

# Chapter 2

## Molecular Epidemiology Database for Sequence Management and Data Mining

**Thessika Hialla Almeida Araújo, Filipe Ferreira de Almeida Rego,  
and Luiz Carlos Junior Alcantara**

### Abstract

A central database to aggregate sequence information from a range of epidemiological aspects including HTLV-1 pathogenesis, origin, and evolutionary dynamic would be useful to scientists and physicians worldwide. This Chapter describes two online tools for studies related to HTLV-1, the HTLV-1 Molecular Epidemiology Database and the HTLV-1 Subtyping Tool. The HTLV-1 Molecular Epidemiology Database is a tool for sequence management and data mining which allows researchers to download sequences with clinical and demographic information. The HTLV-1 Subtyping Tool is an online software used for HTLV-1 genotyping, the algorithm consists in the alignment of a query sequence with a carefully selected set of predefined reference strains, followed by phylogenetic analysis.

**Key words** HTLV-1 genotyping tool, HTLV molecular epidemiology, HTLV-1 database, LTR region, LASP HTLV-1

---

### 1 Introduction

The sequence analysis of HTLV is being a very useful tool with different approaches depending on the analyzed genomic region. The analysis of the LTR and envelope sequences of this virus is being used, in addition to mtDNA and Y chromosome analysis, to study the human migration and virus genotyping [1]. The HTLV pX gene analysis are useful to understand the functions of its generated proteins and their location into the cell [2], while the analysis of env, pol, and gag sequences has improved the knowledge of possible epitopes for vaccine design and helping to improve diagnostic methods [3, 4].

The HTLV nucleotide sequences, as well as from other viruses, are usually stored in the GenBank (<http://www.ncbi.nlm.nih.gov>). However, the GenBank lacks the task of giving appropriated sequences with clinical and demographic data for analysis and does not genotype viruses. Because of this was developed the HTLV-1 Molecular

Epidemiology Database, an online data mining tool that retrieves and stores annotated HTLV-1 proviral sequences from clinical, epidemiological, and geographical studies and the HTLV-1 Subtyping Tool [5–7].

---

## 2 Materials

In this case, all is needed is a computer with Internet connection and a spreadsheet reader. For analysis beyond the scope of this section, most of the software available can be viewed at <http://www.bioinformaticssoftwareandtools.co.in/index.php>.

---

## 3 Methods

### 3.1 HTLV-1 Molecular Epidemiology Database

The website interfaces were developed in HTML and server-side scripting written in PHP. The website interface contains specific search fields to allow various data combinations. User queries create a form (form tag) containing the values (variables) selected. This form generates a script that retrieves the data stored in the MySQL database. A second script organizes the data for display on the website, allowing for visualization of the information with the option to download the organized data. The developed database provides information regarding the indexed sequences in GenBank (*see Note 1*). In addition, all the sequences were genotyped using the LASP HTLV-1 Automated Subtyping Tool [7]. The user is able to choose search criteria and perform a query to generate an output of relevant sequences and information. The sequence output may be downloaded in FASTA format and the information table in Microsoft Excel spreadsheets .xls format. The HTLV-1 Molecular Epidemiology database is hosted on the Gonçalo Moniz Research Center/Oswaldo Cruz Foundation Research Center server with access at <http://htlv1db.bahia.fiocruz.br/>.

1. The HTLV-1 Molecular Epidemiology Database homepage displays interface (Fig. 1) that contains numerous fields for refining database queries.
2. One or more fields may be selected.
3. The remaining unused fields will be ignored when searching the database, but their values will be presented in the final result.
4. In the fields identified by the asterisk “\*,” the user may select multiple choices to make the search more efficient.
5. After choosing the search criteria (Fig. 2), click Run.
6. Next, the browser will be directed to a new page containing the results in table format (Fig. 3).

Choose a criteria and make a search in our HTLV-1 Database.

Genomic Region\*

env-pX  
gag-pol-env-pX  
pX  
LTR  
env

Sampling Date

Subtype

Continent\*

Africa  
Asia  
Central America  
Europe  
North America

Geographic Origin\*

Africa  
Afro-caribbean  
Algeria  
Argentina  
Bolivia

Gender

Ethnicity

Proviral Load

CD4 Count

CD8 Count

Age

Clinical Status

\*For multiple selections hold "ctrl"

Clear Run

**Fig. 1** HTLV-1 Molecular Epidemiology Database homepage displays interface

Choose a criteria and make a search in our HTLV-1 Database.

