Electronic Supplementary Information

A targeted covalent small molecule inhibitor of HIV-1 fusion

Guangyan Zhou,[a] Li He,[a] Kathy H. Li, [b] Cássio C. S. Pedroso [c] and Miriam Gochin* [a,b]

[a.] *Department of Basic Sciences, Touro University California, 1310 Club Drive, Mare Island, Vallejo CA 94592*
[b.] *Department of Pharmaceutical Chemistry, UCSF School of Pharmacy, San Francisco, CA 9414*
[c.] *Lawrence Berkeley National Laboratory, The Molecular Foundry, 1 Cyclotron Road, 67R5114, Berkeley, CA 94720*
*corresponding author, e-mail: miriam.gochin@tu.edu

1. Experimental Methods

Synthesis and characterization of **2**

To a solution of **1** (51.4 mg, 0.1 mmol), prepared as previously described,[1] and sodium 2,3,5,6-tetrafluoro-4-hydroxy benzenesulfonate (26.8 mg, 0.1 mmol) in anhydrous DMF (2.0 mL) was added silicone gel - linked DCC (30.6 mg, 0.11 mmol). The mixture was stirred at room temperature overnight under N_2. TLC indicated that all starting material was consumed. The mixture was centrifuged at 14000 rpm for 2 minutes, the supernatant was collected and dried under high vacuum to remove DMF. Then 150 µL methanol was added to dissolve the product. The methanol solution was added dropwise with stirring to ether. Precipitants were formed, washed twice with ether, collected and dried. Afforded 15.2 mg product, yield: 33.8%. LC-MS data was obtained on an Agilent 1100/Bruker microTOF-Q LC-MS at the Molecular Foundry, Lawrence Berkeley National Laboratory.

Data for **2**: yellow powder; ^{1}H NMR (CD_3OD) δ 8.14 (s, 1H), 8.07 (d, J = 7.2 Hz, 1H), 7.90 (s, 1H), 7.86 (m, 1H), 7.58 (t, J = 8.2 Hz, 2H), 7.51- 7.45 (m, 4H), 7.38- 7.34 (m, 2H), 7.30 (t, J = 1.2 Hz, 1H), 7.28 (t, J = 1.2 Hz, 1H), 6.52 (d, J = 2.6 Hz, 1H), 6.49 (d, J = 2.6 Hz, 1H), 5.56 (s, 2H), 5.46 (s, 2H), 3.82 (s, 3H). MS calcd for $C_{39}H_{25}F_4N_2O_7SNa$: 764.12; found: 743.14 $(M+2H-Na)^+$. ^{19}F NMR(CD_3OD) δ -141.16 (q, 2H), -155.47 (q, 2H), referenced to external TFA (-76.55 ppm) (Figure S1)

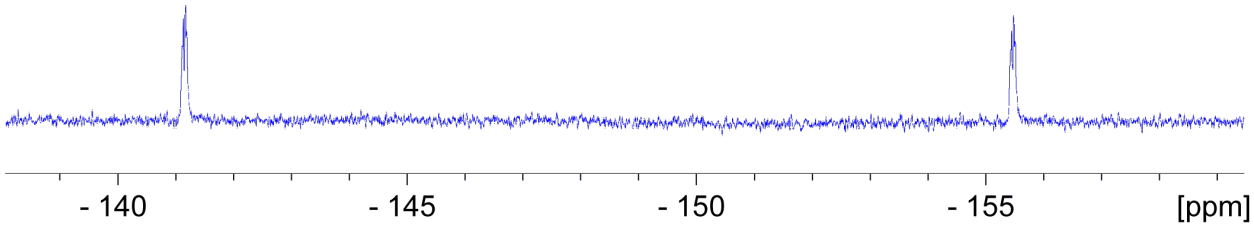

Figure S1. 19F NMR spectrum of **2**, referenced to external TFA at -76.55 ppm. Spectra were obtained on a Bruker Avance 400 MHz spectrometer using a QNP probe, and processed using Topspin.

Reaction of **2** with N-acetyl lysine.

2 (1.7 mM) was added to N-acetyl-lysine (13.3 mM) in sodium phosphate buffer, pH 8 and incubated at 37°C for 1 – 24 hours. Aliquots were withdrawn at several time points and tested by HPLC (Figure S2). Reduction of intensity of the peak of **2** at 14.3 min was accompanied by the appearance and growth of a new peak at 12.9 min, which was confirmed to be the lysine adduct by ESI-MS (MS calcd 684.29; found: 685.31 $(M+H)^+$). The reaction was also tested in PBS at pH 7.4 with similar results (50% conversion after 5 hours).

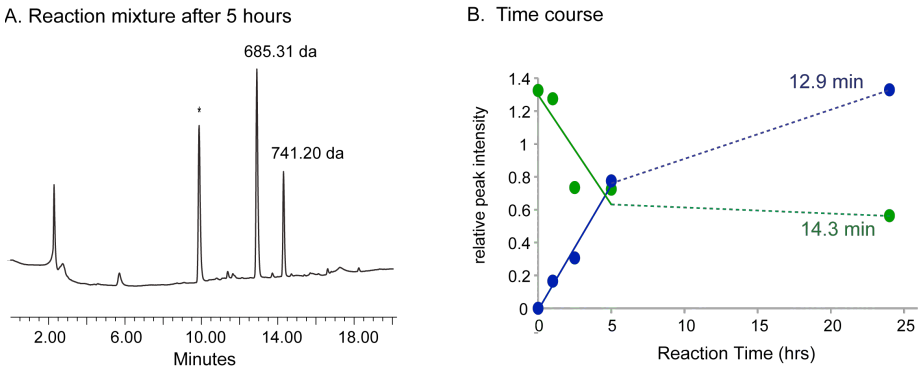

Figure S2. Reaction of **2** with N-acetyl lysine. A HPLC trace after 5 hours.. *A small non-reactive compound (6-bromo-2-methyl-indazole) appearing at 9.9 min was included as a reference. B. Intensities of peaks at 12.9 and 14.3 minutes relative to the reference peak as a function of reaction time.

Synthesis and characterization of gp41 protein construct and reaction with **2**

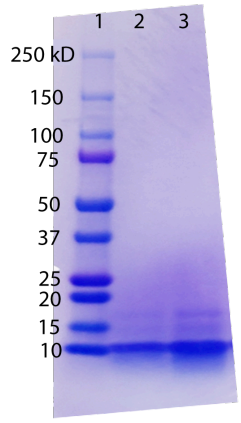

Figure S3. SDS Page gel of C26(L4)N50 (lane 2, 2µg, lane 3, 6µg). The molecular weight ladder is shown in Lane 1.

