AID-2016-0079-ver9-Lamers_1P
Type: other
SEQUENCE NOTES

AIDS RESEARCH AND HUMAN RETROVIRUSES
Volume 00, Number 00, 2016
[a] Mary Ann Liebert, Inc.
DOI: 10.1089/aid.2016.0079

HIV-1 Sequence Data Coverage in Central East Africa from 1959 to 2013

Susanna L. Lamers,[1] Andrew E. Barbier,[1] Oliver Ratmann,[2] Christophe Fraser,[2] Rebecca Rose,[1] Oliver Laeyendecker,[3,4] and Mary K. Grabowski[5]

Abstract

Central and Eastern African HIV sequence data have been most critical in understanding the establishment and evolution of the global HIV pandemic. Here we report on the extent of publicly available HIV genetic sequence data in the Los Alamos National Laboratory Sequence Database sampled from 1959 to 2013 from six African countries: Uganda, Kenya, Tanzania, Burundi, the Democratic Republic of Congo, and Rwanda. We have summarized these data, including HIV subtypes, the years sampled, and the genomic regions sequenced. We also provide curated alignments for this important geographic area in five HIV genomic regions with substantial coverage.

Subsaharan Africa (SSA) accounts for more cases of HIV than any other geographic region worldwide. However, there are relatively few published HIV phylogenetic studies from SSA compared with other regions of the world, including Europe and the United States. Recent recognition of this critical data deficit led to the establishment of the Phylogenetics and Networks for Generalized HIV Epidemics in Africa consortium (PANGEA-HIV) in 2014, which is now generating HIV sequences from 20,000 HIV-infected persons in countries in Eastern and Southern Africa.

Comprehensive data sets that include viral sequences from a wide range of geographic locations collected over many years can help phylogenetic studies achieve more representative samples,[1] improve resolution of focused phylogenetic studies by resolving discrete subepidemics, and provide novel information on the introduction and ongoing spread of HIV in human populations.[2] To compile such a data set, detailed knowledge of existing HIV sequence data is useful. Here, we reviewed the HIV Sequence Database at Los Alamos (www.hiv.lanl.gov) for historical HIV sequence data from the countries of Uganda, Tanzania, Kenya, the Democratic Republic of the Congo (DRC), Burundi, and Rwanda (RW). Our overall objective was to generate, as a first step, high-quality reference alignments of publicly available HIV sequence data from Eastern Africa. These reference alignments could be combined with newer HIV sequence data, such as that from PANGEA-HIV.

All HIV sequences from the six Central and East African countries of interest (n = 18,424) were downloaded along with the following annotations: Genbank accession number, sequence name, subtype, country, year, and sequence start and stop location according to HXB2 numbering and sequence length. We identified multiple clones within a gene, as well as data from multiple genomic regions of the same patient. One clone per subject was retained for a given gene region. A summary of the unique sequence data available and their sequence locations in the HIV genome is shown in Figure 1. Using these sequence data, we defined five HIV genomic regions with relatively high coverage: Region 1, gag (HXB2 location 700–2,100); Region 2, 5¢-pol domain (protease and RT genes, HXB2 location 2,240–3,900); Region 3, gp120 (HXB2 location 6,100–7,900); Region 4, gp41 (HXB2 location 7,900–8,200); and Region 5, the nearly full-length genome (HXB2 location 1–9,720). In comparison, the regulatory and accessory genes vif, vpr, vpu, 5¢-tat, 5¢-rev, nef, and the 3¢ pol regions covering the RNase and integrase genes had limited coverage.

We next downloaded all sequences from each of the six countries that were at least 250 nucleotides in length using the ''one sequence per patient'' option and defined sequence coordinates for each genomic region. The program ElimDupes (www.hiv.lanl.gov/content/sequence/ELIMDUPES/elimdupes.html) was used to identify and remove identical sequences that may have been submitted to the public

[1]Bioinfoexperts, LLC, Thibodaux, Louisiana.
[2]Medical Research Council Centre for Outbreak Analysis and Modeling, Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom.
[3]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Baltimore, Maryland.

