SCIENTIFIC DATA

OPEN

Data Descriptor: Geographically-stratified HIV-1 group M *pol* subtype and circulating recombinant form sequences

Received: 3 October 2017
Accepted: 21 May 2018
Published: 31 July 2018

Soo-Yon Rhee[1] & Robert W. Shafer[1]

Accurate classification of HIV-1 group M lineages, henceforth referred to as subtyping, is essential for understanding global HIV-1 molecular epidemiology. Because most HIV-1 sequencing is done for genotypic resistance testing *pol* gene, we sought to develop a set of geographically-stratified *pol* sequences that represent HIV-1 group M sequence diversity. Representative *pol* sequences differ from representative complete genome sequences because not all CRFs have *pol* recombination points and because complete genome sequences may not faithfully reflect HIV-1 *pol* diversity. We developed a software pipeline that compiled 6,034 one-per-person complete HIV-1 *pol* sequences annotated by country and year belonging to 11 pure subtypes and 70 CRFs and selected a set of sequences whose average distance to the remaining sequences is minimized for each subtype/CRF and country to generate a Geographically-Stratified set of 716 *Pol* Subtype/CRF (GSPS) reference sequences. We provide extensive data on *pol* diversity within each subtype/CRF and country combination. The GSPS reference set will also be useful for HIV-1 *pol* subtyping.

[1]Division of Infectious Diseases, Department of Medicine, Stanford University, Stanford, CA 94301, USA. Correspondence and requests for materials should be addressed to S.-Y.R. (email: syrhee@stanford.edu).

Design Type(s)	gene family classification objective • sequence analysis objective • sequence data transformation objective • source-based data analysis objective • data integration objective
Measurement Type(s)	Molecular Genetic Variation
Technology Type(s)	computational phylogenetic analysis
Factor Type(s)	geographic location • Genetic Variation • biological replicate
Sample Characteristic(s)	Human immunodeficiency virus 1 • Kenya • Cyprus • Tanzania • Uganda • Ukraine • Cameroon • Pakistan • Russia • Rwanda • Uzbekistan • India • Kazakhstan • Republic of South Africa • Democratic Republic of the Congo • Sweden • Kingdom of Spain • Senegal • Nigeria • Australia • Belarus • Georgia • Italy • Somalia • United States of America • Japan • China • Brazil • South Korea • Germany • Thailand • Kingdom of Denmark • Argentina • French Republic • Peru • United Kingdom • Jamaica • Canada • Colombia • Cuba • Haiti • Trinidad and Tobago • Uruguay • Dominican Republic • Hong Kong • Myanmar • Kingdom of the Netherlands • The Philippines • Paraguay • Yemen • Switzerland • Ecuador • Taiwan Province • Bolivia • Gabon • Mexico • Poland • Zambia • Botswana • Malawi • Ethiopia • Nepal • Angola • Israel • Chad • Romania • Finland • Portuguese Republic • Ghana • Guinea-Bissau • Belgium • Central African Republic • Viet Nam • Afghanistan • Indonesia • Iran • Estonia • Niger • Greece • Mali • Burkina Faso • Benin • Cote d'Ivoire • Saudi Arabia • Malaysia • Luxembourg • Chile • Gambia • Singapore

Background & Summary

Accurate classification of HIV-1 group M lineages, henceforth referred to as subtyping, has been essential for understanding the evolution of divergent HIV-1 in the context of the global pandemic. The classification of such sequences is complicated by the HIV-1's high mutation rate and propensity to develop new recombinant forms when two different virus strains infect the same cell. Indeed, HIV-1 group M sequences can be classified into many different lineages referred to as pure subtypes and circulating recombinant forms (CRFs).

The HIV research community led by researchers at the Los Alamos National Laboratories (LANL) HIV Sequence Database established criteria for the taxonomic recognition of the pure subtypes A, B, C, D, F, G, H, J, and K and for an increasing number of CRFs[1,2]. A new CRF has generally been established and assigned a number when a recombinant virus with unique breakpoints has been sequenced in its entirety and identified in three or more epidemiologically unlinked individuals. Recombinant viruses that do not share breakpoints with an established CRF are classified as unique recombinant forms (URFs)[3,4].

Several online programs use phylogenetic or bioinformatic approaches to determine the subtype/CRF of a submitted sequence. Each of these programs relies on a set of ~9,700 bp full-genome reference sequences for each subtype and for a large proportion of CRFs[5–9]. The vast majority of HIV-1 sequencing, however, is done for genotypic resistance testing and is therefore confined to the *pol* gene (2,841 bps; 947 amino acids) which encodes the protease, reverse transcriptase (RT), and integrase enzymes. We therefore sought to develop a set of representative *pol* sequences for subtyping of HIV-1 *pol* sequences. Selecting representative *pol* sequences for each subtype/CRF based solely on *pol* will differ from selecting them based on complete genome sequences because not all CRFs have breakpoints in *pol*. Furthermore, sequences selected based on their complete genome may not reflect the diversity of HIV-1 *pol* because there are fewer complete genome sequences than *pol* sequences.

As many subtypes and CRFs exhibit regional divergence[10–13], we developed a systematic approach for selecting a Geographically-Stratified *Pol* Subtype/CRF (GSPS) reference dataset. For each distinct subtype/CRF and country combination, we characterized the extent of diversity in the *pol* gene and applied a partitioning around medoids (PAM) algorithm to identify the smallest number of centrally located sequences that would minimize the average distance to the closest leaf (ADCL) of the complete set of subtype/CRF/country sequences[14]. This approach is designed to select a subset of sequences that is designed to represent the diversity of a larger collection of sequences[14].

It should be noted, however, that the collection of published HIV-1 sequences may not perfectly represent the entirety of global HIV-1 sequences. Indeed, as we indicate in this study, many CRFs are likely over-represented in public sequence databases compared with their prevalence in HIV-1 infected persons. To counter this fact, we supplemented our GSPS reference dataset with a set of additional sequences that were closest to the progenitors of many of the reported CRFs.

