

HIV Database Workshop

www.hiv.lanl.gov

seq-info@lanl.gov

Presenters: Will Fischer & Bette Korber

Database PIs: Bette Korber, Thomas Leitner,
Karina Yusim

Additional database staff: Werner Abfalterer, Will Fischer,
Brian Foley, Peter Hraber, Elisabeth Sharon Fung,
Robert Funkhouser, Kumkum Ganguly, Jenni Macke,
James Szinger, and Hyejin Yoon



Project Officer: Stuart Shapiro, NIAID, NIH

Editors: Christian Apetrei, Beatrice Hahn, Ilene Mizrahi,
James Mullins, Andrew Rambaut, Steve Wolinsky,
Dan Barouch, Christian Brander, Rob De Boer, Bart Haynes,
Richard Koup, John Moore, Bruce Walker, David Watkins



*Theoretical Biology and Biophysics, T-6
Los Alamos National Laboratory*



Los Alamos HIV Sequence Database Overview

Will Fischer

**Summer School on Quantitative Systems Immunology,
Boston University
June 10-15, 2013**

slides available as PDF documents:

<http://www.hiv.lanl.gov/content/sequence/HIV/HIVWORKSHOP/index.html>

Workshop Topics

HIV Sequence Database and Immunology Database

General introduction

Sequence search interface – alignments and basic trees

Geography search interface

Database Alignments

Tools:

- *Genecutter – processing nucleotide sequences*
- *Treemaker – phylogenetic trees via neighbor-joining*
- *HIV/SIV sequence locator tool*
- *Hypermut – detection of APOBEC-mediated hypermutation*
- *Highlighter – visualization of mutations in related sequences*
- *Protein Feature Accent*

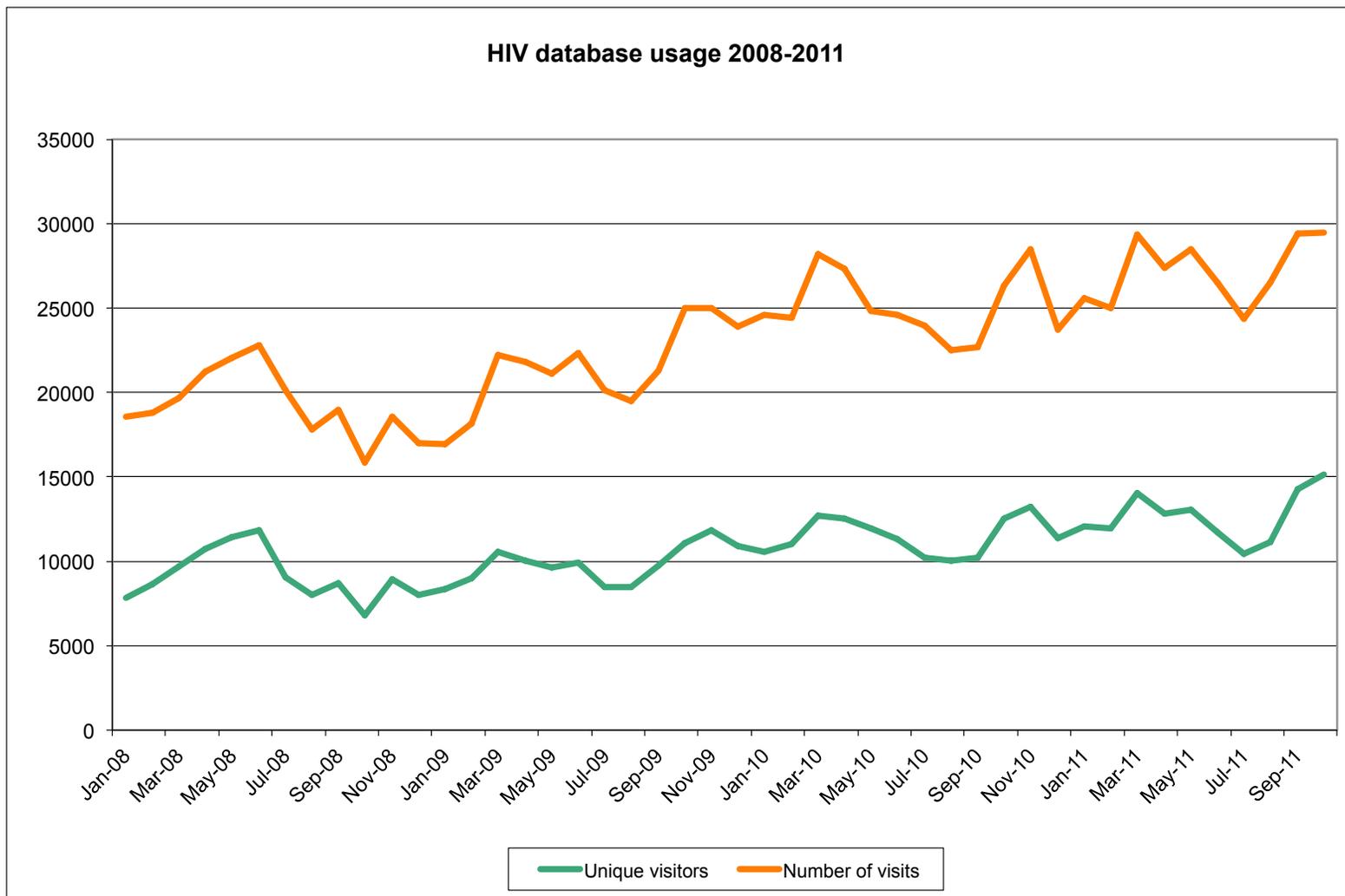
Workshop Goals

- Understanding the database content, how information was obtained, and what is available
- Database searching
- Examples of using tools for analyses

The HIV Databases

- HIV Sequence database – founded 1986, G. Myers
 - Relational database, data from GenBank with added fields from the literature
 - Alignments – align indels and reduce multiple sequences per person
 - Annual hard copy and reviews
 - Web search interfaces: subtype, phenotype, geographic, sampling year...
 - Analysis tools
- HIV Immunology database – founded 1995, B. Korber
 - Comprehensive HIV epitope database, > 300-400 new papers per year
 - Integrate HIV immunological and sequence data
 - Annual hard copy and reviews
 - Web search interfaces: epitope, protein, HLA type, immunogen, keywords
 - Analysis tools for immunologists
- HIV Vaccine database – founded 2003, J. Mokili
 - A searchable relational database of published primate vaccine trials

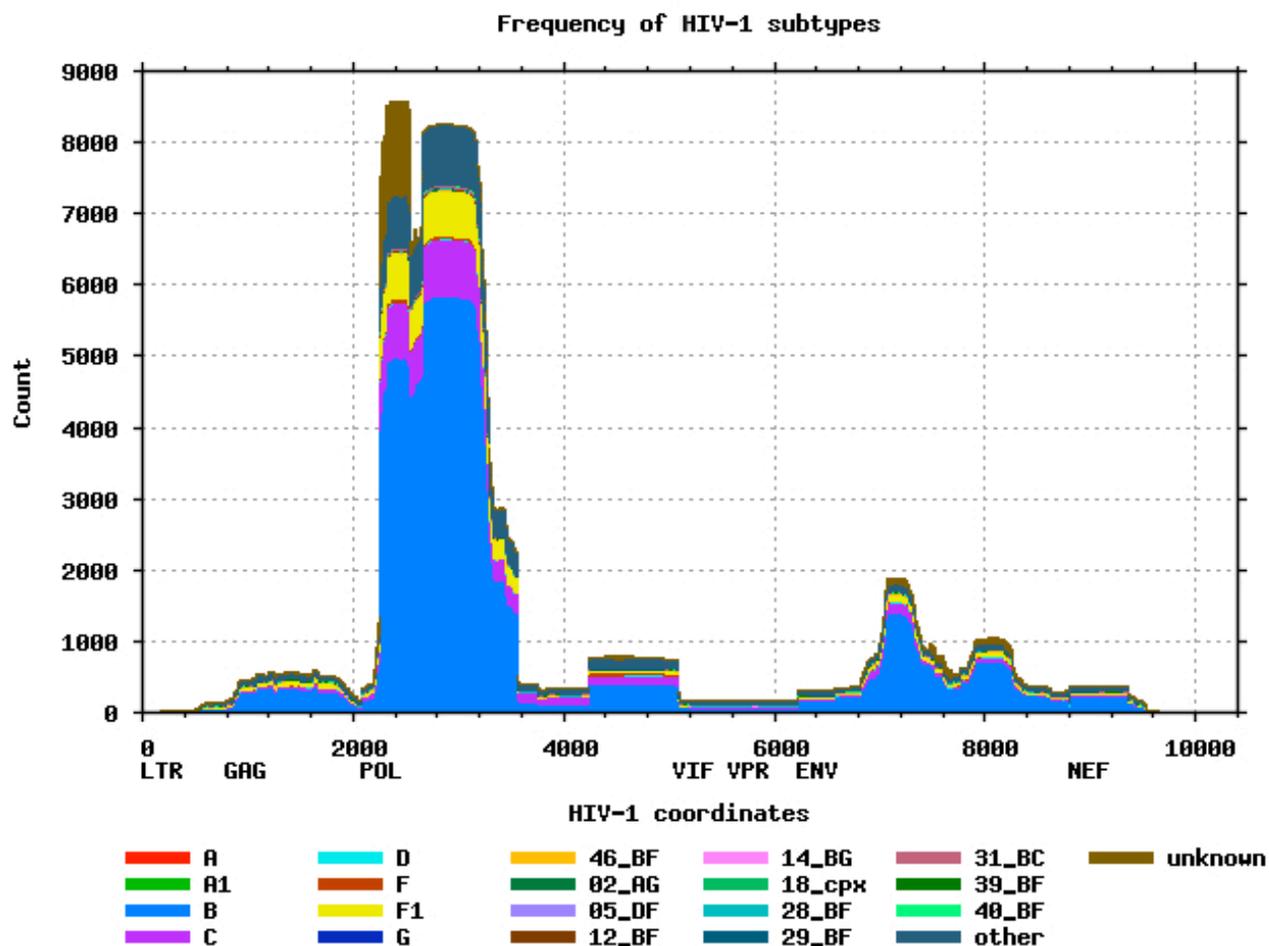
Database usage over time



2012 statistics

Hits: ~12 million downloads: 537 Gb visits: 364,000 visitors: 165,000
<http://www.hiv.lanl.gov/awstats/awstats.pl>

