Supporting Information

Gifford *et al.* 10.1073/pnas.0807873105

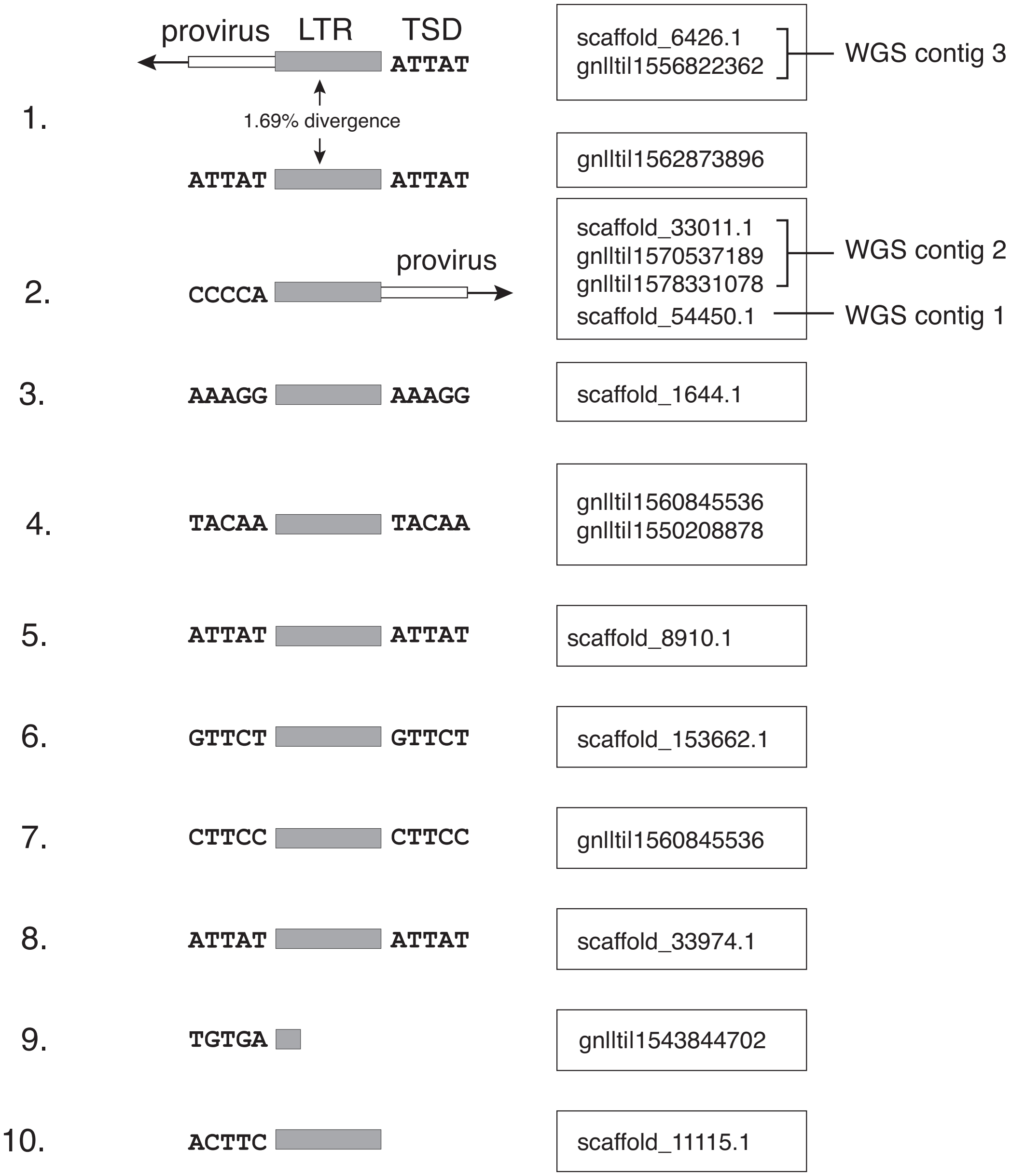


Fig. S1. pSIVgml insertion sites. Ten distinct pSIVgml insertions identified in the low coverage *M. murinus* genome. The ten insertions shown included two distinct full-length insertions (i.e., insertions encoding internal regions), along with nine solo LTRs (one of which was identified at the same locus as a full-length insertion). The IDs of sequences from WGS sequence assembly and trace archives are shown in boxes to the right (those that begin 'scaffold' are from WGS data, all others are from trace archive data). Sequences used to create the WGS contigs (1, 2, and 3) illustrated in Fig. 1 are indicated. Distinct insertions were identified through comparison of genomic DNA and target site duplication (TSD) sequences flanking viral insertions. For each of the insertions shown, at least 30 bp of unambiguously distinct genomic flanking sequence was present. TSD sequences—5-bp stretches of DNA flanking viral insertions that are generated during integration—are shown for each insertion. Where no flanking sequences were available, sequences that could be assembled into contigs were conservatively assumed to belong to the same pSIVgml insertion. At locus (1) both a solo LTR and full-length version of pSIVgml were identified, indicating that solo LTR formation had occurred on one chromosome, but not the other. The divergence between the solo LTR and the 3′ LTR of the full-length insertion at this locus was 1.69%.