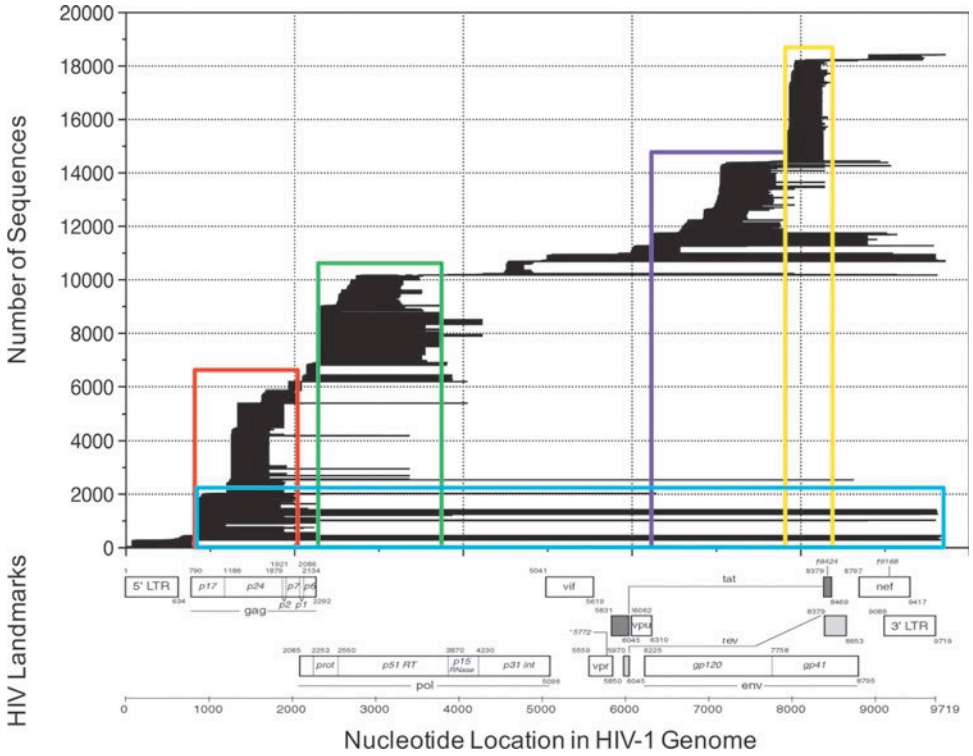

FIG. 1. Combined sequence coverage for Central East Africa. Individual sequences are represented by black horizontal bars that span their HIV genomic location (x-axis). Sequences are stacked by their start location according to HXB2-HIV genomic coordinates. The number of sequences is displayed on the y-axis. Landmarks of the HIV genome are shown below the graph (image from www.lanl.gov) along with their nucleotide location in the HXB2-HIV genome. Genomic regions with deep coverage are boxed as follows: Region 1 is red (700–2,100 bp), Region 2 is green (2,200–3,900 bp), Region 3 is purple (6,100–7,900 bp), Region 4 is yellow (7,900–8,200 bp), and Region 5 is blue (near full length HIV genomes).

databases under different accession numbers. The sequences for each country were then merged into a single file for Regions 1–5 (Supplementary Data are available online at www.liebertpub.com/aid) and codon aligned using Gene Cutter (www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html). The alignments were subsequently manually improved in Geneious software (version 7.0.6, Copyright©2005–2013 Biomatters Ltd.) to correct for obvious automatic alignment errors or nonhomologous data. No alignment was produced in four of the gp120 variable domains in Region 3 (V1, V2, V4, and V5).

The final regional alignments, which include Genbank accession numbers, sampling years, sequence coordinates, and country embedded in the sequence name, are available at www.hiv.lanl.gov/content/sequence/HIV/SI_alignments/datasets.html. The alignments are sorted by pure subtypes (A, B, C, D, F, G, H, J, K, U), followed by sub-subtypes (i.e., A1 and A2) and recombinant sequences. We also provide summary information for each regional alignment, including the number of sequences along with their subtypes (Table 1), the number of sequences originating from each country (Table 2), an alignment graphic that includes overall percentage pairwise identity, percentage identical sites, and coverage (Fig. 2), and the number sequences available for each year (Fig. 3).