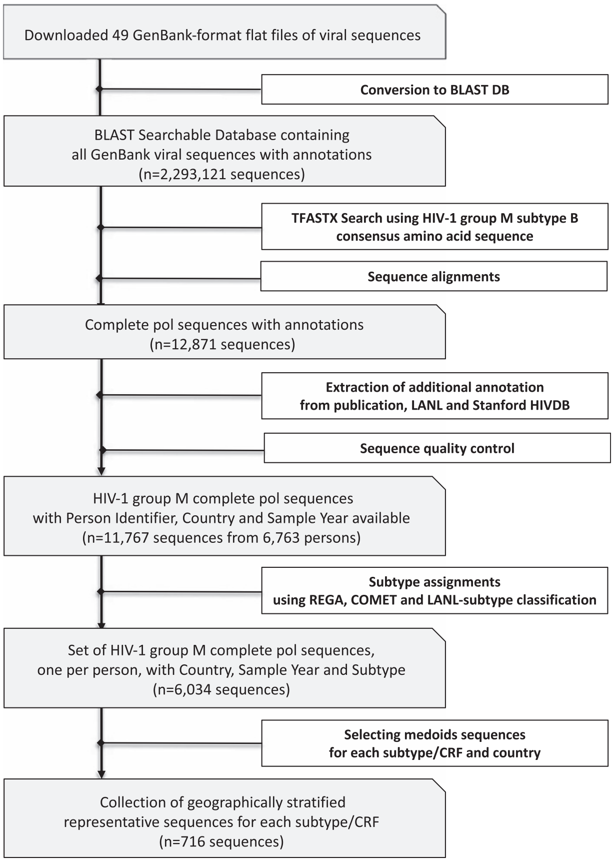
Downloaded 49 GenBank-format flat files of viral sequences
Conversion to BLAST DB
BLAST Searchable Database containing all GenBank viral sequences with annotations (n=2,293,121 sequences)
TFASTX Search using HIV-1 group M subtype B consensus amino acid sequence
Sequence alignments
Complete pol sequences with annotations (n=12,871 sequences)
Extraction of additional annotation from publication, LANL and Stanford HIVDB
Sequence quality control
HIV-1 group M complete pol sequences with Person Identifier, Country and Sample Year available (n=11,767 sequences from 6,763 persons)
Subtype assignments using REGA, COMET and LANL-subtype classification
Set of HIV-1 group M complete pol sequences, one per person, with Country, Sample Year and Subtype (n=6,034 sequences)
Selecting medoids sequences for each subtype/CRF and country
Collection of geographically stratified representative sequences for each subtype/CRF (n=716 sequences)

Figure 1. Schematic overview of the process of creating a set of Geographically-Stratified *Pol* Subtype/CRF (GSPS) reference sequences.

Figure 1 shows the complete process of (1) Identifying complete *pol* sequences from GenBank; (2) Extracting annotation from GenBank, LANL[1], the HIV Drug Resistance Database[15], and associated PubMed publications; (3) Performing sequence quality control; (4) Assigning a consensus subtype to those *pol* sequences for which there was an agreement using two or more of the following subtyping classifications: LANL[1], Rega subtyping tool[5], and COMET[6]; and (5) Selecting the sequences that comprise the GSPS reference set. In addition, the manuscript also characterizes the extent of *pol* diversity within each subtype/CRF and country combination and demonstrates the potential usefulness of the GSPS set of sequences for subtyping studies.

Methods

Compilation of HIV-1 group M complete *pol* sequences

We accessed the repository of NCBI-GenBank Flat File Release 220.0 (release date of June 15, 2017)[16] and downloaded all files containing GenBank virus sequence records (gbvrl1.seq.gz – gbvrl49.seq.gz). Using a custom Perl script (GB_to_Fasta.pl), we parsed each of the 2,293,121 virus sequence records and extracted data from the following GenBank fields: “TITLE”, “AUTHORS”, “TaxonID”, and, if available, “PUBMED”, “collection_date”, “country”, and “notes”. We created an fasta sequence file in which each sequence header contained the extracted sequence annotations. We then converted the fasta sequence file to a BLAST-searchable database using MAKEBLASTDB (NCBI blast+ package)[17].

To retrieve sequences homologous to the HIV-1 subtype B *pol* consensus amino acid sequence, we searched the database using TFASTX (FASTA package)[18]. TFASTX is a local sequence alignment program which searches translated nucleotide sequences using an amino acid query sequence. Unlike TBLASTN (NCBI blast+ package), it can extend an alignment across reading frames[19]. Using an e-value of 0.0000001, TFASTX returned 54,781 High-scoring Segment Pairs (HSPs). We then selected the 12,871 HSPs with a similarity score >80 and 947 amino acids. To remove frameshifts and insertions and to generate a nucleotide sequence alignment of 2,841 bps in length, we used the global nucleotide sequence alignment program, GGSEARCH (FASTA package). A second custom Perl script (Gene_to_Sequences.pl) was developed to streamline the above processes.

The 12,871 HIV-1 complete *pol* sequences were grouped into 631 submission sets containing the same submission “Title” and “Authors”. Approximately 85% of the 631 submission sets were linked to one or more publications containing a PubMed ID. To describe the provenance of each sequence (i.e., whether it

Subtype/CRF	Global Distribution, %[a]	Complete *pol* Sequences, % (n = 6,763)[b]
C	48	18.0
A	12	4.6
B	11	32.2
CRF02_AG	8	1.9
CRF01_AE	5	17.9
G	5	1.1
D	2	1.3
F+H+J+K	1	1.4
Other CRFs	4	11.2
URFs	4	10.5

Table 1. Proportions of each HIV-1 group M subtype/CRF in the set of 6,763 one-per-person complete *pol* sequences in GenBank compared with the reported global distribution of each subtype/CRF. [a]Global distribution of HIV-1 obtained from the WHO-UNAIDS Network for HIV Isolation and Characterisation[22]. The data was collected between 2000 and 2007 from researchers and literature review. [b]Proportion of complete *pol* sequences in GenBank (one per person) as of June 15, 2017.

was the only sequence obtained from an individual vs. whether it was one of multiple sequences obtained from an individual), we created a Person ID descriptor for each sequence using annotations extracted from the GenBank record, LANL, HIVDB, and published papers.

From the set of 12,871 complete HIV-1 *pol* sequences, we excluded 1,104 sequences including 55 non-group M sequences and 1,049 sequences that either lacked annotation for country or sample year (n = 339), lacked a Person ID descriptor (n = 234), or had poor sequence quality (n = 476) defined as the absence of ≥10 amino acid positions or the presence of ≥3 stop codons, ≥20 highly unusual mutations, or ≥5 signature APOBEC-associated mutations[20]. These exclusions left us with 11,767 well-characterized *pol* sequences from 6,763 persons. One sequence per person was then selected for subtyping. If one complete *pol* sequence was available from a person, we selected the complete *pol* sequence from that person. Otherwise, we used the random number generator implemented in mysql with a seed and chose one random sequence from that person.

Subtype assignment

For the 6,763 one-per-person complete *pol* sequences, we recorded the subtype classification in the LANL database, which is based on the author-defined subtype and additional analyses performed by LANL. This classification is based on the entire sequence in a GenBank record and is therefore influenced by sequence outside of the *pol* region. We also submitted each of the *pol* sequences to the online subtyping programs Rega and COMET[5,6]. Sequences for which two or more of three classifications agreed were assigned a consensus subtype/CRF.

