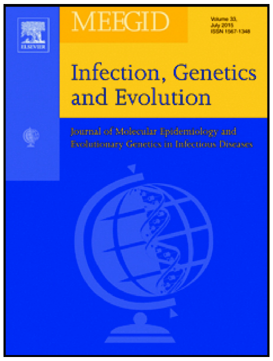
Complete genome sequence of human T-cell lymphotropic type 1 from patients with different clinical profiles, including infective dermatitis


MEEGID
Infection, Genetics and Evolution
Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases

Thessika Hialla Almeida Araújo, Fernanda Khouri Barreto, Aline Dórea Luz Menezes, Clayton Pereira Silva de Lima, Rodrigo Santos de Oliveira, Poliana da Silva Lemos, Bernardo Galvão-Castro, Simone Kashima, Lourdes Farre, Achilea Lisboa Bittencourt, Edgar Marcelino de Carvalho, Luciane Amorim Santos, Filipe Ferreira de Almeida Rego, Aline Cristina Andrade Mota-Miranda, Márcio Roberto Teixeira Nunes, Luiz Carlos Júnior Alcântara

PII: S1567-1348(19)30392-2

DOI: https://doi.org/10.1016/j.meegid.2019.104166

Reference: MEEGID 104166

To appear in: *Infection, Genetics and Evolution*

Received date: 20 September 2019

Revised date: 16 December 2019

Accepted date: 24 December 2019

Please cite this article as: T.H.A. Araújo, F.K. Barreto, A.D.L. Menezes, et al., Complete genome sequence of human T-cell lymphotropic type 1 from patients with different clinical profiles, including infective dermatitis, *Infection, Genetics and Evolution*(2019), https://doi.org/10.1016/j.meegid.2019.104166

Complete genome sequence of human T-cell lymphotropic type 1 from patients with different clinical profiles, including infective dermatitis

Thessika Hialla Almeida Araújo[1,3 ¶]; Fernanda Khouri Barreto[1,5 ¶]; Aline Dórea Luz Menezes[1]; Clayton Pereira Silva de Lima[2]; Rodrigo Santos de Oliveira[2]; Poliana da Silva Lemos[2]; Bernardo Galvão-Castro[3]; Simone Kashima[4]; Lourdes Farre [1,7]; Achilea Lisboa Bittencourt[5]; Edgar Marcelino de Carvalho[1,5]; Luciane Amorim Santos[1,3]; Filipe Ferreira de Almeida Rego[6]; Aline Cristina Andrade Mota-Miranda[5]; Márcio Roberto Teixeira Nunes[2]; Luiz Carlos Júnior Alcântara[1,*].

[1] Fundação Oswaldo Cruz, Brasil

[2] Instituto Evandro Chagas, Brasil

[3] Escola Bahiana de Medicina e Saúde Pública Salvador, Brasil /

[4] Fundação Hemocentro de Ribeirão Preto, Brasil

[5] Universidade Federal da Bahia, Brasil

[6] Universidade Católica do Salvador, Brasil

[7] Catalan Institute of Oncology. Bellvitge Biomedical Research Institute, Barcelona, Spain

[*]Corresponding author

E-mail: luiz.alcantara@ioc.fiocruz.br

Phone: +55 21 25621707
Fax: +55 21 25621779

[¶]These authors contributed equally to this work.

Abstract

The HTLV-1 is the first human retrovirus and is associated with several clinical syndromes, however, the pathogenesis of these clinical manifestations is still not fully understood. Furthermore, there are few complete genomes publicly available, about 0.12 complete genomes per 10,000 infected individuals and the databases have a major deficiency of sequences information. This study generated and characterized 31 HTLV-1 complete genomes sequences derived from individuals with Tropical Spastic Paraparesis/HTLV-1-Associated Myelopathy (TSP/HAM), Adult T-cell leukemia/lymphoma (ATL), infective dermatitis associated to HTLV-1 (IDH) and asymptomatic patients. These sequences are associated to clinical and epidemiological information about the patients. The sequencing data generated on Ion Torrent PGM platform were assembled and mapped against the reference HTLV-1 genome. These sequences were genotyped as Cosmopolitan subtype, Transcontinental subgroup. We identified the variants in the coding regions of the genome of the different clinical profiles, however, no statistical relation was detected. This study contributed to increase of HTLV-1 complete genomes in the world. Furthermore, to better investigate the contribution of HTLV-1 mutations for the disease outcome it is necessary to evaluate the interaction of the viral genome and characteristics of the human host.

Keywords: HTLV-1; complete genome; nucleotide variations

Introduction

Nearly four decades ago, Human T-Cell Lymphotropic Virus Type 1 (HTLV-1) was isolated (Poiesz et al., 1980). HTLV-1 was the first human retrovirus to be described and has been associated with several diseases: Tropical Spastic Paraparesis/HTLV-1-Associated Myelopathy (TSP/HAM) (Gessain et al., 1985), Adult T-cell leukemia/lymphoma (ATL) (Yoshida et al., 1982) and other inflammatory diseases, such as infective dermatitis associated to HTLV-1 (IDH) (La Grenade et al., 1998).

It has been estimated that approximately 5-10 million people are infected with HTLV-1 worldwide. Its distribution is heterogeneous and is mainly concentrated in endemic regions, such as Southwest Japan, the Caribbean Islands, South America and Central Africa (Gessain and Cassar, 2012). Studies have shown that HTLV-1 is endemic in some areas of Brazil, with variable prevalence according to geographic region and population (Carneiro-Proietti et al., 2012). A population-based study revealed that Salvador, the capital of the state of Bahia, has a prevalence of 1.76% (reaching 8.4% in women aged 51 years or older). Besides an elevated prevalence in women, the virus has also been associated with lower education and depressed income levels, in addition to areas with poor indicators of socioeconomic status (Dourado et al., 2003).

The HTLV-1 genome containing the genes *gag*, *pol* and *env*, in addition to the pX region, is flanked by two long terminal repeat (LTR) regions at both 5' and 3' ends. Many enhancer/promoter genetic elements are found in these regions, which are critical in viral RNA transcription. High genetic variability of LTR regions has proven useful in molecular epidemiology studies to classify HTLV-1 isolates in seven subtypes (a-g) described until this date (Neto et al., 2011). The Cosmopolitan subtype (a) is distributed worldwide with its molecular diversity giving rise to five subgroups: Transcontinental

(A), Japanese (B), West African (C), North African (D), Black Peruvian (E) and Ethiopian (F) (Gessain and Cassar, 2012).