1. Leitner T, *et al.*, eds (2005) *HIV Sequence Compendium 2005* (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, New Mexico).

```
        5' LTR
1       TGGAAGGAGATTTTGGGTGCAGCTGAGCACCCTAGTGACTGTGTGGGTGCAGCTGAGCACCCAGGTGACTGTGTACCAAC  80

81      GAGGGGTGCGAAAAGCCAACGAGAGGTGCGAAAGGTGCTGTGTGAGCATACAGGAAACCACAACCGCAGACTGTCTTTCA  160

161     CACCTTTACAGCTTACACAAGGACTTTGCTTTCTATTTGGGGAAGGGGGGCTACTTCAGTACTTGGGGCTTGGGGAGGGC  240
                                                TAR
241     TTGGGGAGCATATATAAGCCTGAGGTTGCCTAACCTCGAGGCCCCCTCACACATCTCTGGGTCCGGCCATCACCCAGACT  320

321     CCAGAGTGTGGATCCACAATAAAGCTGTGCATCTTGGCCCAGAGCCGTGTGGGTATGCCAAGTCTTCTTGCCCTGGGGAA  400

401     GGCAATGCAAGTTGGCCCTTCCAGCTGGCGCCCAACGTGGGGCTGGACTTGATTTCTGCAAACGGTGAGTTTGKGGGAGT  480
                                 -------------------
                                        PBS

                   → Gag (MA)
                    M  E  K  G  K  M  G  D  G  Q  C  L  G  R  D  T  G  K  K  A  Q  R  V
481     TGCAGGGCAAGATGGAAAAGGGGAAGATGGGTGATGGACAGTGTCTAGGGAGAGACACGGKGAAAAAGGCACAGCGGGTA  560

         R  V  R  G  T  G  K  P  M  H  K  L  G  N  F  V  W  A  V  K  I  A  A  A  V  A  E
561     AGAGTGAGAGGTACCGGAAAGCCCATGCACAAGTTAGGGAACTTTGTCTGGGCAGTAAAGATAGCTGCAGCAGTCGCAGA  640

          R  S  I  D  K  T  A  V  G  V  F  R  R  W  P  P  K  G  G  A  G  D  I  N  V  A
641     AAGATCTATCGACAAAACAGCTGTGGGGGTTTTTAGGCGCTGGCCACCAAAGGGGGGGGCAGGGGACATTAATGTAGCAC  720

        L  D  T  L  L  C  Y  G  L  Q  R  K  G  S  Q  W  K  V  Q  R  V  P  Q  M  W  Q  R
721     TAGACACCCTGCTATGTTATGGGTTACAGCGAAAGGGTTCACAGTGGAAAGTACAGAGGGTGCCGCAGATGTGGCAAAGG  800

                                                                   MA end  → CA start
         W  V  G  W  Q  E  Q  L  Y  G  K  K  E  Q  D  P  E  A  E  A  A  A  Y |P  V  V  N
801     TGGGTAGGATGGCAAGAACAGCTCTATGGGAAGAAAGAGCAGGATCCGGAGGCAGAAGCAGCTGCCTACCCGGTTGTAAA  880

          R  G  Q  G  W  A  Y  E  P  M  S  T  R  T  V  A  A  W  I  R  Q  T  R  E  K  G
881     TAGGGGACAGGGGTGGGCTTATGAGCCTATGAGTACCAGAACTGTCGCAGCATGGATTCGACAGACTAGAGAGAAGGGAC  960

        L  T  S  P  E  T  I  T  Y  W  G  L  I  S  Q  D  L  S  S  R  E  Q  V  Q  L  L  E
961     TTACTAGTCCAGAGACAATCACATATTGGGGTTTAATATCCCAAGATCTGTCCAGCAGGGAACAGGTCCAACTGCTGGAA  1040

         V  I  P  G  L  Q  A  D  K  D  M  L  G  A  Y  L  G  E  R  A  R  E  W  D  A  Q  P
1041    GTCATTCCAGGACTTCAGGCAGACAAGGATATGCTGGGGGCATATCTAGGAGAAAGGGCACGTGAGTGGGATGCGCAACC  1120

          Q  Q  P  L  P  Y  T  F  C  S  Y  W  G  I  W  Q  G  I  R  P  F  A  I  S  A  Q
1121    ACAACAGCCATTGCCCTATACTTTCTGCTCATATTGGGGGATTTGGCAGGGGATCAGGCCTTTTGCCATATCAGCGCAAG  1200
```

Fig. S2. pSIVgml consensus sequence. Locations of the proteins encoded by the *gag*, *pol* and *env* genes were determined via homology to the HIV-1 reference sequence HXB2 (1), and by searches against the pFAM database (2). Tat, Rev, and Vif were identified by genomic location, and by the identification of the conserved 'SQV' motif in Vif and a predicted NLS in Rev (3). Also shown is a putative ORF extending into the 3′ LTR. Putative promoter and polyadenylation signals are indicated in bold type. All lentiviruses have PBS sequences specific for tRNALys; however, pSIVgml is unique amongst primate lentiviruses in utilizing tRNALys1,2 rather than tRNALys3. Two regions of nucleic acid secondary structure, TAR and the RRE, are highlighted in dark gray. Black lines adjacent to the corresponding nucleotide sequences indicate the PBS and PPT sequences.

2. Finn RD, *et al*. (2006) Pfam: Clans, web tools and services. *Nucleic Acids Res* 34:D247–D251.
3. Pollard VW, Malim MH (1998) The HIV-1 Rev protein. *Annu Rev Microbiol* 52:491–532.

```
        → Gag (CA)
        G  R  E  A  A  Q  V  F  R  A  W  I  T  Q  G  L  M  S  L  A  Q  L  Q  A  P  H  P
1201    GGCGAGAAGCTGCACAGGTCTTTAGGGCCTGGATAACGCAGGGCTTAATGAGCTTGGCCCAGTTGCAGGCGCCACACCCA    1280

         G  S  T  K  I  L  Q  G  P  K  E  P  Y  G  E  F  I  N  R  L  F  L  Q  I  N  Q  E
1281    GGGTCAACAAAGATCTTGCAGGGACCAAAAGAGCCTTATGGGGAATTTATTAATAGACTGTTTCTCCAGATTAATCAGGA    1360

          G  A  P  E  E  V  K  T  Y  L  K  G  H  L  S  I  E  N  A  N  A  D  C  Q  K  A
1361    AGGAGCCCCAGAGGAAGTAAAGACATATCTTAAGGGGCATCTCAGCATTGAAAACGCTAATGCAGATTGCCAAAGGCCA    1440

 CA truncated ‖→ NC (start truncated)
                     → Pol (start 5’ truncated)
                      A  R  D  D  T  S  A  A  F  C  T  G  T  S  P  R  G  R  D  G  V  V
        M  S  H  L |R  L  E  M  T  P  V  Q  P  S  A  L  E  P  P  Q  G  G  E  M  G  L  Y
1441    TGAGTCATCTTAGGCTAGAGATGACACCAGTGCAGCCTTCTGCACTGGAACCTCCCCAAGGGGGAGAGATGGGGTTGTAC    1520

                   → PR   NC ‖
        P  H  P  Q |G  I  I  K  L |D  H  R  P  L  V  T  L  K  V  G  R  Q  S  V  T  L  L
         P  T  L  K  E  L  L  N  *
1521    CCCACCCTCAAGGAATTATTAAATTAGATCACCGGCCCCTAGTGACTCTGAAGGTAGGGAGACAATCGGTCACCCTCCTG    1600

         I  D  T  G  A  D  N  T  I  I  H  P  K  D  W  K  P  V  G  M  E  E  G  I  I  N  I
1601    ATAGACACCGGGGCTGACAATACAATTATCCATCCTAAAGATTGGAAACCAGTAGGAATGGAAGAGGGGATAATTAACAT    1680

          G  G  I  G  G  S  Q  K  G  I  L  Y  K  Q  V  P  I  S  L  A  D  R  Q  I  R  R
1681    AGGGGGAATTGGAGGTTCTCAAAAGGGGATATTATACAAACAAGTACCTATTAGTCTAGCAGATAGGCAGATACGGAGAA    1760

        T  V  I  R  A  V  T  P  I  N  L  L  G  R  D  N  L  V  S  L  G  I  G  V  V  M  L
1761    CTGTCATAAGGGCAGTGACCCCTATAAATTTACTAGGGAGGGACAATTTAGTATCACTAGGAATTGGAGTAGTGATGCTA    1840

 PR ‖→ RT
         M |A  Q  M  S  V  K  I  V  P  L  P  V  E  L  M  P  G  C  D  G  P  R  V  K  Q  W
1841    ATGGCACAAATGTCAGTAAAAATAGTGCCGCTGCCAGTTGAGTTAATGCCTGGCTGTGATGGGCCAAGAGTAAAACAGTG    1920

          P  L  T  Q  E  K  Y  Q  A  L  A  E  I  V  S  K  L  E  K  E  G  K  V  S  I  A
1921    GCCCCTTACGCAAGAGAAATATCAGGCTCTTGCTGAAATAGTATCTAAATTAGAAAAAGAGGGAAAAGTCAGTATAGCAG    2000

        E  V  S  N  P  Y  N  T  P  V  F  A  I  K  K  K  S  G  K  W  R  M  L  I  D  F  R
2001    AGGTAAGTAATCCCTACAACACTCCGGTGTTTGCCATTAAGAAAAAATCAGGCAAATGGAGAATGCTCATTGACTTTCGA    2080

         V  L  N  A  R  T  K  K  G  A  E  F  Q  L  G  L  P  H  P  A  G  L  Q  K  K  D  N
2081    GTGCTAAATGCTCGAACCAAAAAGGGAGCTGAATTTCAACTGGGCTTGCCTCACCCCGCCGGCCTACAAAAAAAAGATAA    2160
```

