Inferences about the global scenario of human T-cell lymphotropic virus type 1 infection using data mining of viral sequences

Thessika Hialla Almeida Araujo[1], Fernanda Khouri Barreto[1], Luiz Carlos Júnior Alcântara[1]/+, Aline Cristina Andrade Mota Miranda[2]

[1]Centro de Pesquisa Gonçalo Moniz-Fiocruz, Salvador, BA, Brasil
[2]Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador, BA, Brasil

Human T-cell lymphotropic virus type 1 (HTLV-1) is mainly associated with two diseases: tropical spastic paraparesis/HTLV-1-associated myelopathy (TSP/HAM) and adult T-cell leukaemia/lymphoma. This retrovirus infects five-10 million individuals throughout the world. Previously, we developed a database that annotates sequence data from GenBank and the present study aimed to describe the clinical, molecular and epidemiological scenarios of HTLV-1 infection through the stored sequences in this database. A total of 2,545 registered complete and partial sequences of HTLV-1 were collected and 1,967 (77.3%) of those sequences represented unique isolates. Among these isolates, 93% contained geographic origin information and only 39% were related to any clinical status. A total of 1,091 sequences contained information about the geographic origin and viral subtype and 93% of these sequences were identified as subtype "a". Ethnicity data are very scarce. Regarding clinical status data, 29% of the sequences were generated from TSP/HAM and 67.8% from healthy carrier individuals. Although the data mining enabled some inferences about specific aspects of HTLV-1 infection to be made, due to the relative scarcity of data of available sequences, it was not possible to delineate a global scenario of HTLV-1 infection.

Key words: HTLV-1 - data mining - HTLV-1 database

Human T-cell lymphotropic virus type 1 (HTLV-1) was the first described human retrovirus (Poiesz et al. 1980). This retrovirus is the causative agent of tropical spastic paraparesis/HTLV-1-associated myelopathy (TSP/HAM) (Gessain et al. 1985), adult T-cell leukaemia/lymphoma (ATL) (Yoshida et al. 1982) and other inflammatory diseases such as HTLV-1-associated infectious dermatitis (La Grenade et al. 1998) and HTLV-1-associated uveitis (Mochizuki et al. 1992). However, the pathogenesis of some clinical manifestations is not yet fully understood.

Epidemiological data show that HTLV-1 has a worldwide distribution and it is estimated that five-10 million people are infected (Gessain & Cassar 2012). This infection is endemic in southwestern Japan (Mueller et al. 1996), sub-Saharan Africa (Gessain & de Thé 1996), regions of the Caribbean (Hanchard et al. 1990) and minor areas in Iran, Melanesia (Mueller 1991) and Brazil (Galvão-Castro et al. 1997).

Regardless, HTLV-1 epidemiology still presents many challenges. Virus prevalence rates have been correlated with geographic characteristics and the social setting of destitute populations. However, these populations are not frequently the target of great public and government interest (Galvão-Castro et al. 1997). Molecular studies, especially during the late decade, have contributed to the acquisition of knowledge about virus epidemiology and the molecular characteristics. Furthermore, a great amount of viral sequences are generated from these molecular studies because of this appropriate data management and data mining can provide additional consistent information about HTLV-1 infection.

In response to the need of obtaining more information about the already generated and available HTLV-1 sequences, HTLV-1 Molecular Epidemiology Database (htlv1db.fiocruz.bahia.br) was developed (Araujo et al. 2012). This database contains information that can support our understanding of viral pathogenesis, the route of transmission, polymorphisms, epidemiology, genotype-phenotype relationships, geographic distribution and viral evolution. Therefore, the purpose of the present study was to assess the different types of information deposited in HTLV-1 Molecular Epidemiology Database to describe clinical, molecular and epidemiological scenarios about HTLV-1 infection.

MATERIALS AND METHODS

This is a descriptive study about the clinical, molecular and epidemiological data of HTLV-1 infection that are associated with the stored genetic sequences in HTLV-1 Molecular Epidemiology Database.