Although the HTLV-1 has been the first described human retrovirus, the pathogenesis underlying some clinical manifestations of this infection remains obscure. Although many studies have attempted to associate specific gene mutations with infection outcomes, there are few complete genomes publicly available (0.12 complete genomes per 10,000 infected individuals) (Barreto et al., 2016; Mota-Miranda et al., 2013; Neto et al., 2011). As in some associated diseases such as IDH, there is no record of complete genomes. This holds true even in areas considered endemic, such as Salvador. In addition, databases containing genomic sequence data are rather deficient with regard to highly relevant patient information, including proviral load, clinical profile and sampling date data. Here, we generated complete HTLV-1 genome sequences derived from 31 individuals with four distinct clinical profiles and perform molecular analysis to identify variations/mutations in the genome that are associated with these clinical profiles.

Material and Methods

Patients

The present study included 31 peripheral blood samples from HTLV-1 infected patients with defined clinical profiles: 10 samples from patients with TSP/HAM, seven samples from ATL patients, nine samples from asymptomatic patients and five samples from patients with IDH. All patients were recruited by means of convenience sampling at either the HTLV Reference Center, Bahiana School of Medicine and Public Health – Salvador, Bahia, the Prof. Edgar Santos University Hospital (HUPES) – Salvador, Bahia, or at the Regional Blood Center of Ribeirão Preto, São Paulo. The inclusion criteria were the positive result to HTLV-1 infection, and information about clinical classification performed by medical experts in accordance with the criteria established by the World Health Organization (WHO). The exclusion criteria were the presence of these infections such as HIC, HCV. All samples were anonymous and unlinked, and informed written consent was obtained from each study subject. The present research proposal received approval from the Institutional Review Board of the Gonçalo Moniz Research Center (Fiocruz-Bahia).

Sample collection and DNA extraction

Peripheral blood samples (10 mL in EDTA) for each participant were collected at diagnosis and during follow-up. Peripheral blood mononuclear cells (PBMC) were separated by density gradient centrifugation using Histopaque 1077 (Sigma Aldrich). DNA was extracted using a QIAamp DNA blood Mini kit according to the manufacturer's instructions (QIAGEN, Germany).

Amplification of the proviral genome

To design primers for HTLV-1 product amplification, all complete HTLV-1 genome information available in Genbank was obtained to create a consensus sequence. Using IDT SciTools® Web Tools online software, we have generated 32 pairs of primers capable of amplifying the most prevalent variant of HTLV-1 (Supplementary Table 1). Each pair of primers had the capacity to amplify a 400-base pair fragment with 50 overlapping base pairs.

The PCR conditions were the same for all fragments, as follows: denaturation at 94°C for 3 minutes, annealing steps at 94°C (15 sec), 65 °C (45 sec) and 72 °C (1 minute), performed for 35 cycles, followed by a final extension at 72°C for 8 minutes. All amplified products were analyzed on 1% agarose gel and quantified using a QubitTM dsDNA HS Assay Kit (Invitrogen).

Library preparation and sequencing

The amplicons of each patient were pooled and later fragmented in 200 pb using a Bioruptor ® standard sonication device (Diagenode). Libraries were built using the automated AB Library Builder System (Thermo Fisher Scientific), using *Ion ™ Plus Library Kit*. To normalizing the number of molecules required for emulsion PCR (ePCR), a quantitation step was performed in LightCycler® 480 Instrument II (Roche Life Science). The ePCR was performed in an automated Ion OneTouch 2 platform (Thermo Fisher Scientific) in accordance with manufacturer protocols. Finally, the mix was loaded on an *Ion 318™ Chip Kit- Ion Torrent ™* (Thermo Fisher Scientific) and sequencing reactions were performed on an Ion Personal Genome Machine™ (PGM™) System using *Ion PGM™ HI-Q™ View Sequencing* (Thermo Fisher Scientific) kit.

Data analysis

To perform molecular analysis, the fastq files, from each generated viral genome, were trimmed and mapped to the HTLV-1 reference sequence ATK-1 (J02029) using the highest sensitivity and fine-tuning settings, followed by editing and manual alignment of the mapped reads, these steps were performed using the Geneious software (Kearse et al., 2012).

The genetic distances among 31 new complete genomes and others 69 reference sequences were measured using MEGA 7.0 software (Kumar et al., 2016), both within and among the four distinct groups: asymptomatic, ATL, IDH and TSP/HAM. Next, the LTR sequences were submitted to the HTLV-1 Automated Subtyping Tool Version 1.0 to identify the HTLV-1 subtype of each genome (Alcantara et al., 2009).

The non-coding region (LTR) was used to identify motifs for transcription factors using the TFScan plugin in Geneious R6 software. After annotating the major motifs: Tax responsible elements - TxRE (three 21- base pair imperfect repetitions), TATA BOX (TATAA) and E-Box (CANNTG), we have searched for any nucleotide variants that could possibly influence the binding of transcription factors at these motifs.

Coding region analysis was carried out by searching for variants using Geneious R6 Software. All statistical analyses were conducted using STATA software 14.0 using Fisher's nonparametric test, with P values under 0.05 considered statistically significant. Any possible influences exerted by these variants on protein structures was evaluated by amino acid family change analysis, physico-chemical analysis and by searching for post-translational modification sites (Exarchos et al., 2009). Physico-chemical analysis was performed using Network Protein Sequence Analysis (NPSA) (http://npsa-pbil.ibcp.fr/) and the search for post-translational modifications sites was carried out using GeneDoc software (Nicholas K.B.; Nicholas Jr., 1997) and the Prosite tool (Sigrist et al., 2005).

Next, the 31 complete HTLV-1 genome sequences, generated herein, were compared to all available HTLV-1 Cosmopolitan Transcontinental complete and partial sequences found at the GenBank database containing clinical form data (accession numbers): AF042071.1, AF259264.1, AB979451.1, M86840.1, L36905.1, KF797815.1 - KF797835.1, KF797839.1 - KF797850.1, KF797852.1, KF797853.1, KF797855.1, KF797857.1, KF797860.1 - KF797873.1, KF797875.1 - KF797882.1, KF797884.1 - KF797886.1, KF797888.1, KF797889.1, DQ005564.1 - DQ005566.1, AY342300.1 - AY342302.1, AY342305.1, AY342306.1, DQ512875.1, DQ512874.1, DQ471187.1 - DQ471209.1, AY499185.1, AB036351.1, AB036350.1, AB036348.1, L42255.1, AB211217.1, AB211216.1, AB211214.1, AF124043.1, AF014664.1 - AF014670.1, AF014662.1, AF014659.1, AF014653.1 - AF014657.1, GU126475.1, GU126478.1, EU622583.1, EU622582.1, EU392164.1, DQ512873.1, DQ897682.1 - DQ897686.1, AB036349.1, AF014644.1- AF014646.1, U25140.1, U25139.1, U25062.1, L42225.1, AB211220.1, AB211218.1, AB211215.1, AF097525.1 - AF097528.1, AF014647.1 - AF014652.1, GU126474.1, GU126476.1, GU126479.1 - GU126483.1, L42253.1.1, GU126487.1, DQ663637.1, GU126484.1, GU126488.1, GU126489.1, DQ663642.1, GU126485.1, X56949.2, AB036377.1.1, AB211204.1, AB211201.1 and AF485381.1