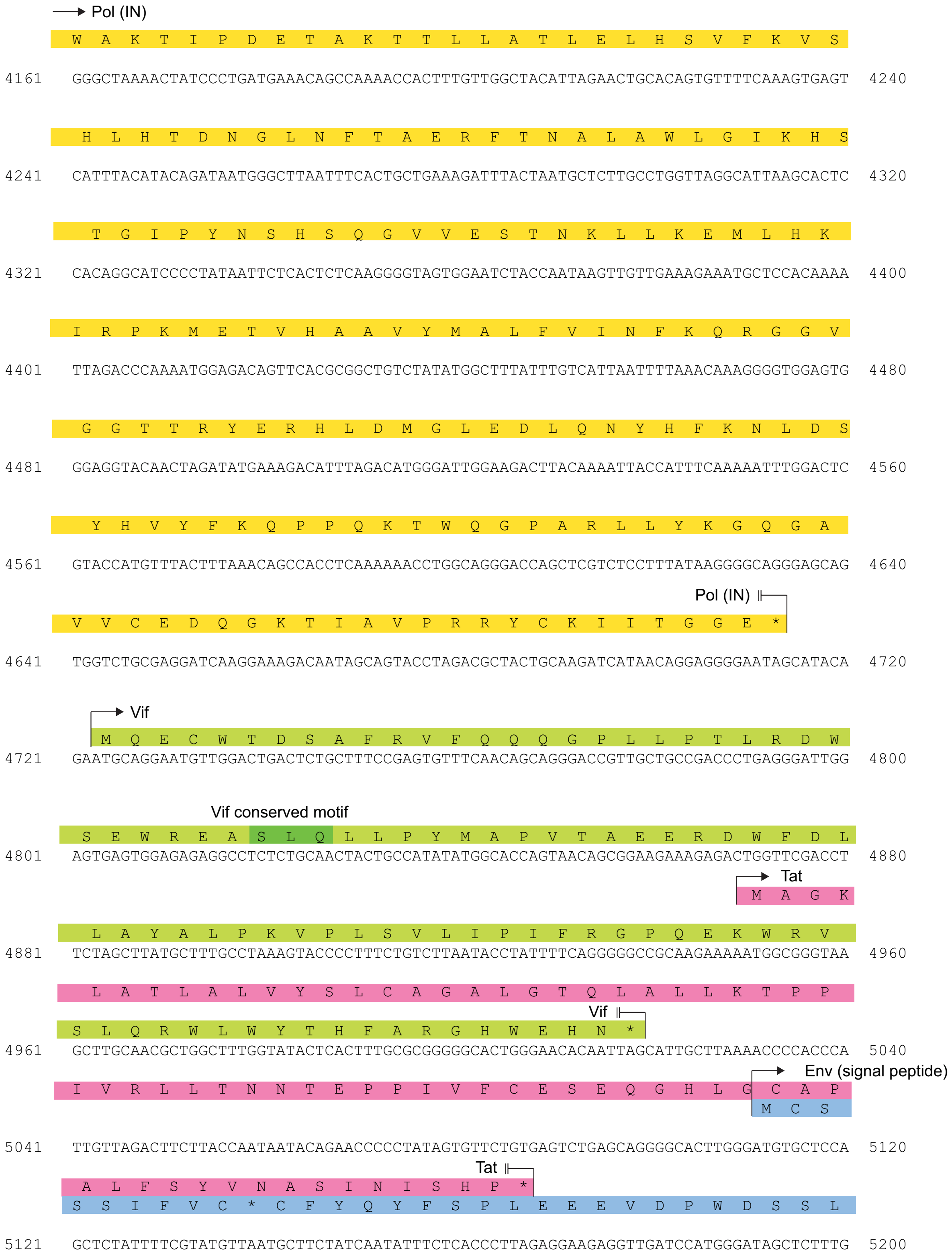
Fig. S2. continued.

```
         ──► Pol (RT)
           V  T  I  L  D  I  G  D  A  Y  F  T  I  P  S  D  P  T  F  K  K  Y  T  A  F  T

2161     TGTCACCATACTAGATATTGGTGATGCTTATTTTACCATACCATCGGACCCCACATTTAAAAAGTACACTGCCTTTACTC  2240

         L  I  P  P  N  N  Q  G  P  A  K  R  F  V  F  N  C  L  P  Q  G  W  V  C  S  P  A

2241     TAATACCACCTAATAATCAGGGACCAGCCAAGAGATTTGTGTTTAATTGTCTACCACAAGGGTGGGTGTGTAGTCCAGCC  2320

          F  Y  Q  R  T  M  S  D  I  L  Q  L  W  K  Q  A  H  L  E  V  M  L  Y  V  Y  M  D

2321     TTTTACCAAAGGACTATGAGTGACATCTTACAACTATGGAAACAAGCTCATCTTGAGGTCATGTTATATGTCTACATGGA  2400

           D  L  L  I  G  T  D  L  P  L  R  E  H  R  R  L  V  Q  E  L  R  S  M  L  L  G

2401     TGACCTTCTAATCGGGACAGACCTCCCCTTGAGGGAGCATAGAAGGCTGGTCCAAGAGCTTAGGAGTATGCTTCTTGGCT  2480

         W  S  F  E  T  P  E  E  N  V  Q  D  Q  W  P  L  Q  *  M  G  Y  E  L  H  P  N  N

2481     GGAGCTTTGAAACTCCTGAAGAAAACGTGCAGGACCAGTGGCCGCTACAGTAGATGGGATATGAGTTACACCCTAATAAT  2560

          W  Q  L  Q  V  R  K  L  E  L  P  D  H  P  T  L  N  E  V  P  K  L  V  G  I  I  N

2561     TGGCAGTTGCAGGTCCGAAAATTAGAGTTACCAGATCACCCCACTTTAAATGAAGTCCCAAAACTGGTGGGAATTATTAA  2640

           W  A  S  Q  I  I  S  G  L  K  I  K  K  L  T  A  M  M  A  G  N  Q  D  L  N  R

2641     TTGGGCTAGTCAAATTATTTCAGGGCTTAAAATAAAGAAGCTTACTGCTATGATGGCAGGGAATCAGGATCTCAATAGAA  2720

         K  I  E  W  T  K  E  A  R  K  E  A  E  E  A  A  K  L  L  *  E  L  P  A  G  G  Y

2721     AGATAGAATGGACTAAGGAAGCTAGAAAGGAAGCTGAAGAGGCAGCTAAGCTGCTCTAGGAGCTCCCAGCAGGGGGGTAT  2800

          V  D  P  L  K  Q  V  E  A  R  I  A  F  V  G  F  L  E  V  T  Y  D  V  H  Q  E  N

2801     GTCGACCCCTTGAAACAGGTGGAGGCTAGAATAGCTTTTGTAGGTTTCCTCGAGGTAACCTACGACGTCCATCAGGAGAA  2880

           I  I  L  W  C  G  R  V  G  S  S  K  A  Y  C  N  R  W  I  C  M  ?  K  A  A  I

2881     TATCATCCTTTGGTGTGGCAGAGTAGGGTCCAGCAAAGCTTATTGTAACCGGTGGATCTGTATG-TAAAAGCAGCAATAA  2960

         K  I  G  R  E  T  L  I  R  L  G  V  V  P  I  L  T  V  P  V  A  K  D  I  W  D  S

2961     AAATAGGCAGGGAAACCTTGATTAGGCTAGGAGTGGTCCCCATACTCACTGTCCCAGTAGCAAAAGATATCTGGGATAGT  3040

          F  A  W  T  A  G  E  V  A  W  F  P  E  V  R  H  Q  A  P  P  L  I  V  E  R  W  M

3041     TTTGCATGGACTGCAGGGGAGGTAGCATGGTTCCCAGAAGTAAGACATCAGGCACCTCCTCTAATAGTTGAGAGATGGAT  3120
                                              RT ├──► RNaseH
           K  M  V  S  E  P  I  K  G  A  R  T  W| Y  I  D  G  S  K  K  R  A  Q  K  A  R

3121     GAAAATGGTATCAGAACCCATTAAGGGGGCAAGAACTTGGTATATCGATGGATCTAAGAAACGGGCCCAAAAAGCTAGAG  3200
```