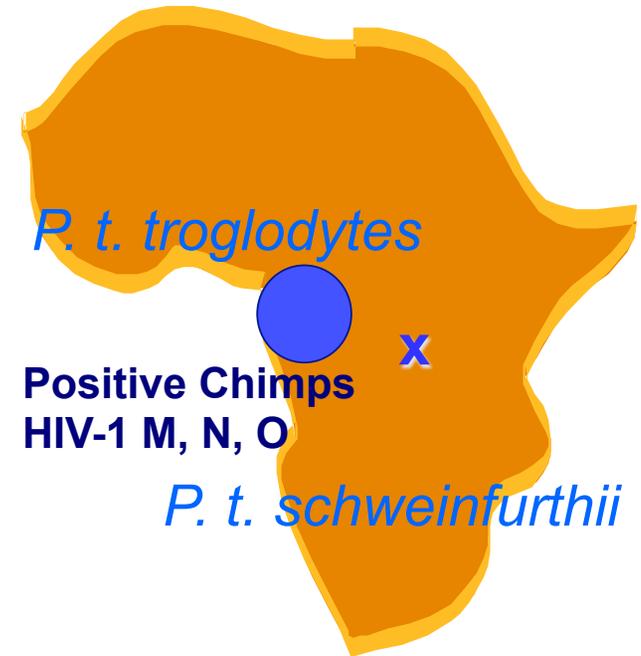
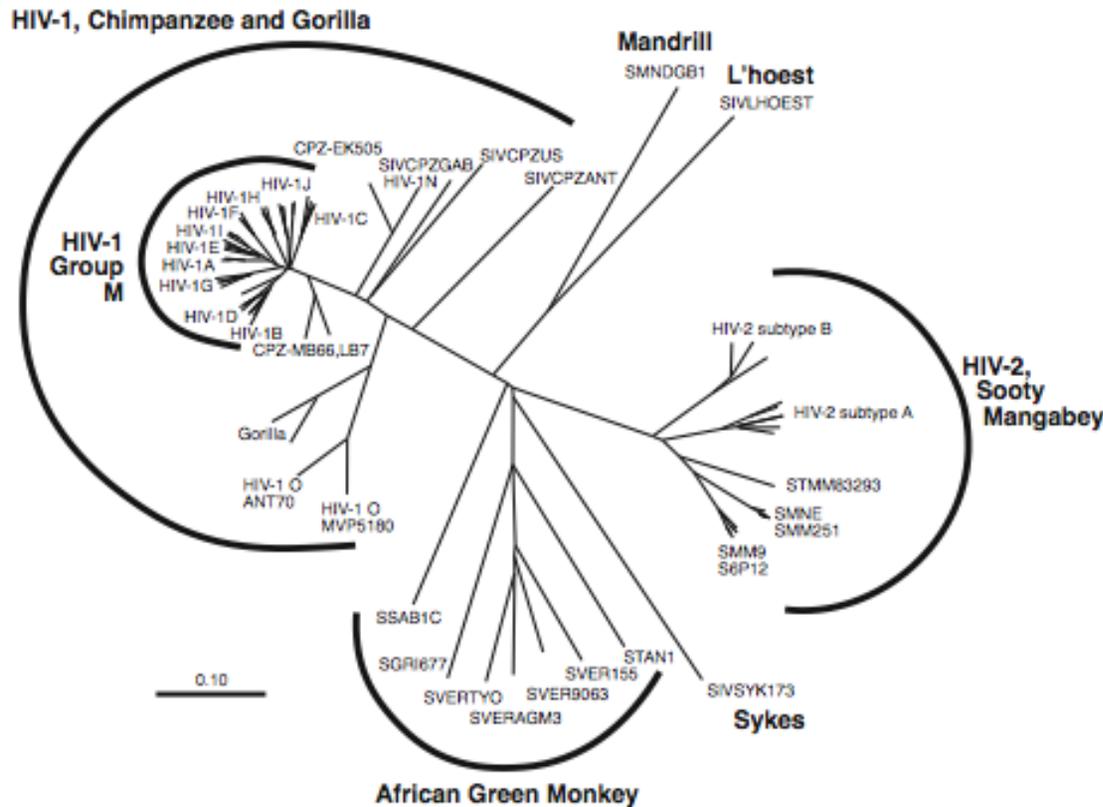
Histogram output



This histogram shows the distribution of sequences from your query across the entire HIV-1 genome. At each position across the genome, the number of sequences overlapping with that position is plotted. The colors represent different subtypes.

Primate Lentiviruses

Alignments: http://www.hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html



Van Heuverswyn, Nature 2006
 Keele, Science 2006
 Corbet, J. Virol 2000
 Foley, HIV database



HIV DATABASES

Entry page at <http://www.hiv.lanl.gov/>

The HIV databases contain data on HIV genetic sequences, immunological epitopes, drug resistance-associated mutations, and vaccine trials. The website also gives access to a large number of tools that can be used to analyze these data. This project is funded by the Division of AIDS of the National Institute of Allergy and Infectious Diseases (NIAID), a part of the National Institutes of Health (NIH). Click on any of the links below to access a database. [Editorial Board](#)

[SEQUENCE DATABASE ▶](#)

[VACCINE DATABASE ▶](#)

[IMMUNOLOGY DATABASE ▶](#)

[OTHER VIRUSES ▶](#)

News:

[Archived News ▶](#)

[New Features for Epitope Location Finder \(ELF\)](#)

ELF displays known and predicted epitopes found within a protein sequence query. ELF results now include both Class I (CTL) and Class II (helper) epitopes. In addition to predicting epitopes based on anchor residues, ELF now includes predictions from the Class I and Class II Binding Predictions tools at the Immune Epitope Database (IEDB). *13 March 2012*

[New Features for HIV BLAST](#)

HIV BLAST has new features. It now allows the user to find best matches among only subtyped sequences, or sequences of a specific subtype. It allows the resulting sequences to be downloaded fully aligned. *01 March 2012*

[New Option for N-GlycoSite](#)

The N-GlycoSite tool predicts N-linked glycosylation sites in amino acid sequences. A new option allows the user to exclude sites with a second-position proline, which is disfavored for N-linked glycosylation. *29 February 2012*

[HIV Antibody Search Results More Specific](#)

The antibody search interface in the HIV Immunology database is now more specific. Searches from the Author, Keyword, and Note fields now display only those notes and references that relate directly to the search. The user may still opt to display all, if desired *09 February 2012*

[New Options for Quickalign](#)

The Quickalign tool aligns any short protein or nucleotide sequence with database sequences. New options provide additional ways to calculate and display frequency by position, and allow the user to include the surrounding region in the alignment. *08 February 2012*

Questions or comments? Contact us at seq-info@lanl.gov



HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES Search Site

Search DB
Advanced Search
Intra-patient Search
Next-gen Sequences
Geography

HIV Sequence Database

Programs and Tools

[Search Interface](#) retrieves HIV and SIV sequences, which can then be aligned and used to build trees

[Geography Search Interface](#) retrieves HIV sequences based on geographical distribution

[Tools for working with sequences](#) lists all our online tools, organized by function

Alignments

[HIV Premade Alignments](#) includes Consensus and Ancestral Sequences, Subtype Reference Alignments, and Complete Alignments

Information

[HIV Sequence Compendium](#) print or order our annual publication

[Tutorials and other information](#) unpublished web-based content

[Links](#) to other HIV/AIDS tools and information

About this website

[FAQ](#) general information about this website

[Site Statistics](#) usage information for www.hiv.lanl.gov

[How to Cite this Database](#)

[Editorial Board](#)

News:

[Archived News](#)

[New Features for HIV BLAST](#)

HIV BLAST has new features. It now allows the user to find best matches among only subtyped sequences, or sequences of a specific subtype. It allows the resulting sequences to be downloaded fully aligned. *01 March 2012*

[New Option for N-GlycoSite](#)

The N-GlycoSite tool predicts N-linked glycosylation sites in amino acid sequences. A new option allows the user to exclude sites with a second-position proline, which is disfavored for N-linked glycosylation. *29 February 2012*

[HIV Antibody Search Results More Specific](#)

The antibody search interface in the HIV Immunology database is now more specific. Searches from the Author, Keyword, and Note fields now display only those notes and references that relate directly to the search. The user may still opt to display all, if desired. *09 February 2012*

[New Options for Quickalign](#)

The Quickalign tool aligns any short protein or nucleotide sequence with database sequences. New options provide additional ways to calculate and display frequency by position, and allow the user to include the surrounding region in the alignment. *08 February 2012*

last modified: Tue Jan 26 10:10 2010

Questions or comments? Contact us at seq-info@lanl.gov.