All the descriptive analyses were performed using the search algorithm implemented at the HTLV-1 database (Araujo et al. 2012). Initially, we made a list with the variables (age, gender, clinical status, subtype, subgroup, geographic origin) that were more frequent in the database. We then performed combinations with the listed variables; for example, we searched for sequences with information about geographic origin, viral subtype and clinical status. These combinations constitute the

doi: 10.1590/0074-0276130587
+ Corresponding author: alcan@bahia.fiocruz.br
Received 18 December 2013
Accepted 7 March 2014

section subheadings of the Results and Discussion section. The HTLV-1 database allows all the search results to be organised as spread sheets for further analyses.

After the search step and the generation of spread sheets containing the results, we used the Excel program to perform descriptive analyses.

RESULTS AND DISCUSSION

Currently, HTLV-1 Molecular Epidemiology Database stores 2,545 HTLV-1 sequences, 1,967 (77.3%) of which represent different isolates. These 1,967 sequences were selected for this study and 91 (3.6%) other distinct sequences, which did not have information about the viral isolate, were also included. Ultimately, 2,058 sequences and their data were analysed.

Geographic origin, viral subtype and clinical status among viral sequences - Among the 2,058 viral sequences, 1,914 (93%) were associated with geographic origin in the GenBank notes. Fig. 1 shows the distribution of HTLV-1 sequences among different geographic regions: 1.6% of the sequences originated from HTLV-1 isolates from North America, 2.4% from Oceania, 3% from Europe, 3.3% from Central America, 17.7% from Africa, 32% from Asia and 40% from South America. With regard to the South America sequences, most of the isolates were from HTLV-1 infections in Brazil (55%) and Argentina (22.1%), as shown in Fig. 2.

Although HTLV-1 infection has a cosmopolitan geographic distribution, it has a heterogeneous distribution, such that Asia and South America are characterised as endemic areas (Proietti et al. 2005, Carneiro-Proietti et al. 2006). This heterogeneous distribution is also represented in the distribution of sequences available in GenBank and in the number of exploratory studies of HTLV-1 infection developed in each geographical region.

Several studies have reported a high prevalence of HTLV infection in Africa (Proietti et al. 2005); however, there are few sequences about this geographic region deposited in GenBank. This profile is frequent and emphasise that it is necessary to increase the use of molecular data as an important tool of epidemiology investigation. This result suggests that it is necessary to create new health units that could be able to perform molecular diagnosis and, therefore, could generate and record molecular data about new HTLV-1 cases. Nonetheless, the high prevalence of HTLV infection in Asia and parts of South America (Carneiro-Proietti et al. 2006, Sonoda et al. 2011) corroborates the high amount of sequences assembled in the database. However, it is possible to observe a lack of clinical and epidemiological information in the GenBank annotations. This observation shows that the authors should provide the maximum amount of information as possible because the molecular data could be useful for many different inferences about HTLV-1 infection and, therefore, useful for encouraging the politics of prevention.

The search about the number of HTLV-1 sequences with geographic origin and viral subtype showed that 1,091 contained information for both variables. The results showed that 1,019 (93.4%) sequences were classified as subtype "a" and that this subtype had a worldwide distribution among the sequences deposited in the database. This higher prevalence can be attributed to the fact that it is the worldwide subtype found especially in Japan, the Caribbean, South America and Africa. Subtypes "b" (4.9%), "c" (0.5%), "d" (0.6%), "e" (0.1%), "f" (0.2%) and "g" (0.2%) were distributed in specific regions (Fig. 3). These subtypes are usually restricted to certain areas, such as subtype "c", found in Australia-Melanesia (Galvão-Castro et al. 1997).

Finally, using the geographic origin and viral subtype, we performed a search for clinical status. It was possible to identify that 279 sequences had information for these three variables in the GenBank annotations. Regarding these sequences, 35.8% originated from Asia, 32.2% from South America, 14.3% from Africa, 10% from Central America, 2.8% from Europe, 4.3% from North America and 0.3% from Oceania. Our analyses showed that 86.3%