Genomic Region\*

env-pX  
gag-pol-env-pX  
pX  
LTR  
env

Sampling Date

Subtype

Continent\*

Africa  
Asia  
Central America  
Europe  
North America

Geographic Origin\*

Africa  
Afro-caribbean  
Algeria  
Argentina  
Bolivia

Gender

Ethnicity

Proviral Load

CD4 Count

CD8 Count

Age

Clinical Status

Asymptomatic

\*For multiple selections hold "ctrl"

Clear Run

**Fig. 2** Selection criteria to search in HTLV-1 Molecular Epidemiology Database homepage (example)

7. Next, select the sequences of interest (Fig. 4). If the user wants to select all sequences, the user should click in the first checkbox.
8. Next, the browser will be directed to the download page (Fig. 5), where the sequence can be downloaded as FASTA or CSV file (*see* **Notes 2** and **3**).
9. The downloaded sequences in FASTA format can be opened in almost all bioinformatics programs.
10. The CSV file can be imported into softwares (STATA, R) for statistical analysis of clinical and demographic information.

HTLV-1 Database output

Information

Your search has 7 results | Order by: 

Accession Number

Go

Please, select the sequences you need before clicking the download button. To select all the sequences, click in the first checkbox

<input type="checkbox"/>	Accession Number	Genomic Region	Size (bp)	Gender	Age	Ethnicity	Geographic Origin	Continent	Clinical Status	Proviral Load	CD4 Count (cells/ml)
<input type="checkbox"/>	<a href="#">S80215.1</a>	LTR	292				Japan	Asia	Asymptomatic		
<input type="checkbox"/>	<a href="#">S80213.1</a>	LTR	158				Japan	Asia	Asymptomatic		
<input type="checkbox"/>	<a href="#">S80212.1</a>	LTR	184				Japan	Asia	Asymptomatic		
<input type="checkbox"/>	<a href="#">L42255.1</a>	LTR	704	female	55		Kuwait	Asia	Asymptomatic		
<input type="checkbox"/>	<a href="#">AB211217.1</a>	LTR	510				Iran	Asia	Asymptomatic		
<input type="checkbox"/>	<a href="#">AB211216.1</a>	LTR	505				Iran	Asia	Asymptomatic		
<input type="checkbox"/>	<a href="#">AB211214.1</a>	LTR	498				Iran	Asia	Asymptomatic		

Download selected sequences

Fig. 3 Data presentation containing the search results

<input type="checkbox"/>	Accession Number	Genomic Region	Size (bp)	Gender	Age	Ethnicity	Geographic Origin	Continent	Clinical Status
<input checked="" type="checkbox"/>	<a href="#">S80215.1</a>	LTR	292				Japan	Asia	Asymptomatic
<input checked="" type="checkbox"/>	<a href="#">S80213.1</a>	LTR	158				Japan	Asia	Asymptomatic
<input checked="" type="checkbox"/>	<a href="#">S80212.1</a>	LTR	184				Japan	Asia	Asymptomatic
<input checked="" type="checkbox"/>	<a href="#">L42255.1</a>	LTR	704	female	55		Kuwait	Asia	Asymptomatic

Fig. 4 Selection of sequences of interest (*Checkbox marking*)

HTLV-1 Download page

Information	
<a href="#">Download FASTA file</a>	<a href="#">Download CSV file</a>
<a href="#">Back to home</a>	<a href="#">Back to previous page</a>

Fig. 5 Page to download the sequences in Fasta format and CSV file containing clinical, phylogenetics and epidemiological data

3.2 HTLV-1 Subtyping Tool

The HTLV-1 Subtyping Tool was developed using Java programming and PHP scripts. This tool accepts up to 1000 sequences at a time (*see* **Notes 4** and **5**). In the first step of the analysis, the genomic region of reference sequence (ATK1 for HTLV-1) is identified using BLAST software. The second step involves the alignment of the query sequence with a complete reference dataset composed of all subtypes. The final step involves the construction of a phylogenetic tree using Tamura-Nei or HKY distance methods with a gamma distribution among site rate heterogeneity, as implemented in PAUP software [8].