Operated by Los Alamos National Security, LLC, for the U.S. Department of Energy's National Nuclear Security Administration
Copyright © 2005-2006 LANSLC. All rights reserved | [Disclaimer/Privacy](#)



Search Interface

■ Help

- Tips at the top of the page are often overlooked
 - Ranges, operators, wildcards, logical groupings
- Mouse-over provides brief descriptions; click field names for details in Help file

■ Searches

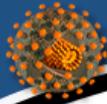
- Searches are case-insensitive
- Records are searchable through sequence, patient, genomic region, or publication information and can be matched to the genomic region of a user-provided alignment
- First seven fields will appear in search results page by default
- A "*" in a textbox will cause that field to be included in the results page
- Patient information (Infection year, Infection country) is different than sequence information (Sampling year and Sampling country)
- Problematic sequence filters (hypermutation, frequent ambiguities, potential contamination)

■ Analysis

- Build a tree with user alignment, search results and subtype reference sequences combined

■ Results

- Can download aligned or unaligned sequences
- Alignments are based on multiple pairwise alignments – alignments are good, but need hand editing for an optimal alignment
- Select all or a subset of sequences for download
- Sequences can be re-ordered by clicking on fields at the top of the page



Sequence Search Interface

Tips

- Click or mouse over the field name for specific tips
- The *italicized fields* are listed in output by default
- To list fields that are not listed by default or included in the search, put an asterisk (*) in the input box
- Use the + and - to see more or fewer search fields
- For other details about each field, see [Help](#) or [Data Dictionary](#)

Last [GenBank](#) update: 2012-02-08

[Advanced Search](#)

Sequence Information

Accession number

Sequence name

Sequence length

exact *Sampling year*

Sampling country

Virus

Subtype
No subtype
A
A1
A2
B

Include [recombinants](#)

More sequence information

We will search for country = Brazil (BR)



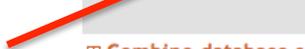
Find all sequences for a specific gene or region (HIV-1 and SIVcpz)

Genomic region
complete genome
5' LTR
5' LTR R
5' LTR U3
5' LTR U5
TAR

Or define *start* and *end*

Include [fragments](#) of minimum length

We will search for complete genomes.



Combine database sequences with your own sequence alignment (HIV-1 and SIVcpz)

Publication Information

Patient Information

Geographical Information

Output

Include [problematic](#) sequences % of non-ACGT

List records per page Show results selected Show SQL

[Advanced Search](#)

last modified: Wed Dec 7 14:05 2011

Results for HIV-1 complete genomes from Brazil

HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES Search Site

Make Tree Download Sequences Save Background Info Make Histogram Geography Clear

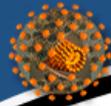
Displaying 1 - 100 of 151 sequences found:
 Note: 6 problematic sequences were removed from this result. Click here to repeat search to [include problematic sequences](#).

[Select all](#) [Unselect all](#) [Invert selection](#) [Show all](#) [One sequence/patient](#) [Select](#) record to [List](#) records per page

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism	
1	<input type="checkbox"/>	Blast BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990		9644	HIV-1
2	<input type="checkbox"/>	Blast BZ167(10007)	AB485642	BZ167	B	BRAZIL	1990		9662	HIV-1
3	<input type="checkbox"/>	Blast BZ163(4569)	AB485656	BZ163	F1	BRAZIL	1990		9602	HIV-1
4	<input type="checkbox"/>	Blast BZ163(4569)	AB485657	BZ163	F1	BRAZIL	1990		9602	HIV-1
5	<input type="checkbox"/>	Blast BR020(143)	AF005494	93BR020_1	F1	BRAZIL	1993		8968	HIV-1
6	<input type="checkbox"/>	Blast BR029(58)	AF005495	93BR029_4	BF1	BRAZIL	1993		8954	HIV-1
7	<input type="checkbox"/>	Blast BR004c(5320)	AF286228	98BR004	C	BRAZIL	1998		9016	HIV-1
8	<input type="checkbox"/>	Blast BZ167(10007)	AY173956	BZ167	B	BRAZIL	1989		8940	HIV-1
9	<input type="checkbox"/>	Blast BZ126(3090)	AY173957	BZ126	F1	BRAZIL	1989		9030	HIV-1
10	<input type="checkbox"/>	Blast BZ163(4569)	AY173958	BZ163	F1	BRAZIL	1989		8991	HIV-1
11	<input type="checkbox"/>	Blast RJ1(10882)	AY455778	99UFRJ_1	29_BF	BRAZIL	1999		8767	HIV-1
12	<input type="checkbox"/>	Blast BR97(10885)	AY455779	94BR_RJ_97	BF	BRAZIL	1994		8962	HIV-1
13	<input type="checkbox"/>	Blast RJ2(10886)	AY455780	99UFRJ_2	BF	BRAZIL	1999		9045	HIV-1
14	<input type="checkbox"/>	Blast BR41(15452)	AY455781	94BR_RJ_41	BF1	BRAZIL	1994		8864	HIV-1
15	<input type="checkbox"/>	Blast RJ16(10887)	AY455782	99UFRJ_16	46_BF	BRAZIL	1999		9002	HIV-1
16	<input type="checkbox"/>	Blast RJ9(10888)	AY455783	99UFRJ_9	BF	BRAZIL	1999		9040	HIV-1
17	<input type="checkbox"/>	Blast BR59(10884)	AY455784	94BR_RJ_59	BF	BRAZIL	1994		8898	HIV-1
18	<input type="checkbox"/>	Blast BR58(10883)	AY455785	94UFRJ_58	BF	BRAZIL	1994		8898	HIV-1

Choose “One sequence/patient” to remove very similar sequences (only available if a region is selected)



HIV sequence database

DATABASES

SEARCH

ALIGNMENTS

TOOLS

PUBLICATIONS

GUIDES

Search Site

Make Tree

Download Sequences

Save Background Info

Make Histogram

Geography

Clear

Tree options (only HIV-1 and SIVcpz)

Include HXB2 Reference Sequence (K03455)

Include subtype reference sequences

Show names as or [compose a label](#)

OK

Reset

Select a few sequences and make tree, allows us to add a reference set to our data and align them

Displaying 1 - 100 of 151 sequences found:

[Exclude problematic sequences](#)

[Select all](#)

[Unselect all](#)

[Invert selection](#)

[Show all](#)

[One sequence/patient](#)

[Select](#) record

to

[List](#)

100

records per page

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession Name	Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism
1	<input checked="" type="checkbox"/>	Blast BZ167(10007)	AB485641	BZ167	B	BRAZIL	1990		9644	HIV-1
2	<input type="checkbox"/>	Blast BZ167(10007)	AB485642	BZ167	B	BRAZIL	1990		9662	HIV-1
3	<input checked="" type="checkbox"/>	Blast BZ163(4569)	AB485656	BZ163	F1	BRAZIL	1990		9602	HIV-1
4	<input type="checkbox"/>	Blast BZ163(4569)	AB485657	BZ163	F1	BRAZIL	1990		9602	HIV-1
5	<input checked="" type="checkbox"/>	Blast BR020(143)	AF005494	93BR020_1	F1	BRAZIL	1993		8968	HIV-1
6	<input checked="" type="checkbox"/>	Blast BR029(58)	AF005495	93BR029_4	BF1	BRAZIL	1993		8954	HIV-1
7	<input checked="" type="checkbox"/>	Blast BR004c(5320)	AF286228	98BR004	C	BRAZIL	1998		9016	HIV-1
8	<input type="checkbox"/>	Blast BZ167(10007)	AY173956	BZ167	B	BRAZIL	1989		8940	HIV-1
9	<input checked="" type="checkbox"/>	Blast BZ126(3090)	AY173957	BZ126	F1	BRAZIL	1989		9030	HIV-1
10	<input type="checkbox"/>	Blast BZ163(4569)	AY173958	BZ163	F1	BRAZIL	1989		8991	HIV-1
11	<input checked="" type="checkbox"/>	Blast RJ1(10882)	AY455778	99UFRJ_1	29_BF	BRAZIL	1999		8767	HIV-1
12	<input checked="" type="checkbox"/>	Blast BR97(10885)	AY455779	94BR_RJ_97	BF	BRAZIL	1994		8962	HIV-1

TreeMaker tool

Choice of outgroup influences the the tree. In general, choose next closest sequences to the “ingroup”. In this case our Brazilian sequences are all HIV-1 M group.

HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES Search Site

Model parameters

Distance model

Gap handling Strip gaps before analysis treat as missing

Site rates Equal Gamma Shape

Reference sequences (TATCDS)

All A-K N, O, CPZ, CRFs Menu select only

A1.KE.1994.Q23_17.AF004885
A1.SE.1994.SE7253.AF069670
A1.UG.1985.U455_U455A.M62320
A1.UG.1992.92UG037.U51190
A1.UG.1998.98UG57136.AF484509

Outgroup

O.BE.1987.ANT70.L20587
 O.CM.1991.MVP5180.L20571
 O.CM.1998.98CMU2901.AY169812
 O.SN.1999.99SE-MP1299.AJ302646
 O.SN.1999.99SE-MP1300.AJ302647

Reference sequences

B.BR.1990.BZ167.AB485641
 B.BR.1990.BZ167.AB485642
 F1.BR.1990.BZ163.AB485656
 F1.BR.1990.BZ163.AB485657
 F1.BR.1993.93BR020_1.AF005494

Database sequences

Results link

Email a link to the results to this address with job title

Submit Reset

These settings minimally influence relative branch lengths, but rarely alter the tree topology.