Overall, 6,034 (89.2%) of the 6,763 sequences were assigned a consensus subtype/CRF: 11 pure subtypes including A2 and F2 (65.6% of 6,034) and 70 CRFs (34.4% of 6,034). The sequences assigned a subtype included (i) 5,091 sequences (84.4% of 6,034) for which the classifications assigned by Rega, COMET, and LANL yielded the same classification; (ii) 875 (14.5% of 6,034) for which two of the three approaches yielded the same classifications; and (iii) 68 (1.1% of 6,034) for recently recognized CRFs in LANL but for which reference sequences did not exist in Rega and COMET.

Thirty-nine sequences assigned to seven CRFs were reclassified because they did not have recombination breakpoints in *pol*: CRF14_BG, CRF15_01B, CRF32_06A1, CRF41_CD, CRF46_BF1, CRF57_BC and CRF88_BC were reclassified based on their *pol* sequence to subtype G, CRF01_AE, CRF06_cpx, subtype D, subtype F1, subtype C, and subtype C, respectively. Twenty sequences assigned to CRF70_BF1, CRF71_BF1 and CRF72_CRF1 were not included because the breakpoints in their *pol* genes were inconsistent[21]. The majority of the remaining 709 sequences (10.5% of 6,763), for which the three approaches diverged, were URFs typically reported as unassigned by Rega and COMET. They were excluded from subsequent analyses because URFs are not recognized established lineages. Most of these URFs contained sequence segments consistent with CRF01 and B (Southeast Asia), B and C (South East Asia), B and F (South America), and A and D (East Africa).

The overall distribution of subtypes in the 6,763 *pol* sequences in order of frequency was as follows: subtype B (32.2%), subtype C (18.0%), CRF01_AE (17.9%), Other CRFs (11.2%), URFs (10.5%), subtype A (4.6%), CRF02_AG (1.9%), subtype D (1.3%), and subtype G (1.1%). The remaining four uncommon subtypes F, H, J, and K comprised 1.4% of the dataset. Table 1 compares this distribution to the estimated global distribution of subtypes from a 2011 publication describing the WHO/UNAIDS Network characterizing the global and regional distribution of subtypes[22]. This comparison shows that our dataset of complete *pol* sequences was more likely to contain viruses belonging to subtype B, CRF01_AE, other

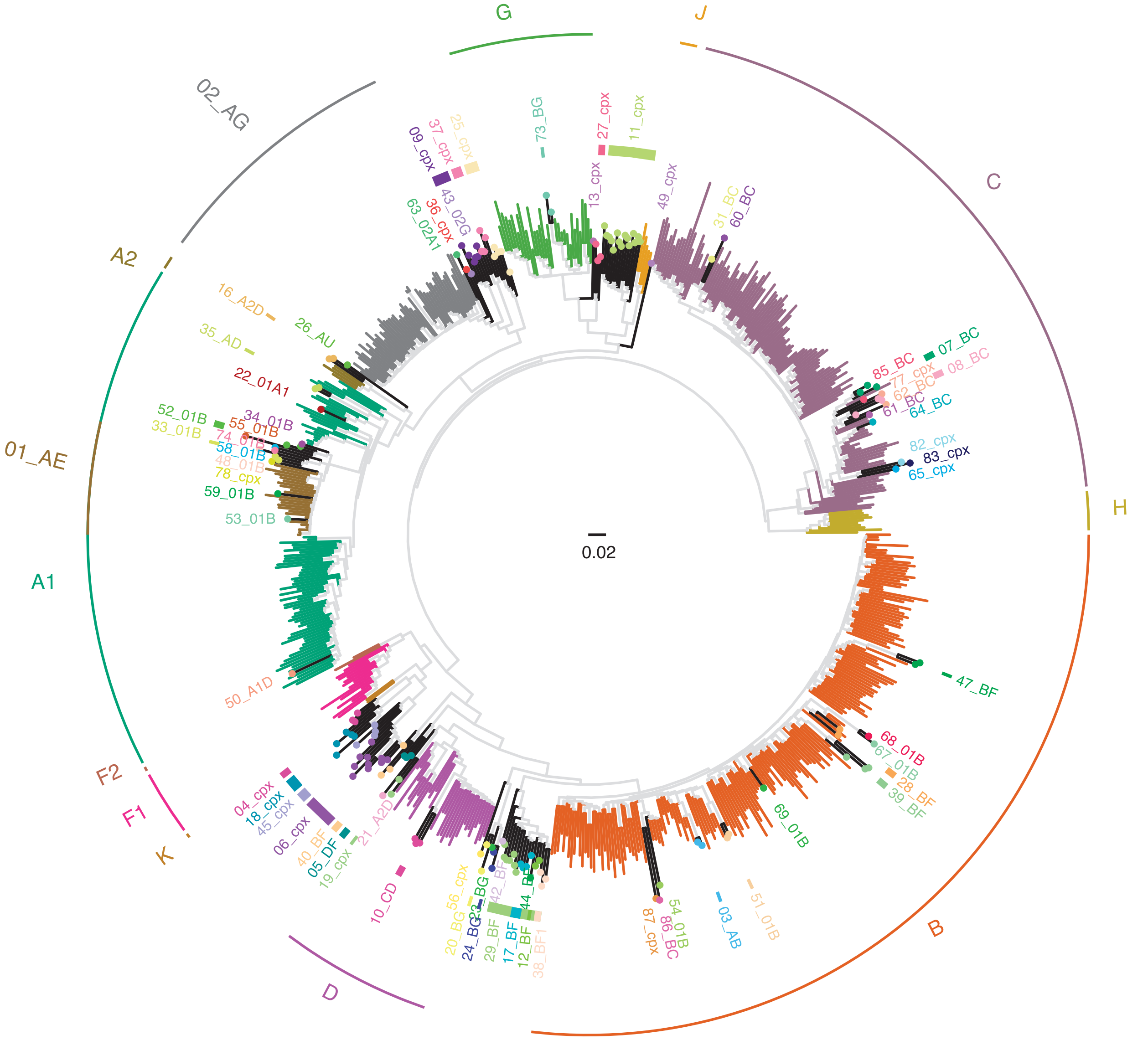

Figure 2. Phylogenetic tree of 716 Geographically-Stratified *Pol* Subtype/CRF (GSPS) reference sequences representing 11 pure subtypes and 70 CRFs. Branches of GSPS sequences belonging to the pure subtypes A1, A2, B, C, D, F1, F2, G, H, J, and K and the highly prevalent CRFs, CRF01_AE and CRF02_AG are color-coded and their clades are indicated in the outer ring. GSPS sequences belonging to the remaining 68 CRFs are indicated by black branches with color-coded circles on the branch tips. The tree was constructed using neighbour joining with branch length optimized by maximum likelihood method with GTR evolution model using the R package phangorn and then rooted at the mid-point. The tree was illustrated using the R package ggtree.

CRFs, and URFs and less likely contain viruses belonging to subtypes C, A, and CRF02_AG compared those in the WHO/UNAIDS publication.