Fig. S2. continued.

```
      → Pol (RT)
        A  G  I  W  T  E  G  E  K  A  Q  V  Q  E  L  E  G  S  N  Q  K  A  E  L  A  A  L

3201    CAGGAATTTGGACAGAGGGAGAGAAGGCACAAGTACAGGAACTGGAGGGCTCAAATCAAAAAGCAGAATTGGCAGCCTTA    3280

         L  Y  A  L  Q  Q  E  D  Q  E  L  N  I  I  T  D  S  Q  Y  V  M  K  V  L  R  L  V

3281    TTGTATGCCTTACAGCAGGAAGACCAAGAATTAAACATTATCACTGATTCTCAATATGTAATGAAAGTGCTGCGACTCGT    3360

          P  W  V  S  D  S  P  L  V  Q  S  I  I  Q  A  V  E  K  K  Q  A  I  Y  L  D  W

3361    GCCATGGGTTAGCGATTCTCCCTTGGTGCAGAGCATCATACAAGCAGTAGAGAAAAAACAGGCTATCTATTTAGATTGGG    3440
                                                                            RNaseH ⊩→ dUTPase
        V  P  G  H  K  G  I  P  G  N  H  K  I  D  E  E  I  Q  Y  W  Q  G  L  V  I |Q  G

3441    TGCCAGGTCATAAGGGAATCCCAGGAAATCATAAAATTGATGAAGAAATTCAATATTGGCAAGGTTTGGTTATCCAAGGC    3520

         T  G  I  L  P  K  R  E  E  D  V  G  Y  D  L  Q  I  P  E  D  V  Y  L  Q  G  L  E

3521    ACAGGTATCCTTCCTAAAAGAGAAGAGGATGTAGGCTATGATTTACAAATTCCAGAAGATGTGTACCTGCAGGGCTTGGA    3600

          R  R  S  V  P  L  N  L  *  V  Q  W  E  K  D  Q  W  G  L  I  V  A  K  S  S  M

3601    AAGGCGGTCCGTTCCGTTGAACTTGTGAGTTCAATGGGAAAAAGACCAATGGGGGTTGATTGTGGCAAAGTCCTCTATGG    3680

        A  Q  M  G  V  I  P  L  G  G  V  I  D  S  G  Y  R  G  P  I  I  I  I  L  W  N  L

3681    CTCAGATGGGGGTGATTCCTTTAGGTGGAGTCATAGATTCTGGGTATAGAGGACCCATCATCATCATCCTATGGAATCTT    3760

         N  R  K  A  V  L  L  K  A  G  K  R  V  A  Q  L  V  I  M  S  L  L  H  E  E  L  Q

3761    AATAGAAAGGCAGTACTCCTTAAAGCCGGAAAAAGAGTGGCTCAACTAGTTATAATGTCTCTACTTCATGAGGAGTTGCA    3840
                                                                     dUTPase ⊩→ IN
          Q  V  Q  Q  V  K  I  D  T  A  R  G  E  G  A  F  G  S  T  G  T  Y |F  L  E  A

3841    ACAAGTTCAGCAGGTCAAAATTGACACGGCCCGAGGTGAAGGAGCATTTGGTTCCACTGGAACCTATTTCTTGGAGGCCA    3920

        I  P  R  A  E  S  D  H  E  L  W  H  S  G  V  K  A  L  M  Q  D  F  G  I  S  Q  M

3921    TCCCTAGAGCAGAAAGTGATCATGAACTATGGCACTCGGGGGTTAAAGCTCTCATGCAGGATTTTGGAATATCTCAAATG    4000

         V  A  K  A  I  V  H  K  C  P  N  C  Q  G  K  G  S  A  I  T  G  V  V  D  Y  T  P

4001    GTGGCTAAAGCCATCGTGCATAAATGTCCTAATTGCCAAGGGAAAGGGTCTGCCATTACAGGGGTGGTGGATTACACCCC    4080

          G  T  W  Q  M  D  V  T  H  W  E  G  H  K  L  L  V  A  V  E  T  A  S  G  L  T

4081    GGGGACATGGCAGATGGATGTTACCCACTGGGAAGGACATAAACTGTTAGTAGCAGTTGAGACTGCTTCTGGGTTAACAT    4160
```