Our Brazilian sequences

Optional mailback, and tree title

ATV java-based view
for quick look, cannot
save/print

HIV sequence

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES

Download Your Tree Results

This tree contains 59 sequences and is 7897 characters long, including insertions.

Phenogram:

- View Tree in ATV (a Java-based phylogenetic tree viewer)
- Download Phenogram (pdf)
- View Phenogram (png)

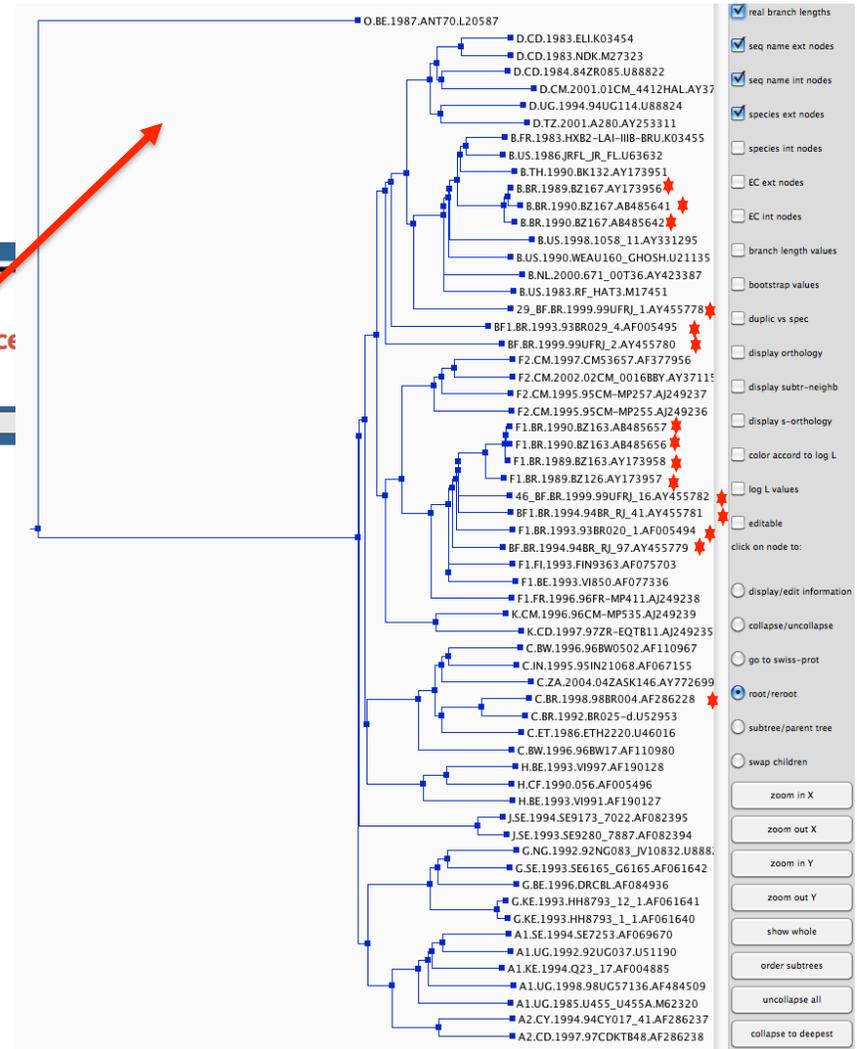
Radial:

- Download radial (unrooted) tree (pdf)
- View radial (unrooted) tree (png)

Alignment used for tree building

- Download fasta alignment (before gapstripping)
- Download fasta alignment in tree order (before gapstripping)
- Download fasta alignment (after gapstripping)
- Download Newick Tree File

last modified: Thu May 7 07:39 2009



Obtaining your sequences of interest and having them aligned to a good reference set was the whole point of this. The tree was just a first check on data and alignment quality.

Save alignment, run GeneCutter or use BioEdit or SeAl to view/adjust.

Save alignment, use BioEdit or SeAl to view/adjust.

Send alignment to GeneCutter or HIV-Align first, is usually best.

Download Your Tree Results

This tree contains 59 sequences and is 7897 characters long, including insertions.

Phenogram:

http://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html

- View Tree in ATV (a Java-based phylogenetic tree viewer)

- Download Phenogram

- View Phenogram

Radial:

- Download radial

- View radial (unrc)

Alignment used for tree

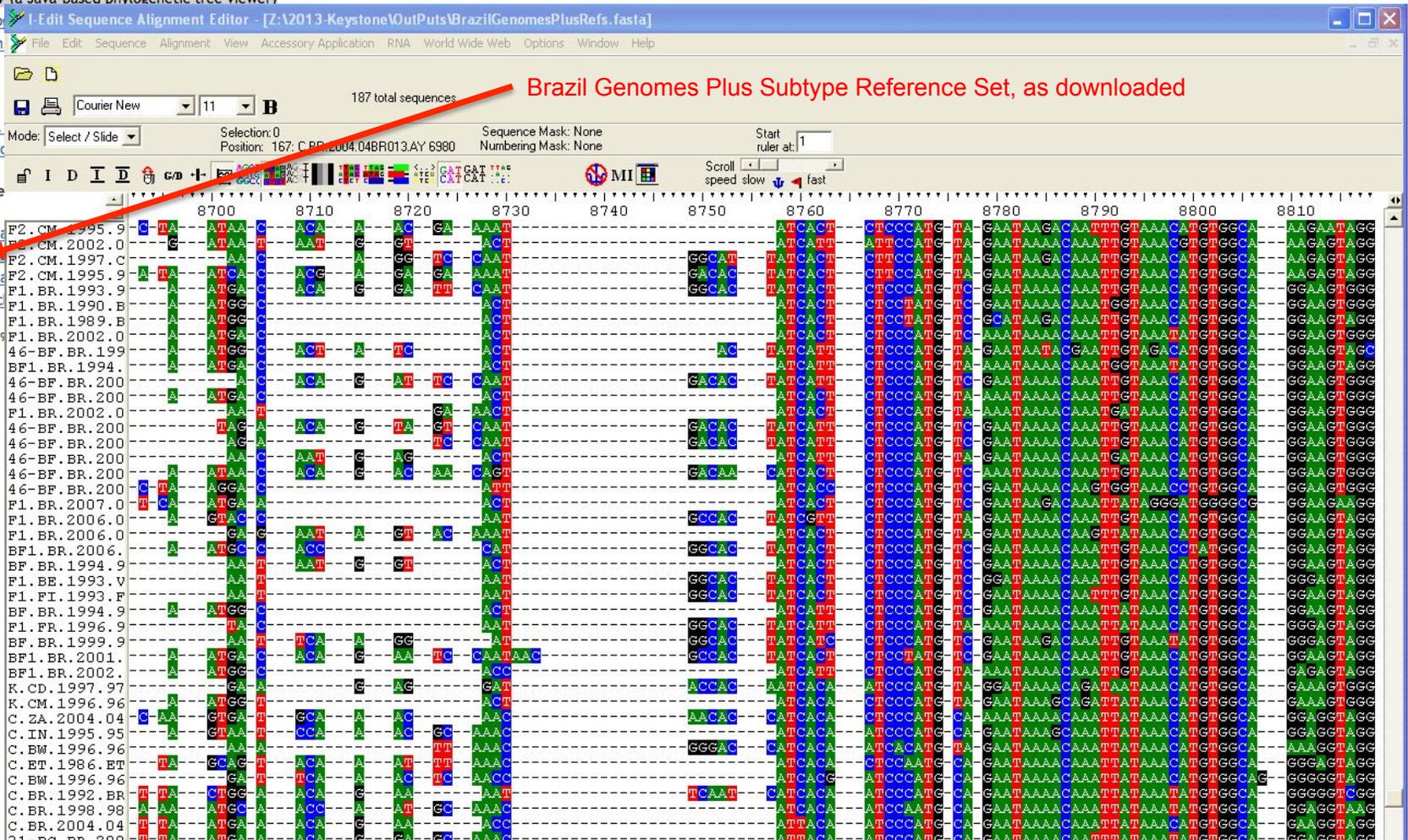
- Download fasta

- Download fasta

- Download fasta

- Download Newick

last modified: Thu May 7 07:39



Brazil Genomes Plus Subtype Reference Set, as downloaded

New search:
all complete
genomes;
then look at
geographic
and subtype
distribution of
the
sequences

HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES Search Site

Make Tree Download Sequences Save Background Info Make Histogram **Geography** Clear

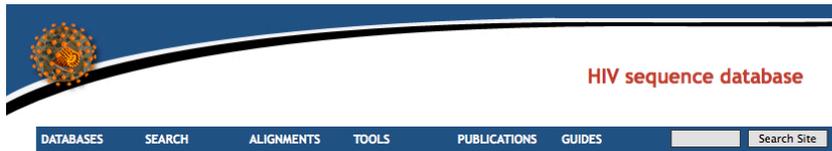
Displaying 1 - 100 of 5338 sequences found:
Note: 478 problematic sequences were removed from this result. Click here to repeat search to [include problematic sequences](#).

