**Table S1**. Primers used to create new restriction sites and CA mutations in pNL4-3 plasmid.

|  |  |
| --- | --- |
| **CA mutant/restriction sites** | **5'-3' (forward) sequence (mutations in bold)** |
| *sfi*I | CAAGCAGAAG**G**C**C**CAGCAGGC**C**GCGGCCGACA |
| *bst*EII | CGAGGCGATGAGCCA**G**GTG**A**C**C**AACACGACGATCATG |
| L6I | CCGATCGTGCAGAAC**A**TCCAGGGACAGATGG |
| V11I | TCCAGGGACAGATG**A**TCCACCAGGCCATC |
| H12Y | CAGGGACAGATGGTC**T**A**T**CAGGCCATCTCCCCA |
| I15L | CACCAGGCC**C**TCTCCCCACGGACGCTTAAC |
| L20I | CTCCCCACGGACG**A**TTAACGCGTGGGT |
| I27V | TAACGCGTGGGTCAAAGTA**G**TCGAGGAGAAGGC |
| A42D | TCCCCATGTTCTCGG**AT**CTTTCCGAGGGAGCC |
| S44A | CATGTTCTCGGCACTT**G**CCGAGGGAGCCAC |
| E45D | GTTCTCGGCACTTTCCGA**T**GGAGCCACCC |
| T48A | CCGAGGGAGCC**G**CCCCGCAGGAC |
| T54A | CCGCAGGACCTGAAC**G**CGATGTTGAACACCG |
| T54M | CCGCAGGACCTGAACA**T**GATGTTGAACACCGTC |
| T58I | CTGAACACGATGTTGAACA**T**CGTCGGCGGG |
| A64G | CGGGCACCAGG**G**GGCCATGCAGA |
| M68I | GGCGGCCATGCAGAT**A**CTTAAGGACACCATC |
| D71E | CAGATGCTTAAGGA**G**ACCATCAACGAGGAG |
| E98D | GGGCCAGATGAGAGA**T**CCGCGGGG |
| T110N | CGGGAACCACCAGCA**A**CTTGCAGGAGCAAAT |
| I124V | TGACTTCGAACCCGCCA**G**TCCCGGTCG |
| E128D | CAATCCCGGTCGGGGA**T**ATCTACAAGAGATGGA |
| V148T | GGATGTACAGCCCT**AC**CAGCATCCTGGAC |
| R154K | CAGCATCCTGGACATC**AA**ACAGGGACCGAAGGAG |
| F161S | GGGACCGAAGGAGCCG**AG**CAGAGACTACGTAGAC |
| D166G | GTTCAGAGACTACGTAG**G**CCGGTTCTTCAAGACTC |
| F169Y | TACGTAGACCGGTTCT**AT**AAGACTCTCCGGGCGG |
| F169A | CTACGTAGACCGGTTC**GC**CAAGACTCTCCGGGCG |
| R173K | GTTCTTCAAGACTCTC**AA**GGCGGAGCAGGCGACG |
| T200S | CCCGGACTGCAAGA**G**CATCCTGAAGGCTC |
| K203R | GACTGCAAGACCATCCTGA**G**GGCTCTCGGC |
| T116S | TGGAAGAGATGATG**T**CGGCGTGCCAGGG |
| G225S | AGGGAGTCGGGGGACCC**A**GCCACAAGGCG |
| V230I | CACAAGGCGCGG**A**TCTTGGCCGAGG |

**Table S2**. Distribution of HIV-1 subtypes in the dataset used to calculate amino acid sequence conservation and derive the COTM sequence.

|  |  |  |
| --- | --- | --- |
| **HIV-1 Subtypes in Group M** | **Number of sequences** | **Percentage** |
| Subtype A | 39 | 3.83 |
| Subtype B | 411 | 40.33 |
| Subtype C | 408 | 40.04 |
| Subtype D | 18 | 1.77 |
| Subtype F | 24 | 2.36 |
| Subtype G | 8 | 0.79 |
| Subtype H | 2 | 0.20 |
| Subtype J | 1 | 0.10 |
| Subtype K | 2 | 0.20 |
| CRFs | 106 | 10.40 |
| **Total** | **1019** | **100.00** |

**Table S3.** Database frequency of the consensus amino acid of group M HIV-1 CA

|  |  |
| --- | --- |
| **Sites (# of sites)** | **Average database frequency of consensus residue** |
| All CA sites (231) | 0.941 |
| All interface sites (48) | 0.936 |
| Interface sites in this study (12) | 0.872 |
| Non-interface sites in this study (21) | 0.810 |

**Table S4.** Subtype B, subtype C consensus and COTM-CA amino acid at the co-evolving residual pair. Taken from (7).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Co-evolving residues** | **Sitea** | **B consensus** | **C consensus** | **COTM-CA** |
| 27-148 | 27 | V | **I** | **I** |
|  | 148 | T | **V** | **V** |
| 41-120\* | 41 | **S** | T | **S** |
|  | 120 | N | **S** | **S** |
| 71-169 | 71 | E | **D** | **D** |
|  | 169 | Y | **F** | **F** |
| 116-124 | 116 | **G** | A | **G** |
|  | 124 | **I** | V | **I** |
| 148-169 | 148 | T | **V** | **V** |
|  | 169 | Y | **F** | **F** |
| 187-208 | 187 | E | **D** | **D** |
|  | 208 | A | **G** | **G** |

a Site numbering correspond to Gag protein of HIV-1 HXB2 strain

\* Indicates subtype specific co-evolving residual pair that is not maintained in our COTM-CA sequence