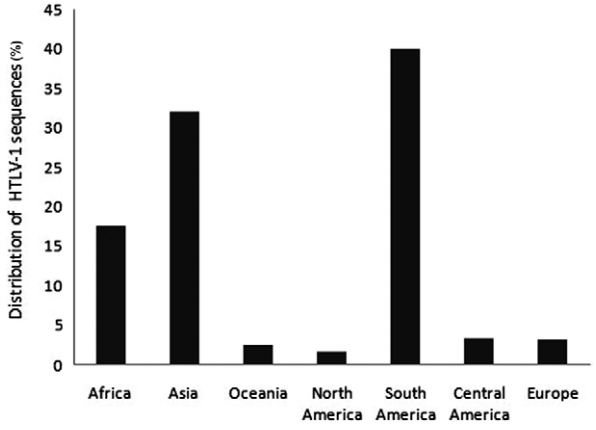

Fig. 1: geographical distribution of 2,058 sequences stored in the HTLV-1 Molecular Epidemiology Database.

Fig. 2: distribution (%) of human T-cell lymphotropic virus type 1 sequences among the countries in South America.

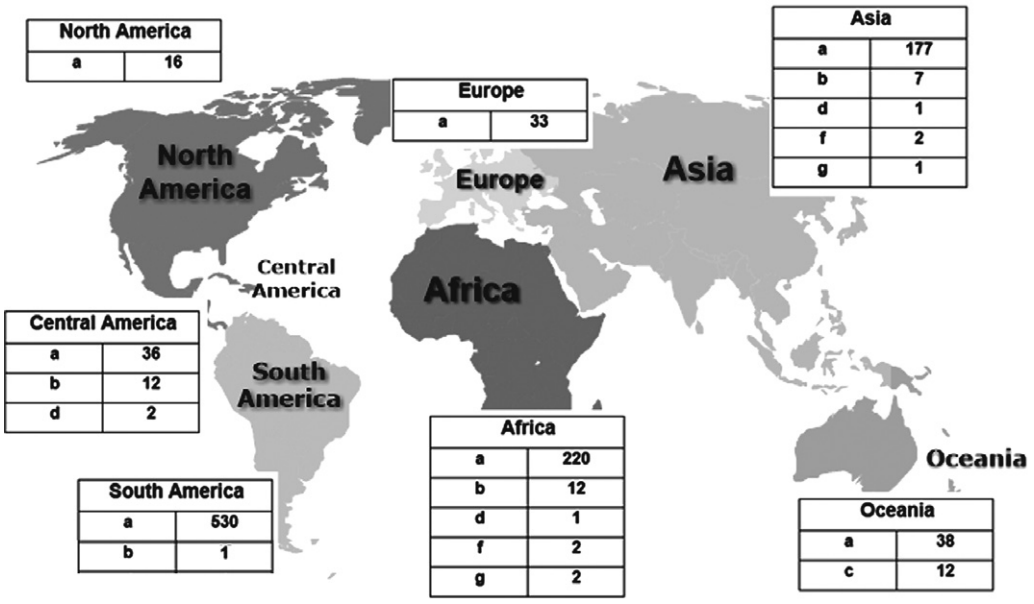

Fig. 3: geographical distribution of absolute number of human T-cell lymphotropic virus type 1 (HTLV-1) sequences stored in the HTLV-1 Molecular Epidemiology Database, using the subtype variable.

sequences were identified as subtype "a", 12.5% sequences were subtype "b" and 1.1% were subtype "d". Concerning the clinical status, 158 sequences (subtypes "a", 88.6%; "b", 10.7%; "d", 0.6%) were derived from HTLV-1 infection in healthy carrier (HC) individuals, 71 sequences (subtypes "a", 80.2%; "b", 18.3%; "c", 1.5%) were from TSP/HAM individuals and 34 sequences (subtypes "a", 91.2%; "b", 8.8%) were from ATL individuals; 19 viral sequences (subtypes "a", 84.2%; "b", 10.5%; "d", 5.3%) were related to other diseases not yet described.

Clinical status and viral subtype and subgroup of viral sequences from South America - Using HTLV-1 Molecular Epidemiology Database enabled the observation of the overall global scenario of HTLV-1 sequences from South America. From all the collected sequences from South America, 77 were from HC individuals, 63 from TSP/HAM individuals and 12 from ATL individuals; six viral sequences from individuals with other diseases. However, 609 sequences did not have any information about the clinical status in the GenBank annotations. As most of the studies of HTLV-1 infection in South America have been developed in Argentina and Brazil, the greatest number of sequences and, therefore, the greatest number of molecular and epidemiological data are also generated from these regions. Because of this fact, it is important to emphasise the need of developing new exploratory studies about HTLV-1 infection in other countries in South America.