[Select all](#) [Unselect all](#) [Invert selection](#) [Show all](#) [Select](#) record to [List](#) records per page

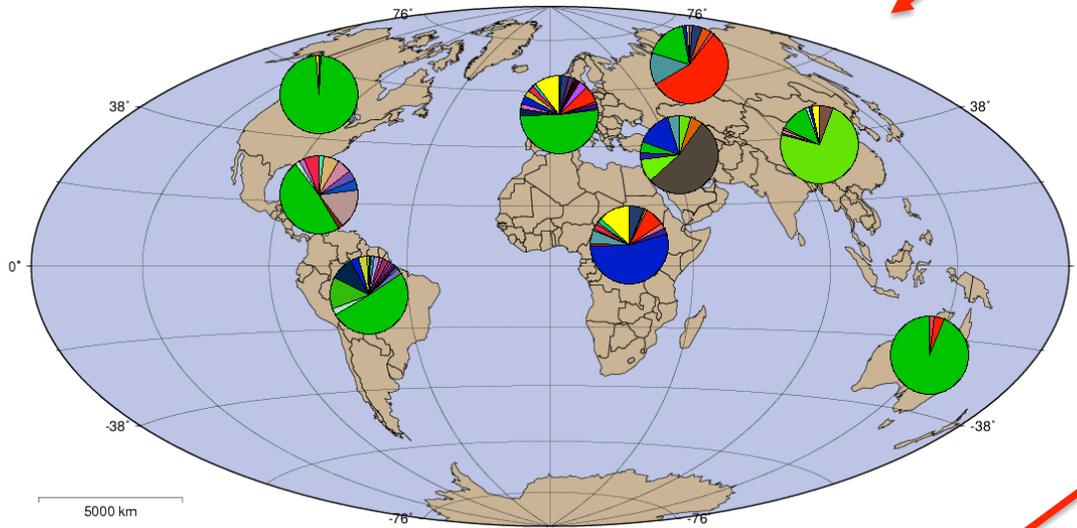
Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism
1	<input type="checkbox"/>	Blast LAI(19535)	A04321 IIB_LAI	B	FRANCE	1983		9193	HIV-1
2	<input type="checkbox"/>	Blast ELI(580)	A07108 ELI_patent	D	DEM REP OF CONGO	1983		9176	HIV-1
3	<input type="checkbox"/>	Blast MAL(578)	A07116 MAL_patent	A1DK	DEM REP OF CONGO	1985		9229	HIV-1
4	<input type="checkbox"/>	Blast LAI(19535)	A07867 LAI-J19	B	FRANCE	1983		9193	HIV-1
5	<input type="checkbox"/>	Blast ELI(580)	A14116 ELI_patent	D	DEM REP OF CONGO	1983		9176	HIV-1
6	<input type="checkbox"/>	Blast NDK(13796)	A34828 NDK_patent	D	DEM REP OF CONGO	1983		9143	HIV-1
7	<input type="checkbox"/>	Blast IN101(14294)	AB023804 93IN101	C	INDIA	1993		9680	HIV-1
8	<input type="checkbox"/>	Blast C1_husband(15892)	AB032740 95TNIH022	01_AE	THAILAND	1995		9427	HIV-1
9	<input type="checkbox"/>	Blast 47(881)	AB032741 95TNIH047	01_AE	THAILAND	1995		9430	HIV-1
10	<input type="checkbox"/>	Blast NJ97-42(24045)	AB049811 97GH-AG1	02_AG	GHANA	1997		9748	HIV-1
11	<input type="checkbox"/>	Blast	AB052867 97GH-AG2	02A1	GHANA	1997		9708	HIV-1
12	<input type="checkbox"/>	Blast NH1(717)	AB052995 93JP_NH1	01_AE	JAPAN	1993		9720	HIV-1
13	<input type="checkbox"/>	Blast NH2(715)	AB070352 NH25_93JPNH25T_93JP_NH2_5T	01_AE	JAPAN	1993		9731	HIV-1
14	<input type="checkbox"/>	Blast CS2(9760)	AB078005 ARES2	B	UNITED STATES	1997		9637	HIV-1
15	<input type="checkbox"/>	Blast 502(3272)	AB097865 mIDU502	01B	MYANMAR	2000		9046	HIV-1

Geography output

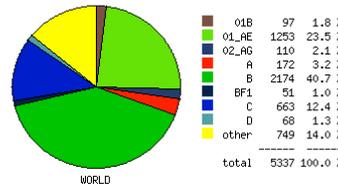


Distribution of all HIV-1 sequences: WORLD
 Please note that this map only includes sequences for which the sampling country is known.



GMT 2003 Mar 11 09:23:31 108M 1.2

Subtype distributions represent the frequency in the HIV Database and not the population
 About this geography site.
 Select organism:
 Select (if a country is selected, it supersedes region)
 or
 Show sequences
 Table [\(html\)](#) of the compiled subtype distribution.

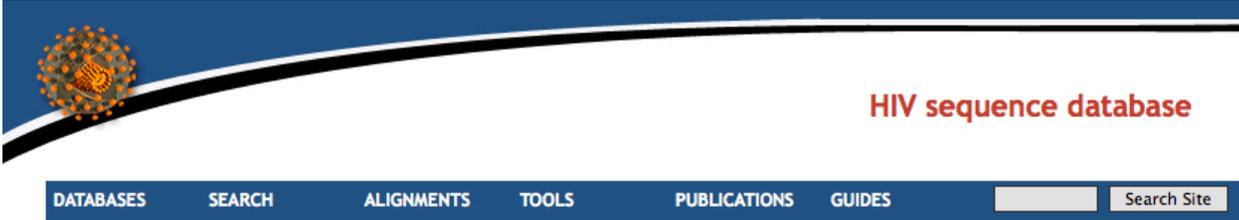


Each continent's pie chart is clickable to "zoom in" on that continent.

Likewise for each country once you are zoomed in to the continent level.

Most complete genomes in the HIV database are subtype B. But subtype C is more prevalent in human infections. Beware of this type of sampling bias.

Click through pie to get (e.g.,) all sequences from Brazil



HIV sequence database

DATABASES SEARCH ALIGNMENTS TOOLS PUBLICATIONS GUIDES Search Site

Displaying 1 - 100 of 15489 sequences found:

Note: 87 problematic sequences were removed from this result. Click here to repeat search to [include problematic sequences](#).

 record to
 records per page

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession Name	Subtype	Country	Sampling Year	Genomic Region	Sequence Length	Organism
1	<input type="checkbox"/>	Blast BZ167(10007)	AB485641 BZ167	B	BRAZIL	1990		9644	HIV-1
2	<input type="checkbox"/>	Blast BZ167(10007)	AB485642 BZ167	B	BRAZIL	1990		9662	HIV-1
3	<input type="checkbox"/>	Blast BZ163(4569)	AB485656 BZ163	F1	BRAZIL	1990		9602	HIV-1
4	<input type="checkbox"/>	Blast BZ163(4569)	AB485657 BZ163	F1	BRAZIL	1990		9602	HIV-1
5	<input type="checkbox"/>	Blast RJ100(4)	AF000238 RJ100	D	BRAZIL	1996		424	HIV-1
6	<input type="checkbox"/>	Blast BR020(143)	AF005494 93BR020_1	F1	BRAZIL	1993		8968	HIV-1
7	<input type="checkbox"/>	Blast BR029(58)	AF005495 93BR029_4	BF1	BRAZIL	1993		8954	HIV-1
8	<input type="checkbox"/>	Blast BR003(655)	AF009369 92BR003	B	BRAZIL	1992		1176	HIV-1
9	<input type="checkbox"/>	Blast BR004a(656)	AF009370 92BR004	B	BRAZIL	1992		1175	HIV-1
10	<input type="checkbox"/>	Blast BR017(657)	AF009371 92BR017_A	B	BRAZIL	1992		1174	HIV-1
11	<input type="checkbox"/>	Blast BR018(658)	AF009372 92BR018_A	B	BRAZIL	1992		1174	HIV-1
12	<input type="checkbox"/>	Blast 92BR019(72)	AF009373 92BR019_A	B	BRAZIL	1992		1176	HIV-1
13	<input type="checkbox"/>	Blast 92BR020(8574)	AF009374 92BR020_A	B	BRAZIL	1992		1176	HIV-1
14	<input type="checkbox"/>	Blast BR021(8563)	AF009375 92BR021a	B	BRAZIL	1992		1173	HIV-1

Tools

■ Analysis and Quality Control

- **HIV BLAST** finds sequences similar to yours in the HIV database.
- **N-Glycosite** finds potential N-linked glycosylation sites.
- **RIP 3.0** (Recombinant Identification Program) detects HIV-1 subtypes and recombination.

■ Alignment and sequence manipulation

- **HIValign** uses our HMM alignment models to align your sequences.
- **Gapstreeze** removes columns with more than a given % of gaps.
- **ElimDupes** Given an alignment or set of unaligned nucleotide or protein sequences, this tool compares the sequences and eliminates any duplicates.

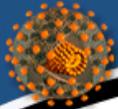
■ Phylogenetics

- **TreeMaker** generates a neighbor-joining phylogenetic tree.
- **PhyML** generates a maximum likelihood phylogenetic tree.
- **TreeRate** finds the phylogenetic root of a tree and calculates evolutionary rate.