A series of PHP scripts are used to read the XML output format produced by the JAVA program and create HTML report pages. The batch report contains information on the sequence name, length, assigned subtype and subgroup, and an illustration of the virus' genome. The HTLV-1 subtyping tool is hosted on the Africa Centre for Health and Population Studies bioinformatics

## LASP HTLV-1 Automated Subtyping Tool (Version 1.0)

This tool uses phylogenetic methods to identify the subtype of query sequences.

*Please note:* The HTLV-1 subtyping is based only in the LTR region of the genome.

*Note for batch analysis:* The LASP HTLV-1 subtype tool accepts up to 1000 sequences at a time.

Enter here your input data as FASTA format.

[Choose a mirror to subtype your sequences](#) or [choose another virus to genotype](#).

```
>HTLV24_DQ005565_
CGGGGGCTTAGAGCCTCCAGTGAAAAACATTTCCGCGAAACAGAAGTCTGAAAAGG
TCA
GGGCCCAGACTAAGGCTCTGACGTCTCCCCCGGAGGGACAGCTCAGCACCGGCTC
AGG
CTAGGCCCTGACGTGTCCCTGAAAGACAAATCATAAGCTCAGACCTCCGGGAAGCC
```

**Fig. 6** HTLV-1 Subtyping Tool homepage interface: adding the sequence in FASTA format to run

group, UKZN, South Africa server with access at <http://www.bioafrica.net/regenotype/html/indexhtlv.html/> (*see Note 1*).

1. This tool uses phylogenetic methods to identify the subtype of query sequences.
2. Enter here your input data as FASTA format, after, click Run (Fig. 6).
3. The batch report (Fig. 7) will be the batch report and contain information on the sequence name, length, assigned subtype, and a figure of the HTLV-1 genome. Accessing the report link will take the user to a report to each submitted sequence.
4. The sequence report (Fig. 8) will be composed of three areas named: sequence assignment, analysis details, and phylogenetic analyses.
5. The Sequence assignment contains information on (Fig. 8):
  - (a) The sequence submitted (name and length).
  - (b) The classification assignment (subtype, subgroup, and bootstrap support).
  - (c) A graphical representation of the HTLV-1 genome showing the genomic region of the query sequence with the start and end positions related to the ATK1 genome.
  - (d) The motivation of the classification (this is based on the decision tree).
6. The Phylogenetic analysis section contains (Fig. 8):
  - (a) The phylogenetic tree in PDF and Nexus format,
  - (b) The log file generated by PAUP (contains info on the model of evolution and its parameters).
  - (c) The alignment used.

### HTLV Genotyping Tool Results

Name	Length	Report	Assignment	Support	Genome
HTLV24_DQ005565_	719bp	<a href="#">Report</a>	subtype_a(subgroup_A)	99.0	
Ni3_Y16497_	719bp	<a href="#">Report</a>	subtype_a(subgroup_B)	99.0	
IDUSSA_DQ005555_	757bp	<a href="#">Report</a>	subtype_a(subgroup_A)	100.0	
BCI1-2_U32552_	693bp	<a href="#">Report</a>	subtype_a(subgroup_A)	100.0	
test2	1190bp	<a href="#">Report</a>	subtype_a(subgroup_B)	100.0	
test	1318bp	<a href="#">Report</a>	HTLV1	NA	

Download results: [XML format](#), [CSV table](#)

**Fig. 7** The batch report that contains information on the sequence name, length, assigned subtype, and a figure of the HTLV-1 genome. Accessing the report link will take the user to a report to each submitted sequence

### Sequence Assignment

**Sequence name :** HTLV24\_DQ005565\_, length: 719 bps

**Assignment:** subtype\_a(subgroup\_A), Bootstrap: 99.0%

HTLV-1 Genome location ATK1 © LASP/FIOCRUZ HTLV1 Subtyping Tool

Your sequence start at position 53 and finish at position 769 in the ATK1 genome.  
 Motivation: Subtype assigned based on sequence located in the LTR clustering with a HTLV1 subtype and/or subgroup with bootstrap > 60%

Developed in cooperation with the [Evolutionary Biology Group](#) at University of Oxford, UK, the [REGA Institute](#) at the Katholieke Universiteit Leuven, Belgium and the [Laboratório Avançado de Saúde Pública \(LASP\)](#) CPqGM/FIOCRUZ, Brazil.