Phenotype Identification by Phylogenetic Analysis

The 31 complete HTLV-1 genome sequences and the sequence J02029 were aligned using Geneious software in order to verify the phenotype relationship. The reference sequence J02029 was used as outgroup. The dataset was submitted to JModeltest to verify the suitable model for analysis (Posada, 2008). The Tamura–Nei with gamma distribution and proportion of invariable sites was selected as the best evolutionary model for the data. Bayesian analysis was performed in Beast software using the evolutionary model described above as prior (Posada, 2008; Suchard et al., 2018). The tree reliability was analyzed using the effective sample size (ESS) parameter.

Data mining

The Bayesian network (BN) learning algorithm (using B-course software adapted by (Deforche et al., 2006) was used to describe and visualize conditional dependencies among the variables of interest: clinical profiles (Asymptomatic, ATL, TSP/HAM and IDH) and mutations identified in the coding regions of the genome (Friedman N, 2013). Dependencies are qualitatively represented by a directed acyclic graph in which each node corresponds to a variable, and a direct arc between nodes represents a direct influence. Robustness of the arcs was scored using a nonparametric bootstrap test (100× replicates), and only arcs with 30% or more support were depicted.

Results

The present study population consisted of 31 HTLV-1 infected individuals with different clinical profile: ATL (n=7), TSP/HAM (n=10), IDH (n=5) and asymptomatic HTLV-1 infection (n=9). Patient age ranged between 12 and 67 years, with a median of 41 years. Females represented 61% of the sample. Median proviral loads were measured: ATL (419 copies per 10^3 PBMCs), TSP/HAM (117 copies per 10^3 PBMCs), IDH (99 copies per 10^3 PBMCs) and asymptomatic (41 copies per 10^3 PBMCs).

Proviral DNA samples, from each individual, were successfully amplified and submitted to Ion Torrent Platform to generate 31 complete HTLV-1 genomes sequences (8,998 base pairs each). Posteriorly the HTLV-1 Automated Subtyping Tool revealed that all sequences were classified as subtype a (cosmopolitan), subgroup A (transcontinental). The overall genetic distance observed on genomes was 0.005 and the genetic distances both within and among groups ranged from 0.006 to 0.008 (Table 1).

The sequences were submitted to molecular characterization through the comparative analysis between them, this made possible the identification of variants in the non-coding region (LTR) and coding regions (gag, pro, pol, env and the pX region). No phylogenetic relationship was found related to the different groups studied (Supplementary Figure 1).

Table 1: Genetic distances in different clinical forms of HTLV-1 complete genomes.

	IDH	TSP/HAM	ATL	Asymptomatic
IDH	0.006	0.007	0.006	0.006
TSP/HAM	0.007	0.008	0.007	0.007
ATL	0.006	0.007	0.007	0.006
Asymptomatic	0.006	0.007	0.006	0.006

IDH, infective dermatitis associated to HTLV-1; TSP/HAM, Tropical Spastic Paraparesis /HTLV-1-Associated Myelopathy; ATL, Adult T-cell leukemia/lymphoma. P- values were calculated by the Fisher's exact test. Were tested 31 new complete genomes and others 69 reference sequences.

Non-coding region analysis

About LTR analysis, we specifically have investigated variants in the TxRE, TATA BOX and E-Box motifs, considering that the transcription factor motifs are important in regulating the transcription of HTLV-1 provirus and that mutations associated with the loss or gain of any of these motifs could be responsible for changes in viral gene expression. Of the motifs evaluated, we have identified only two mutations in TxRE that demonstrated potential to abrogate the Sph1 binding site. The A125G mutation, which could abrogate TxRE1, was found in four isolates from TSP/HAM individuals, three isolates from ATL patients and four isolates from asymptomatic carriers, while the G174A mutation, which could abrogate TxRE2, was found in four isolates from TSP/HAM individuals, two isolates from ATL individuals and three isolates from asymptomatic carriers.

We then have searched for these two mutations in the GenBank database, considering all the LTR Cosmopolitan Transcontinental sequences with available clinical profile data. The A125G mutation was found in 15 isolates from TSP/HAM individuals, in 26 isolates from asymptomatic carriers and in one isolate from ATL patient. While the G174A mutation was found in 17 isolates from TSP/HAM individuals, in 31 isolates from asymptomatic carriers and in one isolate from ATL patient. Similarly, to our sample sequences, the A125G and G174A mutations were not detected in any sequences of the IDH clinical profile. Despite these findings, we have found no statistical association between these mutations and the four clinical profiles available.

Coding region analysis

We have identified 377 mutations in the coding regions, 226 of which were synonymous (i.e. no amino acid alterations). A total of 151 non-synonymous mutations were identified in the *gag* (n=28), *pro* (n=9), *pol* (n=26), *env* (n=37) genes and the pX region (n=51). To evaluate the possible influence of these mutations on HTLV-1 proteins, we have analyzed whether these non-synonymous mutations would introduce amino acid family changes. Accordingly, a total of 63 mutations capable of inducing amino acid family alterations were found, distributed throughout the entire genome. The frequency of these mutations in each coding region was then analyzed with respect to the four clinical profiles and no statistical difference was observed between them (Fig 1A).

Completing the search for functional relation between mutations and clinical profiles, we have decided to evaluate the possible conditional dependencies between the variables of interest: clinical profiles and mutations identified in the coding regions of the genome. For this analysis, 12 mutations were considered, representing a frequency higher than 5% in the data set. We have identified only the dependence relationship between the mutations, with robustness varying from 33 to 100%. Confirming the previous result, there was no relation between mutations and clinical profile (Fig 1B).

Considering the twelve variants, 14 amino acid changes were identified in the coding genes and were submitted to the characterization of the post-translational modification sites and the physical-chemical profiles. Seven of these alterations (P8S, P77R, Q44R, W88S, P92S, G307R, D264G) in the corresponding proteins were found to be capable of creating post-translational modification sites, while four mutations (G201D, T40A, R47C, R37C) were determined to abrogate post-translational modification sites. These changes were found in all clinical profiles at random, with no statistical significance. About physical-chemical analysis, two mutations were found to reduce antigenicity (T106I and V247I) and one mutation (I153V) was associated to increase the antigenicity profile. The T106I mutation was found in viral isolates from TSP/HAM and ATL clinical profile, while the V247I was found in almost all clinical profiles except for IDH. The I153V mutation was found in viral isolates from asymptomatic carriers and IDH individuals. However, we did not find statistical significance between theses mutations and clinical profiles (Table 2).