Only 77 of the sequences from HTLV-1 cases of infection in South America had information about the clinical status, viral subtype and subgroup, at the same time, in the GenBank annotation. Among these sequences, 18 were from TSP/HAM individuals, four sequences were from ATL individuals and six sequences were from individuals with other HTLV-associated diseases. The greatest number of sequences was generated from HC individuals (n = 49). All of the 77 sequences were identified, in the GenBank annotation, as subtype "a", which was the subtype most found among infected individuals in South America. With regard to the subgroup classification, 70 sequences were identified as subgroup "a" (n = 48 HC, n = 2 ATL, n = 15 TSP/HAM, n = 5 other diseases), three sequences as subgroup "b" (n = 2 TSP/HAM, n = 1 ATL) and three sequences as subgroup "c" (n = 1 ATL, n = 1 HC, n = 1 TSP/HAM); only one sequence was identified as subgroup "e" (lymphoma).

Clinical status, age, gender, viral subtype and ethnicity among viral sequences - The investigation about clinical status separately showed that 797 of the stored sequences were related to one HTLV-1-associated clinical status (TSP/HAM, 43%; ATL, 19%; HC, 32.39%; other diseases, 5.61%). The other HTLV-1-associated diseases, such as dermatitis and sicca syndrome, were not reported in any annotation of the stored viral sequences. Using the information about the clinical status, we searched for additional data such as gender, age, viral subtype and ethnicity (Table). Approximately 15.2% of the sequences contained information about the infected patient's gender and 10.8% of the sequences provided the age of the infected patient.

The data about ethnicity were very scarce, as only 41 (5%) of the 797 stored sequences had information about ethnic origin in the GenBank annotation. Approximately 78% of the sequences (n = 41) had information about gender, clinical status, viral subtype, ethnicity and age at the same time. All originated from women infected by one HTLV-1 subtype "a" isolate and 27.5% of these women were younger than 40 years old. Regarding clinical status, 29% of the sequences originated from TSP/HAM women and 67.8% sequences from HC women; 3.2% sequences were generated from infected women with other HTLV-1-associated diseases.

TABLE

Distribution of gender, subtype and geographic origin among the clinical status: tropical spastic paraparesis/human T-cell lymphotropic virus type 1 (HTLV-1)-associated myelopathy (TSP/HAM), adult T-cell leukaemia/lymphoma (ATL), healthy carrier (HC) and other HTLV-1 associated disease

Clinical status	TSP/HAM n = 343 n (%)	ATL n = 137 n (%)	HC n = 261 n (%)	Others[a] n = 45 n (%)
Sequences with information about sex	40 (11.6)	13 (9.4)	89 (34)	16 (35.5)
Male	15 (37.5)	6 (46.2)	33 (37)	8 (50)
Female	25 (62.5)	7 (53.8)	56 (63)	8 (50)
Sequences with information about subtype	71 (20.7)	36 (26.3)	158 (60.5)	19 (42.2)
Subtype				
"a"	57 (80.2)	33 (91.6)	140 (88.6)	16 (84.2)
"b"	13 (18.4)	3 (8.4)	17 (10.8)	2 (10.5)
Other	1 (1.4)	0 (0)	1 (0.6)	1 (5.3)
Sequences with information about geographic origin	320 (93.4)	128 (93.4)	260 (99.6)	45 (100)

a: sequences from patients with either infective dermatitis, histoplasmosis, stroke, seborrhoea dermatitis, leprosy, nonspastic paraparesis, facial palsy or leg paresis.

A great number of studies show that some ethnically defined factors are likely to be associated with HTLV-1 persistence and the development of ATL or TSP/HAM among HTLV-1 endemic populations. Therefore, this information should be further investigated in HTLV-1 infection cases. Furthermore, new studies about the genetic background of infected individuals by the analysing polymorphic determinants of human leukocyte antigen alleles and their immune responsiveness to HTLV-1 are important points in the approach of the ethnic factors involved in HTLV-1 clustering and the disease segregation of ATL and TSP/HAM (Sonoda et al. 2011).