The 6,034 one-per-person, subtype-assigned *pol* sequences were from 88 countries in the following seven regions: Asia (43.5%), Sub-Saharan Africa (26.4%), North America (12.9%), Europe (8.2%), Latin America and Caribbean (7.2%), former Soviet Union (1.5%) and North Africa and Middle East (0.4%) (Complete.Set.txt, Data Citation 1). The countries with the greatest number of subtype/CRFs were China (18), Cameroon (16), DRC (15), Spain (12) and Brazil (11). Twenty-nine countries had 3 to 9 subtype/CRFs and 55 countries had 1 to 2 subtype/CRFs. The number of subtype/CRFs per country was influenced in large part by the frequency with which laboratories in a country performed complete *pol* sequencing.

The subtypes present in the largest number of countries were B (42 countries), C (30 countries), A1 (24 countries), CRF02_AG (20 countries), CRF01_AE (15 countries), D (13 countries), G (13 countries),

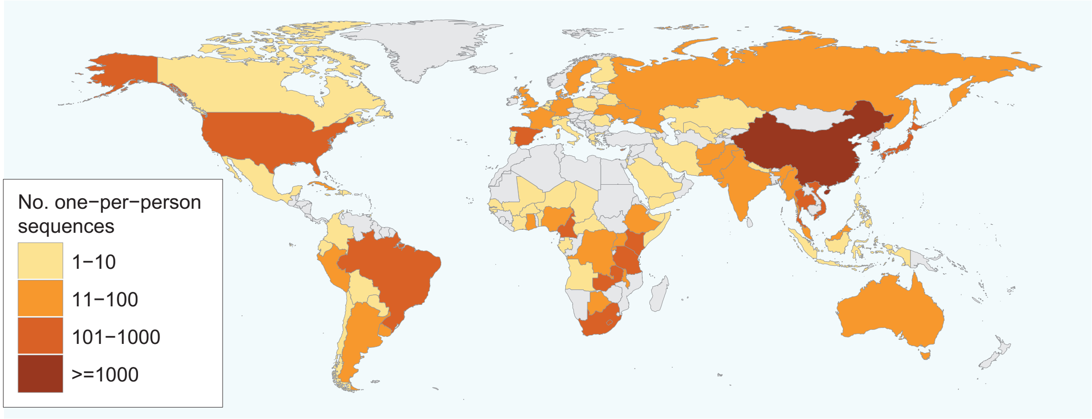

Figure 3. Number of sequences by country in the complete set of 6,034 one-per-person complete HIV-1 group M *pol* sequences. The large number of sequences from China is consistent with the frequent sequencing of complete genome and complete *pol* sequences in research and public health laboratories in this country.

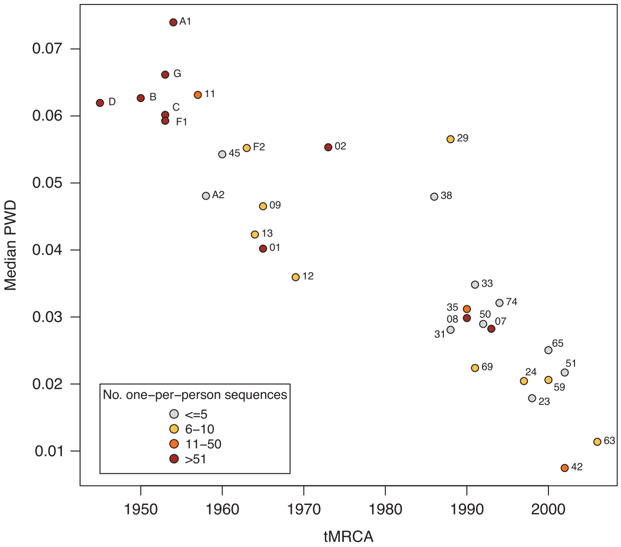

Figure 4. Correlation between the estimated year of the most recent common ancestor (MRCA) and the intra-subtype/CRF diversity in the complete set of 6,034 one-per-person complete HIV-1 group M *pol* sequences. Median PWD: median of the intra-subtype/CRF pairwise distances using TN93 substitution model. Each point indicates a subtype or CRF with CRFs labelled by their number alone. Points have been manually jittered to minimize overlap. Year of the MRCA was obtained from the references listed in Table 4.

F1 (10 countries) and CRF06_cpx (10 countries). Sequences from more than 100 persons were available for five subtype/CRFs including B (n = 2,175), C (n = 1,215), CRF01_AE (n = 1,210), A1 (n = 306), CRF07_BC (n = 300) and CRF02_AG (n = 130). The median sampling year was 2007 (IQR: 2003-2009). Fifty-four percent of the sequences were complete genome sequences.

Creating a set of Geographically-Stratified *Pol* Subtype (GSPS) reference sequences

To identify a set of representative sequences, we applied the k-medoids method using the partitioning around medoids (PAM) algorithm implemented in the Reference Package PReparare (rppr) program in the pplacer package[14]. Medoids are representative members of a dataset whose average distance from all the members in the dataset is minimized. The rppr program takes a phylogenetic tree and the number (k) of medoids (leaves or sequences) to be subsampled that minimize the average distance to the closest leaf (ADCL). We performed this analysis for each set of three or more sequences from the same country within a subtype/CRF.

Subtype/CRF	Reference	tMRCA	CI[1]
A1	Wertheim, J. O. *et al*[25]	1946	1936–1956
A2	Wertheim, J. O. *et al*[25]	1952	1941–1936
B	Wertheim, J. O. *et al*[25]	1955	1946–1964
C	Wertheim, J. O. *et al*[25]	1939	1926–1951
D	Abecasis, A.B. *et al*[26]	1945	1935–1955
F1	Wertheim, J. O. *et al*[25]	1950	1940–1959
F2	Wertheim, J. O. *et al*[25]	1961	1954–1968
G	Tongo, M. *et al*[27]	1953	1939–1963
01_AE	Liao, H. *et al*[28]	1967	1963–1973
02_AG	Faria, N.R. *et al*[29]	1973	1972–1975
07_BC	Tee, K.K. *et al*[30]	1993	1991–1995
08_BC	Tee, K.K. *et al*[30]	1990	1988–1991
09_cpx	Delatorre, E. *et al*[31]	1966	
11_cpx	Delatorre, E. *et al*[31]	1957	
12_BF	Dilernia, D. A. *et al*[32]	1969	1946–1981
13_cpx	Delatorre, E. *et al*[31]	1965	
20_BG	Delatorre, E. *et al*[33]	1996	1994–1998
23_BG	Delatorre, E. *et al*[33]	1998	1996–2000
24_BG	Delatorre, E. *et al*[33]	1997	1996–2000
28_BF	Ristic, N. *et al*[34]	1988	1984–1993
29_BF	Ristic, N. *et al*[34]	1988	1984–1993
31_BC	Passaes, C.P. *et al*[24]	1988	1982–1992
33_01B	Tee, K.K. *et al*[35]	1992	1987–1997
35_AD	Eybpoosh, S. *et al*[36]	1991	
38_BF1	Bello, G. *et al*[37]	1986	1981–1990
42_BF	Struck, D. *et al*[38]	2002	2001–2003
45_cpx	Delatorre, E. *et al*[39]	1965	
48_01B	Li, Y. *et al*[40]	2001	1998–2004
50_A1D	Foster, G. M. *et al*[41]	1992	1966–2007
51_01B	Ng, K. T. *et al*[42]	2000	1992–2006
59_01B	Zhang, W. *et al*[43]	2000	1994–2005
63_02A1	Shcherbakova, N. S. *et al*[44]	2006	2005–2007
65_cpx	Liu, Y. *et al*[45]	2000	1997–2003
69_01B	Hosaka, M. *et al*[46]	1993	1978–1999
74_01B	Cheong, H. T. *et al*[47]	1995	

Table 4. Time to the most recent common ancestors (tMRCA) of HIV-1 group M subtypes and CRFs.