Table 2: Physical-chemical and post translational analysis of 31 HTLV-1 complete genomes.

Gene (protein)	**Physical-chemical analysis**	**Post translational modifications**	**Clinical Profile** Frequency (%)				**p-value**
			ASY n=9	**IDH** n=5	**TSP/HAM** n=10	**ATL** n=7	
Gag (p15)							
P8S	-	Creation cAMP- and cGMP-dependent protein kinase phosphorylation site	1	1	-	-	0.389
P77R	-	Creation cAMP- and cGMP-dependent protein kinase phosphorylation site	-	1	-	-	1.0
Gag (p19)							
T106I	Reduced antigenicity	-	-	-	1	3	0.052
Gag (p24)							
I153V	Increased antigenicity	-	1	1	-	-	0.146
Pro							
G201D	-	Abrogation N-myristoylation site	-	-	1	-	0.538
Pol							
T40A	-	Abrogation protein kinase C phosphorylation site	-	-	-	1	0.315
Q446R	-	Creation protein kinase C phosphorylation site	3	1	1	1	0.614
Env (gp46)							
W88S	-	Creation protein kinase C phosphorylation site	-	-	-	1	0.315
P92S	-	Creation protein kinase C phosphorylation site	-	-	1	-	0.538
V247I	Reduced antigenicity	-	5	-	4	2	0.206
G307R	-	Creation protein kinase C phosphorylation site	-	1	-	-	0.146
pX (Tax)							
D264G	-	Creation N-myristoylation site	-	-	1	-	0.538
pX (p13)							
R47C	-	Abrogation protein kinase C phosphorylation site	1	-	-	-	0.471
pX (p30)							
R37C	-	Abrogation protein kinase C phosphorylation site	-	1	5	3	0.083

Position related to J02029. Asy, asymptomatic; IDH, HTLV-1 associated infective dermatitis; TSP/HAM, Tropical Spastic Paraparesis /HTLV-1-Associated Myelopathy; ATL, Adult T-cell leukemia/lymphoma. P- values were calculated by the Fisher's exact test.

From the above results, we expanded our analysis by searching all HTLV-1 Cosmopolitan Transcontinental sequences with data about clinical forms available using the GenBank database, considering the mutations capable of creating or abrogating post-translational modification sites, as well as the mutations determined to alter physical-chemical properties, the results were similar to those found in our sequences (Supplementary Table 2).

Discussion

Five to ten million individuals are infected worldwide by HTLV-1, however this infection is a neglected public health condition (Cao et al., 2000; Chou et al., 1995; Gessain and Cassar, 2012). The low number of complete genomes reflects reduced investment in HTLV-1-related research. Except Japan, most countries endemic for this infection are underdeveloped or undeveloped.

The viral infection of HTLV-1 carriers is quantified by the number of infected cells expressed as the proviral load, which has been shown to have a wide range of peripheral blood among infected individuals (Nagai et al., 1998; Yakova et al., 2005). The mean proviral load in patients with ATL, TSP/HAM, or other inflammatory syndromes is significantly higher than asymptomatic (Nagai et al., 1998; Yakova et al., 2005). These results indicate that proviral burden may be an important risk marker for the development of associated diseases. Among the diseases associated with HTLV-1, the pathophysiology of ATL favors lymphoproliferation and consequently the highest value of proviral burden in infected individuals (Iwanaga et al., 2010; Okayama et al., 2004). Their main associated clinical manifestations or do not have treatment or it is not efficient.

HTLV-1-infected cells are resistant to apoptosis-inducing agents, so treatment of ATL patients using conventional chemotherapy has very limited benefit (Ohsugi et al., 2004). Some chronic and acute ATL patients are, however, efficiently treated with a combination of interferon α and zidovudine (IFN-α/AZT), to which arsenic trioxide is added in some cases (Zanella et al., 2012). On the other hand, no efficient treatment for IDH and TSP/HAM patients has been described yet.

Therefore, genomic studies with the objective of investigating the mechanisms involved at pathology manifestation and the development of therapies that could either prevent the occurrence of HTLV-1-associated diseases or at least block the evolution of the disease in the early stages are very appropriate and should be very explanatory.

It is important to note however, that the results showed in this report, as others obtained in our group or others already published, indicate that those genomic studies only will be forceful if they bring together some aspects such as: 1- more appropriate methodology; 2- the greatest possible number of individuals per clinical group; 3- different clinical forms; 4-analyses of viral complete genomes; 5- and evaluation of different markers, not only the viral ones, but host (genetic) also.

In this point of view, next generation sequencing (NGS) employs high throughput sequencing technology capable of generating data from millions of base pairs in a single run. NGS platforms generate a greater amount of data, in a shorter period of time and at a lower cost per base sequence in comparison to Sanger sequencing (Cassar et al., 2013; Kuramitsu et al., 2015). Therefore, the fact that we have used the NGS is a positive argument, especially for evaluation of low frequency SNPs.

Researchers with expertise in the area of HTLV-1 investigation have cited the need for comprehensive studies involving the genetic characterization of complete genomes to aid in the understanding of virus pathogenicity and disease development (Cassar et al., 2013; Kuramitsu et al., 2015; Ma et al., 2013; Martin et al., 2011). Here, we generated and characterized 31 complete HTLV-1 genome sequences with four clinical profile classifications: ATL, TSP/HAM, IDH and asymptomatic infection.

In aspect of studying different clinical forms, besides the three more common: ATL, TSP/HAM and asymptomatic carriers, this report could generate seven complete genomes of IDH. The sequences obtained from patients with IDH represent the first complete genome sequences of this clinical type to be made publicly available.

All new sequences generated in this work in the clusters belonging to the Cosmopolitan subtype and subgroup Transcontinental confirms previous data that indicate that there is a predominance of this subtype/subgroup circulating in the Bahia state (Durkin et al., 2006; Lodewick et al., 2011; Lodewick et al., 2009). The genetic distances between intragroup and intergroup were of low diversity, these results confirm the low variability of HTLV-1.