HTLV-1 Molecular Epidemiology Database enabled some inferences about the specific aspects of HTLV-1 infection. However, due to the relative scarcity of data about the available sequences, it was not possible to delineate a global scenario of HTLV-1 infection. Molecular and epidemiological data for viral sequences should be offered more frequently because this information can be used for planning public health policies.

REFERENCES

Araujo THA, Souza-Brito LI, Libin P, Deforche K, Edwards D, Albuquerque-Junior AE, Vandamme AM, Galvão-Castro B, Alcântara LCJ 2012. A public HTLV-1 molecular epidemiology database for sequence management and data mining. *PLoS ONE 7*: e42123.

Carneiro-Proietti ABF, Catalan-Soares BC, Castro-Costa CM, Murphy EL, Sabino EC, Hisada M, Galvão-Castro B, Alcântara LCJ, Remondegui C, Verdonck K, Proietti FA 2006. HTLV in the Americas: challenges and perspectives. *Rev Panam Salud Publica 19*: 44-53.

Galvão-Castro B, Loures L, Rodrigues LG, Sereno A, Ferreira OC, Franco LGP, Muller M, Sampaio DA, Santana A, Passos LM, Proietti F 1997. Distribution of human T-lymphotropic virus type I among blood donors: a nationwide Brazilian study. *Transfusion 37*: 242-243.

Gessain A, Barin F, Vernant JC, Gout O, Maurs L, Calender A, de Thé G 1985. Antibodies to human T lymphotropic virus type I in patients with tropical spastic paraparesis. *Lancet 2*: 407-409.

Gessain A, Cassar O 2012. Epidemiological aspects and world distribution of HTLV-1 infection. *Front Microbiol 3*: 388.

Gessain A, de Thé G 1996. Geographic and molecular epidemiology of primate T lymphotropic retroviruses: HTLV-I, HTLV-II, STLV-I, STLVP and PTLV-L. *Adv Virus Res 47*: 377-426.

Hanchard B, Gibbs WN, Lofters W, Campbell M, Williams E, Williams N, Jaffe E, Cranston B, Panchoosingh LD, Blattner WA, Manns A 1990. *Human retrovirology: HTLV*, Raven Press, New York, p. 173-183.

La Grenade L, Manns A, Fletcher V, Derm D, Carberry C, Hanchard B, Maloney EM, Cranston B, Williams NP, Wilks R, Kang EC, Blattner WA 1998. Clinical, pathologic and immunologic features of human T lymphotropic virus type I-associated infective dermatitis in children. *Arch Dermatol 134*: 439-444.

Mochizuki M, Watanabe T, Yamaguchi K, Tajima K, Yoshimura K, Nakashima S, Shirao M, Araki S, Miyata N, Mori S 1992. Uveitis associated with human T lymphotropic virus type I: seroepidemiologic, clinical and virologic studies. *J Infect Dis 166*: 943-944.

Mueller N 1991. The epidemiology of HTLV-1 infection. *Cancer Causes Control 2*: 37-52.

Mueller N, Okayama A, Stuver S, Tachibana N 1996. Findings from the Miyazaki Cohort Study. *J Acquir Immune Defic Syndr Hum Retrovirol 13* (Suppl. 1): S2-S7.

Poiesz BJ, Ruscetti FW, Gazdar AF, Bunn PA, Minna JD, Gallo RC 1980. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc Natl Acad Sci USA 77*: 7415-7419.

Proietti FA, Carneiro-Proietti ABF, Catalan-Soares BC, Murphy EL 2005. Global epidemiology of HTLV-I infection and associated diseases. *Oncogene 24*: 6058-6068.

Sonoda S, Li HC, Tajima K 2011. Ethnoepidemiology of HTLV-1 related diseases: ethnic determinants of HTLV-1 susceptibility and its worldwide dispersal. *Cancer Sci 102*: 295-301.

Yoshida M, Miyoshi I, Hinuma Y 1982. Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. *Proc Natl Acad Sci USA 79*: 2031-2035.