Fig. S2. continued.

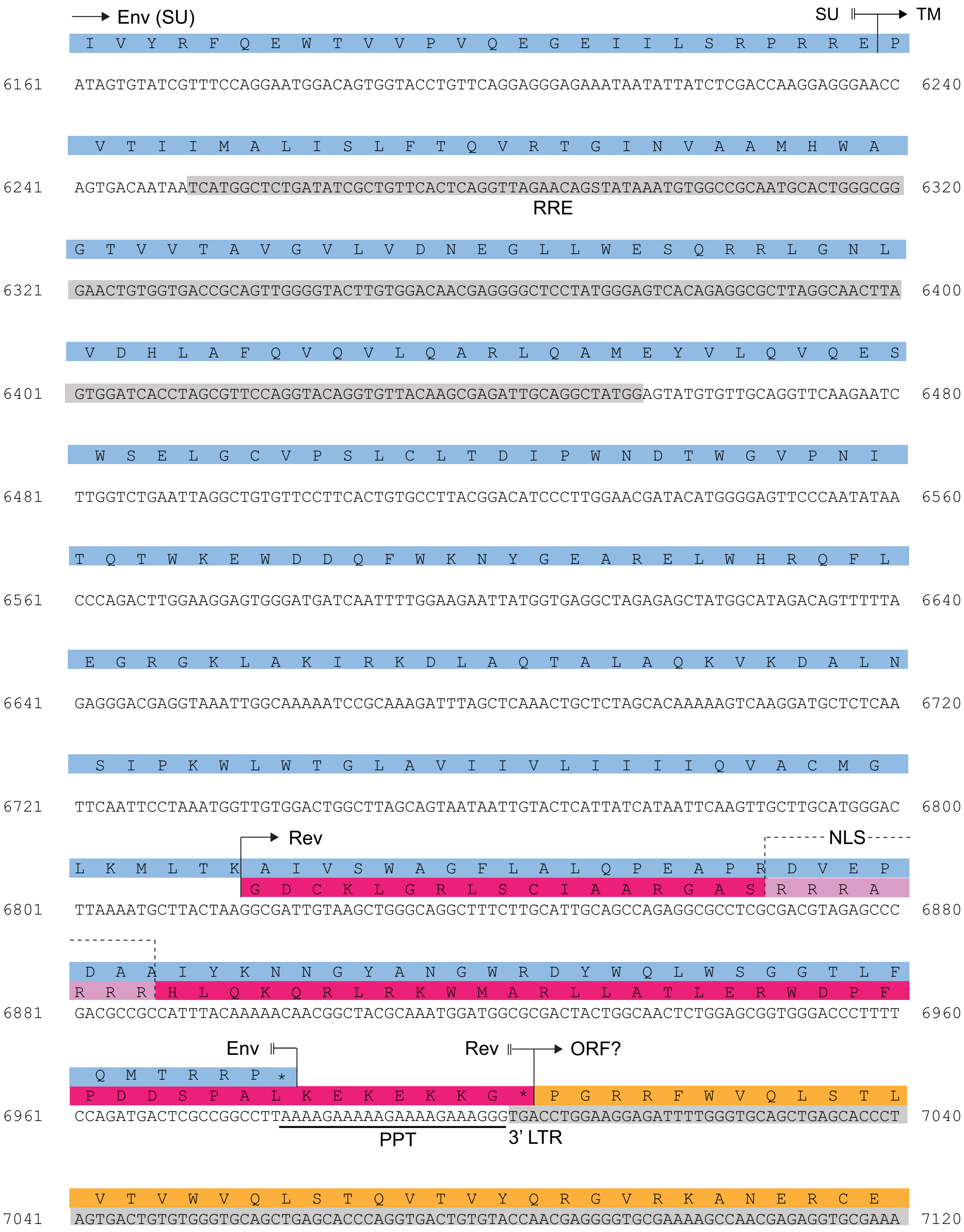


```
       Pol (IN)
       W  A  K  T  I  P  D  E  T  A  K  T  T  L  L  A  T  L  E  L  H  S  V  F  K  V  S

4161   GGGCTAAAACTATCCCTGATGAAACAGCCAAAACCACTTTGTTGGCTACATTAGAACTGCACAGTGTTTTCAAAGTGAGT   4240

        H  L  H  T  D  N  G  L  N  F  T  A  E  R  F  T  N  A  L  A  W  L  G  I  K  H  S

4241   CATTTACATACAGATAATGGGCTTAATTTCACTGCTGAAAGATTTACTAATGCTCTTGCCTGGTTAGGCATTAAGCACTC   4320

         T  G  I  P  Y  N  S  H  S  Q  G  V  V  E  S  T  N  K  L  L  K  E  M  L  H  K

4321   CACAGGCATCCCCTATAATTCTCACTCTCAAGGGGTAGTGGAATCTACCAATAAGTTGTTGAAAGAAATGCTCCACAAAA   4400

       I  R  P  K  M  E  T  V  H  A  A  V  Y  M  A  L  F  V  I  N  F  K  Q  R  G  G  V

4401   TTAGACCCAAAATGGAGACAGTTCACGCGGCTGTCTATATGGCTTTATTTGTCATTAATTTTAAACAAAGGGGTGGAGTG   4480

        G  G  T  T  R  Y  E  R  H  L  D  M  G  L  E  D  L  Q  N  Y  H  F  K  N  L  D  S

4481   GGAGGTACAACTAGATATGAAAGACATTTAGACATGGGATTGGAAGACTTACAAAATTACCATTTCAAAAATTTGGACTC   4560

         Y  H  V  Y  F  K  Q  P  P  Q  K  T  W  Q  G  P  A  R  L  L  Y  K  G  Q  G  A

4561   GTACCATGTTTACTTTAAACAGCCACCTCAAAAAACCTGGCAGGGACCAGCTCGTCTCCTTTATAAGGGGCAGGGAGCAG   4640
                                                                        Pol (IN)
       V  V  C  E  D  Q  G  K  T  I  A  V  P  R  R  Y  C  K  I  I  T  G  G  E  *

4641   TGGTCTGCGAGGATCAAGGAAAGACAATAGCAGTACCTAGACGCTACTGCAAGATCATAACAGGAGGGGAATAGCATACA   4720

          Vif
          M  Q  E  C  W  T  D  S  A  F  R  V  F  Q  Q  Q  G  P  L  L  P  T  L  R  D  W
4721   GAATGCAGGAATGTTGGACTGACTCTGCTTTCCGAGTGTTTCAACAGCAGGGACCGTTGCTGCCGACCCTGAGGGATTGG   4800

                       Vif conserved motif
        S  E  W  R  E  A  S  L  Q  L  L  P  Y  M  A  P  V  T  A  E  E  R  D  W  F  D  L
4801   AGTGAGTGGAGAGAGGCCTCTCTGCAACTACTGCCATATATGGCACCAGTAACAGCGGAAGAAAGAGACTGGTTCGACCT   4880
                                                                             Tat
                                                                             M  A  G  K

         L  A  Y  A  L  P  K  V  P  L  S  V  L  I  P  I  F  R  G  P  Q  E  K  W  R  V
4881   TCTAGCTTATGCTTTGCCTAAAGTACCCCTTTCTGTCTTAATACCTATTTTCAGGGGGCCGCAAGAAAAATGGCGGGTAA   4960

         L  A  T  L  A  L  V  Y  S  L  C  A  G  A  L  G  T  Q  L  A  L  L  K  T  P  P
                                                            Vif
       S  L  Q  R  W  L  W  Y  T  H  F  A  R  G  H  W  E  H  N  *
4961   GCTTGCAACGCTGGCTTTGGTATACTCACTTTGCGCGGGGGCACTGGGAACACAATTAGCATTGCTTAAAACCCCACCCA   5040
                                                                              Env (signal peptide)
       I  V  R  L  L  T  N  N  T  E  P  P  I  V  F  C  E  S  E  Q  G  H  L  G  C  A  P
                                                                              M  C  S

5041   TTGTTAGACTTCTTACCAATAATACAGAACCCCCTATAGTGTTCTGTGAGTCTGAGCAGGGGCACTTGGGATGTGCTCCA   5120
                                                 Tat
        A  L  F  S  Y  V  N  A  S  I  N  I  S  H  P  *
       S  S  I  F  V  C  *  C  F  Y  Q  Y  F  S  P  L  E  E  E  V  D  P  W  D  S  S  L

5121   GCTCTATTTTCGTATGTTAATGCTTCTATCAATATTTCTCACCCTTAGAGGAAGAGGTTGATCCATGGGATAGCTCTTTG   5200
```