Regarding to non-coding region (LTR) analysis, two variants abrogate Sp1 binding sites in specific TxRE regions. Sp1 transcription factor is critical for basic leucine zipper factor (HBZ) transcription, which plays a significant role in the proliferation of infected cells (Peloponese et al., 2004) . These variants were not associated, in our casuistic, with the different clinical profiles analyzed, maybe because they do not have clinical and functional importance, in separated. However, the lack of statistical significance can be due to reduced number of isolates of each clinical group. It is interesting to observe that both mutations were also identified at others sequences available at GenBank database, in an increased number, but with insufficient statistical support yet probably because the lack of information about clinical profile associated to sequences available at public databases.

About the analysis of the coding region, among 151 non-synonymous mutations that were identified, almost 34% of them were found in a single genomic region, pX region, which encodes regulatory proteins, while the other mutations were distributed at structural genes. When we have analyzed the 63 found mutations capable of inducing amino acid family alterations, it is possible to observe the same phenomenon, once 66% of them were identified at regulatory proteins, including Tax. Maybe this observation could signal/confirm the importance of these proteins to viral survival and persistence and therefore, contribute to clinical profile. However, the found variants were not associated with the clinical profiles evaluated, corroborating with earlier studies demonstrating that the nucleotide variants in some fragments of the HTLV-1 genome were specific for the geographic origin of the patients rather than for the type of associated pathologies (Kfoury et al., 2008; Kfoury et al., 2012; Lamsoul et al., 2005; Nasr et al., 2006).

We have also identified two mutations in coding regions that were capable of reducing antigenicity, and one of that was shown to increase antigenicity. Although these variants were not associated to any specific clinical profile, they may become biomarkers as they are located in important protein domains.

Post-translational modification events play an important role in viral replication and cell transformation. Our analyses demonstrated a high identity among the sequences, suggesting the influence of these sites on the host immune response, and virus persistence. We have identified a low frequency of mutations associated with the creation or abrogation of these sites in different clinical profiles. Some modifications as phosphorylation, ubiquitination and acetylation are critical for Tax transactivation via both ATF/CREB and NF-κB pathways and this could act in inhibition of DNA repair, cell cycle control and activation of p53 tumor suppressor (Andonov et al., 2012;

Capobianchi et al., 2013; Jeang, 2001; Malik et al., 1988; Radford et al., 2012) . The post-translational modifications in other proteins are not well documented. Therefore, there is a lack of evidence of the importance of these sites for viral infection (Ellerbrok et al., 1998).

And finally, we suggest, once we have not found any association of mutation to clinical profile, that the viral genomic background should be evaluated together with genetic host factors. A study, analyzing the genotypes of 66 infected individuals, between asymptomatic and TSP/HAM, and 192 control individuals, revealed that FAS-670A/G polymorphism may be associated not only with susceptibility to infection, but also with progression to TSP/HAM (Vallinoto et al., 2012). Since the location of this SNP may favor its binding to the transcription factor STAT1, which would be enough to trigger upregulation or negative regulation of FAS gene expression.

In evolutionary processes, genetic variation, whether in the host or in the virus, undergoes selective pressures until it is selected positively or negatively, revealing the functional changes of each gene information. Infectious agents, for example, may represent a powerful selective force, especially on regulatory molecules, such as T. cells. A reported study has tested 15 genes encoding proteins associated with T-cell co-stimulation in 39 mammalian species, which revealed that 9 of these genes were positively selected. When the SNPs evaluated sought association with disease manifestation or susceptibility to infection, it was possible to demonstrate that disease-associated variations in T-cell genes are preferentially the targets of pathogen-directed selection (Forni et al., 2013).

The reduced number of our cohort, especially intra-group is a negative point, but this reflect the difficulty of conducting neglected infection studies, even if it is an infection that can lead the individuals to death. However, this study contributed to increase the number of HTLV-1 complete genomes published with comprehensive information about viral load and clinical profile data. Furthermore, they are the first 31 complete genomes of an endemic region, highlighting the first associated with IDH. The results of molecular characterization not suggest association between mutations and clinical profile. To investigate better the contribution of HTLV-1 mutations for the disease outcome it is necessary to evaluate the interaction of the viral genome and characteristics of the human host.

Supporting data

All sequences newly reported herein were deposited into the NCBI's GenBank database (accession numbers KY007244-KY007274).

Acknowledgments

We would like to thank the FAPESB and CNPq financial support.

Competing Interests

The authors declare that they have no competing interests.

References

Alcantara, L.C., Cassol, S., Libin, P., Deforche, K., Pybus, O.G., Van Ranst, M., Galvao-Castro, B., Vandamme, A.M., de Oliveira, T., 2009. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Res 37, W634-642.

Andonov, A., Coulthart, M.B., Perez-Losada, M., Crandall, K.A., Posada, D., Padmore, R., Giulivi, A., Oger, J.J., Peters, A.A., Dekaban, G.A., 2012. Insights into origins of Human T-cell Lymphotropic Virus Type 1 based on new strains from aboriginal people of Canada. Infect Genet Evol 12, 1822-1830.

Barreto, F.K., Khouri, R., Rego, F.F.A., Santos, L.A., Castro-Amarante, M.F., Bialuk, I., Pise-Masison, C.A., Galvao-Castro, B., Gessain, A., Jacobson, S., Franchini, G., Alcantara, L.C., Jr., 2016. Analyses of HTLV-1 sequences suggest interaction between ORF-I mutations and HAM/TSP outcome. Infect Genet Evol 45, 420-425.

Cao, F., Ji, Y., Huang, R., Zhao, T., Kindt, T.J., 2000. Nucleotide sequence analyses of partial envgp46 gene of human T-lymphotropic virus type I from inhabitants of Fujian Province in Southeast China. AIDS Res Hum Retroviruses 16, 921-923.

Capobianchi, M.R., Giombini, E., Rozera, G., 2013. Next-generation sequencing technology in clinical virology. Clin Microbiol Infect 19, 15-22.

Carneiro-Proietti, A.B., Sabino, E.C., Leao, S., Salles, N.A., Loureiro, P., Sarr, M., Wright, D., Busch, M., Proietti, F.A., Murphy, E.L., Nhlbi Retrovirus Epidemiology Donor Study-Ii, I.C., 2012. Human T-lymphotropic virus type 1 and type 2 seroprevalence, incidence, and residual transfusion risk among blood donors in Brazil during 2007-2009. AIDS Res Hum Retroviruses 28, 1265-1272.

Cassar, O., Einsiedel, L., Afonso, P.V., Gessain, A., 2013. Human T-cell lymphotropic virus type 1 subtype C molecular variants among indigenous australians: new insights into the molecular epidemiology of HTLV-1 in Australo-Melanesia. PLoS Negl Trop Dis 7, e2418.

Chou, K.S., Okayama, A., Tachibana, N., Lee, T.H., Essex, M., 1995. Nucleotide sequence analysis of a full-length human T-cell leukemia virus type I from adult T-cell leukemia cells: a prematurely terminated PX open reading frame II. Int J Cancer 60, 701-706.

