**Supplementary material 1. The distribution of the geographic origin and risk group for 2408 database-derived sequences.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Geographic source** | **Sampling year** | **n** | **Risk group** |
| **BT** | **Hetero** | **IDUs** | **MSM** | **MTCT** | **SU** | **n/a** |
| Anhui | 2007-2009 | 12 | 4 |  |  |  |  |  | 8 |
| Beijing | 2004-2010 | 181 | 10 |  | 2 | 156 |  |  | 13 |
| Fujian | 1999-2007 | 24 | 1 | 20 | 1 |  |  |  | 2 |
| Guangdong | 2004-2009 | 175 | 1 | 19 | 13 | 5 |  |  | 137 |
| Gansu | 1998-2007 | 4 | 1 |  | 1 |  |  |  | 2 |
| Guizhou | 2007-2009 | 45 |  | 5 | 2 | 1 |  |  | 37 |
| Guangxi | 1996-2010 | 382 | 11 | 18 | 151 |  |  |  | 202 |
| Hainan | 2007 | 78 |  |  |  |  |  |  | 78 |
| Hebei | 2004-2009 | 32 | 14 |  |  | 14 |  |  | 4 |
| Henan | 2001-2010 | 218 | 94 | 4 | 1 | 22 |  |  | 97 |
| Hong Kong | 2007 | 2 |  |  | 1 |  |  |  | 1 |
| Heilongjiang | 2003-2011 | 68 |  | 6 | 1 | 27 |  |  | 34 |
| Hubei | 2004-2007 | 19 | 12 | 1 |  |  | 1 |  | 5 |
| Hunan | 2009 | 2 |  |  |  |  |  |  | 2 |
| Jilin | 2004-2010 | 9 | 2 | 3 |  | 2 |  |  | 2 |
| Jiangsu | 2006-2007 | 15 |  | 6 | 1 | 8 |  |  |  |
| Liaoning | 2000-2010 | 82 | 11 | 1 | 3 | 45 |  | 16 | 6 |
| Ningxia | 2002-2007 | 5 |  |  | 5 |  |  |  |  |
| Qinhai | 2005 | 2 |  |  | 2 |  |  |  |  |
| Sichuan | 1998-2009 | 62 |  | 1 | 53 |  |  |  | 8 |
| Shandong | unknown | 8 | 2 |  |  |  |  |  | 6 |
| Shanghai | 2008-2013 | 1 |  |  |  |  |  |  | 1 |
| Shanxi | 2003-2008 | 67 | 63 | 2 |  |  |  |  | 2 |
| Tianjin | 2007 | 1 |  |  |  | 1 |  |  |  |
| Xinjiang | 1996-2010 | 454 | 2 | 4 | 181 |  |  |  | 267 |
| Yunnan | 1992-2010 | 314 | 4 | 77 | 101 |  | 3 | 18 | 111 |
| Unknown | 2001-2011 | 146 | 12 | 1 | 11 | 2 |  |  | 120 |
| Total |  | 2408 | 244 | 168 | 530 | 283 | 4 | 34 | 1145 |

BT, blood transfusion; Hetero, heterosexuals; IDUs, injecting drug users; MSM, men who have sex with men; MTCT, mother-to-child transmission; SU, sexual transmission, unspecified type; n/a, not available.

**Supplementary material 2. RNA extraction, PCR amplification and sequencing for 526 HIV-1-infected MSM individuals.**

HIV-1 genomic RNA was extracted from 200 µl of plasma using the QIAmp Viral RNA Mini kit (Qiagen, Valencia, CA, USA), following Manufacturer’s instructions. Reverse transcription and nested polymerase chain (nPCR) amplification were performed by a home brew PCR procedure for partial genes of *pol* and *env* as described in our previous reports.12 A one-tube reverse transcriptase polymerase chain reaction kit (Gold-Script one-step RT-PCR kit, Life Technologies, USA), and PCR kit (TaKaRa Ex Taq, Takara Biotechnology Co, Ltd; Dalian, China) were used for amplification of the HIV-1 *pol* gene (protease 1–99 amino acids and part of reverse transcriptase 1–254 amino acids) and *env* gene (part of gp120 C2V5, 220 amino acids), according to the manufacture’s recommendations. About 1050 bp *pol* and 660 bp *env* were amplified. The PCR amplification was carried out in a thermal cycler (GeneAmp PCR System 9700, Applied Biosystems, USA). PCR products were directly sequenced in both directions with sequencing primers using ABI 3730 sequencer. Pre-PCR and post-PCR areas are strictly separated in order to avoid contamination from amplicon aerosol.