As of 2013, most published viral sequence data from Eastern Africa come from Kenya and Uganda (73.6%) (Table 2), and 97.4% of all sequence data were collected after 1990 (Fig. 3). The greatest number of HIV sequences generated in a single year was in 1995. The majority (n = 1,517) of the Ugandan sequence data come from a single population-

Table 1. Subtype Distribution in Regional Alignments

Genomic region	A	B	C	D	F	G	H	J	K	U	Others[a]	Total number of sequences
1	2,631	2	515	2,624	20	113	20	10	11	39	547	6,483
2	3,933	6	1,035	1,865	2	43	5	3	1	33	1,260	8,207
3	2,847	3	712	955	22	114	48	20	15	50	540	5,326
4	2,268	3	301	2,142	10	56	6	10	3	17	336	5,152
5	99	0	27	44	0	3	0	0	1	2	143	319

Table 2. Country Distribution in Regional Alignments

Genomic region	Burundi	DRC	Kenya	Rwanda	Tanzania	Uganda
1	36	454	1,959	69	821	3,144
2	384	215	2,561	344	1,095	3,608
3	109	846	2,121	266	1,192	792
4	9	230	989	281	270	3,373
5	0	27	145	11	49	87

based study in Rakai District, Uganda, in 1995, the Rakai Community Cohort Study (RCCS).[3] Additional sequences from the RCCS (and in some cases from the same RCCS participants) are also available in large numbers in more recent years (2002–2003, 2008–2009).[4,5] Many (60.4%) Kenyan sequences were obtained from women and children with high exposure to HIV in the Pumwani area of Nairobi, Kenya.[6–8] Sequences from the DRC include multiple HIV subtypes, as studies in this region have frequently focused on the varied recombining subtypes in the country.[9] A number of sequences (n = 608, 17.2%) from Tanzania were from 428 infected pregnant women.[10] Many of the sequences from Burundi (n = 220, 39%) were obtained from samples collected during a single surveillance study of 119 individuals living in urban and rural districts.[11]

Notably, some of the oldest sequences from SSA were excluded from our final alignments due to their short length. These include multiple sequences from one individual that span <170 bp of env and pol, which were derived from stored plasma from a subject who died in 1959 from AIDS-like illnesses (Accession numbers AF030667–AF030686),[12] and five short sequences (<82 bp) generated from a paraffin-embedded lymph node sample that was collected in 1960 from the DRC (Accession numbers EU580739, EU589218,