Fig. S2. continued.

```
        → Env (signal peptide)   → SU
         G  L  F  T  D  W  V  S |G  A  H  M  Q  W  L  T  Q  R  A  Q  E  W  R  G  Y  C  Q
5201    GGATTGTTCACAGACTGGGTATCCGGAGCCCATATGCAGTGGCTGACTCAAAGAGCCCAGGAATGGAGGGGATATTGCCA  5280

          P  M  N  C  T  Q  A  N  N  F  T  R  N  C  T  R  P  Y  V  D  Y  E  S  R  P  E
5281    ACCAATGAACTGTACGCRGGCTAATAACTTTACTAGAAATTGTACCAGGCCTTATGTGGATTATGAAAGTAGACCTGAAA  5360

        N  I  Q  E  T  I  S  H  M  Q  L  N  C  T  N  S  T  C  V  W  K  E  C  K  Q  R  L
5361    ACATTCAGGAGACAATTTCACATATGCAGTTAAATTGTACTAATTCAACCTGTGTGTGGAAAGAGTGTAAACAAAGATTG  5440

         F  F  R  G  N  P  P  L  D  A  Q  T  F  R  L  C  V  R  P  P  F  A  L  R  R  C  P
5441    TTCTTCCGGGGTAACCCACCTCTTGATGCCCAAACCTTTAGACTTTGTGTTAGACCACCTTTTGCTTTAAGAAGATGTCC  5520

          P  T  N  R  T  D  W  R  Q  P  Y  K  C  S  E  Q  C  L  T  S  C  T  E  A  V  N
5521    ACCAACCAATAGGACGGACTGGAGGCAACCTTATAAGTGCTCTGAGCAATGCCTAACTTCCTGTACAGAGGCAGTAAATA  5600

        I  T  V  E  T  L  W  Q  S  Q  G  V  L  N  P  N  Q  T  G  V  S  C  Y  Q  E  G  M
5601    TAACTGTCGAAACTTTGTGGCAATCACAGGGAGTGTTAAACCCAAATCAAACAGGGGTGAGCTGTTACCAAGAGGGAATG  5680

         R  V  T  V  Q  T  E  H  D  P  I  G  I  K  V  L  Q  T  V  K  I  P  K  M  T  C  N
5681    AGGGTAACAGTCCAAACTGAGCACGACCCCATTGGGATAAAGGTCTTGCAGACTGTGAAAATACCAAAAATGACCTGTAA  5760

          L  T  G  A  Q  N  N  S  G  Q  K  G  I  V  D  P  C  Y  F  L  C  Y  N  A  T  K
5761    TCTTACAGGAGCTCAAAATAACAGTGGTCAAAAGGGCATAGTTGACCCCTGTTATTTCCTTTGCTATAATGCCACAAAGA  5840

        K  G  R  G  G  N  G  N  P  I  V  L  I  S  C  K  Y  N  G  T  S  G  T  L  T  N  C
5841    AAGGAAGAGGAGGCAACGGCAATCCCATAGTGCTTATCTCTTGTAAGTACAATGGGACGTCTGGGACGTTAACCAATTGT  5920

         E  R  V  F  K  V  S  M  P  G  P  Q  D  P  L  Y  Y  P  T  Y  P  G  E  K  W  L  L
5921    GAAAGAGTTTTTAAAGTGTCAATGCCAGGGCCACAAGATCCCCTATATTATCCAACCTATCCTGGGGAAAAGTGGTTACT  6000

          H  L  P  M  E  E  T  G  D  P  V  Q  C  N  A  S  F  Q  W  L  S  R  S  V  A  L
6001    GCATCTCCCAATGGAGGAAACAGGGGATCCAGTACAGTGTAATGCCTCTTTCCAGTGGCTATCAAGGAGTGTGGCGTTAC  6080

        H  D  T  G  V  L  K  P  N  I  Y  S  S  F  G  A  E  E  A  W  R  D  M  V  E  N  Y
6081    ACGACACTGGGGTACTGAAGCCCAATATCTATAGCTCCTTTGGGGCAGAGGAAGCATGGAGAGATATGGTAGAAAACTAC  6160
```