Overall, there were 294 distinct combinations of subtypes/CRFs and country (subtype/CRF/country) of which 155 contained three or more sequences. For these 155 sequence sets, we created a neighbour-joining tree rooted at its midpoint using the TN93 substitution model. We then used rppr to identify the set of sequences that minimized ADCL for each k with k ranging from 1 to the number of sequences in the tree (n)–1.

After plotting, k vs. ADCL, for each subtype/CRF/country combination (ADCL.txt, Data Citation 1), we decided to select the minimum set of k sequences that lowered ADCL below 4.0% because this could be achieved for 72% (111) of the 155 subtype/CRF/country combinations. However, for 14 (9%) subtype/CRF/country combinations for which there was much sequence diversity and for which the k required to lower ADCL below 4.0% was high, we selected the set of k sequences that lowered ADCL below a higher threshold: (i) 4.5% for subtype A1 in Kenya, Tanzania, and Uganda; for subtype B in Cyprus, Germany, Denmark, Spain; for subtype C in Botswana, Ethiopia, Kenya, Tanzania, and South Africa; and for subtype G in Nigeria; and (ii) 5.0% for subtype B in Brazil. For the remaining 30 subtype/CRF/country combinations (19%) for which the ADCL remained above 4.5% regardless of k, we included all 141 sequences.

For 139 subtype/CRF/country combinations with just one or two sequences, we included just one sequence unless the two sequences belonging to that subtype/CRF/country differed by more than 4.0%.

NCBI	GSPS	# Sequences	Country/Region[a]	Notes[b]
CRF01_AE	A1	5	Eastern Africa (3), South Africa (1), Cyprus (1)	CRF01_AE has been reported primarily in Southeast Asia and Central Africa. A1 is common in East Africa
CRF03_AB	B	1	US (1)	CRF03_AB has been reported primarily in Eastern Europe and Central Asia.
CRF08_BC	C	14	**Eastern Africa (n = 7), Southern Africa (n = 4)**, China (1), India (2)	CRF08_BC has been reported primarily in China.
CRF22_01A1	A1	9	Cameroon (9)	CRF22_01A1 has been reported primarily in Cameroon. Subtype A1 is also common in Cameroon
CRF28_BF	B	4	Brazil (4)	CRF28_BF has been reported primarily in Brazil. Subtype B is also common in Brazil.
CRF31_BC	C	15	**East Africa (n = 11), Southern Africa (n = 4)**	CRF31_BC has been reported primarily in Brazil.
CRF51_01B	B	2	Japan (1), Philippine (1)	CRF51_01B has been reported primarily in Southeast Asia and Japan. Subtype B is common in Japan.
CRF64_BC	C	6	China (6)	CRF64_BC has been reported primarily China. Subtype C is also common in China.
CRF69_01B	B	5	Japan (5)	CRF69_01B has been reported primarily in Japan. Subtype B is common in Japan

Table 5. Test Sequences for which the Subtype/CRF of the Closest NCBI Reference Sequence Differed from the Subtype/CRF of the Closest Geographically-Stratified *Pol* Subtype/CRF (GSPS) Reference Sequence (n = 61; 1.2% of 5,185 Test Sequences). [a]Sequences for which the closest NCBI sequence are unlikely to represent the correct subtype/CRF of the sequences are shown in bold. [b]The information in this column was obtained from Los Alamos National Laboratories (LANL) HIV Sequence Database[1].

Overall the above procedures yielded 684 sequences including (i) 523 sequences from the 155 subtype/CRF/country combinations containing three or more sequences and (ii) 161 sequences from the 139 subtype/CRF/country combinations containing just one or two sequences.

Supplementing GSPS sequences with subtype sequences closest to CRF progenitors

The set of publicly available HIV-1 sequences, particularly sequences contributed by research laboratories, is likely to contain a higher relative proportion of CRF than non-CRF sequences compared with their prevalence in HIV-1 infected patients because CRF sequences may be more likely to be published. If a set of reference sequences contains a CRF sequence but does not contain the non-CRF ancestors that gave rise to the CRF, some non-recombinant sequences may appear to be more similar to the CRF reference sequence than to the parent subtype. This is particularly likely to occur when the CRF is dominated by one subtype and contains just a small segment of another subtype.

To counter this effect, we added 32 additional reference sequences from subtypes A1, A2, B, C, D, F1, G, and CRF01_AE using the following approach. We used the LANL CRF gene maps to identify 49 CRFs comprising one or two of the following subtype/CRFs: A1, A2, B, C, D, F1, G, and CRF01_AE in their *pol* genes. We then created 60 neighbour-joining trees containing each of the sequences belonging to 49 CRFs and each of the sequences from that CRF's dominant *pol* subtype according to the CRF gene map. Each of these trees comprised just the region of *pol* shared by the CRF and its parent subtype. For several CRFs, for which more than one subtype comprised nearly one-half of *pol*, we created trees with that CRF and both parent subtypes. Although CRF01_AE is itself a CRF, it is also the parent for several other CRFs. We used R packages including ape and phytools to identify the most recent common ancestor (MRCA) of the CRF and the sequence within the parent subtype with the shortest patristic distance to the MRCA. Overall 48 distinct sequences from 55 trees were identified of which 32 were not already included in our set of 684 representative sequences, yielding a final total of 716 GSPS sequences (GSPS.Set.txt, Data Citation 1).

Table 2 (available online only) describes the complete set of 6,034 one-person HIV-1 group M *pol* sequences and the 716 GSPS sequences. For each subtype/CRF, Table 2 (available online only) contains the number of sequences, the average pairwise distance (PWD) using the TN93 substitution model, and the number of sequences per country. It also describes the number of sequences per country in the GSPS panel and the overall ADCL for each subtype/CRF using the 716 GSPS sequences. Table 3 (available online only) lists the mean and maximum PWDs within each subtype/CRF for both the set of complete *pol* sequences and the GSPS subset. The data in the table indicates that the mean and maximum PWDs in the complete dataset and in the GSPS subset were very similar.