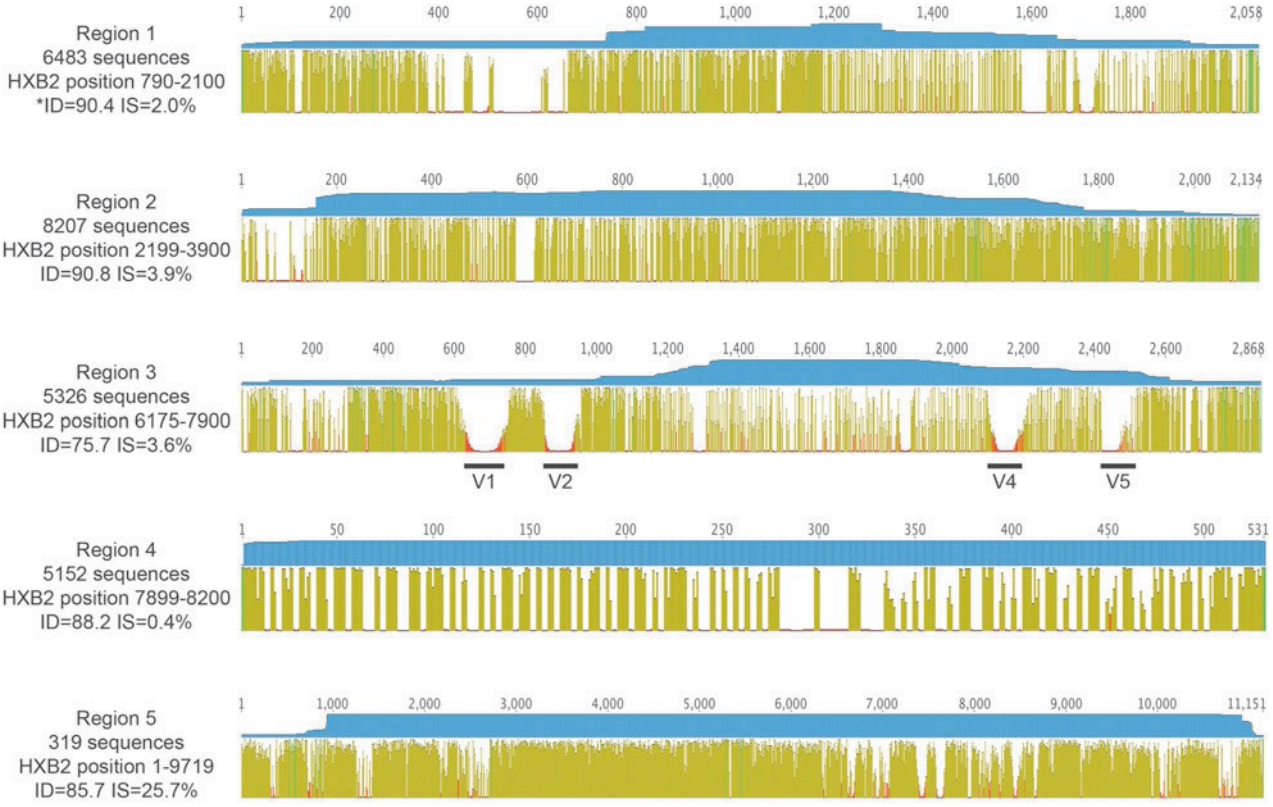

FIG. 2. Alignment coverage and identity. For Regions 1–5 identified in Figure 1, an alignment was generated. The final number of sequences, the HIV region covered by each alignment according to HXB2-HIV numbering, the percentage pairwise sequence identity (ID), and percentage IS for the overall alignment are shown to the left for each region. The graphic is numbered for each region according to the final alignment length. Each alignment is composed of sequences of variable length, as represented by the top blue horizontal coverage bar, which portrays the number of nonend nucleic acid characters at each position along the alignment. For Region 1, the maximum height of the coverage bar at any site is 5,947 nucleotides, which indicates that there are 536 nonoverlapping sequences in this alignment (total number of sequences—minus maximum coverage at any position), for Region 2 the maximum coverage is 7,739 nucleotides (468 nonoverlapping sequences), for Region 3 the maximum coverage is 5,023 (303 nonoverlapping sequences), for Region 4 the maximum coverage is 5,208 nucleotides (all sequences overlap), and for Region 5, the maximum coverage is 5,152 nucleotides (all sequences overlap). Below the coverage graphic, the mean pairwise identity over all pairs in each column of the alignment correlates with the height of each vertical bar along the length of the sequence and is colored as follows: dark green, 100%