**Supplementary material 3. Co-receptor usage of 291 recently HIV-1-infected MSM.**

|  |  |  |
| --- | --- | --- |
| Subtype | **Algorithm I** | **Algorithm II** |
| **CCR5 (%)** | **Non-CCR5 (%)** | ***P* value** | **CCR5 (%)** | **Non-CCR5 (%)** | ***P* value** |
| CRF01\_AE (n=188) | 119(63.3) | 69 (36.7) | <0.001a | 125(66.5) | 63 (33.5) | <0.001a |
| CRF07\_BC (n=81) | 79 (97.5) | 2 (2.5) | <0.064b | 79 (97.5) | 2 (2.5) | <0.064b |
| B (n=22) | 19 (86.4) | 3 (13.6) | 0.031c | 19 (86.4) | 3 (13.6) | 0.057c |

Algorithm I: webPSSM+Geno2pheno (FPR=10%); Algorithm II: webPSSM+Geno2pheno (FPR=5%).

a CRF01\_AE versus CRF07\_BC; b CRF07\_BC versus subtype B; c CRF01\_AE versus subtype B

**Supplementary material 4.** **Viral Tropism at different genetic distance thresholds for phylogenetic transmission analysis.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Maximum distance threshold** | **Subtype** | **Algorithm I a** |  | **Algorithm II a** |
| **Clustered****R5** | **Non-****clustered R5** | **Clustered****X4/DM** | **Non-****clustered X4/DM** | ***P*** |  | **Clustered****R5** | **Non-****clustered****R5** | **Clustered****X4/DM** | **Non-****clustered****X4/DM** | ***P*** |
| 3.0% | Total | 82(37.8) | 135(62.2) | 26(35.1) | 48(64.9) | 0.683 |  | 84(37.7) | 139(62.3) | 24(35.3) | 44(64.7) | 0.723 |
|  | CRF01\_AE | 55(46.2) | 64(53.8) | 25(36.2) | 44(63.8) | 0.182 |  | 57(45.6) | 68(54.4) | 23(36.5) | 40(63.5) | 0.234 |
|  | CRF07\_BC | 21(26.6) | 58(73.4) | 1(50.0) | 1(50.0) | 0.462 |  | 21(26.6) | 58(73.4) | 1(50.0) | 1(50.0) | 0.462 |
|  | B | 6(31.6) | 13(68.4) | 0(0.0) | 3(100.0) | 0.532 |  | 6(31.6) | 13(68.4) | 0(0.0) | 3(100.0) | 0.532 |
| 2.0% | Total | 64(29.5) | 153(70.5) | 21(28.4) | 53(71.6) | 0.856 |  | 66(29.6) | 157(70.4) | 19(27.9) | 49(72.1) | 0.739 |
|  | CRF01\_AE | 39(32.8) | 80(67.2) | 20(29.0) | 49(71.0) | 0.590 |  | 41(32.8) | 84(67.2) | 18(28.6) | 45(71.4) | 0.555 |
|  | CRF07\_BC | 21(26.6) | 58(73.4) | 1(50.0) | 1(50.0) | 0.462 |  | 21(26.6) | 58(73.4) | 1(50.0) | 1(50.0) | 0.462 |
|  | B | 4(21.1) | 15(78.9) | 0(0.0) | 3(100.0) | 1.000 |  | 4(21.1) | 15(78.9) | 0(0.0) | 3(100.0) | 1.000 |
| 1.5% | Total | 52(24.0) | 165(76.0) | 16(21.6) | 58(78.4) | 0.681 |  | 53(23.8) | 170(76.2) | 15(22.1) | 53(77.9) | 0.771 |
|  | CRF01\_AE | 27(22.7) | 92(77.3) | 15(21.7) | 54(78.3) | 0.880 |  | 28(22.4) | 97(77.6) | 14(22.2) | 49(77.8) | 0.978 |
|  | CRF07\_BC | 21(26.6) | 58(73.4) | 1(50.0) | 1(50.0) | 0.462 |  | 21(26.6) | 58(73.4) | 1(50.0) | 1(50.0) | 0.462 |
|  | B | 4(21.1) | 15(78.9) | 0(0.0) | 3(100.0) | 1.000 |  | 4(21.1) | 15(78.9) | 0(0.0) | 3(100.0) | 1.000 |
| 1.0% | Total | 36(16.6) | 181(83.4) | 12(16.2) | 62(83.8) | 0.940 |  | 37(16.6) | 186(83.4) | 11(16.2) | 57(83.8) | 0.936 |
|  | CRF01\_AE | 19(16.0) | 100(84.0) | 11(15.9) | 58(84.1) | 0.996 |  | 20(16.0) | 105(84.0) | 10(15.9) | 53(84.1) | 0.982 |
|  | CRF07\_BC | 17(21.5) | 64(78.5) | 1(50.0) | 1(50.0) | 0.325 |  | 17(21.5) | 64(78.5) | 1(50.0) | 1(50.0) | 0.325 |
|  | B | 0(0.0) | 19(100.0) | 0(0.0) | 3(100.0) |  |  | 0(0.0) | 19(100.0) | 0(0.0) | 3(100.0) |  |
| 0.5% | Total | 20(9.0) | 203(91.0) | 7(9.2) | 69(90.8) | 0.950 |  | 21(9.4) | 202(90.6) | 6(8.8) | 62(91.2) | 0.883 |
|  | CRF01\_AE | 9(7.6) | 110(92.4) | 6(8.7) | 63(91.3) | 0.782 |  | 10(8.0) | 115(92.0) | 5(7.9) | 58(92.1) | 0.988 |
|  | CRF07\_BC | 11(13.9) | 69(86.1) | 1(50.0) | 1(50.0) | 0.276 |  | 11(13.9) | 68(86.1) | 1(50.0) | 1(50.0) | 0.276 |
|  | B | 0(0.0) | 19(100.0) | 0(0.0) | 3(100.0) |  |  | 0(0.0) | 19(100.0) | 0(0.0) | 3(100.0) |  |

**a** Algorithm I: webPSSM+Geno2pheno (FPR=10%); Algorithm II: webPSSM+Geno2pheno (FPR=5%).