Code availability

Scripts used for the present reference datasets generation are available from the site https://github.com/hivdb/Gene_to_Sequences

Data Records

All reference datasets are available from Dryad Digital Repository (Data Citation 1). The complete set of 6,034 one-per-person, subtype-assigned HIV-1 group M complete *pol* sequences is available in a tab-delimited file containing the aligned *pol* sequence, GenBank accession number, GenBank author list,

Consensus Subtype/CRF[a]	GSPS[b]	# Sequences	Country	Notes[c]
CRF02_AG	CRF36_cpx (1)	1	Cameroon (1)	CRF02_AG is common in Cameroon. CRF36_cpx has been reported primarily in Cameroon.
CRF12_BF	CRF17_BF (1), B (3)	4	Argentina (4)	CRF12_BF and CRF17_BF has been reported primarily in South America. Subtype B is also common in South America.
A1	CRF22_01A1 (3)	3	Cameroon (3)	Subtype A1 is common in Cameroon. CRF22_01A1 has been reported primarily in Cameroon.
A1	A2 (1)	1	Republic of the Congo (1)	Subtypes A1 and A2 are common in Republic of the Congo.
B	CRF17_BF(1), CRF38_BF1 (2)	3	Argentina (3)	Subtype B and CRF17_BF are common in South America. CRF38_BF1 has been reported primarily in Uruguay.
B	CRF28_BF (1)	1	Brazil (1)	Subtype B and CRF28_BF are common in Brazil.
C	CRF07_BC (1), CRF85_BC (3)	4	China (4)	Subtype C is common in China. CRF07_BC and CRF85_BC have been reported primarily in China.
D	B (3)	3	Republic of the Congo (1), South Africa (1), Spain (1)	Subtype D is generally only seen in Eastern and Central Africa.
G	CRF43_02G (1)	1	Nigeria (1)	CRF43_02G is primarily reported in Saudi Arabia. Subtype G has been reported commonly in Africa.
G	CRF73_BG (6)	6	Portugal (4), Spain (2)	Subtype G is common in Central Europe. CRF73_BG has been reported in Portugal and Spain.

Table 6. 1300 bp *pol* Test Sequences for which the Consensus Subtype/CRF assigned by COMET and Rega Differed from the Subtype/CRF of the Closest Geographically-Stratified *Pol* Subtype/CRF (GSPS) Reference Sequence (n = 27; 0.4% of the 6,115 Test Sequences with a Consensus Subtype). [a]Subtype/CRF classifications of test sequences that were agreed by COMET and Rega subtyping programs. [b]Subtypes/CRFs of the closest GSPS sequences of the test sequences that differed from their consensus subtype/CRF. [c]The information in this column was obtained from Los Alamos National Laboratories (LANL) HIV Sequence Database[1].

GenBank submission title, PubMed ID, country, sample year, and assigned subtype/CRF (Complete.Set.txt). The GSPS reference set of 716 sequences is also available in a tab-delimited file containing the GenBank accession number, country, sample year, PubMed ID, and assigned subtype/CRF (GSPS.Set.txt). Each of the 716 GSPS sequences is also included in a Fasta format (GSPS.Fasta.txt). The relationship between k and ADCL for each of the 155 subtype/CRF/country combinations with three or more sequences is listed in a tab-delimited file (ADCL.txt).

A phylogenetic tree containing the 716 GSPS sequences is provided in Newick tree format (GSPS.tre) and in pdf format (Figure 2). Phylogenetic trees for each subtype/CRF are also provided in pdf format (GSPS.subtype.tre.pdf.zip). These trees were constructed using neighbour joining with branch length optimized by maximum likelihood method with GTR evolution model using the R package phangorn and then rooted at the mid-point. Figure 3 contains a map in which countries are color-coded according to number of complete *pol* sequences available. For those subtype/CRFs with published time to MRCA (tMRCA), Figure 4 (which is described in the Technical Validation section) plots the tMRCA year vs. the median intra-subtype/CRF PWD in the complete set of 6,034 sequences.

Technical Validation

We performed three analyses to assess the how well the set of 716 GSPS sequences represent the diversity of published *pol* sequences. We also assess the potential usefulness of this set of sequences for subtyping. First, we determined whether the diversity within each subtype/CRF correlated with the age of the subtype/CRF as would be expected by viruses diverging from a common ancestor. Second, we compared the centrality of the GSPS sequences and NCBI Viral Genotyping Program reference sequences[8] within the complete *pol* data set. We also determined whether the subtype/CRF of the closest GSPS sequence was the same as the subtype/CRF of the closest NCBI sequence. Third, we compared the genetic distance of the first 1,300 bps of the GSPS sequences to nearly 7,000 1,300 bp *pol* sequences for which the COMET and Rega subtyping programs assigned the same (i.e., consensus) subtype.

Time to the most recent common ancestor (tMRCA) analysis

A strong correlation is expected between the diversity of sequences and tMRCA for each lineage as less divergence has taken place within lineages that originated more recently. To determine whether the diversity within each subtype/CRF in the 6,034 sequences from which the GSPS set was derived correlated with a published age of the subtype/CRF, we reviewed published papers that used molecular clock methods to estimate subtype and CRF tMRCAs. For those subtype/CRFs with published tMRCAs, we plotted the tMRCA year vs. the median intra-subtype/CRF PWD in the complete set of 6,034 sequences (Figure 4). Although the proportion of subtype/CRF sequences in the complete set is not representative of global epidemiology and the bias might have influenced the diversity of subtype/CRF in this set, there was a strong correlation between the published tMRCAs and the diversity of sequences in the complete set (Pearson's correlation coefficient; $r^2 = 0.78$; $p < 0.001$). Table 4 lists the published papers with the tMRCA estimates used in this analysis.

NCBI reference set comparison

To evaluate the centrality of the GSPS reference sequence set for distance-based subtyping, we compared the genetic distance of the GSPS sequences to the non-GSPS sequences in the complete *pol* data set with the genetic distances of the NCBI Viral Genotyping Program reference sequences[8] to the non-NCBI reference sequences in the complete *pol* data set.

we compared the centrality of the 716 GSPS sequences with the 316 subtype reference sequences used by the NCBI Viral Genotyping Tool[8]. For this comparison, we used 5,185 test sequences–the complete set of 6,034 sequences excluding those sequences that were in either the GSPS or NCBI reference set. We then identified the subtype/CRF of the sequences in GSPS set and in the NCBI set with the fewest nucleotide differences from each test sequence.

For 96.0% of the test sequences, the subtype of the closest GSPS and NCBI reference sequences were the same. The median distance to the closest GSPS and NCBI reference sequences were 2.9% (IQR: 2.1–3.7%) and 4.2% (IQR: 2.9–5.0%), respectively (Wilcoxon rank-sum test: $p < 0.001$). Thus, the GSPS set performed better at minimizing the distance to the test sequences. The increased centrality of the GSPS reference set may reflect the larger number of references in this set and/or the greater representativeness of its sequences.