Deforche, K., Silander, T., Camacho, R., Grossman, Z., Soares, M.A., Van Laethem, K., Kantor, R., Moreau, Y., Vandamme, A.M., non, B.W., 2006. Analysis of HIV-1 pol sequences using Bayesian Networks: implications for drug resistance. Bioinformatics 22, 2975-2979.

Dourado, I., Alcantara, L.C., Barreto, M.L., da Gloria Teixeira, M., Galvao-Castro, B., 2003. HTLV-I in the general population of Salvador, Brazil: a city with African ethnic and sociodemographic characteristics. J Acquir Immune Defic Syndr 34, 527-531.

Durkin, S.S., Ward, M.D., Fryrear, K.A., Semmes, O.J., 2006. Site-specific phosphorylation differentiates active from inactive forms of the human T-cell leukemia virus type 1 Tax oncoprotein. J Biol Chem 281, 31705-31712.

Ellerbrok, H., Fleischer, C., Salemi, M., Reinhardt, P., Ludwig, W.D., Vandamme, A.M., Pauli, G., 1998. Sequence analysis of the first HTLV-I infection in Germany without relations to endemic areas. AIDS Res Hum Retroviruses 14, 1199-1203.

Exarchos, K.P., Exarchos, T.P., Papaloukas, C., Troganis, A.N., Fotiadis, D.I., 2009. Detection of discriminative sequence patterns in the neighborhood of proline cis peptide bonds and their functional annotation. BMC Bioinformatics 10, 113.

Forni, D., Cagliani, R., Pozzoli, U., Colleoni, M., Riva, S., Biasin, M., Filippi, G., De Gioia, L., Gnudi, F., Comi, G.P., Bresolin, N., Clerici, M., Sironi, M., 2013. A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles. Immunity 38, 1129-1141.

Friedman N, G.M., Wyner A. , 2013. Data Analysis with Bayesian Networks: A Bootstrap Approach. . ArXiv13016695 Cs Stat.

Gessain, A., Barin, F., Vernant, J.C., Gout, O., Maurs, L., Calender, A., de The, G., 1985. Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. Lancet 2, 407-410.

Gessain, A., Cassar, O., 2012. Epidemiological Aspects and World Distribution of HTLV-1 Infection. Front Microbiol 3, 388.

Iwanaga, M., Watanabe, T., Utsunomiya, A., Okayama, A., Uchimaru, K., Koh, K.R., Ogata, M., Kikuchi, H., Sagara, Y., Uozumi, K., Mochizuki, M., Tsukasaki, K., Saburi, Y., Yamamura, M., Tanaka, J., Moriuchi, Y., Hino, S., Kamihira, S., Yamaguchi, K., Joint Study on Predisposing Factors of, A.T.L.D.i., 2010. Human T-cell leukemia virus type I (HTLV-1) proviral load and disease progression in asymptomatic HTLV-1 carriers: a nationwide prospective study in Japan. Blood 116, 1211-1219.

Jeang, K.T., 2001. Functional activities of the human T-cell leukemia virus type I Tax oncoprotein: cellular signaling through NF-kappa B. Cytokine Growth Factor Rev 12, 207-217.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647-1649.

Kfoury, Y., Nasr, R., Favre-Bonvin, A., El-Sabban, M., Renault, N., Giron, M.L., Setterblad, N., Hajj, H.E., Chiari, E., Mikati, A.G., Hermine, O., Saib, A., de The, H., Pique, C., Bazarbachi, A., 2008. Ubiquitylated Tax targets and binds the IKK signalosome at the centrosome. Oncogene 27, 1665-1676.

Kfoury, Y., Nasr, R., Journo, C., Mahieux, R., Pique, C., Bazarbachi, A., 2012. The multifaceted oncoprotein Tax: subcellular localization, posttranslational modifications, and NF-kappaB activation. Adv Cancer Res 113, 85-120.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33, 1870-1874.

Kuramitsu, M., Okuma, K., Yamagishi, M., Yamochi, T., Firouzi, S., Momose, H., Mizukami, T., Takizawa, K., Araki, K., Sugamura, K., Yamaguchi, K., Watanabe, T., Hamaguchi, I., 2015. Identification of TL-Om1, an adult T-cell leukemia (ATL) cell line, as reference material for quantitative PCR for human T-lymphotropic virus 1. J Clin Microbiol 53, 587-596.

La Grenade, L., Manns, A., Fletcher, V., Derm, D., Carberry, C., Hanchard, B., Maloney, E.M., Cranston, B., Williams, N.P., Wilks, R., Kang, E.C., Blattner, W.A., 1998. Clinical, pathologic, and immunologic features of human T-lymphotrophic virus type I-associated infective dermatitis in children. Arch Dermatol 134, 439-444.

Lamsoul, I., Lodewick, J., Lebrun, S., Brasseur, R., Burny, A., Gaynor, R.B., Bex, F., 2005. Exclusive ubiquitination and sumoylation on overlapping lysine residues mediate NF-kappaB activation by the human T-cell leukemia virus tax oncoprotein. Mol Cell Biol 25, 10391-10406.

Lodewick, J., Lamsoul, I., Bex, F., 2011. Move or die: the fate of the Tax oncoprotein of HTLV-1. Viruses 3, 829-857.

Lodewick, J., Lamsoul, I., Polania, A., Lebrun, S., Burny, A., Ratner, L., Bex, F., 2009. Acetylation of the human T-cell leukemia virus type 1 Tax oncoprotein by p300 promotes activation of the NF-kappaB pathway. Virology 386, 68-78.

Ma, Y., Zheng, S., Wang, N., Duan, Y., Sun, X., Jin, J., Zang, W., Li, M., Wang, Y., Zhao, G., 2013. Epidemiological analysis of HTLV-1 and HTLV-2 infection among different population in Central China. PLoS One 8, e66795.

Malik, K.T., Even, J., Karpas, A., 1988. Molecular cloning and complete nucleotide sequence of an adult T cell leukaemia virus/human T cell leukaemia virus type I (ATLV/HTLV-I) isolate of Caribbean origin: relationship to other members of the ATLV/HTLV-I subgroup. J Gen Virol 69 (Pt 7), 1695-1710.

Martin, F., Bangham, C.R., Ciminale, V., Lairmore, M.D., Murphy, E.L., Switzer, W.M., Mahieux, R., 2011. Conference highlights of the 15th International Conference on Human Retrovirology: HTLV and related retroviruses, 4-8 June 2011, Leuven, Gembloux, Belgium. Retrovirology 8, 86.