C26(L4)N50, a designed protein displaying the hydrophobic pocket of gp41 (see main text) was synthesized by overexpression in E. coli, and isolated, cleaved and purified into its final form according to previously described methods.[2] In brief, the protein was prepared as a longer construct and cleaved at the N-terminus by formic acid digest, which preferentially selected for a DP in the sequence, leaving proline at position 1. Details regarding the design of this construct will be given elsewhere. The molecular weight (9703 Da) was confirmed by Maldi MS, using an AB SCIEX TF4800 MALDI TOF-TOF Mass Spectrometer (Molecular Foundry, Lawrence Berkeley National Laboratory). SDS page gel (Figure S3) showed a small percentage of a minor impurity produced by cleavage at a DS site at the CHR – loop junction. This was confirmed by Maldi to be a 56 residue protein lacking the CHR residues, but still containing the HP binding site (molecular weight 6367 Da).

2. Computational Calculations and Sequence Alignment

Computational Docking

Flexible docking of **2** in the hydrophobic pocket of PDB structure *2xra* was performed using Schrodinger software (Schrodinger, LLC, New York, NY, 2017). Ligand three dimensional conformations were prepared using LigPrep and docked conformations were calculated using Glide, allowing side chains of residues Trp-571 and Gln-575 to move during the docking. 10 conformers were generated, of which 8 showed an ionic or hydrogen bond interaction with the side chain amino group of Lys 574. Results were viewed using Maestro.

Calculation of the 3D structure of **2**.

The near symmetry of **1**, evident from the NMR spectrum (Figure 2), was disrupted by STP esterification, which was accompanied by downfield shifts of the resonances of one of the two indole groups. Conformational space sampled by **2** was examined by generating 10 low energy conformers using Omega2-2.5.1.4 (OpenEye Scientific Software, Santa Fe, NM. http://www.eyesopen.com).[3, 4] Seven out of the ten low energy conformers were consistent with the observed NMR spectrum in which the sulfotetrafluorophenyl group folds back over the indole groups. Figure S4 shows the lowest energy conformer. Downfield shifts are consistent with an edge on association of the fluorine containing ring with indole C. With the exception of the indole H7' protons, calculated distances to indole C protons were in the range of 6.3 – 7.4 Å, while distances to indole B protons were longer, 7.5 – 9.2 Å. This suggests that most of the downfield shifted resonances belong to indole C. The exceptions are the indole H7' protons with the closest fluorine atom at 5 Å for B-H7' and 5.6 Å for C-H7'. The identity of the downfield shifted methylene protons is less clear, with the methylene between A and B somewhat closer to the fluorine ring than that between C and D (6.7 vs. 7 Å).

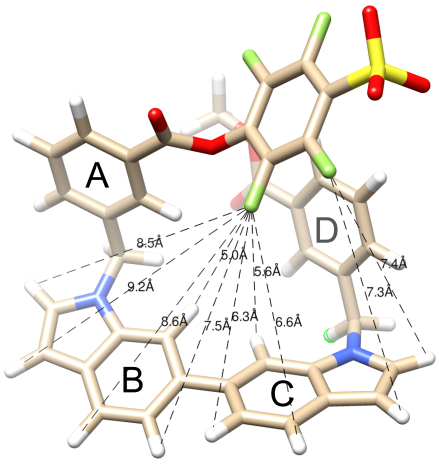

Figure S4. Computed 3D structure of **2**, showing distances from fluorine atoms on one edge of the STP ring to indole protons.