Fig. S2. continued.



```
           I   V   Y   R   F   Q   E   W   T   V   V   P   V   Q   E   G   E   I   I   L   S   R   P   R   R   E   P
6161   ATAGTGTATCGTTTCCAGGAATGGACAGTGGTACCTGTTCAGGAGGGAGAAATAATATTATCTCGACCAAGGAGGGAACC   6240

         V   T   I   I   M   A   L   I   S   L   F   T   Q   V   R   T   G   I   N   V   A   A   M   H   W   A
6241   AGTGACAATAATCATGGCTCTGATATCGCTGTTCACTCAGGTTAGAACAGSTATAAATGTGGCCGCAATGCACTGGGCGG   6320

       G   T   V   V   T   A   V   G   V   L   V   D   N   E   G   L   L   W   E   S   Q   R   R   L   G   N   L
6321   GAACTGTGGTGACCGCAGTTGGGGTACTTGTGGACAACGAGGGGCTCCTATGGGAGTCACAGAGGCGCTTAGGCAACTTA   6400

         V   D   H   L   A   F   Q   V   Q   V   L   Q   A   R   L   Q   A   M   E   Y   V   L   Q   V   Q   E   S
6401   GTGGATCACCTAGCGTTCCAGGTACAGGTGTTACAAGCGAGATTGCAGGCTATGGAGTATGTGTTGCAGGTTCAAGAATC   6480

         W   S   E   L   G   C   V   P   S   L   C   L   T   D   I   P   W   N   D   T   W   G   V   P   N   I
6481   TTGGTCTGAATTAGGCTGTGTTCCTTCACTGTGCCTTACGGACATCCCTTGGAACGATACATGGGGAGTTCCCAATATAA   6560

       T   Q   T   W   K   E   W   D   D   Q   F   W   K   N   Y   G   E   A   R   E   L   W   H   R   Q   F   L
6561   CCCAGACTTGGAAGGAGTGGGATGATCAATTTTGGAAGAATTATGGTGAGGCTAGAGAGCTATGGCATAGACAGTTTTTA   6640

         E   G   R   G   K   L   A   K   I   R   K   D   L   A   Q   T   A   L   A   Q   K   V   K   D   A   L   N
6641   GAGGGACGAGGTAAATTGGCAAAAATCCGCAAAGATTTAGCTCAAACTGCTCTAGCACAAAAAGTCAAGGATGCTCTCAA   6720

         S   I   P   K   W   L   W   T   G   L   A   V   I   I   V   L   I   I   I   I   Q   V   A   C   M   G
6721   TTCAATTCCTAAATGGTTGTGGACTGGCTTAGCAGTAATAATTGTACTCATTATCATAATTCAAGTTGCTTGCATGGGAC   6800

       L   K   M   L   T   K   A   I   V   S   W   A   G   F   L   A   L   Q   P   E   A   P   R   D   V   E   P
                             G   D   C   K   L   G   R   L   S   C   I   A   A   R   G   A   S   R   R   R   A
6801   TTAAAATGCTTACTAAGGCGATTGTAAGCTGGGCAGGCTTTCTTGCATTGCAGCCAGAGGCGCCTCGCGACGTAGAGCCC   6880

         D   A   A   I   Y   K   N   N   G   Y   A   N   G   W   R   D   Y   W   Q   L   W   S   G   G   T   L   F
       R   R   R   H   L   Q   K   Q   R   L   R   K   W   M   A   R   L   L   A   T   L   E   R   W   D   P   F
6881   GACGCCGCCATTTACAAAAACAACGGCTACGCAAATGGATGGCGCGACTACTGGCAACTCTGGAGCGGTGGGACCCTTTT   6960

         Q   M   T   R   R   P   *
         P   D   D   S   P   A   L   K   E   K   E   K   K   G   *   P   G   R   R   F   W   V   Q   L   S   T   L
6961   CCAGATGACTCGCCGGCCTTAAAAGAAAAAGAAAAGAAAGGGTGACCTGGAAGGAGATTTTGGGTGCAGCTGAGCACCCT   7040

         V   T   V   W   V   Q   L   S   T   Q   V   T   V   Y   Q   R   G   V   R   K   A   N   E   R   C   E
7041   AGTGACTGTGTGGGTGCAGCTGAGCACCCAGGTGACTGTGTACCAACGAGGGGTGCGAAAAGCCAACGAGAGGTGCGAAA   7120
```

Fig. S2. continued.

```
        ──▶ ORF?
        R  C  C  V  S  I  Q  E  T  T  T  A  D  C  L  S  H  L  Y  S  L  H  K  D  F  A  F
7121    GGTGCTGTGTGAGCATACAGGAAACCACAACCGCAGACTGTCTTTCACACCTTTACAGCTTACACAAGGACTTTGCTTTC    7200

                                                                                   ORF? ⊩─┐
         Y  L  G  K  G  G  Y  F  S  T  W  G  L  G  R  A  W  G  S  I  Y  K  P  E  V  A  *
7201    TATTTGGGGAAGGGGGGCTACTTCAGTACTTGGGGCTTGGGGAGGGCTTGGGGGAGCATATATAAGCCTGAGGTTGCCTA    7280

7281    ACCTCGAGGTCCCCTCACGCATCTCTGGTTCCGGCCATCACCCAGACTCAGAGTGTGGATCCACAATAAAGCTGTGCATC    7360

7314    TTGGACCCAGAGCCGTGTGGGTGTCTTCTTACC                                                   7393
```

Key

MA	matrix
CA	capsid
NC	nucleocapsid
PR	protease
RT	reverse transcriptase
IN	integrase
SU	surface domain
TM	transmembrane domain
LTR	long terminal repeat
PBS	primer binding site
PPT	polypurine tract
NLS	nuclear localisation signal
TAR	transactivation response element
RRE	rev-responsive element

Fig. S2. continued.