Mota-Miranda, A.C., Barreto, F.K., Amarante, M.F., Batista, E., Monteiro-Cunha, J.P., Farre, L., Galvao-Castro, B., Alcantara, L.C., 2013. Molecular characterization of HTLV-1 gp46 glycoprotein from health carriers and HAM/TSP infected individuals. Virol J 10, 75.

Nagai, M., Usuku, K., Matsumoto, W., Kodama, D., Takenouchi, N., Moritoyo, T., Hashiguchi, S., Ichinose, M., Bangham, C.R., Izumo, S., Osame, M., 1998. Analysis of HTLV-I proviral load in 202 HAM/TSP patients and 243 asymptomatic HTLV-I carriers: high proviral load strongly predisposes to HAM/TSP. J Neurovirol 4, 586-593.

Nasr, R., Chiari, E., El-Sabban, M., Mahieux, R., Kfoury, Y., Abdulhay, M., Yazbeck, V., Hermine, O., de The, H., Pique, C., Bazarbachi, A., 2006. Tax ubiquitylation and sumoylation control critical cytoplasmic and nuclear steps of NF-kappaB activation. Blood 107, 4021-4029.

Neto, W.K., Da-Costa, A.C., de Oliveira, A.C., Martinez, V.P., Nukui, Y., Sabino, E.C., Sanabani, S.S., 2011. Correlation between LTR point mutations and proviral load levels among human T cell lymphotropic virus type 1 (HTLV-1) asymptomatic carriers. Virol J 8, 535.

Nicholas K.B.; Nicholas Jr., H.B.D.I., D.W., 1997. GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEW News 4.

Ohsugi, T., Kumasaka, T., Urano, T., 2004. Construction of a full-length human T cell leukemia virus type I genome from MT-2 cells containing multiple defective proviruses using overlapping polymerase chain reaction. Anal Biochem 329, 281-288.

Okayama, A., Stuver, S., Matsuoka, M., Ishizaki, J., Tanaka, G., Kubuki, Y., Mueller, N., Hsieh, C.C., Tachibana, N., Tsubouchi, H., 2004. Role of HTLV-1 proviral DNA load and clonality in the development of adult T-cell leukemia/lymphoma in asymptomatic carriers. Int J Cancer 110, 621-625.

Peloponese, J.M., Jr., Iha, H., Yedavalli, V.R., Miyazato, A., Li, Y., Haller, K., Benkirane, M., Jeang, K.T., 2004. Ubiquitination of human T-cell leukemia virus type 1 tax modulates its activity. J Virol 78, 11686-11695.

Poiesz, B.J., Ruscetti, F.W., Gazdar, A.F., Bunn, P.A., Minna, J.D., Gallo, R.C., 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 U S A 77, 7415-7419.

Posada, D., 2008. jModelTest: phylogenetic model averaging. Mol Biol Evol 25, 1253-1256.

Radford, A.D., Chapman, D., Dixon, L., Chantrey, J., Darby, A.C., Hall, N., 2012. Application of next-generation sequencing technologies in virology. J Gen Virol 93, 1853-1868.

Sigrist, C.J., De Castro, E., Langendijk-Genevaux, P.S., Le Saux, V., Bairoch, A., Hulo, N., 2005. ProRule: a new database containing functional and structural information on PROSITE profiles. Bioinformatics 21, 4060-4066.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., Rambaut, A., 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4, vey016.

Vallinoto, A.C., Santana, B.B., dos Santos, E.L., Santo, R.R., Hermes, R.B., Sousa, R.C., Cayres-Vallinoto, I., Machado, L.F., Ishak, M.O., Ishak, R., 2012. FAS-670A/G single nucleotide polymorphism may be associated with human T lymphotropic virus-1 infection and clinical evolution to TSP/HAM. Virus Res 163, 178-182.

Yakova, M., Lezin, A., Dantin, F., Lagathu, G., Olindo, S., Jean-Baptiste, G., Arfi, S., Cesaire, R., 2005. Increased proviral load in HTLV-1-infected patients with rheumatoid arthritis or connective tissue disease. Retrovirology 2, 4.

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 U S A 79, 2031-2035.

Zanella, L., Otsuki, K., Marin, M.A., Bendet, I., Vicente, A.C., 2012. Complete genome sequence of Central Africa human T-cell lymphotropic virus subtype 1b. J Virol 86, 12451.

AUTHOR CONTRIBUTIONS

Conception, design and critical revision: Alcântara, L.C.J; Mota-Miranda, A.C.A; Nunes, M.R.T; Galvão-Castro, B; Kashima, S; Farre, L; Bittencourt, A.L; Carvalho, E.M.
Investigations: Araújo, T.H.A; Barreto, F.K; Menezes, A.D.L; Lima, C.P.S; Oliveira, R.S; Lemos, P.S.
Data Curation: Araújo, T.H.A; Barreto, F.K; Mota-Miranda, A.C.A; Rego, F.F.A; Santos, L.A.
Formal Analysis: Araújo, T.H.A; Santos, L.A; Barreto, F.K; Rego, F.F.A.
Writing – Original Draft Preparation: Araújo, T.H.A; Barreto, F.K; Mota-Miranda, A.C.A.
Resources: Alcântara, L.C.J; Nunes, M.R.T.

Figure 1: Analysis of non-synonymous mutations, found in HTLV-1 complete genome, that alter amino acid families.

(A) Distribution of 63 mutations according to the clinical status and genomic region. The Fisher nonparametric test was conducted and no statistical significance was observed. **(B)** Analysis of interactions among 12 mutations found with frequency of at least 5% in HTLV-1 complete genome.

Highlights:

- There are few complete genomes publicly available, about 0.12 complete genomes per 10,000 infected individuals.
- The databases have a major deficiency of HTLV-1 sequences information.
- The project generated and characterized 31 complete HTLV-1 genome sequences with four clinical profile classifications: ATL, TSP/HAM, IDH and asymptomatic infection.