Alignment of gp41 across known clades

Below we show the alignment of gp41 sequences from an HIV sequence database containing approximately four representatives of each group: M (subtypes A-K), N, O and P as well as 15 circulating recombinant forms.[5] Residue positions in the hydrophobic pocket (residues 54 – 69 in the alignment) are either completely or almost completely invariant or involve a conservative substitution.

```
Ref.A1.AU.03.PS1044_Day0.DQ676872          1    AV-G--IGAVFI-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLKLTVWGIKQLQARVLA  71
Ref.A1.RW.92.92RW008.AB253421             1    AV-G--IGAVFI-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.A1.UG.92.92UG037.AB253429             1    AV-T--LGAVFI-G-FLGTAGSTMGAASITLTVQA-RK-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.A2.CD.97.97CDKTB48.AF286238           1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-TGIVQQQ-SNLLKAIEAQQQMLRLTVWGIKQLQARVLA  71
Ref.A2.CM.01.01CM_1445MV.GU201516         1    AV-G--LGAAFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQQLLKLTVWGIKQLQARVLA  71
Ref.A2.CY.94.94CY017_41.AF286237          1    AV-G--LGAVFL-G-FLGAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-SNLLQAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.B.FR.83.HXB2_LAI_IIIB_BRU.K03455      1    AV-G--IGALFL-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.B.NL.00.671_00T36.AY423387            1    ------LGAMFL-G-FLGAAGSTMGAASMALTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  68
Ref.B.TH.90.BK132.AY173951                1    AV-G--IGAVFL-G-FLGAAGSTMGAASVTLTVQA-RL-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.B.US.98.1058_11.AY331295              1    AV-T--LGALFL-G-FLGAAGSTMGAASMTLTVQA-RL-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.C.BR.92.BR025_d.U52953                1    AV-G--IGAVFL-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQTRVLA  71
Ref.C.ET.86.ETH2220.U46016                1    ---#--LGALFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHMLQLTVWGIKQLQTRVLA  69
Ref.C.IN.95.95IN21068.AF067155            1    AV-G--IGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQTRVLA  71
Ref.C.ZA.04.04ZASK146.AY772699            1    AV-G--LGAVLL-G-FLGTAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.D.CD.83.ELI.K03454                    1    AI-G--LGAMFL-G-FLGAAGSTMGARSVTLTVQA-RQ-LM-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.D.CM.01.01CM_4412HAL.AY371157         1    AI-G--LGAMFL-G-FLGAAGSTMGAASVTLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.D.TZ.01.A280.AY253311                 1    AI-G--LGAMFL-G-FLGAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.D.UG.94.94UG114.U88824                1    AI-G--LGALFL-G-FLGTAGSTMGAVSLTLTVQA-RQ-VL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.F1.BE.93.VI850.AF077336               1    AA-G--LGALFLGFLGDSREH--MGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.F1.BR.93.93BR020_1.AF005494           1    AV-G--LGALFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.F1.FI.93.FIN9363.AF075703             1    AV-A--IGAVFL-G-FLSAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-NNLLQAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.F1.FR.96.96FR_MP411.AJ249238          1    AV-G--IGALFL-R-FLGAAGSNIGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.F2.CM.02.02CM_0016BBY.AY371158        1    AV-G--IGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLKAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.F2.CM.95.95CM_MP255.AJ249236          1    AV-G--MGAVLF-G-FLGAAGSTMGAAAITLTAQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.F2.CM.95.95CM_MP257.AJ249237          1    AV-G--MGAMFL-G-FLGAAGSTMGAASITLTVQA-RN-LL-SGIVQQQ-SNLLKAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.F2.CM.97.CM53657.AF377956             1    AV-G--IGAVLL-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLKAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.G.BE.96.DRCBL.AF084936                1    AV-G--VGAIFL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLRARVLA  71
Ref.G.KE.93.HH8793_12_1.AF061641          1    AV-G--VGAVFL-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.G.NG.92.92NG083.U88826                1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTAQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQSRVLA  71
Ref.G.PT.x.PT2695.AY612637                1    AV-G--LGAVLL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.H.BE.93.VI991.AF190127                1    AV-G--MGAFFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIQAQQHMLQLTVWGIKQLQARVLA  71
Ref.H.BE.93.VI997.AF190128                1    -V-G--MGAFFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIQAQQHMLQLTVWGVKQLQARVLA  70
Ref.H.CF.90.056.AF005496                  1    AV-G--MGASFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIQARQHMLQLTVWGIKQLQARVLA  71
Ref.H.GB.00.00GBAC4001.FJ711703           1    AV-G--MGAFFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIQAQQHMLQLTVWGVKQLQARVLA  71
Ref.H.GB.00.00GBAC4001.FJ711703           1    .....................................................................................  71
Ref.H.GB.00.00GBAC4001.FJ711703           1    .....................................................................................  71
Ref.H.GB.00.00GBAC4001.FJ711703           1    .....................................................................................  71
Ref.H.GB.00.00GBAC4001.FJ711703           1    .....................................................................................  71
Ref.J.CD.97.J_97DC_KTB147.EF614151        1    #V-G--MGAVFL-G-FLGTAGSTMGAASIALTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLRLTVWGIKQLQARILA  71
Ref.J.CM.04.04CMU11421.GU237072           1    AL-G--IGAVFL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.J.CM.04.04CMU11421.GU237072           1    .....................................................................................  71
Ref.J.CM.04.04CMU11421.GU237072           1    .....................................................................................  71
Ref.J.SE.93.SE9280_7887.AF082394          1    AV-G--IGAVFL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.K.CD.97.97ZR_EQTB11.AJ249235          1    AV-G--IGALFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQQMLQLTVWGIKQLRARVLA  71
Ref.K.CM.96.96CM_MP535.AJ249239           1    AV-G--LGAVFF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLRARILA  71
Ref.01_AE.AF.07.569M.GQ477441             1    AV-G--IGAMIF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.01_AE.CN.05.05GX001.GU564221          1    AV-G--IGAMIF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.01_AE.TH.90.CM240.U54771              1    AV-G--IGAMIF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.02_AG.CM.99.pBD6_15.AY271690          1    AV-G--LGAFFL-G-FLGVAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHILRLTVWGIKQLQARVLA  71
Ref.02_AG.LR.x.POC44951.AB485636          1    AV-G--LGAVFI-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLRLTVWGIKQLQARVLA  71
Ref.02_AG.NG.x.IBNG.L39106                1    -V-G--LGAVFL-G-FLGAAGSTMGARSITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLKLTVWGIKQLQARVLA  70
Ref.03_AB.RU.97.KAL153_2.AF193276         1    AV-G--IGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.04_cpx.CY.94.94CY032_3.AF049337       1    AV-G--IGAMFL-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLRLTVWGIKQLQARVLA  71
Ref.04_cpx.GR.91.GR11_97PVCH.AF119820     1    AV-G--LGALFL-G-FLGAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.04_cpx.GR.97.GR84_97PVMY.AF119819     1    AV-G--IGAVFP-GLFLGAAGSTMGAASITLTVQA-RQ-LL-FGIVQQQ-SNLLRAIEAQQQLLRLTVWGVKQLQARVLA  72