Of the 4.0% ($n = 208$) of sequences for which the subtype/CRF of the closest matching GSPS and NCBI sequences were not identical, most were not true discordances. For 126 sequences, the GSPS and NCBI subtypes were the same for the *pol* gene (i.e., the NCBI subtype was one of the eight subtype/CRFs that were reclassified by us because they did not contain *pol* recombination points). For 21 sequences, the GSPS subtype/CRF was one of the most recent CRFs that was not yet included in the NCBI set.

The subtype/CRF classifications of the remaining 61 sequences (1.2% of 5,185) with discordant classifications are shown in Table 5. Each of these sequences was closest to a CRF in the NCBI Reference Set but closest to a pure subtype in the GSPS set. For 26 of these sequences, the NCBI classification was likely to be incorrect because the test sequences were from Sub-Saharan Africa but their closest reference sequence was either CRF08_BC that has been reported primarily in China[23] and CRF31_BC that has been reported primarily in Brazil[24]. The fact that NCBI CRF reference sequences were the closest sequence to likely pure subtype test sequences, suggests that the NCBI reference set may not have a sufficient number of representative pure subtype sequences. For the remaining 35 sequences, it was not possible to use isolate data to help infer the likely subtype/CRF.

Genetic distance to 1,300 bp sequences with a consensus COMET/Rega subtype

To evaluate the potential utility of the GSPS reference sequence set for distance-based subtyping for partial *pol* sequences, we downloaded all one-per-patient HIV-1 group M sequences ($n = 50,335$) encompassing the first 1300 bp of *pol* from LANL. We excluded the 6,034 complete *pol* sequences from the download because they were used to derive our GSPS set. We then randomly selected 1500 subtype B and 5500 non-subtype B sequences and aligned the first 1,300 bp of each of these sequences. For 6,115 (87.4%) of sequences, COMET and Rega assigned the same subtype/CRF. For the remaining 885 sequences, either COMET or Rega were discordant or reported that the sequence might have been a URF.

For 6,088 (99.6%) of the 6,115 sequences, the subtype/CRF of the closest sequence in the GSPS sequence set agreed with the consensus subtype/CRF assigned by COMET and Rega. The subtype/CRF classifications of the remaining 27 (0.4%) sequences with a discordance are shown in Table 6. The extraordinarily high concordance between the subtype/CRF of the closest GSPS sequence and the consensus subtype/CRF assigned by COMET/Rega likely reflects the fact that the test dataset contained only established subtypes and CRFs and did not contain URFs. We have not been able to identify any explanation for the few discordances listed in Table 6.

Usage Notes

The set of quality-controlled, one-per-person 6,034 complete *pol* sequences annotated by country and sample year and assigned a subtype/CRF will be useful for HIV-1 molecular epidemiology studies that use *pol* sequence data. The set of 716 GSPS sequences will be useful for HIV-1 *pol* subtyping programs.

The software pipeline used to generate this dataset will be useful to identify, retrieve, align, and organize the annotations for newly submitted HIV-1 *pol* sequences. With different parameters, the software can also be used to identify, retrieve, and organize the annotations for other HIV-1 and virus genes. GB_to_BLASTDB.pl creates a local GenBank database and facilitates the rapid retrieval of GenBank annotations. Gene_to_Sequenes.pl accepts any gene protein sequence and generates an alignment of the gene-coding nucleotide sequences with GenBank annotations. It provides an option for the method of GenBank search, TBLASTN or TFASTX.

References

1. Leitner, T., Korber, B., Daniels, M., Calef, C. & Foley, B. HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences. *Los Alamos National Laboratory* https://www.hiv.lanl.gov/content/sequence/HIV/COMPENDIUM/2005/partI/leitner.pdf (2005).