■ Format and display

- **Protein Feature Accent** provides an interactive 3-D graphic of HIV proteins; the user can map a sequence feature (a short functional domain, epitope, or amino acid) and see where it occurs spatially in the 3D structure.
- **Highlighter** highlights mismatches, matches, transition and transversion mutations, and silent and non-silent mutations in an alignment of nucleotide sequences.
- **SeqPublish** makes alignment publication-ready.
- **Recombinant HIV drawing tool** highlights regions of the genome on a graphical representation

The HIV database sequence analysis tool set



HIV sequence database

DATABASES	SEARCH	ALIGNMENTS	TOOLS	PUBLICATIONS	GUIDES	Search Site
			Index of all tools	HIV BLAST	Quality Control	
			ADRA	HIVAlign	Quick-Align	
			Branchlength	Hypermot	Rainbow Tree	
			Codon Alignment	jpHMM at GOBICS	Recombinant HIV-1 Drawing Tool	
			Consensus Maker	Mosaic Vaccine Tool Suite	RIP	
			ELF	Motif Scan	SeqPublish	publication
			ElimDupes	N-Glycosite	Sequence Locator	ed content
			Entropy	PCOORD	SNAP	
			FindModel	PepMap	SUDI Subtyping	
			Format Converter	PeptGen	SynchAlign	
			Gap Strip/Squeeze	PhyloPlace	Translate	
			GenBank Entry Generation	PhyML	TreeMaker	
			Gene Cutter	Pixel	TreeRate	
			Heatmap	Poisson-Fitter	VESPA	
			Hepitope	Protein Feature Accent	External Tools	
			Highlighter	Protein Structure		

HIV

Click top level to link to full page of tools

Programs and Tools

[Search Interface](#) retrieves HIV and SIV sequences, which can be aligned and used to build trees

[Geography Search Interface](#) retrieves HIV sequences based on geographical distribution

[Tools for working with sequences](#) lists all our online tools, by function

Alignments

[HIV Premade Alignments](#) includes Consensus and Ancestral Sequences, Subtype Reference Alignments, and Complete Alignments

News:

[Archived News](#) ▶

[Sequence Locator improved output for multiple queries](#)

For input of multiple sequences, Sequence Locator now provides links to download the summary information as tab-delimited text files of coordinates. 05 December 2012

last modified: Tue Jan 26 10:10 2010

HIV Database Tools

(alphabetical order within category)

For detailed descriptions, mouse over the links.

Analysis and Quality Control

[Entropy](#) quantifies positional variation in an alignment using Shannon Entropy

[HIV BLAST](#) finds sequences similar to yours in the HIV database

[Hypermut](#) detects hypermutation

[jpHMM at GOBICS](#) detects subtype recombination in HIV-1; hosted at GOBICS as a collaboration between the Department of Bioinformatics, University of Göttingen and the Los Alamos HIV Sequence Database

[N-Glycosite](#) finds potential N-linked glycosylation sites

[PCOORD](#) multidimensional analysis of sequence variation

[Quality Control](#) runs several tools to allow quick QC analysis of HIV-1 sequences; optional step prepares sequence submission for GenBank

[RIP](#) (Recombinant Identification Program) detects HIV-1 subtypes and recombination

[SNAP](#) calculates synonymous/non-synonymous substitution rates

[SUDI Subtyping](#) plots the distance of your sequence to established subtypes

[VESPA](#) (Viral Epidemiology Signature Pattern Analysis) detects residues with different frequencies in two sequence sets

Alignment and sequence manipulation

[Codon Alignment](#) takes a nucleotide alignment and returns a codon alignment and translation

[Consensus Maker](#) computes a customizable consensus

[ElimDupes](#) compares the sequences within an alignment and eliminates any duplicates

[Gap Strip/Squeeze](#) removes columns with more than a given % of gaps

[Gene Cutter](#) clips genes from a nucleotide alignment, codon-aligns, and translates

[HIValign](#) uses our HMM alignment models to align your sequences

Phylogenetics

[Branchlength](#) calculates branch lengths between internal and end nodes

[FindModel](#) finds which evolutionary model best fits your sequences

[PhyloPlace](#) reports phylogenetic relatedness of an HIV-1 sequence with reference sequences

[PhyML](#) generates much better trees than our simple TreeMaker tool

[Poisson-Fitter](#) estimates time since MRCA and star-phylogeny. For use with acute (low diversity) samples.

[TreeMaker](#) generates a quick-and-dirty phylogenetic tree

[TreeRate](#) finds the phylogenetic root of a tree and calculates evolutionary rate

Immunology

[ELF](#) (Epitope Location Finder) identifies known and potential epitopes within peptides

[EpiIgen \(QuickAlien\)](#) aligns a protein sequence (e.g., epitope) to the appropriate protein alignment

[Heatmap](#) displays a table of numbers by using colors to represent the numerical values

[Hepitope](#) identifies potential epitopes based on HLA frequencies

[Mosaic Vaccine Tool Suite](#) designs and assesses polyvalent protein sequences for T-cell vaccines

[Motif Scan](#) finds HLA anchor motifs in protein sequences for specified HLA serotypes, genotypes or supertypes

[PeptGen](#) generates overlapping peptides from a protein sequence

Database search interfaces

[ADRA](#) Antiviral Drug Resistance Analysis, a resistance mutation database

[Advanced Search](#) creates a custom search interface

Tools are organized in groups by function/purpose.

Most tools have explanation pages, and sample data sets.

Many tools were inspired by user comments, please ask for more.

[SynchAlign](#) aligns overlapping alignments to one another

[QuickAlign \(formerly Epilign and Primalign\)](#) aligns a nucleotide or protein sequence (e.g., primer or epitope) to the appropriate genome alignment

[Codon Alignment](#) takes a nucleotide alignment and returns a codon alignment and translation

[ElimDupes](#) compares the sequences within an alignment and eliminates any duplicates

[Pixel](#) generates a PNG image of an alignment using 1 or more colored pixel(s) for each residue

[PepMap](#) can be used to map epitopes, functional domains, or any protein region of interest

Format and display

[Protein Feature Accent](#) provides an interactive 3-D graphic of HIV proteins; can map a sequence feature (a short functional domain, epitope, or amino acid) and see it spatially

[Format Converter](#) converts between alignment formats

[SeqPublish](#) makes publication-ready alignments

[Highlighter](#) highlights mismatches, matches, transitions and transversion mutations and silent and non-silent mutations in an alignment of nucleotide sequences

[Recombinant HIV-1 Drawing Tool](#) creates a graphical representation of your HIV-1 intersubtype recombinant

[Protein Structure Analysis](#) provides a visualization tool for protein sequence properties

[Advanced Search](#) creates a custom search interface

[Geography](#) shows the geographic distribution of sequences in the database

[CTL/CD8+ Search](#) searches for CD8+ epitopes by protein, immunogen, HLA, author, keywords

[T-Helper/CD4+ Search](#) search for CD4+ epitopes by protein, immunogen, HLA, author, keywords

[Antibodies](#) search for HIV antibodies by protein, immunogen, AB type, isotype, author, keywords

[Vaccine Trials Database](#) finds past vaccine trials and their results

[ADRA](#) Antiviral Drug Resistance Analysis, a resistance mutation database

Other tools

[HDent and HDdist](#) perform analysis of heteroduplex mobility shifts

[ODprep and ODfit](#) calculate antibody titers based on concentration and optical density data

External tools

[External tools](#) lists tools and programs on other websites

We tend to list only tools of great use in HIV research. Many of these tools are essential, such as either BioEdit or SeAl for alignment viewing and correction.

<http://www.hiv.lanl.gov/content/sequence/HIV/HIVTools.html>

Pre-Built Sequence alignments

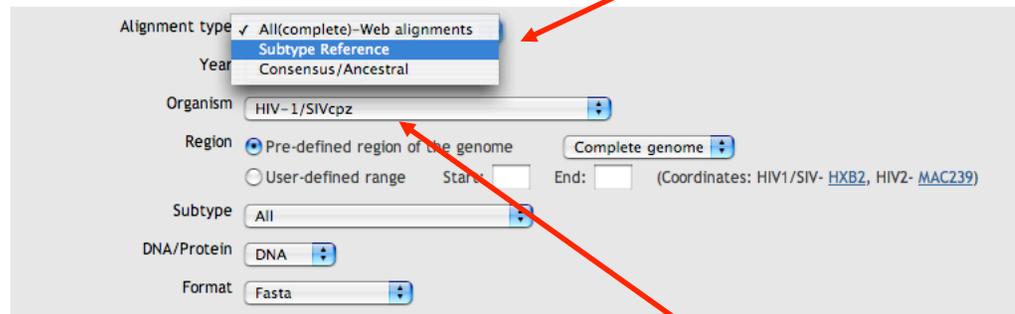
- Originally based on iterations of manual and HMM alignments
- Yearly updates using HMM and manual corrections
- Alignments are in reading frame (codon aligned)
- Contain non-redundant data (one sequence per patient)
- Compendium alignments show fewer sequences than web version
- Reference alignments contain up to four representatives of each subtype. One of each CRF.
- Protein alignments may contain frameshift compensations
- Subtype consensus with ties resolved, as well as maximum likelihood ancestors, are available for reagent production
- Special interest alignments are being added
 - Sequence sets of particular research interest
 - Suggestions welcome to tkl@lanl.gov

HIV Sequence Alignments

- The web alignments provides nucleotide and protein alignments that represent the full spectrum of HIV and SIV sequences in the database.
- The subtype reference alignments contain approximately 4 representatives of each subtype, and are useful for classifying new sequences.
- The consensus/ancestral sequences of genetically associated subsets of HIV-1 sequences include a consensus of each subtype, an M-group consensus-of-consensuses, and some ancestral sequences.

Before use, please read the additional information below.

Options



Alignment type: All(complete)-Web alignments, Subtype Reference, Consensus/Ancstral

Year:

Organism:

Region: Pre-defined region of the genome, User-defined range. Complete genome:

Subtype:

DNA/Protein:

Format:

• Web alignments

What sequences are included

The alignments presented on the web differ from the ones that are printed in the compendia. The whole genome alignments are complete, meaning that they contain all complete genome sequences we have, including very similar ones.