Ref.05_DF.BE.93.VI961.AF076998            1    AV-G--VGALLI-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLQAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.05_DF.BE.x.VI1310.AF193253            1    AM-G--IGAMFL-G-FLAAAGSTMGAASIALTVQA-RQ-LL-SGIVQQQ-NNLLQAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.05_DF.ES.99.X492.AY227107             1    AV-G--LGALFL-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLQAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.06_cpx.AU.96.BFP90.AF064699           1    AV-G--LGAMFL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.06_cpx.EE.01.EE0359.AY535659          1    AV-G--LGAVFL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.06_cpx.GH.03.03GH173_06.AB286851      1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.07_BC.CN.05.XJDC6431_2.EF368372       1    AA-G--LGALLF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQTRVLA  71
Ref.07_BC.CN.05.XJDC6441.EF368370         1    AV-G--IGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLQLTVWGIKQLQTRVLA  71
Ref.07_BC.CN.98.98CN009.AF286230          1    AV-G--IGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQTRVLA  71
Ref.08_BC.CN.06.nx2.HM067748              1    VV-GA#VGAMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQTRVLA  73
Ref.08_BC.CN.97.97CNGX_6F.AY008715        1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQTRVLA  71
Ref.09_cpx.CI.00.00IC_10092.AJ866553      1    AV-G--MGALFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHMLRLTVWGIKQLQARVLA  71
Ref.09_cpx.GH.96.96GH2911.AY093605        1    AV-G--MGALFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.09_cpx.SN.95.95SN1795.AY093603        1    AV-G--MGALFL-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.09_cpx.US.99.99DE4057.AY093607        1    AV-G--MGALFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.10_CD.TZ.96.96TZ_BF061.AF289548       1    AI-G--LGAVFL-G-FLGAAGSTMGAASLTLTVQA-RQ-VL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.10_CD.TZ.96.96TZ_BF071.AF289549       1    AI-G--LGAMFL-G-FLGAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-NNLLKAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.10_CD.TZ.96.96TZ_BF110.AF289550       1    AI-G--LGAVFL-G-FLGAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.11_cpx.CM.95.95CM_1816.AF492624       1    AV-G--IGAVLL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.11_cpx.CM.95.95CM_1816.AF492624       1    .....................................................................................  71
```

```
Ref.11_cpx.CM.95.95CM_1816.AF492624          1    ..........................................................................  71
Ref.11_cpx.CM.96.96CM_4496.AF492623          1    AV-G--IGAVFF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.11_cpx.CM.96.96CM_4496.AF492623          1    ..........................................................................  71
Ref.11_cpx.CM.96.96CM_4496.AF492623          1    ..........................................................................  71
Ref.11_cpx.CM.97.MP818.AJ291718              1    AV-G--IGAVLF-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.12_BF.AR.97.A32879.AF408629              1    AV-G--MGALFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.12_BF.AR.97.A32989.AF408630              1    AV-G--MGALLI-G-FLGATGSTMGAASITLTVQA-KQ-LL-SGIVQQQ-SNLLKAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.12_BF.AR.99.ARMA159.AF385936             1    AV-G--IGALFL-G-FLGAAGSTMGAASVTLTVQA-RQ-LL-SGIVQQQ-NNLLQAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.13_cpx.CM.02.02CM_A1394.DQ845388         1    ---#--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLMAIQAQQHMLKLTVWGIKQLQARVLA  69
Ref.13_cpx.CM.04.04CM_632_28.DQ845387        1    AV-G--IGAVFL-G-FLGAAGSTMGAASLTLTVQA-RN-LL-SGIVQQQ-NNLLRAIEAQQQLLKLTVWGIKQLQARVLA  71
Ref.13_cpx.CM.96.96CM_1849.AF460972          1    #V-G--LGAVFL-G-FLGAAGSTMGVASLTLTVQA-RQ-LL-SGIVQQQ-NNLLMAIEAQQHLLKLTVWGIKQLQARILA  71
Ref.14_BG.ES.00.X605.AF450096                1    AV-G--IGALLF-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHMLQLTVWGIKQLQARILA  71
Ref.14_BG.ES.00.X605.AF450096                1    ..........................................................................  71
Ref.14_BG.ES.00.X605.AF450096                1    ..........................................................................  71
Ref.14_BG.ES.00.X623.AF450097                1    AV-G--IGALLI-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHMLQLTVWGIKQLQARILA  71
Ref.14_BG.ES.00.X623.AF450097                1    ..........................................................................  71
Ref.14_BG.ES.00.X623.AF450097                1    ..........................................................................  71
Ref.14_BG.PT.00.00PTHDE10.GU230137           1    AV-G--IGALLF-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.15_01B.TH.96.M169.DQ354120               1    AV-GT-IGAMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLKAIEAQQHLLQLTVWGIKQLQARVLA  72
Ref.15_01B.TH.99.99TH_MU2079.AF516184        1    AV-GT-IGAMFL-G-FLGTAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  72
Ref.15_01B.TH.99.99TH_R2399.AF530576         1    AV-GA-IGVMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHMLQLTVWGIKQLQARVLA  72
Ref.16_A2D.KE.91.KNH1271.AY945736            1    AV-G--LGAFFL-G-FLGAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLRLTVWGIKQLQARVLA  71
Ref.16_A2D.KR.97.97KR004.AF286239            1    ---------#FL-G-FLGAAGSTMGAASVTLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHMLRLTVWGIKQLQARVLA  65
Ref.17_BF.AR.02.AR02_ARG1139.EU581825        1    AV-G--LGAMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.17_BF.BO.02.BO02_BOL119.EU581827         1    --------#MFL-G-FLGAAGSTMGAASIALTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQARVLA  66
Ref.17_BF.BO.02.BO02_BOL119.EU581827         1    ..........................................................................  66
Ref.17_BF.BO.02.BO02_BOL119.EU581827         1    ..........................................................................  66
Ref.17_BF.PE.02.PE02_PCR0155.EU581828        1    AV-G--MAAMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.18_cpx.CM.97.CM53379.AF377959            1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.18_cpx.CU.99.CU14.AY586541               1    AV-G--REAVFP-G-CLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.18_cpx.CU.99.CU68.AY894993               1    -#-G--LGAFFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-AGIVQQQ-SNLLRAIEAQQHLLKLTVWG#KQLQARVLA  70
Ref.18_cpx.CU.99.CU68.AY894993               1    ..........................................................................  70
Ref.18_cpx.CU.99.CU68.AY894993               1    ..........................................................................  70
Ref.18_cpx.CU.99.CU68.AY894993               1    ..........................................................................  70
Ref.18_cpx.CU.99.CU68.AY894993               1    ..........................................................................  70
Ref.18_cpx.CU.99.CU68.AY894993               1    ..........................................................................  70
Ref.18_cpx.CU.99.CU68.AY894993               1    ..........................................................................  70
Ref.18_cpx.CU.99.CU68.AY894993               1    ..........................................................................  70
Ref.18_cpx.CU.99.CU68.AY894993               1    ..........................................................................  70
Ref.19_cpx.CU.99.CU29.AY588971               1    AV-G--MGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLKLTVWGIKQLQARVLA  71
Ref.19_cpx.CU.99.CU38.AY588970               1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLKLTVWGIKQLQARVLA  71
Ref.19_cpx.CU.99.CU7.AY894994                1    AM-G--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQQLLRLTVWGIKQLQARVLA  71
Ref.19_cpx.CU.99.CU7.AY894994                1    ..........................................................................  71
Ref.19_cpx.CU.99.CU7.AY894994                1    ..........................................................................  71
Ref.20_BG.CU.99.Cu103.AY586545               1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.21_A2D.KE.91.KNH1254.AY945737            1    AI-G--LGAMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.21_A2D.KE.99.KER2003.AF457051            1    AI-G--LGAMFL-G-FLGAAGSTMGAASVTLTVQA-RQ-LL-SGIVQQQ-NNLLMAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.21_A2D.KE.99.KSM4001.AF457072            1    AI-G--LGAMFL-G-FLGAAGSTMGAASVTLAVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.22_01A1.CM.01.01CM_0001BBY.AY371159      1    AL-G--FGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLKLTVWGIKQLQARVLA  71
Ref.22_01A1.CM.02.02CM_3097MN.GQ229529        1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQXMLRLTVWGIKQLQARVLA  71
Ref.22_01A1.CM.02.02CM_3097MN.GQ229529        1    ..........................................................N.......  71
Ref.22_01A1.CM.02.02CM_3097MN.GQ229529        1    ..........................................................H.......  71
Ref.23_BG.CU.03.CB118.AY900571               1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.23_BG.CU.03.CB347.AY900572               1    AV-G--IGAVLF-G-FLGAAGSTMGAASMALTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.24_BG.CU.03.CB378.AY900574               1    AV-G--LGAVFL-G-FLGTAGSTMGAASLTLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.24_BG.CU.03.CB378.AY900574               1    ..........................................................................  71
Ref.24_BG.CU.03.CB378.AY900574               1    ..........................................................................  71
Ref.24_BG.CU.03.CB471.AY900575               1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-GNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.24_BG.ES.08.X2456_2.FJ670526             1    AV-G--LGAVFL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.25_cpx.CM.06.06CM_BA_040.EU693240        1    AV-G--LGAVFL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLKARLLA  71
Ref.25_cpx.CM.06.06CM_BA_040.EU693240        1    ..........................................................................  71
Ref.25_cpx.CM.06.06CM_BA_040.EU693240        1    ..........................................................................  71
Ref.25_cpx.SA.03.J11233.EU697906             1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-NNLLRAIEAXQHLLQLTVWGIKQLQARVLA  71
Ref.25_cpx.SA.03.J11233.EU697906             1    ..........................................................Q.......  71
Ref.25_cpx.SA.03.J11233.EU697906             1    .........................................................R........  71
Ref.25_cpx.SA.03.J11451.EU697908             1    AV-G--LGAVFL-G-XLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.25_cpx.SA.03.J11451.EU697908             1    ..............F...........................................................  71
Ref.25_cpx.SA.03.J11451.EU697908             1    ..............L...........................................................  71
Ref.26_AU.CD.02.02CD_KS069.FM877780          1    AV-G--LGALFL-G-FLGTAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHVLKLTVWGIKQLQARVLA  71
Ref.26_AU.CD.02.02CD_MBTB047.FM877782        1    AV-G--MGALFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLKLTVWGIKQLQARVLA  71
Ref.26_AU.CD.97.97CD_KTB119.FM877777         1    AV-G--IGAMFL-G-FLGTAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.27_cpx.CD.97.97CDKTB49.AJ404325          1    AV-G--IGAMIF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.27_cpx.FR.04.04CD_FR_KZS.AM851091        1    AV-G--IGALIF-G-FLGAAGSTMGAASITLTVHA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.28_BF.BR.99.BREPM12313.DQ085872          1    AV-GT-MGAMFL-G-FLGAAGSTMGAASVTLTVQA-RL-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  72
Ref.28_BF.BR.99.BREPM12609.DQ085873          1    AV-GA-LGALFL-G-FLGAAGSTMGAASMTLTVQA-RQ-LM-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARILA  72
Ref.28_BF.BR.99.BREPM12817.DQ085874          1    AV-GT-IGAMFL-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  72
Ref.29_BF.BR.01.BREPM16704.DQ085876          1    AV-G--IGAMFL-G-FLGAAGSTMGAASVALTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.29_BF.BR.02.BREPM119.AY771590            1    AV-GT-LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHMLQLTVWGIKQLQARVLA  72
Ref.29_BF.BR.99.BREPM11948.DQ085871          1    AL-G--IGAMFL-G-FLGAAGSTMGAASVALTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.31_BC.BR.02.110PA.EF091932               1    ---#--LGAMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQTRVLA  69
Ref.31_BC.BR.02.110PA.EF091932               1    ..........................................................................  69
Ref.31_BC.BR.02.110PA.EF091932               1    ..........................................................................  69
Ref.31_BC.BR.04.04BR137.AY727526             1    AV-G--IGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQTRVLA  71
Ref.31_BC.BR.04.04BR142.AY727527             1    AV-G--LGALFF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQTRVLA  71
Ref.32_06A1.EE.01.EE0369.AY535660            1    AV-G--LGAVFL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.33_01B.ID.07.JKT194_C.AB547464           1    AV-G--IGAMIF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.33_01B.MY.05.05MYKL007_1.DQ366659        1    AV-G--IGVMIF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.33_01B.MY.05.05MYKL045_1.DQ366662        1    AV-G--IGALIF-G-FLGAAGSTMGAASITLTVQT-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.34_01B.TH.99.OUR2478P.EF165541           1    AV-G--IGAMIF-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.35_AD.AF.05.05AF026.EF158043             1    AV-G--LGAVFI-G-FLGAAGSTMGAASVTLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHVLRLTVWGIKQLQARVLA  71
Ref.35_AD.AF.05.05AF094.EF158040             1    AV-G--IGAVFI-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLKLTVWGIKQLQARVLA  71
Ref.35_AD.AF.05.05AF095.EF158041             1    AV-G--LGAMFI-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLKLTVWGIKQLQARVLA  71
Ref.36_cpx.CM.00.00CMNYU1162.EF087995        1    AVVG--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  72
Ref.36_cpx.CM.00.00CMNYU830.EF087994         1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.37_cpx.CM.00.00CMNYU926.EF116594         1    AV-G--LGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.37_cpx.CM.97.CM53392.AF377957            1    AV-G--LGAVFL-G-FLGAAGSTMGAASVTLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.38_BF1.UY.04.UY04_3987.FJ213781          1    AV-G--MAAVFI-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
```

```
Ref.38_BF1.UY.04.UY04_4022.FJ213782          1    AV-G--MAAVFL-G-FLGAAGSTMGAASITLTVQT-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.38_BF1.UY.05.UY05_4752.FJ213780          1    AV-G--MAAMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.39_BF.BR.03.03BRRJ103.EU735534           1    AV-GL-LGAMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARVLA  72