2. Foley, B. T., Leitner, T., Paraskevis, D. & Peeters, M. Primate immunodeficiency virus classification and nomenclature: Review. *Infect Genet Evol* **46,** 150–158 (2016).
3. Carr, J. K. *et al.* HIV-1 recombinants with multiple parental strains in low-prevalence, remote regions of Cameroon: evolutionary relics? *Retrovirology* **7,** 39 (2010).
4. Tongo, M., Dorfman, J. R. & Martin, D. P. High Degree of HIV-1 Group M (HIV-1M) Genetic Diversity within Circulating Recombinant Forms: insight into the Early Events of HIV-1M Evolution. *J Virol* **90,** 2221–2229 (2015).
5. Pineda-Pena, A. C. *et al.* Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infect Genet Evol* **19,** 337–348 (2013).
6. Struck, D., Lawyer, G., Ternes, A. M., Schmit, J. C. & Bercoff, D. P. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Res* **42,** e144 (2014).
7. Kosakovsky Pond, S. L. *et al.* An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS Comput Biol* **5,** e1000581 (2009).
8. Rozanov, M., Plikat, U., Chappey, C., Kochergin, A. & Tatusova, T. A web-based genotyping resource for viral sequences. *Nucleic Acids Res* **32,** W654–W659 (2004).
9. Kim, J., Ahn, Y., Lee, K., Park, S. H. & Kim, S. A classification approach for genotyping viral sequences based on multidimensional scaling and linear discriminant analysis. *BMC Bioinformatics* **11,** 434 (2010).
10. Chevenet, F., Jung, M., Peeters, M., de Oliveira, T & Gascuel, O. Searching for virus phylotypes. *Bioinformatics* **29,** 561–570 (2013).
11. Lapovok, I. *et al.* Short communication: molecular epidemiology of HIV type 1 infection in Kazakhstan: CRF02_AG prevalence is increasing in the southeastern provinces. *AIDS Res Hum Retroviruses* **30,** 769–774 (2014).
12. Junqueira, D. M. & Almeida, S. E. HIV-1 subtype B: Traces of a pandemic. *Virology* **495,** 173–184 (2016).
13. Tongo, M. *et al.* Unravelling the complicated evolutionary and dissemination history of HIV-1M subtype A lineages. *Virus Evol* **4,** vey003 (2018).
14. Matsen, F. A. t., Gallagher, A. & McCoy, C. O. Minimizing the average distance to a closest leaf in a phylogenetic tree. *Syst Biol* **62,** 824–836 (2013).
15. Rhee, S. Y. *et al.* Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res* **31,** 298–303 (2003).
16. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res* **45,** D37–D42 (2017).
17. National Center for Biotechnology Information (NCBI). BLAST https://www.ncbi.nlm.nih.gov/books/NBK1762/ (2016).
18. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85,** 2444–2448 (1988).
19. Gertz, E. M., Yu, Y. K., Agarwala, R., Schaffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol* **4,** 41 (2006).
20. Rhee, S. Y. *et al.* HIV-1 Protease, Reverse Transcriptase, and Integrase Variation. *J Virol* **90,** 6058–6070 (2016).
21. Pessoa, R. *et al.* Deep sequencing of HIV-1 near full-length proviral genomes identifies high rates of BF1 recombinants including two novel circulating recombinant forms (CRF) 70_BF1 and a disseminating 71_BF1 among blood donors in Pernambuco, Brazil. *PLoS ONE* **9,** e112674 (2014).
22. Hemelaar, J. *et al.* Global trends in molecular epidemiology of HIV-1 during 2000-2007. *AIDS* **25,** 679–689 (2011).
23. Feng, Y. *et al.* Geographic origin and evolutionary history of China's two predominant HIV-1 circulating recombinant forms, CRF07_BC and CRF08_BC. *Sci Rep* **6,** 19279 (2016).
24. Passaes, C. P. *et al.* Genetic characterization of HIV-1 BC recombinants and evolutionary history of the CRF31_BC in Southern Brazil. *Infect Genet Evol* **9,** 474–482 (2009).
25. Wertheim, J. O., Fourment, M. & Kosakovsky Pond, S. L. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. *Mol Biol Evol* **29,** 451–456 (2012).
26. Abecasis, A. B., Vandamme, A. M. & Lemey, P. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J Virol* **83,** 12917–12924 (2009).
27. Tongo, M. *et al.* Phylogenetics of HIV-1 subtype G env: Greater complexity and older origins than previously reported. *Infect Genet Evol* **35,** 9–18 (2015).
28. Liao, H. *et al.* Phylodynamic analysis of the dissemination of HIV-1 CRF01_AE in Vietnam. *Virology* **391,** 51–56 (2009).
29. Faria, N. R. *et al.* Phylodynamics of the HIV-1 CRF02_AG clade in Cameroon. *Infect Genet Evol* **12,** 453–460 (2012).
30. Tee, K. K. *et al.* Temporal and spatial dynamics of human immunodeficiency virus type 1 circulating recombinant forms 08_BC and 07_BC in Asia. *J Virol* **82,** 9206–9215 (2008).
31. Delatorre, E. & Bello, G. Time-scale of minor HIV-1 complex circulating recombinant forms from Central and West Africa. *BMC Evol Biol* **16,** 249 (2016).
32. Dilernia, D. A. *et al.* Analysis of HIV type 1 BF recombinant sequences from South America dates the origin of CRF12_BF to a recombination event in the 1970 s. *AIDS Res Hum Retroviruses* **27,** 569–578 (2011).
33. Delatorre, E. & Bello, G. Phylodynamics of the HIV-1 epidemic in Cuba. *PLoS ONE* **8,** e72448 (2013).
34. Ristic, N. *et al.* Analysis of the origin and evolutionary history of HIV-1 CRF28_BF and CRF29_BF reveals a decreasing prevalence in the AIDS epidemic of Brazil. *PLoS ONE* **6,** e17485 (2011).
35. Tee, K. K. *et al.* Estimating the date of origin of an HIV-1 circulating recombinant form. *Virology* **387,** 229–234 (2009).
36. Eybpoosh, S. *et al.* Spatio-Temporal History of HIV-1 CRF35_AD in Afghanistan and Iran. *PLoS ONE* **11,** e0156499 (2016).
37. Bello, G. *et al.* Phylodynamics of HIV-1 circulating recombinant forms 12_BF and 38_BF in Argentina and Uruguay. *Retrovirology* **7,** 22 (2010).
38. Struck, D. *et al.* Near Full-Length Characterization and Population Dynamics of the Human Immunodeficiency Virus Type I Circulating Recombinant Form 42 (CRF42_BF) in Luxembourg. *AIDS Res Hum Retroviruses* **31,** 554–558 (2015).
39. Delatorre, E. *et al.* Tracing the origin of a singular HIV-1 CRF45_cpx clade identified in Brazil. *Infect Genet Evol* **46,** 223–232 (2016).
40. Li, Y. *et al.* Identification of a novel second-generation circulating recombinant form (CRF48_01B) in Malaysia: a descendant of the previously identified CRF33_01B. *J Acquir Immune Defic Syndr* **54,** 129–136 (2010).
41. Foster, G. M. *et al.* Novel HIV-1 recombinants spreading across multiple risk groups in the United Kingdom: the identification and phylogeography of Circulating Recombinant Form (CRF) 50_A1D. *PLoS ONE* **9,** e83337 (2014).
42. Ng, K. T. *et al.* Phylodynamic profile of HIV-1 subtype B, CRF01_AE and the recently emerging CRF51_01B among men who have sex with men (MSM) in Singapore. *PLoS ONE* **8,** e80884 (2013).
43. Zhang, W. *et al.* Identification and characterization of a novel HIV-1 circulating recombinant form (CRF59_01B) identified among men-who-have-sex-with-men in China. *PLoS ONE* **9,** e99693 (2014).
44. Shcherbakova, N. S. *et al.* Short communication: Molecular epidemiology, phylogeny, and phylodynamics of CRF63_02A1, a recently originated HIV-1 circulating recombinant form spreading in Siberia. *AIDS Res Hum Retroviruses* **30,** 912–919 (2014).
45. Liu, Y. *et al.* Phylogenetic Analysis of HIV-1 CRF65_CPX Reveals Yunnan Province Is Still a Source Contributing to the Spread of HIV-1 in China. *J Acquir Immune Defic Syndr* **70,** e120–e122 (2015).

46. Hosaka, M. *et al.* HIV-1 CRF01_AE and Subtype B Transmission Networks Crossover: A New AE/B Recombinant Identified in Japan. *AIDS Res Hum Retroviruses* **32,** 412–419 (2016).

47. Cheong, H. T. *et al.* Genetic Characterization of a Novel HIV-1 Circulating Recombinant Form (CRF74_01B) Identified among Intravenous Drug Users in Malaysia: Recombination History and Phylogenetic Linkage with Previously Defined Recombinant Lineages. *PLoS ONE* **10,** e0133883 (2015).

Data Citation

1. Rhee, S.-Y. & Shafer, R. W. *Dryad Digital Repository* https://doi.org/10.5061/dryad.56j2n (2017).

Acknowledgements

S.-Y.R. and R.W.S. were supported by a grant from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number NIAID AI068581.

Author Contributions

S.-Y.R. and R.W.S. conceived of the study. S.-Y.R. wrote the software and performed the analyses described in the manuscript. S.-Y.R. and R.W.S. wrote the manuscript.

Additional Information

Tables 2 and 3 are available only in the online version of this paper.

Competing interests: The authors declare no competing interests.

How to cite this article: Soo-Yon Rhee & Robert W. Shafer. Geographically-stratified HIV-1 group M pol subtype and circulating recombinant form sequences. *Sci. Data* 5:180148 doi: 10.1038/sdata.2018.148 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.