**Table S5.** Amino acid database frequency of subtype B, subtype C and other group M sequences

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mutation** | **Wild type amino acid frequency** | | | | **Mutant amino acid frequency** | | | |
| **Group M (n=1019)** | **Subtype B (n=411)** | **Subtype C (n=408)** | **Others (n=200)** | **Group M (n=1019)** | **Subtype B (n=411)** | **Subtype C (n=408)** | **Others (n=200)** |
| NTD-NTD | L6I3 | 0.687 | 0.713 | 0.860 | 0.280 | 0.066 | 0.114 | 0.039 | 0.021 |
| V11I1 | 0.939 | 0.976 | 0.973 | 0.795 | 0.037 | 0.015 | 0.012 | 0.135 |
| L20I1 | 1.000 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| A42D1 | 1.000 | 1.000 | 1.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| T54M3 | 0.867 | 0.971 | 0.953 | 0.475 | 0.096 | 0.002 | 0.005 | 0.475 |
| T54A3 | 0.867 | 0.971 | 0.953 | 0.475 | 0.005 | 0.002 | 0.005 | 0.010 |
| T58I3 | 0.876 | 0.978 | 0.968 | 0.480 | 0.103 | 0.002 | 0.010 | 0.500 |
| NTD-CTD | A64G1 | 0.999 | 1.000 | 1.000 | 0.995 | 0.001 | 0.000 | 0.000 | 0.005 |
| M68I1 | 0.981 | 0.968 | 0.990 | 0.990 | 0.013 | 0.024 | 0.002 | 0.010 |
| D71E2 | 0.554 | 0.054 | 0.983 | 0.705 | 0.439 | 0.934 | 0.012 | 0.290 |
| D166G1 | 0.998 | 1.000 | 0.995 | 1.000 | 0.002 | 0.000 | 0.005 | 0.000 |
| F169Y2 | 0.538 | 0.000 | 0.995 | 0.710 | 0.456 | 0.993 | 0.000 | 0.285 |
| F169A2 | 0.538 | 0.000 | 0.995 | 0.710 | 0.000 | 0.000 | 0.000 | 0.000 |
| R173K1 | 0.999 | 1.000 | 0.998 | 1.000 | 0.001 | 0.000 | 0.002 | 0.000 |
| Non-interface | H12Y1 | 0.984 | 1.000 | 0.993 | 0.935 | 0.013 | 0.000 | 0.002 | 0.060 |
| I15L3 | 0.617 | 0.720 | 0.650 | 0.340 | 0.320 | 0.253 | 0.321 | 0.455 |
| I27V2 | 0.575 | 0.251 | 0.880 | 0.620 | 0.419 | 0.742 | 0.113 | 0.380 |
| S44A1 | 0.965 | 0.925 | 0.990 | 0.995 | 0.029 | 0.068 | 0.002 | 0.005 |
| E45D1 | 0.987 | 0.990 | 0.983 | 0.990 | 0.008 | 0.002 | 0.012 | 0.010 |
| T48A1 | 0.984 | 0.988 | 0.983 | 0.980 | 0.006 | 0.000 | 0.007 | 0.015 |
| E98D1 | 0.918 | 0.912 | 0.914 | 0.935 | 0.074 | 0.080 | 0.076 | 0.055 |
| T110N1 | 0.861 | 0.876 | 0.821 | 0.910 | 0.118 | 0.110 | 0.145 | 0.080 |
| I124V1 | 0.739 | 0.929 | 0.556 | 0.720 | 0.246 | 0.056 | 0.422 | 0.280 |
| E128D2 | 0.601 | 0.895 | 0.302 | 0.605 | 0.390 | 0.090 | 0.689 | 0.395 |
| V148T2 | 0.663 | 0.238 | 0.968 | 0.915 | 0.274 | 0.630 | 0.015 | 0.070 |
| R154K2 | 0.555 | 0.698 | 0.373 | 0.635 | 0.438 | 0.294 | 0.618 | 0.365 |
| F161S1 | 0.993 | 0.993 | 0.993 | 0.995 | 0.002 | 0.000 | 0.003 | 0.005 |
| T200S1 | 0.842 | 0.961 | 0.880 | 0.520 | 0.091 | 0.005 | 0.025 | 0.405 |
| K203R2 | 0.588 | 0.983 | 0.196 | 0.575 | 0.406 | 0.010 | 0.799 | 0.420 |
| T216S1 | 0.976 | 0.964 | 0.985 | 0.985 | 0.009 | 0.017 | 0.003 | 0.005 |
| G225S2 | 0.562 | 0.679 | 0.490 | 0.470 | 0.416 | 0.287 | 0.500 | 0.510 |
| V230I1 | 0.925 | 0.869 | 0.971 | 0.950 | 0.066 | 0.119 | 0.020 | 0.050 |
| Co-evolving residues | S41T2 | 0.474 | 0.742 | 0.037 | 0.815 | 0.482 | 0.168 | 0.949 | 0.175 |
| N120S2 | 0.336 | 0.535 | 0.201 | 0.200 | 0.432 | 0.248 | 0.539 | 0.590 |
| G116A2 | 0.391 | 0.681 | 0.032 | 0.525 | 0.443 | 0.226 | 0.789 | 0.180 |
| D187E2 | 0.394 | 0.002 | 0.787 | 0.395 | 0.594 | 0.985 | 0.203 | 0.585 |
| G208A2 | 0.619 | 0.236 | 0.865 | 0.905 | 0.365 | 0.752 | 0.123 | 0.065 |

1 Amino acid site, whose frequency patterns are similar in all group M sequences

2 Amino acid site, whose frequency pattern in subtype B is the opposite of subtype C. The pattern in other group M sequences is either similar to subtype B or subtype C.

3 Amino acid site, whose frequency pattern in subtype B is similar to subtype C, but different from other group M sequences