### Analysis details

- [Phylogenetic analyses](#)  
This section contains the alignments, inferred phylogenetic trees, and detailed results of the evolutionary analysis.

### Phylogenetic analyses

**Phylogenetic analysis with pure subtypes:**

- Export or View the Phylogenetic Tree: [PDF](#), [NEXUS format](#).
- View the [PAUP\\* Log file](#) (Contains bootstrap values for all HIV subtypes)
- Download [the alignment \(Nexus format\)](#).

**Fig. 8** The sequence report will be composed of three areas named: sequence assignment, analysis details, and phylogenetic analyses



## 4 Notes

1. The access numbers in the output page are actual links to the respective sequence in the GenBank page.
2. The FASTA file features a header with the following structure:  
>accession number, serial number, genomic region, isolate, base pair, subtype and subgroup (if available).
3. The CSV file is a table containing all information presented in the search screen, along with a column called Sequence. That column matches the CSV file serial number, making it possible to relate the information in the CSV file with that in the FASTA file.
4. The HTLV-1 subtyping is based only in the LTR region of the genome.
5. The LASP HTLV-1 subtype tool accepts up to 1000 sequences at a time.

## References

1. Brucato N, Cassar O, Tonasso L, Tortevoeye P, Migot-Nabias F, Plancoulaine S, Guitard E, Larrouy G, Gessain A, Dugoujon J-M (2010) The imprint of the Slave Trade in an African American population: mitochondrial DNA, Y chromosome and HTLV-1 analysis in the Noir Marron of French Guiana. *BMC Evol Biol* 10:314
2. Van Prooyen N, Gold H, Andresen V, Schwartz O, Jones K, Ruscetti F, Lockett S, Gudla P, Venzon D, Franchini G (2010) Human T-cell leukemia virus type 1 p8 protein increases cellular conduits and virus transmission. *Proc Natl Acad Sci U S A* 107:20738–20743
3. Filippone C, Bassot S, Betsem E, Tortevoeye P, Guillotte M, Mercereau-Puijalon O, Plancoulaine S, Calattini S, Gessain A (2012) A new and frequent human T-cell leukemia virus indeterminate Western blot pattern: epidemiological determinants and PCR results in central African inhabitants. *J Clin Microbiol* 50:1663–1672
4. Mota-Miranda ACA, Barreto FK, Amarante MFC, Batista E, Monteiro-Cunha JP, Farre L, Galvão-Castro B, Alcantara LCJ (2013) Molecular characterization of HTLV-1 gp46 glycoprotein from health carriers and HAM/TSP infected individuals. *Virol J* 10:75
5. Araujo THA, Barreto FK, Luiz Carlos Júnior A, Miranda ACAM (2014) Inferences about the global scenario of human T-cell lymphotropic virus type 1 infection using data mining of viral sequences. *Mem Inst Oswaldo Cruz* 109:448–451
6. Araujo THA, Souza-Brito LI, Libin P, Deforche K, Edwards D, de Albuquerque-Junior AE, Vandamme A-M, Galvao-Castro B, Alcantara LCJ (2012) A public HTLV-1 molecular epidemiology database for sequence management and data mining. *PLoS One* 7:e42123
7. Alcantara LCJ, Cassol S, Libin P, Deforche K, Pybus OG, Van Ranst M, Galvão-Castro B, Vandamme AM, de Oliveira T (2009) A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. *Nucleic Acids Res* 37:634–642
8. Wilgenbusch JC and Swofford D (2003) Inferring evolutionary trees with PAUP. *Curr Protoc Bioinformatics*, Chapter 6, unit 6.4

Human T-Lymphotropic Viruses

Methods and Protocols

Casoli, C. (Ed.)

2017, XIII, 220 p. 43 illus., 15 illus. in color., Hardcover

ISBN: 978-1-4939-6870-1

A product of Humana Press