Ref.39_BF.BR.03.03BRRJ327.EU735536           1    AV-G--MGAMFI-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-SNLLKAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.39_BF.BR.04.04BRRJ179.EU735535           1    ---#--LGVVFL-G-FLAAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-NNLLKAIDAQQHLLQLTVWGIKQLQARLLA  69
Ref.40_BF.BR.04.04BRRJ115.EU735538           1    AV-T--MGAMFL-G-FLGAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-NNLLKAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.40_BF.BR.04.04BRSQ46.EU735540            1    AV-T--MGAMFL-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLKAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.40_BF.BR.05.05BRRJ200.EU735539           1    ------LGAMFL-G-FLGAAGSTMGAASMTLTVQA-RQ-LL-SGIVQQQ-NNLLKAIEAQQHLLQLTVWGIKQLQARVLA  68
Ref.42_BF.LU.03.luBF_05_03.EU170155          1    --------------------GSTMGAASVTLTVQA-RQ-LL-SGIVQQQ-NNLLKAIEAQQHLLQLTVWGIKQLQARVLA  56
Ref.42_BF.LU.03.luBF_05_03.EU170155          1    ..........................................................................  56
Ref.42_BF.LU.03.luBF_05_03.EU170155          1    ..........................................................................  56
Ref.43_02G.SA.03.J11223.EU697904             1    AV-G--MGAVFL-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.43_02G.SA.03.J11223.EU697904             1    ..........................................................................  71
Ref.43_02G.SA.03.J11223.EU697904             1    ..........................................................................  71
Ref.43_02G.SA.03.J11243.EU697907             1    AV-G--MGAVFL-G-FLGAAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.43_02G.SA.03.J11243.EU697907             1    ..........................................................................  71
Ref.43_02G.SA.03.J11243.EU697907             1    ..........................................................................  71
Ref.43_02G.SA.03.J11243.EU697907             1    ..........................................................................  71
Ref.43_02G.SA.03.J11243.EU697907             1    ..........................................................................  71
Ref.43_02G.SA.03.J11456.EU697909             1    AV-G--LGAFFL-G-FLGTAGSTMGAASITLTVQV-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLQLTVWGIKQLQARVLA  71
Ref.43_02G.SA.03.J11456.EU697909             1    ..........................................................................  71
Ref.43_02G.SA.03.J11456.EU697909             1    ..........................................................................  71
Ref.44_BF.CL.00.CH80.FJ358521                1    AV-T--IGAMFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVQQQ-RNLLMAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.45_cpx.CD.97.97CD_MBFE185.FN392874       1    AV-G--LGAFFL-G-FLGAAGSTMGAAAVTLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.45_cpx.CM.97.97CM_MP814.FN392876         1    AV-G--IGAVFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-SGIVRQQ-NNLLKAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.45_cpx.GA.97.97GA_TB45.FN392877          1    AV-G--LGAVFF-G-FLGAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-SNLLRAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.45_cpx.GA.97.97GA_TB45.FN392877          1    ..........................................................................  71
Ref.45_cpx.GA.97.97GA_TB45.FN392877          1    ..........................................................................  71
Ref.46_BF.BR.01.01BR087.DQ358801             1    AA-G--IAVMLF-G-FLGAAGSTMGAASMALTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.46_BF.BR.01.01BR125.DQ358802             1    AA-G--ITALFL-G-FLGAAGSTMGAASMTLTVQDQTNYWTWPYCS-PREQSAERTIEAQQHLLQLTVWGIKQLQARLLA  74
Ref.46_BF.BR.07.07BR_FPS625.HM026456         1    AV-G--LGALFL-G-FLGAAGSTMGAASITLTVQA-RQ-LL-AGIVQQQ-NNLLRAIEAQQHLLQLTVWGIKQLQARILA  71
Ref.47_BF.ES.08.P1942.GQ372987               1    AV-G--FGALFL-G-FLGAAGSTMGAASVALTVQA-RL-LL-SGIVQQQ-NNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.47_BF.ES.08.X2457_2.FJ670529             1    AV-G--FGALFL-G-FLGAAGSTMGAASMALTGQA-RL-LL-SGIVQQQ-NNLLRAIEAQQHMLQLTVWGIKQLQARVLA  71
Ref.49_cpx.GM.02.N18380.HQ385477             1    AV-G--IGAVFL-G-FLGTARSTMGAASITLTVQA-KQ-LL-SGIVQQQ-SNLLKAIEAQQHLLKLTVWGIKQLQARVLA  71
Ref.49_cpx.GM.03.N26677.HQ385479             1    AV-G--MGAVFL-G-FLGAAGSTMGAASLTLTVQA-RQ-LL-SGIVQQQ-SNLLQAIEAQQHMLKLTVWGIKQLQARVLA  71
Ref.49_cpx.GM.97.N28353.HQ385478             1    AV-G--IGAVFL-G-FLGAAGSTMGAASITLTVHA-RQ-LL-SGILQQQ-SNLLKAIEAQQHLLKLTVWGIKHLQARVLA  71
Ref.O.BE.87.ANT70.L20587                     1    AV-G--LGMLFL-G-VLSAAGSTMGAAATTLAVQT-HT-LL-KGIVQQQ-DNLLRAIQAQQQLLRLSXWGIRQLRARLLA  71
Ref.O.BE.87.ANT70.L20587                     1    ..........................................................................I.....  71
Ref.O.BE.87.ANT70.L20587                     1    ..........................................................................V.....  71
Ref.O.CM.91.MVP5180.L20571                   1    AV-G--LGMLFL-G-VLSAAGSTMGAAATALTVRT-HS-VL-KGIVQQQ-DNLLRAIQAQQHLLRLSVWGIRQLRARLQA  71
Ref.O.CM.98.98CMU2901.AY169812               1    AV-G--LGMLFL-G-VLSAAGSTMGAAATALTVQT-HS-LI-KGIVQQQ-DNLLRAIQAQQELLRLSVWGIRQLRARLLA  71
Ref.O.CM.98.98CMU2901.AY169812               1    ..........................................................................  71
Ref.O.CM.98.98CMU2901.AY169812               1    ..........................................................................  71
Ref.O.SN.99.99SE_MP1300.AJ302647             1    AV-G--LGMLFL-G-VLSAAGSTMGAAATTLAVQT-HT-LM-KGIVQQQ-DNLLRAIQAQQQLLRLSVWGIRQLRARLLA  71
Ref.N.CM.02.DJO0131.AY532635                 1    -#-G--LGALFL-G-FLGAAGSTMGAASITLTVQA-RT-LL-SGIVQQQ-NNLVRAIEAQQHLLQLSIWGIKQLRAKVLA  70
Ref.N.CM.95.YBF30.AJ006022                   1    -#-G--LGALFL-G-FLGAAGSTMGAASITLTVQA-RT-LL-SGIVQQQ-NILLRAIEAQQHLLQLSIWGIKQLQAKVLA  70
Ref.N.CM.97.YBF106.AJ271370                  1    -#-G--LGALFL-G-FLGAAGSTMGAASITLTVQA-RT-LL-SGIVQQQ-NNLLRAIEAQQHLLQLSIWGIKQLRAKVLA  70
Ref.N.CM.97.YBF106.AJ271370                  1    ..........................................................................  70
Ref.N.CM.97.YBF106.AJ271370                  1    ..........................................................................  70
Ref.P.CM.06.U14788.HQ179987                  1    AA-G--LGMLFL-G-FLSAAGSTMGAAGITLTVQA-KQ-LL-HGIVQQQ-NNMLRAIEAQQELLRLSVWGIRQLRARLLA  71
Ref.P.CM.06.U14788.HQ179987                  1    ..........................................................................  71
Ref.P.CM.06.U14788.HQ179987                  1    ..........................................................................  71
Ref.P.FR.09.RBF168.GU111555                  1    #A-G--LGMLFL-G-FMSAAGSTMGAAAVTLTVQA-RQ-LL-HGIVQQQ-NNMLRAIEAQQELLRLSVWGIRQLRARLLA  71
Ref.P.FR.09.RBF168.GU111555                  1    ..........................................................................  71
Ref.P.FR.09.RBF168.GU111555                  1    ..........................................................................  71
Ref.CPZ.CD.90.ANT.U42720                     1    -#-G--IGLFFL-G-LLSAAGSTMGAASIALTAQT-RN-LX-HGIVQQQ-ANLLQAIETQQHLLQLSVWGVKQLQARMLA  70
Ref.CPZ.CD.90.ANT.U42720                     1    .....................................F.....................................  70
Ref.CPZ.CD.90.ANT.U42720                     1    .....................................L.....................................  70
Ref.CPZ.CM.05.SIVcpzMT145.DQ373066           1    -#-G--LGALFL-G-FLGAAGSTMGAASVVLTVQA-RQ-LL-TGIVQQQ-NNLLRAIEAQQHLLQLSVWGIKQLQARVLA  70
Ref.CPZ.US.85.US_Marilyn.AF103818            1    ---G--LGALFL-G-FLGAAGSTMGAASVVLTVQA-RQ-LL-SGIVQQQ-NNLLRAIEAQQHLLQLSVWGIKQLQARVLA  69
```