The gene/protein alignments contain all complete gene sequences we have, with the important exception that very similar sequences (e.g. multiple clones from one isolate, multiple sequences from one person) have been deleted. The selection was made on the basis of phylogenetic trees: from tight clusters of sequences, one representative was retained and the others were removed from the alignment. An exception has been made for HXB2 and LAI, as these are important lab strains that are frequently used in experiments.

All (complete) = one per patient, all sequences for which we have a complete genome.

Subtype Reference = 4 representatives of each subtype, plus one of each circulating intersubtype recombinant form (CRF) of the M group, plus 4 O group, N group, P group and SIV-CPZ

Consensus/Ancstral computed from master alignment periodically.

HIV-2/SIV-SMM and primate lentivirus alignments also available here.

Gene Cutter

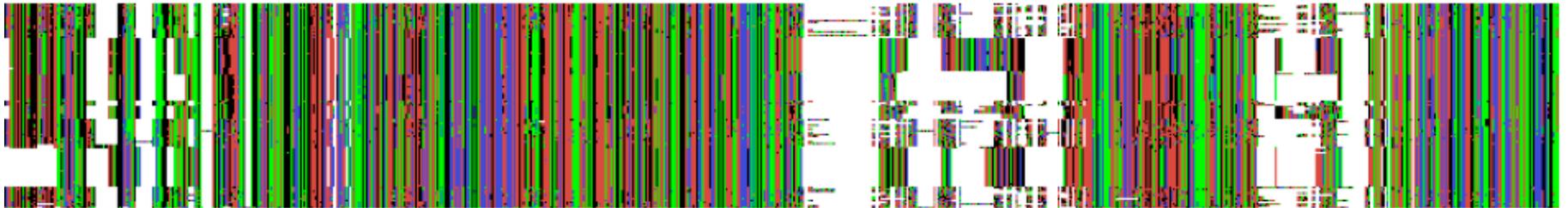
- Unconventional Alignment/Homology program specific for HIV/SIV
- “Cuts out” specified genes and proteins from sets of DNA sequences
 - Aligns to HXB2 via HMMer (or to SIV-Mac239 for HIV-2 and SIV-SMM)
 - Splits input sequences into genes, if desired
 - Aligns DNA sequences by codon, and translates them with interpretation of IUPAC ambiguity codes (e.g. R/Y for purine/pyrimidine)
- Useful for processing new sequence data
 - annotating full length genomes
 - pulling out regions of interest from raw sequence data
- For each gene/region, maintains a list of anomalies
 - stop codons
 - codons containing multi-state characters
 - codons containing indels
- Input sequences may be aligned or unaligned
- Results may be better if the HXB2 sequence is included as a reference in your input file

GeneCutter Result

Alignments viewed with Pixel

<http://www.hiv.lanl.gov/content/sequence/pixel/pixel.html>

Our data aligned to reference set by search tool:
(output of search and tree build was input to GeneCutter)



Our data aligned to reference set by GeneCutter:



Can also be viewed with BioEdit, Se-AL or other multiple
sequence alignment editors.

Treemaker

Check for phylogenetic relatives:

- TreeMaker produces a Neighbor Joining tree for a quick comparison
- TreeMaker uses PAUP* for its calculations; a few model options are available
- Reference sequences can be included, and are aligned to the input automatically
- Trees are displayed using PHYLIP and ATV
- The alignment used for the tree can also be downloaded
- A PhymI interface is also available

<http://www.hiv.lanl.gov/content/sequence/PHYML/interface.html>

http://www.hiv.lanl.gov/components/sequence/HIV/treemaker/treemaker.html

Neighbor TreeMaker

Purpose: This tool takes a nucleotide sequence alignment, converts it to NEXUS format, and uses PAUP to generate a tree, which is displayed using the [PHYLIP](#) programs Drawgram or Drawtree.

Details: After sequence input, the next page will give additional options. Gaps can be treated as missing or stripped. The user can choose from various distance models and select the outgroup sequence. A version of the input alignment in which the sequences have been reordered to match the order in the tree may be downloaded. Trees are calculated using the neighbor-joining method. You can use [FindModel](#) to decide what evolutionary model best fits your data.

Disclaimer: This interface only offers very basic, 'quick-and-dirty' phylogenetic analysis. More in-depth analysis is usually needed. For more information see the [Tree Tutorial](#).

Input

Paste alignment here
[\[Sample Input\]](#)

or upload your file

Paste or type a DNA alignment here.

OR upload an alignment file here.

Tree parameters

Include reference sequences (HIV-1/CPZ only)

HIV/SIV Sequence Locator Tool

- Instantly computes position numbers of DNA or protein fragments relative to a reference strain (HXB2r for HIV-1, SMM239 for SIV)
 - Such numbers, often included in the literature, are frequently incorrect
- Shows the location of the sequence on an HIV map
- Presents protein translations of DNA sequences
- Can be used for input into the search interface, to align a new sequence you have generated with the database set
- Can also retrieve reference sequences
 - by coordinates (range of base or amino-acid positions)
 - by single position (retrieves flanking sequences)

<http://www.hiv.lanl.gov/content/sequence/LOCATE/locate.html>

HIV Sequence Locator Tool

Purpose: This tool has several purposes. It can find the start and end coordinates (relative to the reference strain HXB2) of your input sequence(s) and show which genes or proteins it covers, along with a graphical view of the location of your sequence(s) relative to the reference sequence. The tool will display both the nucleotide sequence and protein translation of your input as it aligns to HXB2. It will also check the reverse complement of your input sequence, and report the orientation with the best match. Another use is to retrieve a section of the HXB2 reference sequence based on its coordinates.

How to use: To find the coordinates for your sequence, either upload or paste your sequence (any format) in the box below, or (for database sequences only) enter GenBank accession numbers. To retrieve the HXB2 sequence for a set of coordinates (see [HIV coordinate map](#)), enter the coordinates and choose the region. To retrieve the entire gene or protein, enter coordinate values of "1" and "end". To retrieve a single nucleotide or range with its surrounding 42-nucleotide sequence, enter the single coordinate in the "from" field and check the box. For more details, see [Sequence Locator Explanation](#).

Useful Links:

[HXB2 numbering](#) | [SIVmm239 numbering](#) (review articles)

[HXB2 spreadsheet](#) | [SIVmm239 spreadsheet](#) (spreadsheets with base-by-base annotation)

Find the location of a sequence

Sequence type Let program decide HIV SIV

Paste your input here
[\[Sample Input\]](#)

or upload your file

Paste or type a DNA or protein sequence here.

-- OR --

Retrieve a region by its coordinates

Enter coordinates: from to (Enter '1' and 'end' to retrieve the entire region.)

Region

Retrieve Nucleotide or protein output

include surrounding region

OR enter numeric coordinates here.

Sequence Locator:

Table of genomic regions touched by query sequence. Query protein translation in blue.				
CDS	NA position relative to CDS start in HXB2	NA position relative to query sequence start	NA position relative to HXB2 genome start	AA position relative to protein start in HXB2
Gag	352 -> 483	1 -> 132	1141 -> 1272	118 -> 161
AAADTGH SNQVSQNYPIVQNIQGQMVHQ AI SPRTL NAWVKV EE				
p17	352 -> 396	1 -> 45	1141 -> 1185	118 -> 132
AAADTGH SNQVSQNY				
p24	1 -> 87	46 -> 132	1186 -> 1272	1 -> 29
PIVQ NIQGQMVHQ AI SPRTL NAWVKV EE				

Sequence below includes up to 42 bases of context surrounding **query sequence**.

Reference Strain	Type	Region	Start	End
HXB2	nuc	complete	1141	1272
Retrieved Sequence:				
GCAGCAGCTGACACAGGACACAGCAATCAGGTCAGCCAAAATTACCCTATAGTGCAGAACATCCAGGGGCAAATGGTACA TCAGGCCATATCACCTAGAACTTTAAATGCATGGGTAAAAGTAGTAGAAGAG				

Organism: HIV

Hypermutation

Hypermur 2.0

Analysis & Detection of APOBEC-induced Hypermur

Purpose: This interface takes a nucleotide alignment and documents the nature and context of nucleotide substitutions in a sequence population relative to a reference sequence.

Details: The first sequence in the input alignment will be used as the reference sequence, and each of the other sequences will be used as a query sequence. Please choose the reference sequence carefully. For example, for an inpatient set, the reference should probably be the most common form in the first sampled time point; for a set of unrelated sequences, the reference should probably be the consensus sequence for the appropriate subtype. Before using, please read:

- [Hypermur Explanation](#)
- [Hypermur 2.0 Details](#)

References: Please reference these articles when using Hypermur:

- Rose, PP and Korber, BT. 2000. Detecting hypermutations in viral sequences with an emphasis on G -> A hypermutation. *Bioinformatics* 16(4): 400-401.
- Bruno, WJ, Abfalaterer, WP, Foley, BT, Leitner, TK and Korber, BT. Detection of hypermutation in HIV sequences using two context positions and avoiding nucleotide content effects. Manuscript submitted.

Input

Indicate [sequence format](#) of input:

Note: Sequences must be aligned, in-frame if possible, and of equal length.

Paste alignment here:

```
>Seq1
CAACTGCTGTTAAATGGCAGTCTAGCAGAAGAAGAGGTAATA
GATCTGAAAATTCACGAATAATG
CTAAAATCATAATAGTACAGTTGAATGAAATCTGTAATAATTG
TATAAGACCCAAACAATAACAAG
AAAAAGTATACATATCGGACCCAGGGAGGCATTTTACACAACAGGA
```

Or upload alignment file: no file selected

Restrict analysis to subregion of alignment from bp to bp (optional)

Hypermur 2.0 Customized Options

These options apply only to Hypermur 2.0 analysis, and have no effect on the Original Hypermur output. For typical analyses of APOBEC-induced hypermutation in HIV, these options should be left in their default settings.