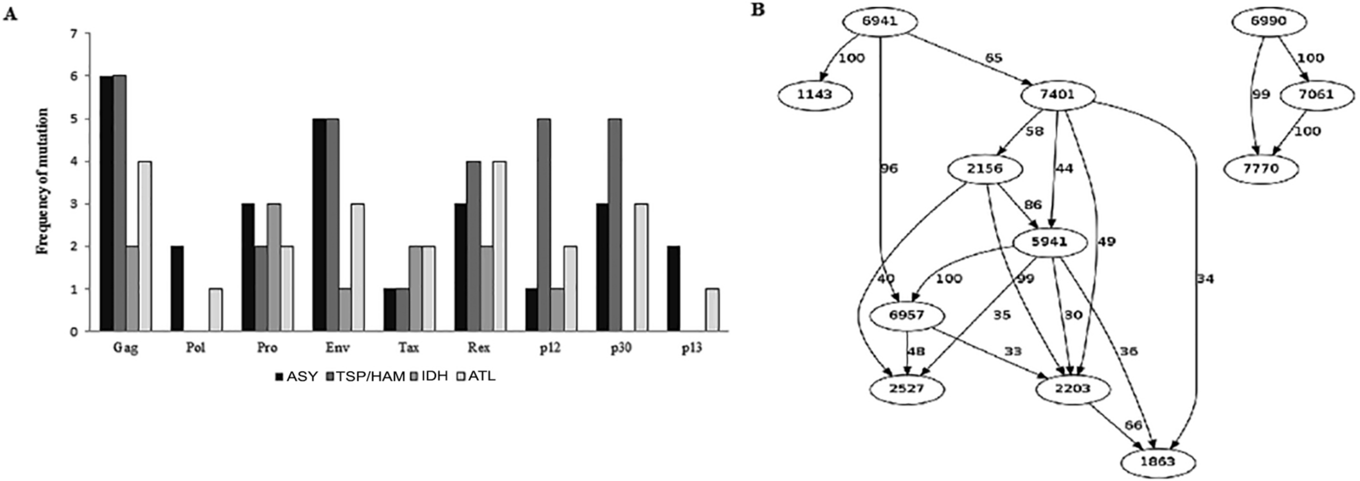

A
Frequency of mutation
0
1
2
3
4
5
6
7
Gag
Pol
Pro
Env
Tax
Rex
p12
p30
p13
■ ASY ■ TSP/HAM ■ IDH □ ATL
B
6941
100
65
1143
7401
96
58
2156
44
86
5941
49
34
40
100
99
6957
35
80
48
33
36
2527
2203
66
1863
6990
100
99
7061
100
7770

Figure 1

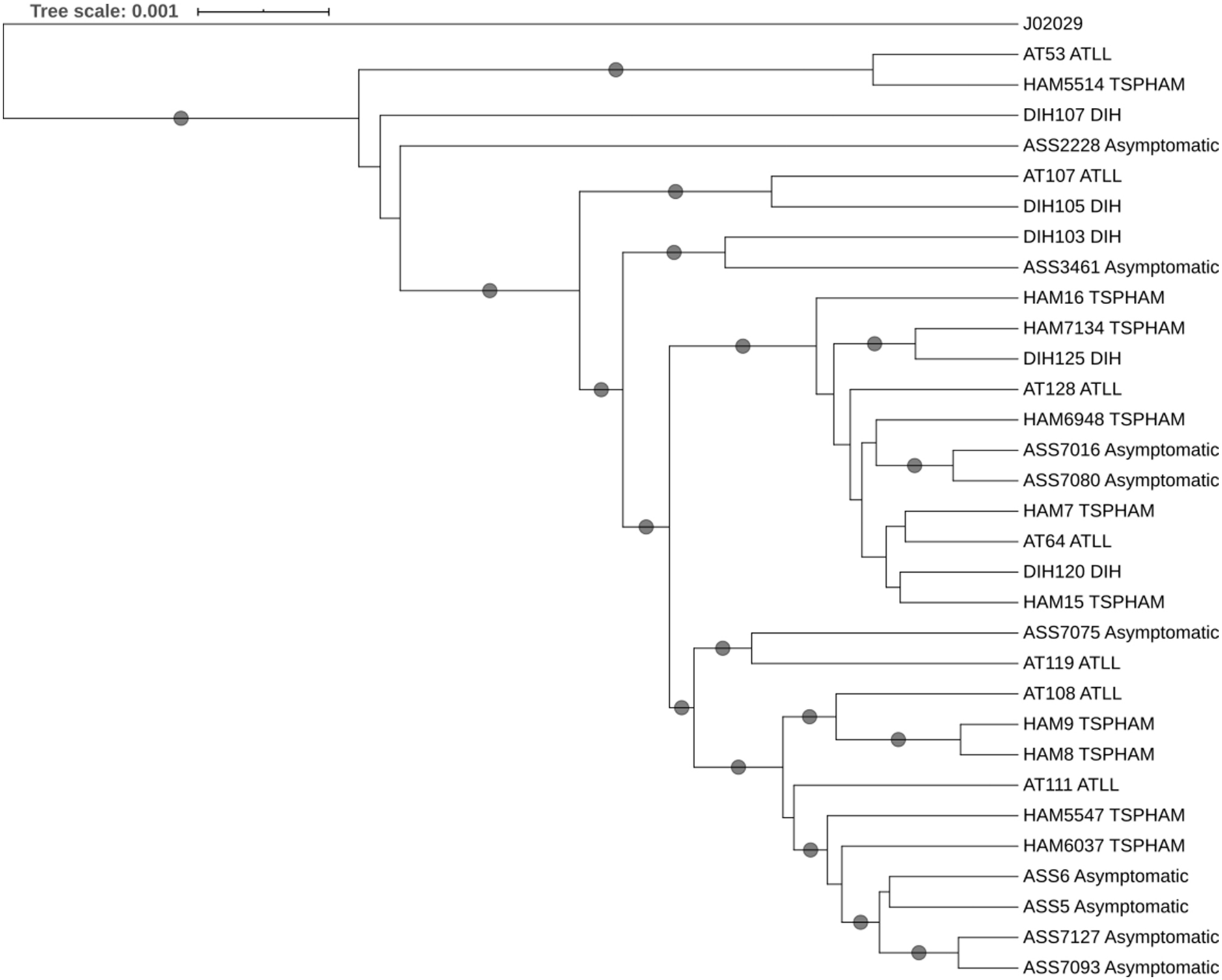

Tree scale: 0.001
J02029
AT53 ATLL
HAM5514 TSPHAM
DIH107 DIH
ASS2228 Asymptomatic
AT107 ATLL
DIH105 DIH
DIH103 DIH
ASS3461 Asymptomatic
HAM16 TSPHAM
HAM7134 TSPHAM
DIH125 DIH
AT128 ATLL
HAM6948 TSPHAM
ASS7016 Asymptomatic
ASS7080 Asymptomatic
HAM7 TSPHAM
AT64 ATLL
DIH120 DIH
HAM15 TSPHAM
ASS7075 Asymptomatic
AT119 ATLL
AT108 ATLL
HAM9 TSPHAM
HAM8 TSPHAM
AT111 ATLL
HAM5547 TSPHAM
HAM6037 TSPHAM
ASS6 Asymptomatic
ASS5 Asymptomatic
ASS7127 Asymptomatic
ASS7093 Asymptomatic

Figure 2