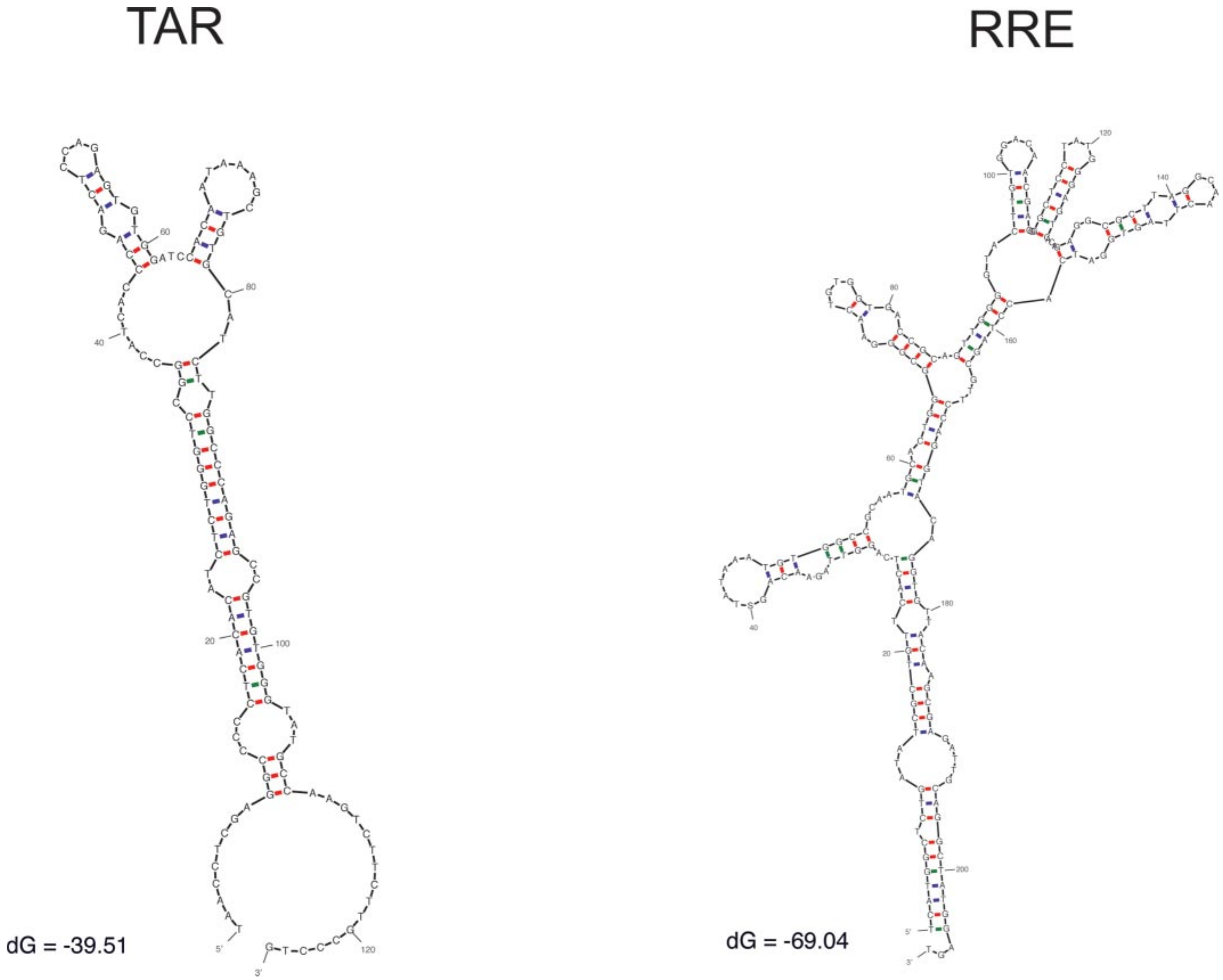


Fig. S3. Putative RNA secondary structure motifs in pSIVgml. Secondary structures were predicted using the MFOLD thermodynamic folding algorithm (4), and assessed by comparison to well-characterized examples in other lentiviruses; (*Left*) the putative TAR (transactivation responsive region) downstream of the viral promoter is a consistently predicted two-finger structure similar to the TAR found in HIV-2 (5); (*Right*) the putative RRE (Rev responsive element) contains a consistently predicted three-finger structure. The precise boundaries of the RRE are uncertain.

4. Zuker M, Mathews DH, Turner DH (1999) Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In *RNA Biochemistry and Biotechnology*, eds Barciszewski J, Clark BFC (Kluwer Academic Publishers, Dordrecht), pp 11–43.

5. Rabson AB, Graves BJ (1997) Retrovirus gene expression: Transcription and RNA processing. In *Retroviruses*, eds Coffin JM, Hughes SH, Varmus HE (CSHL Press, New York), p 226.

Table S1. Complete and low coverage genome sequence data screened

Species	Common name	Release files	Stage*	Geographic range
Complete genome assemblies				
Homo sapiens	human	NCBI build 36 version 2 (2006)	Complete	Worldwide
Pan troglodytes	chimpanzee	NCBI build 2 version 1 (2006)	Complete	Africa
Macaca mulatta	rhesus monkey	NCBI build 2 version 1 (2006)	Complete	Asia
Low coverage genome assemblies				
Microcebus murinus	grey mouse lemur	EMBL Broad Institute Release 1	30% coverage	Madagascar
Trace archive raw sequence files				
Aotus nancymaae	Ma's night monkey	001–002	613023	South America
Ateles geoffroyi	spider monkey	001	28627	South America
Callicebus moloch	dusky titi	001–002	624987	South America
Saimiri sciureus	squirrel monkey	001	1740	South America
Callithrix jacchus	common marmoset	001–058	28216241	South America
Eulemur macaco	black lemur	001	3967	Madagascar
Lemur catta	ring-tailed lemur	001	92444	Madagascar
Microcebus murinus	grey mouse lemur	001–017	8339325	Madagascar
Gorilla gorilla	gorilla	001–008	4119727	Africa
Papio anubis	olive baboon	001–003	1001661	Africa
Papio cynocephalus	yellow baboon	001–002	625576	Africa
Colobus guereza	guereza	001–002	630384	Africa
Cercopithecus aethiops	vervet monkey	001	179976	Africa
Otolemur garnettii	galago	001–018	8859530	Africa
Pongo pygmaeus	orang utan	001–025	12157107	South East Asia
Nomascus leucogenys	white-cheeked gibbon	001–009	4499952	South East Asia
Hylobates concolor	black gibbon	001	2282	South East Asia
Tarsius syrichta	tarsier	001–029	14116211	South East Asia

*For trace archive genome data, the number of sequence reads screened is shown. Sequences reads are typically ≈800 bp in length.