Customize Hypermur pattern:

Mutation

Upstream context: ↓ Downstream context:

Enforce context:

On reference sequence

On both sequences

On query sequence

Customize control pattern:

↓

Output

Analyses to perform: Both Original Hypermur Hypermur 2.0

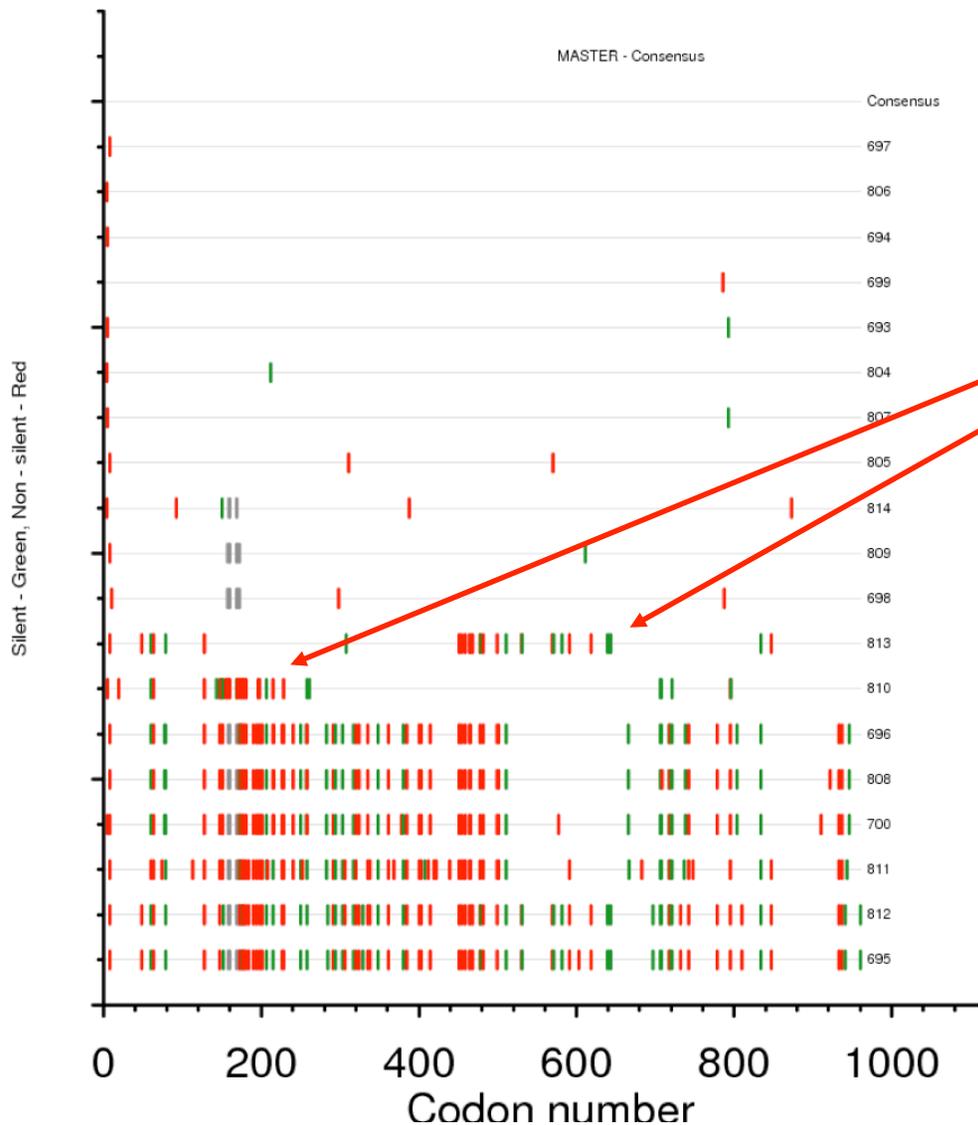
- Detects APOBEC related A->G hypermutation as default

- Can be adapted to detect any fuzzy motif in relation to a control pattern

Highlighter

- Highlights mutations relative to a reference strain, particularly useful for intra-patient analyses.
- Highlights:
 - syn/non-syn
 - transition/transversion
 - Apobec motifs
- Sorts on similarity
- Visualize recombination of closely related sequences

Sequences compared to master

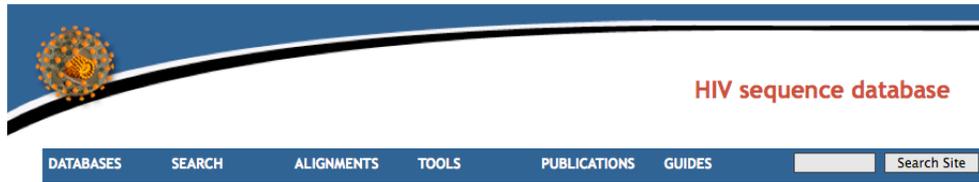


Nonrandom distribution of mutations evident.

Sample Set is from a possible dual Infection, with intra-subtype recombinants evident.

Protein Feature Accent

- Highlights region of interest in an HIV structure
- You can upload a PDB structure, or use one of our annotated Env structures
- You can upload your own alignment and get an entropy map



Protein Feature Accent

This is a beta version!

Some capabilities are not fully implemented, and there may be bugs or other problems. Please use with care and a sense of humor. This tool requires that [Java](#) be installed on your computer.

Purpose: The Protein Feature Accent tool is a quick way to map protein sequence features (for example a short functional domain or an epitope) from a sequence directly on to an interactive graphic of the corresponding 3-D structure of the protein.

How to use: The tool needs only to be directed to use a particular protein structure file in [PDB](#) format. Uploading a sequence is not required; the sequence associated with the chosen structure will always be displayed. Any sequences you do provide will be analyzed (for [entropy](#), etc.), aligned with the structure sequence, and displayed.

If you prefer, you may upload a PDB file for a structure you wish to use instead of those available here. [Click here](#) for a list of all the structures available.

New features:

- Predicted [N-linked glycosylation site](#) highlighting
- User-supplied [alignment entropy](#) color scheme
- [PDB file upload](#) option

We are in the process of adding additional features to the tool.

Select a protein structure: **gp120**
[\[switch to full structure list\]](#)

- 1G9M: HIV-1 HXBC2 GP120 ENVELOPE GLYCOPROTEIN COMPLEXED WITH CD4 A ...
- 2B4C: HIV-1 JR-FL GP120 CORE PROTEIN CONTAINING THE THIRD VARIABLE ...
- 2NY7: HIV-1 GP120 ENVELOPE GLYCOPROTEIN COMPLEXED WITH THE BROADLY ...
- 1RZK: HIV-1 YU2 GP120 ENVELOPE GLYCOPROTEIN COMPLEXED WITH CD4 AND ...

OR

upload a PDB file:

You may provide an amino acid sequence alignment (or a single unaligned sequence) below:

Paste your sequence(s) here

```
>B.BR.99.BREP11931_DQ085869
MRVRETPKKNYQWNRGMLLGLMLICSAEQSWVTYYGVPVWKEASTTLCASDAKAVETEAHNVWAT
HACVPTDPNPQEVVLENVTENFNMMKNNMVEQMHEDIISLWDQSLKPCVKLTFPCETKMSCNVDNATSDT
NSTNSGWEKMAEIRNCSFNVTNINGNKROKEYALFNKLDVVPIDNTSYTLINCNTSVITQACPKISFEP
IPIHYCTPAGFALLKCNDDKFKNGTGPCKNVSTVQCTHGRVVPVSTQLLNGSLAEAEIIVIRSENPTNNAK
TIIIVQLNKTVVINCTRPNNNTRKGIHLGPRTVYATGGIIGNIRQAHCNISGAEWENTLRQIATKLGQCF
KNKTIAFNQSSGDPPEITMHSFNCGGEFFYCNTTQLFNSTWYTWNRNGTNGTITLPCRIKQIINRWQ
```

or upload a sequence file:

List of "recommended" PDB entries

Only a gp120 alignment is provided so far. We hope to add others. You can paste in your own.

<http://www.hiv.lanl.gov/content/sequence/PROTVIS/html/protvis.html>

Jmol window The viewing window below offers [Jmol's interactive features](#), in addition to the control panel at the left.

The screenshot shows the Jmol web interface. On the left is a 'Control Panel' with navigation buttons (arrows, zoom in/out), 're-center', 'spin', 'select display style', 'pick color scheme', and a 'Background' selector. The main window displays a 3D ribbon structure of a protein, with a specific region highlighted in red. Below the structure is a 'Jmol command script:' input field and an 'execute' button. At the bottom is a 'Sequence View' section with checkboxes for 'show PDB file annotation', 'show reference sequence', and 'show reference sequence annotation'. A sequence alignment is shown with a red highlight under the residues 'TLKQIVIKLREQ' in the PDB file sequence. A 'download' button and 'this view as a [jpg] image' option are also present.

Many display options in Jmol are “built in” to this web tool. Use the Jmol command script box below for other commands.

One of the color schemes is “color by entropy” based on diversity in the alignment added below.

Selected region gets highlighted in structure

**Please let us know if you have
questions, comments or
suggestions**

seq-info@lanl.gov

Bette Korber: btk@lanl.gov

Will Fischer: wfischer@lanl.gov