaa ttagataaat gggcaagttt gtggaattgg tttaacataa caaattggct gtggtatata aaattattca taatgatagt aggaggcttg 8300

                                                        Tat, Rev intron end \/ Tat, Rev exon 2 start
gtaggtttaa gaatagtttt tgctgtactt tctatagtga atagagttag gcagggatat tcaccattat cgtttcagac ccacctccca accccgaggg 8400

                        ___ Tat premature stop                  Tat end <
gacccgacag gcccgaagga atagaagaag aaggtggaga gagagacaga gacagatcca ttcgattagt gaacggatcc ttggcactta tctgggacga 8500
tctgcggagc ctgtgcctct tcagctacca ccgcttgaga gacttactct tgattgtaac gaggattgtg gaacttctgg gacgcagggg gtgggaagcc 8600

                                        Rev end <
ctcaaatatt ggtggaatct cctacagtat tggagtcagg aactaaagaa tagtgctgtt agcttgctca atgccacagc catagcagta gctgagggga 8700
```

```
                                                                    Env gp41, gp160 end <    > Nef start
cagatagggt tatagaagta gtacaaggag cttgtagagc tattcgccac atacctagaa gaataagaca gggcttggaa aggattttgc tataagatgg 8800
gtggcaagtg gtcaaaaagt agtgtgattg gatggcctac tgtaagggaa agaatgagac gagctgagcc agcagcagat agggtgggag cagcatctcg 8900
agacctggaa aaacatggag caatcacaag tagcaataca gcagctacca atgctgcttg tgcctggcta gaagcacaag aggaggagga ggtgggtttt 9000

                                                                                  > 3′ LTR U3 region
ccagtcacac ctcaggtacc tttaagacca atgacttaca aggcagctgt agatcttagc cacttttttaa aagaaaaggg gggactggaa gggctaattc 9100

                                                                  ___ Nef premature stop
actcccaaag aagacaagat atccttgatc tgtggatcta ccacacacaa ggctacttcc ctgattagca gaactacaca ccagggccag gggtcagata 9200
tccactgacc tttggatggt gctacaagct agtaccagtt gagccagata agatagaaga ggccaataaa ggagagaaca ccagcttgtt acaccctgtg 9300
agcctgcatg ggatggatga cccggagaga gaagtgttag agtggaggtt tgacagccgc ctagcatttc atcacgtggc ccgagagctg catccggagt 9400

      Nef end <
acttcaagaa ctgctgacat cgagcttgct acaagggact ttccgctggg gactttccag ggaggcgtgg cctgggcggg actggggagt ggcgagccct 9500

                         3′ LTR U3 region \ / 3′ LTR R repeat
cagatcctgc atataagcag ctgctttttg cctgtactgg gtctctctgg ttagaccaga tctgagcctg ggagctctct ggctaactag ggaacccact 9600

                      3′ LTR R repeat \/ 3′ LTR U5 region
gcttaagcct caataaagct tgccttgagt gcttcaagta gtgtgtgccc gtctgttgtg tgactctggt aactagagat ccctcagacc cttttagtca 9700

    3′ LTR U5 end  <
gtgtggaaaa tctctagca                                                                                    9719
```

This numbering was based on previous HIV sequence database annotation, cross-checked with protein structure databases, Tozser et al., *FEBS letters* **281**:77–80 (1991), and R. J. Gorelick and L. E. Henderson, *Human Retroviruses and AIDS 1994*, part III, pages 2–10.