pKa predictions

The online server Rosie[6] was used to predict the pKa of all ionizable residues in C26(L4)N50, using an unpublished crystal structure (to be published) and the neighbor repack option. There are two chains in the asymmetric unit, represented as Chains A and B. Results are listed below. In this construct, K64 corresponds to K574 in gp41.

RESIDUE	IpKa	pKa	RESIDUE	IpKa	pKa
GLU 6 A	4.4	4.1	GLU 6 B	4.4	3.9
GLU 7 A	4.4	3.6	GLU 7 B	4.4	3.1
GLU 13 A	4.4	4.8	GLU 13 B	4.4	4.4
GLU 14 A	4.4	4	GLU 14 B	4.4	3.3
GLU 20 A	4.4	3.9	GLU 20 B	4.4	3.4
GLU 21 A	4.4	3.9	GLU 21 B	4.4	3.8
GLU 50 A	4.4	4.4	GLU 50 B	4.4	5.8
HIS 54 A	6.3	6.3	HIS 54 B	6.3	6.3
LYS 64 A	10.4	10.3	LYS 64 B	10.4	9.8
GLU 74 A	4.4	4.5	GLU 74 B	4.4	3.2
TYR 76 A	10	8.9	TYR 76 B	10	9.8
LYS 78 A	10.4	10.4	LYS 78 B	10.4	10.4
ASP 79 A	4	4	ASP 79 B	4	4

3. LC- MS/MS Analysis

Protein Site Identification Using Liquid Chromatography with Tandem Mass Spectrometry (LC-MS/MS)

To form the covalent adduct, 0.63 mM **2** was incubated with 315 µM C26(L4)N50 in 25mM phosphate buffer, pH 8.at 37°C for 3 hours. The protein was exchanged into formate buffer (pH 2.91) with several rounds of centrifugation using a spin filter (Amicon Ultra – 0.5ml, 3K cutoff, Sigma). 6 ug was then denatured with 8M urea/100 mM ammonium bicarbonate (pH 8), reduced with 100 mM dithiothreitol at 60°C for 30 minutes, followed by alkylation with 100 mM iodoacetamide at room temperature in the dark for 1 hour. The sample was then incubated 4 hours with trypsin (1:20 weight/weight) at 37°C. The peptides formed from the digestion were further purified by C18 ZipTips (Millipore) and analyzed by online LC-MS/MS. The LC separation was carried out on a NanoAcquity UPLC system (Waters) with a linear gradient from 2 - 25% B (0.1% formic acid in acetonitrile) over 48 mins followed by 25 - 37% B over 6 mins and then 37 – 40%

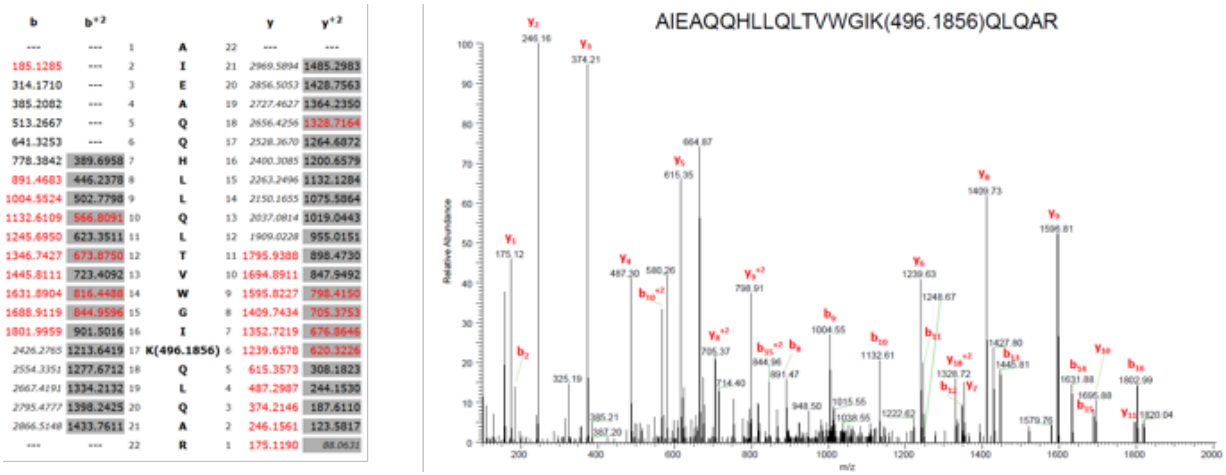

Figure S5. Theoretical sequence ions predicted from the peptide AIEAQQHLLQLTVWGIK(496.1856)QLQAR are listed on the left. The main sequence ions actually detected by Q Exactive Plus Orbitrap mass spectrometer are highlighted in red. The mass spectrum obtained from the peptide AIEAQQHLLQLTVWGIK(496.1856)QLQAR is shown on the right, confirming the formation of the linkage C33H24N2O3 (m/z 496.1856). The actual mass position of some ions in the spectrum was found to be slightly different from their theoretical m/z.