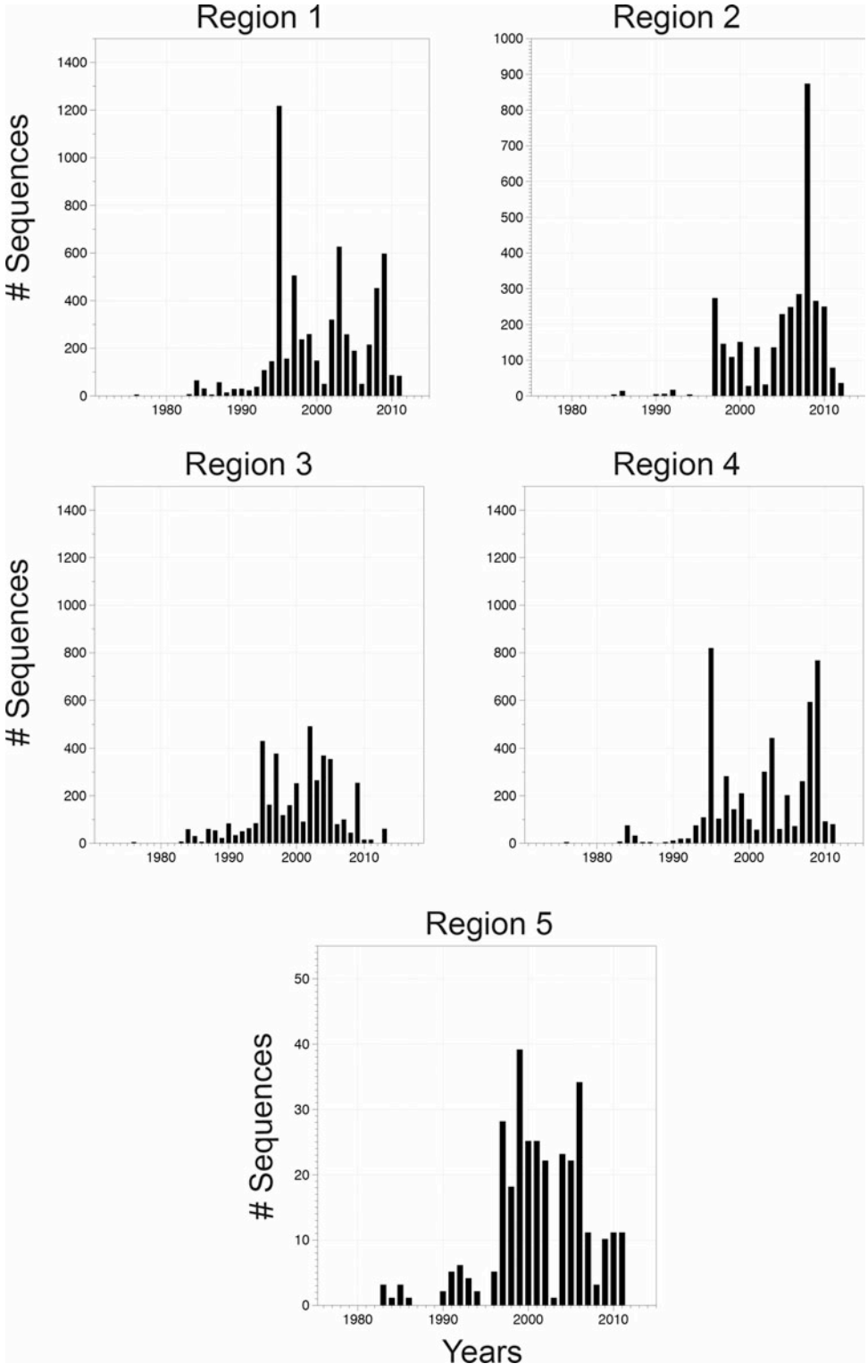

FIG. 3. The number of sequences and years sampled in each regional alignment.

EU580803, EU580840, EU580849).[13] The oldest sequences included in the alignments (Regions 1–4) are molecularly cloned 1976 Zaire isolates, which have been used to study the evolutionary divergence of HIV in Africa previously (Accession numbers U76035, M15896).[14]

In summary, we provide curated alignments of existing countries. In the process of creating these alignments, we identified gaps in HIV sequence information that could be addressed going forward. Notable deficits include data before the 1990s. Data from Tanzania, Burundi, and Rwanda are particularly sparse. In addition, most HIV sequence data within countries come from only a few

PANGEA-HIV may help to address some of these data gaps moving forward.

This research was supported, in part, by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases. The authors would like to thank Brian T. Foley at The Los Alamos HIV Sequence Data base and James J. Dollar at the University of Florida for assistance with codon-based alignments.

Author Disclosure Statement

No competing financial interests exist.

References

1. Frost SD, et al.: Eight challenges in phylodynamic inference. Epidemics 2015;10:88–92.
2. Faria NR, et al.: HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 2014;346:56–61.
3. Collinson-Streng AN, et al.: Geographic HIV type 1 subtype distribution in Rakai district, Uganda. AIDS Res Hum Retroviruses 2009;25:1045–1048.
4. Grabowski MK, et al.: The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: Evidence from spatial clustering, phylogenetics, and egocentric transmission models. PLoS Med 2014;11: e1001610.
5. Conroy SA, et al.: Changes in the distribution of HIV type 1 subtypes D and A in Rakai District, Uganda between 1994 and 2002. AIDS Res Hum Retroviruses 2010;26:1087–1091.
6. Peters HO, et al.: An integrative bioinformatic approach for studying escape mutations in human immunodeficiency virus type 1 gag in the Pumwani Sex Worker Cohort. J Virol 2008;82:1980–1992.
7. Lwembe R, et al.: Changes in the HIV type 1 envelope gene from non-subtype B HIV type 1-infected children in Kenya. AIDS Res Hum Retroviruses 2009;25:141–147.
8. Land AM, et al.: Human immunodeficiency virus (HIV) type 1 proviral hypermutation correlates with CD4 count in HIV-infected women from Kenya. J Virol 2008;82:8172–8182.
9. Kalish ML, et al.: Recombinant viruses and early global HIV-1 epidemic. Emerg Infect Dis 2004;10:1227–1234.
10. VasanA, et al.: Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype. Clin Infect Dis 2006;42:843–852.
11. Vidal N, et al.: HIV type 1 diversity and antiretroviral drug resistance mutations in Burundi. AIDS Res Hum Retroviruses 2007;23:175–180.
12. Zhu T, et al.: An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature 1998;391: 594–597.
13. Worobey M, et al.: Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 2008;455:661–664.
14. Srinivasan A, et al.: Molecular characterization of HIV-1 isolated from a serum collected in 1976: Nucleotide sequence comparison to recent isolates and generation of hybrid HIV. AIDS Res Hum Retroviruses 1989;5:121–129.

Address correspondence to:
Susanna L. Lamers
Bioinfoexperts, LLC
Thiboduax, LA 70301

E-mail: susanna@bioinfox.com