B over 3 mins at a flow rate of 400 nl/min. The MS/MS analysis was performed using Q Exactive Plus Orbitrap mass spectrometer (Thermo). After a survey scan, 10 most intense precursor ions were selected for subsequent fragmentation using HCD activation mode, with mass resolution of 70000 and 17500 for precursor and fragment ions, respectively.

The acquired MS/MS raw data was converted into peaklists using an in-house software PAVA and then analyzed using Protein Prospector search engine. The peaklists were searched against the sequence of interest (Figure S5). The Max. missed cleavages was set to 3 and the precursor and fragment mass tolerances were set at 20 ppm each.

4. Cell-based assays

Cell- cell fusion assay

Cell-cell fusion was measured as previously published[1] using cell lines obtained through the NIH HIV Reagent Program, Division of AIDS, NIAID, NIH. Target cells were TZM-bl cells (#8129, contributed by J.C. Kappes, X. Wu and Tranzyme Inc.) expressing CD4, CCR5 and CXCR4,[7] and containing an integrated reporter gene for firefly luciferase under control of HIV-1 LTR.[8] Effector cells were HL2/3 (#1294, contributed by B.K. Felber and G.N. Pavlakis) which produce HXB2 Env, Tat and Rev.[9] Serially diluted inhibitors were added to 96 well plates containing 20,000 TZM-bl cells per well cultured overnight. 40,000 HL2/3 cells were added per well, and fusion allowed to proceed for 6 hours in reduced serum medium (OptiMEM, Gibco) with a final concentration of 1% DMSO. Cells were lysed (Luciferase Cell Culture Lysis 5x reagent, Promega) and luciferase expression was measured using Luciferase Assay Reagent (Promega) according to the manufacturer's instructions. Controls containing 1µl DMSO with and without HL2/3 cells were measured for each compound

Pseudotyped virus infectivity assay

Pseudotyped viruses were prepared by transfection of 293T cells with 1.6 µg Env deficient HIV-1 backbone vector pSG3Δenv (NIH HIV Reagent Program, Division of AIDS, NIAID, NIH, contributed by Dr's John C. Kappes and Xiaoyun Wu) and 0.8 µg Env expression vector pSM-HXB2-WTgp160, pJRFL-WTgp160, or pAmphoMLV-Env (donated by C. Weiss, FDA), using FuGENE HD transfection reagent (Promega). Pseudovirus containing culture supernatants were harvested 2 days after transfection, passed through 0.45 µm filters and stored in aliquots at −80 °C. For infectivity studies, 20,000 TZM-bl cells per well were plated overnight at 37°C, 5% CO_2 in Dulbecco's modified Eagle's medium (DMEM, Gibco) supplemented with 10% fetal bovine serum (FBS, Atlanta Biologicals) and 50 µg/ml Penicillin-Streptomycin (Cellgro). The following day, the medium was removed and pseudovirus particles in OptiMEM (Gibco) containing 25 µg/ml DEAE dextran (Sigma) were added at 10 tissue culture infective doses per well ($TCID_{50}$), in the absence or presence of serially diluted inhibitors. Positive and negative controls were included by omitting compounds or virus respectively. The final DMSO concentration was 1% in all wells. Cells were incubated at 37°C, 5% CO2 for 5 hours, after which virus and compounds were washed out and the medium replaced with DMEM. After further

incubation for 24 hours, cells were lysed (Luciferase Cell Culture Lysis 5x reagent, Promega) and luminescence measured in a Biotek Synergy 2 plate reader using Luciferase Assay Reagent (Promega, Cat#E153A), following the manufacturer's instructions.

Cell viability assay

The cytotoxic effect of the compounds was determined using the identical cell culture procedure and plates as described above for viral infectivity or cell-cell fusion, but measuring cell viability by fluorescence using a resazurin cell viability reagent (Presto Blue, Invitrogen-Thermo Fisher) following the manufacturer's protocols. Fluorescence was measured using a SpectraMax M5 plate reader (Molecular Devices). The lower bound reading was obtained from wells containing 15% DMSO.

Acknowledgements

The image in Figure S4 was produced using UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.

References

1. Zhou, G.; Sofiyev, V.; Kaur, H.; Snyder, B. A.; Mankowski, M. K.; Hogan, P. A.; Ptak, R. G.; Gochin, M., Structure-Activity Relationship Studies of Indole-Based Compounds as Small Molecule HIV-1 Fusion Inhibitors Targeting Glycoprotein 41. *Journal of Medicinal Chemistry* **2014,** *57* (12), 5270-81.
2. Walsh, J. D.; Chu, S.; Zhang, S. Q.; Gochin, M., Design and characterization of swapped-domain constructs of HIV-1 glycoprotein-41 as receptors for drug discovery. *Protein Eng Des Sel* **2015,** *28* (4), 107-16.
3. Hawkins, P. C.; Nicholls, A., Conformer generation with OMEGA: learning from the data set and the analysis of failures. *J Chem Inf Model* **2012,** *52* (11), 2919-36.
4. Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T., Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J Chem Inf Model* **2010,** *50* (4), 572-84.
5. Kuiken, C.; Korber, B.; Shafer, R. W., HIV sequence databases. *AIDS Rev* **2003,** *5* (1), 52-61.
6. Kilambi, K. P.; Gray, J. J., Rapid calculation of protein pKa values using Rosetta. *Biophys J* **2012,** *103* (3), 587-595.
7. Platt, E. J.; Wehrly, K.; Kuhmann, S. E.; Chesebro, B.; Kabat, D., Effects of CCR5 and CD4 cell surface concentrations on infections by macrophagetropic isolates of human immunodeficiency virus type 1. *J Virol* **1998,** *72* (4), 2855-64.
8. Wei, X.; Decker, J. M.; Liu, H.; Zhang, Z.; Arani, R. B.; Kilby, J. M.; Saag, M. S.; Wu, X.; Shaw, G. M.; Kappes, J. C., Emergence of resistant human immunodeficiency virus type 1 in patients receiving fusion inhibitor (T-20) monotherapy. *Antimicrob Agents Chemother* **2002,** *46* (6), 1896-905.
9. Ciminale, V.; Felber, B. K.; Campbell, M.; Pavlakis, G. N., A bioassay for HIV-1 based on Env-CD4 interaction. *AIDS Res Hum Retroviruses* **1990,** *6* (11), 1281-7.